UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

MODÈLES D'APPRENTISSAGE PROFOND ADAPTATIFS ET GÉNÉRALISABLES POUR LA SEGMENTATION D'IMAGES MÉDICALES

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN SCIENCES ET TECHNOLOGIES DE L'INFORMATION

PAR

MOHAMED LAMINE ALLAOUI

MARS 2026

Cette thèse a été évaluée par un jury composé des personnes suivantes :

Prof. Rezkallah, Miloud (UQO) ..... Président du jury

Prof. Carlos Vazquez (ETS) ..... Examinateur externe

Prof. Ana-Maria Cretu (UQO) ..... Examinateur interne

Prof. Mohand Said Allili (UQO) ..... Directeur de recherche

Thèse acceptée le : 2026

Remerciements

Avant tout, je rends grâce à ALLAH, qui m'a permis d'accomplir ce travail et qui m'a accompagné à chaque étape de ma vie. Sans Ses bénédictions, Sa guidance divine et Sa miséricorde infinie, rien de tout cela n'aurait été possible.

Je tiens à exprimer ma plus profonde gratitude à mon directeur de thèse, le Professeur Mohand Said Allili, pour son soutien constant, sa confiance et son accompagnement généreux tout au long de ce parcours de recherche. Sa disponibilité, ses commentaires constructifs et ses conseils éclairés ont été inestimables pour façonner cette thèse et m'aider à grandir, tant sur le plan académique que personnel. Je lui suis profondément reconnaissant pour l'opportunité de travailler sous sa supervision et pour les connaissances et l'expérience acquises grâce à son mentorat.

J'adresse également mes sincères remerciements à l'Université du Québec en Outaouais (UQO) pour m'avoir offert l'opportunité et les conditions favorables à la réalisation de mes études doctorales dans un environnement stimulant et bienveillant.

Ma plus profonde reconnaissance va aux membres du jury, qui ont aimablement accepté d'évaluer cette thèse. Je vous remercie pour le temps que vous y avez consacré, pour vos commentaires perspicaces et pour vos suggestions constructives, qui ont contribué à améliorer la qualité de ce travail.

À ma bien-aimée épouse, j'exprime toute ma gratitude et mon amour. Merci d'avoir été mon pilier, pour ta patience, tes encouragements et ton soutien inébranlable tout au long de ce long et exigeant parcours. Ta présence a été mon refuge dans les moments les plus difficiles, et ta confiance en moi m'a donné la force de persévérer.

À mes parents et à mes frères, je dois tout. Merci pour votre amour inconditionnel, pour avoir toujours cru en moi et pour m'avoir soutenu à chaque étape du chemin. Vos sacrifices, vos valeurs et vos encouragements sans fin ont été la source de ma force et de ma détermination. Cette réussite est autant la vôtre que la mienne.

Je remercie également mes amis et collègues pour leur amitié, leur soutien moral et les discussions enrichissantes que nous avons partagées. Vous avez rendu ce parcours plus humain, plus joyeux et inoubliable.

Enfin, j'exprime ma gratitude à toutes celles et ceux qui, de près ou de loin, ont contribué à l'achèvement de ce travail.

Table des matières

Liste des figures **6**

Liste des tableaux **11**

Liste des abréviations, sigles et acronymes **14**

Résumé **16**

1 Introduction **20**
- 1.1 Problématique et Contexte de Recherche . . . 20
- 1.2 Conventions Terminologiques et Définition du Domaine d'Application . . 21
- 1.3 Objectifs de la Thèse . . . 22
  - 1.3.1 Objectif de Recherche Principal . . . 22
  - 1.3.2 Objectif Secondaire . . . 23
- 1.4 Contributions . . . 24
- 1.5 Publications . . . 25
- 1.6 Organisation de la Thèse . . . 26

2 Concepts et Fondements Théoriques **27**
- 2.1 Introduction . . . 27
- 2.2 Fondements de l'Apprentissage Profond . . . 28
  - 2.2.1 Réseaux de Neurones Artificiels : Fondements et Principes . . . 28
  - 2.2.2 Extraction de Caractéristiques Hiérarchiques . . . 29
  - 2.2.3 Apprentissage Profond Versus Apprentissage Automatique Traditionnel : Une Analyse Comparative . . . 30
  - 2.2.4 Apprentissage Profond dans l'Analyse de Données Visuelles : Applications Médicales . . . 31

2.3 Réseaux de Neurones Convolutionnels (CNNs) et Extraction de Caractéristiques . . . 32
2.3.1 Introduction aux CNNs : Pourquoi Ils Sont Essentiels pour les Tâches de Vision . . . 32
2.3.2 Architecture CNN et Composants Centraux . . . 32
2.3.3 Forces et Limitations des CNNs en Imagerie Médicale . . . 41
2.4 Architectures Modernes d'Apprentissage Profond pour les Tâches de Vision 42
2.4.1 Évolution au-delà des CNNs . . . 42
2.4.2 Vision Transformers (ViTs) . . . 42
2.4.3 Modèles State-Space (Mamba) . . . 44
2.4.4 Mécanismes d'Attention dans l'Apprentissage Profond . . . 46
2.4.5 Auto-encodeurs et Apprentissage de Représentations . . . 49
2.4.6 Défis et Considérations dans l'Apprentissage Profond pour l'Analyse d'Images Médicales . . . 51
2.4.7 Conclusion . . . 52

3 Revue de Littérature en Segmentation d'Images Médicales 54
3.1 Vue d'Ensemble . . . 54
3.2 Méthodes Traditionnelles . . . 54
3.3 Architectures Basées sur CNN . . . 58
3.4 Architectures Basées sur Transformer . . . 59
3.5 Architectures Hybrides . . . 61
3.6 Modèles State-Space (Vision Mamba) . . . 62
3.7 Méthodes Basées sur Prompts et Universelles . . . 64
3.8 Analyse Comparative . . . 66
3.9 Conclusion . . . 68

4 Mixture of Experts pour la Segmentation de Lésions Cutanées 69
4.1 Introduction aux Défis de la Segmentation de Lésions Cutanées . . . 69
4.2 Vue d'Ensemble du Modèle MEDiXNet . . . 71
4.2.1 Motivation et Principes de Conception . . . 72
4.2.2 Composants Clés de MEDiXNet . . . 73
4.3 Résultats et Analyse . . . 78
4.3.1 Ensembles de Données et Configuration Expérimentale . . . 79
4.3.2 Évaluation Quantitative des Performances . . . 80

4.3.3 Analyse Qualitative et Discussion 81
4.4 Limitations Actuelles et Directions Futures de MEDiXNet 83

5 MixLVMM : Un Mélange de Modèles Vision Mamba Légers pour une Segmentation Robuste de Lésions Cutanées 85
5.1 Introduction et Motivation 85
5.2 Composants du Modèle et Méthodologie 87
5.2.1 Architecture Expert MixLVMM 87
5.2.2 Le Gate Network 89
5.2.3 Adaptive Salient Region Attention Module (ASRAM) 91
5.2.4 Stratégie de Pré-entraînement et Transfer Learning 91
5.3 Protocole Expérimental 92
5.3.1 Ensembles de Données 92
5.3.2 Paramètres d'Entraînement et d'Évaluation 93
5.3.3 Métriques d'Évaluation 94
5.3.4 Résultats et Discussion 95
5.3.5 Études d'Ablation 100
5.4 Conclusion 105

6 HA-U^{3}Net : Un Framework Agnostique aux Modalités pour la Segmentation d'Images Médicales 3D Utilisant une Structure V-Net Imbriquée et Attention Hybride 107
6.1 Introduction 107
6.1.1 Motivation et Contexte 107
6.1.2 Défis de Segmentation Tridimensionnelle 108
6.1.3 Lacune de Recherche et Limitations des Méthodes Existantes . . . 109
6.1.4 Objectifs du Chapitre 110
6.1.5 Contributions Clés 110
6.2 Méthodologie 111
6.2.1 Architecture U^3-Net Imbriquée 111
6.2.2 Variante U^{3}Mamba 116
6.3 Protocole Expérimental et Implémentation 119
6.3.1 Ensembles de Données et Configuration d'Évaluation 119
6.3.2 Détails d'Implémentation 121
6.4 Résultats et Analyse 124

6.4.1 Études d'Ablation 124
6.4.2 Résultats Quantitatifs 128
6.4.3 Résultats Qualitatifs 132
6.5 Conclusion 136

7 TD-DIMB : Text-Driven Dense Inverted Mamba Bottlenecks pour la Segmentation d'Images Médicales Interactive 138
7.1 Introduction 138
7.1.1 Motivation et Contexte de Recherche 138
7.1.2 Défis d'Intégration Cross-Modale en Imagerie Médicale 139
7.1.3 Limitations des Méthodes Actuelles Basées sur Prompts 139
7.1.4 Objectifs de Recherche et Portée 140
7.1.5 Contributions Clés et Innovations 140
7.2 Méthodologie et Conception Architecturale 141
7.2.1 Vue d'Ensemble du Framework TD-DIMB 141
7.2.2 Modules Dense Inverted Mamba Bottleneck (DIMB) 144
7.2.3 Mécanisme Text-Driven Selective Scan 2D (TD-SS2D) 146
7.2.4 Formulation de la Reinforced Gaussian Dice Loss (RGDL) 147
7.3 Conception Expérimentale et Implémentation 149
7.3.1 Protocole d'Évaluation et Ensembles de Données 149
7.3.2 Détails d'Implémentation et Protocoles d'Entraînement 152
7.4 Résultats et Analyse Globale 153
7.4.1 Études d'Ablation et Analyse de Composants 153
7.4.2 Résultats Quantitatifs 157
7.4.3 Évaluation Qualitative et Interprétation Clinique 158
7.4.4 Efficacité de Calcul et Viabilité Clinique 160
7.5 Discussion et Analyse Critique 162
7.5.1 Limitations et Contraintes 162
7.5.2 Directions de Recherche Futures 163
7.6 Résumé du Chapitre et Conclusion 163

8 FUSE-RAG : Few-shot Universal Segmentation avec Retrieval-Augmented Generation pour l'Imagerie Médicale 165
8.1 Introduction 165
8.1.1 Motivation et Contexte de Recherche 165

8.1.2 Le Défi de la Segmentation Universelle . . . 166
8.1.3 Retrieval-Augmented Generation pour l'Imagerie Médicale . . . 167
8.1.4 Objectifs de Recherche et Contributions . . . 167
8.1.5 Innovations Clés et Impact de Recherche . . . 168
8.2 Méthodologie . . . 169
8.2.1 Formulation du Problème et Vue d'Ensemble du Framework . . . 169
8.2.2 Mécanisme de Récupération ROI-Aware . . . 170
8.2.3 Réseau de Segmentation Retrieval-Conditioned . . . 173
8.2.4 Protocole d'Entraînement et Formulation de Perte . . . 177
8.3 Évaluation Expérimentale . . . 179
8.3.1 Configuration Expérimentale . . . 180
8.3.2 Études d'Ablation . . . 180
8.4 Résultats et Discussion . . . 184
8.4.1 Analyse Quantitative . . . 184
8.4.2 Analyse Qualitative . . . 187
8.4.3 Analyse de Signification Statistique . . . 188
8.5 Discussion et Analyse Critique . . . 189
8.5.1 Limitations et Contraintes . . . 190
8.5.2 Directions de Recherche Futures . . . 190
8.6 Résumé de Chapitre et Conclusion . . . 191

9 Conclusion **192**
9.1 Recommandations et Perspectives . . . 194
9.1.1 Extension au Traitement Volumétrique Tridimensionnel Complet . 194
9.1.2 Détection Automatisée des Régions d'Intérêt . . . 195
9.1.3 Segmentation Multiclasse et Multi-organe Unifiée . . . 195
9.1.4 Quantification de l'Incertitude et Fiabilité Clinique . . . 195
9.1.5 Robustesse Inter-domaine et Adaptation aux Domaines Éloignés . 196
9.1.6 Validation Clinique et Déploiement en Environnement Réel . . . 196
9.1.7 Apprentissage Fédéré et Confidentialité des Données . . . 196

Bibliographie **197**

Liste des figures

2.1 Structure d'un perceptron multicouche (MLP), illustrant les couches et la connectivité des neurones. 29
2.2 Illustration de l'extraction de caractéristiques hiérarchiques dans les réseaux de neurones profonds. 30
2.3 Fonctions d'activation les plus populaires utilisées dans les architectures CNN. 36
2.4 Visualisation détaillée de l'opération de max pooling 2×2. Gauche : Image en niveaux de gris originale avec région surlignée. Centre : Vue agrandie de la région sélectionnée 8×8 pixels avec superposition de grille montrant les valeurs de pixels originaux. Droite : Résultat après application du max pooling 2×2, où chaque cellule contient la valeur maximale de sa région 2×2 correspondante dans l'image originale. 37
2.5 Comparaison du max pooling et average pooling avec différentes tailles de kernel. Rangée supérieure : Image originale (gauche), max pooling avec kernel 2×2 (centre), et average pooling avec kernel 2×2 (droite). Rangée médiane : Image originale (gauche), max pooling avec kernel 3×3 (centre), et average pooling avec kernel 3×3 (droite). Rangée inférieure : Image originale (gauche), max pooling avec kernel 4×4 (centre), et average pooling avec kernel 4×4 (droite). Notez comment les tailles de kernel plus grandes réduisent les dimensions spatiales plus agressivement tandis que le max pooling préserve mieux les caractéristiques importantes que l'average pooling. 38
2.6 L'architecture du modèle Transformer [1]. 44

2.7 Aperçu du Selective State Space Model. Le modèle transforme les séquences d'entrée x_t en sorties y_t par des transitions d'états latents, avec une sélection de paramètres dépendante de l'entrée permettant un traitement efficace de l'information [2]. 46
2.8 L'architecture de Vision Mamba démontrant le traitement bidirectionnel et les embeddings positionnels pour une modélisation efficace du contexte visuel [3]. 46
2.9 Structure générale d'un Auto-encodeur illustrant le paradigme encodeur-décodeur. 51

3.1 Vue d'ensemble taxonomique des méthodologies de segmentation d'images médicales. 55

4.1 Présentation pictoriale typique des images cutanées dans l'ensemble de données de test ISIC-2017 avec différentes images difficiles pour la segmentation. 69
4.2 Exemples de lésions cutanées avec apparences contrastées. La rangée supérieure illustre des lésions foncées avec haut contraste et frontières clairement définies, tandis que la rangée inférieure représente des lésions claires caractérisées par un faible contraste et des frontières indistinctes. 70
4.3 Workflow MEDiXNet illustrant l'interaction entre les Réseaux Experts, le Gate Network, et le module ASRAM. 72
4.4 Architecture MEDiXNet illustrant les réseaux experts spécialisés, le Gate Network, et l'intégration du module ASRAM. 74
4.5 Architecture détaillée de l'Adaptive Salient Region Attention Module (ASRAM), illustrant l'attention spatiale via les convolutions dilatées multi-échelles (SSAM), l'attention par canal par le pooling global et MLP (SCAM), et l'intégration d'attention guidée par saillance. 77
4.6 Comparaison qualitative des masques de segmentation sur des images représentatives de l'ensemble de données ISIC. MEDiXNet atteint une précision de segmentation supérieure comparée aux méthodes de pointe. . 82

4.7 Comparaison détaillée des frontières des résultats de segmentation sur des exemples de lésions difficiles. Les lignes vertes indiquent les frontières de vérité terrain, tandis que les lignes rouges montrent les prédictions de chaque méthode. MEDiXNet démontre une précision exceptionnelle en délimitation des frontières. 83

5.1 Architecture Expert Vision Mamba. Chaque expert intègre Patch Embedding, blocs FVM, Patch Merging/Expanding, ASRAM, et connexions skip. 89
5.2 Architecture Gate Network pour le routage basé sur la similarité utilisant triplet loss. 90
5.3 Architecture Adaptive Salient Region Attention Module (ASRAM). . . . 91
5.4 Comparaison de performance avant et après transfer learning pour deux réseaux experts sur les catégories de lésions à tons foncés et clairs. E1 désigne l'Expert 1 (spécialisé pour les lésions foncées) et E2 désigne l'Expert 2 (spécialisé pour les lésions claires). La notation E1 Dark représente la performance de l'Expert 1 évaluée sur l'ensemble de test de lésions foncées, tandis que E1 Light représente la performance de ce même expert évaluée sur l'ensemble de test de lésions claires. De manière similaire, E2 Dark et E2 Light représentent respectivement les performances de l'Expert 2 sur les ensembles de test de lésions foncées et claires. Les résultats montrent que le transfer learning améliore la généralisation de chaque expert sur les deux catégories de lésions. 92
5.5 Comparaison des Scores Dice sur tous les ensembles de données. 98
5.6 Comparaison des résultats qualitatifs sur l'ensemble de test ISIC 2018. . 99
5.7 Résultats d'apprentissage few-shot (ISIC 2017). 100
5.8 Catégorisation des données d'entraînement dans les ensembles de données, excepté les ensembles de données PH2 et DermQuest, qui sont dédiés exclusivement pour l'inférence. 102
5.9 Visualisation box plot du Dice Similarity Coefficient (DSC) pour différents nombres d'experts sur les ensembles de données ISIC 2017 et ISIC 2018. . 103

6.1 L'architecture en forme de U^3 de HA-U^3Net conçue pour une extraction de caractéristiques multi-échelles efficace. La structure encodeur-décodeur exploite la représentation de caractéristiques hiérarchiques avec des connexions skip pour retenir les détails spatiaux dans les résolutions. Une supervision profonde est incorporée à de multiples étapes de décodeur, améliorant le flux de gradient durant l'entraînement et renforçant la précision de segmentation. 112
6.2 Architecture de bloc Hybrid Attention (HA), combinant l'attention de canal et spatiale pour améliorer la représentation de caractéristiques pour l'imagerie médicale 3D. 115
6.3 Bloc Tri-orientated Spatial Mamba dans le module U^3. 117
6.4 Analyse de distribution des structures pathologiques dans les ensembles de données de différentes modalités, démontrant l'hétérogénéité des présentations tumorales dans les modalités d'imagerie. (A) BraTS : distribution volumétrique des sous-régions tumorales cérébrales (ET, TC, WT) dans les scans IRM. (B) AutoPET : caractéristiques de compte et volume des lésions dans FDG-PET/CT corps entier. (C) ABUS : fréquence de volume tumoral en ultrasons 3D du sein. 121
6.5 Distribution de Fréquence d'Organe par Système Anatomique dans l'ensemble de données TotalSegmentator. 122
6.6 Impact de la profondeur des niveaux imbriqués (L) sur la performance de HA-U^3Net sur tous les ensembles de données. 127
6.7 Comparaison qualitative des résultats de segmentation sur les ensembles de données ultrasons. Les deux rangées supérieures correspondent à ABUS, illustrant la performance intra-ensemble. Les deux rangées inférieures montrent les résultats sur BUSI, démontrant la généralisation inter-ensembles en inférence zero-shot. 132
6.8 Comparaison qualitative des résultats de segmentation sur BraTS. La figure affiche les slices IRM dans la modalité T1Gd, illustrant les régions tumorales dans divers cas. 133
6.9 Comparaison de résultats qualitatifs dans Total Segmentator. 134
6.10 Comparaison de résultats qualitatifs dans AutoPET. 135

6.11 Visualisation de cartes de caractéristiques sur diverses étapes du réseau, comparant SRB (haut), U^3-Block (milieu), et U^3-Block avec Hybrid Attention (bas). De gauche à droite : image d'entrée, caractéristiques initiales, caractéristiques mid-encoder, caractéristiques bottleneck, caractéristiques early decoder, caractéristiques mid-decoder, et caractéristiques finales. 136

7.1 Architecture Text-Driven Dense Inverted Mamba Bottleneck Network (TD-DIMB). 142
7.2 Conception du bloc *Text-Driven Selective Scan 2D* (TD-SS2D). 148
7.3 Extraction de caractéristiques guidée par *prompt* dans TD-DIMB. 155
7.4 Comparaison qualitative des performances de segmentation sur différents ensembles de données. 159

8.1 Conception du mécanisme de récupération *ROI-aware*. Haut : le processus d'indexation des données crée la base de connaissances à travers l'extraction d'*embeddings* enrichis par les régions d'intérêt (ROI) et leur indexation via FAISS. Bas : le processus d'inférence emploie le *retrieval-augmented generation* pour améliorer la segmentation en interrogeant la base de connaissances et en sélectionnant des exemples de support anatomiquement pertinents pour le réseau de segmentation. 170
8.2 Conception architecturale du réseau de segmentation FUSE-RAG. 174
8.3 Architecture du Bloc d'Évaluation de la Qualité du Support (SQAB). . . 176
8.4 Comparaison qualitative de qualité de récupération d'ensemble de support. 182
8.5 Résultats de segmentation qualitative sur l'ensemble de données de lésions d'AVC ATLAS 2.0 et l'ensemble de données de pneumonie QaTa-COVID19. 187

Liste des tableaux

2.1 Comparaison entre les Méthodes d'Apprentissage Automatique Traditionnelles et l'Apprentissage Profond 31

4.1 Comparaison de différentes méthodes sur les ensembles de données ISIC. 81

5.1 Comparaison de différentes méthodes sur l'ensemble de données ISIC 2017. 95
5.2 Comparaison de différentes méthodes sur l'ensemble de données ISIC 2018. 96
5.3 Comparaison de différentes méthodes sur l'ensemble de données PH2. . . 97
5.4 Comparaison de performance de MixLVMM avec différents mécanismes d'attention sur l'ensemble de données ISIC 2018. 101
5.5 Comparaison de performance de différentes configurations du modèle MixLVMM sur l'ensemble de données ISIC 2018. 103
5.6 Analyse de Signification Statistique de MixLVMM vs. Baselines les Plus Performants 105

6.1 Résultats de Validation Croisée 4-Fold à Travers les Ensembles de Données (Score Dice Moyen). 124
6.2 Impact des Blocs Hybrid Attention (HA) et Variantes de Modèle sur la Performance (DSC Moyen (%)). 125
6.3 Impact de la Supervision Profonde (DS) sur la Performance de HA-U^{3}Net (DSC Moyen (%)). 126
6.4 Comparaison avec Méthodes Pertinentes sur l'Ensemble de Données ABUS 3D. 128
6.5 Validation cross-dataset : modèles 2D entraînés sur slices ABUS et testés sur l'ensemble de données BUSI. 129
6.6 Comparaison de différentes méthodes sur l'ensemble de données BraTS. . 129

6.7 Comparaison de métriques moyennes sur les anatomies dans l'ensemble de données TotalSegmentator. 130
6.8 Comparaison avec Méthodes Pertinentes sur l'Ensemble de Données AutoPET. 131
6.9 Comparaison d'efficacité de calcul sur volumes d'entrée 128×128×128. . . 131

7.1 Spécifications d'Ensembles de Données d'Entraînement Universel 151
7.2 Étude d'ablation component-wise. Dice (%) ↑ et HD95 (pixels) ↓. 154
7.3 Effet des Stratégies de *Prompting* avec Guidance Multi-*Modal* de MedSigLIP. 155
7.4 Comparaison de différents backbones de prompting avec architecture TD-DIMB. 156
7.5 Comparaison de fonction de perte sur TD-DIMB. Dice (%) ↑ et HD95 (px) ↓. 157
7.6 Quantitative comparison of TD-DIMB and state-of-the-art models on both task-specific and universal generalization settings using CAMUS, autoPET22, ATLAS, and QaTa-COV19 datasets. 158
7.7 Comparaison de complexité de calcul et d'efficacité des méthodes de segmentation sur images d'entrée 256×256 avec entrainement universel. . . . 160
7.8 Résultats de validation croisée four-fold pour TD-DIMB sur évaluation task-specific. 161
7.9 Analyse de signification statistique des améliorations TD-DIMB avec valeurs p et intervalles de confiance 95%. 161

8.1 Étude d'ablation du système de récupération démontrant l'impact de différents backbones d'extraction de caractéristiques sur la performance de segmentation d'images médicales *few-shot* avec $K = 4$ exemples de support. Résultats rapportés comme moyenne ± écart-type à travers 5 *seeds* aléatoires. 181
8.2 Étude d'ablation d'architecture de réseau de segmentation démontrant les contributions progressives de composants à la performance de segmentation d'images médicales few-shot avec $K = 4$ exemples de support. Les composants sont ajoutés incrémentalement, et les résultats sont rapportés comme moyenne ± écart-type à travers 5 graines aléatoires. 182

8.3 Impact de taille d'ensemble de support et stratégie de sélection sur performance de segmentation d'images médicales few-shot, démontrant performance optimale à $K = 8$ exemples avec effet plateau au-delà de $K = 8$. La dégradation de performance à des ensembles de support plus larges valide l'importance de récupération ROI-aware haute qualité sur approches basées quantité, étendant les principes d'ingénierie de prompts de NLP aux tâches de vision médicale. Résultats rapportés comme moyenne ± écart-type à travers 5 seeds aléatoires. 184
8.4 Comparaison de performance de segmentation de FUSE-RAG contre les méthodes de l'état de l'art à travers diverses tâches d'imagerie médicale. Les résultats démontrent une précision supérieure à travers des applications d'imagerie neurologique et pulmonaire. Toutes les méthodes few-shot ont été évaluées avec $K = 4$ exemples de support. 185
8.5 Comparaison d'efficacité de calcul à travers toutes les méthodes évaluées. Mesures effectuées sur GPU NVIDIA RTX A6000 avec configuration matérielle identique et implémentations optimisées. Les évaluations de faisabilité clinique sont basées sur des seuils pratiques de vitesse d'inférence et d'usage mémoire. 185
8.6 Analyse de validation croisée K-fold de FUSE-RAG (Kf=5). La validation croisée a été effectuée sur les ensembles de test, en alternant les données de requête et de support à chaque itération. 188
8.7 Analyse de signification statistique des améliorations FUSE-RAG avec valeurs p et intervalles de confiance 95%. 189

9.1 Synthèse comparative des architectures proposées en termes de complexité computationnelle, performance et portée de généralisation. 193

Liste des abréviations, sigles et acronymes

ANN Artificial Neural Networks

ASRAM Adaptive Salient Region Attention Module

CNN Convolutional Neural Networks

CT Computed Tomography

ELU Exponential Linear Unit

FC Fully Connected Layers

FUSE-RAG Few-shot Universal Segmentation with Retrieval-Augmented Generation

GAP Global Average Pooling

GMP Global Max Pooling

GPU Graphics Processing Unit

HA Hybrid Attention

HA-U^{3}Net Hybrid Attention U^3-Net

HD95 95th Percentile Hausdorff Distance

IT Inference Time

MEDiXNet Mixture of Expert Dermatological Imaging Networks

MixLVMM Mixture of Lightweight Vision Mamba Models

MLP MultiLayer Perceptrons

MoE Mixture of Experts

MRI Magnetic Resonance Imaging

NLP Natural Language Processing

RAG Retrieval-Augmented Generation

ReLU Rectified Linear Units

ROI Region of Interest

SCAM Salient Channel Attention Module

SGD Stochastic Gradient Descent

SSAM Separable Spatial Attention Module

SSM State Space Models

SVM Support Vector Machines

tanh hyperbolic tangent

TD-DIMB Text-Driven Dense Inverted Mamba Bottlenecks

TT Training Time

Vim Vision Mamba

ViT Vision Transformers

Résumé

La segmentation d'images médicales fait face à des défis majeurs en raison de la grande variabilité entre les modalités d'imagerie, les structures anatomiques, les présentations pathologiques et les protocoles d'acquisition, ce qui limite considérablement la capacité de généralisation des approches automatisées traditionnelles. Cette thèse présente une investigation systématique d'architectures d'apprentissage profond adaptatives et généralisables, visant à gérer cette variabilité à des échelles progressivement croissantes. Elle établit un cadre méthodologique complet en quatre étapes, allant de la spécialisation intra-domaine à des capacités d'adaptation universelles.

La recherche propose cinq innovations architecturales majeures abordant la gestion de la variabilité selon une progression structurée. **MEDiXNet** (*Mixture of Expert Dermatological Imaging Networks*) introduit des architectures à *mixture-of-experts* avec routage dynamique et modules d'attention visuelle, traitant spécifiquement la variabilité intra-domaine en imagerie dermoscopique par une gestion spécialisée des différentes présentations de lésions. **MixLVMM** (*Mixture of Lightweight Vision Mamba Models*) fait évoluer cette approche en intégrant des architectures Vision Mamba avec routage par *triplet-loss* et génération automatique d'ancres, atteignant des performances équivalentes avec une réduction significative du nombre de paramètres. **HA-U^{3}Net** (*Hybrid Attention U^3-Net*) étend ces capacités à la variabilité inter-modalités tridimensionnelle grâce à des blocs U^3 imbriqués et des mécanismes d'attention hybrides, démontrant une généralisation robuste à travers les modalités IRM, CT, échographie et PET.

TD-DIMB (*Text-Driven Dense Inverted Mamba Bottlenecks*) progresse vers la gestion de la variabilité sémantique en intégrant des prompts en langage naturel avec des modèles fondamentaux médicaux à l'aide des modules *Dense Inverted Mamba Bottleneck* et des mécanismes *Text-Driven Selective Scan 2D*, permettant une adaptation dynamique des tâches via une optimisation cliniquement informée. **FUSE-RAG** (*Few-shot Universal Segmentation with Retrieval-Augmented Generation*) en constitue l'aboutissement,

atteignant une adaptation universelle grâce à un mécanisme de *génération augmentée par récupération (RAG)* spécifiquement conçu pour l'imagerie médicale. Ce framework intègre des mécanismes de récupération *ROI-aware* qui injectent la connaissance anatomique experte dans les représentations des modèles fondamentaux, démontrant des améliorations substantielles de 10,26 % et 8,86 % du coefficient de Dice sur la segmentation de lésions d'AVC et de pneumonies dans des domaines anatomiques entièrement nouveaux.

La validation expérimentale approfondie établit plusieurs principes clés pour la conception de modèles d'apprentissage profond véritablement adaptatifs et généralisables : une adaptabilité architecturale intrinsèquement pensée pour les données médicales, une généralisation inter-modalités via des mécanismes d'attention sophistiqués et l'intégration de modèles fondamentaux, un conditionnement sémantique reliant le traitement automatisé au raisonnement clinique, et une récupération intelligente de connaissances favorisant une sélection qualitative des exemples de support.

Ce travail établit un nouveau paradigme qui dépasse le compromis traditionnel entre approches spécialisées et universelles, démontrant qu'une gestion systématique de la variabilité permet à des systèmes universels d'atteindre des niveaux de performance proches de ceux des modèles spécialisés, tout en conservant leur adaptabilité face à de nouveaux scénarios cliniques. Il constitue ainsi une base solide pour les systèmes d'IA médicale de nouvelle génération, à la fois adaptatifs aux contextes cliniques variés et généralisables à travers les modalités d'imagerie, favorisant une amélioration du diagnostic et une optimisation des soins aux patients.

Abstract

Medical image segmentation faces major challenges due to the extensive variability across imaging modalities, anatomical structures, pathological presentations, and acquisition protocols, which severely limits the generalization capability of traditional automated approaches. This thesis presents a systematic investigation of adaptive and generalizable deep learning architectures designed to handle such variability at progressively increasing scales. It establishes a comprehensive four-stage methodological framework that progresses from intra-domain specialization to universal few-shot adaptation capabilities.

The research introduces five major architectural innovations that address variability management through a structured progression. **MEDiXNet** (*Mixture of Expert Dermatological Imaging Networks*) introduces *mixture-of-experts* architectures with dynamic routing and visual attention modules, specifically addressing intra-domain variability in dermoscopic imaging through specialized handling of diverse lesion presentations. **MixLVMM** (*Mixture of Lightweight Vision Mamba Models*) extends this approach by integrating Vision Mamba architectures with triplet-loss-based routing and automatic anchor generation, achieving comparable performance with a significant reduction in parameters. **HA-U^{3}Net** (*Hybrid Attention U^3-Net*) expands these capabilities to three-dimensional inter-modality variability using nested U^3-blocks and hybrid attention mechanisms, demonstrating robust generalization across MRI, CT, ultrasound, and PET modalities.

TD-DIMB (*Text-Driven Dense Inverted Mamba Bottlenecks*) advances toward semantic variability management by integrating natural language prompts with medical foundation models through *Dense Inverted Mamba Bottleneck* modules and *Text-Driven Selective Scan 2D* mechanisms, enabling dynamic task adaptation through clinically informed optimization. **FUSE-RAG** (*Few-shot Universal Segmentation with Retrieval-Augmented Generation*) represents the culmination of this progression, achieving uni-

versal adaptation through a *Retrieval-Augmented Generation (RAG)* mechanism specifically designed for medical imaging. This framework integrates *ROI-aware* retrieval mechanisms that embed expert anatomical knowledge into foundation model representations, demonstrating substantial Dice coefficient improvements of 10.26 % and 8.86 % in stroke lesion and pneumonia segmentation across entirely unseen anatomical domains.

Comprehensive experimental validation establishes several key principles for designing truly adaptive and generalizable deep learning models : intrinsic architectural adaptability tailored to medical data characteristics, cross-modality generalization through sophisticated attention mechanisms and foundation model integration, semantic conditioning bridging automated processing with clinical reasoning, and intelligent knowledge retrieval enabling quality-driven support selection.

This work establishes a new paradigm that transcends the traditional trade-off between specialized and universal approaches, demonstrating that systematic variability management enables universal systems to achieve performance levels approaching those of task-specific models while preserving adaptability to novel clinical scenarios. It therefore provides a solid foundation for the next generation of medical AI systems—both adaptive to diverse clinical contexts and generalizable across imaging modalities—advancing diagnostic accuracy and improving patient care.

Chapitre 1

Introduction

1.1 Problématique et Contexte de Recherche

L'imagerie médicale est devenue indispensable dans les soins de santé modernes, améliorant considérablement la précision diagnostique, la planification thérapeutique et la prise en charge des patients dans un large éventail de pathologies. Malgré ces avancées, le processus de délimitation précise des structures anatomiques et pathologiques, connu sous le nom de segmentation d'images médicales, demeure un goulot d'étranglement critique dans les flux de travail cliniques. Le défi fondamental ne réside pas dans la complexité technique des tâches individuelles de segmentation, mais dans la variabilité extraordinaire qui caractérise les données d'imagerie médicale parmi les populations de patients, les structures anatomiques, les présentations pathologiques, les protocoles d'acquisition et les technologies d'imagerie.

Cette variabilité se manifeste à de multiples échelles interconnectées, créant une barrière redoutable au développement de systèmes de segmentation robustes et généralisables. Au niveau intra-domaine, les images au sein de la même catégorie pathologique présentent une hétérogénéité substantielle : les lésions cutanées varient considérablement en pigmentation, texture et morphologie ; les tumeurs cérébrales présentent des formes, tailles et caractéristiques d'intensité diverses ; les structures cardiaques démontrent des variations anatomiques significatives parmi les populations de patients. Au niveau inter-domaine, différentes régions anatomiques et conditions pathologiques exigent des stratégies de segmentation entièrement distinctes, tandis que les variations inter-modalités introduisent une complexité additionnelle car les mêmes structures anatomiques apparaissent fondamentalement différentes lorsqu'elles sont captées par diverses technologies d'imagerie.

La segmentation manuelle, traditionnellement réalisée par des experts cliniques, s'avère inadaptée pour répondre à ce défi de variabilité. Au-delà d'être laborieuse et chronophage, les approches manuelles souffrent d'une variabilité inter-observateur élevée qui amplifie l'hétérogénéité inhérente des données, limitant la fiabilité diagnostique et la reproductibilité. De plus, le volume considérable et la diversité des données d'imagerie

médicale générées dans les systèmes de soins de santé modernes dépassent largement la capacité des approches d'annotation manuelle, nécessitant des solutions automatisées capables de s'adapter à une variabilité sans précédent sans supervision humaine extensive.

Les avancées récentes en apprentissage profond ont démontré des perspectives prometteuses pour automatiser la segmentation d'images médicales, pourtant les approches existantes échouent fondamentalement à traiter le défi de variabilité de manière systématique. Les méthodologies actuelles adoptent typiquement des solutions spécifiques à la tâche, requérant un développement de modèle, un entraînement et un déploiement indépendants pour chaque région anatomique, condition pathologique ou protocole d'imagerie. Cette approche fragmentée aboutit à une prolifération de modèles spécialisés qui manquent d'adaptabilité inter-domaines, exigeant une collecte de données extensive, une annotation experte et des ressources de calcul pour chaque nouvelle application. La complexité computationnelle des architectures existantes exacerbe ce défi en créant des barrières au déploiement généralisé, particulièrement lorsque de multiples modèles spécialisés doivent être maintenus simultanément.

Le défi central traité dans cette thèse est le développement d'architectures d'apprentissage profond adaptatives capables de gérer systématiquement la variabilité extraordinaire inhérente à la segmentation d'images médicales tout en maintenant une efficacité de calcul appropriée pour le déploiement clinique. Plutôt que de développer encore une autre collection de solutions spécifiques à la tâche, cette recherche poursuit un paradigme fondamentalement différent : créer des frameworks unifiés capables de s'adapter à divers scénarios d'imagerie médicale avec une supervision minimale, des exigences d'annotation réduites et une utilisation efficace des ressources. Cette approche traite l'écart critique entre le potentiel théorique de la segmentation d'images médicales automatisée et son implémentation pratique dans des environnements de soins de santé caractérisés par des ressources limitées, des populations de patients diverses et des besoins cliniques en constante évolution.

1.2 Conventions Terminologiques et Définition du Domaine d'Application

Pour assurer la précision et éliminer toute ambiguïté potentielle dans cette thèse, nous établissons des définitions claires pour les concepts terminologiques clés qui présentent des interprétations variées dans la littérature d'imagerie médicale. Le domaine de l'imagerie médicale souffre d'un usage terminologique incohérent, particulièrement concernant le terme « modalité », qui requiert une clarification explicite pour maintenir la rigueur scientifique et faciliter l'interprétation précise de nos contributions.

Dans cette thèse, le terme « modalités d'imagerie » fait référence exclusivement aux technologies d'imagerie médicale distinctes et aux techniques d'acquisition, incluant la tomodensitométrie (CT), l'imagerie par résonance magnétique (IRM), la tomographie

par émission de positons (PET), l'imagerie échographique et la radiographie conventionnelle (rayons X). Chacune représente une approche physique fondamentalement différente pour l'acquisition d'images médicales avec des caractéristiques distinctes concernant la résolution spatiale, les mécanismes de contraste et les capacités de différentiation tissulaire.

Au sein de l'imagerie IRM, nous distinguons entre différentes séquences d'impulsions et pondérations de contraste, spécifiquement les images pondérées T1, T2, T1 avec rehaussement de contraste (T1ce) et les images FLAIR (Fluid-Attenuated Inversion Recovery). Suivant la terminologie radiologique établie, nous nous référons à celles-ci comme « séquences de contraste » ou « contrastes IRM » plutôt que « modalités » pour éviter la confusion avec la catégorie plus large des modalités d'imagerie.

Dans le contexte de l'intelligence artificielle et des architectures d'apprentissage profond, « multi-modal » se réfère spécifiquement aux modèles qui traitent simultanément plusieurs types de données d'entrée, telles que la combinaison de descriptions textuelles avec des données d'images visuelles, comme illustré dans notre framework TD-DIMB. Le traitement inter-modal décrit la capacité des modèles à établir des correspondances et transférer la connaissance entre différents types d'entrée, tel qu'utiliser la compréhension sémantique textuelle pour guider l'extraction de caractéristiques visuelles.

Nos évaluations expérimentales sur « plusieurs modalités » se réfèrent spécifiquement à la généralisation entre les technologies d'imagerie CT, IRM, échographie et PET, tandis que les capacités « multi-modales » dans TD-DIMB et FUSE-RAG se réfèrent au traitement simultané d'entrées textuelles et visuelles. Ces conventions terminologiques demeurent cohérentes dans tous les chapitres, descriptions expérimentales et discussions techniques, assurant que les affirmations concernant l'adaptabilité universelle, la généralisation inter-modalités et le traitement multi-modal soient interprétées avec précision dans leurs contextes techniques intentionnels.

1.3 Objectifs de la Thèse

1.3.1 Objectif de Recherche Principal

L'objectif principal de cette thèse est de traiter systématiquement le défi de l'hétérogénéité et de la variabilité en segmentation d'images médicales à travers le développement d'architectures d'apprentissage profond progressivement adaptatives. Spécifiquement, nous visons à répondre à la question de recherche fondamentale : *Comment pouvons-nous concevoir des frameworks d'apprentissage profond adaptatifs qui gèrent systématiquement la variabilité en imagerie médicale, progressant des variations intra-domaine vers la généralisation universelle inter-modalités tout en maintenant la faisabilité de déploiement clinique ?*

Pour traiter cette question, cette thèse présente une approche compréhensive en quatre étapes qui étend progressivement la portée de la gestion de variabilité :

1. **Gestion de la Variabilité Intra-Domaine :** Nous commençons par traiter l'hétérogénéité au sein de domaines pathologiques spécifiques à travers des architectures basées sur des experts qui peuvent gérer des variations d'apparence distinctes, telles que les présentations des lésions à contraste élevé versus faible contraste en imagerie dermoscopique.
2. **Généralisation Inter-Modalités :** Nous étendons notre approche pour développer des architectures capables de maintenir des performances élevées entre différentes technologies d'imagerie (IRM, CT, échographie, PET) et régions anatomiques tout en préservant l'efficacité de calcul grâce à des structures imbriquées innovantes et des mécanismes d'attention hybrides.
3. **Adaptation Guidée Sémantiquement :** Nous avançons vers des capacités universelles en incorporant des mécanismes de conditionnement en langage naturel qui permettent l'adaptation à des tâches précédemment non vues à travers des guidances textuelles plutôt que le réentraînement architectural, exploitant des modèles fondamentaux médicaux pour une interprétabilité améliorée.
4. **Adaptation Universelle Few-Shot :** Nous atteignons une adaptabilité compréhensive à travers des frameworks augmentés par récupération qui peuvent rapidement s'adapter à de nouvelles structures anatomiques, pathologies et scénarios d'imagerie avec une supervision minimale grâce à une sélection d'exemples anatomiquement informée.

Cette progression systématique est réalisée à travers cinq contributions architecturales distinctes : MEDiXNet et MixLVMM traitent la variabilité intra-domaine en imagerie dermoscopique, HA-U^{3}Net permet la généralisation inter-modalités dans les données médicales volumétriques, TD-DIMB introduit le conditionnement sémantique à travers des prompts en langage naturel, et FUSE-RAG atteint l'adaptation universelle few-shot à travers la génération augmentée par récupération.

1.3.2 Objectif Secondaire

Bien que la performance de segmentation demeure notre priorité absolue, cette thèse poursuit un équilibre stratégique entre précision et efficacité computationnelle, reconnaissant que le déploiement clinique réaliste nécessite des architectures qui ne sacrifient pas la performance au profit de la légèreté, ni ne compromettent la praticité au profit de gains marginaux de précision. Plutôt que de maximiser uniquement la performance ou minimiser uniquement la complexité, toutes les architectures proposées recherchent le point d'équilibre optimal où des performances élevées sont atteintes avec une complexité modérée ou légère. Ceci implique le développement d'architectures innovantes qui réduisent substantiellement les exigences paramétriques et de calcul comparées aux approches existantes tout en maintenant ou améliorant la précision de segmentation. L'objectif n'est pas la légèreté pour elle-même, mais plutôt l'efficacité intelligente : atteindre les meilleures performances possibles avec des ressources computationnelles raisonnables appropriées pour le déploiement clinique à grande échelle. Cette approche

équilibrée reconnaît que l'impact clinique réel nécessite des modèles qui peuvent être déployés largement et maintenus efficacement, tout en délivrant la précision diagnostique nécessaire pour supporter les décisions cliniques critiques.

1.4 Contributions

Cette thèse présente cinq contributions substantielles accomplies au sein d'une investigation systématique des architectures d'apprentissage profond adaptatives pour gérer la variabilité en segmentation d'images médicales.

1. **MEDiXNet : Mixture of Experts pour la Segmentation de Lésions Cutanées** : Un modèle Mixture of Experts innovant traitant la dégradation de performance dans les lésions de couleur claire à travers des réseaux d'experts spécialisés et des mécanismes de routage dynamique. Le Module d'Attention de Région Saillante Adaptatif améliore la délimitation des frontières dans les régions à faible contraste, établissant un pont entre les méthodes de segmentation généralisées et la classification spécialisée des lésions.
2. **MixLVMM : Un Mélange de Modèles Vision Mamba Légers pour une Segmentation Robuste de Lésions Cutanées** : Une évolution de MEDiXNet remplaçant les experts basés sur CNN par des blocs Vision Mamba à efficacité paramétrique. Le framework incorpore un routage entraîné par triplet-loss et une génération automatique d'ancres via clustering UMAP, atteignant une précision supérieure avec une réduction paramétrique comparée aux approches traditionnelles.
3. **HA-U^{3}Net : Un Framework Agnostique aux Modalités pour la Segmentation d'Images Médicales 3D Utilisant une Structure V-Net Imbriquée et Attention Hybride** : Une architecture de blocs U^3 imbriqués innovante conçue pour l'extraction de caractéristiques profondes multi-échelles à travers un traitement hiérarchique à de multiples niveaux imbriqués. Le framework extrait des représentations de caractéristiques riches à chaque échelle de traitement, permettant des performances robustes parmi diverses conditions d'imagerie médicale et structures anatomiques allant des petites lésions aux grands organes. Les mécanismes d'attention hybrides optimisent le traitement de données 3D entre les modalités d'imagerie IRM, CT, échographie et PET, tandis que la variante légère U^{3}Mamba permet le déploiement dans des environnements à ressources contraintes.
4. **TD-DIMB : Text-Driven Dense Inverted Mamba Bottlenecks pour la Segmentation d'Images Médicales Interactive** : Une architecture intermodale innovante intégrant des prompts en langage naturel avec le traitement d'images médicales à travers des modules Dense Inverted Mamba Bottleneck qui atteignent une fusion de caractéristiques à complexité linéaire. Les mécanismes Text-Driven Selective Scan 2D permettent l'adaptation dynamique de tâches à

travers le conditionnement sémantique, tandis que l'intégration du modèle fondamental MedSigLIP fournit une compréhension robuste du domaine médical au-delà de la correspondance exacte de vocabulaire. La Reinforced Gaussian Dice Loss incorpore les priorités cliniques dans l'optimisation pour une précision de frontières améliorée.

5. **FUSE-RAG : Few-shot Universal Segmentation avec Retrieval-Augmented Generation pour l'Imagerie Médicale** : À notre connaissance, le premier framework de génération augmentée par récupération spécifiquement conçu pour la segmentation d'images médicales few-shot anatomiquement informée. Les mécanismes de récupération ROI-aware incorporent la connaissance anatomique experte dans les représentations de modèles fondamentaux, permettant une sélection de support privilégiant la qualité sur la quantité qui atteint des améliorations de coefficient de Dice de 10,26% et 8,86% sur la segmentation de lésions d'AVC et de pneumonie, respectivement, tout en démontrant l'efficacité des prompts visuels pour les applications d'imagerie médicale.

1.5 Publications

Voici la liste des publications réalisées jusqu'à présent dans le cadre de cette recherche :

1. M. L. Allaoui, M. S. Allili, "MEDiXNet : A Robust Mixture of Expert Dermatological Imaging Networks for Skin Lesion Segmentation", *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024. **Link**
2. **Collaboration** : A. Nouboukpo, M. L. Allaoui, and M. S. Allili, "MSCNet : Multi-scale spatial consistency for deep semi-supervised skin lesion segmentation", *Engineering Applications of Artificial Intelligence*, 2024. **Link**
3. M. L. Allaoui, M. S. Allili, "MixLVMM : A Mixture of Lightweight Vision Mamba Models for Enhancing Skin Lesion Segmentation Across High Tone Variability", *IEEE Access* , 2025. **Link**
4. M. L. Allaoui, M. S. Allili, A. Belaid, "HA-U^{3}Net : A Modality-Agnostic Framework for 3D Medical Image Segmentation Using Nested V-Net Structure and Hybrid Attention", *knowledge-based systems*, 2025. **Link**
5. M. L. Allaoui, M. S. Allili, "TD-DIMB : Text-Driven Dense Inverted Mamba Bottlenecks for Interactive Medical Image Segmentation", *Under Review*, 2026.
6. M. L. Allaoui, M. S. Allili, A. Belaid, "FUSE-RAG : Few-shot Universal Segmentation with Retrieval-Augmented Generation for Medical Imaging", *Under Review*, 2026.

1.6 Organisation de la Thèse

Le Chapitre 2 fournit les fondements théoriques essentiels pour comprendre les méthodologies d'apprentissage profond employées dans cette recherche. Le chapitre introduit les concepts clés des architectures de réseaux de neurones, se concentrant sur les réseaux de neurones convolutionnels, les mécanismes d'attention, les Vision Transformers et les modèles state-space qui servent de blocs de construction pour les contributions ultérieures.

Le Chapitre 3 présente une revue de littérature compréhensive examinant les avancées en segmentation d'images médicales. L'analyse évalue de manière critique les approches existantes parmi les méthodes traditionnelles, les architectures basées sur CNN, les modèles basés sur transformer, les approches hybrides et les modèles state-space, soulignant les forces, limitations et défis non traités qui motivent cette recherche.

Les contributions techniques commencent avec le Chapitre 4, qui introduit notre première contribution majeure traitant la variabilité intra-domaine dans les images dermoscopiques. S'appuyant sur cette fondation, le Chapitre 5 étend le paradigme basé sur des experts à travers des architectures Vision Mamba légères. La portée de recherche s'étend ensuite à l'imagerie médicale tridimensionnelle dans le Chapitre 6, qui traite les défis de généralisation inter-modalités entre diverses technologies d'imagerie. Progressant vers le conditionnement sémantique, le Chapitre 7 fait avancer notre framework en incorporant le guidage en langage naturel pour les tâches de segmentation interactive. La progression culmine avec le Chapitre 8, qui présente notre approche d'adaptation universelle few-shot à travers la génération augmentée par récupération.

La conclusion 9 synthétise les découvertes clés dans toutes les contributions, discute des implications plus larges pour la segmentation d'images médicales, analyse les limitations et esquisse les directions de recherche futures vers des solutions d'IA clinique pratiques et évolutives.

Chapitre 2

Concepts et Fondements Théoriques

2.1 Introduction

L'analyse d'images médicales a connu des avancées significatives stimulées par les percées en intelligence artificielle, particulièrement l'émergence et l'évolution rapide des techniques d'apprentissage profond. L'apprentissage profond, caractérisé par des réseaux de neurones hiérarchiques capables d'apprendre des motifs complexes directement à partir de données brutes, est devenu transformateur en soins de santé en améliorant substantiellement la précision diagnostique, en automatisant des tâches complexes et répétitives, et en permettant des applications de médecine de précision. Ce chapitre vise à fournir une présentation approfondie des fondements de l'apprentissage profond critique pour comprendre ses applications spécialisées en segmentation et analyse d'images médicales.

Contrairement aux méthodes conventionnelles, qui s'appuient extensivement sur des caractéristiques conçues manuellement par des experts, l'apprentissage profond automatise le processus d'extraction de caractéristiques, capturant des représentations hiérarchiques allant des textures et contours de base aux structures anatomiques et pathologiques complexes. Cette capacité a significativement fait progresser les applications cliniques telles que la détection de lésions, la délimitation tumorale et la segmentation multi-organes, soulignant le rôle central de l'apprentissage profond dans les diagnostics médicaux contemporains et la planification thérapeutique.

Ce chapitre introduira et définira méthodiquement la terminologie essentielle, les architectures et les mécanismes sous-jacents aux frameworks d'apprentissage profond modernes. Partant des concepts fondamentaux, incluant les réseaux de neurones artificiels et l'apprentissage de caractéristiques hiérarchiques, nous approfondirons progressivement les architectures avancées telles que les réseaux de neurones convolutionnels (CNNs), les Vision Transformers (ViTs), et les modèles hybrides intégrant des mécanismes d'attention. De plus, le chapitre discutera de l'apprentissage de représentations à travers les auto-encodeurs et traitera les défis pratiques rencontrés lors de l'intégration de l'apprentissage profond dans la pratique clinique. Collectivement, ces fondements théoriques

fournissent la compréhension nécessaire pour contextualiser et évaluer les contributions de recherche élaborées dans les chapitres ultérieurs.

2.2 Fondements de l'Apprentissage Profond

L'apprentissage profond a émergé comme l'une des avancées les plus significatives en intelligence artificielle, remodelant divers domaines incluant la vision par ordinateur, le traitement du langage naturel, et particulièrement l'analyse d'images médicales. Sa force distinctive réside dans l'apprentissage de représentations complexes directement à partir de données à grande échelle, éliminant ainsi le besoin d'extraction manuelle extensive de caractéristiques. Cette section fournit une exploration structurée et compréhensive des concepts fondamentaux sous-jacents à l'apprentissage profond, établissant les bases pour comprendre les architectures sophistiquées de réseaux de neurones discutées plus tard dans cette thèse.

2.2.1 Réseaux de Neurones Artificiels : Fondements et Principes

Les réseaux de neurones artificiels (ANNs) sont des modèles computationnels inspirés par la structure et la fonctionnalité des neurones biologiques. Les neurones biologiques intègrent les signaux d'entrée reçus d'autres neurones, déclenchant des sorties lors de l'atteinte d'un seuil. De manière analogue, les neurones artificiels combinent les entrées numériques au moyen de poids appris, sommant ces entrées et appliquant des fonctions d'activation non linéaires telles que sigmoïde, tangente hyperbolique (tanh), ou unités linéaires rectifiées (ReLU). Ces fonctions introduisent la non-linéarité cruciale pour modéliser des relations complexes au sein des données médicales [4].

Un perceptron, le modèle de réseau de neurones le plus simple, est restreint aux frontières de décision linéaires et donc limité à résoudre des problèmes linéairement séparables. Pour modéliser des relations plus complexes, de multiples perceptrons sont organisés en couches, formant des perceptrons multicouches (MLPs). Dans les MLPs, les sorties de chaque couche servent d'entrées aux couches ultérieures, permettant la transformation progressive des entrées brutes en représentations de plus en plus abstraites [5].

Le processus d'entraînement des ANNs implique l'optimisation des poids pour minimiser une fonction de perte prédéfinie, quantifiant les écarts entre les prédictions et les vraies étiquettes. Les algorithmes d'optimisation basés sur le gradient, notamment la descente de gradient stochastique (SGD) [6] et les méthodes adaptatives telles qu'Adam [7], ajustent itérativement les poids pour atteindre des performances prédictives optimales, particulièrement critiques dans les tâches médicales supervisées comme la segmentation et la classification d'images [5].

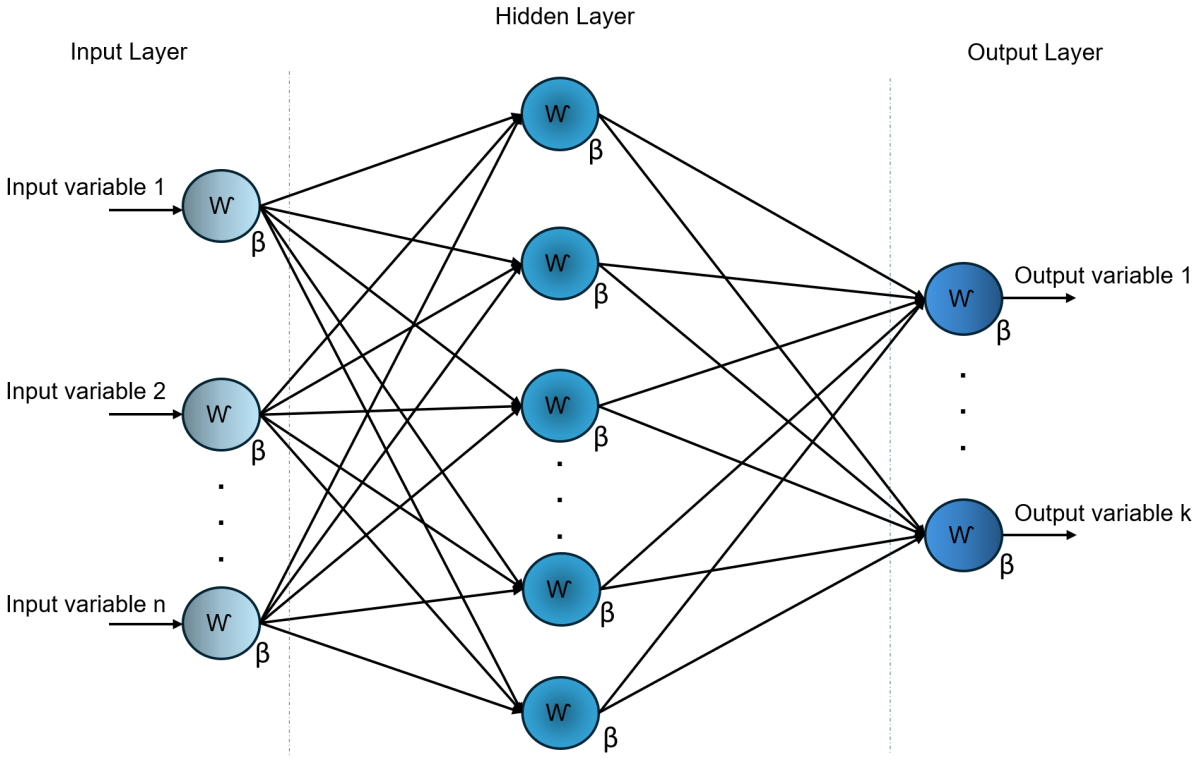

FIGURE 2.1 – Structure d'un perceptron multicouche (MLP), illustrant les couches et la connectivité des neurones.

2.2.2 Extraction de Caractéristiques Hiérarchiques

Une force centrale de l'apprentissage profond réside dans sa capacité d'extraction de caractéristiques hiérarchiques. Les techniques d'apprentissage automatique traditionnelles nécessitent typiquement une ingénierie de caractéristiques explicite, exigeant une expertise de domaine substantielle. À l'inverse, l'apprentissage profond apprend de manière autonome des représentations hiérarchiques à partir de données brutes, les couches inférieures détectant des caractéristiques visuelles élémentaires telles que les contours ou les textures, et les couches supérieures synthétisant celles-ci en structures complexes pertinentes pour des tâches spécifiques.

Par exemple, dans des tâches d'imagerie médicale telles que la segmentation tumorale à partir de scans IRM, les couches initiales du réseau peuvent identifier des motifs visuels fondamentaux tels que les gradients d'intensité et les variations de texture. Les couches intermédiaires interprètent ces caractéristiques de base en structures anatomiques, tandis que les couches plus profondes les intègrent en motifs cliniquement pertinents, permettant l'identification et la délimitation précises de régions pathologiques telles que les tumeurs [8].

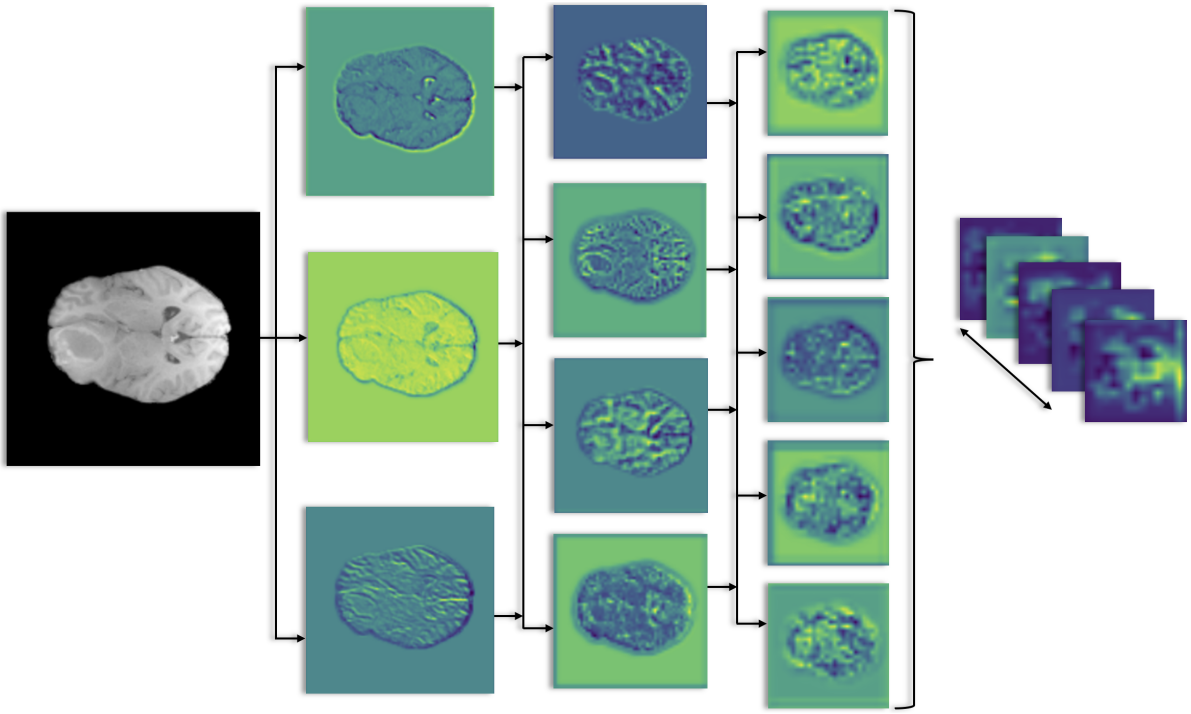

FIGURE 2.2 – Illustration de l'extraction de caractéristiques hiérarchiques dans les réseaux de neurones profonds.

2.2.3 Apprentissage Profond Versus Apprentissage Automatique Traditionnel : Une Analyse Comparative

Une distinction claire entre les méthodes d'apprentissage automatique traditionnelles et l'apprentissage profond réside dans les approches d'extraction de caractéristiques. Les approches d'apprentissage automatique conventionnelles, incluant la régression logistique [9], les machines à vecteurs de support (SVM) [10], et les forêts aléatoires [11], s'appuient significativement sur des caractéristiques artisanales adaptées à des problèmes spécifiques, nécessitant une intervention manuelle extensive et une expertise de domaine. En conséquence, leurs performances atteignent souvent un plateau lorsqu'elles rencontrent des données à grande échelle ou hautement variables.

Par contre, les techniques d'apprentissage profond découvrent automatiquement des caractéristiques significatives directement à partir de données brutes à travers l'apprentissage bout-à-bout, réduisant les biais induits par l'humain et améliorant l'évolutivité et la généralisation. Cette capacité d'apprentissage automatique de caractéristiques permet aux modèles d'apprentissage profond de gérer efficacement des ensembles de données de plus en plus complexes [4].

Par exemple, dans la classification de lésions cutanées, les approches traditionnelles dépendent typiquement de descripteurs définis manuellement tels que l'analyse de texture ou les histogrammes de couleur. En comparaison, les modèles d'apprentissage pro-

fond, particulièrement les CNNs, extraient automatiquement des caractéristiques discriminatives directement à partir de données au niveau pixel, produisant une précision et une robustesse améliorées, impactant ainsi profondément les diagnostics cliniques et les résultats patients.

TABLE 2.1 – Comparaison entre les Méthodes d'Apprentissage Automatique Traditionnelles et l'Apprentissage Profond

Critères	**Apprentissage Automatique Traditionnel**	**Apprentissage Profond**
Ingénierie de Caractéristiques	Nécessite une extraction manuelle de caractéristiques	Extraction automatique de caractéristiques bout-à-bout
Évolutivité	Limitée avec des données volumineuses ou haute-dimensionnelles	Évolutif vers des ensembles de données étendus et des structures complexes
Performance	Dépendante de la qualité des caractéristiques et de l'expertise domaine	Les performances s'améliorent substantiellement avec la quantité et qualité des données
Interprétabilité	Typiquement transparente et explicable	Considérée comme moins interprétable en raison des architectures profondes
Application	Ensembles de données structurés de taille modérée	Ensembles de données complexes, non-structurés, à grande échelle

2.2.4 Apprentissage Profond dans l'Analyse de Données Visuelles : Applications Médicales

L'apprentissage profond a significativement fait progresser l'analyse de données visuelles, particulièrement en imagerie médicale où les méthodes traditionnelles ont rencontré des difficultés avec les motifs spatiaux complexes et les caractéristiques diagnostiques subtiles. Les réseaux de neurones profonds, particulièrement les CNNs, modélisent intrinsèquement les relations spatiales, capturant les variations subtiles et les structures anatomiques détaillées cruciales dans les diagnostics cliniques.

Les modalités d'imagerie médicale telles que les rayons X, la tomodensitométrie, l'IRM et l'échographie génèrent des données haute-dimensionnelles nécessitant des techniques analytiques sophistiquées. L'extraction de caractéristiques hiérarchiques des CNNs

à partir des intensités de pixels permet une gestion efficace de cette complexité de données, améliorant la précision diagnostique pour des conditions telles que les nodules pulmonaires dans les scans CT, le cancer du sein dans les mammographies, et les maladies neurologiques en IRM. Ces avancées soulignent le potentiel transformateur de l'apprentissage profond dans les flux de travail cliniques, améliorant substantiellement les résultats patients et les efficacités opérationnelles [8].

2.3 Réseaux de Neurones Convolutionnels (CNNs) et Extraction de Caractéristiques

2.3.1 Introduction aux CNNs : Pourquoi Ils Sont Essentiels pour les Tâches de Vision

Les Réseaux de Neurones Convolutionnels (CNNs) ont émergé comme une architecture neuronale puissante explicitement conçue pour traiter des données spatiales telles que les images. Contrairement aux réseaux de neurones traditionnels (MLPs), qui échouent à préserver les relations spatiales en aplatissant les données d'images en vecteurs, les CNNs utilisent intrinsèquement la localité spatiale, les poids partagés et les représentations hiérarchiques pour identifier efficacement des caractéristiques significatives. Cette capacité de traitement tenant compte de l'information spatiale rend les CNNs particulièrement adaptés aux tâches liées aux images, où la reconnaissance de structures et de motifs est vitale [4]. En imagerie médicale, les CNNs facilitent l'extraction automatique de caractéristiques, améliorant significativement la précision diagnostique pour des images médicales complexes.

2.3.2 Architecture CNN et Composants Centraux

Les architectures CNN sont principalement composées de couches convolutionnelles, de fonctions d'activation, de couches de pooling et de couches entièrement connectées. Chaque composant remplit un rôle unique dans l'extraction de caractéristiques, la représentation et la prise de décision.

2.3.2.1 Couches Convolutionnelles

Les couches convolutionnelles constituent le composant fondamental des CNNs et sont spécifiquement conçues pour traiter des données structurées en grille, telles que les images. L'opération primaire au sein de ces couches est la convolution, qui implique l'application systématique de petits filtres apprenables connus sous le nom de kernels sur les données d'entrée pour générer des cartes de caractéristiques. Mathématiquement, l'opération de convolution bidimensionnelle standard entre une image d'entrée I et un

filtre F est définie par :

$$(F * I)(x, y) = \sum_{m,n} F(m, n) \cdot I(x - m, y - n) \tag{2.1}$$

où I représente l'image d'entrée, F dénote le kernel (filtre), (x, y) indique la position spatiale au sein de la carte de caractéristiques de sortie, et (m, n) représentent les indices itérant sur les dimensions du kernel, définissant les positions relatives au sein du filtre qui sont multipliées avec les pixels d'entrée correspondants lors de l'opération de convolution.

Chaque kernel convolutionnel (poids), typiquement initialisé au moyen de techniques telles que l'initialisation Xavier [12] ou He [13], apprend progressivement à reconnaître des motifs locaux spécifiques. Durant la phase d'entraînement, ces kernels sont ajustés via le mécanisme de rétropropagation, où les méthodes d'optimisation basées sur le gradient mettent à jour itérativement les poids des kernels pour minimiser l'écart entre les prédictions du réseau et les étiquettes de vérité terrain. Par conséquent, les filtres évoluent de la détection de motifs simplistes, tels que les contours et gradients de couleur, vers la capture de textures complexes et de structures spécifiques aux objets.

S'étendant au-delà des opérations de convolution standard, les CNNs incorporent des variantes convolutionnelles avancées, notamment les convolutions dilatées et les convolutions séparables en profondeur, chacune conçue pour traiter des défis distincts dans l'extraction de caractéristiques. Les convolutions dilatées, également connues sous le nom de convolutions atrous, introduisent des espaces entre les éléments du kernel, élargissant ainsi le champ récepteur sans ajouter de paramètres additionnels ou de charge de calcul. Une convolution dilatée avec un facteur de dilatation d est définie formellement par :

$$(F *_d I)(x, y) = \sum_{m,n} F(m, n) \cdot I(x - d \cdot m, y - d \cdot n) \tag{2.2}$$

Cette méthode permet aux CNNs d'intégrer un contexte multi-échelle, un avantage significatif pour segmenter des structures anatomiques de tailles variées ou capturer des lésions à différentes échelles dans les images médicales [5].

Les convolutions séparables en profondeur, introduites de manière proéminente par les architectures MobileNet [14], décomposent la convolution conventionnelle en deux opérations séparées : la convolution en profondeur et la convolution ponctuelle. Dans la convolution en profondeur, les kernels sont appliqués indépendamment pour chaque canal de l'image d'entrée, réduisant drastiquement la charge de calcul comparée aux convolutions standard. Par la suite, les convolutions ponctuelles (convolutions 1×1) combinent ces caractéristiques par canal, mélangeant l'information entre les canaux. Cette décomposition non seulement diminue le nombre de paramètres significativement, mais réduit aussi la complexité de calcul, facilitant l'entraînement et l'inférence efficaces. Une telle efficacité est particulièrement avantageuse dans les déploiements cliniques, où les ressources de calcul et l'efficacité temporelle sont des contraintes cruciales.

De plus, ces opérations de convolution spécialisées facilitent une représentation de caractéristiques améliorée, permettant aux CNNs de mieux discriminer parmi des carac-

téristiques cliniques étroitement liées, améliorant ainsi les performances du modèle sur des ensembles de données médicales sensibles et difficiles.

2.3.2.2 Fonctions d'Activation

Les fonctions d'activation constituent un composant essentiel des CNNs en introduisant des transformations non linéaires critiques qui permettent au réseau de modéliser des relations complexes inhérentes aux données. Sans fonctions d'activation, les réseaux de neurones ne feraient que des transformations linéaires, restreignant significativement leur capacité à représenter des motifs sophistiqués et non linéaires présents dans les images médicales. Ainsi, les fonctions d'activation jouent un rôle pivot dans l'amélioration de l'expressivité et de l'adaptabilité des CNNs.

Historiquement, la fonction sigmoïde était parmi les premières fonctions d'activation largement utilisées dans les réseaux de neurones. Mathématiquement, la fonction sigmoïde est définie par :

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

Bien que la sigmoïde produise des valeurs bornées entre 0 et 1, la rendant naturellement adaptée aux tâches de classification binaire, elle présente des inconvénients significatifs. Spécifiquement, les fonctions sigmoïdes sont susceptibles de saturation, où les sorties approchent leurs limites asymptotiques (0 ou 1) lorsque les valeurs d'entrée deviennent très grandes ou très petites. Dans ces régions saturées, les gradients approchent zéro, causant ce qui est communément connu sous le nom de problème du gradient évanescent [15]. Par conséquent, l'entraînement de réseaux de neurones profonds avec des fonctions d'activation sigmoïdes devient difficile en raison de la propagation inefficace des gradients dans les couches réseau plus profondes.

Pour atténuer certains problèmes de saturation, la fonction tangente hyperbolique (tanh) a émergé comme alternative. La fonction d'activation tanh est définie mathématiquement comme :

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{x}} \tag{2.4}$$

Contrairement à la sigmoïde, tanh mappe les entrées vers la plage $(-1, 1)$, offrant des sorties centrées sur zéro bénéfiques pour la convergence d'entraînement. Malgré cet avantage, tanh connaît toujours des problèmes de saturation aux valeurs d'entrée extrêmes, la rendant problématique dans les architectures de réseaux profonds en raison du risque continué d'évanescence du gradient.

Pour traiter ces limitations, l'Unité Linéaire Rectifiée (ReLU) est devenue la fonction d'activation standard dans les CNNs profonds, particulièrement favorisée pour les couches cachées. Définie par :

$$\mathrm{ReLU}(x) = \max(0, x) \tag{2.5}$$

ReLU introduit une efficacité de calcul, des taux de convergence plus rapides et atténue le problème du gradient évanescent rencontré avec les fonctions sigmoïde et tanh. En raison

de sa forme linéaire par morceaux, ReLU facilite une propagation efficace des gradients durant la rétropropagation. Cependant, ReLU n'est pas sans limitations ; les neurones peuvent devenir inactifs durant l'entraînement s'ils reçoivent constamment des entrées négatives, un scénario référé comme le problème du "ReLU mourant" [13]. Pour traiter ceci, des versions alternatives telles que Leaky ReLU et Exponential Linear Unit (ELU) ont émergé. Leaky ReLU introduit une petite pente négative aux entrées négatives, mathématiquement exprimée comme :

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{si } x \geq 0 \\ \alpha x, & \text{sinon} \end{cases}, \quad \alpha \text{ typiquement petit (e.g., 0.01)} \tag{2.6}$$

Similairement, ELU lisse l'espace d'entrée négatif pour maintenir la propagation du gradient :

$$\text{ELU}(x) = \begin{cases} x, & \text{si } x \geq 0 \\ \alpha(e^x - 1), & \text{sinon} \end{cases} \tag{2.7}$$

où α est un hyperparamètre positif, typiquement fixé à 1.

Tandis que les fonctions sigmoïde et tanh demeurent utiles dans les couches de sortie pour des tâches spécifiques, les architectures CNN modernes utilisent prédominairement ReLU ou ses variantes dans les couches cachées en raison de leur robustesse et performance supérieure. Le choix attentif des fonctions d'activation impacte critiquement les dynamiques d'entraînement, l'efficacité d'extraction de caractéristiques et la précision globale du modèle, particulièrement dans des applications sophistiquées telles que la segmentation d'images médicales. La Figure 2.3 présente une vue d'ensemble comparative des fonctions d'activation les plus couramment utilisées dans les architectures CNN, illustrant leurs comportements respectifs sur le domaine d'entrée.

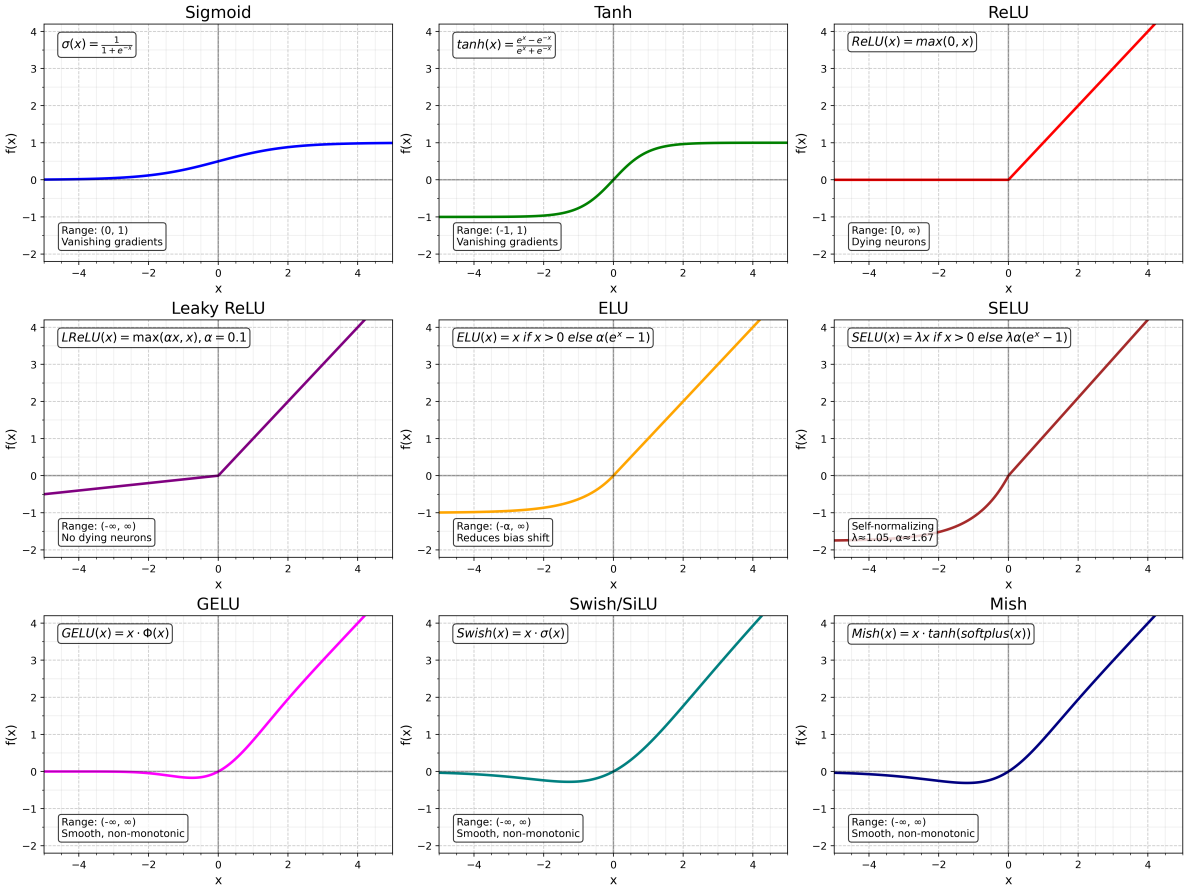

FIGURE 2.3 – Fonctions d'activation les plus populaires utilisées dans les architectures CNN.

2.3.2.3 Couches de Pooling

Les couches de pooling jouent un rôle crucial dans les CNNs, réduisant systématiquement la dimensionnalité des cartes de caractéristiques extraites pour améliorer l'efficacité de calcul et l'invariance du modèle aux transformations d'entrée. L'objectif primaire du pooling est de consolider l'information spatiale au sein des cartes de caractéristiques, réduisant ainsi la complexité des couches réseau ultérieures et rendant les CNNs robustes contre les translations spatiales mineures et les distorsions dans les données d'entrée. En résumant la sortie des couches convolutionnelles au sein d'un voisinage défini, le pooling assure que de petites variations en position ou orientation n'altèrent pas significativement les représentations de plus haut niveau, ce qui est particulièrement avantageux en analyse d'images médicales, où des variations subtiles dans les conditions d'imagerie sont communes [8]. La Figure 2.4 illustre le fonctionnement détaillé de l'opération de max pooling 2×2 appliquée à une image en niveaux de gris.

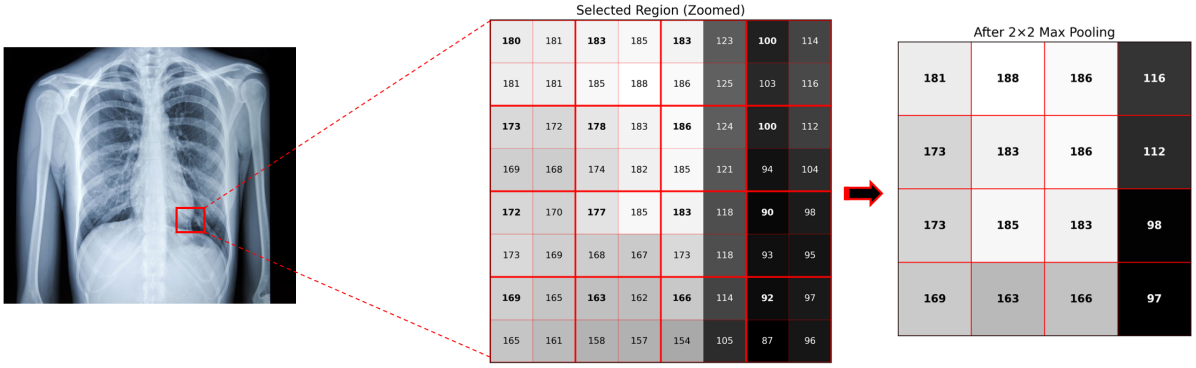

FIGURE 2.4 – Visualisation détaillée de l'opération de max pooling 2×2. Gauche : Image en niveaux de gris originale avec région surlignée. Centre : Vue agrandie de la région sélectionnée 8×8 pixels avec superposition de grille montrant les valeurs de pixels originaux. Droite : Résultat après application du max pooling 2×2, où chaque cellule contient la valeur maximale de sa région 2×2 correspondante dans l'image originale.

La forme la plus commune de pooling est le max pooling, qui sélectionne l'activation maximale au sein d'une région spécifiée, typiquement une fenêtre rectangulaire. Formellement, le max pooling est défini mathématiquement comme :

$$P_{x,y}^{\max} = \max_{(i,j)\in R_{x,y}} (a_{i,j}) \tag{2.8}$$

où $R_{x,y}$ représente la région de pooling locale autour de la position (x, y). Le max pooling est particulièrement efficace pour faire ressortir les caractéristiques les plus saillantes, telles que les contours proéminents ou les motifs structurels distincts, tout en écartant les activations moins significatives ou bruyantes. Ainsi, il renforce les caractéristiques les plus discriminatives dans différentes localisations de la carte de caractéristiques.

Une alternative au max pooling est l'average pooling, qui calcule la moyenne arithmétique des activations au sein de la région de pooling. L'average pooling est défini comme suit :

$$P_{x,y}^{\text{avg}} = \frac{1}{|R_{x,y}|} \sum_{(i,j)\in R_{x,y}} (a_{i,j}) \tag{2.9}$$

Contrairement au max pooling, l'average pooling préserve une information contextuelle plus compréhensive en capturant l'intensité globale au sein de la région de pooling. Bien que l'average pooling soit moins sensible aux pics aigus dans les activations, il peut potentiellement diluer les signaux saillants avec ceux moins informatifs, affectant la capacité du réseau à détecter les détails fins critiques dans des tâches spécifiques de segmentation d'images médicales. Néanmoins, l'average pooling peut être avantageux quand le lissage du bruit ou des artefacts est désirable.

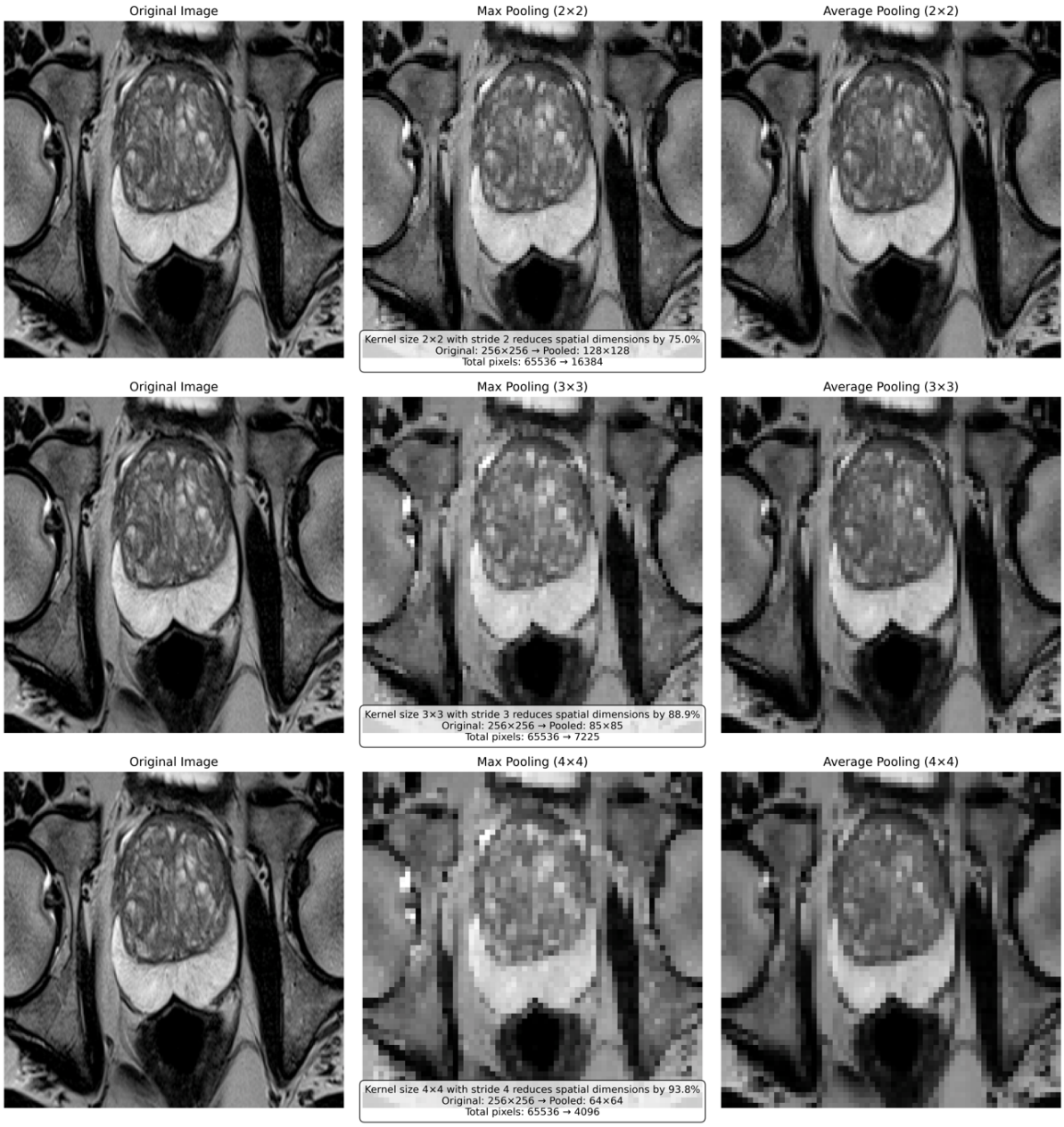

FIGURE 2.5 – Comparaison du max pooling et average pooling avec différentes tailles de kernel. Rangée supérieure : Image originale (gauche), max pooling avec kernel 2×2 (centre), et average pooling avec kernel 2×2 (droite). Rangée médiane : Image originale (gauche), max pooling avec kernel 3×3 (centre), et average pooling avec kernel 3×3 (droite). Rangée inférieure : Image originale (gauche), max pooling avec kernel 4×4 (centre), et average pooling avec kernel 4×4 (droite). Notez comment les tailles de kernel plus grandes réduisent les dimensions spatiales plus agressivement tandis que le max pooling préserve mieux les caractéristiques importantes que l'average pooling.

Au-delà des méthodes de pooling local, les architectures CNN implémentent fréquemment des techniques de pooling global, telles que le global average pooling et le global max pooling. Ces méthodes agrègent des cartes de caractéristiques spatiales entières en valeurs scalaires simples, réduisant ainsi substantiellement la complexité du modèle et le nombre de paramètres requis par les couches entièrement connectées. De plus, le pooling global contraint le réseau à développer des représentations généralisées et holistiques, améliorant la robustesse et la généralisation du modèle, particulièrement bénéfiques lors du traitement d'ensembles de données médicales limités ou bruités.

Les architectures CNN avancées peuvent aussi intégrer des stratégies de pooling alternatives, telles que le Lp pooling et le stochastic pooling. Le Lp pooling généralise les opérations de pooling en appliquant des Lp-normes au sein des régions de pooling, introduisant de la flexibilité et capturant potentiellement des statistiques spatiales complexes. Le stochastic pooling sélectionne les activations de manière probabiliste basée sur leurs magnitudes, introduisant un mécanisme de régularisation stochastique qui peut aider à atténuer le surapprentissage.

Néanmoins, les opérations de pooling viennent avec des limitations inhérentes, notamment la perte potentielle de résolution spatiale. Cette perte d'information spatiale peut affecter négativement les tâches requérant une localisation précise, telle que la segmentation d'images médicales. Pour contrecarrer cette limitation, les architectures CNN contemporaines réduisent souvent ou éliminent les couches de pooling traditionnelles, les remplaçant par des convolutions à pas ou des convolutions dilatées, équilibrant ainsi la préservation d'information spatiale détaillée avec l'efficacité de calcul.

Le choix et le placement des opérations de pooling demeurent ainsi des considérations essentielles dans la conception d'architectures CNN, particulièrement dans le domaine de l'imagerie médicale, où maintenir l'information anatomique détaillée tout en atteignant l'efficacité de calcul est crucial.

2.3.2.4 Couches Entièrement Connectées

Les couches entièrement connectées (FC) constituent un composant critique des CNNs, servant particulièrement comme étapes de traitement finales suivant les opérations convolutionnelles et de pooling. Contrairement aux couches convolutionnelles, qui exploitent la localité spatiale et les poids partagés, les couches entièrement connectées comprennent de multiples neurones interconnectés, chaque neurone étant connecté à chaque neurone de la couche précédente. Cette structure de connectivité dense ressemble à celle des perceptrons multicouches (MLPs) traditionnels et permet aux couches FC d'effectuer un raisonnement complexe de haut niveau en intégrant les caractéristiques compréhensives et abstraites extraites des couches antérieures.

Dans les couches entièrement connectées, chaque neurone calcule une somme pondérée des entrées de tous les neurones de la couche précédente et applique par la suite une fonction d'activation non linéaire. Formellement, le calcul effectué par une couche FC peut être exprimé mathématiquement comme :

$$y = f(Wx + b) \tag{2.10}$$

où x dénote le vecteur de caractéristiques d'entrée, typiquement obtenu en aplatissant les cartes de caractéristiques multidimensionnelles produites par les couches convolutionnelles et de pooling précédentes ; W est la matrice de poids contenant les paramètres appris qui quantifient l'importance de chaque caractéristique d'entrée ; b représente le vecteur de biais, consistant aussi de paramètres appris qui permettent au réseau de décaler les fonctions d'activation pour mieux s'ajuster aux données ; et f est une fonction d'activation non linéaire, communément softmax pour les tâches de classification ou sigmoïde pour les scénarios de classification binaire [5]. Le rôle primaire des couches FC au sein d'une architecture CNN est de synthétiser l'information capturée par les couches convolutionnelles et de pooling en une forme appropriée pour la prise de décision ou la classification. Chaque neurone dans ces couches agit comme un intégrateur de caractéristiques, combinant de multiples caractéristiques de haut niveau en représentations unifiées qui sont par la suite utilisées pour la tâche de prédiction finale. En raison de leurs connexions denses, les couches FC sont hautement efficaces pour capturer le contexte global et les relations parmi les caractéristiques extraites, facilitant une classification précise et une prise de décision robuste.

Cependant, la connectivité dense inhérente aux couches FC résulte en nombres significativement plus élevés de paramètres comparés aux couches convolutionnelles, augmentant substantiellement la complexité de calcul et le risque de surapprentissage, particulièrement lors du traitement d'ensembles de données médicales relativement petits. Pour traiter ces défis, diverses techniques de régularisation, telles que dropout [16], weight decay [17], et batch normalization [18], sont typiquement appliquées aux couches FC pour améliorer la généralisation et atténuer les risques de surapprentissage. Dropout, par exemple, désactive aléatoirement un sous-ensemble de neurones durant l'entraînement, prévenant ainsi la co-adaptation parmi les neurones et encourageant le réseau à développer des représentations de caractéristiques plus robustes et généralisées. Weight decay ajoute un terme de pénalité à la fonction de perte proportionnel à la magnitude des poids du réseau, contraignant les valeurs des paramètres et prévenant une complexité excessive du modèle. Batch normalization standardise les entrées des couches en normalisant les activations sur les mini-batches, stabilisant les dynamiques d'entraînement et réduisant la sensibilité à l'initialisation des poids tout en fournissant des effets de régularisation implicites. De plus, les architectures CNN contemporaines emploient parfois des couches de pooling global comme remplacement ou complément aux couches entièrement connectées pour réduire le nombre de paramètres et la complexité de calcul, particulièrement dans les applications où l'efficacité et la généralisation sont primordiales. Néanmoins, les couches entièrement connectées demeurent indispensables dans les scénarios requérant une intégration explicite de caractéristiques et des décisions de classification détaillées, telles que les tâches diagnostiques médicales précises où la synthèse de caractéristiques diverses extraites de données d'imagerie complexes est critique.

2.3.3 Forces et Limitations des CNNs en Imagerie Médicale

Les Réseaux de Neurones Convolutionnels ont révolutionné l'analyse d'images médicales en raison de leurs capacités d'extraction de caractéristiques hiérarchiques automatisées, leur invariance spatiale et leur performance remarquable dans diverses applications cliniques. En apprenant directement à partir d'entrées au niveau pixel, les CNNs réduisent significativement la dépendance à l'ingénierie manuelle de caractéristiques, permettant au modèle de découvrir des motifs complexes cruciaux pour la précision diagnostique.

Malgré ces forces, les CNNs souffrent intrinsèquement de limitations critiques. La limitation primaire provient de la nature locale des opérations de convolution, qui capturent l'information exclusivement au sein de champs récepteurs limités. Cette contrainte de localité restreint les CNNs de modéliser efficacement les dépendances à long terme et les relations contextuelles globales au sein d'une image. De telles contraintes deviennent particulièrement problématiques en imagerie médicale, où une compréhension compréhensive du contexte global, tel que reconnaître les relations anatomiques et la cohérence spatiale des lésions, est essentielle pour des diagnostics précis. Par exemple, distinguer des structures pathologiques dans des scans radiologiques nécessite souvent d'intégrer l'information de régions disparates de l'image, une capacité intrinsèquement limitée par les opérations locales des CNNs.

De plus, les caractéristiques intrinsèques des images médicales amplifient les défis associés aux CNNs. Les images cliniques exhibent fréquemment une variabilité considérable en termes de qualité, contraste, résolution et présence d'artefacts en raison des différences dans les protocoles d'imagerie, mouvements patients, ou limitations d'équipement. Additionnellement, les modalités d'imagerie médicale telles que l'échographie, l'IRM, ou les scans CT produisent des images intrinsèquement bruyantes et parfois ambiguës. La localité des opérations de convolution exacerbe davantage ces défis, car les CNNs peuvent avoir du mal à contextualiser le bruit ou les artefacts quand ils manquent de contexte spatial plus large nécessaire pour les différencier des caractéristiques diagnostiquement pertinentes. Un autre défi critique pour les CNNs en imagerie médicale concerne l'interprétabilité. Les architectures CNN, fonctionnant largement comme des boîtes noires, offrent un aperçu limité de leurs processus de prise de décision. Cette opacité complique la validation clinique et diminue la confiance parmi les professionnels de santé, qui requièrent des prédictions explicables pour informer des décisions cliniques critiques. Pour atténuer ces limitations inhérentes, la recherche actuelle intègre les CNNs avec des mécanismes d'attention globaux et des architectures basées sur les transformers. Ces approches hybrides étendent le champ récepteur et améliorent la capacité du réseau à capturer efficacement les dépendances à long terme et les indices contextuels globaux. En introduisant des modules d'attention explicites, les réseaux peuvent pondérer sélectivement l'information contextuelle globale, améliorant ainsi l'interprétabilité et les qualités de robustesse indispensables pour les applications cliniques à enjeux élevés [8].

En résumé, bien que les CNNs aient significativement fait progresser les applications d'imagerie médicale, comprendre leurs limitations de localité et les traiter par des inno-

vations architecturales avancées demeure crucial pour exploiter leur plein potentiel dans les diagnostics cliniques.

2.4 Architectures Modernes d'Apprentissage Profond pour les Tâches de Vision

2.4.1 Évolution au-delà des CNNs

Les CNNs ont indubitablement marqué un tournant significatif dans les tâches de reconnaissance et d'analyse visuelles. Malgré leur succès profond, les architectures CNN souffrent intrinsèquement des limitations discutées dans la section précédente. Pour traiter ces limitations fondamentales, les architectures modernes ont évolué au-delà des CNNs classiques, incorporant des paradigmes de calcul avancés capables de modéliser efficacement les dépendances à long terme et l'information contextuelle globale. L'évolution dans les architectures d'apprentissage profond a été motivée par la nécessité de capturer plus efficacement les données hiérarchiques et relationnelles dans diverses tâches de reconnaissance visuelle.

L'exploration au-delà des CNNs a introduit deux paradigmes primaires. Premièrement, les architectures transformer, initialement conçues pour le traitement du langage naturel, ont significativement influencé la vision par ordinateur par les Vision Transformers (ViTs), employant des mécanismes de self-attention qui capturent les dépendances à long terme en pondérant dynamiquement l'importance de chaque élément dans les données d'entrée [5]. Deuxièmement, les modèles state-space, particulièrement les réseaux Mamba, exploitent leur capacité à modéliser les dépendances à long terme avec une efficacité de calcul améliorée, contournant les limitations traditionnelles à la fois des CNNs et des transformers [2].

Additionnellement, les architectures hybrides combinant les CNNs avec les ViTs ou SSMs représentent une approche convaincante, exploitant les forces complémentaires des paradigmes de modélisation convolutionnels et basés sur l'attention ou les modèles state-space. Ces modèles intégrés équilibrent efficacement l'extraction de caractéristiques locales avec la compréhension contextuelle globale, fournissant une performance, robustesse et interprétabilité supérieures dans diverses tâches de reconnaissance visuelle [19, 20].

En résumé, l'évolution au-delà des CNNs reflète une avancée critique dans l'apprentissage profond, motivée par le besoin de modèles plus puissants, flexibles et efficaces capables de traiter les défis intrinsèques posés par les données visuelles complexes.

2.4.2 Vision Transformers (ViTs)

Les Vision Transformers (ViTs) représentent un départ significatif des CNNs traditionnels, altérant fondamentalement l'approche de l'analyse de données visuelles en

exploitant les mécanismes d'attention originellement introduits dans le domaine du traitement du langage naturel (NLP) [1]. Contrairement aux CNNs, qui construisent progressivement des caractéristiques visuelles hiérarchiques par des opérations de convolution locales, les Vision Transformers traitent les images globalement dès les premières étapes en utilisant des mécanismes de self-attention. Ce changement radical permet aux ViTs de capturer intrinsèquement le contexte global et les dépendances spatiales à long terme, traitant une limitation critique des architectures CNN dans les tâches visuelles complexes.

L'architecture ViT, introduite par Dosovitskiy et al. [21], remodèle fondamentalement l'approche de la reconnaissance d'images en traitant les images similairement aux tokens dans le traitement du langage. Spécifiquement, une image est d'abord partitionnée en patches non-chevauchants, chacun aplati en vecteurs, appelés patch embeddings. Ces embeddings sont linéairement projetés et augmentés avec des embeddings positionnels, préservant ainsi l'information spatiale essentielle. Formellement, étant donnée une image représentée par $I \in \mathbb{R}^{H\times W\times C}$ (où H, W, et C dénotent la hauteur, largeur et canaux respectivement), l'image est divisée en N patches non-chevauchants, chacun de taille $(P\times P\times C)$, produisant une séquence de patches aplatis $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, où $N = HW/P^2$ est la longueur de séquence des patch embeddings [21].

Central à l'efficacité des ViTs est le mécanisme de self-attention, qui calcule la pertinence entre tous les patches d'image simultanément. Le self-attention calcule les queries (Q), keys (K), et values (V) à partir des patches embarqués et détermine dynamiquement les poids basés sur la pertinence inter-patch. Cette opération, connue sous le nom de scaled dot-product attention, est mathématiquement définie comme :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\Bigg(V \tag{2.11}$$

où Q, K, et V représentent des projections linéaires apprises des patches embarqués, et d_k dénote la dimension des queries et keys [1]. L'introduction de la multi-head self-attention enrichit davantage la capacité représentationnelle du modèle en traitant concurremment l'information de multiples sous-espaces de représentation. Ceci permet au modèle de capturer des motifs divers au sein des données visuelles simultanément, améliorant ainsi significativement sa conscience du contexte global et son pouvoir interprétatif.

Malgré leurs biais inductifs moindres comparés aux CNNs, particulièrement concernant la localité, l'invariance spatiale et l'équivariance de translation, les ViTs ont démontré une performance supérieure quand pré-entraînés sur des ensembles de données suffisamment grands. Leur capacité à modéliser le contexte global sans biais inductifs explicites leur permet d'exceller dans diverses tâches de reconnaissance visuelle où le contexte global influence significativement la précision. Néanmoins, les ViTs ne sont pas sans défis ; leur dépendance au pré-entraînement à grande échelle pour une performance optimale demeure une considération. Cependant, des innovations telles que les architectures hybrides combinant CNNs avec ViTs offrent des avenues prometteuses pour atténuer ces contraintes.

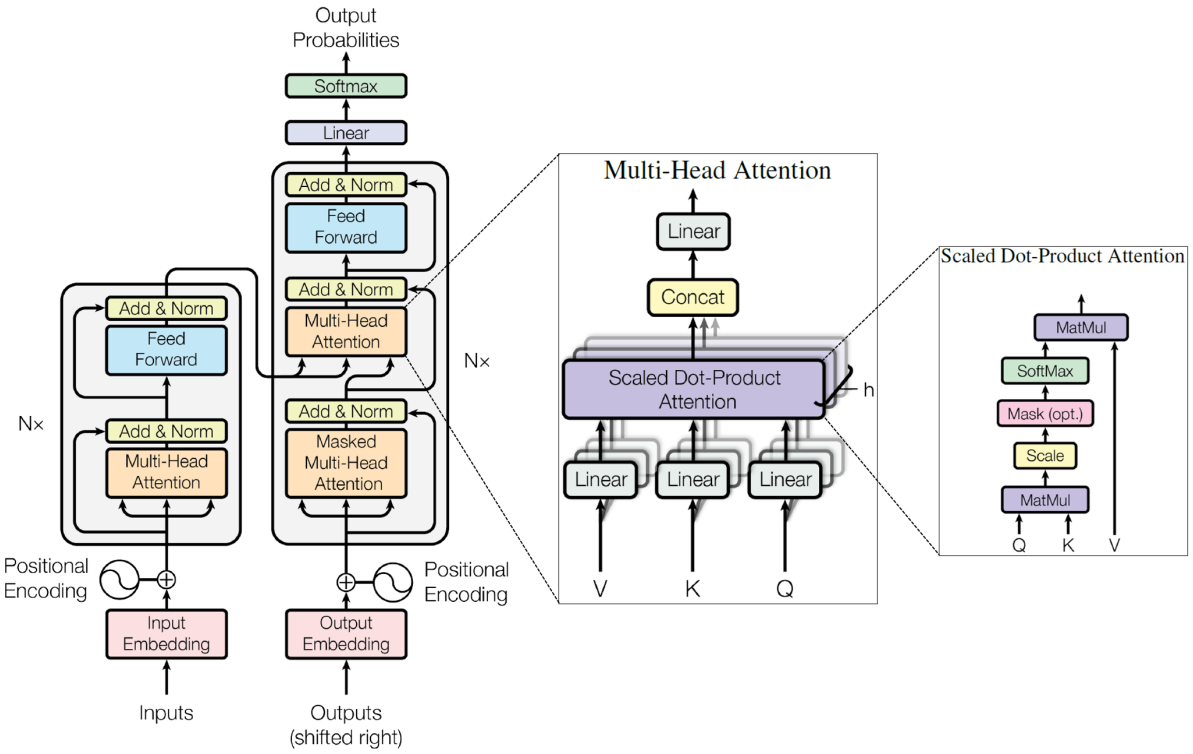

FIGURE 2.6 – L'architecture du modèle Transformer [1].

Globalement, l'architecture Vision Transformer représente une avancée pivot dans l'apprentissage de représentations visuelles, permettant une modélisation compréhensive du contexte global dans diverses applications de vision par ordinateur.

2.4.3 Modèles State-Space (Mamba)

Récemment, les Modèles State-Space (SSMs), particulièrement les architectures telles que Mamba, ont émergé comme alternatives convaincantes aux CNNs et aux transformers, offrant des stratégies de calcul innovantes pour traiter les données visuelles efficacement et effectivement. Issus de la théorie classique du contrôle et des systèmes, les modèles state-space représentent les motifs visuels par des formulations algébriques linéaires, où l'état interne du modèle évolue récursivement basé sur des transformations linéaires des entrées actuelles et des états précédents. Cette modélisation séquentielle capture intrinsèquement les dépendances à long terme avec une efficacité remarquable, surmontant les limitations des champs récepteurs restreints des CNNs et la complexité quadratique des mécanismes d'attention transformer [2].

Formellement, les modèles state-space opèrent en maintenant un vecteur d'état caché mis à jour par des transitions linéaires définies comme suit :

$$s_{t+1} = As_t + Bx_t, \quad y_t = Cs_t + Dx_t \tag{2.12}$$

Dans cette équation, le vecteur d'état s_t encode la mémoire du modèle des entrées précédentes, mis à jour itérativement par la matrice de transition A et influencé par l'entrée actuelle x_t via la matrice de projection d'entrée B. La sortie y_t, calculée via les matrices C et D, représente directement les prédictions ou caractéristiques extraites, résumant ainsi efficacement l'information contextuelle pertinente sur de longues séquences.

Les modèles state-space ont récemment été revitalisés par des architectures telles que le modèle Mamba, qui améliorent significativement leur applicabilité dans les contextes modernes d'apprentissage profond. L'innovation clé de Mamba réside dans la modélisation state-space sélective, introduisant des transformations state-space linéaires structurées optimisées explicitement pour la parallélisation matérielle moderne. En employant des techniques telles que le routage sélectif et les optimisations orientées matériel, Mamba atteint une complexité de calcul quasi-linéaire avec la longueur de séquence, un avantage considérable sur les transformers dont les calculs de self-attention évoluent quadratiquement [2].

Étendant ce concept dans le traitement de données visuelles, Vision Mamba (Vim) a introduit des modifications notables adaptées explicitement pour les tâches d'analyse d'images. Contrairement aux CNNs traditionnels, qui s'appuient sur des convolutions spatialement locales, ou aux transformers, qui utilisent la self-attention globale, Vision Mamba traite l'information visuelle comme une séquence de patches d'images tokenisés de manière bidirectionnelle, améliorant significativement sa capacité à capturer efficacement les relations contextuelles globales. Le modèle Vim partitionne une image d'entrée en patches, chacun linéairement embarqué et positionnellement encodé pour maintenir le contexte spatial [3].

Au niveau du calcul, Vision Mamba démontre des améliorations d'efficacité substantielles sur les CNNs et transformers. La complexité linéaire des calculs state-space, combinée avec les stratégies d'optimisation orientées matériel, permet à Vim d'atteindre des vitesses d'inférence substantiellement plus élevées et un usage réduit de mémoire GPU. Cette efficacité de calcul positionne Vision Mamba comme un backbone hautement attractif pour le déploiement dans des environnements à ressources contraintes ou des applications en temps réel où l'inférence rapide est cruciale.

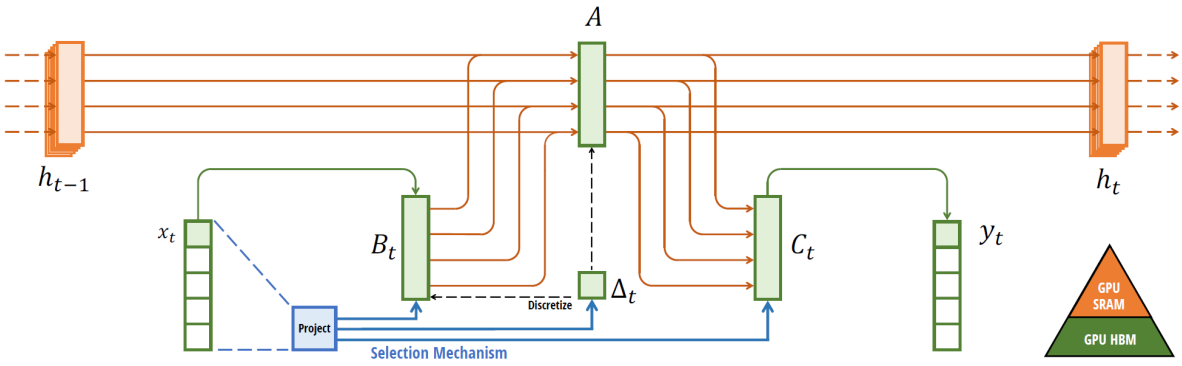

FIGURE 2.7 – Aperçu du Selective State Space Model. Le modèle transforme les séquences d'entrée x_t en sorties y_t par des transitions d'états latents, avec une sélection de paramètres dépendante de l'entrée permettant un traitement efficace de l'information [2].

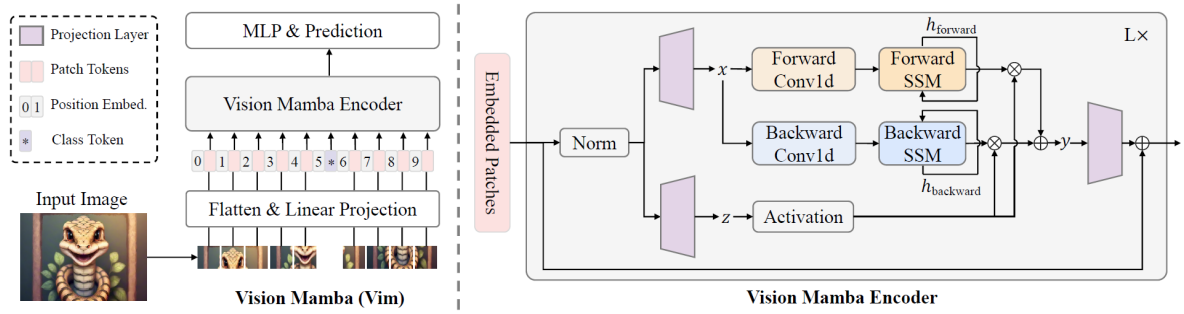

FIGURE 2.8 – L'architecture de Vision Mamba démontrant le traitement bidirectionnel et les embeddings positionnels pour une modélisation efficace du contexte visuel [3].

2.4.4 Mécanismes d'Attention dans l'Apprentissage Profond

Les mécanismes d'attention ont émergé comme composants indispensables dans les frameworks d'apprentissage profond modernes, améliorant significativement la capacité représentationnelle et l'interprétabilité des modèles en se concentrant sélectivement sur l'information critique tout en supprimant les données non pertinentes ou redondantes. Inspirés par le système cognitif humain, qui concentre naturellement l'attention sur les entrées sensorielles saillantes pour traiter efficacement le flot d'information écrasant, les modèles d'apprentissage profond utilisent l'attention pour allouer préférentiellement les ressources de calcul vers les caractéristiques les plus pertinentes au sein d'une entrée, améliorant la précision et la robustesse des processus de prise de décision [22].

2.4.4.1 Le Rôle de l'Attention dans l'Amélioration de l'Extraction de Caractéristiques

Dans les réseaux de neurones convolutionnels, les caractéristiques sont traditionnellement extraites par des convolutions spatiales fixes qui traitent uniformément toutes les régions d'entrée. Cependant, dans de nombreux scénarios, particulièrement en imagerie médicale, toutes les régions d'entrée ne contribuent pas également aux insights cliniques. Les mécanismes d'attention pondèrent adaptivement l'importance des caractéristiques, permettant aux modèles de mettre l'accent sélectivement sur les régions diagnostiquement pertinentes, de supprimer l'information non pertinente ou redondante, et d'améliorer la qualité des représentations extraites. Par conséquent, les mécanismes d'attention sont devenus instrumentaux dans l'affinement des cartes de caractéristiques pour mieux représenter l'information critique nécessaire pour une interprétation et une prise de décisions cliniques précises.

2.4.4.2 Types de Mécanismes d'Attention

Les mécanismes d'attention peuvent être largement catégorisés en attention spatiale, attention de canal, et self-attention, chacun adapté à des aspects distincts d'affinement et de modélisation de caractéristiques.

Attention Spatiale. Les mécanismes d'attention spatiale priorisent les régions au sein des cartes de caractéristiques basées sur leur importance informative, concentrant efficacement les ressources du modèle sur les localisations spatiales saillantes critiques pour des prédictions précises. L'attention spatiale améliore les CNNs en pondérant adaptivement les caractéristiques basées sur leur pertinence au sein de contextes spatiaux spécifiques. Typiquement, cette attention est calculée en utilisant des descripteurs spatiaux poolés tels que le global max pooling et le global average pooling suivis d'opérations convolutionnelles pour générer une carte d'attention spatiale. Cette carte met l'accent sur les régions importantes et supprime celles non pertinentes, menant à une précision de localisation améliorée essentielle dans les tâches de segmentation [22].

Attention de Canal L'attention de canal vise à identifier la signification des canaux de caractéristiques individuelles en considérant les interdépendances entre canaux. Des méthodes comme le bloc Squeeze-and-Excitation (SE) recalibrent dynamiquement les caractéristiques par canal en modélisant explicitement les dépendances inter-canaux par le global average pooling et les couches entièrement connectées [22]. Une telle recalibration par canal aide les réseaux à se concentrer sur les canaux portant les signaux les plus informatifs pertinents à des caractéristiques cliniques spécifiques, améliorant ainsi l'efficacité du modèle dans la capture de caractéristiques pathologiques complexes. Des variantes telles que Efficient Channel Attention (ECA) optimisent davantage ce processus en réduisant la complexité de calcul, maintenant une efficacité de performance appropriée pour les applications médicales.

Self-Attention (Transformers) Le self-attention, extensivement popularisé par les transformers, traite la limitation des champs récepteurs locaux en capturant des dépendances globales étendues entre les positions spatiales. Elle calcule les poids d'attention basés sur les interactions entre toutes les paires de positions au sein d'une séquence d'entrée, ajustant dynamiquement la représentation de chaque caractéristique basée sur sa pertinence contextuelle à toute autre caractéristique.

Détection de Saillance comme Mécanisme d'Attention. Les mécanismes de détection de saillance ont été efficacement intégrés dans les frameworks d'attention pour affiner davantage la capacité des modèles à se concentrer sur les régions diagnostiquement significatives au sein des images. Les méthodes basées sur la saillance visent à émuler l'attention visuelle humaine en identifiant les régions d'une image qui se distinguent visuellement ou sont les plus informatives, guidant ainsi l'attention du modèle vers ces caractéristiques saillantes. Ces méthodes s'appuient sur l'analyse de caractéristiques visuelles de bas niveau telles que le contraste, la luminance et l'orientation des contours pour produire des cartes de saillance, mettant en évidence les régions qui capturent l'intérêt et la pertinence visuels [23]. Des implémentations récentes d'attention basée sur la saillance, telles que graph-based visual saliency (GBVS), utilisent des modèles graphiques pour mettre en évidence systématiquement les zones visuellement proéminentes, guidant l'exploration et l'analyse visuelles ultérieures [24]. En imagerie médicale, les cartes de saillance facilitent la détection et la segmentation de régions pathologiques en pondérant dynamiquement les zones susceptibles de contenir de l'information cliniquement pertinente, améliorant significativement l'efficacité et la précision des interprétations diagnostiques. De plus, l'évidence suggère que le guidage visuel basé sur la saillance ressemble étroitement aux comportements naturels de regard humain durant l'exploration de scène, supportant son adoption dans les contextes cliniques pour améliorer la précision diagnostique et l'interprétabilité [24, 25].

2.4.4.3 Signification dans les Modèles de Segmentation

Les mécanismes d'attention ont significativement élevé la performance des modèles de segmentation d'images médicales en combinant efficacement l'information locale et globale. En mettant explicitement en évidence les régions, canaux et structures relationnelles cruciaux au sein des images médicales, les modèles améliorés par l'attention démontrent une précision, une interprétabilité et des capacités de généralisation supérieures comparées aux architectures traditionnelles CNN uniquement. Particulièrement, les mécanismes d'attention permettent aux modèles profonds de délimiter plus précisément les frontières et structures pathologiques, une exigence fondamentale pour les diagnostics cliniques et la planification thérapeutique. De plus, les mécanismes d'attention traitent les problèmes communs rencontrés en imagerie médicale, tels que les tailles de lésions variées, les structures anatomiques irrégulières et les marqueurs patho-

logiques subtils, renforçant ainsi leur rôle pivot dans les systèmes d'imagerie médicale modernes [22].

2.4.5 Auto-encodeurs et Apprentissage de Représentations

Les auto-encodeurs représentent un paradigme puissant au sein de l'apprentissage profond, se concentrant fondamentalement sur l'apprentissage de représentations en reconstruisant les données d'entrée par une représentation interne compressée. L'architecture primaire d'un auto-encodeur comprend deux parties intégrales : un encodeur et un décodeur. La fonction encodeur transforme les données d'entrée en une représentation latente, communément référée comme un code, tandis que le décodeur vise à reconstruire l'entrée originale à partir de cette représentation condensée. Idéalement, l'auto-encodeur apprend à capturer les caractéristiques les plus saillantes des données, permettant efficacement la réduction de dimensionnalité et le débruitage [26]. Initialement proposés comme outils de réduction de dimensionnalité, les auto-encodeurs ont évolué significativement, exploitant les avancées dans les architectures de réseaux de neurones et les méthodologies d'entraînement. Les premiers auto-encodeurs étaient typiquement sous-complets, signifiant que la dimension du code (représentation latente) était explicitement contrainte à être plus petite que la dimension de l'entrée. Cette contrainte forçait les auto-encodeurs à prioriser et capturer seulement les caractéristiques les plus essentielles et discriminatives au sein des données d'entrée. Les auto-encodeurs non linéaires offraient particulièrement des avancées substantielles en étendant les méthodes de réduction de dimensionnalité linéaires telles que l'Analyse en Composantes Principales (PCA), fournissant une généralisation flexible et non linéaire capable de capturer des structures de données complexes non atteignables avec des méthodes linéaires comme PCA [27].

En analyse d'images médicales, les auto-encodeurs ont trouvé des applications répandues en raison de leur capacité inhérente pour l'extraction de caractéristiques et la réduction de dimensionnalité sans s'appuyer sur des données étiquetées, traitant la rareté commune des annotations. Les auto-encodeurs peuvent capturer les motifs sous-jacents et les structures complexes dans les ensembles de données d'imagerie médicale, permettant des tâches telles que le débruitage d'images, la détection d'anomalies, et la détection de lésions non supervisées. Par exemple, les denoising auto-encodeurs (DAEs) ont été particulièrement valables en imagerie médicale, entraînés explicitement pour reconstruire des images propres à partir de versions bruyantes. En forçant le réseau à apprendre une représentation interne robuste de la structure d'image sous-jacente, les DAEs améliorent la robustesse des systèmes diagnostiques contre diverses sources de bruit et artefacts communément rencontrés dans les scénarios d'imagerie clinique [28, 29]. De plus, plusieurs variantes d'auto-encodeur spécialisées ont émergé, chacune adaptée pour des scénarios spécifiques au sein de l'analyse d'images médicales. Les sparse auto-encodeurs, qui imposent des contraintes de sparsité sur la représentation latente, encouragent le réseau à activer moins de neurones, produisant ainsi des caractéristiques plus interprétables et sémantiquement significatives, souvent utilisées dans des tâches requérant des

représentations de caractéristiques distinctes telles que la détection d'anomalies ou la segmentation de lésions. De même, les contractifs auto-encodeurs (CAE) introduisent des pénalités de régularisation basées sur le Jacobien de l'activation de l'encodeur, promouvant des représentations stables et invariantes en rendant le modèle insensible aux petites perturbations dans les données d'entrée. Cette propriété est particulièrement bénéfique dans les contextes cliniques, où la consistance et la stabilité des caractéristiques extraites sont critiques pour des résultats diagnostiques fiables [30, 31].

Les Auto-encodeurs Variationnels (VAEs) représentent une autre avancée significative au sein des frameworks d'auto-encodeur, étendant le concept en embarquant des variables latentes au sein d'un framework de modélisation probabiliste. Les auto-encodeurs variationnels encodent les entrées en distributions de probabilité plutôt qu'en vecteurs déterministes, permettant ainsi la génération d'images médicales nouvelles et synthétiques, aidant l'augmentation de données, la détection d'anomalies, et les tâches de modélisation générative. Cette capacité de modélisation probabiliste des VAEs facilite un apprentissage et une compréhension plus approfondis des distributions d'images médicales complexes, les rendant outils précieux dans la recherche médicale et les diagnostics cliniques [32].

Malgré leurs forces, les auto-encodeurs font intrinsèquement face à certaines limitations. Sans contraintes ou régularisations appropriées, les auto-encodeurs haute-capacité peuvent apprendre des mappings d'identité triviaux, mémorisant essentiellement les exemples d'entraînement plutôt que d'apprendre des caractéristiques généralisables. De plus, la performance des auto-encodeurs est hautement sensible à leurs choix architecturaux et d'hyperparamètres, nécessitant un ajustement attentif pour atteindre des résultats optimaux dans des tâches spécifiques d'imagerie médicale. De plus, bien que puissants dans la capture de caractéristiques essentielles, les auto-encodeurs manquent typiquement de mécanismes d'interprétabilité explicites, présentant des défis dans la validation et l'adoption cliniques.

En résumé, les auto-encodeurs fournissent une fondation robuste et polyvalente pour l'apprentissage de représentations, contribuant significativement aux avancées en analyse d'images médicales. Leur évolution continuelle, particulièrement par l'intégration de stratégies de régularisation nouvelles et la modélisation probabiliste, souligne leur potentiel dans le traitement des défis complexes d'imagerie clinique, malgré les domaines restants pour amélioration et les efforts de recherche en cours.

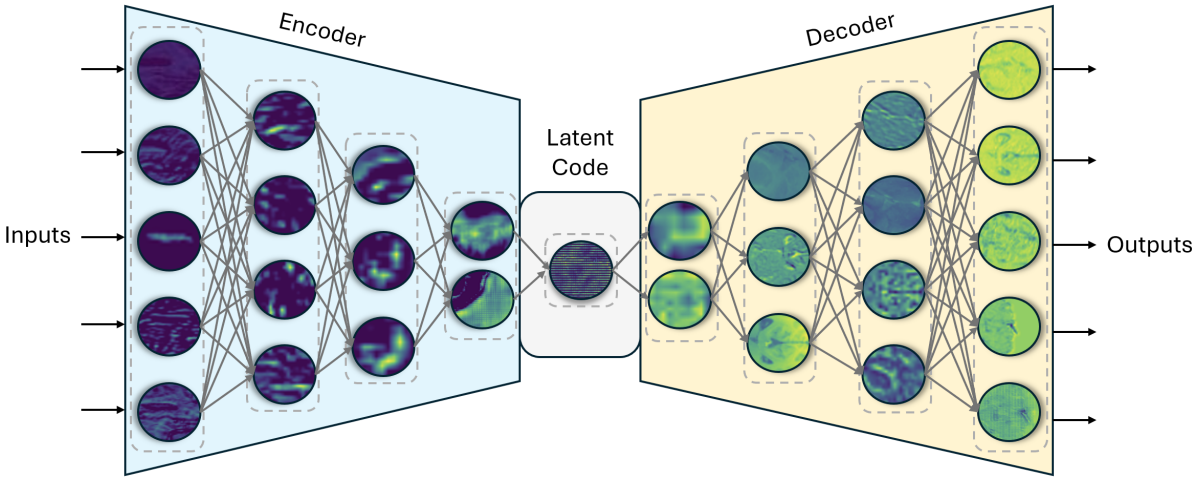

FIGURE 2.9 – Structure générale d'un Auto-encodeur illustrant le paradigme encodeur-décodeur.

2.4.6 Défis et Considérations dans l'Apprentissage Profond pour l'Analyse d'Images Médicales

L'apprentissage profond a significativement fait progresser les capacités de l'analyse d'images médicales, pourtant plusieurs défis et considérations inhérents doivent être traités pour que ces méthodes soient efficacement et sûrement intégrées dans la pratique clinique. Un défi proéminent en analyse d'images médicales est la disponibilité limitée de données médicales annotées. Les annotations de haute qualité nécessitent typiquement une expertise de domaine significative, sont chronophages et coûteuses à obtenir, résultant souvent en ensembles de données petits et déséquilibrés qui représentent inadéquatement la variabilité rencontrée dans les contextes cliniques. Par conséquent, les modèles entraînés sur de tels ensembles de données restreints exhibent fréquemment une généralisation pauvre quand déployés dans des environnements du monde réel [33]. De plus, les modèles d'apprentissage profond doivent exhiber une généralisabilité robuste entre diverses modalités d'imagerie, qui diffèrent significativement en résolution spatiale, caractéristiques de contraste et paramètres d'acquisition. L'imagerie médicale englobe une large gamme de modalités, telles que l'Imagerie par Résonance Magnétique (IRM), la Tomodensitométrie (CT), l'échographie et les lames histopathologiques, chacune présentant des physiques d'imagerie uniques et des représentations de caractéristiques différentes. Les modèles spécifiquement entraînés sur une modalité peinent souvent à généraliser à une autre, nécessitant des stratégies innovantes telles que l'adaptation de domaine, le transfer learning, ou les techniques d'intégration multi-modale pour assurer une performance consistante entre différentes plateformes d'imagerie [33].

Les contraintes de calcul représentent une autre considération critique, particulièrement concernant le déploiement clinique. Les environnements de soins de santé ont souvent des exigences strictes pour l'efficacité de calcul et les réponses à faible latence,

particulièrement dans les contextes d'urgence ou à ressources limitées. Malgré les avancées dans le matériel de calcul, de nombreuses architectures sophistiquées d'apprentissage profond demeurent exigeantes en calcul, rendant difficile leur déploiement en temps réel. Traiter ces limitations implique d'optimiser la complexité du modèle, d'utiliser des architectures efficaces en calcul telles que les réseaux convolutionnels légers, ou les modèles state-space, et d'exploiter les technologies d'accélération matérielle pour atteindre des temps d'inférence cliniquement viables [33].

Les considérations éthiques et l'interprétabilité sont également vitales en analyse d'images médicales. Les modèles d'apprentissage profond, particulièrement les réseaux de neurones convolutionnels et les modèles basés sur transformer, fonctionnent typiquement comme des "boîtes noires", fournissant des insights limités dans les processus de prise de décision qui sous-tendent leurs prédictions. Cette opacité pose des défis éthiques et pratiques substantiels, incluant la responsabilité pour les décisions diagnostiques et maintenir la confiance des patients. Les mécanismes d'interprétabilité et d'explicabilité, tels que la cartographie de saillance et la visualisation d'attention, sont devenus essentiels pour élucider les prédictions de modèle et faciliter leur acceptation par les parties prenantes cliniques. Les modèles transparents non seulement construisent la confiance parmi les professionnels de santé mais permettent aussi la détection de biais potentiels ou de décisions erronées, améliorant significativement la sécurité des patients et la fiabilité clinique [34].

En résumé, bien que l'apprentissage profond offre un potentiel transformateur pour l'analyse d'images médicales, le traitement de plusieurs défis critiques demeure impératif : la limitation des ensembles de données annotées, la généralisabilité entre modalités, les contraintes de calcul, ainsi que les considérations éthiques et d'interprétabilité. Surmonter ces obstacles requiert une recherche continue, une collaboration interdisciplinaire et des innovations méthodologiques rigoureuses, pavant ainsi la voie vers des solutions diagnostiques fiables, éthiques et cliniquement bénéfiques fondées sur l'intelligence artificielle.

2.4.7 Conclusion

Ce chapitre a fourni une exploration compréhensive des principes d'apprentissage profond fondamentaux et avancés qui constituent les bases théoriques nécessaires pour comprendre les développements ultérieurs de cette thèse. La discussion a commencé en établissant les concepts essentiels de l'apprentissage profond, mettant l'accent sur l'évolution depuis les méthodes d'apprentissage automatique traditionnelles vers des modèles de réseaux de neurones plus sophistiqués capables de gérer des données médicales complexes.

Un examen détaillé des architectures, incluant les CNNs, Vision Transformers, et modèles State-Space, a illustré les avancées progressives dans l'apprentissage profond, soulignant particulièrement leurs capacités dans la capture d'information spatiale et contextuelle critique pour le diagnostic médical. Les mécanismes d'attention ont été pré-

sentés, soulignant leur rôle dans l'amélioration de la précision diagnostique en dirigeant efficacement les ressources de calcul et en améliorant l'interprétabilité. L'importance des auto-encodeurs dans le traitement des défis pratiques tels que les données annotées limitées a également été mise en évidence. Ces méthodologies contribuent significativement à construire des modèles robustes, efficaces et précis, même en présence de rareté de données.

Le chapitre a aussi traité les défis clés rencontrés dans le déploiement de l'apprentissage profond dans les contextes cliniques, incluant la variabilité des données, les demandes de calcul et les considérations éthiques. Cette compréhension des fondements théoriques et des défis pratiques est essentielle pour contextualiser et évaluer de manière critique les approches existantes dans la littérature ainsi que pour comprendre les innovations architecturales présentées dans les contributions de cette thèse.

Avec cette fondation conceptuelle établie, le chapitre suivant présentera une revue approfondie de l'état de l'art en segmentation d'images médicales, permettant de situer les contributions de cette recherche au sein des tendances actuelles et d'identifier les lacunes qui motivent les développements méthodologiques proposés dans les chapitres ultérieurs.

Chapitre 3

Revue de Littérature en Segmentation d'Images Médicales

3.1 Vue d'Ensemble

Étant donné l'importance de la segmentation d'images médicales, une recherche substantielle a été dédiée au développement de méthodologies de segmentation précises, efficaces et généralisables. Ce chapitre présente une revue approfondie des méthodologies évolutives en segmentation d'images médicales, commençant par les approches traditionnelles et progressant vers les paradigmes modernes d'apprentissage profond. Initialement, nous discutons les techniques de segmentation classiques qui s'appuient sur les principes fondamentaux du traitement d'images et les méthodes traditionnelles de reconnaissance de motifs. Par la suite, nous approfondissons les architectures avancées d'apprentissage profond, soulignant leur évolution progressive depuis les CNNs vers les modèles basés sur Transformer, les architectures hybrides CNN-Transformer, et les SSMs émergents représentés par les frameworks Vision Mamba. De plus, nous explorons les innovations récentes dans les approches de segmentation basées sur les prompts et l'universalité, soulignant leur potentiel à surmonter les limitations de spécificité modale et les contraintes d'annotation. Pour synthétiser ces développements, une analyse comparative compréhensive est conduite, discutant les forces, limitations et applicabilité de chaque méthodologie. Cette analyse permet d'identifier les lacunes critiques dans la recherche existante et de délimiter les directions prometteuses pour l'investigation future.

3.2 Méthodes Traditionnelles

Les méthodes traditionnelles de segmentation d'images médicales ont longtemps été fondamentales dans l'analyse clinique, exploitant les techniques de traitement d'images et les méthodes classiques de reconnaissance de motifs pour partitionner les images médicales en régions diagnostiquement significatives. Malgré les avancées rapides intro-

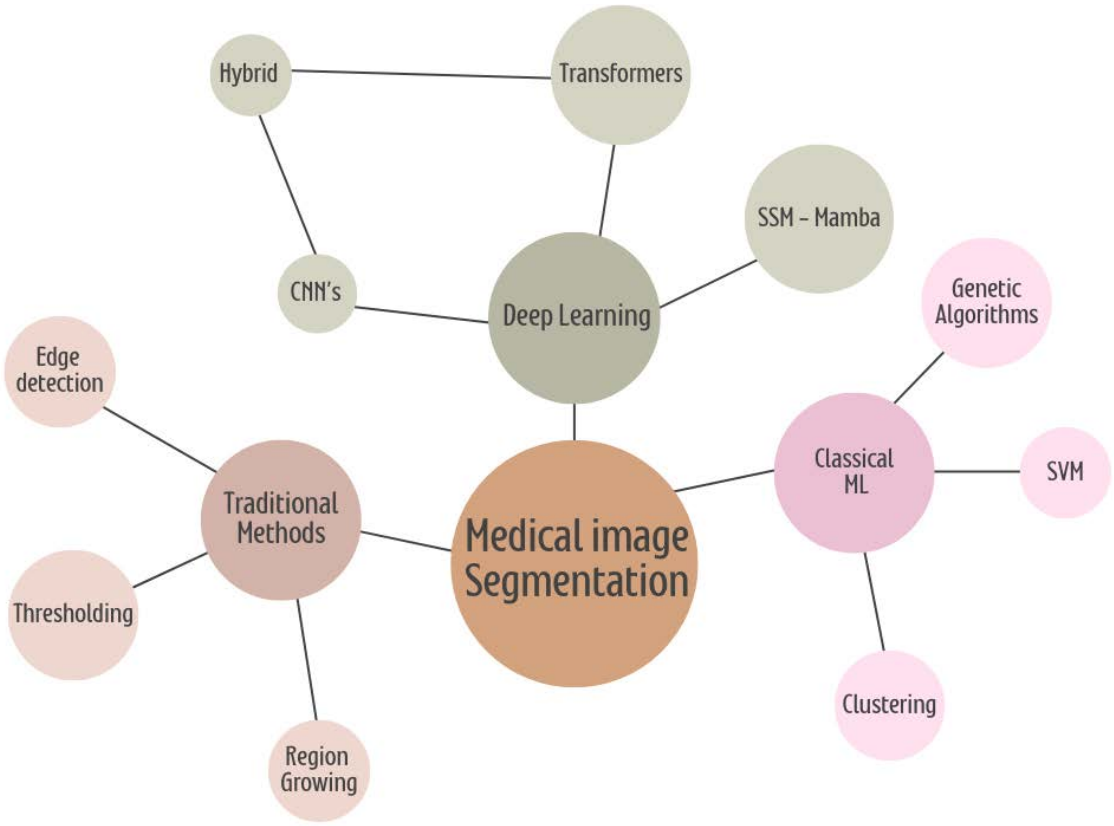

FIGURE 3.1 – Vue d'ensemble taxonomique des méthodologies de segmentation d'images médicales.

duites par l'apprentissage profond, ces approches classiques demeurent essentielles en raison de leur interprétabilité, efficacité de calcul et fiabilité dans des contextes cliniques spécifiques.

Parmi les approches les plus fondamentales, les techniques de seuillage qui classifient les pixels basés sur des seuils d'intensité. Les méthodes de seuillage adaptatif établissent dynamiquement les seuils basés sur les caractéristiques d'intensité locales, améliorant significativement la précision de segmentation dans les images avec variations d'intensité spatiales [35]. S'appuyant sur ces principes, les méthodes basées sur les régions étendent l'approche de seuillage en regroupant les pixels en zones homogènes en utilisant des critères tels que la similarité d'intensité ou la cohérence statistique. Notamment, les approches de croissance de région adaptative raffinent itérativement la segmentation basée sur les propriétés d'intensité locales, les rendant particulièrement efficaces pour segmenter les structures anatomiques avec des distributions d'intensité bien définies, telles que l'IRM cardiaque et l'imagerie cérébrale [36, 37, 38].

Complémentairement aux approches basées sur les régions, les méthodes basées sur les contours segmentent les structures anatomiques en identifiant les discontinuités d'intensité ou les frontières au sein des images. Les techniques comme la détection de contours multi-résolution ont été utilisées avec succès dans la segmentation d'images IRM cérébrales, fournissant une information de frontière précise et détaillée. Bien qu'intrinsèquement susceptibles au bruit et aux contours fragmentés, les méthodes basées sur les contours continuent de délivrer des résultats précis lorsqu'elles sont combinées avec des stratégies de prétraitement telles que le filtrage anisotrope et les algorithmes robustes de liaison de contours [39, 40].

Dans une perspective non supervisée, les approches basées sur le clustering classifient les pixels en groupes basés sur leurs similarités d'intensité sans s'appuyer sur des étiquettes préalables, excellant ainsi dans les scénarios de segmentation non supervisée. L'algorithme Fuzzy C-Means (FCM) est notamment répandu en imagerie médicale en raison de son efficacité dans la gestion de l'incertitude et des effets de volume partiel [41, 42]. Les variations avancées, incluant FCM kernelized et FCM spatialement contraint, ont incorporé des informations a priori spatiales et d'intensité pour améliorer la robustesse contre le bruit et l'inhomogénéité d'intensité communément rencontrés dans les scans IRM [43, 44, 45]. De plus, les méthodes combinant de multiples kernels améliorent davantage la performance de segmentation en intégrant adaptivement diverses représentations de caractéristiques [46]. Les approches de clustering telles que k-means, lorsqu'elles sont augmentées avec des techniques de post-traitement comme l'algorithme watershed amélioré, atténuent aussi les problèmes communs de sur-segmentation et raffinent la précision des frontières [47].

Adoptant une approche guidée par la connaissance anatomique, les méthodes de segmentation basées sur atlas utilisent des templates anatomiques comme références pour segmenter les images médicales. Ces approches alignent les images spécifiques aux patients avec des atlas standardisés, propageant les étiquettes d'atlas sur de nouvelles images par des processus de recalage. Les innovations telles que les level sets compétitifs et les mécanismes de contrôle flou améliorent significativement la précision de la segmentation basée sur atlas, incorporant efficacement les connaissances a priori spatiales et d'intensité [48, 49, 50].

Parallèlement aux approches basées sur atlas, les méthodes de segmentation basées sur les graphes traitent les tâches de segmentation comme des problèmes de partitionnement de graphes, utilisant des algorithmes de théorie des graphes tels que les arbres couvrants minimaux, la fusion de régions basée sur des comparaisons par paires, et les formulations graph-cut. Ces méthodes ont démontré une performance robuste dans la segmentation de structures complexes telles que l'IRM de prostate, PET, images cellulaires microscopiques, et anatomie cardiaque [51, 52, 53, 54, 55]. En intégrant la connectivité floue dans les méthodologies graph-cut, l'algorithme fuzzy-cuts enrichit davantage les capacités de segmentation en tenant compte de l'incertitude et de la connaissance anatomique partielle [56].

Offrant une flexibilité géométrique supérieure, les modèles déformables capturent les frontières anatomiques complexes par des frameworks adaptatifs. Les contours actifs paramétriques (snakes) et les méthodes géométriques level set font évoluer adaptivement les contours ou surfaces vers les frontières d'objets guidées par les caractéristiques d'images et contraintes. Les versions améliorées de ces méthodes, qui intègrent les connaissances a priori de forme, le remaillage dynamique, et les représentations médiales déformables (M-reps), améliorent significativement la précision de segmentation, particulièrement pour les structures avec des contours ambigus ou manquants dans les modalités d'imagerie médicale telles que l'IRM cardiaque et l'échographie prostatique [57, 58, 59].

Dans une perspective probabiliste, les méthodes statistiques incorporent les distributions d'intensité et les connaissances a priori anatomiques dans les frameworks de segmentation, modélisant les variations d'intensité locales et les dépendances spatiales. Les approches bayésiennes de croissance de région gèrent efficacement le bruit d'image et les inhomogénéités d'intensité, fournissant une segmentation robuste et précise [60]. De même, les approches probabilistes basées sur atlas, utilisant les algorithmes expectation-maximization (EM) et les Markov Random Fields (MRFs), offrent une segmentation temporellement consistante et précise pour les scénarios d'imagerie cardiaque dynamique [61, 62].

Au niveau du traitement géométrique, les opérations morphologiques manipulent les structures géométriques d'images par des opérations basiques telles que la dilatation, l'érosion, l'ouverture, et la fermeture, facilitant la suppression de bruit, la fermeture d'espaces, et la séparation de structures adjacentes. Ces méthodes jouent un rôle significatif dans les étapes de prétraitement et post-traitement des pipelines de segmentation. Par exemple, la reconstruction morphologique combinée avec la détection de ligne centrale extrait efficacement les réseaux de vaisseaux sanguins rétiniens, tandis que les approches morphologiques segmentent précisément les arbres artériels coronaires à partir d'images angiographiques, améliorant significativement la précision de segmentation et l'interprétabilité clinique [63, 64].

Enrichissant ces approches traditionnelles, les méthodologies d'apprentissage automatique classiques, incluant les machines à vecteurs de support (SVM), la factorisation matricielle non négative (NMF), et les algorithmes génétiques (GAs), ont extensivement supporté la segmentation d'images médicales en fournissant des capacités robustes d'extraction de caractéristiques et de classification. Les techniques combinant NMF avec SVM, par exemple, identifient efficacement les caractéristiques pertinentes des données d'imagerie cérébrale fonctionnelle, atteignant une haute précision diagnostique dans les applications de neuroimagerie [65]. De même, les algorithmes génétiques, connus pour leur robustesse et flexibilité, gèrent efficacement les tâches de segmentation complexes impliquant des images bruyantes et des frontières anatomiques complexes [66].

Collectivement, les méthodes de segmentation traditionnelles représentent des solutions robustes, interprétables et efficaces en calcul qui continuent de former la pierre angulaire de l'analyse d'images médicales. Bien qu'étant de plus en plus supplémentées par les approches modernes d'apprentissage profond, ces méthodes classiques fournissent

des connaissances fondamentales et une performance robuste dans des applications cliniques spécifiques. Leur intégration avec les techniques de calcul contemporaines et les paradigmes d'apprentissage avancés offre des voies prometteuses pour la recherche future et l'innovation en segmentation d'images médicales.

3.3 Architectures Basées sur CNN

Les Réseaux de Neurones Convolutionnels ont émergé comme l'une des méthodologies les plus influentes pour la segmentation d'images médicales en raison de leur capacité supérieure à apprendre des représentations de caractéristiques hiérarchiques et discriminatives directement à partir des données. Leurs forces inhérentes, telles que l'invariance par translation, l'extraction de caractéristiques hiérarchiques, et la représentation spatiale efficace, ont significativement fait progresser l'état de l'art dans les tâches de segmentation.

Dans cette perspective d'amélioration architecturale, plusieurs études se sont concentrées sur l'enrichissement des architectures CNN pour atteindre une meilleure précision, particulièrement par des améliorations de l'architecture U-Net classique. Par exemple, des variations comme U-Net++ et DenseUNet ont été proposées, intégrant des connexions denses et des structures imbriquées pour enrichir les représentations de caractéristiques et fournir une performance robuste entre différentes modalités d'imagerie médicale [67, 68].

Parallèlement à ces avancées structurelles, les mécanismes d'attention ont été extensivement explorés au sein des frameworks CNN pour améliorer la performance de segmentation. Des modèles tels qu'Attention U-Net, Squeeze-and-Excitation U-Net, et autres ont intégré avec succès des modules d'attention dans les CNNs pour raffiner les cartes de caractéristiques, permettant aux réseaux de se concentrer plus efficacement sur les structures anatomiques pertinentes et les régions des lésions [69, 70].

Complémentairement aux mécanismes d'attention, des modules convolutionnels avancés comme les convolutions dilatées et l'atrous spatial pyramid pooling (ASPP) ont été incorporés dans les architectures CNN pour améliorer les capacités d'extraction de caractéristiques multi-échelles. Ces approches capturent efficacement l'information contextuelle à diverses résolutions, traitant les défis posés par les lésions de différentes tailles et formes. Les études introduisant ces techniques convolutionnelles avancées ont démontré une précision de segmentation supérieure, particulièrement dans les cas caractérisés par des frontières de lésions irrégulières et des structures anatomiques diverses [71, 72, 73].

Face aux défis spécifiques de délimitation, d'autres efforts de recherche se sont concentrés sur la gestion des scénarios de segmentation difficiles caractérisés par des frontières indistinctes et des artefacts. Par exemple, les chercheurs ont proposé des architectures basées sur CNN qui incorporent des modules de raffinement de frontières et des mécanismes tenant compte des contours, ciblant explicitement la délimitation précise des frontières de lésions. Ces méthodes améliorent significativement la segmentation de lésions avec des contrastes subtils et des contours irréguliers, surmontant les limitations

rencontrées par les architectures CNN traditionnelles dans la gestion des frontières de lésions complexes [74, 75, 76].

Au niveau de l'optimisation de l'entraînement, plusieurs études ont introduit des fonctions de perte innovantes et des stratégies d'entraînement adaptées aux CNNs pour atténuer les défis communs de segmentation tels que le déséquilibre de classes et les petites régions cibles. Des fonctions de perte personnalisées, incluant focal loss, des variantes de Dice loss, et des formulations de perte hybrides combinant de multiples métriques, ont été proposées pour améliorer l'efficacité d'entraînement et la précision de segmentation des modèles CNN [77, 78, 79, 80].

Traitant la problématique des données annotées limitées, certaines recherches ont investigué les méthodes CNN basées sur l'apprentissage semi-supervisé et faiblement supervisé. Ces approches exploitent de petites quantités de données étiquetées aux côtés de larges ensembles de données non étiquetées, étendant efficacement l'applicabilité des CNNs dans des scénarios où l'annotation extensive n'est pas faisable. Les architectures CNN semi-supervisées ont démontré une performance encourageante dans les tâches avec des données annotées rares, soulignant leur potentiel pour un déploiement pratique dans les contextes cliniques [81, 82, 83].

Finalement, dans une perspective de déploiement pratique, les chercheurs ont proposé des architectures CNN optimisées explicitement pour l'efficacité de calcul et la segmentation en temps réel. Les modèles CNN légers, tels que les variantes d'architectures basées sur MobileNet et EfficientNet, ont démontré des résultats prometteurs dans l'atteinte d'une segmentation précise avec des demandes de calcul significativement réduites. Ces avancées sont particulièrement critiques pour le déploiement dans des environnements à ressources contraintes, améliorant l'utilité pratique et l'évolutivité des systèmes de segmentation basés sur CNN [84, 85, 86].

Globalement, ces méthodes basées sur CNN soulignent l'évolution continue des architectures convolutionnelles, mettant en évidence leur adaptabilité et robustesse dans le traitement des défis divers de segmentation au sein de l'imagerie médicale. Par les innovations en conception architecturale, mécanismes d'attention, techniques convolutionnelles avancées, fonctions de perte spécialisées, et stratégies d'apprentissage semi-supervisé, les CNNs continuent de jouer un rôle pivot dans l'avancement des capacités de segmentation d'images médicales, supportant l'amélioration des diagnostics et la prise de décision clinique.

3.4 Architectures Basées sur Transformer

Les architectures Transformer ont émergé comme des solutions puissantes en segmentation d'images médicales en raison de leur capacité à capturer les dépendances à long terme par les mécanismes self-attention. Dans cette perspective d'intégration multi-échelle, plusieurs méthodes basées sur les transformers ciblent spécifiquement l'intégration de caractéristiques multi-échelles et hiérarchiques pour améliorer la précision de segmentation. MISSFormer [87] introduit une structure encodeur-décodeur hiérar-

chique présentant le bloc ReMix-FFN, combinant efficacement les contextes globaux et locaux. De même, MS-Former [88] emploie une conception dual-branch avec des stratégies d'apprentissage auto-supervisé pour maintenir la consistance sémantique, améliorant significativement la précision de segmentation pour des structures complexes telles que les lésions cutanées et les poumons.

Face aux contraintes de calcul inhérentes aux transformers, des méthodes comme MAXFormer [89] reconfigurent le self-attention en voies locales-globales et externes, réduisant substantiellement les demandes de calcul tout en maintenant une extraction de caractéristiques robuste. De même, Slim UNETR [90] adopte des blocs transformer légers utilisant le self-attention décomposée, les rendant particulièrement adaptés aux contextes cliniques à ressources limitées. En outre, LeViT-UNet [91] intègre des encodeurs transformer multi-étapes légers dans un framework U-Net, équilibrant efficacement la vitesse et la précision de segmentation.

Au niveau des mécanismes d'attention avancés, des approches adaptées spécifiquement pour les tâches de segmentation médicale raffinent davantage la performance des transformers. H2Former [92] incorpore l'attention de canal multi-échelle hiérarchique pour modéliser efficacement les dépendances à long terme, tandis que SMAFormer [93] intègre l'attention pixel, canal, et spatiale, améliorant la segmentation de tumeurs et organes petits et de forme irrégulière.

Traitant spécifiquement les défis de délimitation des frontières, certaines architectures Transformer se distinguent par leurs approches innovantes. BATFormer [94] introduit des modules transformer locaux tenant compte des frontières et un partitionnement de fenêtre adaptatif guidé par l'entropie, améliorant significativement la précision de segmentation avec une charge de calcul réduite. APFormer [95] emploie l'élagage adaptatif combiné avec l'attention auto-régularisée, offrant une haute précision avec une complexité de modèle dramatiquement réduite.

Enrichissant ces approches par des caractéristiques inspirées de la radiomique, des innovations exploitant les filtres avancés ont davantage amélioré les architectures basées sur transformer. GLoG-CSUnet [96] intègre des filtres Gabor adaptatifs et Laplacien de Gaussien (LoG) dans les architectures transformer, augmentant significativement la précision de segmentation en capturant les textures et frontières détaillées. De même, PCCTrans [97] combine des modules convolution-transformer parallèles pour optimiser efficacement la fusion des détails spatiaux et du contexte global.

S'étendant au domaine tridimensionnel, les variantes transformer spécialisées adaptées pour les scénarios de segmentation 3D démontrent également des améliorations notables. SwinMM [98] utilise l'apprentissage auto-supervisé multi-vues pour améliorer la précision de segmentation et l'efficacité des données en imagerie médicale volumétrique, tandis que SegFormer3D [99] offre une conception transformer hiérarchique efficace, réduisant grandement la charge de calcul sans compromettre la précision.

Collectivement, ces développements soulignent la polyvalence et l'efficacité du transformer en segmentation d'images médicales, mettant en évidence ses forces uniques dans la modélisation des relations spatiales complexes et l'information contextuelle multi-

échelle. Leur raffinement continué souligne leur rôle crucial dans l'avancement des méthodologies de segmentation pour les applications cliniques.

3.5 Architectures Hybrides

Les architectures hybrides, intégrant les CNNs et ViTs ou SSMs, représentent une avancée significative en segmentation d'images médicales en combinant efficacement les forces complémentaires de chaque approche. Comme discuté précédemment, les CNNs excellent dans la capture de détails locaux par l'extraction de caractéristiques spatialement invariantes, tandis que les Transformers sont particulièrement efficaces dans la modélisation des dépendances contextuelles à long terme. En fusionnant ces capacités distinctes, les modèles hybrides fournissent une performance de segmentation améliorée, particulièrement dans les tâches d'imagerie clinique complexes requérant une délimitation précise de structures anatomiques complexes et une compréhension contextuelle approfondie.

Illustrant les avantages de cette intégration, plusieurs méthodes ont démontré des améliorations significatives. Par exemple, ScribFormer [100] emploie une architecture triple-branch, combinant efficacement une branche CNN dédiée à l'extraction de caractéristiques locales avec une branche Transformer adaptée pour la représentation de contexte global, complétée par une branche attention-guided class activation map (ACAM). Cette conception excelle particulièrement dans les scénarios où les annotations sont limitées (par exemple, segmentation scribble-supervised), améliorant significativement la précision de segmentation en équilibrant le détail local et la cohérence sémantique globale.

Dans une perspective de gestion d'incertitude, UCTNet [101] introduit une architecture hybride guidée par l'incertitude, dirigeant les modules Transformer spécifiquement vers les régions identifiées par l'estimation d'incertitude CNN. Cette fusion sélective assure que les ressources de calcul sont optimalement utilisées, améliorant la précision de segmentation et la stabilité de convergence, particulièrement dans les conditions d'imagerie médicale difficiles.

Approfondissant l'intégration architecturale, le framework MCBTNet [102] intègre les CNNs et Transformers au sein d'une architecture encodeur-décodeur unifiée, employant un bi-level routing attention transformer qui capture dynamiquement le contexte global et fusionne efficacement les caractéristiques multi-échelles via un mécanisme d'attention Frequency-Channel-Spatial. De même, CTC Net [103] exploite des encodeurs duaux, des backbones CNN et Transformer, intégrés par un Cross-domain Fusion Block (CFB), atteignant une segmentation supérieure en fusionnant de manière harmonieuse les caractéristiques locales détaillées et les contextes globaux étendus.

Poursuivant cette évolution, une autre avancée remarquable est présentée par GL-MambaNet [104], qui raffine davantage l'intégration hybride par des modules de fusion de caractéristiques global-local spécialisés et des améliorations d'attention multi-étapes,

atteignant une précision de segmentation remarquable sur des ensembles de données complexes comme GlaS et PH2.

En résumé, les architectures hybrides traitent efficacement les limitations des modèles autonomes en exploitant leurs forces respectives de manière complémentaire. Cette intégration stratégique permet une gestion robuste à la fois de l'information locale et globale, rendant les méthodes hybrides particulièrement avantageuses pour les scénarios de segmentation complexes.

3.6 Modèles State-Space (Vision Mamba)

Les SSMs, particulièrement les architectures Vision Mamba, ont récemment émergé comme alternatives efficaces aux méthodes basées sur CNN et Transformer, offrant une complexité de calcul linéaire dans la modélisation des dépendances à long terme. Les architectures Vision Mamba capitalisent sur les calculs state-space structurés, fournissant une modélisation de contexte global robuste cruciale pour la segmentation d'images médicales précise sans encourir les lourdes charges de calcul typiques des Transformers.

Posant les fondements de cette approche, les contributions initiales, telles que Vision Mamba [3], ont introduit des couches state-space à complexité linéaire adaptées explicitement pour la segmentation d'images médicales, réduisant efficacement la surcharge de calcul tout en modélisant précisément les dépendances globales. Étendant ce travail fondamental, des approches hiérarchiques intégrant des caractéristiques multi-échelles ont été développées. Notamment, U-Mamba [105] et MambaUNet [106] ont incorporé des structures hiérarchiques au sein de frameworks en forme de U, utilisant des couches state-space et des mécanismes d'attention selective kernel. Ces modèles gèrent efficacement les scénarios de segmentation impliquant un faible contraste, des arrière-plans bruités, et des frontières de lésions diffuses en agrégeant efficacement le contexte multi-échelle et les détails de frontière locaux.

Renforçant ces capacités par des composants convolutionnels, des améliorations supplémentaires impliquent l'intégration de modules convolutionnels pour compléter la modélisation de contexte global de Vision Mamba. Des méthodes comme VMAXL-UNet [107] et CaVMamba [108] emploient des augmentations convolutionnelles, combinant la capacité d'extraction de caractéristiques locales des CNNs avec la modélisation state-space. CNS-UNet [109] utilise une stratégie dual-encoder et des portes d'attention légères pour équilibrer l'information globale et locale, atteignant efficacement une performance de segmentation robuste.

Dans une perspective d'optimisation de l'efficacité, des variantes légères telles que MCI-Net [110] et EM-Net [111] ont été développées pour réduire les exigences de calcul. MCI-Net réduit dramatiquement la complexité de calcul par la modélisation de séquence linéaire, la rendant particulièrement adaptée aux environnements à ressources contraintes. EM-Net et FMamba [112] intègrent l'apprentissage dans le domaine fréquentiel pour améliorer l'extraction de caractéristiques entre les échelles, améliorant significativement la précision et la vitesse d'entraînement.

Adoptant une approche d'intégration hybride, les architectures combinant Vision Mamba avec des composants CNN démontrent d'excellentes capacités de segmentation. MambaVesselNet [113] intègre des caractéristiques locales pilotées par CNN et des couches bottleneck state-space, excellant dans la segmentation cérébrovasculaire 3D complexe. De même, MSVM-UNet [114] traite efficacement les défis de représentation multi-échelle et de sensibilité directionnelle utilisant des convolutions spécialisées au sein des blocs state-space, résultant en une performance supérieure dans diverses tâches de segmentation.

Étendant l'applicabilité au-delà de la segmentation supervisée complète, les méthodes interactives et semi-supervisées basées sur Vision Mamba explorent des paradigmes d'apprentissage alternatifs. ESM-Net [115] introduit la convolution spatialement augmentée pour la segmentation d'images médicales interactive, rationalisant les interactions utilisateur et améliorant la qualité de segmentation. Semi-Mamba-UNet [116] incorpore l'apprentissage contrastif au niveau pixel, démontrant des gains substantiels dans les scénarios avec des données étiquetées limitées, le rendant hautement pratique pour l'usage clinique.

Au niveau de l'agrégation contextuelle avancée, les modules spécialisés se concentrant sur l'agrégation de contexte et la fusion de caractéristiques améliorent également les frameworks Vision Mamba. PHMamba [117] présente une approche enrichie contextuellement, améliorant significativement l'extraction d'information contextuelle pour les tâches de segmentation cardiaque et multi-organes. Polyp-Mamba [118] incorpore des modules sémantiques scale-aware et des techniques d'injection sémantique globale, établissant de nouveaux benchmarks de performance pour la segmentation de polypes. En outre, SK-VM++ [119] redéfinit les connexions skip avec des modules basés sur Mamba, améliorant la fusion de caractéristiques entre niveaux élevés et bas et atteignant des améliorations de segmentation considérables, tandis que RM-UNet [120] optimise l'efficacité des paramètres par les blocs Residual Visual State Space et les modules state-space rotationnels.

Démontrant une polyvalence supplémentaire, les frameworks multi-tâches utilisant Vision Mamba ont exploré des applications élargies. UBGM [121] introduit l'apprentissage multi-tâches bidirectionnel guidé par l'incertitude, combinant efficacement les tâches de classification et de segmentation pour améliorer significativement la fiabilité diagnostique. Les adaptations aux modèles à usage général comme le Segment Anything Model (SAM) [122] soulignent davantage la capacité de Vision Mamba en segmentation médicale. Tri-Plane Mamba (TP-Mamba) [123] et MambaSAM [124] adaptent efficacement SAM pour la segmentation médicale 3D utilisant des adaptateurs spatiaux légers et l'attention cross-branch, atteignant une précision de segmentation remarquable même avec des données d'entraînement minimales.

Globalement, les architectures basées sur Vision Mamba représentent des avancées méthodologiques substantielles en segmentation d'images médicales, fournissant des solutions efficaces, précises et polyvalentes capables de traiter efficacement divers défis d'imagerie clinique.

3.7 Méthodes Basées sur Prompts et Universelles

Les méthodes de segmentation basées sur prompts et universelles introduisent un paradigme transformateur en segmentation d'images médicales en utilisant des prompts textuels et visuels pour guider les prédictions de modèles. Contrairement aux méthodes traditionnelles et d'apprentissage profond, qui nécessitent typiquement un ré-entraînement extensif ou un fine-tuning pour chaque nouvelle tâche de segmentation, ces approches guidées par prompts visent l'adaptabilité et la généralisation entre diverses structures anatomiques et modalités d'imagerie. Cette flexibilité rationalise significativement leur intégration dans les workflows cliniques et réduit la dépendance aux larges ensembles de données annotées.

Exploitant la synergie vision-langage, plusieurs modèles ont exploité les prompts textuels dérivés des modèles vision-language pour guider efficacement la segmentation. Le CLIP-Driven Universal Model [125] utilise les embeddings sémantiques du Contrastive Language-Image Pre-training (CLIP) pour capturer le contexte anatomique, permettant une segmentation précise d'organes multiples et tumeurs avec une généralisation améliorée. De même, le Universal Model présenté par [126] incorpore la génération de paramètres guidée par le langage, facilitant la segmentation de structures diverses, gérant efficacement les annotations partielles, et s'étendant harmonieusement aux classes nouvelles. Ces modèles ont atteint des performances benchmark sur des ensembles de données significatifs tels que le Medical Segmentation Decathlon (MSD) et BTCV, soulignant leur efficacité pratique.

Renforçant cette synergie vision-langage, MedCLIP-SAM [127] combine CLIP avec SAM, démontrant des capacités de segmentation zero-shot et faiblement supervisées robustes entre diverses modalités, incluant échographie, IRM, et images radiographiques. De plus, TP-DRSeg [128] emploie des prompts textuels explicites enrichis avec des connaissances a priori médicales spécifiques pour améliorer la précision de segmentation des lésions subtiles de rétinopathie diabétique, démontrant comment la connaissance spécifique au domaine embarquée au sein des prompts linguistiques peut significativement améliorer la fiabilité clinique.

Favorisant l'interaction clinicien-machine, les méthodes de segmentation interactive pilotées par des prompts intuitifs ont davantage étendu l'applicabilité clinique des approches basées sur les prompts. ScribblePrompt [129] permet aux cliniciens de segmenter interactivement des images biomédicales par des annotations simples telles que scribbles ou boîtes englobantes (Bounding Box), réduisant substantiellement l'effort d'étiquetage manuel tout en atteignant une précision supérieure comparée aux méthodes interactives traditionnelles. L'efficacité de cette méthode dans les contextes cliniques souligne le potentiel des approches interactives guidées par prompts pour traiter efficacement la rareté d'annotations.

Dans une perspective d'optimisation de l'efficacité paramétrique, les stratégies de prompt adaptatif ont également été développées pour améliorer l'efficacité et la performance dans les tâches de segmentation avec des données annotées limitées. Par exemple,

PUNETR [130] introduit le prompt-tuning parameter-efficient, intégrant des prompts class-specific dans une architecture UNETR pré-entraînée, atteignant une haute performance de segmentation avec des ajustements de paramètres minimaux. Cette approche ferme significativement l'écart entre les modèles prompt-tuned et les homologues fully fine-tuned, démontrant son utilité clinique sous des conditions de données contraintes.

S'étendant au domaine tridimensionnel, les méthodes basées sur les prompts ont également été adaptées pour les tâches de segmentation 3D complexes. Le 3DSAM-adapter [131] étend efficacement le framework SAM des contextes 2D vers 3D, maintenant une excellente performance de segmentation entre les ensembles de données médicales volumétriques avec un fine-tuning minimal. De même, ProMISe [132] emploie des adaptateurs légers et des prompts single-point pour améliorer les modèles 2D pré-entraînés pour une segmentation 3D précise, capturant efficacement les contextes spatiaux complexes critiques pour des tâches telles que la délimitation tumorale.

Traitant les défis d'adaptation multi-modale, les approches guidées par prompts ont efficacement traité les défis dans les scénarios d'adaptation multi-modale et de domaine. MAVP [133] utilise des prompts visuels modality-aware pour gérer les données d'imagerie multi-modale incomplètes, maintenant efficacement une performance de segmentation robuste. FVP [134] introduit le Fourier visual prompting pour l'adaptation de domaine non supervisée source-free, permettant aux modèles d'atteindre une haute performance dans de nouveaux domaines d'imagerie sans entraînement supplémentaire, soulignant davantage la flexibilité et l'efficacité de données des modèles guidés par les prompts.

Intégrant des connaissances anatomiques détaillées, les architectures spécialisées intégrant des prompts anatomiques et pathologiques détaillés ont également émergé. PC-Net [135] emploie des prompts anatomiques hiérarchiques améliorés par les embeddings CLIP, renforçant significativement la robustesse et généralisation de segmentation. De même, UniSeg [136] emploie des prompts universels pour guider dynamiquement la segmentation entre diverses régions anatomiques, atteignant une performance supérieure entre de multiples benchmarks de segmentation médicale.

Finalement, explorant les approches génératives, les stratégies de modélisation générative, guidées par des prompts textuels et visuels, ont davantage démontré le potentiel des approches basées sur les prompts. Les modèles de diffusion [137] exploitent efficacement les prompts spécifiques aux lésions pour générer des images dermoscopiques de haute qualité, améliorant simultanément la précision de segmentation et le réalisme des images.

En résumé, les méthodes de segmentation basées sur prompts et universelles offrent une adaptabilité, efficacité et généralisation sans précédent, les positionnant comme outils prometteurs pour la pratique clinique. Cependant, une recherche supplémentaire dans l'optimisation de prompts, l'intégration spécifique au domaine, et l'harmonisation de workflow demeure essentielle pour réaliser leur plein potentiel.

3.8 Analyse Comparative

La progression des architectures de segmentation d'images médicales, depuis les heuristiques traditionnelles vers les paradigmes modernes d'apprentissage profond, reflète un effort continu pour combler l'écart entre la faisabilité de calcul, la complexité anatomique, et la fiabilité clinique. Cette évolution n'est pas seulement technologique, mais conceptuelle ; chaque changement méthodologique représente une redéfinition de comment nous modélisons, comprenons, et interprétons les images médicales.

Posant les fondements historiques, les techniques de segmentation traditionnelles, bien que maintenant limitées en portée, ont fourni des éléments fondamentaux essentiels qui ont façonné notre compréhension de la représentation d'images. Leur interprétabilité, leurs demandes en ressources faibles, et leur comportement déterministe conservent encore de la valeur dans des scénarios cliniques contraints. Pourtant, leur incapacité à s'adapter aux distributions de données hétérogènes, frontières subtiles, et variations inter-modalités les ont rendues inadéquates pour les besoins diagnostiques modernes.

Marquant une rupture paradigmatique, les CNNs ont représenté une avancée majeure en permettant l'apprentissage de caractéristiques hiérarchiques guidées par les données. Les architectures comme U-Net ont apporté une viabilité pratique à la segmentation automatisée. Leur force réside dans la reconnaissance de motifs locaux et la robustesse à la variabilité anatomique. Cependant, la conception fondamentale des CNNs, construite sur la localité et l'invariance par translation, contraint inévitablement leur capacité à modéliser le contexte anatomique global. Les améliorations comme les convolutions dilatées et les mécanismes d'attention traitent partiellement ceci, mais souvent par des extensions architecturales plutôt qu'un changement structurel.

Redéfinissant la modélisation contextuelle, les modèles basés sur Transformer, particulièrement les ViTs, ont révolutionné la représentation de caractéristiques en permettant aux modèles de considérer globalement l'ensemble d'une image. Ceci s'est avéré inestimable pour les anatomies avec des dépendances spatialement dispersées ou frontières ambiguës. Pourtant, leur complexité quadratique par rapport à la résolution d'entrée soulève des questions critiques concernant l'évolutivité, particulièrement en imagerie médicale 3D. De plus, leur besoin important en données pose des défis dans les domaines où les ensembles de données annotés sont rares. Les avancées récentes dans les Transformers efficaces, par l'attention sparse, l'élagage adaptatif, et la tokenisation hiérarchique, sont prometteuses, mais beaucoup de ces améliorations demeurent déconnectées, manquant de principes de conception unifiés ancrés dans les contraintes d'imagerie médicale.

En explorant des synergies complémentaires, les architectures hybrides intégrant CNNs et Transformers représentent un compromis attrayant. Elles reconnaissent la nature complémentaire des représentations locales et globales : les CNNs excellent en détail, les Transformers en contexte. Cependant, les modèles hybrides souffrent souvent de complexité architecturale et d'inefficacité de calcul. De plus, beaucoup de conceptions actuelles suivent une fusion naïve plutôt qu'une intégration méthodique tenant compte de l'anatomie. Basés sur notre analyse, nous argumentons que l'avenir des modèles hybrides

réside dans les pipelines adaptatifs aux tâches et dynamiquement reconfigurables, où les interactions local-global sont modulées basées sur le contexte anatomique et l'intention diagnostique.

Proposant une alternative efficace en calcul, Vision Mamba introduit une approche convaincante en offrant une complexité linéaire pour la modélisation de dépendance globale. Ceci traite le goulot d'étranglement majeur des architectures Transformer sans sacrifier la compréhension contextuelle. Cependant, la nature séquentielle de Mamba introduit ses propres limitations dans le raisonnement spatial lorsqu'appliquée directement aux données d'images denses. Une opportunité centrale réside dans la conception de SSMs tenant compte de l'information spatiale qui intègre les forces de convolution et de récurrence structurée dans une formulation unifiée. En outre, la modularité et l'efficacité paramétrique des modèles basés sur Mamba présentent un avantage stratégique pour le déploiement embarqué, où l'inférence en temps réel et la faible consommation énergétique sont primordiales.

Ouvrant vers des systèmes généralistes, les modèles de segmentation basés sur les prompts et l'universalité signalent un changement vers des systèmes généralistes et adaptatifs à l'utilisateur, capables de segmentation entre anatomies, modalités, et tâches avec supervision minimale. Ces modèles, inspirés par les modèles de langage et de vision à grande échelle, sont attrayants sur le plan conceptuel mais immatures sur le plan pratique. Leur succès dépend de la résolution de plusieurs défis ouverts : conception de prompt robuste, interprétabilité clinique, et intégration de connaissances a priori du domaine sans surapprentissage vers des représentations de données biaisées. Le potentiel, cependant, est substantiel, particulièrement lorsqu'associé avec des stratégies few-shot learning et d'adaptation de domaine pour supporter la variabilité du monde réel.

Synthétisant ces observations, de notre point de vue, le chemin vers l'avant ne réside pas dans la déclaration d'un paradigme supérieur à un autre, mais dans l'orchestration de leurs forces. Une architecture de segmentation moderne devrait tenir compte de l'anatomie, être adaptative au contexte, efficace en données, et interprétable par conception. Ceci implique de dépasser les pipelines statiques vers des systèmes modulaires et reconfigurables qui peuvent sélectivement activer les CNNs, Transformers, SSMs, ou composants guidés par prompts basés sur les caractéristiques de données et exigences cliniques.

En conclusion, le domaine de la segmentation d'images médicales présente aujourd'hui un paysage riche en approches diverses, mais manque d'intégration cohérente. Le véritable progrès ne résidera pas dans le développement isolé de nouvelles architectures, mais dans la convergence méthodique de paradigmes complémentaires. Cette convergence doit être guidée par quatre principes fondamentaux : premièrement, l'intégration de la connaissance anatomique dans la conception des modèles ; deuxièmement, le respect des contraintes imposées par les réalités cliniques ; troisièmement, la garantie de l'interprétabilité des prédictions ; et quatrièmement, l'optimisation de l'efficacité de calcul pour permettre un déploiement pratique.

Cette thèse s'inscrit dans cette vision en adoptant une approche double. D'une part, nous évaluons de manière critique les méthodes existantes, identifiant leurs forces et limitations respectives. D'autre part, nous proposons des architectures innovantes qui intègrent les enseignements tirés de cette analyse comparative, visant ainsi à combler les lacunes identifiées dans l'état de l'art actuel.

3.9 Conclusion

Ce chapitre a fourni une vue d'ensemble large et structurée des avancées méthodologiques en segmentation d'images médicales, depuis les approches traditionnelles basées sur des règles vers les paradigmes contemporains d'apprentissage profond. Par cette revue, il devient évident que chaque génération de modèles a traité des défis spécifiques, tels que la complexité anatomique, les limitations de calcul, ou le besoin de généralisabilité entre tâches et modalités.

Cette évolution établit le contexte pour les contributions présentées dans les chapitres suivants. S'appuyant sur les forces et traitant les limitations identifiées ici, le chapitre suivant introduit notre premier modèle proposé, enraciné dans une architecture basée sur CNN. Il est conçu pour améliorer la précision de segmentation, particulièrement dans les scénarios marqués par l'ambiguïté structurelle et les artefacts d'imagerie. Cette contribution s'inscrit dans un effort plus large visant à développer des solutions de segmentation adaptables, efficaces et cliniquement viables.

Chapitre 4

Mixture of Experts pour la Segmentation de Lésions Cutanées

4.1 Introduction aux Défis de la Segmentation de Lésions Cutanées

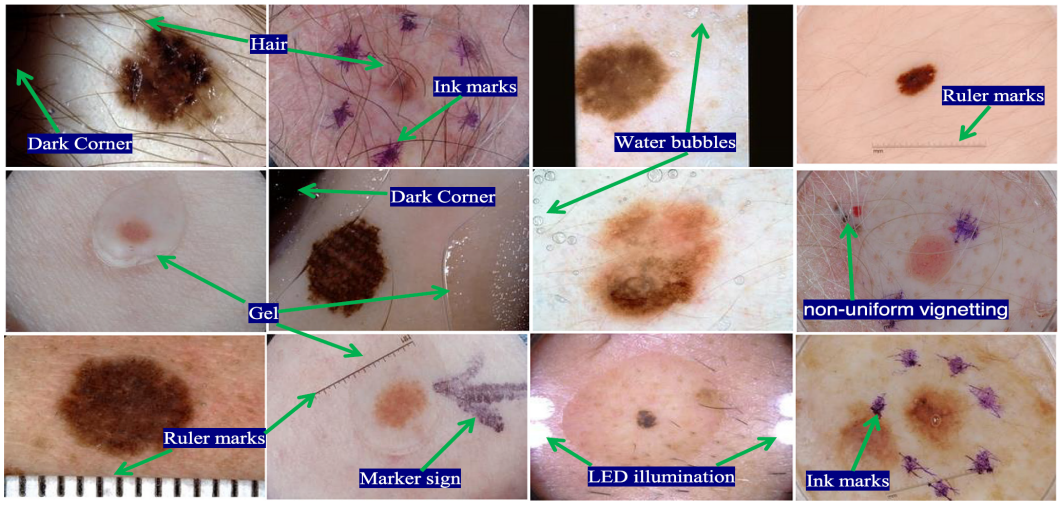

FIGURE 4.1 – Présentation pictoriale typique des images cutanées dans l'ensemble de données de test ISIC-2017 avec différentes images difficiles pour la segmentation.

La segmentation précise des lésions cutanées dans les images médicales joue un rôle pivot dans la détection précoce et le diagnostic précis des cancers de la peau, particulièrement le mélanome, qui demeure parmi les cancers les plus fréquents et diagnostiqués de plus en plus souvent dans le monde. La segmentation précoce et précise des lésions améliore significativement la précision diagnostique, guide la planification thérapeutique efficace, et facilite le suivi des patients. Malgré des progrès substantiels, la segmentation

automatisée des lésions cutanées continue de poser des défis considérables au sein de l'analyse d'images médicales.

Concernant la variabilité morphologique, un défi critique est la variabilité inhérente parmi les différents types de lésions. Les lésions cutanées exhibent une diversité substantielle dans leur apparence, caractérisée par des tailles, couleurs, formes, textures, et localisations anatomiques variées. Par exemple, certaines lésions apparaissent comme des régions distinctes et fortement pigmentées avec des frontières claires, tandis que d'autres se manifestent comme des zones subtiles à faible contraste se fondant progressivement avec la peau saine environnante (Figure 4.1). Ce large spectre de caractéristiques des lésions complique significativement les tâches de segmentation, posant des difficultés tant pour les experts humains que pour les systèmes automatisés.

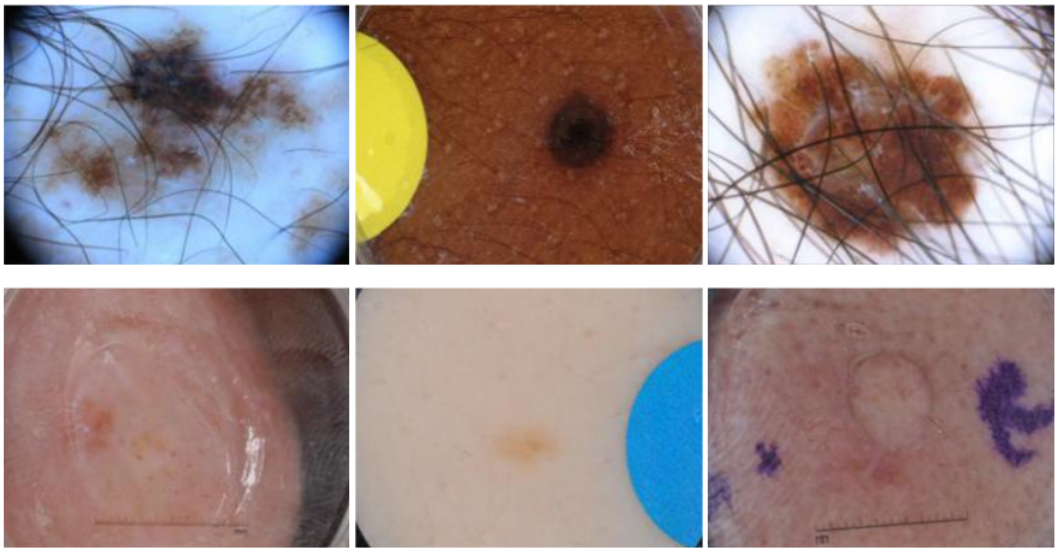

FIGURE 4.2 – Exemples de lésions cutanées avec apparences contrastées. La rangée supérieure illustre des lésions foncées avec haut contraste et frontières clairement définies, tandis que la rangée inférieure représente des lésions claires caractérisées par un faible contraste et des frontières indistinctes.

Au niveau des perturbations d'imagerie, le processus de segmentation est davantage compliqué par divers artefacts d'images et bruits, tels que cheveux, ombres, reflets, et conditions d'éclairage inconsistantes. Ces artefacts obscurcissent fréquemment les frontières des lésions ou créent de faux indices visuels, augmentant la probabilité d'erreurs de segmentation ou de classification erronée. Les variations dans la qualité d'image, provenant de différents dispositifs d'imagerie et environnements cliniques, exacerbent ces difficultés, impactant particulièrement les méthodes de segmentation traditionnelles dépendantes de caractéristiques visuelles consistantes.

Traitant spécifiquement la délimitation des contours, un défi critique supplémentaire en segmentation concerne les frontières des lésions irrégulières et indistinctes, particulièrement évidentes dans le mélanome invasif. Ces lésions exhibent souvent des contours

diffus en raison de la propagation imprévisible des cellules cancéreuses dans les tissus sains adjacents, compliquant la délimitation entre la lésion et la peau normale. La détection précise de frontières dans de tels scénarios est essentielle, influençant directement les décisions cliniques concernant le diagnostic patient et la gestion ultérieure.

Face à ces limitations méthodologiques, les approches de segmentation traditionnelles, telles que les méthodes de seuillage et de détection de contours, ont historiquement montré une efficacité limitée dans la gestion de ces défis de segmentation complexes, motivant l'adoption de méthodes avancées d'apprentissage profond, particulièrement les CNNs, qui ont significativement amélioré la performance de segmentation en exploitant les capacités d'extraction de caractéristiques hiérarchiques pour capturer efficacement l'information contextuelle locale. Cependant, les approches basées sur CNN font encore face à des limitations inhérentes, notamment leur dépendance aux champs récepteurs localisés, restreignant leur capacité à capturer les contextes globaux extensifs vitaux pour segmenter précisément les lésions avec des structures de frontières complexes. De plus, les modèles CNN nécessitent souvent des quantités substantielles de données annotées pour généraliser efficacement, présentant un autre défi étant donné la disponibilité limitée d'ensembles de données de lésions cutanées étiquetées par des experts.

En explorant des approches hybrides avancées, la recherche récente a de plus en plus favorisé les approches de segmentation hybrides qui intègrent les CNNs avec des mécanismes d'attention avancés et des techniques basées sur la saillance. Les mécanismes d'attention améliorent la concentration du modèle sur les régions d'images diagnostiquement pertinentes, gérant efficacement la variabilité dans l'apparence des lésions et améliorant la délimitation de frontières complexes. Les approches basées sur la saillance améliorent davantage la robustesse de segmentation en mettant en évidence les régions spécifiques aux lésions et réduisant l'impact des caractéristiques de fond non pertinentes et des artefacts.

S'appuyant sur ces avancées, nous proposons le framework MEDiXNet, une architecture Mixture of Experts (MoE) innovante spécifiquement conçue pour traiter les défis de segmentation identifiés. MEDiXNet intègre uniquement des réseaux experts spécialisés adaptés pour des catégories de lésions distinctes, combinés avec un Adaptive Salient Region Attention Module (ASRAM) innovant. Cette conception assure une performance de segmentation robuste et précise, améliorant significativement la précision dans les cas difficiles, incluant ceux avec des contrastes subtils et frontières indistinctes, tout en minimisant efficacement la sensibilité aux artefacts d'images et au bruit.

4.2 Vue d'Ensemble du Modèle MEDiXNet

Comme illustré en Figure 4.3, le modèle MEDiXNet introduit un framework adaptatif et robuste spécifiquement conçu pour traiter les défis de segmentation de lésions cutanées. En exploitant une architecture Mixture of Experts (MoE), MEDiXNet segmente efficacement les types de lésions diverses, traitant particulièrement les limitations rencontrées par les modèles traditionnels et basés sur l'apprentissage profond standard.

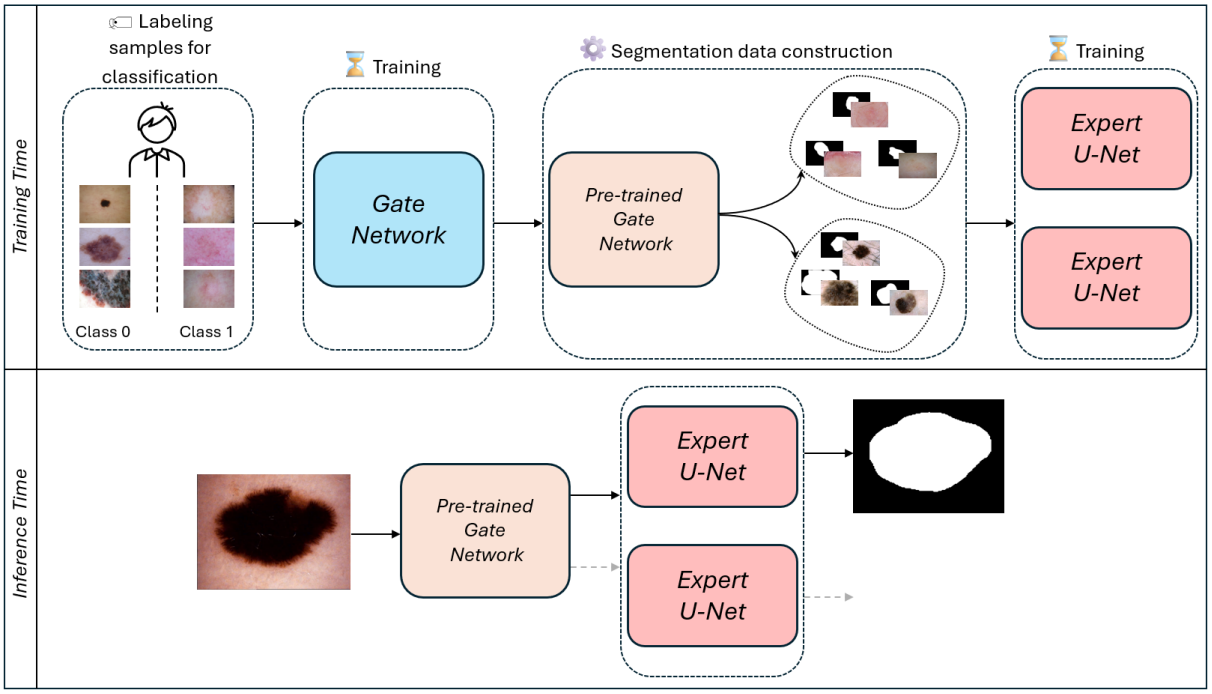

FIGURE 4.3 – Workflow MEDiXNet illustrant l'interaction entre les Réseaux Experts, le Gate Network, et le module ASRAM.

L'innovation clé de MEDiXNet réside dans ses réseaux experts spécialisés, son mécanisme de routage dynamique, et l'intégration de l'Adaptive Salient Region Attention Module (ASRAM), améliorant collectivement la précision des frontières des lésions et l'adaptabilité du modèle.

4.2.1 Motivation et Principes de Conception

La motivation derrière MEDiXNet provient d'observations critiques des limitations dans les approches de segmentation existantes, particulièrement dans la distinction entre lésions foncées et claires. Les lésions foncées, caractérisées par un contraste plus élevé contre la peau saine, sont généralement plus faciles à segmenter avec précision. En revanche, les lésions claires exhibent des contrastes subtils et des frontières ambiguës, présentant des défis significatifs même pour les cliniciens expérimentés. Les modèles généralisés existants basés sur CNN et Transformer sous-performent souvent sur ces lésions à faible contraste, indiquant la nécessité d'approches de modélisation spécialisées.

Adoptant une stratégie de spécialisation, reconnaître cette distinction fondamentale entre types de lésions a inspiré l'adoption d'une stratégie "divide and conquer", se manifestant dans l'architecture MoE de MEDiXNet. Contrairement aux modèles généralisés traditionnels, MEDiXNet emploie de multiples réseaux experts spécialisés, chacun entraîné pour gérer optimalement soit les lésions foncées soit claires. Le réseau de routage

dirige dynamiquement chaque image d'entrée vers l'expert le plus approprié, assurant une segmentation ciblée et précise. L'intégration d'ASRAM raffine davantage cette approche ciblée en mettant l'accent sur les caractéristiques spécifiques aux lésions et réduisant l'interférence de fond.

Garantissant la viabilité pratique, l'efficacité de calcul demeure un composant critique de MEDiXNet, assurant l'aptitude pour les applications cliniques réelles. Cette efficacité est atteinte par l'incorporation stratégique de composants de modèles légers et de mécanismes d'attention optimisés, assurant un déploiement clinique réactif sans compromettre la précision de segmentation.

Ainsi, les principes architecturaux de MEDiXNet englobent :

1. Spécialisation par de multiples réseaux experts spécifiques aux lésions pour améliorer la précision de segmentation.
2. Mécanismes d'attention adaptative avancés pour délimiter précisément les frontières des lésions et supprimer l'information non pertinente.
3. Efficacité de calcul optimale, assurant l'applicabilité pratique dans les contextes cliniques.

4.2.2 Composants Clés de MEDiXNet

MEDiXNet comprend trois composants primaires : les Réseaux Experts, le Gate Network, et l'Adaptive Salient Region Attention Module (ASRAM). Chaque composant contribue distinctement à la robustesse globale et à l'adaptabilité du modèle de segmentation.

4.2.2.1 Réseaux Experts

Au sein de MEDiXNet, les réseaux experts sont méticuleusement conçus pour traiter les attributs visuels distincts des lésions cutanées, différenciant particulièrement entre les catégories de lésions foncées et claires. Chaque réseau expert emploie une architecture en forme de U consistante caractérisée par une structure encodeur-décodeur optimisée pour une segmentation détaillée et précise des lésions.

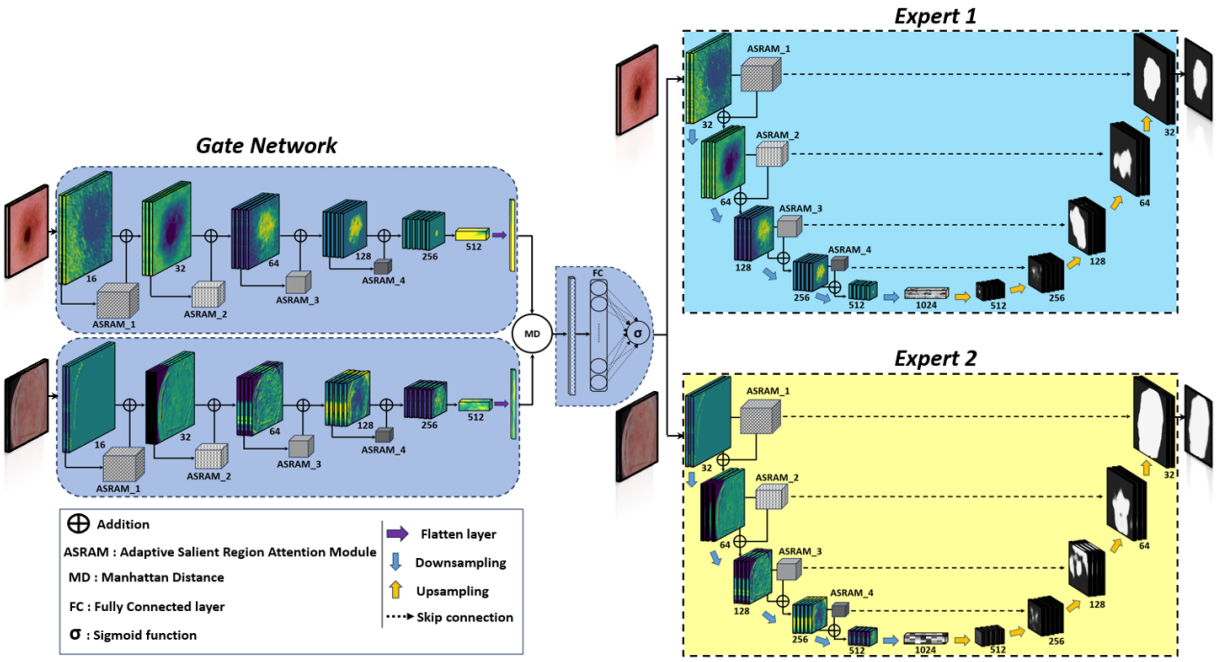

FIGURE 4.4 – Architecture MEDiXNet illustrant les réseaux experts spécialisés, le Gate Network, et l'intégration du module ASRAM.

Au niveau de l'extraction de caractéristiques, l'encodeur consiste en quatre étapes d'encodage consécutives, chacune conçue pour extraire progressivement des caractéristiques spatiales hiérarchiques. Chaque étape d'encodage comprend une pile de couches convolutionnelles, chacune suivie par une normalisation par batch et une activation Rectified Linear Unit (ReLU), culminant en une couche max-pooling pour réduire systématiquement les dimensions spatiales. Cette extraction de caractéristiques progressive capture efficacement les détails texturaux et structurels complexes des lésions, cruciaux pour une segmentation précise.

Concernant la reconstruction spatiale, le décodeur reflète l'encodeur avec quatre étapes de décodage correspondantes, mais utilise un traitement asymétrique pour reconstruire des masques de segmentation haute résolution à partir des caractéristiques encodées. Chaque étape de décodage commence avec des opérations d'upsampling, suivies par des couches convolutionnelles, normalisation par batch, et activations ReLU. Ces opérations améliorent la résolution spatiale et reconstruisent efficacement des frontières de segmentation détaillées.

Préservant l'information fine, les connexions skip relient directement les étapes correspondantes d'encodeur et de décodeur, critiques pour préserver le détail spatial. Ces connexions fournissent des voies directes pour l'information spatiale, atténuant la perte de détails cruciaux typiquement rencontrée durant l'encodage.

Renforçant les capacités d'attention, chaque réseau expert intègre l'Adaptive Salient Region Attention Module (ASRAM), améliorant la représentation contextuelle et

mettant l'accent sur les caractéristiques diagnostiquement pertinentes. La combinaison d'attention multi-échelle et de connexions skip améliore significativement la capacité de segmentation, particulièrement dans les cas difficiles caractérisés par des frontières des lésions subtiles ou ambiguës.

Dans une perspective d'optimisation spécialisée, chaque réseau expert est indépendamment entraîné en utilisant des ensembles de données spécialisés sélectionnés pour des catégories de lésions spécifiques. Cette approche d'entraînement ciblée assure que chaque réseau capture et modélise avec précision les distinctions visuelles subtiles entre types de lésions, améliorant substantiellement la précision et la fiabilité globales de segmentation.

Ancré dans le paradigme Mixture of Experts (MoE), MEDiXNet exploite stratégiquement de multiples réseaux experts spécialisés pour gérer optimalement les caractéristiques diverses des lésions. Les couches convolutionnelles initiales traitent efficacement les artefacts d'imagerie communs, tandis que l'intégration ultérieure des modules d'attention ASRAM fournit un raffinement de caractéristiques complet, combinant la préservation de détails locaux avec l'amélioration contextuelle globale. Cette approche structurée assure des résultats de segmentation précis et une adaptabilité robuste aux scénarios d'imagerie clinique.

4.2.2.2 Gate Network pour la Classification de Lésions

Le Gate Network classifie dynamiquement et dirige les images d'entrée vers les réseaux experts les plus appropriés utilisant une architecture de réseau de neurones siamois adaptée pour l'apprentissage few-shot. Cette approche traite efficacement la disponibilité limitée d'images de lésions cutanées étiquetées, permettant au Gate Network de généraliser précisément à partir de données annotées minimales.

Structurellement, le Gate Network consiste en deux sous-réseaux parallèles identiques, chacun responsable d'extraire les embeddings de caractéristiques à partir de paires d'images : une image query, et un échantillon prototype représentatif d'une catégorie de lésion (soit lésions foncées ou claires). Chaque sous-réseau comprend une séquence de blocs convolutionnels résiduels, méticuleusement conçus pour une extraction de caractéristiques efficace et effective. Spécifiquement, chaque bloc convolutionnel résiduel contient une couche convolutionnelle, suivie par la Normalisation par Batch et la fonction d'activation ReLU. Cette combinaison assure un entraînement stable, une convergence efficace, et l'extraction de caractéristiques riches et discriminatives essentielles pour identifier précisément les distinctions subtiles entre lésions.

Intégrant des mécanismes d'attention intermédiaires, un Adaptive Salient Region Attention Module (ASRAM) est stratégiquement intégré entre les blocs résiduels avant d'appliquer le max-pooling, exactement comme utilisé dans la partie encodeur Expert. Ce mécanisme d'attention intermédiaire met en évidence dynamiquement les zones saillantes des lésions, raffinant significativement la représentation de caractéristiques spatiales en réduisant le bruit et l'information de fond non pertinente. Un tel raffinement adaptatif spatial et par canal améliore dramatiquement la capacité du Gate Network à différencier

entre les apparences similaires de lésions, améliorant ainsi la précision de classification globale.

Au niveau de la représentation vectorielle, à la sortie des piles convolutionnelles, chaque sous-réseau parallèle produit un vecteur d'embedding dense de 512 dimensions représentant les caractéristiques de haut niveau des lésions. Pour quantifier la similarité entre ces vecteurs d'embedding, la distance de Manhattan (norme L1) est calculée comme :

$$\text{Distance de Manhattan} = \sum_{i=1}^{512} \Big(x_i - y_i| \tag{4.1}$$

où x_i et y_i représentent les éléments correspondants des deux vecteurs d'embedding obtenus des sous-réseaux.

Produisant le score de routage final, le vecteur de distance calculé est par la suite passé par une couche entièrement connectée, raffinant davantage la représentation et réduisant la dimensionnalité à une valeur scalaire. Finalement, une fonction d'activation sigmoïde est appliquée, produisant un score de similarité dans la plage [0, 1], indiquant efficacement la probabilité que les images appariées appartiennent à la même catégorie de lésion. Cette métrique de similarité explicite assure des décisions de routage précises et confiantes.

L'architecture et la stratégie d'entraînement employées dans le Gate Network supportent directement son rôle critique dans le framework MEDiXNet. En exploitant les blocs convolutionnels résiduels avancés, les mécanismes d'attention ASRAM intermédiaires, et les métriques de similarité précises, le Gate Network distingue robustement les variations subtiles entre lésions, assignant de manière optimale chaque image de lésion à son réseau expert correspondant. Par conséquent, le modèle atteint une précision de segmentation améliorée et une adaptabilité améliorée dans divers scénarios cliniques.

4.2.2.3 Adaptive Salient Region Attention Module (ASRAM)

L'Adaptive Salient Region Attention Module (ASRAM) raffine la segmentation des lésions par un mécanisme d'attention hybride sophistiqué qui intègre l'attention spatiale, l'attention par canal, et l'amélioration guidée par saillance. L'architecture détaillée, illustrée en Figure 4.5, améliore systématiquement la représentation de caractéristiques et met l'accent sur les zones diagnostiquement pertinentes au sein des images de lésions.

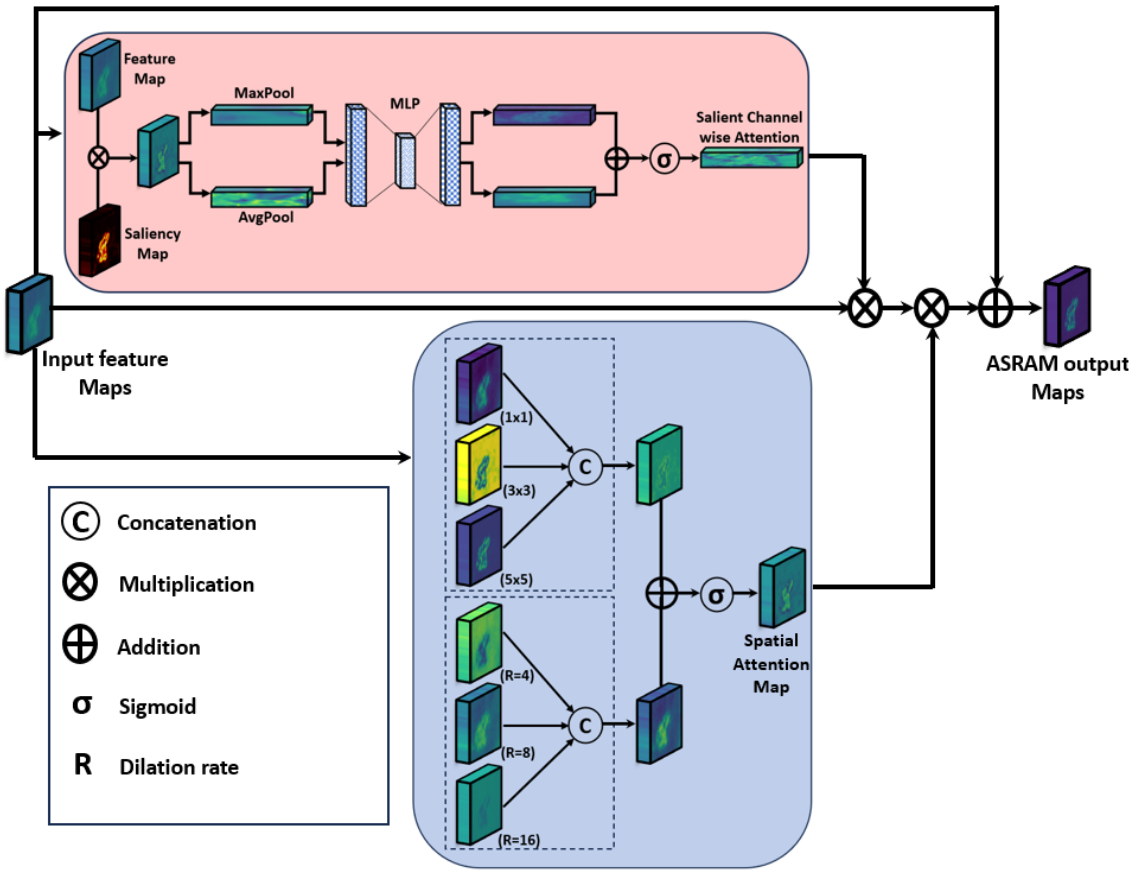

FIGURE 4.5 – Architecture détaillée de l'Adaptive Salient Region Attention Module (ASRAM), illustrant l'attention spatiale via les convolutions dilatées multi-échelles (SSAM), l'attention par canal par le pooling global et MLP (SCAM), et l'intégration d'attention guidée par saillance.

L'architecture d'ASRAM repose sur deux branches parallèles mais complémentaires : l'attention spatiale et l'attention par canal saillante, dont les sorties sont combinées pour produire des cartes d'attention raffinées.

Traitant les caractéristiques spatiales, la branche d'attention spatiale, appelée *Separable Spatial Attention Module* (SSAM), est spécifiquement conçue pour identifier et mettre l'accent sur les caractéristiques spatiales significatives pertinentes aux frontières et structures des lésions. Initialement, les cartes de caractéristiques d'entrée subissent un traitement via des convolutions séparables en profondeur avec des tailles de kernel 1×1, 3×3, et 5×5, aux côtés de convolutions dilatées avec des taux de dilatation de 4, 8, et 16. Cette approche multi-échelle capture efficacement les détails spatiaux à diverses résolutions. Les sorties de convolution de chaque groupe sont concaténées, formant deux cartes de caractéristiques primaires : $F_1 = [f_1(F), f_3(F), f_5(F)]$ et $F_2 = [f_{d=4}(F), f_{d=8}(F), f_{d=16}(F)]$. Ces cartes concaténées sont sommées et par la suite passées par une fonction d'activation sigmoïde, générant la carte d'attention spatiale

raffinée :

$$M_{\text{SSAM}}(F) = \sigma\left(F_1 + F_2\right) \tag{4.2}$$

Cette attention spatiale met en évidence sélectivement les régions saillantes, supprimant efficacement les caractéristiques de fond non pertinentes.

Complémentant l'attention spatiale, le *Salient Channel Attention Module* (SCAM) priorise dynamiquement les canaux de caractéristiques basés sur leur importance diagnostique. Étant donné une carte de caractéristiques d'entrée $F \in \mathbb{R}^{C\times H\times W}$ et une carte de saillance S générée par l'algorithme FASA [138], SCAM effectue d'abord une multiplication element-wise entre la carte de caractéristiques et la carte de saillance, produisant une carte de caractéristiques améliorée par saillance $F_S = F \otimes S$. Cette carte améliorée subit un global average pooling (GAP) et un global max pooling (GMP), générant des descripteurs de canaux distincts. Les deux descripteurs sont traités par un perceptron multicouche (MLP) partagé, dont les sorties sont sommées et passées par une activation sigmoïde pour produire la carte d'attention de canal :

$$M_{\text{SCAM}}(F) = \sigma(\text{MLP}(\text{GAP}(F_S)) + \text{MLP}(\text{GMP}(F_S))) \tag{4.3}$$

Intégrant les deux mécanismes d'attention, la sortie finale d'ASRAM intègre les modules SCAM et SSAM par une stratégie de connexion résiduelle. Initialement, la carte de caractéristiques d'entrée F est modulée par la carte d'attention de canal de SCAM, produisant une représentation de caractéristiques intermédiaire $F' = M_{\text{SCAM}}(F) \otimes F$. Par la suite, F' est raffinée en appliquant l'attention spatiale de SSAM, générant une représentation davantage améliorée $F'' = M_{\text{SSAM}}(F') \otimes F'$. La sortie finale d'ASRAM combine cette carte raffinée avec l'entrée originale via une addition element-wise, assurant un raffinement de caractéristiques complet sans perte d'information :

$$F_{\text{ASRAM}} = F + F'' \tag{4.4}$$

En intégrant efficacement l'attention spatiale et par canal avec le raffinement guidé par la saillance, ASRAM améliore significativement la capacité de MEDiXNet à délimiter précisément les frontières subtiles des lésions, particulièrement bénéfique dans les scénarios impliquant un faible contraste ou des contours indistincts. Par conséquent, MEDiXNet démontre une précision de segmentation améliorée et une applicabilité clinique robuste en diagnostics dermatologiques.

4.3 Résultats et Analyse

Cette section présente une évaluation complète du framework MEDiXNet proposé pour la segmentation de lésions cutanées. Les résultats expérimentaux sont structurés en plusieurs parties, commençant par la description des ensembles de données utilisés pour l'évaluation, suivie d'une analyse détaillée des résultats quantitatifs et qualitatifs obtenus. De plus, une étude d'ablation est conduite pour évaluer les contributions des

différents composants de MEDiXNet, particulièrement le framework Mixture of Experts, le Gate Network, et l'Adaptive Salient Region Attention Module (ASRAM). Une analyse comparative de performance contre les modèles de segmentation de pointe est également incluse pour démontrer la supériorité de l'approche proposée.

4.3.1 Ensembles de Données et Configuration Expérimentale

Le MEDiXNet proposé a été évalué en utilisant les ensembles de données de challenge ISIC 2017 [139] et ISIC 2018 [140] publiquement disponibles, qui fournissent un ensemble diversifié d'images dermoscopiques. Ces ensembles de données sont des benchmarks largement utilisés en recherche de segmentation de lésions cutanées, contenant des lésions de tailles, formes, et niveaux de contraste variés, les rendant idéaux pour évaluer la capacité de généralisation des modèles d'apprentissage profond.

Concernant la composition des ensembles, l'ensemble de données ISIC 2017 comprend 2000 images d'entraînement, 150 images de validation, et 600 images de test, avec les masques de segmentation de référence correspondants. L'ensemble de données ISIC 2018 est plus large, contenant 2594 images d'entraînement, 100 images de validation, et 1000 images de test, fournissant un ensemble d'évaluation plus extensif. Pour assurer la consistance en entraînement, toutes les images ont été redimensionnées à 256×256 pixels, et des techniques standard de prétraitement d'images, incluant la normalisation, l'amélioration de contraste, et l'augmentation de données (rotations aléatoires, flipping, et déformations élastiques), ont été appliquées.

Concernant l'entraînement du Gate Network (dispatcher), ce composant critique de MEDiXNet a été entraîné en utilisant 10 images de lésions foncées et 10 images de lésions claires, validé contre deux images query (une de chaque catégorie). Cette approche a assuré que le dispatcher classifiait précisément les lésions basées sur les variations de contraste, permettant au framework de diriger chaque image vers le modèle de segmentation approprié.

Concernant l'implémentation technique, le modèle a été implémenté en PyTorch et entraîné en utilisant un GPU NVIDIA RTX A6000 pour 200 époques. L'optimiseur Adam a été utilisé pour les mises à jour de paramètres, et le taux d'apprentissage a suivi une décroissance par annealing cosinus pour assurer une convergence stable. La fonction de perte Dice, largement utilisée pour les tâches de segmentation, a été employée comme fonction de perte primaire :

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \varepsilon} \tag{4.5}$$

où p_i et g_i représentent les valeurs prédites et de vérité terrain, respectivement, sur l'ensemble des N pixels, et ε assure la stabilité numérique.

4.3.2 Évaluation Quantitative des Performances

Pour évaluer de manière complète l'efficacité de MEDiXNet, nous avons conduit des expérimentations extensives comparant sa performance avec plusieurs modèles de segmentation d'apprentissage profond de pointe, incluant DSNet[141], EGE-UNet [142], FATNet [143], MFSNet [144], MsRED [145], et MSCANet [146]. Ces analyses comparatives ont employé des métriques d'évaluation largement reconnues : Dice Similarity Coefficient (DSC), Jaccard Index (JC), Sensitivity (Sens), Specificity (Spec), et Accuracy (Acc). Ces métriques ont été choisies en raison de leur capacité à fournir une évaluation holistique de la qualité de segmentation, englobant la précision de chevauchement, la précision des frontières, et la robustesse contre les faux positifs et faux négatifs.

Présentant les résultats comparatifs, le tableau 4.1 résume les métriques de performance quantitatives obtenues des évaluations conduites sur les ensembles de données de test ISIC 2017 et ISIC 2018. Notamment, MEDiXNet a consistamment surpassé les méthodes concurrentes sur toutes les métriques, démontrant sa précision et robustesse supérieures.

Analysant le Dice Similarity Coefficient, celui-ci reflète le chevauchement entre les frontières des lésions prédites et de référence, mesurant directement la précision de segmentation. MEDiXNet a atteint des scores DSC de 0.9433 et 0.9487 sur les ensembles de données ISIC 2017 et ISIC 2018, respectivement, surpassant significativement toutes les autres méthodes. Ceci indique la capacité exceptionnelle du modèle à produire des masques de segmentation précis, capturant précisément les frontières des lésions même dans des scénarios complexes.

Concernant le Jaccard Index, qui mesure la similarité entre les segmentations prédites et de référence en tenant compte de l'intersection et union des zones, cet indice souligne davantage la performance supérieure de MEDiXNet, atteignant les plus hautes valeurs JC (0.8756 et 0.8902 sur ISIC 2017 et 2018 respectivement). Les hauts scores JC indiquent l'efficacité de MEDiXNet à minimiser les erreurs de segmentation et à représenter précisément les zones des lésions.

Au niveau de la sensibilité diagnostique, Sensitivity (Sens) mesure la capacité du modèle à détecter précisément les régions des lésions, une métrique essentielle en diagnostics médicaux où les faux négatifs peuvent mener à des conséquences sévères. MEDiXNet a enregistré des scores de sensibilité de 0.9545 et 0.9601 pour ISIC 2017 et ISIC 2018, respectivement, démontrant sa fiabilité en identification des zones des lésions et minimisation des détections manquées.

Traitant la spécificité, Specificity (Spec), reflétant la capacité du modèle à identifier correctement les régions saines, est également critique pour éviter les faux positifs. MEDiXNet a atteint des scores de spécificité impressionnants de 0.9524 et 0.9565, indiquant sa performance robuste en distinction des lésions de la peau saine, réduisant significativement les fausses alertes.

Finalement, Accuracy (Acc) fournit une vue d'ensemble équilibrée de la performance globale du modèle. Les scores de précision de MEDiXNet de 0.9673 (ISIC 2017) et 0.9686

(ISIC 2018) ont clairement surpassé les méthodes concurrentes, confirmant son efficacité à segmenter de manière fiable les lésions dans diverses conditions d'imagerie et types de lésions.

Les résultats présentés dans le Tableau 4.1 soulignent les améliorations remarquables de MEDiXNet sur les modèles existants, démontrant clairement son potentiel pour l'intégration clinique et les applications dermatologiques pratiques. Cette analyse quantitative établit fermement MEDiXNet comme une méthode supérieure capable de gérer une large gamme de défis de segmentation avec une précision et fiabilité exceptionnelles.

TABLE 4.1 – Comparaison de différentes méthodes sur les ensembles de données ISIC.

ISIC 2017								
Method	DSC	JC	Sens	Spec	Acc	Params (M)	TT (h)	IT (ms)
DSNet	0.8586	0.8134	0.8674	0.8603	0.9082	10	1.7	185.4
EGE-UNet	0.8671	0.8223	0.8894	0.8692	0.8935	**0.053**	**0.94**	**12.8**
FATNet	0.9059	0.8554	0.9027	0.9261	0.9430	30	8.2	485.2
MFSNet	0.9150	0.8513	0.9062	0.9450	0.9580	31	8.4	498.7
MsRED	0.8684	0.8122	0.8971	0.8588	0.9014	3.8	1.5	45.3
MSCANet	0.8861	0.8328	0.9308	0.8392	0.9136	27	7.8	421.8
MEDiXNet	**0.9433**	**0.8756**	**0.9545**	**0.9524**	**0.9673**	8.2	3.3	**28.9**
ISIC 2018								
Method	DSC	JC	Sens	Spec	Acc	Params (M)	TT (h)	IT (ms)
DSNet	0.8953	0.8447	0.9251	0.8713	0.9447	10	1.7	185.4
EGE-UNet	0.8837	0.8551	0.8927	0.9080	0.9239	**0.053**	**0.94**	**12.8**
FATNet	0.9273	0.8776	0.9090	0.9550	0.9534	30	8.2	485.2
MFSNet	0.9190	0.8730	0.8840	0.9590	0.9460	31	8.4	498.7
MsRED	0.8924	0.8252	0.9281	0.8878	0.9344	3.8	1.5	45.3
MSCANet	0.9050	0.8683	0.8929	0.9096	0.9443	27	7.8	421.8
MEDiXNet	**0.9487**	**0.8902**	**0.9601**	**0.9565**	**0.9686**	8.2	3.3	**28.9**

4.3.3 Analyse Qualitative et Discussion

Pour compléter l'évaluation quantitative, nous avons conduit des analyses comparatives visuelles utilisant des échantillons représentatifs des ensembles de données ISIC. Les Figures 4.6 et 4.7 illustrent les résultats de segmentation pour divers cas difficiles : lésions à faible contraste, lésions avec frontières irrégulières et diffuses, et images affectées par des artefacts tels que cheveux et encres de marqueur.

Analysant la précision des masques de segmentation, la Figure 4.6 montre que MEDiXNet produit des masques de segmentation étroitement alignés avec les annotations de vérité terrain. Les modèles concurrents tels que DSNet, EGE-UNet, et MsRED présentent des limitations notables dans la délimitation précise des frontières, particulièrement pour les lésions avec gradients de couleur subtils et faible contraste contre la peau saine environnante. En revanche, MEDiXNet segmente avec succès ces lésions difficiles, capturant leur étendue complète tout en préservant les détails complexes des frontières.

Examinant la délimitation des contours, les comparaisons détaillées de la Figure 4.7 révèlent la performance supérieure de MEDiXNet. Les contours rouges indiquent les prédictions du modèle, tandis que les contours verts représentent les annotations de vérité terrain. MEDiXNet gère efficacement les cas complexes où d'autres méthodes ont échoué : lésions obscurcies par des brins de cheveux ou caractérisées par des frontières diffuses et irrégulières. L'intégration de réseaux experts spécialisés et du module ASRAM améliore la capacité de MEDiXNet à se concentrer précisément sur les caractéristiques cliniquement pertinentes, produisant une segmentation hautement précise et détaillée.

Ces résultats qualitatifs renforcent les observations quantitatives, confirmant la capacité de MEDiXNet à segmenter de manière fiable les lésions sous des conditions d'imagerie clinique variées et exigeantes. Sa performance robuste la rend particulièrement appropriée pour l'intégration dans les workflows diagnostiques cliniques, offrant aux dermatologues un outil précieux pour l'évaluation précoce et précise des lésions cutanées.

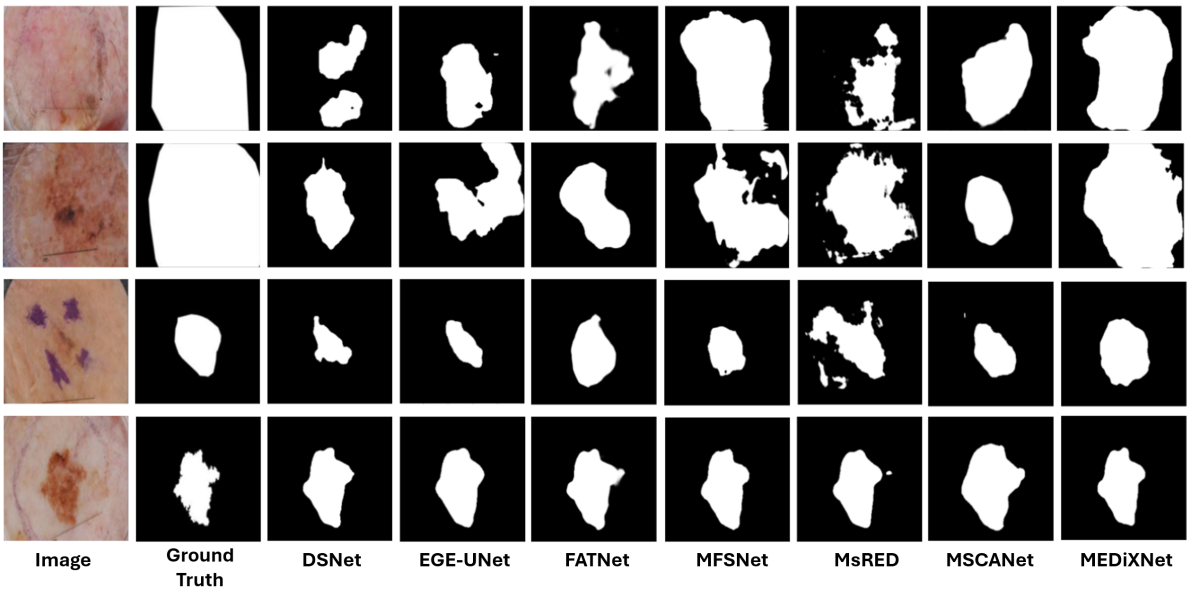

FIGURE 4.6 – Comparaison qualitative des masques de segmentation sur des images représentatives de l'ensemble de données ISIC. MEDiXNet atteint une précision de segmentation supérieure comparée aux méthodes de pointe.

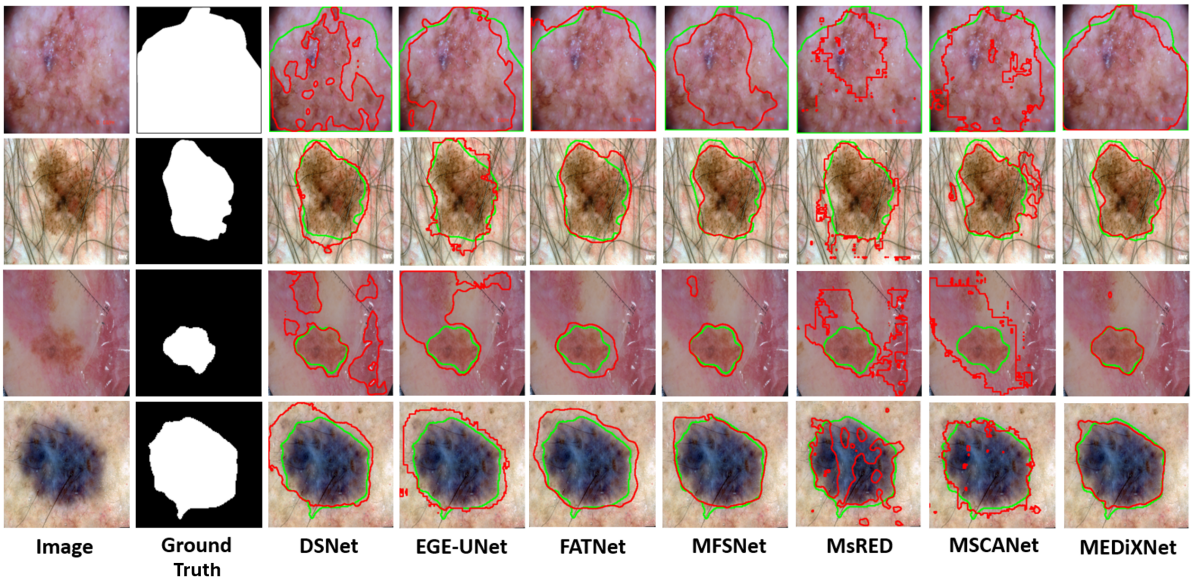

FIGURE 4.7 – Comparaison détaillée des frontières des résultats de segmentation sur des exemples de lésions difficiles. Les lignes vertes indiquent les frontières de vérité terrain, tandis que les lignes rouges montrent les prédictions de chaque méthode. MEDiXNet démontre une précision exceptionnelle en délimitation des frontières.

4.4 Limitations Actuelles et Directions Futures de MEDiXNet

Malgré les avancées substantielles démontrées par le framework MEDiXNet en segmentation de lésions cutanées, plusieurs limitations notables persistent et méritent discussion.

Concernant la sélection des données d'entraînement, la sélection actuelle d'images anchor et la construction correspondante d'ensembles de données pour entraîner le Gate Network sont conduites manuellement. Ce processus manuel s'avère chronophage et potentiellement sujet à des biais et inconsistances, impactant la fiabilité et l'évolutivité du modèle. De plus, les anchors et ensembles de données d'entraînement sont actuellement sélectionnés principalement sur la base des caractéristiques de couleur des lésions, négligeant d'autres caractéristiques visuelles et morphologiques critiques qui pourraient contribuer de manière significative à distinguer les catégories de lésions.

Au niveau de la complexité architecturale, MEDiXNet emploie une architecture basée sur CNN, où l'extension du framework au-delà de deux réseaux experts augmente substantiellement les paramètres du modèle. Cette augmentation mène à une inefficacité de calcul et pose des défis pour le déploiement pratique.

Concernant le mécanisme de routage, l'approche existante implémente une classification basée sur un seuil dur pour diriger les images vers les experts. Cette stratégie

s'avère problématique pour les cas ambigus avec des scores de similarité proches de 0.5, qui sont forcément classifiés dans une catégorie unique. Cette classification rigide peut compromettre la précision de segmentation pour les lésions affichant des caractéristiques visuelles intermédiaires.

Concernant la spécialisation des experts, chaque réseau expert est entraîné exclusivement sur son sous-ensemble de données respectif. Cette approche améliore la spécialisation mais limite la capacité de généralisation. Une erreur de classification par le Gate Network peut ainsi entraîner des résultats de segmentation sous-optimaux, soulignant une dépendance critique sur la précision du mécanisme de routage.

Les recherches futures devront traiter explicitement ces limitations par plusieurs axes d'amélioration : automatiser les processus de sélection d'anchor, incorporer de multiples caractéristiques visuelles au-delà de la couleur, optimiser l'architecture pour une efficacité de calcul améliorée, raffiner les stratégies de classification pour mieux gérer les cas ambigus, et améliorer la capacité de généralisation des réseaux experts. Ces améliorations permettront d'accroître la robustesse et la précision globales de segmentation du framework MEDiXNet.

Chapitre 5

MixLVMM : Un Mélange de Modèles Vision Mamba Légers pour une Segmentation Robuste de Lésions Cutanées

5.1 Introduction et Motivation

La segmentation précise des lésions cutanées demeure une pierre angulaire dans la détection précoce et le diagnostic des cancers de la peau, notamment le mélanome. S'appuyant sur les fondations établies par les approches d'apprentissage profond précédentes, incluant les CNNs, les modèles hybrides basés sur Transformer, et notre travail antérieur avec le framework MEDiXNet, des progrès substantiels ont été réalisés dans l'amélioration de la précision de segmentation. Cependant, des défis persistants tels que la haute variabilité de couleur parmi les lésions, la prévalence d'artefacts d'imagerie, et la subtilité des frontières des lésions continuent de limiter la performance de segmentation dans les environnements cliniques réels.

Le modèle MEDiXNet a introduit une architecture Mixture of Experts (MoE) pour traiter la variabilité entre les types de lésions, ciblant particulièrement les différences entre lésions foncées et claires. Bien que cette spécialisation ait amélioré la précision de segmentation, MEDiXNet a conservé certaines limitations notables. Le modèle est demeuré efficace avec seulement deux réseaux experts, résultant en une taille compacte d'approximativement 8.2 millions de paramètres. Toutefois, il devient de plus en plus exigeant en calcul lorsque le nombre d'experts augmente. Cette contrainte architecturale limite son évolutivité dans les scénarios nécessitant une gestion plus large de la diversité des lésions.

Face au problème de spécialisation excessive, MEDiXNet souffrait également d'un problème de "vision étroite" : les réseaux experts étaient strictement spécialisés soit pour les lésions foncées soit claires, et le Gate Network était entraîné uniquement pour

distinguer entre ces deux catégories. Lorsqu'une erreur de classification survenait, par exemple, diriger une lésion à ton clair vers l'expert à ton foncé, les résultats de segmentation s'avéraient souvent catastrophiques, produisant parfois des masques de segmentation presque vides. Dans les cas ambigus où les lésions exhibaient des caractéristiques des deux catégories, l'entropie de classification approchait 0.5, menant à une assignation d'expert effectivement aléatoire et dégradant la fiabilité de segmentation.

Proposant une solution intégrée, MixLVMM est conçu pour surmonter ces limitations critiques. Il introduit quatre avancées clés :

1. **Experts Basés sur Vision Mamba Légers** : Remplacer les réseaux experts basés sur CNN avec des architectures Vision Mamba (VMM) légères améliore la capacité du modèle à capturer efficacement les dépendances à long terme, et réduit de manière significative l'empreinte paramétrique. Cette approche assure une efficacité de calcul sans sacrifier la performance de segmentation.
2. **Génération d'Anchors Automatisée et Mécanisme de Routage Amélioré** : Au lieu de sélectionner manuellement les anchors, MixLVMM exploite une approche non supervisée utilisant Uniform Manifold Approximation and Projection (UMAP) combiné avec le clustering k-means++. Cette stratégie identifie de manière systématique des anchors représentatifs basés sur les distributions d'espace latent, améliorant la robustesse et la généralisabilité du Gate Network. Le Gate Network est entraîné en utilisant une formulation triplet loss, raffinant ses capacités discriminatives.
3. **Stratégie d'Apprentissage à Deux Étapes pour une Spécialisation d'Expert Plus Large** : MixLVMM introduit une approche d'apprentissage à deux étapes pour traiter la limitation de vision étroite. La première phase consiste en un pré-entraînement global sur l'ensemble de données complet pour permettre une compréhension générale de la diversité des lésions. La seconde phase effectue un fine-tuning spécifique aux experts, permettant à chaque expert de se spécialiser non seulement sur les tons de couleur mais également sur des caractéristiques plus nuancées des lésions : texture, forme, et complexité des frontières. Cette spécialisation plus large permet à MixLVMM de gérer de manière robuste un large spectre d'apparences et d'atténuer les risques associés aux erreurs de classification.
4. **Raffinement d'Attention Préservé via ASRAM** : Reconnaissant l'efficacité de l'Adaptive Salient Region Attention Module (ASRAM) introduit dans MEDiX-Net, MixLVMM conserve ASRAM sans modification. ASRAM continue de jouer un rôle crucial dans le raffinement des caractéristiques spatiales et par canal au sein des réseaux experts légers, assurant une délimitation précise des frontières des lésions.

Par ces améliorations, MixLVMM traite les défis centraux de la segmentation de lésions cutanées : gérer la variabilité visuelle, améliorer la faisabilité de calcul pour le déploiement, et automatiser les étapes critiques dans le workflow du modèle. En combinant

des architectures d'experts légères mais puissantes avec un mécanisme de dispatching entièrement automatisé et optimisé, MixLVMM représente une avancée substantielle vers des systèmes de diagnostic assistés par ordinateur pratiques, évolutifs et cliniquement viables pour la dermatologie.

Les sections suivantes détaillent l'architecture MixLVMM, ses composants clés, la validation expérimentale, et l'analyse comparative de performance contre les modèles de segmentation de pointe.

5.2 Composants du Modèle et Méthodologie

Cette section présente une exposition détaillée des composants architecturaux et méthodologiques centraux qui constituent le framework MixLVMM. Chaque composant est soigneusement conçu pour traiter un défi spécifique en segmentation de lésions cutanées, contribuant collectivement à la robustesse, l'efficacité, et l'adaptabilité globales du modèle.

5.2.1 Architecture Expert MixLVMM

Chaque expert dans MixLVMM adopte une architecture encodeur-décodeur basée sur Vision Mamba unifiée, raffinée pour supporter la spécialisation entre divers types de lésions. Bien que tous les experts partagent une structure identique, leur force réside dans leur entraînement dédié sur des sous-groupes visuellement distincts de lésions, menant à une spécialisation et généralisation améliorées.

Au niveau de l'encodage des caractéristiques, l'encodeur initie avec un bloc convolutionnel résiduel, suivi par une couche patch embedding qui divise l'image d'entrée $x \in \mathbb{R}^{H\times W\times 3}$ en patches plus petits et mappe les caractéristiques vers une représentation $2C$-dimensionnelle, résultant en $x' \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 2C}$. Ces tokens sont ensuite normalisés et passés séquentiellement dans des blocs *Focused Vision Mamba (FVM)* empilés. Le processus d'encodage implique de multiples étapes de downsampling via Patch Merging, chacune suivie d'une paire de blocs FVM.

Concernant la reconstruction spatiale, le décodeur reflète l'encodeur avec des couches Patch Expanding symétriques entrelacées avec des blocs FVM, facilitant l'upsampling progressif et le raffinement de caractéristiques. Les connexions skip relient les blocs encodeur et décodeur aux échelles correspondantes via de simples opérations d'addition, assurant la préservation des détails spatiaux. Le masque de segmentation final $y \in \mathbb{R}^{H\times W\times 1}$ est généré via une couche de projection convolutionnelle 1×1.

Au cœur de l'architecture, le bloc FVM représente l'unité de calcul fondamentale dans MixLVMM, incorporant deux composants principaux : le bloc *Visual State Space (VSS)* et l'*Adaptive Salient Region Attention Module (ASRAM)*. Chaque bloc VSS inclut une convolution séparable en profondeur (DS-Conv), un mécanisme de mise à jour state-space linéaire, et le 2D Selective Scan (SS2D). Ces blocs implémentent le modèle state-space structuré, décrit par les ODEs linéaires suivantes :

$$\frac{d}{dt}h(t) = Ah(t) + Bx(t) \tag{5.1}$$
$$y(t) = Ch(t) \tag{5.2}$$

où A, B, et C sont des matrices apprenables représentant les mappings de transition, entrée, et sortie, respectivement. Ces équations sont discrétisées utilisant la transformée Zero-Order Hold (ZOH) :

$$\overline{A} = \exp(\Delta A) \tag{5.3}$$
$$\overline{B} = A^{-1}\left[\exp(\Delta A) - I\right] B \tag{5.4}$$

produisant la règle de mise à jour en temps discret :

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t \tag{5.5}$$
$$y_t = Ch_t \tag{5.6}$$

En termes pratiques, l'implémentation commence avec la Layer Normalization et divise l'entrée en deux branches. Une branche applique une projection linéaire et l'activation SiLU, tandis que l'autre procède par un pipeline DS-Conv et SS2D. Les deux voies sont fusionnées par une multiplication element-wise et une addition résiduelle :

$$\mathbf{y} = L(\text{SS2D}(\text{DS-Conv}(L(\mathbf{x})))) \odot L(\mathbf{x}) + \mathbf{x}, \tag{5.7}$$

où L dénote une couche linéaire. Cette combinaison permet au modèle de capturer efficacement les dépendances à court et à long terme.

La Figure 5.1 illustre l'architecture experte complète.

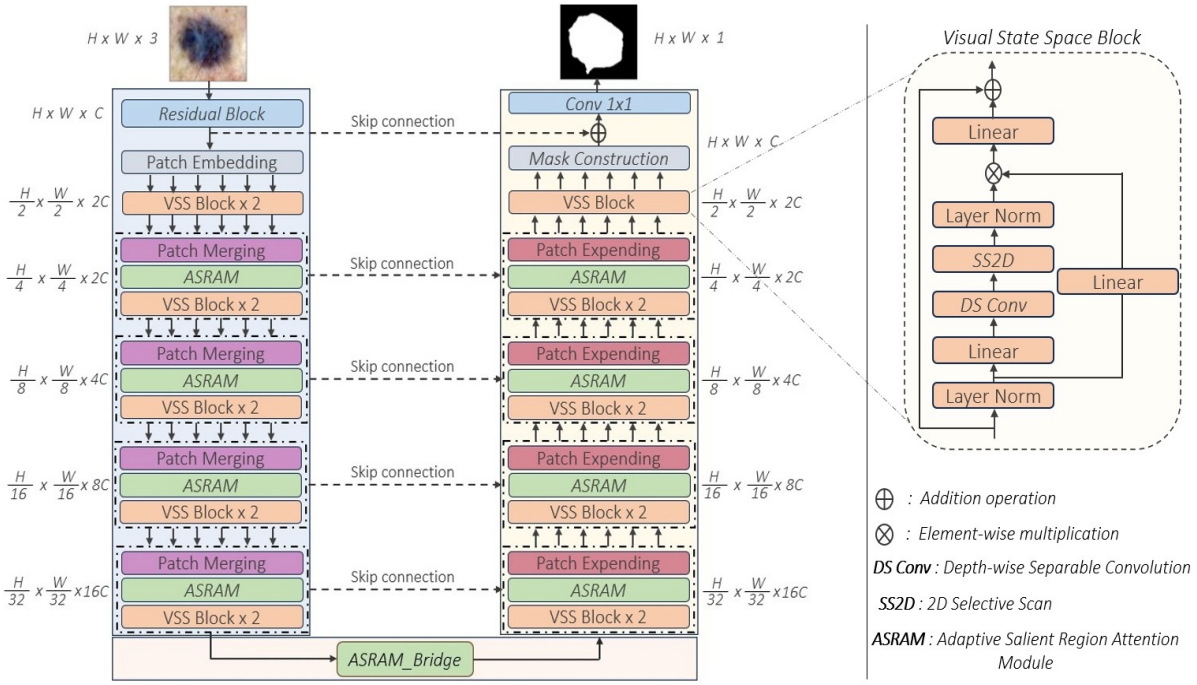

FIGURE 5.1 – Architecture Expert Vision Mamba. Chaque expert intègre Patch Embedding, blocs FVM, Patch Merging/Expanding, ASRAM, et connexions skip.

5.2.2 Le Gate Network

Le Gate Network est central au mécanisme de dispatching dynamique de MixLVMM. Il utilise un paradigme d'apprentissage few-shot pour diriger les images vers l'expert le plus approprié en apprenant un espace d'embedding tenant compte des distances.

L'architecture (Figure 5.2) suit une configuration siamoise, où des triplets d'images (anchor a_i, positive p_i, et negative n_i) sont passés dans un encodeur fait de blocs résiduels utilisant des convolutions séparables en profondeur. UMAP est d'abord appliqué à l'ensemble de données d'entraînement pour réduire sa dimensionnalité, suivi par le clustering K-means++ pour identifier des sous-groupes significatifs. Les images anchor sont ensuite sélectionnées de ces clusters.

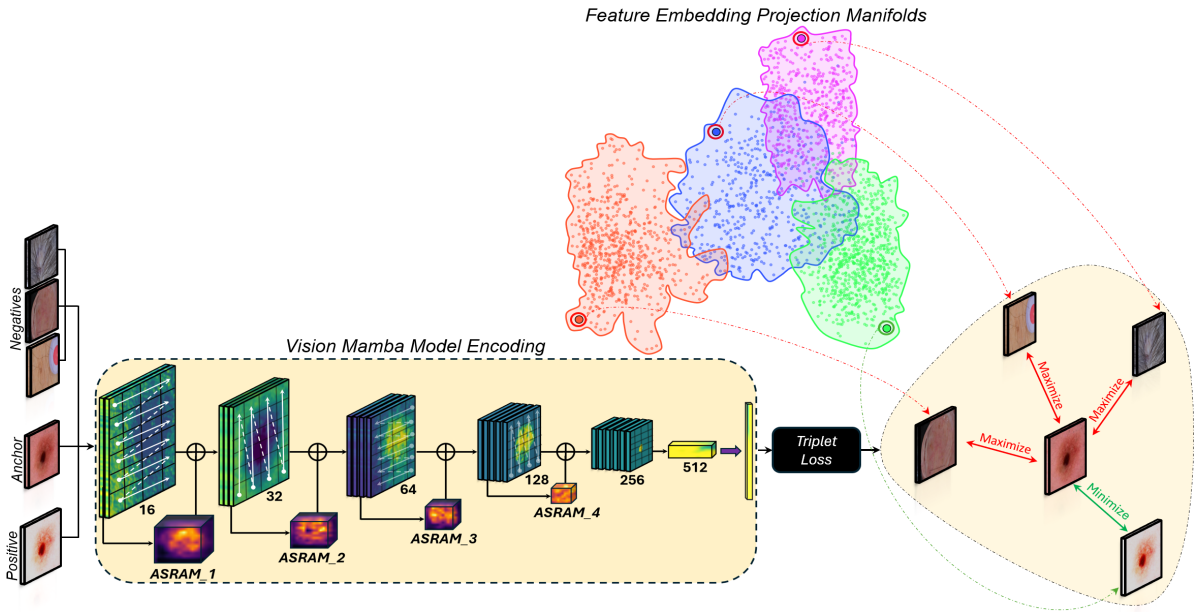

FIGURE 5.2 – Architecture Gate Network pour le routage basé sur la similarité utilisant triplet loss.

Au niveau de l'optimisation, le modèle est entraîné en utilisant la triplet loss contrastive :

$$L_i = \max\left(0, |g(a_i) - g(p_i)| - |g(a_i) - g(n_i)| + \varepsilon\right) \tag{5.8}$$

où ε est la marge. Une fois entraîné, la distance cosinus entre une image test et chaque anchor est calculée. Ces distances sont converties en poids de sélection d'expert via la fonction de routage basée sur softmax :

$$w_i(x) = \frac{\exp\left(-\frac{d_i(x)}{T}\right)}{\sum_{j=1}^{K} \exp\left(-\frac{d_j(x)}{T}\right)}, \quad i = 1, 2, \ldots, K \tag{5.9}$$

avec T contrôlant la douceur de l'assignation d'expert. La sortie de segmentation finale est calculée comme :

$$y(x) = \sum_{i=1}^{K} w_i(x).f_i(x), \tag{5.10}$$

où $f_i(x)$ est la sortie du i-ème expert. Cette stratégie assure un routage adaptatif et robuste même dans les cas ambigus.

5.2.3 Adaptive Salient Region Attention Module (ASRAM)

ASRAM est préservé sans modification de l'architecture MEDiXNet et demeure une pierre angulaire des réseaux experts dans MixLVMM. Il intègre à la fois l'attention par canal et spatiale en utilisant les modules SCAM et SSAM. Les cartes d'attention sont générées à partir de caractéristiques améliorées par saillance et fusionnées en utilisant un schéma résiduel pour préserver le contexte original.

Une explication détaillée et l'architecture visuelle d'ASRAM peuvent être trouvées au Chapitre 4. Pour la complétude, nous incluons sa vue d'ensemble schématique en Figure 5.3.

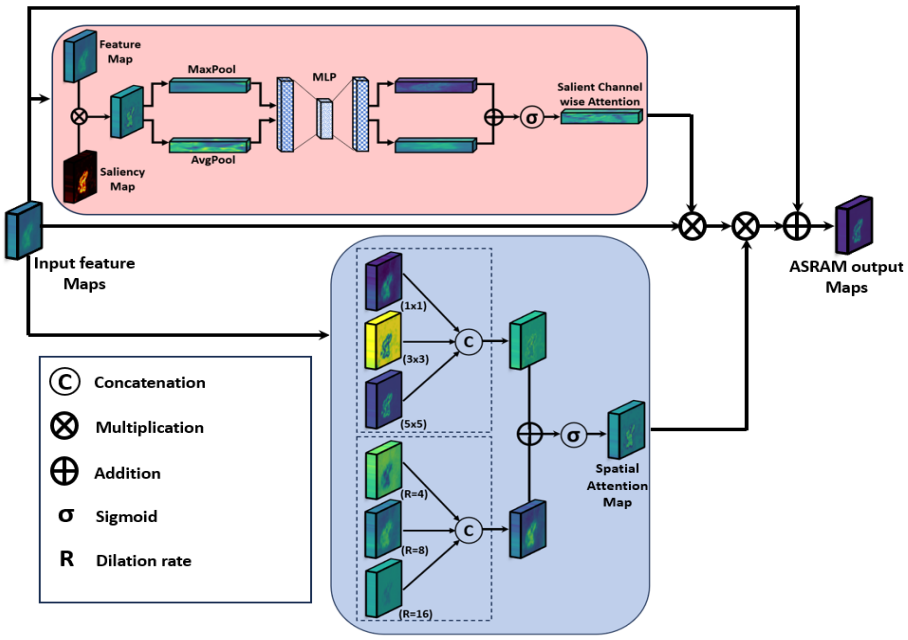

FIGURE 5.3 – Architecture Adaptive Salient Region Attention Module (ASRAM).

5.2.4 Stratégie de Pré-entraînement et Transfer Learning

Pour atténuer le problème de vision étroite et promouvoir la généralisation, une stratégie d'entraînement à deux phases est implémentée. Dans la première phase, tous les experts sont pré-entraînés sur l'ensemble de données complet, leur permettant de capturer les sémantiques visuelles partagées entre les types de lésions. Une fois la convergence atteinte, les poids de l'encodeur sont gelés pour prévenir la perte des représentations globales.

Dans la seconde phase, chaque expert subit un fine-tuning sur sa sous-catégorie respective de lésions. Cette procédure permet la spécialisation en caractéristiques de haut niveau sans écraser les connaissances apprises antérieurement, réduisant ainsi le risque d'oubli catastrophique. L'efficacité de cette approche est illustrée en Figure 5.4,

qui compare la performance des experts sur différents types de lésions avant et après transfer learning.

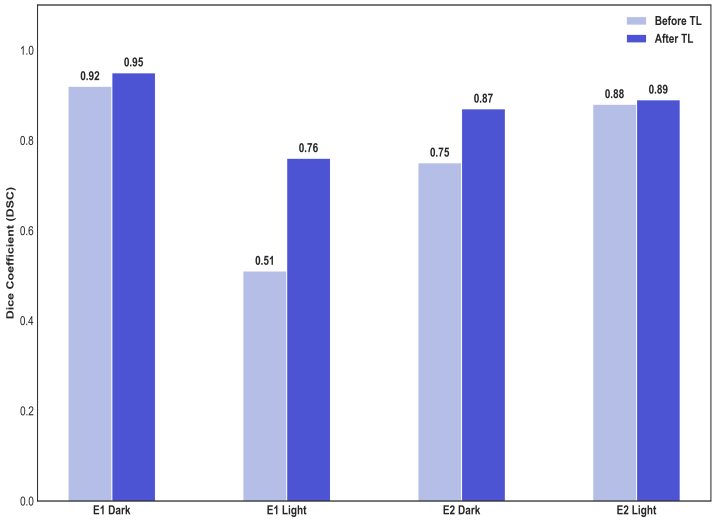

FIGURE 5.4 – Comparaison de performance avant et après transfer learning pour deux réseaux experts sur les catégories de lésions à tons foncés et clairs. E1 désigne l'Expert 1 (spécialisé pour les lésions foncées) et E2 désigne l'Expert 2 (spécialisé pour les lésions claires). La notation E1 Dark représente la performance de l'Expert 1 évaluée sur l'ensemble de test de lésions foncées, tandis que E1 Light représente la performance de ce même expert évaluée sur l'ensemble de test de lésions claires. De manière similaire, E2 Dark et E2 Light représentent respectivement les performances de l'Expert 2 sur les ensembles de test de lésions foncées et claires. Les résultats montrent que le transfer learning améliore la généralisation de chaque expert sur les deux catégories de lésions.

5.3 Protocole Expérimental

Pour évaluer de manière rigoureuse la performance et robustesse du framework MixLVMM proposé, nous avons conçu un protocole expérimental complet comprenant des ensembles de données benchmark divers, des procédures d'entraînement cohérentes, et des critères d'évaluation robustes. Cette section décrit les ensembles de données employés, les paramètres d'entraînement adoptés pour les experts de segmentation et le mécanisme de routage, et les métriques utilisées pour quantifier la performance de segmentation dans une gamme variée de types de lésions et de conditions d'imagerie.

5.3.1 Ensembles de Données

L'évaluation de MixLVMM a été conduite utilisant cinq ensembles de données publics largement reconnus pour la segmentation de lésions cutanées : **ISIC 2016**, **ISIC 2017**,

ISIC 2018, **PH2**, et **DermQuest**. Ces ensembles de données englobent une gamme diverse de morphologies de lésions, variations de pigmentation, et artefacts d'imagerie, les rendant bien adaptés pour valider la performance de généralisation dans des contextes cliniques réels.

Concernant les ensembles ISIC, l'ensemble de données **ISIC 2016** consiste en 900 images d'entraînement et 379 images de test. L'ensemble de données **ISIC 2017** inclut 2,000 images d'entraînement, 150 images de validation, et 600 images de test. Pour **ISIC 2018**, le plus large des trois benchmarks ISIC, 2,594 images d'entraînement, 100 images de validation, et 1,000 images de test sont fournis. Ces trois ensembles de données proviennent de l'International Skin Imaging Collaboration et servent comme benchmarks standardisés pour les modèles de segmentation dermoscopique.

Pour évaluer la généralisation, nous avons évalué la capacité de généralisation de notre modèle en utilisant l'ensemble de données **PH2** (200 images) et l'ensemble de données **DermQuest** (274 images), ce dernier ayant été fusionné avec DermIS pour fournir une diversité de validation supplémentaire. Contrairement à la série ISIC, PH2 et DermQuest ont été utilisés exclusivement pour des fins de validation, nous permettant de tester la capacité de MixLVMM à segmenter des données précédemment non vues provenant de distributions indépendantes.

5.3.2 Paramètres d'Entraînement et d'Évaluation

Le pipeline d'entraînement MixLVMM consiste en deux composants primaires : les réseaux experts de segmentation et le Gate Network. Chaque composant suit une stratégie d'entraînement adaptée appropriée à son rôle au sein de l'architecture.

Prétraitement d'Images. Toutes les images d'entrée des ensembles de données d'entraînement et de validation ont été redimensionnées à une résolution fixe de pixels pour assurer l'uniformité dans le pipeline. Des étapes de prétraitement standard, incluant la normalisation et l'augmentation de données (e.g., flips et rotations aléatoires), ont été appliquées pour améliorer la généralisation.

Experts de Segmentation. Chaque modèle expert a été entraîné sur les données d'entraînement ISIC complètes avant la spécialisation de sous-catégorie. L'entraînement a été conduit en utilisant l'**optimiseur Adam** avec un taux d'apprentissage initial de 0.001. Les modèles ont été entraînés pour jusqu'à 300 époques, avec un arrêt précoce employé basé sur la performance de validation pour prévenir le surapprentissage. L'entraînement a été effectué sur un **GPU NVIDIA RTX A6000**, permettant un entraînement parallèle efficace de multiples experts.

Gate Network. Le Gate Network a été entraîné en utilisant une stratégie d'apprentissage few-shot exploitant la procédure de génération d'anchor basée sur UMAP-k-means.

Seulement un petit ensemble d'images anchor représentatives de chaque cluster a été utilisé pour l'entraînement. Le même optimiseur et taux d'apprentissage ont été appliqués comme dans la phase d'entraînement des experts. La triplet loss a été utilisée pour superviser l'apprentissage d'embedding, assurant une séparation robuste des sous-catégories de lésions dans l'espace latent.

Fonction de Perte. Pour les tâches de segmentation, nous avons employé une fonction de perte composée, combinant *Binary Cross Entropy* (BCE) et *Dice loss*, définie comme :

$$L_{\text{BCE-Dice}} = \alpha L_{\text{BCE}} + \beta L_{\text{Dice}} \tag{5.11}$$

où L_{Dice} est calculée comme :

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \varepsilon} \tag{5.12}$$

Ici, p_i dénote la probabilité prédite pour chaque pixel, g_i l'étiquette de vérité terrain correspondante, N le nombre total de pixels, et ε une constante petite pour la stabilité numérique. Les coefficients α et β équilibrent la contribution des termes BCE et Dice ; nous avons fixé empiriquement les deux à 0.5 dans nos expérimentations.

Cette formulation de perte hybride a été choisie pour optimiser conjointement la classification pixel-wise et la précision de segmentation au niveau de la région, capturant efficacement les propriétés locales et globales des lésions.

Validation Croisée. Pour améliorer la robustesse statistique et réduire le biais d'évaluation, un schéma de validation croisée 4-fold a été employé. La performance rapportée correspond aux scores moyennés sur tous les folds, assurant que les capacités du modèle ne sont pas liées à une partition de données spécifique.

5.3.3 Métriques d'Évaluation

Pour évaluer de manière complète la performance de segmentation de MixLVMM, nous avons adopté une suite de métriques quantitatives largement utilisées, chacune ciblant différents aspects de l'efficacité du modèle :

- **Dice Similarity Coefficient (DSC) :** Mesure le chevauchement entre les masques prédits et de vérité terrain.
- **Jaccard Index (JC) :** Évalue l'intersection-over-union, une mesure plus stricte que DSC.
- **Sensitivity (Sens) :** Évalue la capacité du modèle à identifier correctement les pixels des lésions.
- **Specificity (Spec) :** Quantifie dans quelle mesure le modèle évite les faux positifs.
- **Accuracy (Acc) :** Reflète la justesse de classification globale.

– **Hausdorff Distance (HD95) :** Capture le 95ème percentile des erreurs de distance de surface en pixels, indiquant les déviations maximales pour les images d'entrée 256×256.

Ces métriques fournissent une évaluation équilibrée et multi-perspective de la qualité de segmentation au niveau région et frontière.

5.3.4 Résultats et Discussion

5.3.4.1 Résultats Quantitatifs

Nous présentons les résultats quantitatifs du modèle MixLVMM comparé à 17 modèles de pointe, catégorisés en méthodes basées sur CNN, Transformer, Hybrides, et SSM. Cette évaluation couvre les ensembles de données ISIC 2017, ISIC 2018 et PH2.

Method	Year	DSC (%)	JC (%)	Sens (%)	Spec (%)	Acc (%)	HD 95	Params (M)	TT (h)	IT (ms)
CNN based methods										
UNet [147]	2015	85.11	75.41	87.67	91.33	91.09	31.87	34.4	2.1	527.17
nnUNet [148]	2018	88.15	83.63	91.21	95.39	93.73	27.57	11.2	1.8	438.76
CA-Net [149]	2021	87.37	81.83	90.86	92.13	92.51	28.58	2.8	1.4	21.1
EIU-Net [150]	2023	88.18	82.84	90.34	92.69	93.13	26.88	14.1	1.9	201.43
I^2UNet [151]	2024	90.94	85.76	93.17	95.26	94.41	25.63	7.0	1.6	71.98
Transformer based methods										
SwinUNet [152]	2021	88.22	81.61	90.16	93.29	92.08	30.52	43.9	8.7	615.94
TransUNet [153]	2021	88.74	82.55	90.37	94.06	93.61	30.52	91.4	9.2	1044.1
SLT-Net [154]	2022	87.02	81.70	89.35	92.72	92.46	31.14	60.3	8.9	735.21
FTN [155]	2022	88.25	83.22	91.41	93.34	93.69	28.68	19.9	7.8	281.66
XBound-Former [156]	2023	91.41	85.84	93.24	95.18	94.44	25.59	35.3	8.4	748.99
Hybrid based methods										
TransFuse [157]	2021	90.72	85.29	90.52	95.17	93.88	26.34	34.5	8.1	621.43
FATNet [158]	2022	90.59	85.54	90.27	94.61	93.30	28.11	28.8	7.9	229.23
DEUNet [159]	2023	88.52	83.46	89.83	91.46	92.53	26.54	31.7	8.0	130.99
UCM-NET [160]	2024	86.62	78.61	90.12	88.42	91.55	36.18	**0.049**	**0.91**	**20.97**
SSM's based methods										
Swin-UMamba† [161]	2024	90.32	84.33	91.76	93.71	92.48	29.22	28	5.1	38.5
VM-UNet-V2 [162]	2024	89.85	83.31	91.88	92.90	92.39	29.58	17.9	4.8	526.42
H-vmunet [163]	2025	91.57	85.09	92.87	95.31	94.58	28.96	8.97	4.5	152.79
MixLVMM (Ours)	2025	**92.29**	**86.19**	**93.72**	**95.94**	**94.98**	**25.15**	2.5	4.1	22.73

TABLE 5.1 – Comparaison de différentes méthodes sur l'ensemble de données ISIC 2017.

Les Tableaux 5.1, 5.2, et 5.3 montrent que MixLVMM a surpassé les autres méthodes sur tous les ensembles de données et métriques. Ceci indique une excellente précision de

Method	Year	DSC (%)	JC (%)	Sens (%)	Spec (%)	Acc (%)	HD 95	Params (M)	TT (h)	IT (ms)
CNN based methods										
UNet [147]	2015	86.87	79.25	88.08	95.59	92.44	33.61	34.4	2.1	527.17
nnUNet [148]	2018	89.91	84.00	91.35	93.19	93.06	31.93	11.2	1.8	438.76
CA-Net [149]	2021	89.23	81.44	90.38	96.21	94.90	34.26	2.8	1.4	21.1
EIU-Net [150]	2023	90.12	84.35	93.11	95.04	93.72	32.52	14.1	1.9	201.43
I^2UNet [151]	2024	91.64	85.34	92.84	95.29	94.32	29.02	7.0	1.6	71.98
Transformer based methods										
SwinUNet [152]	2021	89.32	83.09	92.35	95.18	93.91	33.58	43.9	8.7	615.94
TransUNet [153]	2021	89.05	83.49	93.67	95.78	93.28	35.40	91.4	9.2	1044.1
SLT-Net [154]	2022	88.10	82.84	92.33	93.36	91.41	30.10	60.3	8.9	735.21
FTN [155]	2022	89.78	84.72	92.13	95.56	93.24	31.03	19.9	7.8	281.66
XBound-Former [156]	2023	92.42	85.44	93.50	95.91	95.37	25.92	35.3	8.4	748.99
Hybrid based methods										
TransFuse [157]	2021	90.95	84.89	93.61	95.22	94.78	27.95	34.5	8.1	621.43
FATNet [158]	2022	90.06	84.40	92.48	95.70	94.21	34.12	28.8	7.9	229.23
DEUNet [159]	2023	90.51	84.81	92.57	95.98	94.43	33.58	31.7	8.0	130.99
UCM-NET [160]	2024	87.20	81.63	90.96	92.39	91.83	40.16	**0.049**	**0.91**	**20.97**
SSM's based methods										
Swin-UMamba† [161]	2024	90.03	84.27	91.83	95.27	94.65	30.20	28	5.1	38.5
VM-UNet-V2 [162]	2024	89.82	83.65	90.54	94.45	93.76	33.62	17.9	4.8	526.42
H-vmunet [163]	2025	91.17	84.62	93.00	95.14	94.57	28.31	8.97	4.5	152.79
MixLVMM (Ours)	2025	**93.24**	**85.62**	**94.11**	**96.26**	**95.42**	**25.55**	2.5	4.1	22.73

TABLE 5.2 – Comparaison de différentes méthodes sur l'ensemble de données ISIC 2018.

Method	Year	DSC (%)	JC (%)	Sens (%)	Spec (%)	Acc (%)	HD 95	Params (M)	IT (ms)
CNN based methods									
UNet [147]	2015	89.51	82.72	90.14	93.63	92.25	32.66	34.4	527.17
nnUNet [148]	2018	92.27	84.22	93.21	96.45	95.03	30.86	11.2	438.76
CA-Net [149]	2021	92.48	85.51	91.12	96.43	94.24	33.49	2.8	21.1
EIU-Net [150]	2023	93.19	86.64	92.33	96.87	94.54	30.67	14.1	201.43
I^2UNet [151]	2024	94.69	87.88	94.28	96.51	96.11	29.59	7.0	71.98
Transformer based methods									
SwinUNet [152]	2021	93.59	87.37	93.67	94.21	93.83	31.01	43.9	615.94
TransUNet [153]	2021	93.94	86.20	93.73	92.41	93.01	33.15	91.4	1044.1
SLT-Net [154]	2022	91.45	85.24	90.55	93.17	91.68	30.91	60.3	735.21
FTN [155]	2022	91.40	85.13	92.52	95.19	94.87	30.41	19.9	281.66
XBound-Former [156]	2023	94.85	87.64	94.58	97.22	96.81	27.61	35.3	748.99
Hybrid based methods									
TransFuse [157]	2021	94.50	87.09	92.84	95.82	93.94	29.04	34.5	621.43
FATNet [158]	2022	93.10	87.63	92.41	95.72	93.33	34.82	28.8	229.23
DEUNet [159]	2023	92.88	86.67	93.24	94.81	93.55	33.19	31.7	130.99
UCM-NET [160]	2024	89.32	84.37	90.12	90.88	91.23	40.10	**0.049**	**20.97**
SSM's based methods									
Swin-UMamba† [161]	2024	93.60	87.21	93.24	96.57	95.78	30.66	28	38.5
VM-UNet-V2 [162]	2024	92.53	85.91	94.29	95.51	95.28	32.42	17.9	526.42
H-vmunet [163]	2025	94.18	87.52	94.28	97.10	96.33	29.42	8.97	152.79
MixLVMM (Ours)	2025	**95.67**	**88.24**	**94.93**	**97.89**	**96.96**	**26.62**	2.5	22.73

TABLE 5.3 – Comparaison de différentes méthodes sur l'ensemble de données PH2.

segmentation globale et une délimitation de frontières précise, cruciale pour l'analyse d'images médicales.

Analysant les résultats par famille architecturale, un examen plus approfondi révèle des observations clés. Par exemple, sur l'ensemble de données ISIC 2018, parmi les méthodes basées sur CNN, I^{2}UNet performe bien avec un DSC (91.64%) et un JC (85.34%), surpassant UNet et nnUNet. Les méthodes basées sur Transformer comme SwinUNet et TransUNet capturent les dépendances à long terme, avec XBound-Former atteignant la performance la plus haute avec DSC (92.42%) et JC (85.44%). Cependant, ces modèles ont souvent une complexité de calcul élevée et des temps d'inférence plus longs. Les modèles hybrides comme TransFuse et FATNet équilibrent la précision de segmentation et l'efficacité, avec TransFuse atteignant 90.95% sur DSC et 84.89% sur JC.

Concernant les méthodes basées sur SSM, incluant MixLVMM, elles ont montré une performance exceptionnelle. MixLVMM atteint les scores DSC et JC les plus élevés à 93.24% et 85.62%, respectivement, et le score HD95 le plus bas à 25.55. MixLVMM dispose également de moins de paramètres (2.5M) et de temps d'inférence rapides (IT = 22.73ms), le rendant approprié pour les tâches haute-précision et temps-réel. Ces résultats soulignent la performance robuste et l'efficacité de calcul de MixLVMM, comme illustré en Fig.5.5.

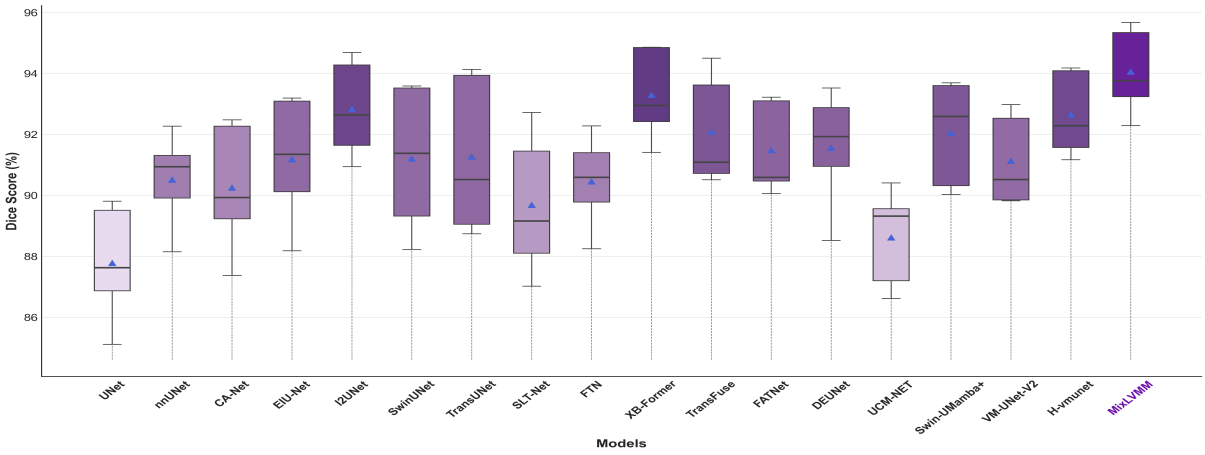

FIGURE 5.5 – Comparaison des Scores Dice sur tous les ensembles de données.

5.3.4.2 Résultats Qualitatifs

Nous démontrons la performance de segmentation de MixLVMM sur des cas difficiles, soulignant sa robustesse et efficacité dans la segmentation précise de lésions cutanées avec des caractéristiques et complexités diverses. La Fig. (5.6) présente quelques exemples difficiles démontrant la performance de MixLVMM comparée à d'autres modèles. Ces résultats visuels soulignent les capacités de segmentation avancées de MixLVMM, en faisant un choix fiable pour les applications cliniques où la précision et l'exactitude sont primordiales.

Examinant les cas complexes, nous avons noté que le modèle MixLVMM a démontré une performance de segmentation supérieure sur tous les ensembles de données dans les cas complexes, tels que les lésions foncées avec la présence d'artefacts (e.g., cheveux, bulles), et les lésions claires se fondant dans la peau environnante. MixLVMM a maintenu de manière consistante des marges précises des lésions et a démontré une spécificité élevée, délimitant efficacement les véritables frontières des lésions et surpassant souvent les autres modèles. Sa performance stable dans différents ensembles de données souligne sa robustesse dans diverses conditions d'imagerie. MixLVMM a géré les lésions foncées et claires avec une efficacité similaire, maintenant une qualité de segmentation élevée indépendamment de la couleur ou texture des lésions.

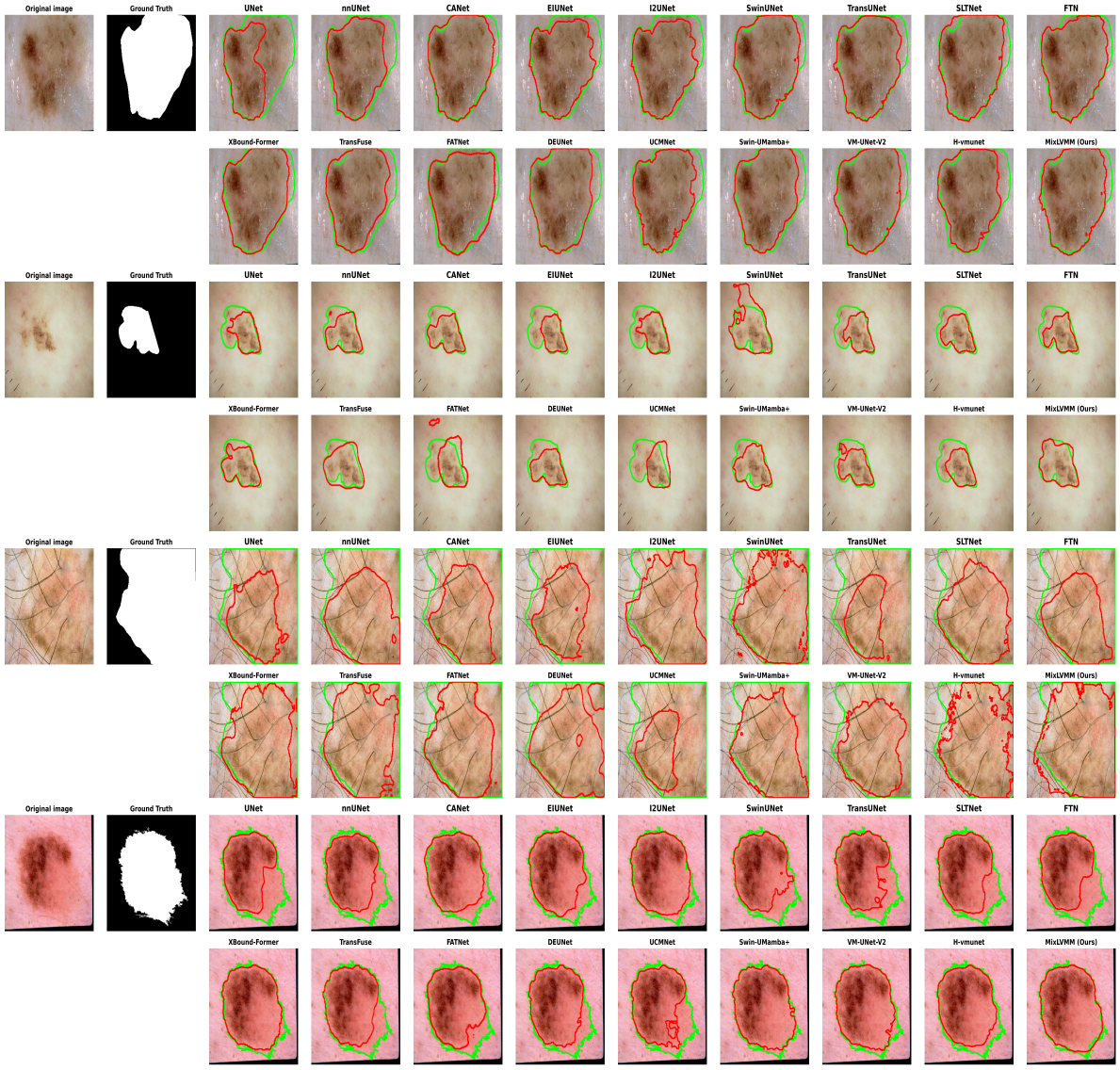

FIGURE 5.6 – Comparaison des résultats qualitatifs sur l'ensemble de test ISIC 2018.

5.3.5 Études d'Ablation

5.3.5.1 Étude d'ablation du Gate Network

Pour évaluer le Gate Network (GN), nous avons conduit une étude d'ablation utilisant l'ensemble de données ISIC 2017 pour un mélange de deux experts. L'étude vise à évaluer la performance de différentes mesures de similarité : distance MSE (L_2), contrastive loss, et triplet loss, et l'effet de la variation du nombre d'exemples d'entraînement (100, 50, 20, 10, et apprentissage 1-shot). Les résultats sont illustrés en Fig.5.7.

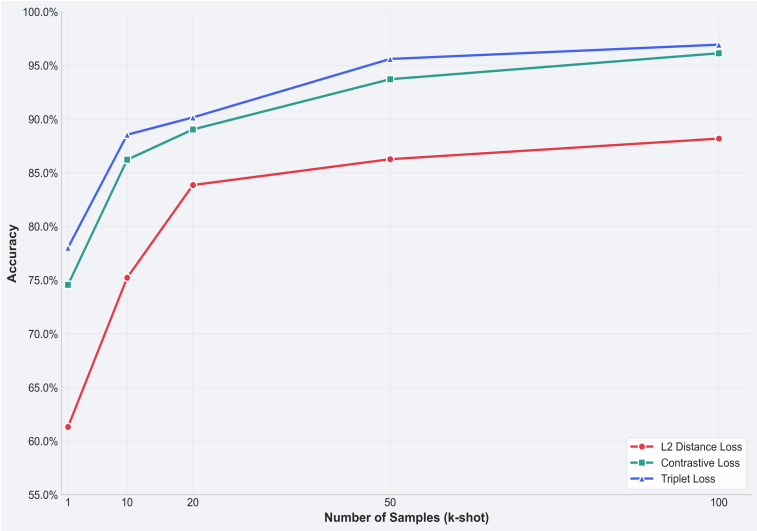

FIGURE 5.7 – Résultats d'apprentissage few-shot (ISIC 2017).

À partir des courbes obtenues, nous observons que l'utilisation de la triplet loss a surpassé les autres pertes dans toutes les tailles d'échantillon. La performance s'est améliorée avec un nombre croissant d'exemples d'entraînement. Cependant, même avec un nombre limité d'exemples (10 ou 20), la triplet loss a montré une précision substantielle, indiquant son potentiel pour les scénarios d'apprentissage few-shot. À l'opposé, les configurations utilisant les pertes L_2 et contrastive ont montré une précision plus faible, particulièrement dans les paramètres one-shot et few-shot. Ceci suggère que bien que ces mesures de similarité puissent être efficaces, elles peuvent nécessiter plus de données d'entraînement pour atteindre une performance comparable à la triplet loss. Globalement, cette étude d'ablation souligne l'importance du choix de la fonction de perte et du nombre d'exemples d'entraînement dans la performance du GN. La fonction triplet loss, en particulier, se distingue comme un choix optimal pour entraîner le GN.

5.3.5.2 Étude d'ablation de l'attention ASRAM

Pour évaluer l'efficacité du module ASRAM, nous l'avons comparé avec d'autres mécanismes d'attention utilisant un mélange de 2 experts et l'ensemble de données ISIC 2018. Les résultats sont résumés dans le Tableau 5.4.

TABLE 5.4 – Comparaison de performance de MixLVMM avec différents mécanismes d'attention sur l'ensemble de données ISIC 2018.

Attention	DSC (%)	JC (%)	Sens (%)	Spec (%)	Acc (%)	HD95
BAM [164]	92.39	85.28	93.51	95.27	94.30	25.87
CBAM [165]	93.12	85.41	93.87	95.73	94.84	25.63
SE [166]	92.86	85.36	93.70	95.51	94.63	25.79
GCT [167]	93.22	85.41	94.03	95.20	95.34	25.57
Triplet.A [168]	92.73	85.19	93.62	95.49	95.26	25.81
ASRAM (Ours)	**93.24**	**85.62**	**94.11**	**96.26**	**95.42**	**25.55**

Nous observons clairement que l'ASRAM a surpassé les autres mécanismes d'attention sur toutes les métriques. Il a atteint les scores les plus élevés pour DSC (93.24%), JC (85.62%), et des améliorations substantielles en Sens (94.11%) et Spec (96.26%), respectivement. De plus, il a obtenu le HD95 le plus bas (25.55), indiquant des frontières de segmentation plus précises. Ces études d'ablation démontrent l'efficacité d'utiliser des réseaux experts spécialisés en conjonction avec le module ASRAM. La configuration MixLVMM Mix-δ (Voir Tableau 5.5), qui incorpore ces deux éléments, fournit une performance de segmentation supérieure, particulièrement pour les cas difficiles avec des frontières complexes et des caractéristiques variées des lésions.

5.3.5.3 Étude d'Ablation des Experts Vision Mamba

Dans la première partie de cette étude, nous nous concentrons sur un mélange de deux experts pour évaluer la performance du modèle MixLVMM sous différents paramètres : Mix-α (un modèle unique pour toutes les données sans ASRAM), Mix-β (un modèle unique pour toutes les données avec ASRAM), Mix-γ (MixLVMM avec version experts sans ASRAM), et Mix-δ (MixLVMM avec version experts avec ASRAM).

Caractérisant les distributions de données, pour montrer les distributions statistiques des données dans le cas de deux experts, nous avons utilisé la procédure discutée en section 5.2.2 pour diviser les lésions en deux sous-catégories qui distinguent majoritairement les lésions "foncées" et "claires". Le Gate Network (GN) est ensuite entraîné avec 10 exemples de triplets. Pour réduire le déséquilibre de données entre les deux catégories, nous avons employé une stratégie d'augmentation de données à deux niveaux. Premièrement, l'augmentation de données statiques est appliquée seulement aux images de lésions claires pour augmenter leur représentation. Ensuite, les augmentations aléatoires en ligne

sont utilisées pour les deux catégories. Étant donné l'importance critique de la couleur en imagerie dermoscopique, nous avons restreint l'augmentation aux transformations géométriques. De plus, nous avons utilisé les masques de vérité terrain pour recadrer les zones de peau saine des images originales, créant une troisième catégorie d'augmentation, incluant les régions de peau saine. Ces échantillons sont annotés avec des masques noirs pour les différencier des images contenant des lésions. La Fig.5.8 présente la distribution des deux catégories dans les ensembles de données.

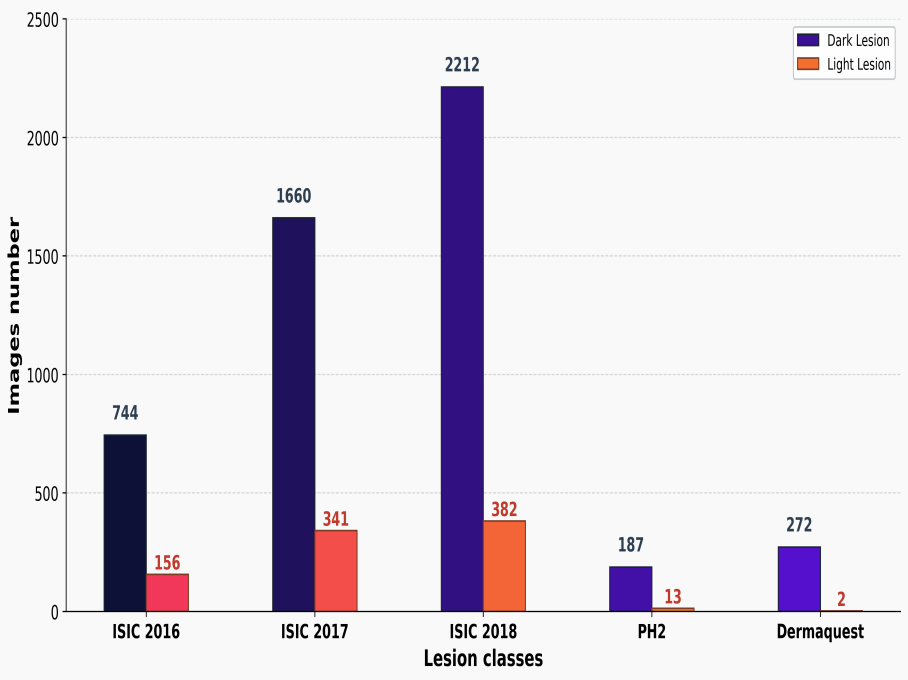

FIGURE 5.8 – Catégorisation des données d'entraînement dans les ensembles de données, excepté les ensembles de données PH2 et DermQuest, qui sont dédiés exclusivement pour l'inférence.

Analysant les configurations, les résultats obtenus pour les modèles comparés sont résumés dans le Tableau 5.5. Nous observons que l'inclusion du module ASRAM améliore de manière substantielle la performance des configurations à modèle unique et expert. Par exemple, la métrique DSC pour Mix-α a augmenté de 90.76% à 91.51% avec l'addition d'ASRAM (Mix-β). De plus, lors de l'incorporation de réseaux de segmentation experts sans ASRAM (Mix-γ), le DSC s'est davantage amélioré à 92.22%. La combinaison de la segmentation expert et du module ASRAM (Mix-δ) a atteint la meilleure performance sur toutes les métriques, avec un DSC de 93.24%, une métrique JC de 85.62%, et un HD95 de 25.55, indiquant une délimitation de frontières plus précise.

Explorant le nombre optimal d'experts, au-delà de l'évaluation de l'efficacité de l'approche basée sur des experts, nous avons conduit des expérimentations supplémentaires pour déterminer le nombre optimal d'experts au sein du framework MixLVMM. L'objec-

TABLE 5.5 – Comparaison de performance de différentes configurations du modèle MixLVMM sur l'ensemble de données ISIC 2018.

Config	DSC	JC	Sens	Spec	Acc	HD95
Mix-α	90.76	83.34	91.55	93.75	93.84	27.49
Mix-β	91.51	84.43	92.66	93.81	94.30	26.71
Mix-γ	92.22	85.12	93.32	95.10	95.18	25.94
Mix-δ	**93.24**	**85.62**	**94.11**	**96.26**	**95.42**	**25.55**

tif était d'évaluer comment la variation du nombre d'experts influence la performance de segmentation et si un nombre croissant de modèles spécialisés mène à des améliorations continues. À cette fin, nous avons entraîné des modèles MixLVMM avec des nombres variés d'experts, allant d'un expert unique à des configurations avec jusqu'à 32 experts. La performance de segmentation a été évaluée sur les ensembles de données ISIC 2017 et ISIC 2018 utilisant la métrique Dice Similarity Coefficient (DSC). La Figure 5.9 présente la distribution des scores DSC pour différents nombres d'experts.

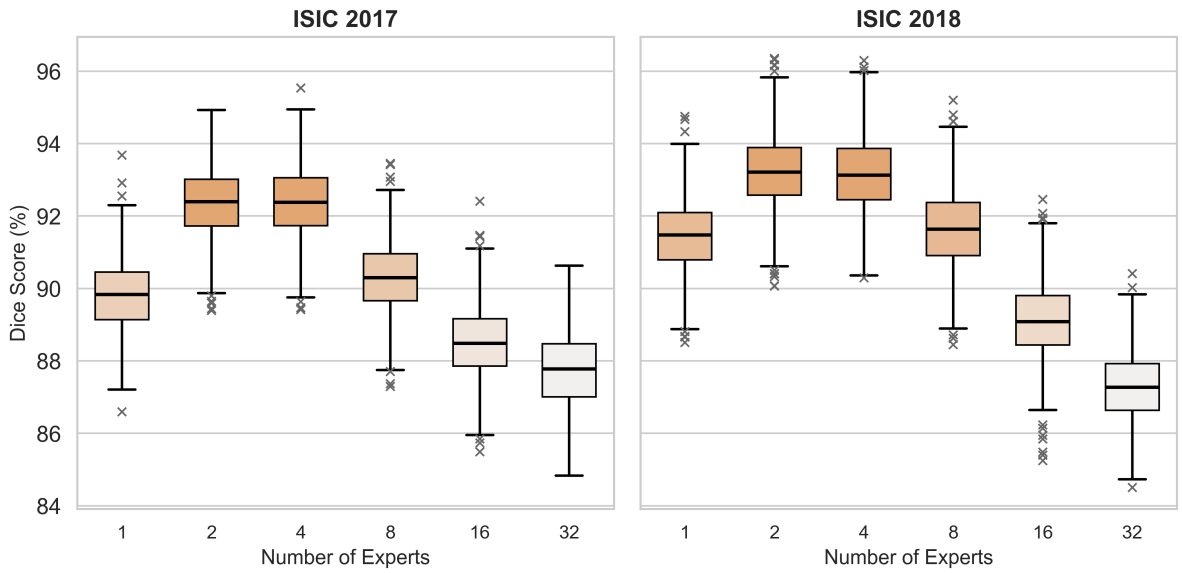

FIGURE 5.9 – Visualisation box plot du Dice Similarity Coefficient (DSC) pour différents nombres d'experts sur les ensembles de données ISIC 2017 et ISIC 2018.

À partir des résultats, nous observons que l'augmentation du nombre d'experts de un à deux mène à une amélioration substantielle en DSC, soulignant l'avantage de la spécialisation de tâche. Cependant, cette tendance ne continue pas indéfiniment. Bien qu'augmenter le nombre d'experts améliore initialement la précision de segmentation, au-delà d'un certain seuil, la performance commence à se dégrader.

Un facteur critique influençant cette dégradation est la taille effective de l'échantillon d'entraînement par expert. Lorsque le nombre d'experts augmente, l'ensemble de données est partitionné en sous-ensembles plus petits, résultant en une réduction du nombre d'échantillons d'entraînement disponibles par expert. Ceci impacte négativement la capacité du modèle à généraliser, car chaque expert reçoit une quantité limitée de données d'entraînement, menant au surapprentissage ou à une incapacité à capturer la variabilité complète de l'ensemble de données. Ce phénomène est évident en Figure 5.9, où les modèles avec 16 et 32 experts exhibent un déclin en performance comparés aux configurations avec moins d'experts.

Ces observations soulignent le compromis entre spécialisation et suffisance de données. Bien que la segmentation basée sur des experts offre des avantages en adaptant l'extraction de caractéristiques à des sous-espaces spécifiques de la distribution de données, la fragmentation excessive des données d'entraînement diminue ces bénéfices. Un équilibre optimal doit donc être maintenu pour assurer que chaque expert ait suffisamment d'échantillons d'entraînement pour apprendre des représentations robustes tout en exploitant la spécialisation. Sur la base de nos expérimentations, la meilleure performance de segmentation a été atteinte avec 2 à 4 experts, où les améliorations DSC ont été maximisées sans perte notable de données d'entraînement par expert. Au-delà de cette plage, des rendements décroissants et une dégradation de performance éventuelle ont été observés en raison de la rareté d'échantillons d'entraînement pour chaque expert. Cette observation est cruciale pour concevoir de futures architectures Mixture-of-Experts (MoE) pour la segmentation d'images médicales, car elle souligne l'importance de maintenir une taille d'échantillon d'entraînement suffisante par expert tout en exploitant les bénéfices de la spécialisation expert.

5.3.5.4 Analyse de Signification Statistique

Pour évaluer la signification statistique des améliorations de performance et renforcer la fiabilité de nos résultats, nous avons conduit des tests statistiques comparant MixLVMM aux baselines les plus performants sur tous les ensembles de données. Le tableau 5.6 présente l'analyse statistique avec les p-values et intervalles de confiance 95% pour les améliorations de coefficient Dice.

L'analyse statistique démontre que MixLVMM atteint des améliorations statistiquement significatives ($p < 0.05$) sur toutes les méthodes baseline les plus performantes dans les trois ensembles de données. Les niveaux de signification stables ($p < 0.001$ pour la plupart des comparaisons) et les intervalles de confiance non-chevauchants fournissent une évidence forte de la robustesse et fiabilité de notre méthode proposée. L'amélioration moyenne de 1.19% en coefficient Dice avec IC 95% [0.86, 1.52] dans tous les ensembles de données confirme la supériorité stable de MixLVMM dans la gestion de présentations diverses de lésions et conditions d'imagerie variées.

TABLE 5.6 – Analyse de Signification Statistique de MixLVMM vs. Baselines les Plus Performants

Datasets	Méthodes	Amélioration Dice (%)	p-value	IC 95%
ISIC 2017	vs. H-vmunet	+0.72	<0.001	[0.34, 1.10]
	vs. XBound-Former	+0.88	<0.001	[0.51, 1.25]
	vs. I²UNet	+1.35	<0.001	[0.95, 1.75]
ISIC 2018	vs. H-vmunet	+2.07	<0.001	[1.62, 2.52]
	vs. XBound-Former	+0.82	<0.001	[0.43, 1.21]
	vs. I²UNet	+1.60	<0.001	[1.18, 2.02]
PH2	vs. H-vmunet	+1.49	<0.001	[0.97, 2.01]
	vs. XBound-Former	+0.82	0.002	[0.31, 1.33]
	vs. I²UNet	+0.98	<0.001	[0.54, 1.42]
Global	**Amélioration Moyenne**	**+1.19**	**<0.001**	**[0.86, 1.52]**

5.4 Conclusion

Ce chapitre a présenté **MixLVMM**, un Mélange innovant et efficace de Modèles Vision Mamba Légers pour une segmentation robuste de lésions cutanées. S'appuyant sur les limitations observées dans notre travail précédent (MEDiXNet), cette approche introduit plusieurs innovations architecturales et méthodologiques visant à traiter les défis critiques en analyse d'images dermoscopiques : haute variabilité inter-lésions, sensibilité aux erreurs de classification d'experts, et contraintes de calcul des systèmes experts traditionnels basés sur CNN.

Les contributions principales reposent sur trois piliers fondamentaux. Premièrement, le remplacement des backbones convolutionnels par des réseaux experts basés sur Vision Mamba légers permet d'atteindre un compromis favorable entre précision et efficacité, autorisant le déploiement d'experts plus spécialisés sans accroître la complexité du modèle. Deuxièmement, l'utilisation d'un Gate Network basé sur triplet-loss avec des anchors automatiquement générés améliore la fiabilité du routage expert et assure une plus grande robustesse dans les cas ambigus ou complexes. Troisièmement, une stratégie d'apprentissage à deux étapes - pré-entraînement global suivi d'un fine-tuning spécialisé, atténue le risque de surapprentissage et d'oubli catastrophique tout en maintenant les capacités de généralisation et discrimination. Dans nos expérimentations sur de multiples ensembles de données, MixLVMM a surpassé un large spectre de modèles de pointe avec des gains particuliers en précision de frontières des lésions et robustesse sous des conditions chargées d'artefacts ou à faible contraste.

Au-delà de ses performances, MixLVMM offre un framework flexible et évolutif pour les systèmes de diagnostic assistés par ordinateur. Sa modularité et interprétabilité le rendent bien adapté aux workflows cliniques temps-réel et ouvrent des perspectives prometteuses pour des extensions futures. La flexibilité architecturale permet une généra-

lisation aisée à d'autres conditions dermatologiques telles qu'eczéma, brûlures et psoriasis, qui exhibent des apparences hétérogènes bénéficiant de la décomposition basée sur experts. Le paradigme mixture of experts peut également être adapté de manière systématique pour gérer la variabilité entre différentes modalités d'imagerie médicale. Pour les applications 2D, le mécanisme de routage peut diriger les images sur la base de caractéristiques spécifiques à la modalité, tandis que pour l'imagerie 3D, l'architecture peut être étendue en remplaçant les blocs Vision Mamba 2D par des contreparties 3D, permettant le traitement volumétrique pour les scans CT, volumes IRM, et données ultrasonores 3D.

Une direction particulièrement prometteuse concerne l'application aux ensembles de données IRM multi-contraste, où différents experts peuvent se spécialiser en séquences de contraste spécifiques (T1, T2, T1-contrast enhanced, et FLAIR). Le réseau de routage apprendrait à diriger chaque type de contraste vers son expert spécialisé, améliorant potentiellement la précision de segmentation pour les tumeurs cérébrales, les lésions de substance blanche, et autres pathologies neurologiques. Au-delà de l'adaptation de modalité, la stratégie d'assignation expert peut être redessinée pour gérer différentes conditions pathologiques au sein de la même modalité d'imagerie, chaque expert apprenant des caractéristiques spécifiques à la pathologie tout en partageant le framework architectural commun.

Les extensions futures pourraient également intégrer de multiples modalités d'entrée (dermoscopie, photos cliniques, imagerie thermique) ou traiter simultanément de multiples tâches (segmentation et classification). Le faible nombre de paramètres du modèle et l'inférence rapide le rendent bien adapté à l'intégration dans les dispositifs portables utilisés dans les contextes à ressources limitées, avec une validation clinique nécessaire pour évaluer son utilisation du point de vue des professionnels de santé.

Pour conclure, MixLVMM constitue une étape substantielle vers la conception de modèles de segmentation efficaces, précis, et généralisables adaptés aux défis d'imagerie médicale.

Chapitre 6

HA-U^{3}Net : Un Framework Agnostique aux Modalités pour la Segmentation d'Images Médicales 3D Utilisant une Structure V-Net Imbriquée et Attention Hybride

6.1 Introduction

6.1.1 Motivation et Contexte

S'appuyant sur le travail fondamental présenté dans les chapitres précédents traitant les défis de variabilité en segmentation d'images médicales, ce chapitre étend notre framework de recherche au domaine complexe de l'imagerie médicale tridimensionnelle. Bien que nos contributions antérieures aient démontré l'efficacité d'architectures adaptatives dans la gestion de l'hétérogénéité au sein d'images dermoscopiques bidimensionnelles, la transition vers les données médicales volumétriques introduit des défis fondamentalement différents qui nécessitent des innovations méthodologiques sophistiquées.

La segmentation d'images médicales tridimensionnelles représente une avancée critique en pratique clinique moderne, permettant l'identification précise de régions pathologiques, le diagnostic complet de maladies, et la délimitation anatomique détaillée dans les ensembles de données volumétriques. Contrairement aux méthodes de segmentation bidimensionnelles, les approches 3D capturent la continuité spatiale essentielle et le contexte volumétrique qui sont indispensables pour les évaluations cliniques complètes. Cette capacité s'avère particulièrement vitale dans les workflows cliniques englobant le diagnostic, la planification de traitement, et la préparation chirurgicale, où l'analyse volumétrique précise impacte directement les résultats patients.

La signification clinique de la segmentation 3D robuste est exemplifiée dans de multiples domaines médicaux. En oncologie, où 316,950 nouveaux cas de cancer du sein et 42,170 décès associés sont projetés pour 2025, l'analyse volumétrique précise devient essentielle pour la détection précoce et le monitoring thérapeutique. De même, l'analyse de tumeurs cérébrales, malgré des taux d'incidence plus faibles, demande une précision exceptionnelle en raison de la nature agressive de ces tumeurs malignes et les complexités inhérentes à leurs protocoles de traitement. Ces réalités cliniques soulignent le besoin urgent de développer des systèmes de segmentation tridimensionnelle robustes, rapides, et précis capables d'améliorer la précision diagnostique et de guider des interventions cliniques efficaces.

6.1.2 Défis de Segmentation Tridimensionnelle

La progression de la segmentation d'images médicales 2D vers 3D introduit une multitude de défis interconnectés qui remodèlent fondamentalement les exigences de calcul et méthodologiques des systèmes de segmentation. Les approches de segmentation 3D traditionnelles s'appuient souvent sur des caractéristiques conçues manuellement et des heuristiques spécifiques au domaine, limitant de manière substantielle leur généralisabilité dans diverses applications médicales et modalités d'imagerie. Ces méthodes conventionnelles exhibent une spécificité de tâche inhérente, nécessitant des connaissances expertes substantielles tout en offrant une évolutivité limitée pour traiter les défis cliniques émergents.

Au niveau de la complexité de calcul, le traitement de données volumétriques représente un obstacle primaire en segmentation tridimensionnelle. Bien que les CNNs 2D démontrent une efficacité pour les images médicales planaires, ils éprouvent des difficultés à capturer les relations inter-slice critiques qui définissent les ensembles de données volumétriques tels que l'imagerie par résonance magnétique ou les scans CT. Les réseaux convolutionnels 3D traitent cette limitation par des kernels volumétriques qui capturent la continuité spatiale entre les slices, améliorant ainsi la qualité de segmentation et la compréhension structurelle. Cependant, cette amélioration nécessite des augmentations substantielles en exigences de calcul et en mémoire, créant des défis de déploiement significatifs dans les environnements cliniques à ressources contraintes.

Concernant la variabilité inter-modalités, les variations entre modalités présentent un autre défi fondamental en segmentation d'images médicales 3D. Différentes modalités d'imagerie exhibent des caractéristiques distinctes en termes de résolution, profils de contraste, patterns de bruit, et distributions d'artefacts. Par exemple, la segmentation de tumeurs cérébrales à partir d'imagerie par résonance magnétique demeure difficile en raison des frontières tumorales irrégulières, patterns d'infiltration tissulaire, et problèmes spécifiques à la modalité, incluant l'inhomogénéité d'intensité, le contraste tissulaire variable, et les artefacts de bruit. De même, les scans CT peuvent exhiber un contraste limité pour la différenciation de tissus mous, tandis que l'imagerie par ultrasons souffre

de bruit de granularité, d'ombre acoustique, et d'autres artefacts qui compromettent la précision de segmentation.

La complexité anatomique inhérente à l'imagerie médicale 3D aggrave ces défis. Segmenter diverses régions anatomiques introduit une variabilité substantielle en forme, taille, et arrangement spatial, aggravée par les inconsistances provenant de différents protocoles d'imagerie et paramètres d'acquisition. Cette complexité devient particulièrement prononcée dans les régions abdominales et thoraciques, où les organes chevauchants et les relations spatiales complexes entravent de manière substantielle la délimitation précise des frontières, souvent amplifiée par le mouvement patient, les implants métalliques, ou d'autres facteurs confondants.

6.1.3 Lacune de Recherche et Limitations des Méthodes Existantes

Malgré des progrès substantiels en segmentation d'images médicales basée sur l'apprentissage profond, les modèles de segmentation 3D de pointe actuels souffrent de limitations architecturales fondamentales qui restreignent leur applicabilité clinique et évolutivité. La limitation primaire réside dans leur philosophie de conception spécifique à la tâche, où les modèles optimisés pour des scénarios de segmentation particuliers démontrent une transférabilité pauvre lorsqu'appliqués à différentes cibles anatomiques ou contextes d'imagerie.

Les architectures de segmentation 3D contemporaines exhibent des exigences d'adaptation sévères lors de la transition entre tâches de segmentation de complexité variée. Par exemple, les modèles conçus pour la segmentation de petites lésions (telle que la détection de tumeurs du sein) nécessitent des modifications architecturales substantielles, couches additionnelles, ou composants de post-traitement spécialisés pour segmenter efficacement de grandes structures anatomiques comme les poumons ou organes abdominaux. Cette limitation provient de leurs configurations de champ récepteur fixes et mécanismes d'extraction de caractéristiques spécifiques à l'échelle qui ne peuvent s'adapter dynamiquement aux caractéristiques spatiales diverses inhérentes aux différentes structures anatomiques.

L'optimisation spécifique à la modalité des modèles existants représente une autre limitation critique. Les architectures actuelles nécessitent typiquement des modifications d'ingénierie extensives, incluant des couches de normalisation spécialisées, pipelines de prétraitement spécifiques au domaine, ou ajustements architecturaux lorsqu'appliqués à différentes modalités d'imagerie (IRM vers CT, ultrason vers PET). Ceci nécessite des efforts de re-engineering substantiels plutôt qu'un simple réentraînement, créant des barrières significatives au déploiement clinique généralisé où de multiples modalités d'imagerie sont couramment employées.

Les modèles existants présentent également une polyvalence limitée dans l'extraction de caractéristiques, s'appuyant souvent sur des représentations de caractéristiques superficielles ou étroites qui échouent à capturer les patterns anatomiques multi-échelles

riches requis pour une généralisation robuste entre tâches. Leur rigidité architecturale prévient le transfert de connaissances efficace dans différents scénarios de segmentation, forçant les chercheurs et cliniciens à développer des modèles spécialisés pour chaque nouvelle application plutôt que d'exploiter des frameworks unifiés et adaptables.

Ce manque fondamental d'agnosticité architecturale entraîne une prolifération de modèles spécialisés, chacun nécessitant des ressources de calcul dédiées, une surcharge de maintenance importante, et des connaissances expertes pour le déploiement. L'absence d'architectures véritablement agnostiques aux modalités et adaptatives à l'échelle représente une lacune majeure qui limite l'évolutivité pratique de la segmentation d'images médicales 3D dans des environnements cliniques divers.

6.1.4 Objectifs du Chapitre

Ce chapitre traite les limitations identifiées en étendant notre framework de thèse pour gérer la variabilité en imagerie médicale au domaine 3D. Notre objectif primaire consiste à développer un framework de segmentation agnostique aux modalités capable de maintenir une haute performance dans diverses modalités d'imagerie tout en traitant les défis de calcul et méthodologiques inhérents au traitement de données volumétriques.

Plus précisément, ce travail vise à démontrer comment nos principes établis pour gérer la variabilité d'imagerie médicale peuvent être efficacement adaptés pour gérer la complexité accrue des données tridimensionnelles. En développant le framework HA-U^{3}Net, nous cherchons à fournir une solution unifiée qui maintient une performance robuste dans les ensembles de données IRM, CT, ultrasons, et tomographie par émission de positrons, validant ainsi la généralisabilité de notre approche de gestion de variabilité.

Les objectifs supplémentaires du chapitre incluent l'établissement de standards d'efficacité de calcul pour la segmentation 3D qui permettent un déploiement clinique pratique, tout en atteignant simultanément une précision de segmentation de pointe dans de multiples régions anatomiques et conditions pathologiques.

6.1.5 Contributions Clés

Ce chapitre présente plusieurs contributions majeures au domaine de la segmentation d'images médicales 3D. Nous introduisons l'architecture HA-U^{3}Net, un framework 3D innovant qui incorpore l'extraction de caractéristiques multi-échelles directement au sein de blocs de traitement individuels imbriqués. Cette conception traite fondamentalement les limitations d'architectures convolutionnelles traditionnelles en permettant la capture simultanée de détails fins et d'informations contextuelles de haut niveau essentielles pour une segmentation volumétrique précise.

Notre seconde contribution majeure concerne le développement d'un mécanisme d'attention hybride qui combine l'attention spatiale et par canal, spécifiquement optimisé pour les données médicales 3D. Contrairement aux mécanismes d'attention existants qui se concentrent sur les relations spatiales ou de canaux indépendamment, notre ap-

proche hybride priorise dynamiquement les caractéristiques volumétriques sur la base des variations de structures anatomiques dans différentes modalités d'imagerie.

De plus, nous présentons une validation expérimentale complète démontrant des capacités de généralisation inter-modalités exceptionnelles, avec des améliorations de performance constantes dans diverses modalités d'imagerie, incluant l'IRM, le CT, les ultrasons, et la tomographie par émission de positrons. Finalement, nous introduisons U^{3}Mamba, une variante légère qui maintient une performance de segmentation élevée tout en réduisant de manière substantielle la complexité de calcul, améliorant ainsi la faisabilité de déploiement clinique pratique.

6.2 Méthodologie

L'architecture HA-U^{3}Net comprend deux innovations fondamentales qui traitent les limitations identifiées dans les approches de segmentation 3D existantes. Le premier composant consiste en des blocs U^3 imbriqués qui intègrent des structures encodeur-décodeur hiérarchiques au sein d'unités de traitement individuelles, permettant l'extraction simultanée de caractéristiques multi-échelles dans les volumes tridimensionnels. Le second composant implémente un mécanisme d'attention hybride combinant l'attention spatiale et par canal pour mettre l'accent de manière adaptative sur les structures cliniquement pertinentes dans diverses modalités d'imagerie.

Cette conception architecturale traite directement les limitations fondamentales des approches 3D conventionnelles, qui opèrent typiquement à résolutions fixes et peinent avec la généralisation inter-modalités, en fournissant des capacités d'extraction de caractéristiques robustes qui maintiennent une performance stable dans de multiples modalités d'imagerie sans nécessiter de modifications architecturales.

6.2.1 Architecture U^3-Net Imbriquée

L'architecture fondamentale de HA-U^{3}Net s'appuie sur le succès établi de U^{2}Net [169] dans la gestion de scénarios à faible contraste, étendant ses capacités pour capturer efficacement tant les détails fins locaux que l'information contextuelle globale au sein de données médicales volumétriques 3D. Cette extension représente une avancée architecturale majeure, spécifiquement conçue pour traiter les hiérarchies spatiales complexes et les exigences de caractéristiques multi-échelles inhérentes à la segmentation d'images médicales volumétriques dans diverses modalités d'imagerie.

Le framework architectural central emploie une adaptation 3D sophistiquée de blocs U résiduels, englobant tant les variantes RSU standard que RSU4F dilatées, qui permettent collectivement l'extraction efficace de caractéristiques spatiales multi-échelles à partir de volumes médicaux complexes. L'architecture initie le traitement par une opération de convolution 3D initiale qui transforme le volume d'entrée en représentations de caractéristiques intermédiaires, établissant ainsi des cartes de caractéristiques compactes

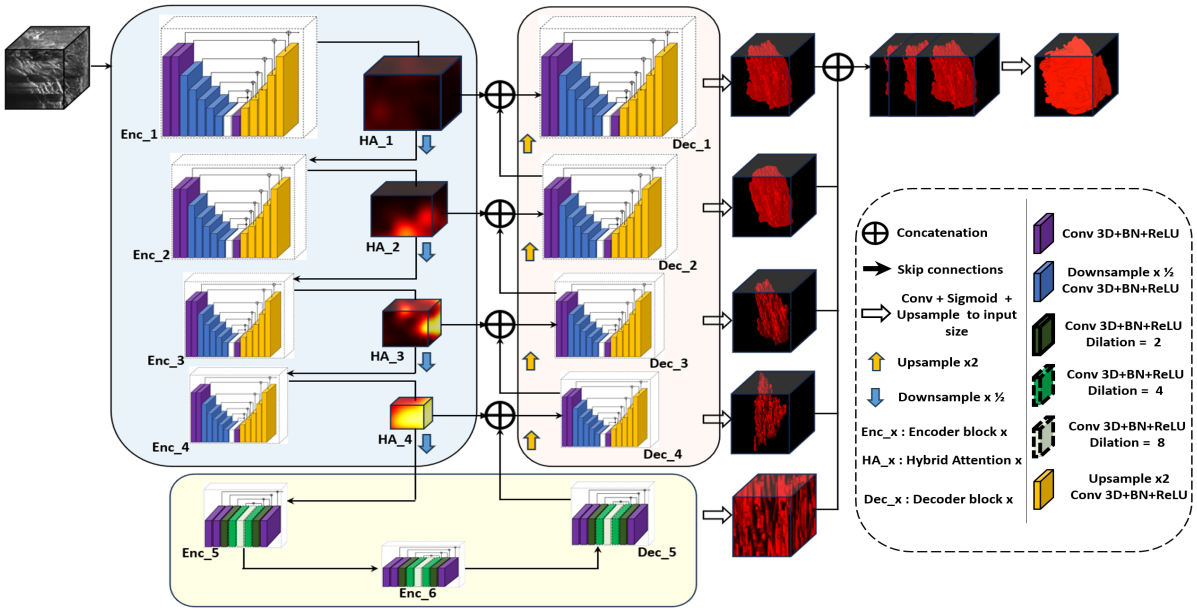

FIGURE 6.1 – L'architecture en forme de U^3 de HA-U^3Net conçue pour une extraction de caractéristiques multi-échelles efficace. La structure encodeur-décodeur exploite la représentation de caractéristiques hiérarchiques avec des connexions skip pour retenir les détails spatiaux dans les résolutions. Une supervision profonde est incorporée à de multiples étapes de décodeur, améliorant le flux de gradient durant l'entraînement et renforçant la précision de segmentation.

mais hautement expressives qui servent de fondation pour le traitement hiérarchique ultérieur.

Chaque bloc architectural au sein du framework transforme un volume d'entrée V de dimensions $C_{\text{in}} \times H \times W \times D$ en une carte de caractéristiques intermédiaire raffinée $F(V)$ caractérisée par C_{out} canaux, où C_{in} représente le nombre de canaux d'entrée (typiquement 1 pour l'imagerie médicale en niveaux de gris ou de multiples canaux pour les acquisitions multi-contraste), et D, H, W correspondent aux dimensions de profondeur, hauteur, et largeur du volume d'entrée, respectivement. Ce processus de transformation permet à l'architecture de maintenir la cohérence spatiale tout en extrayant progressivement des représentations de caractéristiques de plus en plus abstraites dans de multiples échelles de résolution.

Pour améliorer les capacités de discrimination de caractéristiques, particulièrement autour de frontières tumorales subtiles et structures anatomiques complexes qui exhibent souvent un faible contraste ou une délimitation ambiguë, des blocs d'attention hybrides sont intégrés stratégiquement dans l'architecture. Ces mécanismes d'attention fournissent un raffinement de caractéristiques ciblé qui s'avère particulièrement précieux lors du traitement de régions difficiles caractérisées par l'hétérogénéité tissulaire, les variations d'intensité, ou les artefacts spécifiques à la modalité qui peuvent compromettre la précision de segmentation dans les approches conventionnelles.

6.2.1.1 Structure de Bloc U^3

Le bloc U^3 constitue l'unité de construction fondamentale de l'architecture HA-U^{3}Net, représentant une extension sophistiquée du framework conceptuel U^{2}Net bidimensionnel en une structure U imbriquée tridimensionnelle complète. Spécifiquement conçu pour le traitement de données volumétriques, ce composant architectural intègre des voies encodeur-décodeur hiérarchiques complètes au sein de blocs de traitement individuels, permettant ainsi l'extraction simultanée de caractéristiques multi-échelles, la fusion intelligente de caractéristiques, et le raffinement progressif de caractéristiques. Cette philosophie de conception distingue fondamentalement le bloc U^3 des blocs convolutionnels conventionnels qui opèrent à résolutions spatiales fixes, car il démontre une capacité supérieure à capturer tant les détails anatomiques fins locaux que l'information contextuelle plus large essentielle pour modéliser efficacement les hiérarchies spatiales complexes et relations inter-structurelles qui caractérisent les données d'imagerie médicale 3D.

Limitations du Traitement Convolutionnel Conventionnel : Les architectures de blocs convolutionnels traditionnels traitent les cartes de caractéristiques d'entrée $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times H \times W \times D}$ par des séquences de T couches convolutionnelles empilées, où chaque opération de convolution individuelle peut être mathématiquement exprimée comme :

$$\mathbf{X}_{l+1} = \sigma(\mathbf{W}_l * \mathbf{X}_l + \mathbf{b}_l), \tag{6.1}$$

où $\mathbf{W}_l$ et $\mathbf{b}_l$ représentent les poids du kernel de convolution et paramètres de biais de la l-ème couche respectivement, et $\sigma(\cdot)$ dénote la fonction d'activation non-linéaire

telle que ReLU. Ces opérations de convolution empilées opèrent intrinsèquement à résolutions spatiales fixes, transformant et raffinant progressivement les caractéristiques d'entrée tout en maintenant des dimensions spatiales cohérentes dans le pipeline de traitement, à moins que des opérations explicites de down-sampling ou up-sampling (telles que pooling ou convolutions transposées) soient délibérément incorporées. Cependant, cette approche conventionnelle traite fondamentalement les caractéristiques dans des espaces à résolution unique par couche, ce qui contraint de manière substantielle la capacité du réseau à capturer des représentations hiérarchiques et relations spatiales multi-échelles efficacement.

Conception de Bloc U^3 Imbriqué : L'architecture de bloc U^3 traite ces limitations fondamentales par sa conception innovante de bloc en forme de U imbriqué, où les capacités de traitement multi-échelles et l'organisation structurelle imbriquée fournissent une compréhension complète des hiérarchies spatiales et contextuelles inhérentes aux données médicales volumétriques. Le bloc U^3 incorpore des voies encodeur-décodeur opérant dans $L = 4$ échelles distinctes, utilisant des opérations de down-sampling coordonnées ($\mathcal{D}$), opérations d'up-sampling ($\mathcal{U}$), et transformations convolutionnelles ($\mathcal{F}_{\text{conv}}$) pour traiter les données d'entrée dans de multiples niveaux de résolution simultanément.

La phase d'encodage du bloc U^3 traite les échelles d'entrée selon la formulation mathématique suivante :

$$\begin{aligned} \mathbf{Z}_0 &= \mathcal{F}_{\text{conv}}(\mathbf{X}_l), \\ \mathbf{Z}_1 &= \mathcal{F}_{\text{conv}}(\mathcal{D}(\mathbf{Z}_0)), \\ &\vdots \\ \mathbf{Z}_{L-1} &= \mathcal{F}_{\text{conv}}(\mathcal{D}(\mathbf{Z}_{L-2})), \end{aligned} \tag{6.2}$$

où L représente le nombre total de niveaux de traitement imbriqués au sein de l'architecture de bloc. Lors de l'atteinte du niveau de représentation de caractéristiques le plus grossier, le bloc initie la phase de décodage par des opérations systématiques d'upsampling et de fusion intelligente de caractéristiques :

$$\begin{aligned} \mathbf{Z}'_{L-2} &= \mathcal{F}_{\text{conv}}(\mathcal{U}(\mathbf{Z}_{L-1})) + \mathbf{Z}_{L-2}, \\ &\vdots \\ \mathbf{Z}'_1 &= \mathcal{F}_{\text{conv}}(\mathcal{U}(\mathbf{Z}'_2)) + \mathbf{Z}_1 \\ \mathbf{X}_{l+1} &= \mathcal{F}_{\text{conv}}(\mathcal{U}(\mathbf{Z}'_1)) + \mathbf{Z}_0 \end{aligned} \tag{6.3}$$

Intégration de Caractéristiques Multi-Échelles : Pour toute carte de caractéristiques d'entrée donnée $\mathbf{X}_l$, le bloc U^3 traite les données dans de multiples échelles spatiales via des voies coordonnées de downsampling et upsampling. Durant chaque étape de downsampling, les caractéristiques subissent des opérations de compression conçues pour capturer l'information contextuelle globale, tandis que des connexions skip positionnées stratégiquement transfèrent les représentations de caractéristiques intermédiaires vers les

étapes d'upsampling correspondantes, assurant que les détails anatomiques haute résolution sont systématiquement réintégrés dans le pipeline de traitement. Ces connexions skip facilitent la réutilisation directe de caractéristiques dans les niveaux de traitement imbriqués, produisant ultimement une sortie $\mathbf{X}_{l+1}$ qui combine efficacement le détail anatomique fin local avec l'information contextuelle plus large essentielle pour une segmentation volumétrique précise. Cette architecture imbriquée permet le traitement simultané dans de multiples échelles tout en maintenant l'efficacité paramétrique par la réutilisation systématique de poids dans les niveaux hiérarchiques.

6.2.1.2 Mécanisme d'Attention Hybride

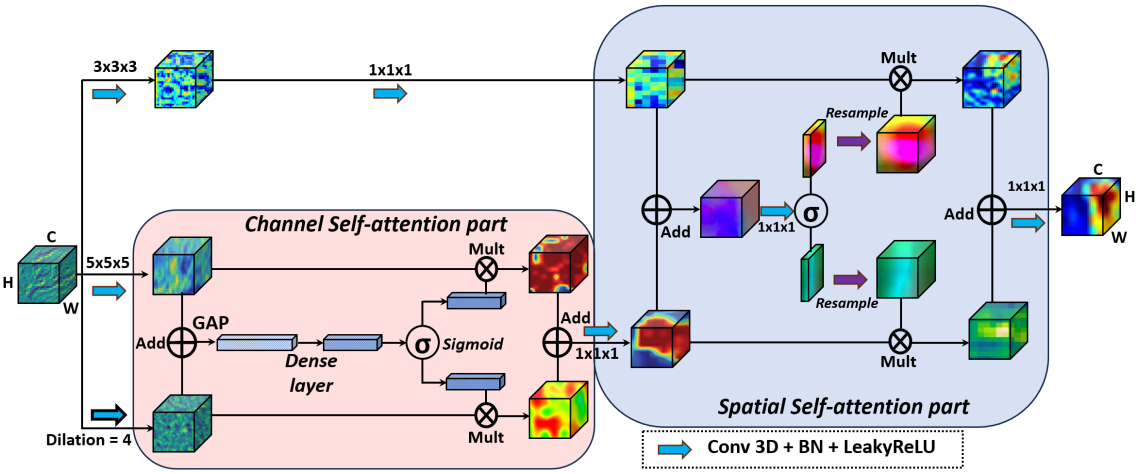

FIGURE 6.2 – Architecture de bloc Hybrid Attention (HA), combinant l'attention de canal et spatiale pour améliorer la représentation de caractéristiques pour l'imagerie médicale 3D.

Une innovation architecturale fondamentale de HA-U^{3}Net réside dans le bloc sophistiqué *Hybrid Attention* (HA), stratégiquement positionné à la conclusion de chaque étape d'encodeur au sein de l'architecture. Ce mécanisme d'attention fonctionne comme un filtre intermédiaire intelligent qui construit systématiquement des descriptions multi-échelles complètes des cartes de caractéristiques d'entrée avant d'appliquer des opérations séquentielles de self-attention de canal suivies d'opérations spatiales spécifiquement conçues pour mettre l'accent sur les structures anatomiques diagnostiquement pertinentes et régions pathologiques au sein de données volumétriques.

Extraction de Caractéristiques Multi-Échelles : Étant donné un tenseur de caractéristiques d'encodeur $F \in \mathbb{R}^{C\times D\times H\times W}$, le mécanisme d'attention hybride initie le traitement par une extraction de caractéristiques multi-échelles sophistiquée utilisant quatre voies convolutionnelles parallèles conçues pour capturer des caractéristiques à différentes échelles spatiales et tailles de champ récepteur. Ces voies emploient des opérations de convolution $3 \times 3 \times 3$, $5 \times 5 \times 5$, dilatées ($3 \times 3 \times 3$, dilation=4), et $1 \times 1 \times 1$,

mathématiquement dénotées comme ϕ_3, ϕ_5, ϕ_d, et ϕ_1, respectivement. Cette stratégie de traitement parallèle génère des cartes de caractéristiques correspondantes F_3, F_5, F_d, et F_1, chacune capturant des aspects distincts de détail structurel et information contextuelle essentielle pour une représentation de caractéristiques complète.

Calcul d'Attention de Canal : Le composant d'attention de canal du mécanisme hybride opère en fusionnant intelligemment les caractéristiques à grande échelle dérivées des voies F_5 et F_d, qui sont ensuite traitées par des opérations de global average pooling suivies d'un réseau perceptron multicouche (MLP) léger. Ce calcul d'attention de canal peut être mathématiquement exprimé comme :

$$M_c = \sigma\left(\mathrm{MLP}(\mathrm{GAP}(F_5 + F_d))\right), \quad F_c = \phi_1\left(M_c \otimes (F_5 + F_d)\right),$$

où $\otimes$ représente les opérations de multiplication élément par élément et σ dénote la fonction d'activation sigmoïde. Ce mécanisme de pondération de canal priorise efficacement les canaux sur la base de leur importance relative pour la tâche de segmentation, produisant un tenseur de caractéristiques raffiné F_c qui met l'accent sur les représentations de caractéristiques par canal les plus diagnostiquement pertinentes.

Intégration d'Attention Spatiale : Suite au traitement d'attention de canal, l'attention spatiale est appliquée par la fusion intelligente du tenseur raffiné par canal F_c avec les caractéristiques de voie haute résolution F_1. Ce mécanisme d'attention spatiale est formulé comme :

$$M_s = \sigma\left(\phi_1(F_c + F_1)\right), \quad F_s = \phi_1\left(M_s \otimes (F_c + F_1)\right),$$

produisant le tenseur de sortie final avec attention hybride : $Y_{\mathrm{HA}} = F_s$. Cette représentation de tenseur enrichie par attention est ensuite transmise à l'étape de décodeur correspondante par des connexions skip positionnées stratégiquement, assurant que les caractéristiques guidées par attention raffinées sont efficacement intégrées dans la voie de décodage.

Applications Cliniques et Bénéfices Inter-Modalités : Par la combinaison séquentielle de mécanismes d'attention de canal et spatial, le bloc HA permet à HA-U^{3}Net de démontrer une capacité exceptionnelle à se concentrer sur les frontières tumorales subtiles et régions à faible contraste dans diverses modalités d'imagerie. Cette stratégie d'attention double améliore de manière substantielle la performance de segmentation dans les zones anatomiques complexes ou ambiguës où les approches conventionnelles montrent leurs limites, notamment dans les régions caractérisées par l'hétérogénéité tissulaire, les variations d'intensité, ou les artefacts spécifiques à la modalité. Le mécanisme d'attention hybride représente ainsi un composant crucial pour atteindre une performance inter-modalités robuste tout en maintenant une efficacité de calcul appropriée pour les scénarios de déploiement clinique.

6.2.2 Variante U^{3}Mamba

S'appuyant sur ces avancées fondamentales en modélisation state-space des modèles vision mamba présentés dans les chapitres précédents, nous introduisons U^{3}Mamba,

une variante légère mais puissante de HA-U^{3}Net spécifiquement conçue pour réduire la complexité de calcul tout en maintenant des capacités d'extraction de caractéristiques robustes essentielles pour une segmentation d'images médicales 3D précise. Dans cette conception architecturale innovante, les blocs résiduels convolutionnels (incluant tant les variantes RSU que RSU4F) sont systématiquement remplacés par des blocs Residual U Mamba spécialisés, qui intègrent la couche sophistiquée Tri-orientated Spatial Mamba (ToM) pour fournir des capacités de traitement directionnelles complètes optimisées pour les données médicales volumétriques.

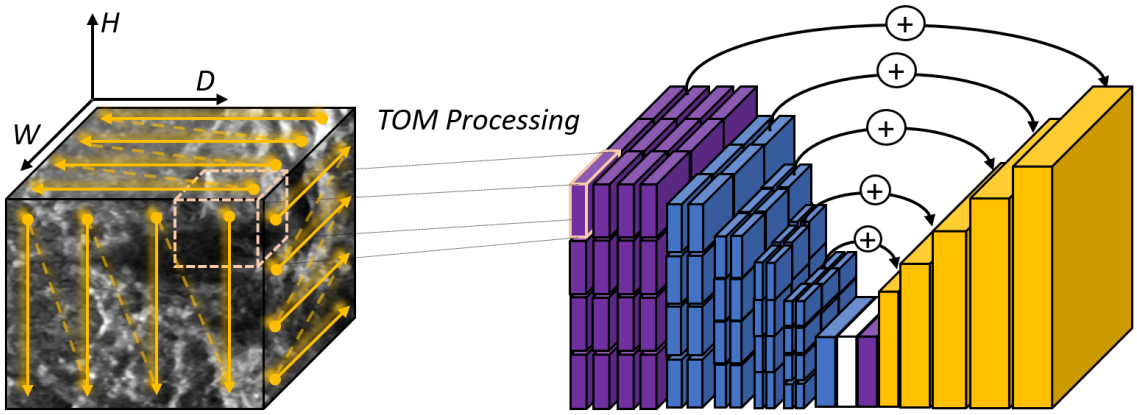

FIGURE 6.3 – Bloc Tri-orientated Spatial Mamba dans le module U^3.

Formulation State-Space pour Vision 3D : L'architecture U^{3}Mamba s'appuie sur le framework établi du modèle Vision Mamba en intégrant une formulation state-space sophistiquée spécifiquement adaptée pour les tâches de vision tridimensionnelle. À son cœur mathématique fondamental, les séquences de caractéristiques d'entrée $\mathbf{x}_t$ sont modélisées par les équations state-space suivantes :

$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}\mathbf{h}_t,$$

où $\mathbf{h}_t \in \mathbb{R}^N$ représente le vecteur d'état caché au pas de temps t, et $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times D}$, $\mathbf{C} \in \mathbb{R}^{D\times N}$ constituent des matrices de projection apprenables qui permettent la transformation adaptative de caractéristiques et l'évolution d'état. La sortie $\mathbf{y}_t$ dénote la représentation de caractéristiques traitées au pas de temps t, permettant le traitement séquentiel d'information spatiale tout en maintenant la conscience de contexte global.

Module Gated Spatial Convolution : Pour préserver la structure spatiale 3D essentielle dans le pipeline de traitement state-space, chaque bloc U^{3}Mamba initie le traitement par un module sophistiqué Gated Spatial Convolution (GSC) qui fusionne intelligemment les caractéristiques spatiales via des opérations de multiplication élément par élément entre des voies de convolution parallèles $3\times3\times3$ et $1\times1\times1$. Ce mécanisme de gating peut être mathématiquement exprimé comme :

$$\mathrm{GSC}(\mathbf{Z}) = \mathbf{Z} + \phi_3(\phi_3(\mathbf{Z}) \odot \phi_1(\mathbf{Z})),$$

où $\odot$ dénote les opérations de gating élément par élément qui permettent l'amélioration sélective de caractéristiques sur la base des caractéristiques spatiales. Cette stratégie assure que l'information spatiale est préservée et améliorée dans les étapes de traitement state-space ultérieures.

Traitement Tri-orientated Mamba : Suite au traitement de convolution spatiale initial, la couche Tri-orientated Mamba (ToM), inspirée du travail SegMamba [170], capture l'information contextuelle directionnelle complète en décomposant systématiquement le tenseur de caractéristiques d'entrée $\mathbf{Z}$ en trois vues anatomiques correspondant aux orientations axiales, coronales, et sagittales. Chaque orientation est traitée indépendamment utilisant des opérations Mamba spécialisées, avec le traitement global formulé comme :

$$\Psi_{\text{ToM}}(\mathbf{Z}) = \sum_{k=1}^{3} \Omega_k(\mathcal{F}_k(\mathbf{Z})),$$

où $\mathcal{F}_k$ représente les opérations d'aplatissement de tenseur le long de l'orientation $k \in \{\text{axial}, \text{coronal}, \text{sagittal}\}$, et Ω_k applique la transformation state-space correspondante le long de l'orientation anatomique spécifiée. Cette stratégie de traitement tri-orientée assure la capture complète de dépendances spatiales dans tous les plans de visualisation anatomiques majeurs.

Pipeline de Traitement Complet : Pour un bloc U^{3}Mamba traitant l'entrée $\mathbf{X}_l$ au niveau d'imbrication l, le pipeline de traitement hiérarchique complet opère selon la formulation mathématique suivante :

$$\begin{aligned}
\mathbf{Z}_0 &= \text{GSC}(\mathbf{X}_l), \\
\mathbf{Z}_1 &= \Psi_{\text{ToM}}(\mathcal{D}(\text{GSC}(\mathbf{Z}_0))), \\
&\vdots \\
\mathbf{Z}_{L-1} &= \Psi_{\text{ToM}}(\mathcal{D}(\text{GSC}(\mathbf{Z}_{L-2}))),
\end{aligned} \tag{6.4}$$

où L représente le nombre total de niveaux de traitement imbriqués. Après avoir atteint le niveau de représentation de caractéristiques le plus grossier, le bloc effectue des opérations systématiques d'upsampling et de fusion intelligente de caractéristiques :

$$\begin{aligned}
\mathbf{Z}'_{L-2} &= \Psi_{\text{ToM}}(\mathcal{U}(\mathbf{Z}_{L-1})) + \mathbf{Z}_{L-2}, \\
&\vdots \\
\mathbf{Z}'_1 &= \Psi_{\text{ToM}}(\mathcal{U}(\mathbf{Z}'_2)) + \mathbf{Z}_1 \\
\mathbf{X}_{l+1} &= \Psi_{\text{ToM}}(\mathcal{U}(\mathbf{Z}'_1)) + \mathbf{Z}_0
\end{aligned} \tag{6.5}$$

où $\mathcal{D}$ et $\mathcal{U}$ représentent les opérations de downsampling et upsampling au sein du framework structurel imbriqué, et les connexions résiduelles préservent l'information de caractéristiques multi-échelles essentielle dans le pipeline de traitement hiérarchique.

Efficacité de Calcul et Applications Cliniques : Cette formulation mathématique démontre que l'intégration harmonieuse des couches GSC et ToM capture de manière efficace les dépendances spatiales multi-directionnelles par des méthodologies de traitement séquentiel anatomiquement informées, tout en préservant simultanément l'efficacité de calcul par les caractéristiques de complexité linéaire inhérentes à la modélisation state-space. En étendant le champ récepteur effectif tout en réduisant la complexité de calcul globale, cette approche fournit une alternative hautement évolutive et légère aux mécanismes d'attention conventionnels, particulièrement bénéfique pour traiter les grands volumes médicaux 3D rencontrés en pratique clinique. S'appuyant sur cette fondation théorique robuste, U^{3}Mamba combine avec succès les avantages de calcul de la modélisation state-space avec l'efficacité prouvée du framework structurel U^3 imbriqué pour délivrer une solution robuste et efficace pour la segmentation d'images médicales 3D dans diverses modalités d'imagerie et applications cliniques.

6.3 Protocole Expérimental et Implémentation

Pour évaluer l'efficacité et les capacités de généralisation inter-modalités du framework HA-U^{3}Net proposé, nous avons conçu un protocole expérimental approfondi englobant de multiples ensembles de données d'imagerie 3D représentant diverses modalités d'imagerie médicale et applications cliniques. Cette conception expérimentale valide notre framework pour gérer la variabilité en imagerie médicale en démontrant une performance robuste dans tant les tâches de segmentation tumorale que de délimitation de structures anatomiques.

6.3.1 Ensembles de Données et Configuration d'Évaluation

La validation expérimentale emploie quatre ensembles de données soigneusement sélectionnés qui représentent collectivement l'étendue des applications d'imagerie médicale 3D et démontrent les défis inter-modalités que HA-U^{3}Net est conçu pour traiter :

Ensemble de Données ABUS pour l'Analyse par Ultrasons du Sein : L'ensemble de données Automated Breast Ultrasound (ABUS) [171] comprend 200 images tumorales volumétriques du sein par ultrasons provenant du challenge TDSC 2023, acquises en utilisant le système Invenia ABUS au Harbin Medical University Cancer Hospital. Chaque acquisition volumétrique inclut des images en niveaux de gris haute résolution accompagnées de masques tumoraux binaires précisément annotés, avec des résolutions spatiales cohérentes variant entre $843 \times 546 \times 270$ et $865 \times 682 \times 354$ voxels. L'ensemble de données a subi un prétraitement systématique, incluant la normalisation d'intensité et des procédures de recadrage focalisées sur la tumeur pour réduire les régions de fond non pertinentes tout en préservant le contexte anatomique essentiel. Suivant le protocole de partitionnement original de l'ensemble de données, nous avons alloué 100 échantillons pour l'entraînement, 30 échantillons pour la validation, et 70 échantillons pour le test. Les techniques d'augmentation de données en ligne, incluant les ajustements d'intensité,

retournements spatiaux, et déformations élastiques, ont été appliquées durant les phases d'entraînement pour améliorer les capacités de généralisation du modèle et refléter précisément la variabilité clinique.

Ensemble de Données BraTS 2023 pour la Segmentation de Tumeurs Cérébrales : L'ensemble de données Brain Tumor Segmentation (BraTS) 2023 [172] fournit des scans IRM multi-paramétriques complets obtenus de sources cliniques diverses, spécifiquement conçu pour les défis de segmentation tumorale cérébrale tridimensionnelle. Chaque cas patient inclut quatre modalités d'imagerie co-enregistrées (séquences pondérées T1, T1-gadolinium enhanced, T2, et T2-FLAIR) qui fournissent collectivement une information de contraste tissulaire complète essentielle pour une délimitation tumorale précise. Les procédures de prétraitement standardisées assurent l'uniformité spatiale et d'intensité dans tous les cas, tandis que les annotations expertes délimitent précisément trois sous-régions tumorales distinctes : tumeur rehaussante (ET), œdème (ED), et noyau nécrotique/non-rehaussant (NCR), qui sont groupées en cibles de segmentation cliniquement pertinentes incluant les régions tumeur rehaussante (ET), noyau tumoral (TC), et tumeur entière (WT). Pour notre protocole expérimental, 1,251 cas ont été divisés en partitions d'entraînement 80% et validation 20%, avec l'ensemble de validation officiel BraTS comprenant 219 cas réservés pour l'évaluation de test finale.

Ensemble de Données TotalSegmentator pour l'Analyse Multi-Organe : L'ensemble de données TotalSegmentator [173] englobe 1,228 scans CT avec des annotations complètes couvrant 117 structures anatomiques distinctes, incluant organes majeurs, composants squelettiques, systèmes musculaires, et structures vasculaires. Cet ensemble de données complet supporte diverses applications cliniques, incluant l'analyse de volume d'organe, caractérisation de maladie, planification chirurgicale, et conception de traitement radiothérapeutique. L'ensemble de données reflète la variabilité clinique réelle par l'inclusion de cas provenant de contextes cliniques variés, incorporant des variations naturelles en présentations pathologiques, technologies de scanner, et protocoles d'acquisition. Les procédures de prétraitement systématiques incluent le rééchantillonnage à espacement voxel isotropique 1mm^3 et normalisation d'intensité z-score pour assurer la cohérence dans différents paramètres d'acquisition. Le partitionnement de l'ensemble de données suit une division 85% entraînement, 5% validation, et 10% test pour fournir une évaluation robuste tout en maintenant des données d'entraînement suffisantes pour la complexité des tâches de segmentation multi-organe. Les méthodologies d'augmentation de données, incluant les rotations spatiales aléatoires et perturbations d'intensité, ont été appliquées pour améliorer la généralisation du modèle dans la variabilité anatomique et variations de conditions d'imagerie.

Ensemble de Données AutoPET pour l'Analyse d'Imagerie Métabolique : L'ensemble de données AutoPET [174] comprend des scans de tomographie par émission de positrons et CT fluorodésoxyglucose corps entier co-enregistrés (FDG-PET/CT) obtenus de patients avec diverses conditions oncologiques, incluant mélanome, lymphome, cancer pulmonaire, et sujets contrôles sains. L'acquisition de données a été effectuée dans deux centres médicaux majeurs utilisant des protocoles d'imagerie standardisés

pour assurer la cohérence et la pertinence clinique. L'ensemble de données inclut 1,014 études d'entraînement et 150 études de test, permettant une évaluation complète des caractéristiques de généralisabilité inter-institutionnelle essentielles pour le déploiement clinique. Les annotations tumorales ont été effectuées manuellement par des radiologues expérimentés utilisant des critères PET/CT établis pour l'identification de lésions métaboliques. Pour notre implémentation expérimentale, 90% de l'ensemble d'entraînement a été alloué pour l'entraînement du modèle, 10% pour la validation, avec l'ensemble de test officiel réservé pour l'évaluation de performance finale. Les mêmes techniques d'augmentation de données ont été utilisées pour ce jeu de données.

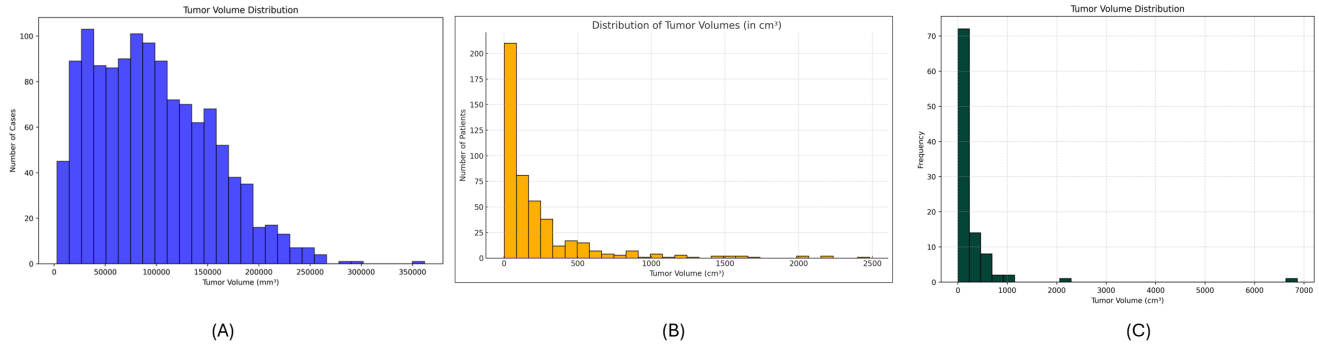

FIGURE 6.4 – Analyse de distribution des structures pathologiques dans les ensembles de données de différentes modalités, démontrant l'hétérogénéité des présentations tumorales dans les modalités d'imagerie. (A) BraTS : distribution volumétrique des sous-régions tumorales cérébrales (ET, TC, WT) dans les scans IRM. (B) AutoPET : caractéristiques de compte et volume des lésions dans FDG-PET/CT corps entier. (C) ABUS : fréquence de volume tumoral en ultrasons 3D du sein.

La Figure 6.4 présente les distributions volumétriques et de fréquence des structures pathologiques dans les ensembles de données BraTS, ABUS, et AutoPET. Ces données illustrent la forte variabilité en volumes tumoraux et présentations morphologiques, allant de petites lésions mammaires en imagerie par ultrasons aux larges tumeurs cérébrales visibles en IRM, ainsi qu'aux lésions métaboliques détectées en PET/CT. La Figure 6.5 fournit une illustration détaillée de la distribution des structures anatomiques dans six systèmes d'organes majeurs représentés dans l'ensemble de données TotalSegmentator, soulignant la couverture complète de régions anatomiques diverses s'étendant dans les systèmes cardiovasculaire, respiratoire, gastro-intestinal, génito-urinaire, nerveux, et musculo-squelettique essentiels pour valider les capacités de généralisation de notre framework proposé.

6.3.2 Détails d'Implémentation

L'entraînement et l'évaluation de nos modèles de segmentation tridimensionnelle ont été conduits sur une infrastructure de calcul haute performance comprenant deux GPUs

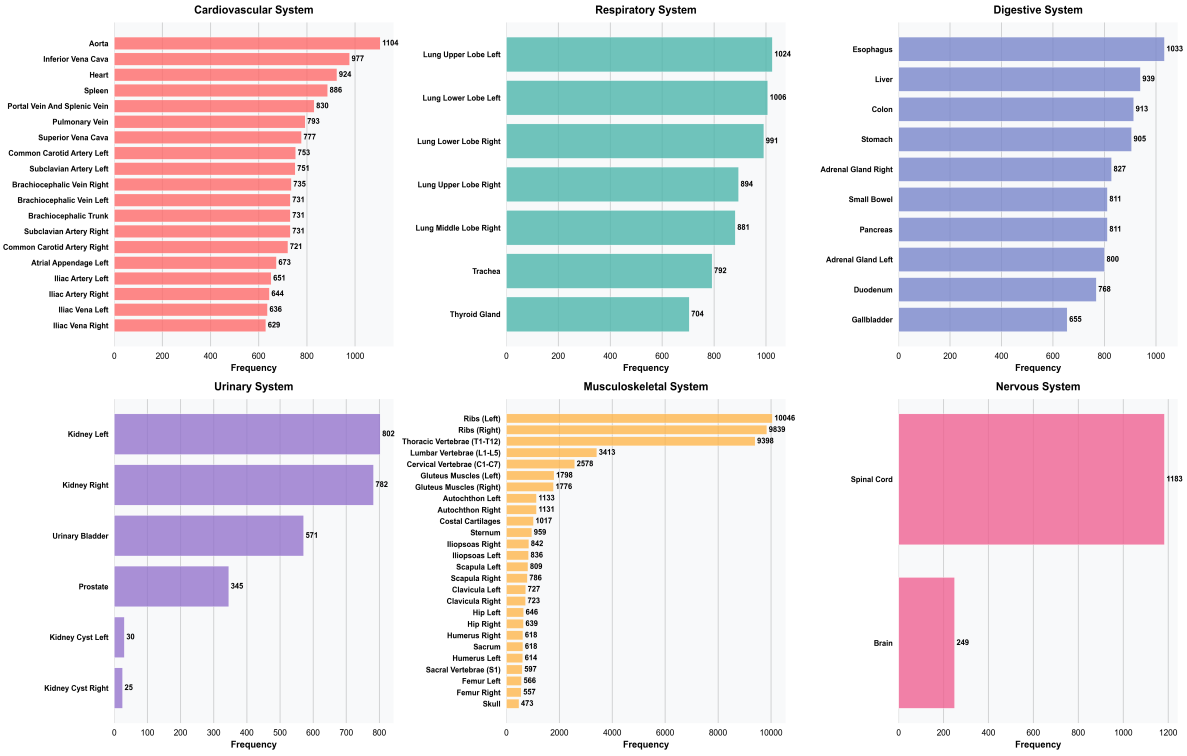

FIGURE 6.5 – Distribution de Fréquence d'Organe par Système Anatomique dans l'ensemble de données TotalSegmentator.

NVIDIA RTX A6000, chacun équipé de 48GB de mémoire. Cette configuration permet le traitement simultané de multiples volumes 3D tout en supportant les composants architecturaux complexes de HA-U^{3}Net, incluant les blocs U^3 imbriqués et mécanismes d'attention hybrides.

Configuration d'Optimisation : Le processus d'entraînement a employé l'optimiseur Nesterov Adam, combinant les avantages de momentum du gradient accéléré Nesterov avec les bénéfices de taux d'apprentissage adaptatif de l'algorithme Adam. Le taux d'apprentissage initial a été fixé à 10^{-4} après une exploration d'hyperparamètres pour assurer une convergence stable dans diverses modalités d'imagerie tout en prévenant les problèmes d'explosion ou de disparition de gradient.

Prétraitement de Données et Standardisation : Pour assurer un traitement cohérent dans diverses modalités d'imagerie, tous les volumes 3D ont été rééchantillonnés et redimensionnés à des dimensions uniformes de $128 \times 128 \times 128$ voxels. Cette standardisation permet un traitement par batch efficace tout en maintenant une résolution spatiale suffisante pour une délimitation précise de structures anatomiques. Ces dimensions représentent un équilibre optimal entre efficacité de calcul et préservation d'information spatiale.

Protocole d'Entraînement et Régularisation : Le processus d'entraînement s'est étendu sur 10^3 époques, avec un mécanisme d'arrêt précoce implémenté pour prévenir le surapprentissage. Cette stratégie monitore les métriques de performance de validation et termine automatiquement l'entraînement lorsqu'aucune amélioration n'est observée sur des époques consécutives.

Pour assurer une évaluation robuste et minimiser le biais potentiel du partitionnement, nous avons employé une validation croisée 4-fold dans toutes les évaluations expérimentales. Cette approche fournit des estimations de performance statistiquement fiables tout en maximisant l'utilisation des données d'entraînement disponibles.

Formulation de Fonction de Perte : Pour optimiser la performance de segmentation dans diverses structures anatomiques et présentations pathologiques, nous avons implémenté une fonction de perte combinée intégrant les composants Binary Cross Entropy (BCE) et Dice loss, équilibrant la précision de classification voxel-wise avec la performance de segmentation basée sur le chevauchement. La fonction de perte globale est formulée comme :

$$L_{\text{BCE-Dice}} = \alpha L_{\text{BCE}} + \beta L_{\text{Dice}}, \tag{6.6}$$

où les paramètres α et β permettent le réglage de l'importance relative de la précision de classification au niveau pixel versus l'optimisation de chevauchement au niveau de la région. Le composant Dice loss est calculé selon :

$$L_{\text{Dice}} = 1 - \frac{2 \times \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2 + \varepsilon}, \tag{6.7}$$

où p_i représente la probabilité prédite pour chaque voxel, g_i dénote l'étiquette de vérité terrain correspondante, N indique le nombre total de voxels dans le volume, et ε constitue une constante petite pour assurer la stabilité numérique et prévenir la division par zéro.

Métriques d'Évaluation : Pour évaluer la performance de segmentation, nous avons utilisé deux métriques primaires capturant tant les caractéristiques de précision de chevauchement que de précision de frontière. Le Dice Similarity Coefficient (DSC) quantifie le chevauchement volumétrique entre les régions de segmentation prédites et de vérité terrain, fournissant une mesure de précision globale particulièrement sensible aux erreurs de faux positifs et faux négatifs. La Hausdorff Distance 95ème percentile (HD95) évalue la discordance de frontière entre les volumes prédits et de vérité terrain, offrant un aperçu détaillé de la précision d'alignement de frontières et de la qualité de correspondance de surface.

Ces métriques sont particulièrement bien adaptées pour l'évaluation de segmentation tridimensionnelle, capturant tant la qualité volumétrique globale que les caractéristiques de fidélité de frontière essentielles pour les applications cliniques.

6.4 Résultats et Analyse

Cette section présente les résultats expérimentaux et l'analyse détaillée de la performance du framework HA-U^{3}Net dans diverses modalités d'imagerie médicale.

6.4.1 Études d'Ablation

Pour évaluer les contributions individuelles des innovations architecturales de HA-U^{3}Net, nous avons conduit une série d'expérimentations d'ablation sur nos quatre ensembles de données d'évaluation : ABUS, BraTS, TotalSegmentator, et AutoPET. Ces expérimentations ont été conçues pour analyser l'impact quantitatif de composants architecturaux clés, incluant les blocs d'attention hybride (HA), le rôle des mécanismes de supervision profonde, et l'influence des configurations de niveaux imbriqués sur la performance de segmentation. Pour assurer la fiabilité statistique de nos résultats, nous avons employé une stratégie de validation croisée 4-fold, évaluant la cohérence de performance sur plusieurs entraînements et folds de validation.

Le Tableau 6.1 détaille les résultats de validation croisée 4-fold de HA-U^{3}Net et sa variante U^{3}Mamba. Ces résultats démontrent une cohérence exceptionnelle sur les différents folds et ensembles de données, établissant ainsi la fiabilité statistique et la robustesse de notre framework.

Table 6.1 – Résultats de Validation Croisée 4-Fold à Travers les Ensembles de Données (Score Dice Moyen).

Fold / Méthode	ABUS 3D		BraTS		Total Segmentor		AutoPET	
	HA-U^{3}Net	U^{3}Mamba	HA-U^{3}Net	U^{3}Mamba	HA-U^{3}Net	U^{3}Mamba	HA-U^{3}Net	U^{3}Mamba
Fold 1	86.42	84.12	92.61	90.81	94.83	93.31	95.09	94.23
Fold 2	85.13	84.76	91.73	89.75	93.44	92.91	94.31	93.47
Fold 3	84.77	83.95	92.82	89.92	93.18	93.20	95.02	94.15
Fold 4	86.12	84.69	92.54	90.64	93.25	93.08	94.62	93.79
Moyenne	85.61	84.38	92.43	90.28	93.68	93.13	94.76	93.91

6.4.1.1 Impact des Mécanismes d'Attention Hybride

Le Tableau 6.2 fournit une analyse comparative de nos modèles avec et sans les blocs d'attention hybride, structure architecturale en forme de U^3, et implémentations de variantes Mamba. Les résultats soulignent l'importance de ces composants architecturaux pour atteindre une performance de segmentation supérieure sur tous les ensembles de données évalués, démontrant leur rôle critique dans la généralisation inter-modalités robuste.

TABLE 6.2 – Impact des Blocs Hybrid Attention (HA) et Variantes de Modèle sur la Performance (DSC Moyen (%)).

Modèle	ABUS	BraTS	Total Seg	AutoPET
UNet3D	74.54	88.11	76.51	88.75
UNet3D + HA	76.69	89.86	76.62	89.12
U^3Mamba - HA	79.44	90.18	91.95	91.71
U^3Mamba	81.51	90.26	92.88	92.13
U^3Net	80.68	89.15	92.29	91.84
HA-U^3Net	**83.46**	**90.92**	**93.37**	**92.57**

L'inclusion de blocs d'attention hybride a amélioré les scores de coefficient Dice sur tous les ensembles de données évalués. Cette amélioration était particulièrement prononcée dans l'ensemble de données ABUS, qui présente des défis uniques dus aux caractéristiques bruyantes de la modalité ultrasons. Les blocs d'attention hybride ont démontré une capacité remarquable à traiter tant les caractéristiques spatiales que par canal, permettant d'atténuer efficacement les artefacts de bruit, ombre acoustique, et patterns de granularité.

L'architecture U^3 imbriquée a contribué de manière substantielle à la performance globale par sa capacité à capturer des caractéristiques multi-échelles essentielles pour une délimitation précise de frontières. Cette innovation s'est avérée précieuse pour délimiter les frontières tumorales dans l'ensemble de données BraTS, où la segmentation de sous-régions tumorales (tumeur rehaussante, noyau tumoral, et tumeur entière) nécessite l'intégration tant d'information contextuelle globale que de détails fins locaux.

La variante U^3Mamba offre une alternative légère à l'architecture HA-U^3Net, atteignant des résultats compétitifs tout en réduisant le nombre de paramètres. Cette efficacité la rend bien adaptée aux applications cliniques temps-réel et scénarios de déploiement à ressources contraintes, comme en témoigne sa performance sur l'ensemble de données TotalSegmentator comprenant 117 structures anatomiques distinctes.

6.4.1.2 Analyse du Mécanisme de Supervision Profonde

L'intégration de mécanismes de supervision profonde dans l'architecture HA-U^{3}Net permet de résoudre les problèmes de flux de gradient rencontrés dans les réseaux 3D profonds traitant des données médicales volumétriques. Cette stratégie consiste à extraire les masques de segmentation à chaque niveau du décodeur, puis à les combiner avec la sortie finale. Cette agrégation multi-niveaux améliore l'apprentissage de caractéristiques dans l'ensemble du décodeur et renforce la stabilité d'entraînement en assurant une propagation de gradient efficace dans toute l'architecture.

Cette approche multi-niveaux remplit deux objectifs : assurer une propagation de gradient robuste dans la structure de réseau profond, prévenant les problèmes de disparition de gradient communs dans les architectures 3D profondes ; et imposer un apprentissage cohérent sur différentes échelles spatiales, assurant que tant les détails anatomiques fins que l'information contextuelle plus large contribuent à la performance de segmentation.

Validation Empirique : Les études d'ablation évaluent la contribution de la supervision profonde dans notre framework, comparant la performance de HA-U^{3}Net avec et sans supervision profonde sur quatre modalités d'imagerie distinctes.

TABLE 6.3 – Impact de la Supervision Profonde (DS) sur la Performance de HA-U^{3}Net (DSC Moyen (%)).

Configuration	ABUS	BraTS	Total Seg	AutoPET
HA-U^{3}Net - DS	82.58	90.15	92.58	91.84
HA-U^{3}Net + DS	**83.46**	**90.92**	**93.37**	**92.57**
Improvement	+0.88	+0.77	+0.79	+0.73
U^{3}Mamba - DS	80.72	89.61	92.19	91.48
U^{3}Mamba + DS	**81.51**	**90.26**	**92.88**	**92.13**
Improvement	+0.79	+0.65	+0.69	+0.65

Les résultats démontrent des améliorations cohérentes sur tous les ensembles de données et variantes de modèles, allant de +0.73% sur AutoPET à +0.88% sur ABUS, avec une amélioration moyenne de +0.79%. U^{3}Mamba exhibe des tendances similaires, avec des gains moyens de +0.70% sur tous les ensembles de données. La supervision profonde s'avère particulièrement efficace sur l'ensemble de données ABUS, où la complexité de l'imagerie par ultrasons, caractérisée par le bruit de granularité et l'ombre acoustique, bénéficie d'un flux de gradient amélioré et d'un apprentissage de caractéristiques renforcé par la supervision intermédiaire.

6.4.1.3 Analyse des Niveaux Imbriqués de HA-U^{3}Net

La détermination de la profondeur optimale de niveau imbriqué dans l'architecture HA-U^{3}Net représente une décision de conception critique affectant tant la performance

de segmentation que l'efficacité de calcul. Cette analyse investigue la relation entre profondeur architecturale (paramètre L, indiquant les niveaux imbriqués) et performance de segmentation dans notre framework.

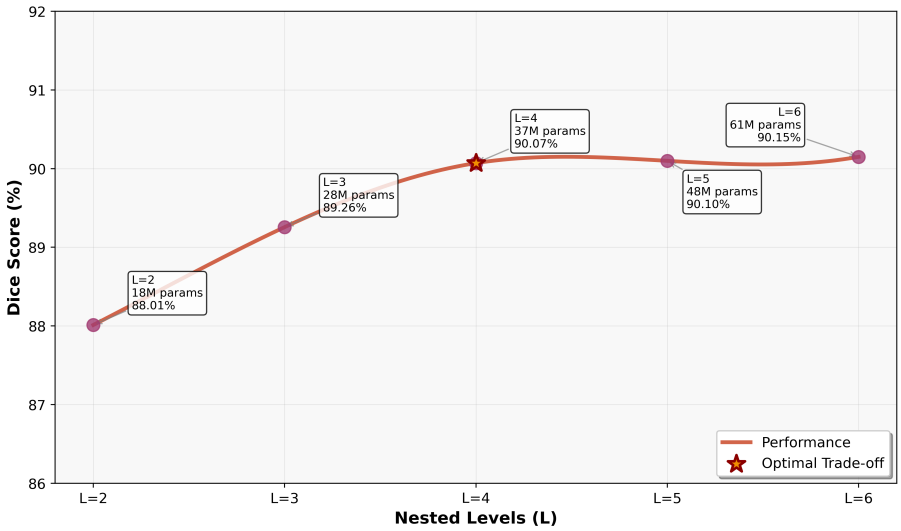

FIGURE 6.6 – Impact de la profondeur des niveaux imbriqués (L) sur la performance de HA-U^{3}Net sur tous les ensembles de données.

L'investigation révèle des améliorations avec une profondeur architecturale croissante : 88.01% de coefficient Dice moyen pour $L = 2$, 89.26% pour $L = 3$, et 90.08% pour $L = 4$. Ce pattern valide notre conception d'architecture imbriquée, confirmant que des structures hiérarchiques plus profondes capturent efficacement les relations spatiales complexes dans des données médicales volumétriques.

Au-delà de $L = 4$, l'architecture présente une saturation de performance avec des améliorations minimales : 90.10% pour $L = 5$ (+0.02%) et 90.15% pour $L = 6$ (+0.07%). L'analyse révèle des augmentations importantes de paramètres : de 37M paramètres à $L = 4$ à 48M à $L = 5$ et 61M à $L = 6$. Ces exigences de calcul accrues, couplées avec des améliorations de performance négligeables, établissent $L = 4$ comme la configuration optimale.

Optimisation de Déploiement Clinique : La sélection de $L = 4$ représente un équilibre optimal entre précision de segmentation et efficacité de calcul, s'alignant avec notre objectif de développer des solutions cliniquement viables. Cette configuration fournit une performance proche de l'optimum tout en maintenant des exigences de calcul appropriées pour le déploiement clinique, incluant les applications temps-réel et environnements à ressources contraintes.

6.4.2 Résultats Quantitatifs

6.4.2.1 Résultats sur l'Ensemble de Données ABUS

HA-U^{3}Net a atteint une performance supérieure sur toutes les métriques d'évaluation, avec un Dice Similarity Coefficient (DSC) de 83.46%, sensitivité de 92.82%, spécificité de 99.89%, et Distance de Hausdorff 95ème percentile (HD95) de 14.07 mm. Ces résultats (voir Tableau 6.4) démontrent des améliorations importantes comparées aux méthodes de pointe établies, incluant une amélioration DSC de 4.15% sur nnUNet et des améliorations notables en précision de délimitation de frontières comme en témoigne le score HD95 réduit.

TABLE 6.4 – Comparaison avec Méthodes Pertinentes sur l'Ensemble de Données ABUS 3D.

Méthode	Année	Params(M)	DSC ↑	Sens ↑	Spec ↑	95HD (mm) ↓
UNet 3D [175]	2016	79	74.54	83.47	98.81	21.92
nnUNet [176]	2021	32	79.31	88.68	98.96	19.55
CKD-TransBTS [177]	2023	82	80.33	89.12	98.64	19.02
SegMamba [170]	2024	67	79.61	88.82	99.29	19.26
MA-SAM [178]	2024	97	80.39	90.40	99.63	17.45
P-Former [179]	2025	46	79.91	89.51	98.67	18.64
U^{3}Mamba	2025	**6**	81.51	91.42	99.55	17.16
HA-U^{3}Net	2025	37	**83.46**	**92.82**	**99.89**	**14.07**

L'analyse comparative révèle les avantages architecturaux de notre conception de bloc U^3 imbriqué dans la gestion des défis d'images ultrasons. Notamment, notre variante légère U^{3}Mamba atteint une performance compétitive avec seulement 6M paramètres, représentant une réduction significative en exigences computationnelles tout en maintenant la précision.

Analyse de Généralisation : Pour valider les capacités de généralisation du framework dans le domaine ultrasons, nous avons conduit une validation inter-ensembles en utilisant des variantes bidimensionnelles de notre architecture. L'évaluation a consisté à entraîner sur des slices ABUS 2D et valider en zero-shot sur l'ensemble de données BUSI [180], fournissant un aperçu de la capacité du modèle à gérer le transfert entre ensembles sans réentraînement.

Les résultats de validation inter-ensembles démontrent des capacités de généralisation remarquables, avec HA-U^{2}Net atteignant 76.65% DSC sur l'ensemble de données BUSI sans réentraînement, représentant seulement une dégradation de performance de 10.64% malgré un transfert entre ensembles important. Cette performance dépasse toutes les méthodes de référence, avec des améliorations de 1.73% sur U^{2}Mamba et 5.52% sur MA-SAM 2D. Ces résultats, présentés dans le Tableau 6.5, valident l'efficacité de notre mécanisme d'attention hybride dans l'apprentissage de caractéristiques robustes qui généralisent dans différents protocoles d'imagerie par ultrasons et populations de patients.

TABLE 6.5 – Validation cross-dataset : modèles 2D entraînés sur slices ABUS et testés sur l'ensemble de données BUSI.

Méthode	ABUS 2D		BUSI (Zero-shot)	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNet 2D [175]	78.65	19.82	58.34	31.25
nnUNet 2D [176]	83.41	16.87	67.52	26.48
CKD-TransBTS 2D [177]	82.73	17.54	67.25	27.93
SegMamba 2D [170]	83.19	17.21	66.86	27.35
MA-SAM 2D [178]	84.56	15.92	71.13	25.18
P-Former 2D [179]	83.82	16.54	68.97	26.12
U^2Mamba	85.78	14.65	74.92	23.42
HA-U^2Net	**87.29**	**13.18**	**76.65**	**21.67**

6.4.2.2 Résultats sur l'Ensemble de Données BraTS

L'ensemble de données Brain Tumor Segmentation (BraTS) représente l'un des benchmarks les plus difficiles en segmentation d'images médicales en raison de la nature complexe des sous-régions tumorales de glioblastome, incluant la délimitation de tumeur rehaussante (ET), noyau tumoral (TC), et tumeur entière (WT). Ces sous-régions exhibent des caractéristiques radiologiques distinctes et frontières irrégulières, nécessitant des capacités d'extraction de caractéristiques sophistiquées pour distinguer précisément entre les composants tumoraux et le tissu cérébral sain.

HA-U^3Net démontre une performance remarquable sur toutes les sous-régions tumorales comme le montre le tableau 6.6, atteignant un score Dice moyen de 90.92% et score HD95 moyen de 6.04 mm. L'architecture de bloc U^3 imbriqué du framework s'avère particulièrement efficace dans la gestion des frontières tumorales complexes et structures hétérogènes caractéristiques de la pathologie glioblastome, avec des améliorations cohérentes dans les tâches de segmentation de tumeur rehaussante (88.34%), noyau tumoral (90.76%), et tumeur entière (93.66%).

TABLE 6.6 – Comparaison de différentes méthodes sur l'ensemble de données BraTS.

Méthode	Params	Dice ↑ (%)				Sens/Spec Moyen ↑ (%)		HD95 (mm) ↓		
		ET	TC	WT	Moyen	Sens	Spec	ET	TC	WT
UNet 3D [175]	79	86.57	87.94	89.81	88.11	85.94	89.12	13.35	9.23	10.29
nnUNet [176]	56	87.69	89.31	92.49	89.83	87.25	89.90	12.54	7.23	8.45
CKD-TransBTS [177]	82	88.22	90.14	93.22	90.53	89.74	92.12	6.64	6.97	6.45
SegMamba [170]	67	87.98	89.16	92.61	89.92	88.12	91.34	10.26	7.55	7.32
MA-SAM [178]	97	88.19	90.22	92.99	90.46	88.76	92.04	7.59	7.64	7.53
P-Former [179]	46	87.52	88.62	91.73	89.29	88.41	90.83	10.51	7.89	7.72
U^3Mamba	6	88.03	89.97	92.79	90.26	89.18	91.92	7.28	7.31	7.11
HA-U^3Net	37	**88.34**	**90.76**	**93.66**	**90.92**	**90.30**	**93.00**	**5.74**	**6.05**	**6.34**

Les scores HD95 supérieurs dans toutes les régions tumorales (5.74 mm pour ET, 6.05 mm pour TC, 6.34 mm pour WT) démontrent la précision du framework en délimitation de frontières, particulièrement critique pour les applications de planification chirurgicale et radiothérapie.

6.4.2.3 Résultats sur l'Ensemble de Données TotalSegmentator

L'évaluation sur TotalSegmentator valide la capacité du framework à segmenter plus de 100 structures anatomiques en imagerie CT, présentant des défis de variation d'échelle, frontières chevauchantes, et relations spatiales inter-organes.

HA-U^{3}Net atteint un DSC de 93.37% et HD95 de 16.91 mm (Tableau 6.7). L'architecture imbriquée gère efficacement les variations d'échelle de l'anatomie abdominale et thoracique, des petites structures vasculaires aux grands systèmes d'organes.

TABLE 6.7 – Comparaison de métriques moyennes sur les anatomies dans l'ensemble de données TotalSegmentator.

Méthode	Année	Params(M)	DSC ↑	Sens ↑	Spec ↑	95HD (mm) ↓
UNet 3D [175]	2016	79	76.51	79.03	85.61	26.94
nnUNet [176]	2021	35	92.89	93.49	95.55	20.34
CKD-TransBTS [177]	2023	82	91.85	92.61	94.30	24.13
SegMamba [170]	2024	67	91.13	93.62	95.14	20.76
MA-SAM [178]	2024	97	92.76	93.84	94.70	20.48
P-Former [179]	2025	46	91.91	93.08	95.02	20.61
U^{3}Mamba	2025	**6**	92.88	**95.61**	95.23	18.46
HA-U^{3}Net	2025	37	**93.37**	94.52	**96.07**	**16.91**

6.4.2.4 Résultats sur l'Ensemble de Données AutoPET

L'évaluation sur AutoPET traite les défis de l'imagerie de tomographie par émission de positrons, incluant résolution spatiale faible, ambiguïté métabolique, et artefacts de bruit sur des données fonctionnelles haute résolution. HA-U^{3}Net atteint un DSC de 92.57%, sensitivité de 94.39%, spécificité de 95.32%, et HD95 de 4.79 mm (Tableau 6.8), surpassant MA-SAM (+1.14% DSC) et CKD-TransBTS (+1.35% DSC) avec une précision de frontières nettement améliorée.

La performance robuste sur cet ensemble de données multi-institutionnel valide les capacités de généralisation du framework dans différents protocoles d'acquisition et types de scanners. L'efficacité du mécanisme d'attention hybride dans le traitement des défis spécifiques au PET démontre la flexibilité de notre approche de gestion de variabilité, s'adaptant efficacement tant aux modalités d'imagerie anatomiques qu'à l'imagerie métabolique fonctionnelle avec une précision maintenue.

TABLE 6.8 – Comparaison avec Méthodes Pertinentes sur l'Ensemble de Données AutoPET.

Méthode	Année	Params(M)	DSC ↑	Sens ↑	Spec ↑	95HD (mm) ↓
UNet 3D [175]	2016	79	88.75	90.52	93.05	8.40
nnUNet [176]	2021	32	90.34	92.48	93.31	7.14
CKD-TransBTS [177]	2023	82	91.22	93.65	94.81	5.59
SegMamba [170]	2024	67	90.64	93.18	94.17	5.93
MA-SAM [178]	2024	97	91.43	93.84	94.83	5.10
P-Former [179]	2025	46	90.95	93.09	93.98	5.91
U^{3}Mamba	2025	**6**	92.13	94.07	95.30	4.96
HA-U^{3}Net	2025	37	**92.57**	**94.39**	**95.32**	**4.79**

6.4.2.5 Analyse d'Efficacité de Calcul

Pour évaluer la faisabilité de déploiement clinique, nous avons mesuré les exigences de calcul sur des volumes standardisés $128 \times 128 \times 128$ avec un GPU NVIDIA RTX A6000 : compte de paramètres, FLOPs, temps d'inférence (IT), utilisation mémoire pic, et temps d'entraînement (TT).

TABLE 6.9 – Comparaison d'efficacité de calcul sur volumes d'entrée 128×128×128.

Méthode	Params (M)	FLOPs (G)	IT (ms)	Mémoire Pic (MB)	TT (h)
UNet 3D [175]	79	1900.07	344.50	3749.66	27.3
nnUNet [176]	32	1649.10	731.66	3876.31	18.7
CKD-TransBTS [177]	82	423.79	608.67	2893.08	48.9
SegMamba [170]	67	1563.80	201.60	2337.61	31.4
MA-SAM [178]	97	4561.40	8421.23	5947.69	49.1
P-Former [179]	46	**79.85**	433.09	3540.18	23.6
U^{3}Mamba	**6**	102.67	**172.02**	**741.75**	**18.2**
HA-U^{3}Net	37	662.38	545.70	3143.11	21.6

U^{3}Mamba atteint une efficacité remarquable avec 6M paramètres, 172.02 ms temps d'inférence, et 18.2 heures temps d'entraînement, adapté pour les applications à ressources contraintes et temps-réel. HA-U^{3}Net, avec 37M de paramètres, 545.70 ms de temps d'inférence, et 21.6 heures de temps d'entraînement, fournit une précision supérieure pour les applications cliniques haute précision. Ces deux variantes couvrent diverses exigences de déploiement tout en préservant une performance inter-modalités robuste.

6.4.3 Résultats Qualitatifs

L'évaluation qualitative complète les métriques quantitatives par une évaluation visuelle de qualité de segmentation, précision de frontières, et pertinence clinique. Cette analyse permet un examen détaillé du comportement du framework dans des scénarios difficiles, incluant les régions à faible contraste, frontières anatomiques complexes, et artefacts spécifiques à la modalité.

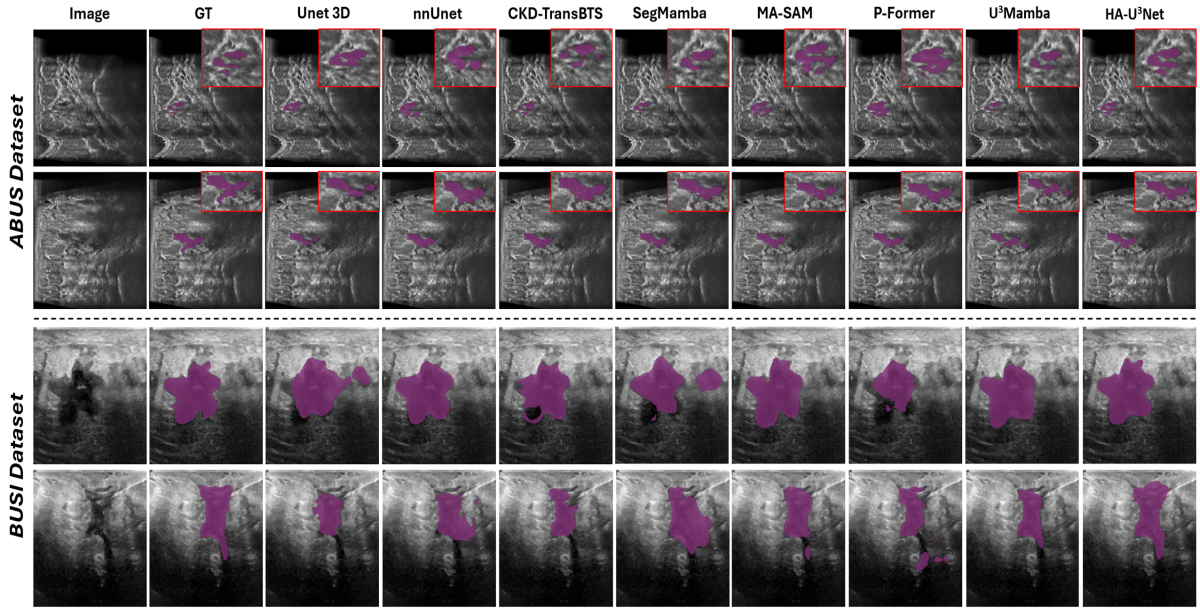

FIGURE 6.7 – Comparaison qualitative des résultats de segmentation sur les ensembles de données ultrasons. Les deux rangées supérieures correspondent à ABUS, illustrant la performance intra-ensemble. Les deux rangées inférieures montrent les résultats sur BUSI, démontrant la généralisation inter-ensembles en inférence zero-shot.

La Figure 6.7 montre la performance de HA-U^{3}Net en segmentation tumorale mammaire par ultrasons, surpassant SegMamba, CKD-TransBTS, P-Former, et MA-SAM par une délimitation précise de frontières. Le framework capture les détails structurels fins, particulièrement dans les régions caractérisées par des artefacts de granularité et ombre acoustique, résultant en masques de segmentation lisses et précis alignés avec les annotations de vérité terrain.

La Figure 6.8 illustre la performance de HA-U^{3}Net en segmentation tumorale cérébrale, délimitant précisément les sous-régions tumorales (tumeur rehaussante, noyau tumoral, tumeur entière) avec des frontières précises et erreurs minimales. La visualisation met en évidence les relations spatiales parmi les composants tumoraux, démontrant la capacité du framework à capturer les variations morphologiques subtiles essentielles pour la planification neurochirurgicale et radiothérapie.

La Figure 6.9 présente les capacités de segmentation 3D de HA-U^{3}Net sur six systèmes anatomiques majeurs : cardiovasculaire, digestif, musculo-squelettique, nerveux,

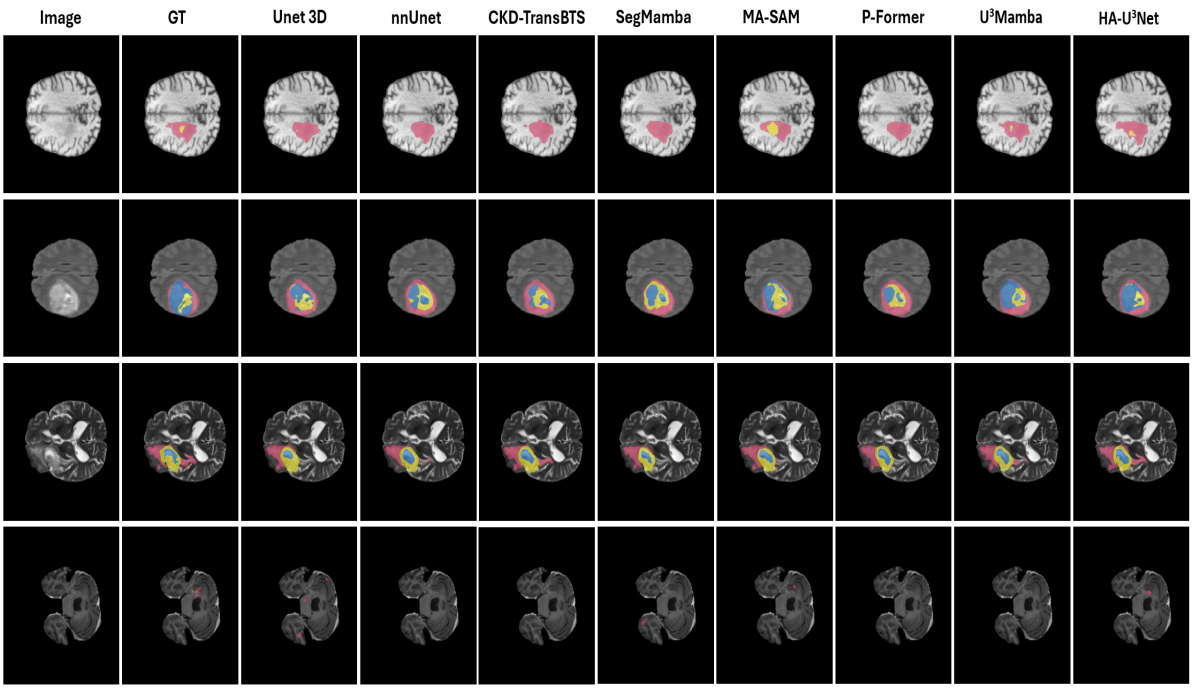

FIGURE 6.8 – Comparaison qualitative des résultats de segmentation sur BraTS. La figure affiche les slices IRM dans la modalité T1Gd, illustrant les régions tumorales dans divers cas.

FIGURE 6.9 – Comparaison de résultats qualitatifs dans Total Segmentator.

respiratoire, et urinaire. Le framework capture avec précision les structures anatomiques détaillées englobant tant les composants de tissus mous qu'osseux, réduisant les erreurs de frontières comparées aux méthodes alternatives.

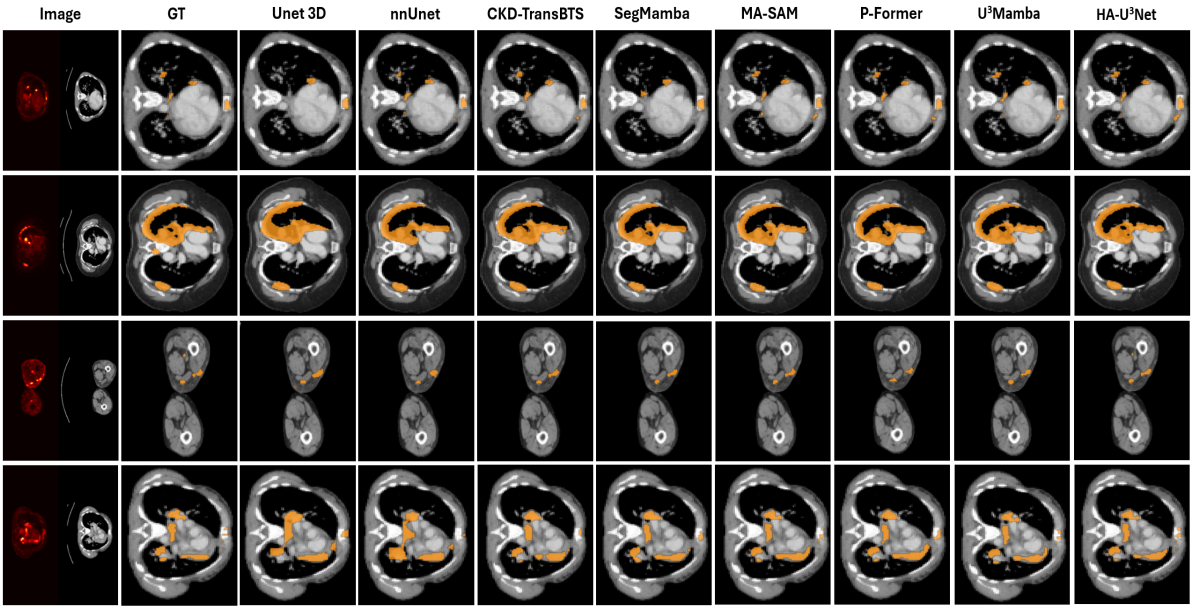

FIGURE 6.10 – Comparaison de résultats qualitatifs dans AutoPET.

La Figure 6.10 montre la performance de HA-U^{3}Net en imagerie PET/CT, segmentant avec précision les lésions métaboliquement actives de tailles variées tout en différenciant la captation pathologique de l'activité FDG physiologique. Le framework atteint une meilleure détection des lésions avec moins de faux positifs, notamment dans les zones thoraciques et abdominales.

Analyse de Capacité d'Extraction de Caractéristiques : Pour évaluer l'efficacité de conception architecturale, nous avons comparé les blocs Standard Residual Blocks (SRB), U^3-Blocks, et U^3-Blocks avec Hybrid Attention (HA) sur diverses étapes de traitement dans le pipeline de segmentation. Cette analyse, conduite sur BraTS, révèle des différences fondamentales en capacités de traitement et raffinement de caractéristiques.

La figure 6.11 révèle des caractéristiques de traitement distinctes sur les variantes architecturales. L'extraction de caractéristiques initiales démontre que les deux variantes U^3-Block exhibent un focus précoce sur les régions tumorales, tandis que SRB produit des activations plus larges et moins discriminatives. Dans les étapes progressives (mid-encoder, bottleneck, early decoder), le U^3-Block avec Hybrid Attention maintient des représentations précises et focalisées sur la tumeur par une suppression efficace de détails anatomiques non pertinents. Le mécanisme d'attention hybride assure la pertinence clinique des caractéristiques extraites, surpassant tant SRB que les U^3-Blocks sans attention en capacités de discrimination.

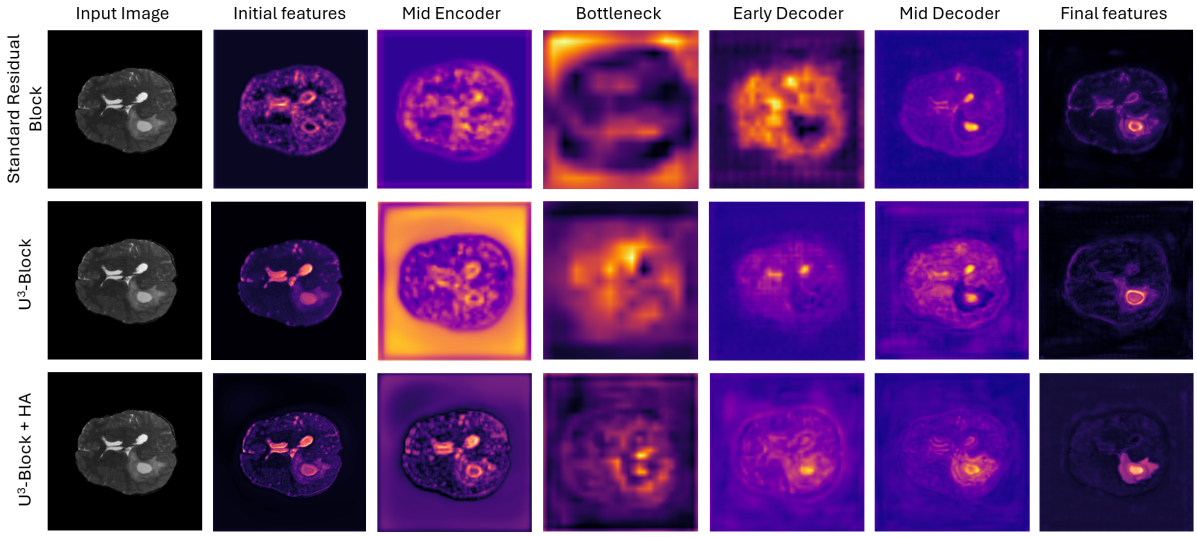

FIGURE 6.11 – Visualisation de cartes de caractéristiques sur diverses étapes du réseau, comparant SRB (haut), U^3-Block (milieu), et U^3-Block avec Hybrid Attention (bas). De gauche à droite : image d'entrée, caractéristiques initiales, caractéristiques mid-encoder, caractéristiques bottleneck, caractéristiques early decoder, caractéristiques mid-decoder, et caractéristiques finales.

Aux étapes finales, les U^3-Blocks avec Hybrid Attention délivrent une segmentation tumorale précise par l'exclusion efficace de structures anatomiques non pertinentes. En revanche, les approches alternatives capturent des régions non-cibles additionnelles, incluant des contours anatomiques plus larges non reliés aux objectifs de segmentation, démontrant l'importance de la conception architecturale avancée et des mécanismes d'attention pour une segmentation précise.

6.5 Conclusion

Ce chapitre démontre l'extension de notre framework de thèse pour gérer la variabilité d'imagerie médicale au domaine volumétrique par le développement de HA-U^{3}Net. L'architecture proposée traite les limitations fondamentales en segmentation d'images médicales 3D en intégrant des blocs U^3 imbriqués avec des mécanismes d'attention hybrides, permettant une performance robuste dans diverses modalités d'imagerie. Les réalisations clés incluent une validation démontrant une performance inter-modalités supérieure et la création de U^{3}Mamba, une variante légère atteignant une précision compétitive avec une réduction significative de paramètres pour les environnements à ressources contraintes.

La validation sur multiples modalités d'imagerie démontre une performance de segmentation supérieure et confirme nos principes de gestion de variabilité dans le domaine volumétrique. La performance inter-modalités cohérente valide la généralisabilité de notre approche au-delà du contexte bidimensionnel établi dans les chapitres précé-

dents, fournissant un framework unifié pour gérer la variabilité d'imagerie médicale selon les dimensions et modalités. Ce travail avance le domaine en fournissant un framework indépendant de la modalité qui répond aux exigences de déploiement clinique, positionnant l'architecture pour une adoption clinique étendue avec des améliorations potentielles en précision diagnostique et planification de traitement dans diverses spécialités médicales.

Les extensions futures incluent l'intégration avec quantification d'incertitude pour une fiabilité clinique améliorée, optimisation pour les modalités d'imagerie émergentes, et développement de capacités de traitement temps-réel pour les applications intraopératoires. L'architecture légère U^{3}Mamba présente des opportunités pour le déploiement mobile et au point de soin, étendant potentiellement l'accès aux capacités de segmentation avancées dans des contextes de soins de santé à ressources limitées et avançant l'impact pratique d'architectures deep learning adaptatives en pratique clinique.

Chapitre 7

TD-DIMB : Text-Driven Dense Inverted Mamba Bottlenecks pour la Segmentation d'Images Médicales Interactive

7.1 Introduction

7.1.1 Motivation et Contexte de Recherche

Les chapitres précédents ont traité la variabilité en analyse d'images médicales, progressant de solutions spécifiques au domaine vers des frameworks généralisables gérant diverses conditions d'imagerie et structures anatomiques. Bien que nos contributions aient démontré des avancées dans la gestion de variabilité intra-modalitées par l'imagerie dermoscopique et segmentation volumétrique, une dimension critique demeure non traitée : l'intégration d'information sémantique cross-modale qui s'étend au-delà de la reconnaissance de patterns purement visuels.

La pratique clinique contemporaine demande des systèmes de segmentation qui comblent l'écart sémantique entre traitement visuel automatisé et raisonnement clinique. Les radiologues exploitent des descriptions textuelles, terminologie anatomique, et connaissances contextuelles pour guider l'interprétation d'images, employant des vocabulaires sémantiques riches englobant tant les descriptions structurelles que les caractéristiques fonctionnelles. Ce workflow suggère que les systèmes de segmentation médicale robustes doivent transcender les approches uniquement visuelles pour incorporer la richesse sémantique inhérente à la connaissance du domaine médical.

Ce chapitre traite ces limitations en développant une architecture de segmentation nouvelle, cross-modale et guidée par le texte. En incorporant des prompts de langage naturel et des modèles fondamentaux spécifiques au domaine médical, nous proposons une

approche unifiée qui maintient une performance robuste tout en atteignant la flexibilité en adaptation de tâche et en généralisation cross-modale.

7.1.2 Défis d'Intégration Cross-Modale en Imagerie Médicale

L'intégration cross-modale en imagerie médicale présente des défis uniques, la distinguant des tâches vision-langage générales. L'imagerie médicale nécessite un alignement sémantique précis entre descriptions linguistiques et structures anatomiques hautement spécifiques, compliquée par une terminologie médicale spécialisée demandant une compréhension spécifique au domaine au-delà de la compréhension linguistique générale.

L'intégration de workflow clinique présente des défis fondamentaux car les professionnels médicaux emploient des terminologies variables selon la spécialisation, protocoles institutionnels, et contextes. Un cardiologue décrivant la fonction ventriculaire gauche met l'accent sur des repères anatomiques différents par rapport à un radiologue effectuant la même évaluation. Les systèmes cross-modaux doivent démontrer une robustesse dans ce spectre de communication tout en maintenant la précision.

Les défis de calcul de fusion efficace d'information cross-modale tout en maintenant une performance temps-réel représentent des obstacles importants. Les mécanismes de fusion basés sur l'attention traditionnels imposent une complexité quadratique qui évolue mal avec les images médicales haute résolution, particulièrement problématique en analyse volumétrique, où l'efficacité de traitement affecte directement les workflows de soins patient.

7.1.3 Limitations des Méthodes Actuelles Basées sur Prompts

Les approches de segmentation existantes basées sur les prompts présentent des limitations critiques restreignant l'utilité clinique. La plupart des frameworks démontrent une adaptabilité limitée aux variations terminologiques cliniques, nécessitant des correspondances de vocabulaire exactes et présentant une généralisation faible aux termes synonymes ou terminologie spécifique à l'institution. Cette fragilité limite le déploiement où la communication médicale qui varie selon les spécialités et les contextes.

La rigidité architecturale présente une autre limitation fondamentale. De nombreux modèles intègrent l'information textuelle seulement à des couches spécifiques ou s'appuient sur un conditionnement a posteriori, échouant à maintenir la cohérence sémantique dans le pipeline d'extraction de caractéristiques. Cette intégration superficielle résulte en un alignement sous-optimal entre intention textuelle et traitement visuel, prévenant la compréhension cross-modale efficace dans le processus de segmentation.

Les stratégies d'entraînement s'appuient sur des fonctions de perte standard qui traitent les erreurs de segmentation uniformément, indépendamment de la signification clinique. Cette approche échoue à prendre en compte l'importance différentielle des régions anatomiques ; les faux négatifs dans les structures critiques peuvent avoir un impact clinique plus grand que l'imprécision de frontières dans les zones moins critiques. L'ab-

sence de stratégies d'optimisation cliniquement informées limite l'applicabilité pratique de ces systèmes dans les scénarios médicaux réels.

7.1.4 Objectifs de Recherche et Portée

Ce chapitre développe un framework de segmentation text-driven qui traite les limitations identifiées dans les méthodes actuelles basées sur les prompts. Notre objectif primaire consiste à créer une architecture unifiée qui maintient une haute performance dans diverses modalités tout en fournissant une adaptabilité dynamique par des prompts de langage naturel.

Les objectifs spécifiques incluent : (1) développer des mécanismes de fusion cross-modale à complexité linéaire qui intègrent les prompts textuels directement dans le traitement de caractéristiques visuelles ; (2) établir une compréhension sémantique robuste qui s'étend au-delà du matching de vocabulaire exact pour gérer les variations terminologiques cliniques ; et (3) incorporer des stratégies d'optimisation cliniquement informées qui priorisent les régions anatomiques importantes durant l'entraînement.

La portée de recherche englobe des contributions théoriques en conception architecturale cross-modale et validation empirique sur de multiples modalités d'imagerie, évaluées par une évaluation tant de la performance spécifique aux tâches que des capacités de généralisation.

7.1.5 Contributions Clés et Innovations

Ce chapitre présente cinq innovations fondamentales avançant la segmentation d'images médicales cross-modale :

1. **Segmentation Guidée par Domaine Médical Sans Réentraînement** : Nous proposons un framework de segmentation entièrement prompt-aware qui intègre la guidance textuelle dans l'encodeur, bottleneck, et décodeur, plutôt qu'à des couches isolées. Ceci assure une propagation fine de sémantique médicale dans la hiérarchie de segmentation, permettant l'adaptation au temps d'inférence sans réentraînement et supportant la personnalisation dirigée par le clinicien qui aligne les sorties avec le raisonnement clinique.
2. **Modélisation State-Space Text-Driven via TD-SS2D** : Nous introduisons TD-SS2D. À notre connaissance, ceci est le premier module state-space prompt-aware adapté pour la segmentation d'images médicales. Contrairement au self-attention conventionnelle, TD-SS2D atteint une complexité linéaire tout en intégrant directement la sémantique médicale et indices textuels dans le processus de mise à jour d'état. Cette conception permet une modélisation efficace de dépendances à long terme, et assure un alignement robuste entre caractéristiques visuelles et les prompts cliniques.

3. **Dense Inverted Mamba Bottlenecks (DIMB)** : Nous proposons DIMB, un bloc architectural qui fusionne des connexions résiduelles inversées, des convolutions séparables en profondeur, et la modélisation state-space dans une conception dense connectée enrichie d'une modulation adaptative au domaine médical. Cette architecture améliore la réutilisation de caractéristiques et le flux de gradient, tout en maintenant l'efficacité de calcul pour le déploiement temps-réel.
4. **Reinforced Gaussian Dice Loss (RGDL)** : Pour traiter l'ambiguïté clinique, nous proposons la Reinforced Gaussian Dice Loss (RGDL), qui combine des étiquettes lissées par gaussienne avec un schéma de pondération inspiré de l'apprentissage par renforcement. Contrairement aux pertes conventionnelles, RGDL privilégie les régions cliniquement critiques, pénalise davantage les faux négatifs dans les zones diagnostiques, et améliore la précision des frontières, résultant en une segmentation plus fiable des structures ambiguës ou sous-représentées.
5. **Évaluation Multi-Modale** : TD-DIMB est évalué sur quatre ensembles de données cliniquement pertinents et de modalités diverses : CAMUS (ultrasons), autoPET 2022 (PET/CT), Atlas v2.0 (IRM), et QaTa-COVID19 (rayon-X). Les résultats montrent des performances supérieures aux méthodes de référence, et une meilleure généralisation entre modalités.

7.2 Méthodologie et Conception Architecturale

7.2.1 Vue d'Ensemble du Framework TD-DIMB

S'appuyant sur les principes établis dans les chapitres précédents pour gérer la variabilité en imagerie médicale, cette section présente le Text-Driven Dense Inverted Mamba Bottleneck Network (TD-DIMB), un modèle conçu pour la segmentation d'images médicales cross-modale.

L'architecture TD-DIMB adopte une conception encodeur-décodeur inspirée de U-Net, adaptée pour intégrer la l'information sémantique dans un pipeline de segmentation efficace. En exploitant les avancées récentes en modélisation state-space et conditionnement basé sur les prompts, TD-DIMB traite les limitations d'architectures spécifiques à la modalité identifiées précédemment, offrant une solution flexible pour la segmentation d'images médicales.

Comme illustré en Fig 7.1, le framework TD-DIMB comprend six composants principaux : 1) Un modèle fondamental MedSigLIP gelé et pré-entraîné [181] qui extrait des caractéristiques vision médicale et texte synchronisées à partir de paires image-prompt, fournissant une compréhension sémantique spécifique au domaine médical, 2) Un encodeur hiérarchique composé de convolutions à stride et blocs Dense Inverted Mamba Bottleneck (DIMB), qui extraient des caractéristiques visuelles multi-échelles conditionnées sur des embeddings de domaine médical, 3) Une couche bottleneck avec un module Text-Driven Selective Scan 2D (TD-SS2D) pour modéliser le contexte sémantique à la

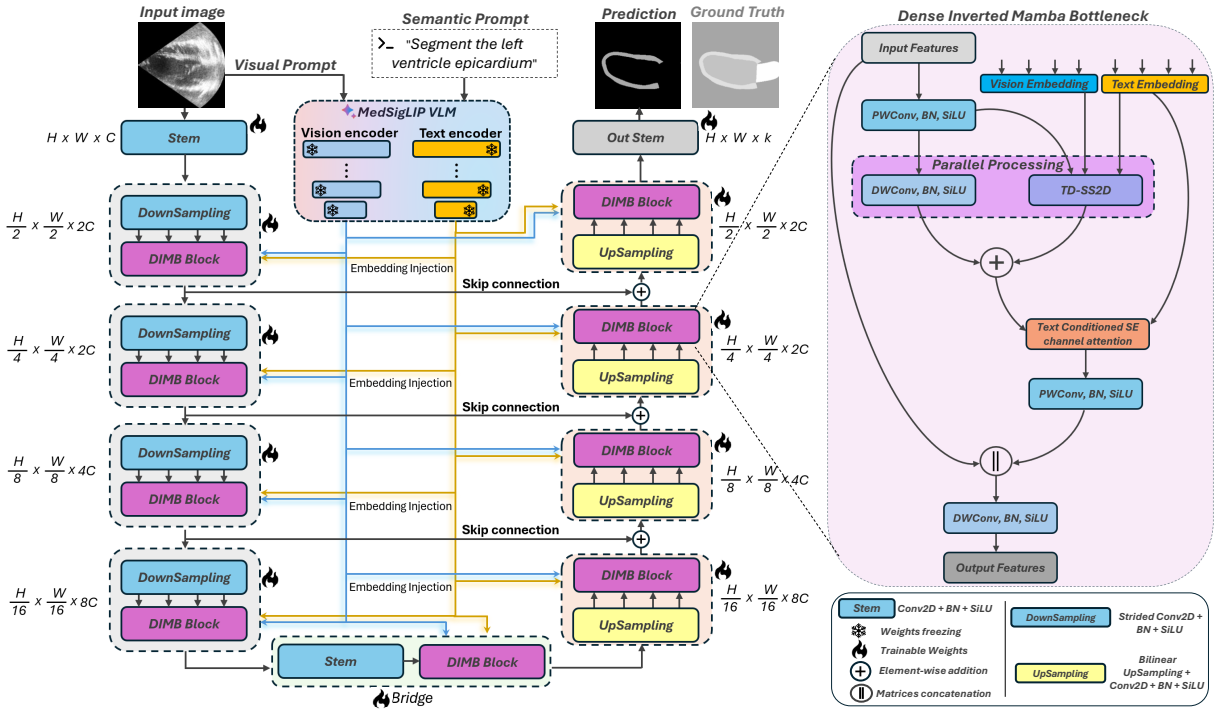

FIGURE 7.1 – Architecture Text-Driven Dense Inverted Mamba Bottleneck Network (TD-DIMB).

résolution la plus grossière, 4) Un décodeur symétrique qui effectue la récupération de résolution via l'upsampling bilinéaire, connexions skip, et raffinement basé sur DIMB, assurant la cohérence sémantique sur les échelles spatiales, 5) Une tête de projection produisant une carte de segmentation multi-classe alignée avec la sémantique du prompt, et 6) Extraction de caractéristiques visuelles médicales multi-échelles qui fournit un guidage hiérarchique dans les étapes encodeur-décodeur, maintenant un conditionnement sémantique continu dans l'ensemble du pipeline.

Soit $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ une image d'entrée, $\mathbf{I}_{\text{med}} \in \mathbb{R}^{3\times 448\times 448}$ l'entrée MedSigLIP correspondante, et T un prompt spécifié par l'utilisateur (e.g., "*segmenter l'épicarde du ventricule gauche*"). L'objectif est de prédire une carte de probabilités de classe $\mathbf{P} \in \mathbb{R}^{K\times H\times W}$, où K dénote le nombre de catégories cibles, incluant le background de l'image. Le processus de segmentation est explicitement conditionné sur le prompt T par des embeddings de domaine médical, qui modulent le traitement de caractéristiques dans toutes les couches, assurant un alignement sémantique cohérent dans l'ensemble du pipeline.

7.2.1.1 Extraction et Intégration de Caractéristiques de Domaine Médical

L'intégration de connaissances spécifiques au domaine médical représente une rupture des approches de segmentation conventionnelles et traite des limitations critiques identifiées dans les méthodes de fusion cross-modale existantes. Pour incorporer efficacement

la connaissance de domaine médical, des caractéristiques vision et texte synchronisées sont extraites en utilisant MedSigLIP, un modèle fondamental vision-langage médical spécialisé pré-entraîné sur des ensembles de données d'imagerie médicale vastes avec des descriptions cliniques correspondantes.

Étant donné la paire d'entrée $(\mathbf{I}_{\text{med}}, T)$, des caractéristiques médicales multi-échelles sont obtenues par le traitement du modèle fondamental :

$$\{\mathbf{e}_{\text{vis}}^{(l)}, \mathbf{e}_{\text{text}}\} = \text{MedSigLIP}(\mathbf{I}_{\text{med}}, T), \tag{7.1}$$

où $\mathbf{e}_{\text{vis}}^{(l)} \in \mathbb{R}^{N_p \times 256}$ représente les caractéristiques visuelles extraites aux couches $l \in \{6, 12, 18, 24\}$ avec $N_p = 1023$ patches, et $\mathbf{e}_{\text{text}} \in \mathbb{R}^{256}$ représente l'embedding du texte médical. Ces caractéristiques sont interpolées spatialement pour correspondre aux résolutions de cartes de caractéristiques TD-DIMB par des mécanismes d'interpolation apprenables qui préservent le contenu sémantique tout en s'adaptant aux exigences spatiales. Les poids MedSigLIP restent gelés durant l'entraînement, préservant les a priori de domaine médical pré-entraînés tout en permettant un fine-tuning efficace de l'architecture de segmentation.

Le modèle fondamental médical spécialisé capture des relations sémantiques spécifiques au domaine, hiérarchies anatomiques, et dépendances de contexte clinique essentielles pour une interprétation précise d'images médicales. Les caractéristiques visuelles hiérarchiques capturent progressivement des concepts médicaux abstraits, des détails anatomiques de bas niveau aux relations structurelles de haut niveau, tandis que l'embedding textuel fournit une interprétation robuste de terminologie clinique dans diverses spécialités médicales et vocabulaires institutionnels.

7.2.1.2 Architecture d'Encodage Hiérarchique Guidée par Prompt

La voie d'encodage de TD-DIMB implémente une stratégie de traitement hiérarchique qui intègre le guidage sémantique durant l'extraction de caractéristiques. L'image d'entrée $\mathbf{I}$ est transformée en une carte de caractéristiques $\mathbf{F}_0$ par une convolution 3×3 avec 16 canaux de sortie, établissant la base pour le traitement hiérarchique ultérieur.

L'encodeur comprend L étapes séquentielles, chacune incluant une convolution à stride 3×3 pour le downsampling spatial suivie d'un bloc DIMB qui encode conjointement l'information visuelle et sémantique. À l'étape l, la carte de caractéristiques de sortie est calculée comme :

$$\mathbf{F}_l = \text{DIMB}_l(\text{Conv}_{\text{stride}=2}(\mathbf{F}_{l-1}), \mathbf{e}_{\text{text}}, \mathbf{e}_{\text{vis}}^{(l)}), \tag{7.2}$$

où la dimension state-space interne croît avec la profondeur, permettant une modélisation efficace de contexte à long terme avec guidage de domaine médical.

À la couche la plus profonde de l'encodeur, un bloc TD-SS2D est employé pour capturer les dépendances sémantiques globales sans contraintes de parallélisme convolutionnel. Ce module sert de pont sémantique entre les voies encodeur et décodeur,

maximisant la modélisation de contexte conditionné par prompt à la résolution spatiale la plus grossière.

7.2.1.3 Décodage Multi-Résolution avec Cohérence Sémantique

L'architecture de décodeur implémente une conception symétrique qui reflète la structure de l'encodeur, récupérant progressivement la résolution spatiale tout en maintenant la cohérence sémantique sur les échelles. Chaque étape de décodeur implémente des stratégies de fusion de caractéristiques qui intègrent de multiples sources d'information pour une qualité de reconstruction optimale.

Chaque étape de décodeur i fusionne les caractéristiques de la sortie décodeur upsample $\mathbf{Y}_{i+1}$ et la connexion skip encodeur correspondante $\mathbf{X}_i$ par une concaténation par canal et raffinement ultérieur basé sur DIMB :

$$\mathbf{Y}_i = \text{DIMB}_i \left(\text{Conv}(\text{Upsample}(\mathbf{Y}_{i+1})) \| \mathbf{X}_i, \mathbf{e}_{\text{text}}, \mathbf{e}_{\text{vis}}^{(i)} \right) \Bigg(\tag{7.3}$$

où $\|$ dénote la concaténation par canal. Chaque DIMB décodeur raffine la représentation fusionnée de manière adaptative au prompt, préservant la cohérence sémantique sur les échelles tout en améliorant progressivement la résolution spatiale.

À l'étape finale de décodeur, une convolution 1×1 projette la sortie vers K canaux correspondant aux catégories cibles, suivie par une activation softmax pour produire les probabilités de segmentation finales $\mathbf{P}$. Cette projection finale maintient l'alignement sémantique avec le prompt original tout en produisant des distributions de probabilités cliniquement interprétables sur les structures anatomiques.

Une caractéristique distinctive de TD-DIMB est l'injection cohérente de sémantiques de prompt via des embeddings de domaine médical dans toutes les couches de traitement. Contrairement aux pipelines conventionnels qui traitent l'information textuelle comme entrée statique ou guidage a posteriori, TD-DIMB intègre le conditionnement sémantique directement dans le processus d'apprentissage, guidant les mécanismes d'attention, stratégies de sélection de caractéristiques, et opérations de décodage dans l'ensemble de la hiérarchie réseau.

7.2.2 Modules Dense Inverted Mamba Bottleneck (DIMB)

Le module Dense Inverted Mamba Bottleneck (DIMB) représente l'innovation architecturale centrale du framework TD-DIMB, conçu pour capturer tant la structure spatiale locale que les dépendances sémantiques globale tout en maintenant l'efficacité architecturale essentielle pour le déploiement clinique.

Inspiré par la conception inverted residual de MobileNetV3, DIMB étend cette formulation légère avec des patterns de connectivité dense, mécanismes de modulation token-aware, et une voie de modélisation state-space guidée par prompt pour atteindre l'alignement sémantique à chaque étape de transformation de caractéristiques.

La conception architecturale des modules DIMB implémente une stratégie de traitement multi-branche qui capture des aspects complémentaires d'information spatiale et sémantique. Chaque bloc DIMB commence par une étape d'expansion de canaux, appliquant une convolution pointwise (1×1) pour augmenter la capacité représentationnelle du tenseur d'entrée. Suivant l'expansion de canaux, les caractéristiques expandues sont traitées par une couche convolutionnelle depthwise, qui capture des patterns spatiaux localisés pertinents aux contours anatomiques, variations texturales, et caractéristiques de frontières. Les caractéristiques expandues sont ensuite divisées en deux branches de traitement parallèles qui capturent des aspects complémentaires de structure spatiale et de contenu sémantique.

La première branche suit une voie de traitement convolutionnelle conventionnelle pour préserver les détails spatiaux haute fréquence essentiels pour une localisation précise de frontières et une discrimination anatomique fine. La seconde branche est dirigée vers le module Text-Driven Selective Scan 2D (TD-SS2D), représentant une innovation en traitement de caractéristiques cross-modal. Cette voie de traitement enrichit la représentation de caractéristiques d'information sémantique adaptative au prompt en modélisant les dépendances à long terme par un bloc Vision Mamba, modulée par des embeddings de domaine médical dérivés de MedSigLIP.

Les sorties de ces deux branches parallèles sont ensuite fusionnées en utilisant un mécanisme de gating appris $\mathcal{G}$, qui repondère adaptativement le tenseur de caractéristiques combiné selon les exigences spécifiques à la tâche et pertinence sémantique. En pratique, $\mathcal{G}$ est implémenté comme une convolution 1×1 suivie d'une fonction d'activation sigmoïde, appliquée par canal sur les caractéristiques visuelles concaténées.

Pour raffiner la représentation fusionnée et améliorer l'alignement clinique, nous introduisons un mécanisme squeeze-and-excitation (SE) conditionné par domaine médical qui représente une avancée sur les approches d'attention de canal conventionnelles. Contrairement aux blocs SE traditionnels qui s'appuient uniquement sur le global average pooling pour l'estimation d'importance de canal, notre variante incorpore tant les caractéristiques texte que vision de MedSigLIP pour accentuer les canaux alignés avec le focus sémantique du prompt clinique.

Après cette étape d'amélioration sémantique, une projection finale $\mathcal{P}$ est appliquée pour restaurer la dimensionnalité de canal originale tout en préservant la capacité représentationnelle améliorée développée dans le pipeline de traitement. Plutôt que d'employer l'addition résiduelle conventionnelle, nous adoptons une stratégie de connectivité dense en concaténant l'entrée du bloc avec sa sortie traitée.

Formellement, la transformation DIMB complète intègre tous les composants architecturaux en un framework de traitement unifié :

$$\mathbf{X}_{\text{out}} = \text{Concat}\left(\mathbf{X}, \mathcal{P}\left(\text{SE}\left(\mathcal{G}\left(\mathcal{C}(\mathbf{X}), \tilde{\mathbf{X}}\right)\left(\mathbf{e}_{\text{text}}, \mathbf{e}_{\text{vis}}^{(l)}\right)\right)\right)\right) \tag{7.4}$$

où $\tilde{\mathbf{X}} = \text{TD-SS2D}(\mathbf{X}, \mathbf{e}_{\text{text}}, \mathbf{e}_{\text{vis}}^{(l)})$, $\mathcal{C}(\cdot)$ dénote le flux convolutionnel, $\mathcal{G}(\cdot)$ est l'unité de gating qui apprend à fusionner les deux branches via modulation feature-wise, $\text{SE}(\cdot)$

est le module squeeze-and-excitation conditionné par le domaine médical, et $\mathcal{P}(\cdot)$ est la couche de projection.

7.2.3 Mécanisme Text-Driven Selective Scan 2D (TD-SS2D)

Une innovation centrale de l'architecture TD-DIMB est le module Text-Driven Selective Scan 2D (TD-SS2D), qui introduit un nouveau mécanisme pour la modélisation de caractéristiques visuelles cross-modale en unifiant le traitement state-space à complexité linéaire avec le conditionnement sémantique guidé par prompt. Ce composant architectural traite des limitations critiques dans les approches de fusion cross-modale existantes tout en fournissant une efficacité de calcul supérieure comparée aux mécanismes traditionnels basés sur l'attention.

Tandis que les mécanismes de self-attention conventionnels ont démontré des capacités de raisonnement global puissantes dans diverses tâches vision-langage, leur complexité de calcul quadratique présente des défis d'évolutivité importants pour les applications d'imagerie médicale haute résolution. En revanche, TD-SS2D fournit une alternative légère mais expressive qui maintient les capacités de modélisation de dépendances à long terme tout en permettant l'intégration efficace de guidage sémantique via des prompts de langage naturel et connaissance de domaine médical.

La base architecturale de TD-SS2D implémente un traitement multi-flux qui capture des aspects complémentaires de structure spatiale et de contenu sémantique. Le module commence par une étape d'expansion de canal, où la carte de caractéristiques d'entrée est projetée linéairement vers un espace de dimension supérieure qui permet une modélisation sémantique plus riche et capacité représentationnelle améliorée. Ce tenseur expandu est ensuite divisé en deux flux de traitement parallèles : un flux spatial pour le mélange de caractéristiques locales et un flux sémantique pour la modulation guidée par prompt.

Le flux spatial est traité au moyen d'une convolution depthwise, permettant de capter des motifs anatomiques fins tels que les contours et les transitions de texture, éléments essentiels à l'interprétation des images médicales. Parallèlement, le flux sémantique est modulé par l'embedding textuel $\mathbf{e}\text{text}$ et par les caractéristiques visuelles $\mathbf{e}\text{vis}^{(l)}$ issues de MedSigLIP, un modèle fondamental spécialisé dans le domaine médical. Cette double stratégie de modulation est réalisée à l'aide de fonctions de gating apprises, qui adaptent dynamiquement les représentations visuelles en fonction de leur pertinence.

Pour améliorer l'alignement cross-modal et la compréhension sémantique, un mécanisme léger de *cross-attention* est appliqué afin de fusionner le flux visuel modulé avec les *embeddings* textuels et visuels issus de MedSigLIP. Dans cette configuration, les caractéristiques visuelles servent de requêtes, tandis que les *embeddings* du domaine médical jouent à la fois le rôle de clés et de valeurs. Ce mécanisme permet une réinterprétation dynamique du contenu visuel en fonction de l'intention clinique et des connaissances spécialisées du domaine.

À la suite de cette étape d'alignement sémantique, la représentation obtenue est transmise à un bloc de modélisation à espace d'états (*State-Space Model, SSM*) inspiré de Mamba, marquant une rupture avec les approches de traitement conventionnelles. Contrairement aux couches convolutionnelles ou basées sur les *transformers*, qui imposent des schémas de traitement fixes, cette formulation conserve un état interne évolutif lors de l'exploration du domaine spatial, guidé par des dynamiques de transition apprises. Ces dernières s'adaptent aussi bien aux relations spatiales qu'au contenu sémantique.

La sortie finale du module TD-SS2D subit une normalisation suivie d'une projection vers la dimensionnalité d'origine des caractéristiques, assurant ainsi la compatibilité avec l'architecture TD-DIMB dans son ensemble, tout en préservant l'enrichissement sémantique acquis au cours du traitement. Le flux de calcul complet s'exprime comme suit :

$$\mathbf{X}_{\text{out}} = \text{Linear}\left(\text{Norm}\left(\text{SSM}\left(\text{Attn}\left(\mathbf{G} \odot \tilde{\mathbf{X}}_v, \mathbf{e}_{\text{text}}, \mathbf{e}_{\text{vis}}^{(l)}\right)\right)\right)\right) \tag{7.5}$$

où $\tilde{\mathbf{X}}_v = \text{DWConv}(\mathbf{X}_v)$ et $\mathbf{X}_v$ désigne le flux visuel d'entrée, $\mathbf{e}_{\text{text}}$ l'*embedding* textuel médical, $\mathbf{e}_{\text{vis}}^{(l)}$ les caractéristiques visuelles issues de MedSigLIP, et $\mathbf{G}$ le vecteur de *gating* spécifique au domaine médical appliqué élément par élément.

Le module TD-SS2D constitue ainsi le moteur de raisonnement sémantique de TD-DIMB, transformant les données visuelles brutes en représentations alignées sur le *prompt* et adaptées à la tâche, avec une efficacité de calcul élevée. L'architecture TD-SS2D est illustrée à la Fig. 7.2.

7.2.4 Formulation de la Reinforced Gaussian Dice Loss (RGDL)

Pour renforcer la capacité de généralisation de TD-DIMB dans divers contextes cliniques et dépasser les limites des approches d'optimisation classiques, nous introduisons la *Reinforced Gaussian Dice Loss* (RGDL). Cette fonction objectif hybride capture simultanément le chevauchement spatial, l'incertitude des frontières et la pénalisation asymétrique des erreurs de prédiction. Elle permet ainsi de traiter plusieurs défis récurrents de la segmentation d'images médicales, tels que l'ambiguïté des tissus mous, le déséquilibre de classes et la variabilité des annotations dans les ensembles de données cliniques réels.

Les fonctions de perte courantes en segmentation, comme la *cross-entropy* ou la Dice loss standard, considèrent toutes les erreurs de manière uniforme, sans prendre en compte ni leur importance clinique relative, ni l'incertitude inhérente aux annotations médicales. Cette limitation devient particulièrement problématique dans des situations où différents types d'erreurs ont des conséquences cliniques inégales sur la prise en charge des patients. De plus, les annotations expertes présentent souvent une incertitude intrinsèque au niveau des frontières anatomiques, liée aux artefacts d'imagerie, aux effets de volume partiel ou encore aux divergences d'interprétation entre spécialistes.

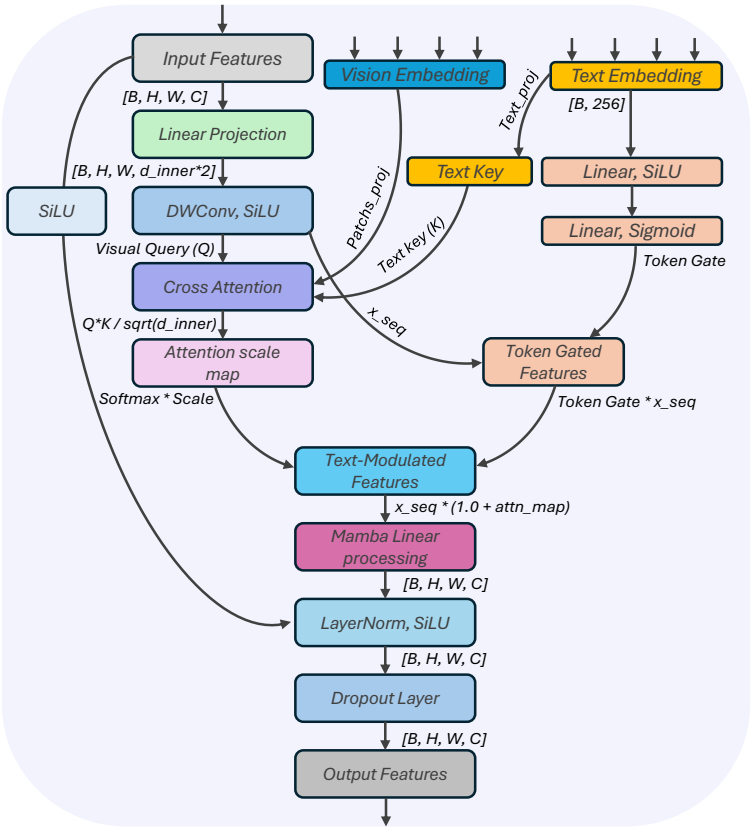

FIGURE 7.2 – Conception du bloc *Text-Driven Selective Scan 2D* (TD-SS2D).

La RGDL surmonte ces limites en modélisant explicitement l'incertitude des annotations et en intégrant les priorités cliniques directement dans la fonction d'optimisation. Dans une tâche de segmentation binaire, l'image d'entrée $X \in \mathbb{R}^{B\times C\times H\times W}$ est transformée en un masque probabiliste $\hat{Y} \in [0,1]^{B\times 1\times H\times W}$, associé à un masque de vérité terrain binaire $Y \in \{0,1\}^{B\times 1\times H\times W}$.

Afin de représenter l'incertitude aux frontières et de tenir compte de l'ambiguïté des annotations médicales, une version lissée $\tilde{Y} \in [0,1]^{B\times 1\times H\times W}$ est construite par convolution gaussienne appliquée au masque original Y :

$$\tilde{Y} = Y * G_\sigma, \quad G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{7.6}$$

où G_σ est un noyau gaussien 2D isotrope de déviation standard σ, contrôlant l'ampleur de la modélisation d'incertitude aux frontières, et $*$ désigne l'opération de convolution.

La perte RGDL combine une version modifiée de la Dice loss, calculée avec le masque lissé, et un terme de renforcement basé sur des poids d'importance apprenables, traduisant directement les priorités cliniques :

$$\mathcal{L}_{\text{RGDL}}(\tilde{Y}, Y, \hat{Y}) = \lambda_1 \cdot \mathcal{L}_{\text{Dice}}(\tilde{Y}, Y, \hat{Y}) + \lambda_2 \cdot \mathcal{L}_{\text{RL}}(Y, \hat{Y}), \tag{7.7}$$

où $\lambda_1 + \lambda_2 = 1$ sont des coefficients qui équilibrent la contribution du composant sensible aux frontières et celle du schéma de pondération inspiré de l'apprentissage par renforcement.

Le terme de Dice modifié améliore la sensibilité aux contours subtils en intégrant le masque lissé dans le numérateur, tout en conservant le masque binaire original dans le dénominateur :

$$\mathcal{L}_{\text{Dice}}(\tilde{Y}, Y, \hat{Y}) = 1 - \frac{2\sum_i \tilde{y}_i \hat{y}_i + \epsilon}{\sum_i y_i + \sum_i \hat{y}_i + \epsilon}, \tag{7.8}$$

où i indexe les pixels du lot d'entraînement et ϵ est une constante de régularisation destinée à éviter la division par zéro et à assurer la stabilité numérique.

Pour refléter l'importance clinique relative des erreurs et introduire des priorités spécifiques au domaine médical, la perte de renforcement attribue des poids apprenables (α, β, γ) aux contributions respectives des vrais positifs, faux positifs et faux négatifs :

$$\begin{aligned}\mathcal{L}_{\text{RL}}(Y, \hat{Y}) = -\frac{1}{\sum_i y_i + \epsilon} \cdot \Big(&\alpha \sum_i \tilde{y}_i \hat{y}_i - \beta \sum_i (1 - y_i)\hat{y}_i \\ &- \gamma \sum_i y_i (1 - \hat{y}_i)\Big).\end{aligned} \tag{7.9}$$

Chaque coefficient est paramétré par une variable apprenable activée par une fonction sigmoïde, par exemple $\alpha = \alpha_{\max} \cdot \sigma(\theta_\alpha)$, et optimisé conjointement avec les paramètres du réseau. Le modèle apprend ainsi à récompenser ou pénaliser de manière adaptative certains résultats de prédiction au cours de l'entraînement.

Pris ensemble, ces deux composantes complémentaires guident le réseau à optimiser à la fois la précision du chevauchement spatial et la pertinence clinique des segmentations. L'objectif RGDL établit donc un compromis entre l'adaptation souple aux frontières et une régularisation spécifique aux erreurs, améliorant la robustesse et la précision de TD-DIMB sur diverses modalités d'imagerie.

7.3 Conception Expérimentale et Implémentation

7.3.1 Protocole d'Évaluation et Ensembles de Données

Afin d'évaluer à la fois la performance spécialisée et les capacités de généralisation entre modalités d'imagerie, nous adoptons un protocole d'évaluation à deux volets. Dans le paradigme *task-specific*, TD-DIMB est entraîné et testé séparément sur deux jeux de données cliniquement distincts : CAMUS (imagerie cardiaque ultrasonore) et autoPET22

(imagerie oncologique PET/CT). CAMUS met en évidence des défis propres à l'échographie, tels que le bruit *speckle* et l'ombre acoustique, tandis qu'autoPET22 introduit des difficultés liées à la fusion de modalités (PET et CT) et à la segmentation de structures oncologiques.

En complément, le paradigme de généralisation universelle consiste à entraîner TD-DIMB sur un ensemble regroupant 14 bases de données d'imagerie médicale couvrant diverses régions anatomiques : structures abdominales, régions neurologiques, systèmes pulmonaires, organes urologiques et reproductifs, imagerie mammaire, structures cardiaques, ainsi que des tâches de segmentation tumorale. Ces ensembles incluent un large éventail de modalités, telles que l'échographie, la tomodensitométrie, l'imagerie par résonance magnétique, la tomographie par émission de positons et la radiographie conventionnelle. Le modèle est ensuite évalué sur deux ensembles jamais utilisés lors de l'entraînement, QaTa-COV19 (imagerie radiographique thoracique) et ATLAS v2.0 (segmentation de lésions d'AVC en IRM), afin de mesurer directement sa capacité de généralisation vers de nouvelles structures anatomiques et des applications cliniques inédites, sans adaptation *task-specific*.

Étant donné que TD-DIMB repose sur un cadre de segmentation bidimensionnel, tous les ensembles de données volumétriques ont été systématiquement décomposés en coupes axiales, puis redimensionnés à une résolution standard de 256×256 pixels. Cette étape assure un compromis entre efficacité computationnelle, couverture anatomique et cohérence inter-protocole d'acquisition.

7.3.1.1 Pipeline de Prétraitement et Standardisation

Le cadre expérimental implémente un pipeline de prétraitement unifié conçu pour traiter les caractéristiques variées des données d'imagerie médicale issues de différentes modalités, tout en préservant l'information clinique essentielle. Cette approche permet de gérer les variations importantes des distributions d'intensité, des résolutions spatiales et des artefacts propres aux différentes techniques d'imagerie.

Nous avons appliqué des stratégies de normalisation spécifiques à chaque modalité. Toutes les images ont été ramenées dans l'intervalle [0,1] afin d'assurer une homogénéité des distributions d'intensité, tout en conservant les contrastes cliniquement pertinents. Les coupes IRM et échographiques ont en outre subi une normalisation *z-score* par coupe, afin de compenser les variations d'intensité inhérentes et les facteurs de mise à l'échelle dépendant de l'acquisition, caractéristiques de ces modalités.

Les étiquettes anatomiques correspondantes ont été découpées axialement selon la même approche que les images, chaque structure étant formulée comme une tâche de segmentation binaire indépendante pour permettre un *conditioning* flexible guidé par *prompt*. Pour renforcer la robustesse du modèle et atténuer les déséquilibres de classes fréquents dans les ensembles d'imagerie médicale, nous avons systématiquement exclu les coupes ne contenant que de l'arrière-plan, tout en imposant un échantillonnage équilibré d'exemples positifs durant l'entraînement.

TABLE 7.1 – Spécifications d'Ensembles de Données d'Entraînement Universel

Ensemble de Données	**Description Clinique**	**Études**	**Modalité**	**Focus Anatomique Primaire**
AbdomenCT-1K [182]	Segmentation abdominale multi-organe avec couverture anatomique comprehensive	361	CT	Foie, reins, rate, pancréas, vésicule biliaire
AMOS [183]	Segmentation d'organes abdominaux cross-modale pour apprentissage modality-invariant	240	CT, IRM	Organes abdominaux avec consistance cross-modale
BraTS [172]	Segmentation de tumeurs cérébrales incluant sous-régions tumorales et structures péritumorales	6,096	IRM	Gliomes, régions rehaussantes, noyau tumoral, œdème
CHAOS [184]	Défi d'organes abdominaux cross-modal à travers protocoles d'imagerie	40	CT, T2-IRM	Foie, reins, rate avec alignement modal
KiTS 2023 [185]	Segmentation de rein et tumeurs rénales avec annotations pathologiques	489	CT	Reins, tumeurs rénales, lésions kystiques
LiTS [186]	Segmentation de tumeurs hépatiques avec délimitation précise de frontières	131	CT	Parenchyme hépatique, lésions hépatiques
LUNA [187]	Analyse de nodules pulmonaires avec couverture pulmonaire comprehensive	888	CT	Tissu pulmonaire, nodules pulmonaires
MSD Collection [188]	Decathlon de Segmentation Médicale s'étendant sur dix tâches anatomiques	3,225	CT, IRM	Cœur, hippocampe, foie, poumon, pancréas, autres
PROMISE12 [189]	Segmentation de prostate utilisant IRM pondérée T2 haute résolution	37	T2-IRM	Glande prostatique, zone périphérique
ABUS [171]	Ultrason mammaire automatisé avec caractérisation tumorale 3D	200	US 3D	Lésions mammaires, changements fibrokystiques
BUSI [190]	Ultrason mammaire avec représentation de lésions bénignes et malignes	1312	US 2D	Masses mammaires, frontières tumorales
Total Segmentator V2 [173]	Segmentation d'atlas anatomique de 117 structures anatomiques	1,228	CT	Cartographie anatomique corps entier comprehensive
CAMUS [191]	Analyse de mouvement cardiaque utilisant imagerie échocardiographique multi-vue	500	US	Ventricule gauche, oreillette gauche, myocarde
autoPET [174]	Segmentation de lésions corps entier en imagerie PET oncologique	1,014	PET/CT	Lésions métaboliquement actives, charge tumorale

Le partitionnement des données au niveau patient a été rigoureusement appliqué afin de prévenir toute fuite d'information et d'assurer une évaluation valide des performances dans tous les scénarios. L'évaluation de généralisation universelle suit une répartition 80%/20% entre entraînement et validation, tandis que l'évaluation *task-specific* adopte une partition 70%/10%/20% pour l'entraînement, la validation et le test, garantissant des ensembles indépendants pour l'ajustement des hyperparamètres et l'évaluation finale.

Les stratégies d'augmentation de données combinent des transformations *online* et *offline* visant à accroître la robustesse du modèle tout en respectant le réalisme anatomique. Les augmentations *online* incluent des déformations élastiques, transformations affines, corrections gamma et l'ajout de bruit gaussien au cours des itérations d'entraînement. Des transformations géométriques *offline* supplémentaires, telles que les inversions horizontales et verticales, les rotations aléatoires et les décalages spatiaux, sont appliquées en particulier sur les ensembles de petite taille afin d'assurer une diversité suffisante des données d'entraînement.

7.3.2 Détails d'Implémentation et Protocoles d'Entraînement

L'implémentation expérimentale de TD-DIMB requiert une infrastructure de calcul performante et des protocoles d'entraînement soigneusement optimisés, capables de gérer les exigences complexes liées à la diversité des modalités d'imagerie, tout en assurant une convergence stable et efficace.

7.3.2.1 Infrastructure de calcul et Configuration Matérielle

Toutes les validations expérimentales ont été menées dans un environnement de calcul haute performance spécifiquement configuré pour les applications d'apprentissage profond. La plateforme principale repose sur deux GPUs NVIDIA RTX A6000 offrant une mémoire combinée de 96 GB, indispensable pour traiter de larges tailles de lot et répondre aux besoins de l'imagerie médicale haute résolution.

L'environnement logiciel utilise PyTorch comme framework d'apprentissage profond principal, exploitant ses graphes computationnels dynamiques et sa gestion optimisée de la mémoire GPU pour implémenter les composants architecturaux complexes de TD-DIMB.

7.3.2.2 Stratégie d'Optimisation et Configuration d'Entraînement

Le protocole d'entraînement adopte une stratégie d'optimisation avancée, adaptée aux objectifs multiples de la segmentation d'images médicales à travers diverses modalités et structures anatomiques, tout en garantissant une convergence stable.

Nous avons utilisé l'optimiseur Adam avec accélération de Nesterov, une taille de lot de 16 et un taux d'apprentissage initial de 10^{-4}, déterminé à l'issue d'une exploration systématique sur des ensembles de validation représentatifs. Le programme de taux d'ap-

prentissage suit un schéma cosinusoïdal d'*annealing*, favorisant une convergence régulière et évitant les arrêts prématurés.

Les entraînements s'étendent jusqu'à 1000 époques, avec arrêt précoce basé sur la stabilisation de la perte de validation, afin de prévenir le surapprentissage tout en assurant un apprentissage suffisant sur l'ensemble des données.

7.3.2.3 Validation Croisée et Robustesse Statistique

Pour l'évaluation *task-specific*, nous avons mis en œuvre une validation croisée en 4 plis, garantissant la robustesse statistique et la fiabilité des mesures de performance. Les partitions respectent le niveau patient afin d'éviter toute fuite d'information, tout en assurant une représentation équilibrée des profils cliniques dans les différents plis.

Dans le cadre de la généralisation universelle, TD-DIMB est entraîné sur le pool multi-organes et multi-modalités avec une stratégie d'échantillonnage dynamique de *prompts*. À chaque itération, un *prompt* sémantique est échantillonné aléatoirement pour conditionner l'identification de la structure cible, permettant une adaptation *task-specific* sans nécessiter ni modification architecturale ni entraînements séparés pour chaque structure.

Cette stratégie de *prompting* dynamique constitue une avancée par rapport aux approches multi-tâches conventionnelles, en permettant au modèle d'apprendre des mécanismes d'alignement généralisables entre modalités et d'adapter ses prédictions à de nouvelles cibles anatomiques et descriptions cliniques sans réentraînement dédié.

7.4 Résultats et Analyse Globale

Cette section présente une évaluation complète du framework TD-DIMB à travers l'analyse systématique de ses composants architecturaux, de ses capacités de généralisation entre modalités d'imagerie et de sa viabilité en contexte clinique. L'analyse suit le protocole d'évaluation à deux volets introduit précédemment, en fournissant à la fois une étude détaillée des différents composants et une évaluation plus large des performances dans divers scénarios d'imagerie médicale.

7.4.1 Études d'Ablation et Analyse de Composants

Afin d'évaluer systématiquement la contribution de chaque composant architectural et des stratégies d'entraînement du framework TD-DIMB, nous avons mené des études d'ablation étendues sur l'ensemble des bases de données considérées. Ces expérimentations apportent des indications essentielles sur l'importance relative de chaque innovation et permettent d'isoler l'impact des modules proposés sur les performances de segmentation ainsi que sur la complexité de calcul.

Toutes les expérimentations d'ablation ont suivi des protocoles d'entraînement homogènes, avec une évaluation reposant sur le Dice Similarity Coefficient (DSC), la dis-

tance de Hausdorff au 95^{e} percentile (HD95), le nombre de paramètres et le volume d'opérations en virgule flottante (FLOPs). Ce choix de métriques assure une évaluation rigoureuse, couvrant à la fois les performances de segmentation et l'efficacité en calcul, deux critères essentiels pour envisager un déploiement clinique.

7.4.1.1 Analyse Architecturale par Composant

En partant d'une architecture U-Net de référence représentant les approches de segmentation conventionnelles, nous avons introduit successivement trois composants centraux : le *conditioning* basé sur des *prompts* textuels, les modules Dense Inverted Mamba Bottleneck (DIMB) et la fonction d'optimisation Reinforced Gaussian Dice Loss (RGDL).

Le Tableau 7.2 présente les résultats complets de cette analyse systématique des composants sur l'ensemble des bases de données d'évaluation.

TABLE 7.2 – Étude d'ablation component-wise. Dice (%) ↑ et HD95 (pixels) ↓.

Variante de Modèle	CAMUS		autoPET22		ATLAS		QaTa-COV19	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
U-Net	91.70	35.42	89.42	25.84	90.24	24.67	84.73	36.89
+Prompting	93.15	28.73	90.58	23.21	91.76	22.45	86.41	31.67
+DIMB	94.09	25.34	91.86	20.89	92.87	20.12	87.68	28.73
TD-DIMB	**97.07**	**19.47**	**93.21**	**16.84**	**94.65**	**16.23**	**89.87**	**22.15**

L'ajout de mécanismes de *conditioning* par *prompt* a entraîné des améliorations immédiates et substantielles sur l'ensemble des bases de données évaluées, avec des gains particulièrement marqués en précision des frontières, comme en témoignent les réductions constantes du HD95 et les améliorations du DSC. Ces résultats apportent une démonstration solide de l'importance fondamentale de la guidance sémantique en segmentation d'images médicales, en montrant que le *conditioning* entre modalités permet de surmonter certaines limites inhérentes aux approches exclusivement visuelles.

L'intégration des modules DIMB a conduit à des gains supplémentaires, soulignant l'efficacité de la modélisation par *state-space*, combinée à une connectivité dense, pour capturer à la fois les structures spatiales locales et les dépendances sémantiques globales. Le modèle TD-DIMB complet, intégrant l'ensemble des innovations proposées, a atteint des performances optimales sur toutes les métriques d'évaluation, confirmant l'effet synergique de la combinaison des avancées architecturales et des stratégies d'optimisation.

7.4.1.2 Analyse de Sensibilité aux Prompts et Multi-Modal

L'analyse de sensibilité aux *prompts* explore la réactivité du modèle face à différentes stratégies de *prompting*, en tirant parti des capacités multi-*modal* de MedSigLIP. Nous avons systématiquement évalué quatre configurations distinctes : (1) configuration **No Prompt**, éliminant toute guidance textuelle et visuelle de MedSigLIP ; (2) configuration

Visual Prompt, exploitant uniquement les caractéristiques visuelles issues de MedSigLIP ; (3) configuration **Text Prompt**, reposant exclusivement sur des descriptions textuelles médicales ; (4) configuration **Multi-modal Prompt**, combinant de manière synchronisée les caractéristiques textuelles et visuelles.

Le Tableau 7.3 présente une analyse complète des performances obtenues selon ces différentes configurations de *prompting*.

TABLE 7.3 – Effet des Stratégies de *Prompting* avec Guidance Multi-*Modal* de MedSigLIP.

Type de Prompt	CAMUS		autoPET22		ATLAS		QaTa-COV19	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
Aucun Prompt	94.84	26.73	91.93	21.45	93.37	20.89	88.09	29.34
Prompt Visuel	95.42	24.67	92.18	19.78	93.68	19.45	88.39	27.56
Prompt Textuel	96.89	21.12	93.07	17.69	94.41	17.54	89.49	23.34
Prompt Multi-modal	**97.07**	**19.47**	**93.21**	**16.84**	**94.65**	**16.23**	**89.87**	**22.15**

L'absence de *prompt* de guidage a entraîné une dégradation constante des performances sur l'ensemble des bases de données évaluées, soulignant l'importance critique de l'apport de connaissances médicales pour atteindre une segmentation optimale. Le

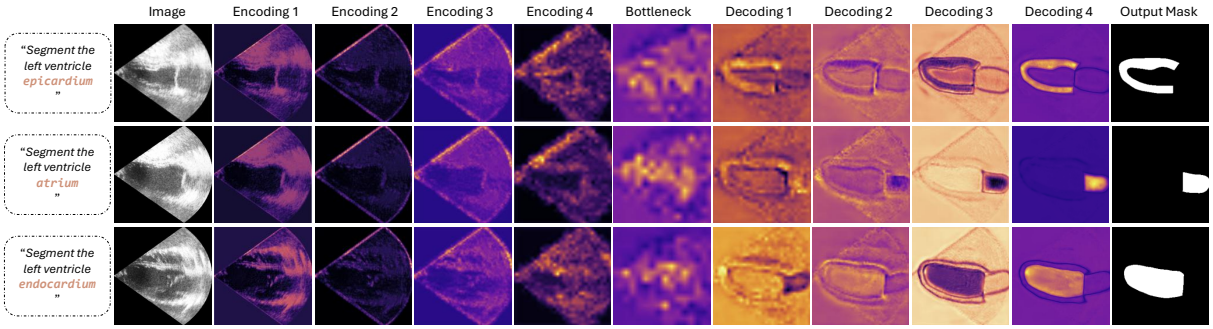

FIGURE 7.3 – Extraction de caractéristiques guidée par *prompt* dans TD-DIMB.

Le *prompting* textuel seul a surpassé le *prompting* visuel dans la plupart des scénarios, avec des gains particulièrement notables sur des ensembles de données difficiles tels que QaTa-COV19. Ce résultat met en évidence l'impact déterminant de la compréhension sémantique clinique pour l'analyse d'images médicales. La configuration multi-*modal prompt* a atteint des performances optimales sur l'ensemble des métriques, en combinant la compréhension du texte médical et la vision pour maximiser à la fois l'alignement clinique et la précision de segmentation.

Afin d'évaluer plus en profondeur la robustesse du système, nous avons examiné les performances de TD-DIMB face à des formulations de *prompt* s'écartant sensiblement

du vocabulaire d'entraînement. Le modèle a démontré une compréhension sémantique robuste allant au-delà du simple *matching* lexical. Bien qu'entraîné sur une terminologie médicale spécifique telle que « tumor », « lesion » et « stroke lesion », des tests réalisés avec des expressions équivalentes sur le plan sémantique, telles que « infected region », « abnormal area » et « damaged tissue », ont conduit à des segmentations réussies. Cela illustre la capacité du modèle à gérer la variabilité terminologique et les synonymes cliniques.

Comme illustré dans la Fig. 7.3, les *prompts* orientent l'activation de régions distinctes à travers les couches encodeur-décodeur lors du ciblage de structures anatomiques telles que l'épicarde, l'oreillette et l'endocarde, démontrant une modulation efficace adaptée au domaine médical.

7.4.1.3 Analyse Comparative des Modèles Fondamentaux Médicaux

Pour valider le choix de MedSigLIP comme *backbone* de *prompting* optimal dans le cadre de TD-DIMB, nous avons mené des comparaisons approfondies avec différents modèles fondamentaux du domaine médical, en les évaluant sur l'ensemble des bases de données utilisées.

Le Tableau 7.4 présente une comparaison de performances couvrant des modèles *text-only*, tels que BioBERT et ClinicalBERT, ainsi que des modèles vision-langage multi-*modaux*, incluant MedCLIP et MedSigLIP.

TABLE 7.4 – Comparaison de différents backbones de prompting avec architecture TD-DIMB.

Backbone	CAMUS		autoPET22		ATLAS		QaTa-COV19	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
BioBERT [192]	96.85	21.45	92.99	17.92	94.40	17.56	89.32	23.78
ClinicalBERT [193]	96.87	21.23	93.09	17.78	94.38	17.67	89.46	23.45
MedCLIP [194]	96.98	20.78	93.15	17.34	94.61	16.89	89.76	22.67
MedSigLIP	**97.07**	**19.47**	**93.21**	**16.84**	**94.65**	**16.23**	**89.87**	**22.15**

Les modèles fondamentaux *text-only* ont montré des performances prometteuses, avec ClinicalBERT surpassant BioBERT sur la majorité des métriques, confirmant ainsi l'importance de la spécialisation clinique du modèle. Les modèles fondamentaux multi-*modaux* ont quant à eux présenté des avantages clairs par rapport aux approches uniquement textuelles, MedCLIP offrant des améliorations substantielles grâce à ses capacités de compréhension vision-texte synchronisée.

MedSigLIP a obtenu les meilleures performances sur l'ensemble des bases de données évaluées, avec des gains particulièrement notables sur des ensembles exigeants tels qu'ATLAS et QaTa-COV19. Ces résultats mettent en évidence l'alignement supérieur de MedSigLIP avec le domaine médical et ses capacités avancées de fusion multi-*modal*, validant ainsi notre choix de ce modèle fondamental comme *backbone* du framework TD-DIMB.

7.4.1.4 Évaluation de l'Impact de la Fonction de Perte RGDL

L'évaluation de la *Reinforced Gaussian Dice Loss* (RGDL) que nous proposons fournit des indications essentielles sur l'efficacité de stratégies d'optimisation informées par les priorités cliniques, en comparaison avec les fonctions de perte conventionnelles.

Le Tableau 7.5 présente une comparaison de performances couvrant plusieurs formulations de fonctions de perte, incluant des approches standards telles que la *binary cross-entropy* et la Dice loss, ainsi que des fonctions de perte spécialisées pour l'imagerie médicale, telles que la *Boundary Loss* et la *DHN-NCE*.

TABLE 7.5 – Comparaison de fonction de perte sur TD-DIMB. Dice (%) ↑ et HD95 (px) ↓.

Perte	CAMUS		autoPET22		ATLAS		QaTa-COV19	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
BCE	95.26	26.89	92.12	20.45	93.31	19.78	88.19	28.67
Dice	95.87	24.73	92.65	18.92	93.88	18.45	88.45	26.34
BCE + Dice	96.17	23.89	92.76	18.67	93.91	18.23	88.51	25.78
Focal	95.58	25.34	92.15	19.67	93.56	19.12	88.43	27.45
Tversky	95.87	24.56	92.32	19.23	93.88	18.67	88.64	26.12
Boundary Loss [195]	96.23	22.67	92.58	18.34	94.05	17.89	88.78	24.89
DHN-NCE [196]	96.38	21.89	92.71	17.92	94.08	17.67	89.11	23.45
RGDL	**97.07**	**19.47**	**93.21**	**16.84**	**94.65**	**16.23**	**89.87**	**22.15**

Les formulations de fonctions de perte standards ont montré des performances de référence raisonnables, la combinaison BCE + Dice offrant des améliorations modestes par rapport aux pertes individuelles. Les fonctions de perte spécifiques au domaine médical, en particulier la *Boundary Loss*, ont présenté des avantages clairs sur les formulations standards, confirmant l'importance de stratégies d'optimisation adaptées au contexte clinique.

La *DHN-NCE* a affiché des performances compétitives, notamment sur des ensembles de données exigeants tels que QaTa-COV19, soulignant l'intérêt des approches d'apprentissage contrastif pour les applications d'imagerie médicale. La RGDL a obtenu des résultats systématiquement supérieurs sur l'ensemble des bases de données et des métriques évaluées, avec des gains particulièrement marqués sur l'ensemble ATLAS, qui comporte des structures lésionnelles de faible contraste et de petite taille. Les améliorations substantielles observées à la fois sur le DSC et le HD95 mettent en évidence l'efficacité de la RGDL pour traiter l'ambiguïté des frontières et les déséquilibres de classes.

7.4.2 Résultats Quantitatifs

Nous avons mené une comparaison quantitative complète de TD-DIMB avec des modèles récents de segmentation de l'état de l'art, en considérant à la fois les protocoles d'évaluation *task-specific* et de généralisation universelle. Le Tableau 7.6 présente une comparaison détaillée des performances à travers des méthodes représentatives de l'état

de l'art, incluant des approches établies telles que nnUNet, des architectures récentes basées sur les *transformers*, des méthodes de modélisation *state-space*, ainsi que des frameworks de segmentation médicale guidés par des *prompts* spécialisés.

TABLE 7.6 – Quantitative comparison of TD-DIMB and state-of-the-art models on both task-specific and universal generalization settings using CAMUS, autoPET22, ATLAS, and QaTa-COV19 datasets.

Task-Specific Evaluation							
Method	Year	CAMUS (US)			autoPET22 (PET/CT)		
		Dice ↑	HD95 ↓	Recall ↑	Dice ↑	HD95 ↓	Recall ↑
nnUNet [176]	2021	92.30±3.4	32.45±4.2	94.56±2.9	90.13±4.2	24.73±2.8	92.21±3.6
SwinUMamba† [197]	2024	92.27±3.6	33.16±4.5	93.96±3.0	91.45±3.9	22.45±2.4	93.62±3.2
LViT [198]	2024	93.39±3.1	28.92±3.8	94.42±2.8	91.32±3.7	21.83±2.2	91.27±3.8
TexD-KD [199]	2025	93.09±3.3	26.78±3.6	94.61±2.8	91.65±3.6	20.67±2.1	92.89±3.5
RecLMIS [200]	2025	94.17±3.0	23.48±3.2	94.13±2.9	92.07±3.4	19.45±1.9	94.44±3.2
MedCLIP-SAMv2 [196]	2025	96.30±2.5	20.67±2.9	96.84±2.4	92.56±3.1	17.89±1.7	94.92±2.8
TD-DIMB (Ours)	2026	**97.07±2.1**	**19.47±2.8**	**97.23±2.3**	**93.21±2.9**	**16.84±1.6**	**95.67±2.7**
Universal Generalization Evaluation							
Method	Year	ATLAS (MRI)			QaTa-COV19 (X-Ray)		
		Dice ↑	HD95 ↓	Recall ↑	Dice ↑	HD95 ↓	Recall ↑
nnUNet [176]	2021	91.05±4.0	23.84±3.1	93.28±3.1	79.63±5.2	34.82±5.3	85.44±4.8
SwinUMamba† [197]	2024	91.75±3.7	22.67±2.9	94.04±3.0	82.73±4.9	32.17±4.9	86.83±4.5
LViT [198]	2024	92.64±3.4	21.34±2.6	94.92±2.7	83.81±4.7	30.41±4.6	87.18±4.2
TexD-KD [199]	2025	92.74±3.3	19.89±2.3	94.02±2.7	88.09±4.1	27.35±3.8	90.92±3.9
RecLMIS [200]	2025	93.09±3.2	18.67±2.1	95.83±2.5	86.13±4.4	25.89±4.2	89.36±3.9
MedCLIP-SAMv2 [196]	2025	93.81±2.8	17.45±1.8	95.67±2.3	88.22±3.9	23.84±3.7	90.85±3.5
TD-DIMB (Ours)	2026	**94.65±2.6**	**16.23±1.7**	**96.31±2.0**	**89.87±3.7**	**22.15±3.5**	**92.14±3.3**

L'évaluation quantitative met en évidence la supériorité constante de TD-DIMB sur l'ensemble des bases de données et métriques considérées. Dans les scénarios *task-specific*, le modèle atteint une précision de segmentation robuste et une meilleure délimitation des frontières à travers diverses modalités d'imagerie. En conditions de généralisation universelle, où les modèles sont évalués sur des ensembles totalement inédits, tels qu'ATLAS (IRM) et QaTa-COV19 (radiographie thoracique), TD-DIMB maintient des avantages substantiels sur toutes les approches de référence.

La performance de rappel élevée observée dans les tâches centrées sur les lésions souligne la capacité de TD-DIMB à détecter efficacement des structures cliniquement pertinentes, tout en réduisant les faux négatifs, y compris dans des scénarios hors distribution particulièrement difficiles.

7.4.3 Évaluation Qualitative et Interprétation Clinique

En complément de l'analyse quantitative, nous avons réalisé une évaluation qualitative systématique sur les mêmes ensembles diversifiés employés dans notre protocole. La Fig. 7.4 illustre une comparaison visuelle des résultats de TD-DIMB face à des méthodes

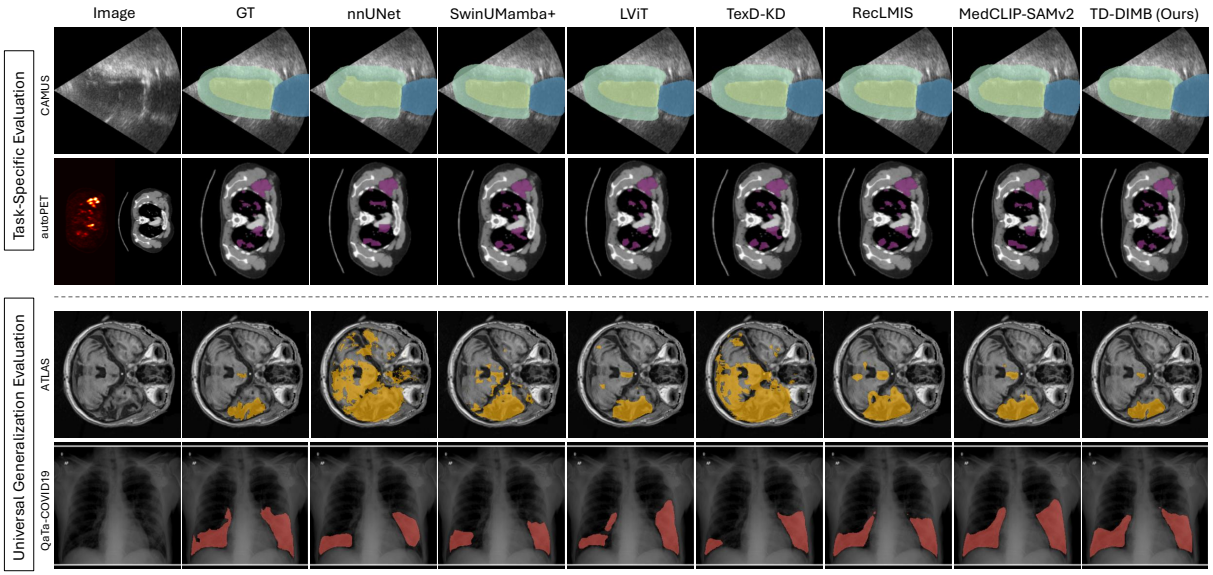

FIGURE 7.4 – Comparaison qualitative des performances de segmentation sur différents ensembles de données.

de référence de l'état de l'art, mettant en évidence des différences cruciales en termes de précision des frontières, de plausibilité anatomique et de résilience aux artefacts propres à chaque modalité.

L'analyse qualitative sur l'échographie cardiaque (CAMUS) montre la capacité de TD-DIMB à générer des masques lisses et anatomiquement cohérents, suivant fidèlement les contours du ventricule gauche tout en supprimant efficacement le bruit *speckle* et les artefacts d'ombre acoustique. Dans autoPET22 (PET/CT), TD-DIMB se distingue par sa précision dans la segmentation de structures tumorales petites et irrégulières, en évitant les fuites de contours et les omissions fréquemment observées dans les approches de référence.

L'évaluation sur ATLAS v2.0 (IRM) met en évidence la sensibilité et la spécificité de TD-DIMB dans la détection de lésions d'AVC petites et diffuses, tout en préservant la cohérence topologique et en maintenant un alignement étroit avec les annotations expertes. Enfin, sur QaTa-COV19 (radiographie thoracique), TD-DIMB génère des masques de lésions COVID cohésifs et pathologiquement consistants, évitant les prédictions fragmentées ou incomplètes typiques des méthodes comparatives.

Ces résultats qualitatifs confirment la capacité exceptionnelle de TD-DIMB à généraliser à travers les modalités d'imagerie et les tâches cliniques, en produisant des segmentations à la fois anatomiquement fidèles et cliniquement interprétables. Les améliorations visuelles constantes dans des contextes variés constituent une preuve solide de l'utilité clinique potentielle de l'approche proposée.

7.4.4 Efficacité de Calcul et Viabilité Clinique

Afin d'évaluer la viabilité pratique du déploiement de TD-DIMB, nous avons conduit une analyse systématique de la complexité de calcul en comparant l'ensemble des méthodes sur des images d'entrée standardisées (256×256) exécutées sur GPU NVIDIA RTX A6000. Le Tableau 7.7 présente une comparaison détaillée de l'efficacité à travers des métriques clés de calcul, essentielles pour évaluer la faisabilité clinique du déploiement.

TABLE 7.7 – Comparaison de complexité de calcul et d'efficacité des méthodes de segmentation sur images d'entrée 256×256 avec entrainement universel.

Méthode	Params (M)	FLOPs (G)	IT (ms)	TT (h)	Mémoire (MB)	FPS
nnUNet [176]	**19.1**	10.87	15.51	22.3	1060	181.5
SwinUMamba† [197]	28.0	18.9	38.5	28.1	**153.6**	**340.03**
LViT [198]	39.9	27.08	18.14	30.4	266.25	122.89
TexD-KD [199]	35.1	**7.8**	19.11	33.2	296.96	109.77
RecLMIS [200]	56.82	23.92	**10.18**	35.8	911.36	98.27
MedCLIP-SAMv2 [196]	251	211.09	125.6	55.6	5529	12.6
TD-DIMB	26.2	34.7	84.99	**31.6**	3584	22.6

TD-DIMB présente une efficacité de calcul compétitive avec 26,2M de paramètres et 34,7G FLOPs, se positionnant favorablement parmi les méthodes contemporaines tout en atteignant une performance de segmentation supérieure. La comparaison avec MedCLIP-SAMv2 est particulièrement éclairante, car TD-DIMB requiert environ 6× moins de FLOPs (34,7G contre 211,09G) tout en offrant une performance de segmentation systématiquement supérieure sur l'ensemble des bases de données évaluées.

L'intégration des caractéristiques issues du modèle fondamental MedSigLIP entraîne un surcoût en calcul, mais apporte des capacités critiques de compréhension du domaine médical qui expliquent les améliorations substantielles observées dans divers scénarios cliniques.

7.4.4.1 Analyse de Signification Statistique et Fiabilité

Nous avons mis en œuvre des protocoles de validation croisée en 4 plis sur les ensembles de données *task-specific*, avec des divisions respectant le niveau patient afin de prévenir toute fuite d'information et d'assurer une évaluation représentative des performances en contexte clinique. Le Tableau 7.8 présente les résultats détaillés de cette validation croisée, montrant des performances constantes à travers les plis et des écarts-types réduits, confirmant la fiabilité et la stabilité du modèle.

Pour évaluer de manière systématique la signification statistique des améliorations de performance, nous avons réalisé des tests t appariés comparant TD-DIMB à l'ensemble des méthodes de référence dans les deux contextes d'évaluation. Le Tableau 7.9 présente une analyse complète de la significativité, montrant que TD-DIMB atteint des améliorations statistiquement significatives par rapport à toutes les approches concurrentes.

TABLE 7.8 – Résultats de validation croisée four-fold pour TD-DIMB sur évaluation task-specific.

Ensemble de Données	**Fold**	**Score Dice (%)** Moyenne ± Écart	**HD95 (pixels)** Moyenne ± Écart
CAMUS	Fold 1	97.89 ± 2.14	18.34 ± 3.2
	Fold 2	97.34 ± 2.38	19.12 ± 3.6
	Fold 3	98.12 ± 2.09	17.89 ± 2.9
	Fold 4	97.45 ± 2.44	18.67 ± 3.4
	Global	**97.70 ± 0.34**	**18.51 ± 0.51**
autoPET22	Fold 1	94.12 ± 3.76	15.89 ± 2.4
	Fold 2	93.56 ± 4.12	16.67 ± 2.8
	Fold 3	94.23 ± 3.89	15.78 ± 2.6
	Fold 4	93.67 ± 3.95	16.23 ± 2.7
	Global	**93.90 ± 0.31**	**16.14 ± 0.39**

TABLE 7.9 – Analyse de signification statistique des améliorations TD-DIMB avec valeurs p et intervalles de confiance 95%.

Comparaison	**Ensemble de Données**	**Amélior. DSC (%)**	**IC 95% (%)**	**valeur p**
Évaluation Task-Specific				
Le nôtre vs nnUNet	CAMUS	+4.77	[4.12, 5.42]	<0.001
	autoPET22	+3.08	[2.31, 3.85]	<0.001
Le nôtre vs SwinUMamba	CAMUS	+4.80	[4.15, 5.45]	<0.001
	autoPET22	+1.76	[1.09, 2.43]	0.003
Le nôtre vs LViT	CAMUS	+3.68	[3.11, 4.25]	<0.001
	autoPET22	+1.89	[1.21, 2.57]	0.002
Le nôtre vs TexD-KD	CAMUS	+3.98	[3.45, 4.51]	<0.001
	autoPET22	+1.56	[0.89, 2.23]	0.004
Le nôtre vs RecLMIS	CAMUS	+2.90	[2.41, 3.39]	<0.001
	autoPET22	+1.14	[0.52, 1.76]	0.012
Le nôtre vs MedCLIP-SAMv2	CAMUS	+0.77	[0.34, 1.20]	0.018
	autoPET22	+0.65	[0.15, 1.15]	0.031
Évaluation Généralisation Universelle				
Le nôtre vs nnUNet	ATLAS	+3.60	[2.89, 4.31]	<0.001
	QaTa-COV19	+10.24	[9.12, 11.36]	<0.001
Le nôtre vs SwinUMamba	ATLAS	+2.90	[2.25, 3.55]	<0.001
	QaTa-COV19	+7.14	[6.18, 8.10]	<0.001
Le nôtre vs LViT	ATLAS	+2.01	[1.41, 2.61]	<0.001
	QaTa-COV19	+6.06	[5.16, 6.96]	<0.001
Le nôtre vs TexD-KD	ATLAS	+1.91	[1.32, 2.50]	<0.001
	QaTa-COV19	+1.78	[0.98, 2.58]	0.007
Le nôtre vs RecLMIS	ATLAS	+1.56	[1.01, 2.11]	0.002
	QaTa-COV19	+3.74	[2.89, 4.59]	<0.001
Le nôtre vs MedCLIP-SAMv2	ATLAS	+0.84	[0.31, 1.37]	0.015
	QaTa-COV19	+1.65	[0.86, 2.44]	0.009

L'analyse de la significativité statistique met en évidence de manière robuste les avantages de performance de TD-DIMB, avec des valeurs de p allant de < 0.001 pour les améliorations substantielles face aux approches de référence les plus faibles, à 0.009–0.031 pour les comparaisons avec la méthode concurrente la plus performante (MedCLIP-SAMv2). Les intervalles de confiance à 95% confirment des gains significatifs, avec des tailles d'effet comprises entre +0.65% et +10.24% sur différents ensembles de données et scénarios d'évaluation.

7.5 Discussion et Analyse Critique

La validation expérimentale exhaustive de TD-DIMB met en évidence des avancées significatives en segmentation d'images médicales entre modalités, établissant de nouveaux standards de performance grâce à l'intégration synergique du *conditioning* sémantique guidé par *prompt*, de la modélisation efficace du contexte à long terme et d'une optimisation informée par les priorités cliniques. Les améliorations constantes observées dans divers scénarios d'évaluation confirment notre hypothèse centrale : les modèles fondamentaux spécialisés pour le domaine médical favorisent une généralisation robuste à travers les modalités d'imagerie, sans nécessiter de modifications architecturales *task-specific*.

Les performances supérieures obtenues découlent de trois innovations clés : l'intégration de MedSigLIP comme modèle fondamental pour le *conditioning* sémantique, l'introduction des modules Dense Inverted Mamba Bottleneck pour un traitement efficace des caractéristiques, et l'emploi de la *Reinforced Gaussian Dice Loss* pour une optimisation sensible aux frontières. Les études d'ablation systématiques confirment la contribution déterminante de chacun de ces composants, avec des gains particulièrement marqués en précision des frontières et en capacités de généralisation entre modalités. Enfin, le protocole d'évaluation à deux volets démontre à la fois l'excellence *task-specific* et le potentiel de généralisation universelle, tandis que la validation statistique rigoureuse apporte une preuve solide de la fiabilité clinique du modèle.

7.5.1 Limitations et Contraintes

Malgré ces résultats prometteurs, plusieurs limitations importantes contraignent encore le cadre actuel. Notre évaluation s'est concentrée principalement sur un traitement basé sur des coupes bidimensionnelles, même pour des ensembles de données intrinsèquement tridimensionnels. Si cette approche facilite le *benchmarking* évolutif sur divers ensembles et modalités d'imagerie, elle abstrait néanmoins l'information de continuité spatiale, essentielle pour une interprétation complète des images médicales. Le traitement de volumes complets ou l'adoption de stratégies 2.5D pourrait améliorer la cohérence des segmentations, en particulier dans des applications nécessitant un contexte spatial, telles que le suivi tumoral ou l'analyse longitudinale.

Les capacités de généralisation universelle demeurent limitées par la proximité entre les domaines d'entraînement et d'évaluation. TD-DIMB obtient des performances remarquables sur des ensembles de test partageant des caractéristiques similaires avec le corpus d'entraînement. Par exemple, ATLAS (segmentation de lésions d'AVC) bénéficie de la présence, dans les données d'apprentissage, d'ensembles cérébraux tels que BraTS (segmentation de gliomes). Bien que les pathologies diffèrent, ces deux tâches impliquent la segmentation de lésions cérébrales avec des modalités et contextes anatomiques comparables. En revanche, les performances se dégradent sur des domaines morphologiquement éloignés, où l'apparence anatomique et les structures de référence divergent fortement de toute représentation apprise lors de l'entraînement.

Enfin, le cadre actuel présente une sensibilité à la précision des *prompts*, ce qui peut limiter sa flexibilité en contexte clinique. Lorsque les *prompts* manquent de spécificité quant aux structures cibles (par exemple : « segmenter les anomalies » ou « segmenter n'importe quoi » sans précision anatomique), le modèle tend à générer des masques binaires globaux englobant toutes les pathologies reconnues, plutôt que de différencier les structures. Ce comportement souligne l'importance de *prompts* anatomiquement précis et suggère qu'un déploiement clinique efficace bénéficierait de l'utilisation de gabarits structurés de *prompts* ou d'une intégration avec des systèmes d'ontologies médicales.

7.5.2 Directions de Recherche Futures

Plusieurs pistes de recherche se dessinent pour dépasser les limitations actuelles et renforcer l'applicabilité clinique. La priorité la plus immédiate est d'étendre TD-DIMB au traitement volumétrique tridimensionnel complet, afin de préserver la continuité spatiale indispensable à de nombreuses applications médicales. Une telle extension implique toutefois de relever les défis liés à l'augmentation des besoins en calcul associés à la modélisation *state-space* 3D, tout en maintenant les avantages d'efficacité déjà démontrés.

Le renforcement de la robustesse inter-domaines constitue une autre orientation essentielle, nécessitant l'intégration d'une plus grande diversité de domaines dans les ensembles d'entraînement et le développement de mécanismes adaptatifs pour traiter les applications d'imagerie morphologiquement éloignées. De plus, l'ajout de capacités de sortie multi-classes, permettant la segmentation simultanée de plusieurs structures anatomiques, améliorerait considérablement l'intégration dans les flux de travail cliniques. Enfin, l'exploration de mécanismes de quantification de l'incertitude pour les systèmes multi-*modaux* représenterait une avancée cruciale pour la sécurité clinique, en fournissant des estimations de confiance indispensables au soutien à la décision médicale.

7.6 Résumé du Chapitre et Conclusion

Ce chapitre a présenté le *Text-Driven Dense Inverted Mamba Bottleneck Network* (TD-DIMB), un framework novateur de segmentation d'images médicales entre modalités, conçu pour surmonter des limitations fondamentales des approches existantes tout

en établissant de nouveaux standards en matière d'efficacité de calcul et de capacités de généralisation universelle. TD-DIMB intègre des modules *Dense Inverted Mamba Bottleneck* qui associent la modélisation *state-space* à une connectivité dense, permettant de capturer efficacement à la fois les caractéristiques spatiales locales et les dépendances sémantiques globales. Le mécanisme *Text-Driven Selective Scan 2D* fournit un alignement sémantique à complexité linéaire entre *prompts* textuels et caractéristiques visuelles, autorisant un déploiement pratique sur des images médicales haute résolution. Parallèlement, la *Reinforced Gaussian Dice Loss* améliore significativement la précision des frontières grâce à une modélisation lissée par noyau gaussien et une pondération des erreurs cliniques adaptative.

L'évaluation exhaustive démontre la capacité supérieure de TD-DIMB à exploiter conjointement la guidance sémantique textuelle et visuelle à travers différentes modalités d'imagerie, dans un protocole d'évaluation dual confirmant à la fois l'excellence des performances *task-specific* et la robustesse de la généralisation universelle. La validation statistique apporte une preuve solide de la fiabilité clinique, le framework atteignant des résultats de l'état de l'art tout en conservant une efficacité de calcul compatible avec un déploiement pratique. Ces résultats montrent que les modèles *state-space* à complexité linéaire peuvent atteindre des performances traditionnellement associées aux architectures *transformer*, souvent coûteuses en calcul, confirmant ainsi l'importance d'intégrer des modèles fondamentaux médicaux spécialisés pour une compréhension robuste entre modalités.

En conclusion, TD-DIMB établit la segmentation guidée par *prompt* comme un paradigme viable pour les systèmes d'IA médicaux de prochaine génération, offrant une base solide pour des solutions évolutives, interprétables et cliniquement adaptables, comblant ainsi l'écart entre les avancées de recherche et les exigences pratiques du déploiement en santé.

Chapitre 8

FUSE-RAG : Few-shot Universal Segmentation avec Retrieval-Augmented Generation pour l'Imagerie Médicale

8.1 Introduction

8.1.1 Motivation et Contexte de Recherche

S'appuyant sur la progression méthodique établie au fil de nos travaux précédents, ce chapitre représente l'aboutissement de notre parcours de recherche vers la segmentation d'images médicales universelle. Alors que nos études antérieures ont démontré la gestion efficace de la variabilité intra-dataset grâce à des architectures spécialisées (MEDiXNet, MixLVMM), l'adaptabilité entre modalités à travers des frameworks tridimensionnels (HA-U^{3}Net), ainsi que le guidage sémantique basé sur des approches à *prompt* (TD-DIMB), le défi ultime de l'analyse d'images médicales demeure le développement de capacités de segmentation véritablement universelles, capables de s'adapter à de nouvelles structures anatomiques, pathologies et modalités d'imagerie, avec une supervision minimale.

La segmentation d'images médicales constitue un pilier essentiel du diagnostic assisté par ordinateur, soutenant des applications cliniques majeures telles que la délimitation de structures anatomiques, l'identification de régions pathologiques et la planification de traitements sur diverses modalités d'imagerie. Une segmentation précise favorise des tâches en aval critiques, notamment la navigation chirurgicale, la planification de radiothérapie et le suivi longitudinal de l'évolution des maladies, influençant directement la prise de décision clinique et les résultats pour les patients. Par exemple, une segmentation exacte des structures cardiaques est cruciale pour évaluer la fonction ventriculaire gauche et planifier des procédures interventionnelles, tandis que la délimitation fine des

frontières tumorales demeure indispensable pour la planification de dose en radiothérapie et le guidage des interventions chirurgicales.

Les avancées récentes en matière de modèles fondamentaux et d'architectures, notamment les *Vision Transformers* et les *State Space Models*, ont considérablement élargi les capacités d'analyse des images médicales. Toutefois, leur application à la segmentation médicale en contexte *few-shot* reste encore peu explorée, en particulier pour le développement de mécanismes de récupération capables d'exploiter la richesse représentationnelle des modèles vision-langage médicaux. Bien que le *Segment Anything Model* (SAM) et ses variantes adaptées au domaine médical aient montré des résultats prometteurs, ils demeurent dépourvus de mécanismes de récupération anatomiquement informés, pourtant essentiels à un apprentissage *few-shot* efficace dans les contextes où la compréhension anatomique prévaut sur la simple similarité visuelle.

8.1.2 Le Défi de la Segmentation Universelle

En pratique clinique, les radiologues sont quotidiennement confrontés à une grande diversité de scénarios d'imagerie, tels que l'évaluation d'un AVC nécessitant la segmentation rapide de lésions cérébrales, l'analyse oncologique demandant une délimitation précise des frontières tumorales à travers plusieurs modalités d'imagerie, ou encore l'évaluation cardiaque impliquant une quantification rigoureuse des cavités à partir de différents protocoles de scanner. Chaque situation présente des structures anatomiques uniques, des manifestations pathologiques spécifiques et des caractéristiques d'imagerie particulières que les modèles actuels peinent à gérer sans un réentraînement intensif.

Dans un contexte de déploiement clinique, les contraintes pratiques sont évidentes : un service de radiologie ne peut raisonnablement maintenir une multitude de modèles de segmentation spécialisés, chacun nécessitant ses propres données d'apprentissage et ressources de calcul. Lorsqu'un nouveau protocole d'imagerie est introduit, qu'une pathologie rare est rencontrée ou qu'un scanner différent est utilisé, les approches existantes exigent la collecte de nouvelles données, leur annotation par des experts et un réentraînement complet du modèle, un processus long, coûteux et peu scalable dans un environnement clinique.

Les approches de segmentation en contexte *few-shot* présentent une limite conceptuelle majeure dans le domaine médical. Alors que les méthodes de vision par ordinateur reposent sur la similarité visuelle globale, en supposant que les images « semblables » partagent des informations pertinentes, l'analyse d'images médicales requiert une compréhension anatomique approfondie. Par exemple, une IRM cardiaque et une IRM cérébrale peuvent paraître similaires visuellement (toutes deux en niveaux de gris, avec des contrastes comparables), mais elles représentent des structures anatomiques totalement différentes nécessitant des stratégies de segmentation distinctes. À l'inverse, deux scanners pulmonaires peuvent sembler dissemblables en raison de variations pathologiques ou de différences d'acquisition, tout en partageant des repères anatomiques essentiels qui permettent le transfert de connaissances. Cette inadéquation entre la similarité visuelle

globale et la pertinence anatomique constitue une limitation majeure que la littérature actuelle n'a pas encore su résoudre de manière satisfaisante.

8.1.3 Retrieval-Augmented Generation pour l'Imagerie Médicale

Le *retrieval-augmented generation* (RAG) incarne le principe de requête dynamique de connaissances externes durant l'inférence, offrant un paradigme prometteur pour la segmentation d'images médicales. Tandis que cette approche a transformé le traitement du langage naturel en permettant aux modèles d'accéder à l'information contextuelle pertinente au-delà de leurs données d'entraînement, son application à l'imagerie médicale demeure largement inexplorée. La lacune critique réside dans l'adaptation des principes RAG aux domaines visuo-anatomiques, où récupérer des connaissances textuelles doit être remplacé par une correspondance visuelle anatomiquement informée. Ceci nécessite de nouveaux mécanismes pour identifier et tirer parti de connaissances anatomiques expertes sans les stocker directement dans les paramètres du modèle, permettant l'accès dynamique à des exemplaires médicaux pertinents durant l'inférence.

8.1.4 Objectifs de Recherche et Contributions

Pour adresser ces limitations fondamentales et compléter notre trajectoire de recherche vers la segmentation d'images médicales universelle, nous introduisons FUSE-RAG (*Few-shot Universal Segmentation avec Retrieval-Augmented Generation*), un nouveau framework qui représente la première approche *retrieval-augmented generation* incorporant des connaissances anatomiques *ROI-aware* pour la segmentation d'images médicales. Notre approche pionnière intègre systématiquement l'expertise anatomique dans les principes RAG en maintenant une base de connaissances dynamique d'exemples médicaux annotés et en récupérant des exemplaires anatomiquement pertinents à travers une correspondance de similarité *ROI-aware*.

Ce mécanisme intègre des connaissances anatomiques directement dans les représentations de modèles fondamentaux durant l'encodage, plutôt qu'à travers des ajustements de similarité post-traitement. Ceci permet au système de se concentrer sur des structures anatomiques spécifiques présentes tant dans les images de requête que de support à travers l'amélioration de caractéristiques guidées par attention, dépassant les caractéristiques d'images globales pour identifier des exemplaires structurellement pertinents. FUSE-RAG opère à travers deux phases majeures : premièrement, le processus de récupération *ROI-aware* crée un ensemble de support pertinent d'exemples anatomiquement similaires qui servent comme *prompts* visuels pour le réseau de segmentation ; deuxièmement, une architecture de segmentation novatrice traite l'image de requête conditionnée sur ces exemplaires anatomiques récupérés pour générer des masques de segmentation précis.

L'architecture de segmentation tire parti de trois innovations techniques clés : les *Anatomical Correspondence Blocks* (ACB), qui utilisent les modules *Selective Scan 2D* (SS2D) de *Vision Mamba* pour une modélisation efficace des dépendances à long terme avec une complexité de calcul linéaire, permettant la capture tant de détails pathologiques *fine-grained* que de contexte anatomique global tout en adressant les limitations de passage à l'échelle quadratique des mécanismes d'attention traditionnels ; les *Support Quality Assessment Blocks* (SQAB), qui pondèrent intelligemment les exemples récupérés selon leur pertinence anatomique ; et les *Support-Conditioned Skip Connections* (SCSC), qui propagent le guidage anatomique à travers la voie du décodeur.

8.1.5 Innovations Clés et Impact de Recherche

Notre mécanisme de récupération *ROI-aware* pionnier, qui intègre des connaissances anatomiques expertes à partir des représentations de modèles fondamentaux, permet à FUSE-RAG d'atteindre des améliorations de performance substantielles de 10,26% et 8,86% en coefficient Dice sur la segmentation de lésions d'AVC et de pneumonie, respectivement, établissant une nouvelle référence de l'état de l'art en segmentation d'images médicales *few-shot*, tout en étendant les principes d'ingénierie de *prompts* issus du traitement du langage naturel vers l'imagerie médicale, selon un paradigme fondé sur la qualité plutôt que la quantité.

Le framework démontre que des exemples anatomiquement pertinents et soigneusement sélectionnés surpassent significativement des ensembles plus larges d'images de support choisies aléatoirement, offrant une approche efficace pour tirer parti d'annotations expertes limitées dans des contextes de déploiement clinique. Les capacités de généralisation entre modalités, validées à travers l'IRM et la radiographie thoracique, établissent l'applicabilité universelle des mécanismes de récupération anatomiquement informés, répondant ainsi à des obstacles majeurs à l'adoption de l'IA dans divers environnements cliniques.

FUSE-RAG représente l'aboutissement de notre progression de recherche systématique, atteignant les capacités de segmentation universelle que les contributions précédentes avaient progressivement approchées à travers des innovations architecturales de plus en plus sophistiquées et des stratégies d'intégration entre modalités. Cette contribution finale établit une base solide pour les systèmes d'IA médicale de prochaine génération, alliant performance robuste et précision anatomique requise pour les applications cliniques, et ouvrant de nouvelles perspectives pour le développement de systèmes de segmentation d'images médicales universels capables de s'adapter efficacement à des scénarios cliniques nouveaux avec une supervision minimale.

8.2 Méthodologie

8.2.1 Formulation du Problème et Vue d'Ensemble du Framework

Soit $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}$ une collection d'entraînement diversifiée englobant de multiples structures anatomiques, pathologies et modalités d'imagerie, où $\mathbf{x}_i \in \mathbb{R}^{H\times W\times C}$ représente une image médicale et $\mathbf{y}_i \in \{0,1\}^{H\times W}$ son masque de segmentation correspondant. H et W désignent respectivement la hauteur et la largeur de l'image, $C = 1$ correspond à la dimension du canal en niveaux de gris, et M est le nombre total d'exemples d'entraînement. Durant l'inférence, étant donnée une image de requête $\mathbf{x}_q$ d'une tâche totalement inédite et un ensemble de support $\mathcal{S} = \{(\mathbf{x}_j^s, \mathbf{y}_j^s)\}_{j=1}^{N}$ contenant tous les exemples annotés disponibles pour cette nouvelle tâche, notre objectif est d'apprendre une fonction de segmentation universelle $f_\theta : (\mathbf{x}_q, \mathcal{S}') \to \hat{\mathbf{y}}_q$ capable de segmenter précisément l'image de requête en tirant parti d'un sous-ensemble stratégiquement sélectionné $\mathcal{S}' \subset \mathcal{S}$, composé des exemples de support les plus pertinents du pool. Ici, $\mathbf{x}_j^s$ et $\mathbf{y}_j^s$ désignent respectivement la j-ième image et son masque de support, N est le nombre total d'exemples de support disponibles, θ représente les paramètres du modèle apprenables, et $\hat{\mathbf{y}}_q$ correspond au masque de segmentation prédit pour l'image de requête.

L'idée clé sous-jacente à FUSE-RAG est que la qualité de la sélection de l'ensemble de support détermine de manière critique la performance de la segmentation *few-shot*. Alors que des approches existantes comme UniverSeg [201] reposent sur une sélection aléatoire de l'ensemble de support, cette stratégie échoue souvent à identifier les exemples anatomiquement les plus pertinents susceptibles d'orienter de manière optimale la segmentation de l'image de requête. Notre approche surmonte cette limitation en introduisant un mécanisme de sélection de support intelligent, capable d'identifier les exemples les plus similaires à l'image de requête, fournissant ainsi un guidage anatomique plus efficace pour la segmentation.

Comme illustré en Figure 8.1, FUSE-RAG fonctionne selon un pipeline en deux étapes, maximisant l'utilité des informations issues de l'ensemble de support. La première étape utilise un mécanisme de récupération *ROI-aware* avancé, qui analyse l'image de requête et sélectionne intelligemment les exemples de support anatomiquement les plus pertinents $\mathcal{S}'$ au sein de l'ensemble $\mathcal{S}$. La seconde étape s'appuie sur une architecture de segmentation novatrice, qui traite l'image de requête conditionnée par les exemples de support récupérés, en exploitant des mécanismes d'attention et une fusion de caractéristiques adaptative pour intégrer efficacement les connaissances anatomiques extraites à partir de l'ensemble de support sélectionné.

Le mécanisme de récupération *ROI-aware* intègre des connaissances anatomiques expertes à partir des représentations de modèles fondamentaux médicaux durant l'encodage, plutôt qu'à travers des ajustements de similarité en post-traitement. Ceci permet au système de se concentrer sur des structures anatomiques spécifiques présentes à la fois dans les images de requête et de support, en renforçant les caractéristiques guidées

par attention, au-delà des informations globales, pour identifier des exemples structurellement pertinents. La formulation mathématique de FUSE-RAG peut être exprimée comme :

$$\hat{\mathbf{y}}_q = f_\theta(\mathbf{x}_q, \mathcal{S}'), \quad \mathcal{S}' = \mathcal{R}(\mathbf{x}_q, \mathcal{S}) \tag{8.1}$$

où $\mathcal{R}(\mathbf{x}_q, \mathcal{S})$ représente la fonction de récupération *ROI-aware* qui sélectionne les exemples de support anatomiquement les plus pertinents à travers des représentations de caractéristiques enrichies par attention, et f_θ désigne la fonction de segmentation générant le masque de requête. Cette formulation permet au modèle d'exploiter les connaissances anatomiques provenant d'exemples de support sélectionnés de manière stratégique, atteignant ainsi une performance de segmentation supérieure sur des tâches totalement inédites, tout en conservant une efficacité de calcul adaptée au déploiement clinique.

8.2.2 Mécanisme de Récupération ROI-Aware

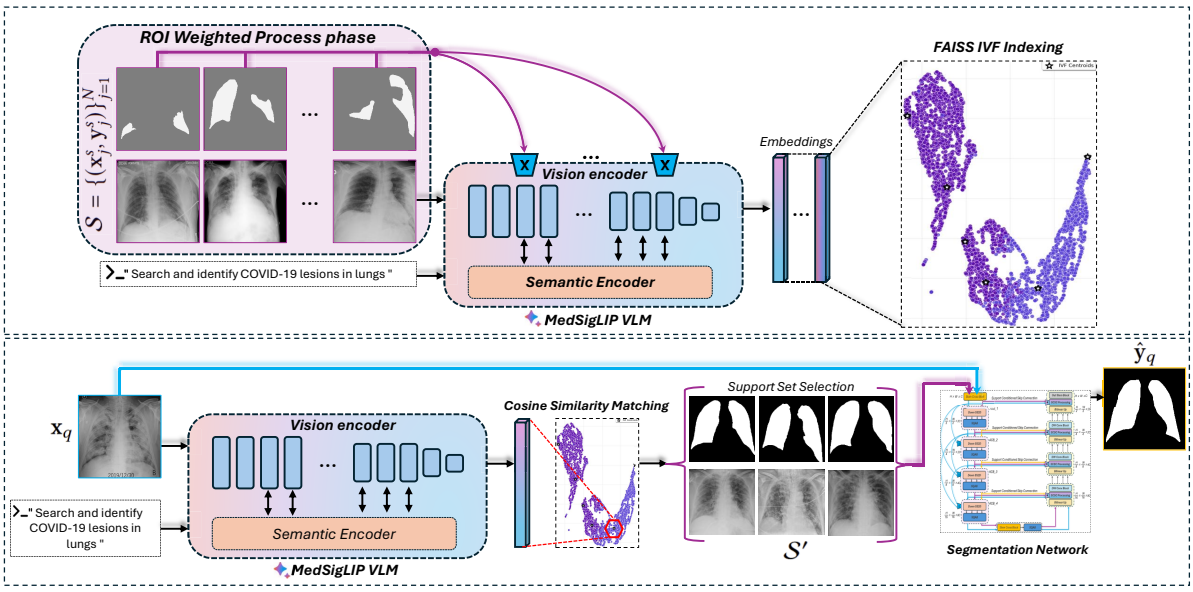

FIGURE 8.1 – Conception du mécanisme de récupération *ROI-aware*. Haut : le processus d'indexation des données crée la base de connaissances à travers l'extraction d'*embeddings* enrichis par les régions d'intérêt (ROI) et leur indexation via FAISS. Bas : le processus d'inférence emploie le *retrieval-augmented generation* pour améliorer la segmentation en interrogeant la base de connaissances et en sélectionnant des exemples de support anatomiquement pertinents pour le réseau de segmentation.

8.2.2.1 Extraction d'*Embeddings* Spécifiques au Domaine Médical

Nous employons MedSigLIP [181], un modèle vision-langage du domaine médical récent introduit par l'équipe Google Research, spécifiquement conçu pour la récupération d'images sémantiques dans les applications de santé, afin d'extraire des *embeddings* sémantiquement riches à la fois pour les images de requête et de support. MedSigLIP a été préentraîné sur diverses modalités d'imagerie médicale, incluant les radiographies thoraciques, les volumes CT/IRM, les images dermatologiques et les lames histopathologiques, ce qui en fait le modèle fondamental idéal pour la compréhension d'images médicales entre modalités.

Notre approche introduit une extraction d'*embedding* asymétrique, où les images de support bénéficient d'une attention guidée par ROI, tandis que les images de requête utilisent un traitement standard. Étant donnée une image de support $\mathbf{x}_i^s \in \mathbb{R}^{H\times W\times C}$ avec son masque expert correspondant $\mathbf{y}_i^s \in \{0,1\}^{H\times W}$ et un *prompt* textuel médical $\mathbf{t}$, nous générons des *embeddings* enrichis par ROI à travers une modification du mécanisme d'attention du *transformer* :

$$\mathbf{e}_i^s = \text{MedSigLIP}_{\text{ROI}}(\mathbf{x}_i^s, \mathbf{y}_i^s, \mathbf{t}), \quad \mathbf{e}_q = \text{MedSigLIP}(\mathbf{x}_q, \mathbf{t}) \tag{8.2}$$

où $\mathbf{e}_i^s \in \mathbb{R}^d$ représente l'*embedding* de support *ROI-aware*, $\mathbf{e}_q \in \mathbb{R}^d$ désigne l'*embedding* de requête standard, $d = 1152$ correspond à la dimension des *embeddings*, $\mathbf{t}$ est le *prompt* textuel spécifique à la tâche (ex. : « segmenter le ventricule gauche cardiaque » ou « identifier le nodule pulmonaire »), $\text{MedSigLIP}_{\text{ROI}}(\cdot)$ désigne le modèle MedSigLIP enrichi par ROI, et $\text{MedSigLIP}(\cdot)$ le modèle standard. Le processus d'*embedding* enrichi par ROI incorpore l'injection d'attention spatiale à travers plusieurs couches du *transformer*, produisant des représentations de caractéristiques anatomiquement focalisées qui priorisent les régions annotées par l'expert.

8.2.2.2 Injection d'Attention ROI au Niveau du Transformer

S'appuyant sur le cadre d'*embedding* asymétrique, notre innovation principale réside dans la modification du passage avant du *vision transformer* de MedSigLIP afin d'injecter l'attention *ROI-aware* durant l'encodage. Le masque spatial $\mathbf{y}_i^s$ est d'abord converti en une représentation au niveau des *patches*, compatible avec la structure séquentielle du *transformer*. Étant donnée la taille de *patch* de MedSigLIP de 14×14 pixels pour des images d'entrée 448×448, nous obtenons une grille de 32×32 = 1024 *patches*, et convertissons le masque binaire en poids d'attention au niveau des *patches* :

$$\mathbf{y}_i^{patch} = \mathcal{P}(\mathbf{y}_i^s), \quad \mathbf{w}_i^{attn} = \alpha_{\text{ROI}} \cdot \mathbf{y}_i^{patch} + \alpha_{\text{bg}} \cdot (1 - \mathbf{y}_i^{patch}) \tag{8.3}$$

où $\mathbf{y}_i^{patch} \in [0,1]^P$ représente le masque au niveau des *patches* avec $P = 1024$, $\mathcal{P}(\cdot)$ désigne l'opérateur de conversion en *patch* qui effectue le sous-échantillonnage spatial par *average pooling* à l'intérieur de chaque région de 14×14 pixels, $\mathbf{w}_i^{attn} \in \mathbb{R}^P$ représente les poids d'attention, $\alpha_{\text{ROI}} = 0.8$ est le poids attribué aux régions d'intérêt

(ROI) et $\alpha_{\text{bg}} = 0.2$ celui des régions de fond. Cette formulation garantit que les régions ROI reçoivent 80 % d'importance, tandis que les régions de fond conservent 20 % de contribution, préservant ainsi le contexte global essentiel à un apprentissage robuste des caractéristiques.

L'injection d'attention ROI est appliquée à plusieurs couches du *transformer* $\mathcal{L}_{inj} = \{6, 9, 12, 15, 18\}$ durant le passage avant, selon la modification suivante :

$$\begin{aligned} \mathbf{H}^{(l)} &= \mathcal{T}^{(l)}(\mathbf{H}^{(l-1)}) \\ \mathbf{h}_{\text{cls}}^{(l)} &= \mathbf{H}^{(l)}[:, 0, :], \quad \mathbf{h}_{\text{patch}}^{(l)} = \mathbf{H}^{(l)}[:, 1:, :] \\ \tilde{\mathbf{h}}_{\text{patch}}^{(l)} &= \mathbf{h}_{\text{patch}}^{(l)} \odot \mathbf{w}_i^{attn}, \quad \mathbf{H}^{(l)} = [\mathbf{h}_{\text{cls}}^{(l)}, \tilde{\mathbf{h}}_{\text{patch}}^{(l)}], \quad l \in \mathcal{L}_{inj} \end{aligned} \tag{8.4}$$

où $\mathbf{H}^{(l)} \in \mathbb{R}^{B \times (P+1) \times d}$ représente les états cachés à la couche l avec une longueur de séquence totale $(P + 1) = 1025$, incluant le *token* CLS, $\mathcal{T}^{(l)}(\cdot)$ désigne l'opération de la l-ième couche du *transformer*, $\mathbf{h}_{\text{cls}}^{(l)} \in \mathbb{R}^{B \times 1 \times d}$ et $\mathbf{h}_{\text{patch}}^{(l)} \in \mathbb{R}^{B \times P \times d}$ représentent respectivement les caractéristiques du *token* CLS et des *patches*, $\tilde{\mathbf{h}}_{\text{patch}}^{(l)}$ désigne les caractéristiques pondérées par attention, $\odot$ représente la multiplication élément par élément avec diffusion le long de la dimension des caractéristiques, B est la taille du lot, et $[\cdot, \cdot]$ indique la concaténation le long de la dimension de séquence. L'injection multi-couche garantit que l'information ROI influence les représentations à plusieurs niveaux d'abstraction, des caractéristiques sémantiques intermédiaires (couches 6–12) jusqu'aux concepts anatomiques de haut niveau (couches 15–18).

8.2.2.3 Matching de Similarité et Vote de Support

Les *embeddings* de support pondérés par ROI permettent un calcul direct de similarité sans étape de post-traitement. Puisque la pertinence anatomique est déjà intégrée au sein des représentations de support elles-mêmes, la similarité est calculée à l'aide de la mesure de similarité cosinus :

$$s_i = \frac{\mathbf{e}_q^T \mathbf{e}_i^s}{||\mathbf{e}_q||_2 \cdot ||\mathbf{e}_i^s||_2} \tag{8.5}$$

où $s_i \in [-1, 1]$ est le score de similarité pour le i-ième exemple de support, $\mathbf{e}_q^T$ désigne la transposée de l'*embedding* de requête, et $|| \cdot ||_2$ représente la norme L2. Cette formulation garantit que les exemples de support anatomiquement pertinents obtiennent naturellement des scores de similarité plus élevés grâce à leurs représentations enrichies par attention. L'approche gère intrinsèquement la variabilité de taille des régions d'intérêt (ROI) à travers la pondération attentionnelle, évitant ainsi le biais lié à la taille des organes ou pathologies.

Sur la base des scores de similarité calculés, le calcul ROI pondéré maintient une efficacité de calcul élevée grâce à l'indexation FAISS, tout en ajoutant un surcoût minimal.

Les *embeddings* de support $\{\mathbf{e}_i^s\}_{i=1}^N$ sont préalablement calculés en utilisant l'injection d'attention ROI, puis indexés via la méthode standard d'*Inverted File* (IVF) pour une récupération efficace. Lors du traitement d'une requête, une recherche de similarité cosinus standard est effectuée, exploitant l'information ROI préalablement intégrée sans nécessiter d'étapes de calcul supplémentaires :

$$\mathcal{S}' = \{(\mathbf{x}_j^s, \mathbf{y}_j^s) : j \in \underset{K}{\text{argmax}}(s_1, s_2, \ldots, s_N)\} \tag{8.6}$$

où $\mathcal{S}'$ est l'ensemble de support sélectionné contenant les exemples anatomiquement les plus pertinents, $\mathbf{x}_j^s$ et $\mathbf{y}_j^s$ sont respectivement la j-ième image et le masque de support, j représente l'indice d'exemple de support, $K = 5$ est le nombre d'exemples à récupérer, et $\underset{K}{\text{argmax}}(\cdot)$ sélectionne les indices des K scores de similarité les plus élevés parmi $\{s_1, s_2, \ldots, s_N\}$. L'injection d'attention ROI est réalisée uniquement lors de l'indexation de l'ensemble de support, tandis que le traitement de la requête utilise l'extraction d'*embedding* MedSigLIP standard, garantissant une correspondance anatomique supérieure et une performance améliorée en segmentation *few-shot*.

8.2.3 Réseau de Segmentation Retrieval-Conditioned

Le réseau de segmentation *Retrieval-Conditioned* traite les exemples de support récupérés $\mathcal{S}'$ pour générer des masques de segmentation précis à partir des images de requête. L'architecture repose sur une U-Net symétrique intégrant des composants optimisés pour l'interaction *cross-modale* entre l'information de requête et celle du support enrichi par ROI, tout en maintenant une efficacité de calcul élevée.

8.2.3.1 Description d'Architecture Globale

Le réseau adopte une conception U-Net symétrique comprenant quatre niveaux d'encodeur et quatre niveaux de décodeur correspondants, assurant un traitement des caractéristiques multi-échelles complet, essentiel à une segmentation d'images médicales précise. Cette structure facilite la modélisation des structures anatomiques à travers différentes résolutions spatiales, allant des détails de frontières fins à haute résolution jusqu'à l'information contextuelle globale aux échelles plus grossières.

La voie d'encodeur traite les caractéristiques de requête à travers des représentations progressivement sous-échantillonnées, chaque niveau capturant des concepts anatomiques de plus en plus abstraits. À chaque étage de l'encodeur, les exemples de support récupérés sont simultanément traités et intégrés à l'aide de mécanismes d'interaction *cross-modale*. L'architecture utilise un sous-échantillonnage apprenable à l'aide de convolutions *depthwise separable* avec *stride*, optimisant ainsi l'efficacité de calcul tout en préservant les détails anatomiques fins et la précision de segmentation, essentiels pour les applications cliniques.

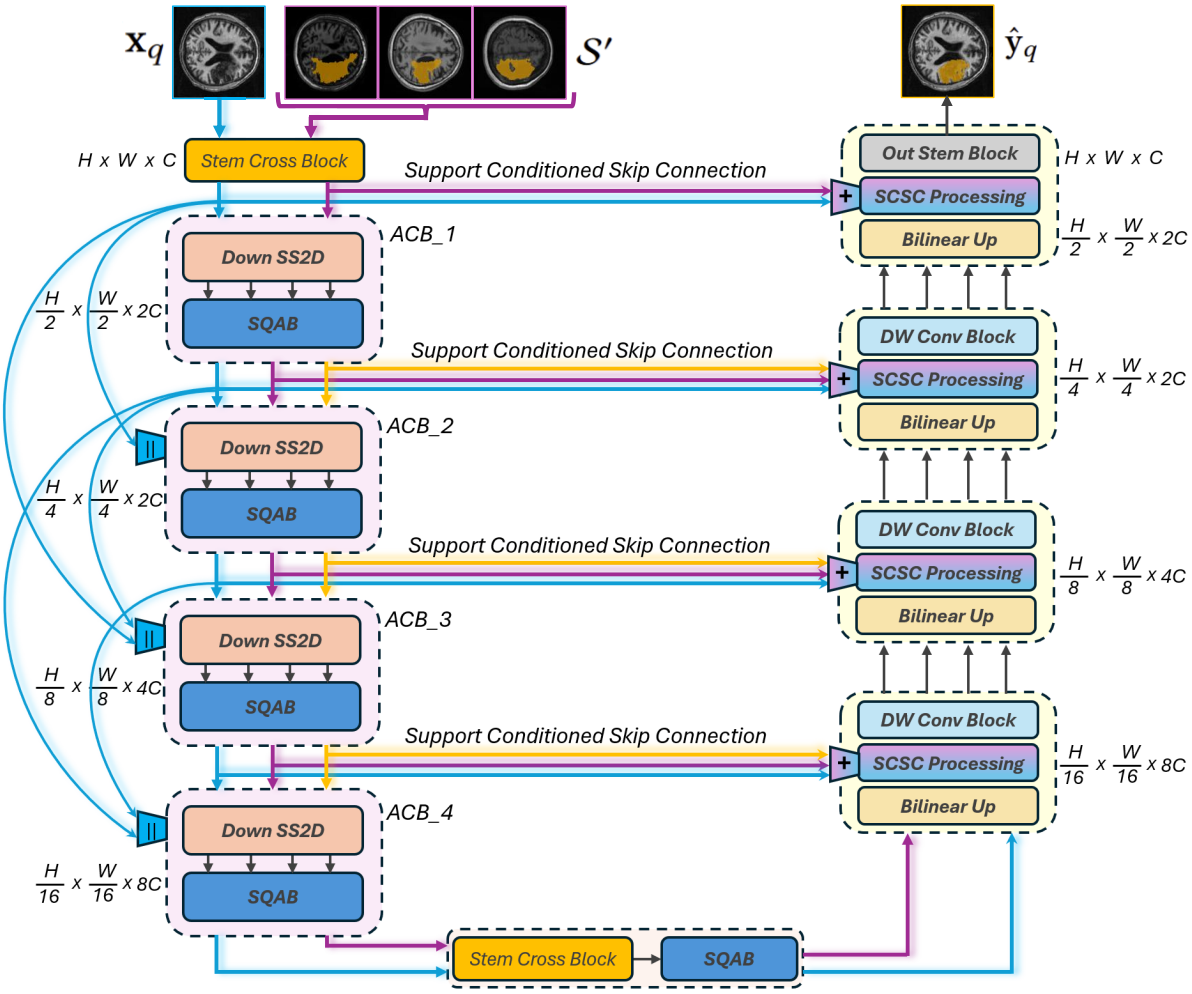

FIGURE 8.2 – Conception architecturale du réseau de segmentation FUSE-RAG.

Le traitement des caractéristiques commence par l'extraction initiale appliquée à la fois aux entrées de requête et de support, au moyen de blocs *stem convolutionnels* identiques. L'image de requête est traitée directement, tandis que chaque exemple de support subit une concaténation par canaux avec son masque correspondant avant traitement :

$$\mathbf{F}_q^{(0)} = \Psi(\mathbf{x}_q), \quad \mathbf{F}_s^{(0)} = \{\Psi([\mathbf{x}_i^s, \mathbf{y}_i^s])\}_{i=1}^K \tag{8.7}$$

où $\mathbf{F}_q^{(0)} \in \mathbb{R}^{B\times C_0\times H\times W}$ représente les caractéristiques initiales de la requête, $\mathbf{F}_s^{(0)}$ désigne la pile de caractéristiques initiales de support, $\Psi(\cdot)$ représente le bloc *stem convolutionnel* comprenant une convolution, une activation SiLU, une normalisation en lot, et des connexions résiduelles appliquées de manière identique aux deux voies. $\mathbf{x}_i^s$ et $\mathbf{y}_i^s$ désignent respectivement la i-ième image et son masque de support (sélectionnés via la récupération *ROI-aware*), B correspond à la taille du lot, $C_0 = 32$ à la dimension initiale des caractéristiques, et $[\cdot,\cdot]$ à la concaténation par canaux. Les exemples de support dans $\mathcal{S}'$ sont spécifiquement choisis en fonction de leur pertinence anatomique grâce au mécanisme de récupération *ROI-aware*, assurant ainsi un guidage anatomique optimal pour le réseau de segmentation.

8.2.3.2 Bloc de Correspondance Anatomique

Le Bloc de Correspondance Anatomique (ACB) constitue le bloc de construction fondamental de la voie d'encodeur, intégrant trois composants de traitement essentiels : la modélisation des dépendances à long terme via les modules SS2D, la pondération intelligente des exemples de support au moyen des *Support Quality Assessment Blocks* (SQAB), et l'intégration de caractéristiques *cross-modale*.

Étant données les caractéristiques de requête d'entrée $\mathbf{F}_q^{(l)}$ et les caractéristiques de support $\mathbf{F}_s^{(l)}$ au niveau d'encodeur l, l'ACB applique des opérations de projection suivies d'un traitement SS2D pour une capture efficace des dépendances anatomiques globales :

$$\mathbf{G}_q^{(l)} = \mathrm{SS2D}_q(\mathcal{W}_q^{(l)}\mathbf{F}_q^{(l)}), \quad \mathbf{G}_{s,i}^{(l)} = \mathrm{SS2D}_s(\mathcal{W}_s^{(l)}\mathbf{F}_{s,i}^{(l)}), \quad i = 1,\ldots,K \tag{8.8}$$

où $\mathbf{G}_q^{(l)}$ représente les caractéristiques de requête améliorées au niveau l, $\mathbf{G}_{s,i}^{(l)}$ désigne les caractéristiques améliorées pour le i-ième exemple de support au même niveau, $\mathbf{F}_q^{(l)}$ et $\mathbf{F}_{s,i}^{(l)}$ sont respectivement les caractéristiques d'entrée de requête et de support, $\mathcal{W}_q^{(l)}, \mathcal{W}_s^{(l)} : \mathbb{R}^{B\times C_l\times H_l\times W_l} \to \mathbb{R}^{B\times C_l\times H_l\times W_l}$ sont des opérateurs de projection apprenables implémentés par convolutions *depthwise separable*, $\mathrm{SS2D}_q$ et $\mathrm{SS2D}_s$ désignent les opérateurs *Selective Scan 2D*, C_l est le nombre de canaux au niveau l, et H_l, W_l sont les dimensions de la carte de caractéristiques à ce même niveau. À la suite du traitement SS2D, les caractéristiques améliorées sont transmises au *Support Quality Assessment Block* pour la pondération et la fusion adaptatives.

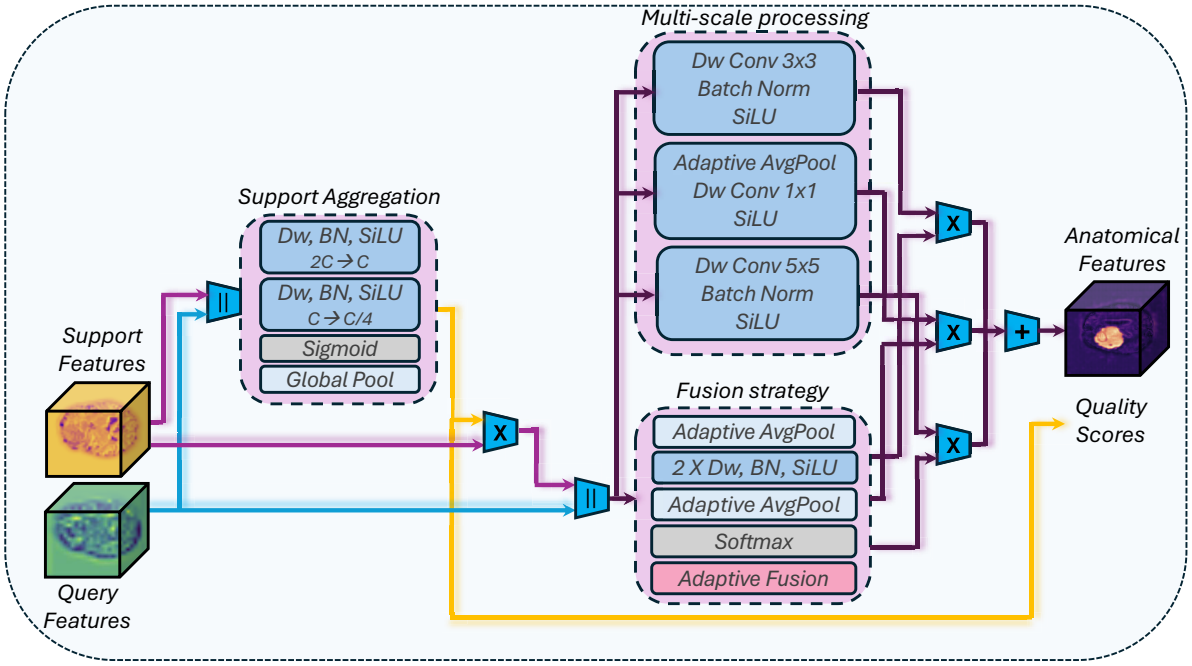

FIGURE 8.3 – Architecture du Bloc d'Évaluation de la Qualité du Support (SQAB).

8.2.3.3 Bloc d'Évaluation de la Qualité du Support

Le Bloc d'Évaluation de la Qualité du Support (SQAB) effectue à la fois la pondération unifiée des exemples de support et la fusion de caractéristiques adaptative au contenu, cette dernière se référant à la capacité du module à ajuster sa stratégie de fusion en fonction des propriétés spécifiques des paires de caractéristiques requête–support. Le SQAB calcule d'abord les scores de qualité à travers une analyse de similarité *cross-modale*, puis procède à une agrégation pondérée par la qualité et à une fusion adaptative :

$$
\begin{aligned}
q_i^{(l)} &= \sigma(\mathrm{MLP}([\mathrm{GAP}(\mathbf{G}_q^{(l)}), \mathrm{GAP}(\mathbf{G}_{s,i}^{(l)})])), \quad \mathbf{G}_s^{(l)} = \sum_{i=1}^{K} q_i^{(l)} \cdot \mathbf{G}_{s,i}^{(l)} \\
\boldsymbol{\alpha}^{(l)} &= \mathrm{Softmax}(\mathcal{W}_a([\mathbf{G}_q^{(l)}, \mathbf{G}_s^{(l)}])), \quad \mathbf{V}^{(l)} = \boldsymbol{\alpha}^{(l)} \odot \mathbf{G}_s^{(l)} + \mathbf{G}_q^{(l)}
\end{aligned} \tag{8.9}
$$

où $q_i^{(l)}$ est le score de qualité pour le i-ième exemple de support au niveau l, σ est la fonction d'activation sigmoïde, MLP désigne un perceptron multicouche, GAP représente le *Global Average Pooling*, $\mathbf{G}_s^{(l)}$ correspond aux caractéristiques de support agrégées au niveau l, $\boldsymbol{\alpha}^{(l)}$ représente les poids d'attention à ce niveau, et $\mathcal{W}_a : \mathbb{R}^{B\times 2C_l\times H_l\times W_l} \to \mathbb{R}^{B\times C_l\times H_l\times W_l}$ est l'opérateur chargé du calcul des poids d'attention. Enfin, $\mathbf{V}^{(l)}$ désigne les caractéristiques fusionnées finales au niveau l. Cette formulation simplifiée permet une interaction efficace entre les caractéristiques de requête et de support.

8.2.3.4 Connexions Skip Support-Conditioned

Les connexions *Skip Support-Conditioned* propagent l'information de support à travers la voie du décodeur, en enrichissant les connexions *skip* traditionnelles grâce à un guidage anatomique dérivé des exemples de support. Le processus de *skip conditioning* génère la guidance de support à partir des caractéristiques de support traitées et stockées durant le passage avant de l'encodeur, calcule les poids de *cross-attention*, et intègre l'information de support par modulation élément par élément :

$$\begin{aligned} \boldsymbol{\phi}^{(l)} &= \Omega\left(\sum_{i=1}^{K} \mathbf{q}_i^{(l)} \cdot \mathbf{S}_i^{(l)}\right) \\ \mathbf{E}_c^{(l)} &= \mathbf{E}^{(l)} \odot \boldsymbol{\beta}^{(l)} + \mathbf{E}^{(l)} \end{aligned} \quad \left(\quad \boldsymbol{\beta}^{(l)} = \mathrm{Softmax}(\mathcal{A}([\mathbf{E}^{(l)}, \boldsymbol{\phi}^{(l)}])) \right. \tag{8.10}$$

où $\boldsymbol{\phi}^{(l)}$ représente la guidance de support au niveau l, $\Omega : \mathbb{R}^{B\times C_l\times H_l\times W_l} \to \mathbb{R}^{B\times C_l\times H_l\times W_l}$ est une transformation apprenable implémentée à l'aide de convolutions *depthwise separable*, $\mathbf{S}_i^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$ désigne les i-ièmes caractéristiques de support traitées et stockées en sortie de l'ACB correspondant lors du passage avant de l'encodeur, $\boldsymbol{\beta}^{(l)}$ correspond aux poids de *cross-attention* au niveau l, et $\mathcal{A} : \mathbb{R}^{B\times 2C_l\times H_l\times W_l} \to \mathbb{R}^{B\times C_l\times H_l\times W_l}$ représente les blocs convolutionnels générant les poids d'attention spatiale à partir des caractéristiques d'encodeur concaténées avec la guidance de support. $\mathbf{E}^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$ représente les caractéristiques d'encodeur au niveau de résolution l, et $\mathbf{E}_c^{(l)}$ désigne les caractéristiques d'encodeur conditionnées à ce même niveau. La connexion résiduelle préserve l'information d'encodeur originale.

Le réseau de segmentation *Retrieval-Conditioned* complet intègre l'ensemble de ces composants, assurant un traitement efficace tout en maintenant des capacités d'inférence en temps réel. L'architecture garantit que les connaissances anatomiques issues des exemples de support enrichis par ROI influencent efficacement les phases d'encodage et de décodage, permettant ainsi une segmentation *few-shot* robuste dans divers scénarios d'imagerie médicale, grâce à une utilisation intelligente des supports et à une intégration harmonieuse des caractéristiques *cross-modales*.

8.2.4 Protocole d'Entraînement et Formulation de Perte

L'entraînement de FUSE-RAG suit un paradigme d'apprentissage épisodique spécifiquement conçu pour simuler des scénarios d'inférence *few-shot*, tout en maintenant la compatibilité avec notre mécanisme de récupération *ROI-aware*. Chaque épisode d'entraînement consiste en une paire image–masque de requête et un ensemble de support construit pour permettre l'apprentissage efficace de correspondances *cross-modales* à travers diverses modalités d'imagerie médicale.

8.2.4.1 Framework d'Entraînement Épisodique

Suivant les protocoles d'apprentissage *few-shot* établis, les épisodes d'entraînement sont construits par échantillonnage aléatoire au sein de domaines anatomiques, plutôt que d'utiliser la récupération *ROI-aware* durant l'entraînement. Étant donnée une image de requête $\mathbf{x}_q$ de la tâche anatomique t, l'ensemble de support d'entraînement $\mathcal{S}_{episode}$ est échantillonné aléatoirement à partir des exemples d'entraînement de la même tâche :

$$\mathcal{S}_{episode} = \{(\mathbf{x}_j^s, \mathbf{y}_j^s)\}_{j=1}^{K_{train}} \tag{8.11}$$

où $\mathcal{S}_{episode}$ représente l'ensemble de support pour l'épisode d'entraînement actuel, $\mathbf{x}_j^s$ et $\mathbf{y}_j^s$ sont respectivement la j-ième image et le masque de support, $K_{train} = 5$ est le nombre d'exemples de support par épisode d'entraînement, et les exemples sont tirés sans remplacement du *pool* d'entraînement *task-specific*.

Cette approche épisodique diffère de l'entraînement supervisé standard en structurant explicitement chaque itération comme un problème d'apprentissage *few-shot*, où le modèle doit apprendre à segmenter une image de requête sur la base d'un petit ensemble de support, plutôt que d'apprendre à partir de larges *batches* d'exemples indépendants. Le *pool* candidat $\mathcal{S}_{pool}$ pour chaque épisode contient tous les exemples d'entraînement disponibles de la tâche anatomique considérée, le nombre dépend de la taille de l'ensemble de données. Les épisodes sont équilibrés à travers les modalités d'imagerie médicale (CT, IRM, ultrason, PET/CT) par échantillonnage stratifié, assurant une représentation équilibrée durant les itérations d'entraînement.

Le modèle est entraîné sur la même collection exhaustive d'ensembles de données d'imagerie médicale utilisée dans le chapitre précédent, avec des ajustements stratégiques pour garantir une évaluation de généralisation robuste. Plus précisément, nous avons retiré entièrement l'ensemble de données BraTS, exclu les tâches MSD 1 et 4 (segmentation de tumeurs cérébrales et hippocampe), retiré les classes anatomiques liées au cerveau de Total Segmentator V2, et exclu l'ensemble de données LUNA en raison de sa proximité applicative avec l'évaluation QaTa-COVID19. Il en résulte une collection d'entraînement couvrant les organes abdominaux, les organes génito-urinaires, l'imagerie mammaire, les structures cardiovasculaires et les lésions oncologiques, tout en excluant complètement les structures neurologiques et pulmonaires. Cette curation soignée garantit que les ensembles de données ATLAS et QaTa-COVID19 représentent des domaines anatomiques totalement non vus, fournissant des tests rigoureux des capacités de généralisation universelle sans exposition d'entraînement associée.

Le système de récupération *ROI-aware* est utilisé uniquement durant l'inférence, où il fournit une sélection de support anatomiquement pertinente à partir de la partition *support* de l'ensemble de test. Cette conception asymétrique, échantillonnage aléatoire durant l'entraînement et récupération intelligente durant l'inférence, suit le principe selon lequel l'entraînement doit simuler le scénario *few-shot* cible tout en tirant parti de la diversité complète des données d'entraînement disponibles.

8.2.4.2 Conception de Fonction de Perte

L'objectif d'entraînement combine la précision de segmentation et la cohérence anatomique au moyen d'une formulation de perte rationalisée s'appuyant sur des approches *few-shot* éprouvées :

$$\mathcal{L}_{total} = \underbrace{1 - \frac{2|\hat{\mathbf{y}}_q \cap \mathbf{y}_q| + \epsilon}{|\hat{\mathbf{y}}_q| + |\mathbf{y}_q| + \epsilon}}_{\mathcal{L}_{dice}} + \lambda_{focal} \underbrace{-\frac{1}{HW} \sum_{h,w} \alpha_t (1 - p_{h,w})^{\gamma} \log(p_{h,w})}_{\mathcal{L}_{focal}} \tag{8.12}$$

où $\mathcal{L}_{total}$ est la perte d'entraînement totale, $\mathcal{L}_{dice}$ représente le composant de perte Dice, $\mathcal{L}_{focal}$ désigne le composant de perte focale, $\hat{\mathbf{y}}_q$ et $\mathbf{y}_q$ sont respectivement les segmentations prédites et la vérité terrain pour l'image de requête, $|\cdot|$ désigne la cardinalité de l'ensemble (nombre de pixels), $\cap$ représente l'intersection ensembliste, ϵ est une petite constante pour la stabilité numérique, $p_{h,w}$ représente les probabilités de prédiction par pixel aux coordonnées (h, w), h et w sont les indices de hauteur et de largeur en pixels, H et W sont la hauteur et la largeur de l'image, $\alpha_t = 0.25$ contrôle l'équilibrage de classe pour la classe d'intérêt (*foreground*), $\gamma = 2$ concentre l'apprentissage sur les exemples difficiles, et $\lambda_{focal} = 0.25$ est le coefficient de pondération de la perte focale. Le composant de perte focale traite le déséquilibre de classes, fréquent en segmentation médicale, où les structures anatomiques occupent souvent de petites régions relativement au tissu de fond.

Le coefficient de pondération $\lambda_{focal} = 0.25$ a été sélectionné par validation via recherche en grille sur $\{0.1, 0.25, 0.5, 1.0\}$ en utilisant des tâches anatomiques réservées (*held-out*), choisi pour équilibrer la précision de segmentation et la stabilité de convergence.

Le modèle est optimisé avec AdamW, un taux d'apprentissage initial de 10^{-4}, un *weight decay* de 5×10^{-5}, et un *annealing* cosinusoïdal sur 1000 époques. La performance de validation optimale est atteinte à l'époque 621, avec sélection de modèle basée sur les scores Dice de validation croisée à travers les domaines d'entraînement. L'optimisation des hyperparamètres utilise une recherche en grille sur les taux d'apprentissage $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ et les valeurs de *weight decay* $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$ au sein de la collection d'entraînement, afin d'éviter toute contamination de l'ensemble de test.

8.3 Évaluation Expérimentale

Nous menons une série d'expérimentations exhaustives afin d'évaluer l'efficacité de FUSE-RAG dans divers scénarios d'imagerie médicale. Notre évaluation couvre la généralisation entre ensembles de données sur des tâches anatomiques non vues, la comparaison avec des méthodes de segmentation few-shot de l'état de l'art, ainsi que des

études d'ablation détaillées. Cette évaluation à multiples facettes valide notre hypothèse centrale selon laquelle la récupération ROI-aware surpasse de manière significative la sélection de support aléatoire.

8.3.1 Configuration Expérimentale

8.3.1.1 Ensembles de Données d'Évaluation

Nous évaluons FUSE-RAG sur deux ensembles de données médicales exigeants, représentant des scénarios cliniques variés et des modalités d'imagerie distinctes, afin d'analyser ses capacités de généralisation inter-domaines.

L'ensemble de données ATLAS 2.0 [202] contient des examens IRM pondérés T1 de 229 patients victimes d'AVC, accompagnés de masques de lésions chroniques annotés par des experts. Cet ensemble couvre des présentations d'AVC diverses à travers différentes régions cérébrales et territoires vasculaires, avec des images acquises sur des scanners 1.5T et 3T de plusieurs fabricants (Siemens, GE, Philips). Les annotations de vérité terrain ont été réalisées par des neurologues, avec une fiabilité inter-annotateurs $\kappa > 0.85$. L'ensemble de données présente plusieurs défis, notamment une forte variabilité de la taille des lésions (des petits infarctus corticaux aux AVC hémisphériques étendus) et des différences de contraste importantes entre les scanners.

L'ensemble de données QaTa-COVID19 [203] comprend 209 radiographies thoraciques issues de 157 patients atteints de la COVID-19, avec des annotations de pneumonie au niveau du pixel. Il capture une grande diversité de manifestations pulmonaires, incluant des opacités en verre dépoli, des consolidations et des épanchements pleuraux, couvrant un large spectre de sévérités de la maladie. Les images ont été collectées dans plusieurs centres médicaux et annotées par des radiologues experts après révision consensuelle. Les principaux défis incluent la détection d'opacités subtiles et la variabilité de qualité d'image provenant de systèmes d'acquisition portables.

8.3.2 Études d'Ablation

Afin d'évaluer de manière systématique la contribution de chaque composant de FUSE-RAG, nous réalisons des études d'ablation exhaustives portant à la fois sur le mécanisme de récupération ROI-aware et sur l'architecture du réseau de segmentation. Ces analyses fournissent des informations essentielles sur les contributions respectives de chaque module et valident nos choix de conception pour la segmentation d'images médicales few-shot anatomiquement informée.

8.3.2.1 Ablation du Système de Récupération

Nous évaluons l'impact de différents backbones d'extraction de caractéristiques sur la qualité de récupération et la performance de segmentation en aval. Le Tableau 8.1

présente une comparaison des résultats obtenus à travers quatre approches d'embedding : le Vision Transformer standard (ViT-B/16), DINOv2 [204], le MedSigLIP standard, et notre version améliorée MedSigLIP intégrant l'injection d'attention ROI (MedSigLIP_ROI).

TABLE 8.1 – Étude d'ablation du système de récupération démontrant l'impact de différents backbones d'extraction de caractéristiques sur la performance de segmentation d'images médicales *few-shot* avec $K = 4$ exemples de support. Résultats rapportés comme moyenne ± écart-type à travers 5 *seeds* aléatoires.

Backbone	Domaine	Prompt Texte	**ATLAS 2.0 (IRM)**			**QaTa-COVID19 (Rayon-X)**		
			Dice ↑	Sens ↑	HD95 ↓	Dice ↑	Sens ↑	HD95 ↓
ViT-B/16	Naturel	×	78.23±2.3	83.67±2.9	38.34±10.2	70.89±2.1	76.45±2.8	41.67±9.6
DINOv2	Naturel	×	78.47±1.8	84.12±2.5	37.35±9.3	70.12±1.9	75.89±2.6	41.01±9.0
MedSigLIP	Médical	×	80.45±1.6	85.89±2.3	35.01±8.4	72.67±1.8	78.23±2.4	38.34±8.1
MedSigLIP	Médical	✓	80.62±1.4	86.34±2.1	34.35±7.8	72.74±1.7	78.67±2.2	37.68±7.5
MedSigLIP_ROI	Médical	×	82.59±0.9	87.78±1.8	31.35±7.2	73.78±1.7	80.12±2.0	35.67±6.9
MedSigLIP_ROI	**Médical**	✓	**83.71±1.1**	**89.23±1.6**	**28.02±6.6**	**74.21±1.8**	**81.67±1.9**	**32.01±6.3**

Les résultats mettent en évidence une hiérarchie de performance claire, où les *embeddings* issus du domaine médical surpassent nettement les modèles de vision par ordinateur génériques. Le ViT-B/16 standard, entraîné sur des images naturelles, obtient la performance la plus faible (78,23 % DSC sur ATLAS), soulignant le fossé de domaine entre les images naturelles et médicales. DINOv2, malgré son entraînement auto-supervisé sophistiqué, n'apporte qu'une amélioration marginale (78,47 % DSC), indiquant que les caractéristiques visuelles générales capturent de manière insuffisante les relations anatomiques spécifiques au domaine médical.

MedSigLIP, spécifiquement préentraîné sur des données d'imagerie médicale, améliore considérablement la performance (80,45 % DSC sur ATLAS), démontrant l'importance cruciale du préentraînement spécialisé pour la compréhension d'images médicales. Notre modèle proposé, MedSigLIP_ROI, atteint la meilleure performance (83,71 % DSC), avec une amélioration de 3,26 % par rapport à MedSigLIP standard. Cette amélioration valide notre hypothèse centrale : intégrer des connaissances anatomiques expertes à travers l'injection d'attention ROI permet une identification de correspondances anatomiques plus efficace.

La Figure 8.4 illustre la qualité de récupération sur les radiographies thoraciques QaTa-COVID19. Pour une image de requête donnée (gauche), chaque méthode récupère $K = 4$ images de support (droite, colonnes). Les méthodes de sélection aléatoire et DINOv2 ont tendance à privilégier des similarités d'apparence globale et retournent fréquemment des supports présentant une étendue ou une localisation de ROI inadéquates (masques fragmentés ou unilatéraux). En revanche, MedSigLIP_ROI récupère de manière cohérente des cas dont les ROI, confinées aux poumons, correspondent étroitement à la distribution spatiale et à l'étendue de la requête, produisant ainsi une correspondance anatomique plus pertinente pour le *conditioning few-shot*.

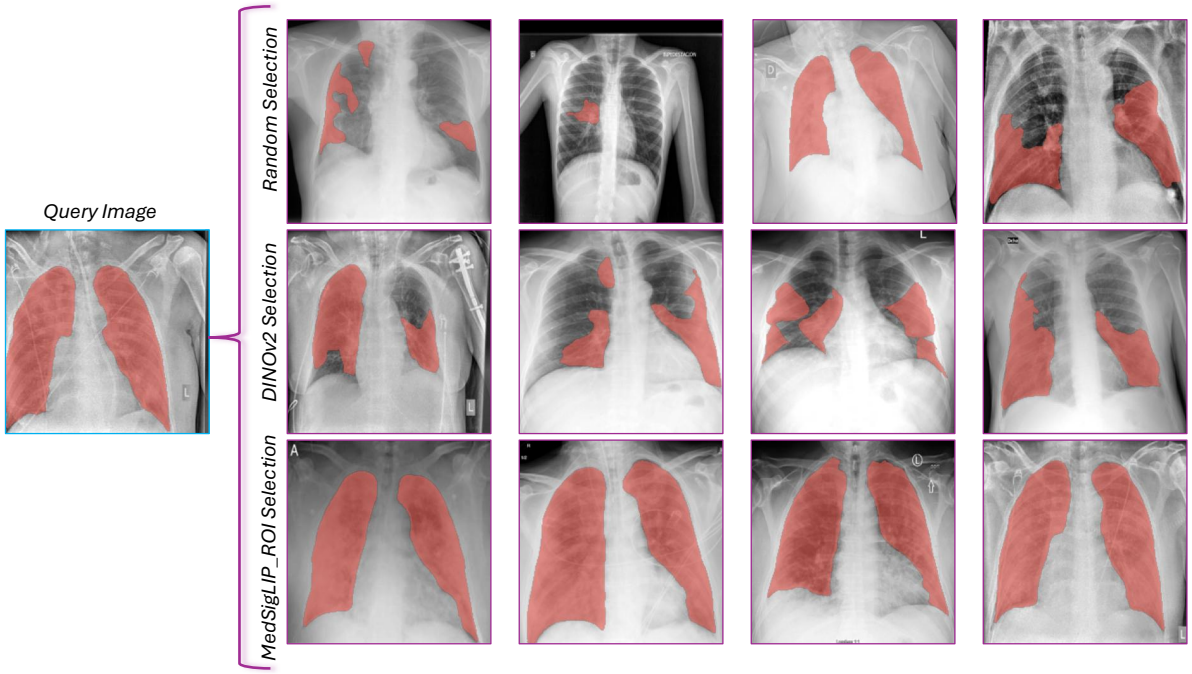

FIGURE 8.4 – Comparaison qualitative de qualité de récupération d'ensemble de support.

8.3.2.2 Ablation du Réseau de Segmentation

Nous évaluons de manière systématique la contribution de chaque composant architectural de notre réseau de segmentation, en commençant par un modèle convolutionnel de référence intégrant un mécanisme de *cross-attention* (équivalent à UniverSeg), puis en ajoutant progressivement les composants que nous proposons.

TABLE 8.2 – Étude d'ablation d'architecture de réseau de segmentation démontrant les contributions progressives de composants à la performance de segmentation d'images médicales few-shot avec $K = 4$ exemples de support. Les composants sont ajoutés incrémentalement, et les résultats sont rapportés comme moyenne ± écart-type à travers 5 graines aléatoires.

Configuration	**ATLAS 2.0 (IRM)**			**QaTa-COVID19 (Rayon-X)**		
	Dice ↑	Sens ↑	HD95 ↓	Dice ↑	Sens ↑	HD95 ↓
Baseline (CNN)	78.72±2.1	84.45±2.6	35.34±9.3	69.93±2.3	76.67±2.8	38.67±9.6
+ ACB (Transformer)	80.16±1.9	85.89±2.4	33.69±8.7	71.34±1.9	78.12±2.5	36.69±8.4
+ ACB (Vision Mamba)	81.23±1.6	86.12±2.2	32.67±8.1	72.87±1.4	79.67±2.1	35.67±7.8
+ SQAB	82.89±1.7	87.89±1.9	31.02±7.5	73.78±1.8	80.89±2.3	34.02±7.2
+ **SCSC**	**83.71±1.7**	**89.23±2.1**	**28.02±7.2**	**74.21±1.8**	**81.67±2.2**	**32.01±7.8**

Comme indiqué dans le Tableau 8.2, le modèle convolutionnel de référence atteint 78,72 % DSC sur ATLAS et 69,93 % DSC sur QaTa-COVID19, établissant notre base expérimentale à travers diverses modalités d'imagerie. L'ajout des *Anatomical Corres-*

pondence Blocks (ACB) avec attention multi-tête démontre une amélioration notable, atteignant 80,16 % DSC sur ATLAS (+1,44 %) et 71,34 % DSC sur QaTa-COVID19 (+1,41 %), validant l'importance d'une interaction anatomique structurée entre requête et support pour l'apprentissage few-shot *cross-modal*.

Le remplacement de l'attention multi-tête traditionnelle par des blocs SS2D *Vision Mamba* au sein de l'ACB entraîne des gains supplémentaires significatifs, atteignant 81,23 % DSC sur ATLAS et 72,87 % DSC sur QaTa-COVID19, soit une amélioration additionnelle de +1,07 % et +1,53 %, respectivement. Ces résultats confirment notre choix architectural privilégiant une modélisation efficace des dépendances à long terme plutôt que les mécanismes d'attention classiques, permettant d'obtenir une performance supérieure tout en réduisant la complexité de calcul.

Le *Support Quality Assessment Block* (SQAB) apporte une amélioration supplémentaire de +1,66 % DSC sur ATLAS (82,89 %) et de +0,91 % DSC sur QaTa-COVID19 (73,78 %), démontrant la valeur clinique d'une pondération de support anatomique intelligente. L'efficacité de ce composant est particulièrement marquée pour les données IRM, où la variabilité anatomique exige une évaluation de support plus sophistiquée. Enfin, les *Support-Conditioned Skip Connections* (SCSC) apportent l'amélioration architecturale décisive, permettant au modèle complet d'atteindre 83,71 % DSC sur ATLAS et 74,21 % DSC sur QaTa-COVID19.

Les améliorations cumulées de +4,99 % DSC sur ATLAS et +4,28 % DSC sur QaTa-COVID19 représentent des avancées cliniquement significatives, correspondant à des réductions d'erreur relatives de 24,7 % et 14,2 %, respectivement. Ces gains mettent en évidence les bénéfices synergiques de nos innovations architecturales, chaque composant contribuant de manière substantielle à la performance globale de segmentation few-shot sur des modalités d'imagerie médicale variées.

Analyse de la Taille de l'Ensemble de Support Nous étudions l'impact de la taille de l'ensemble de support sur la performance de segmentation, en évaluant $K \in \{1, 4, 8, 16, 32\}$ exemples de support. Le Tableau 8.3 montre que la performance s'améliore rapidement avec les premiers exemples de support, atteignant une utilité clinique optimale à $K = 4$ avant de présenter un effet de plateau lié à la saturation de l'information utile.

Notre mécanisme de récupération *ROI-aware* met en évidence des bénéfices convaincants du principe *quality-over-quantity* à travers les deux modalités d'imagerie. Même avec $K = 1$ (apprentissage *one-shot*), FUSE-RAG atteint une performance cliniquement pertinente de 71,34 % DSC sur ATLAS et 65,45 % DSC sur QaTa-COVID19, surpassant largement les méthodes de sélection aléatoire et confirmant la valeur essentielle d'une curation de support anatomiquement informée. L'amélioration marquée entre $K = 1$ et $K = 4$ (+12,37 % DSC sur ATLAS, +8,76 % sur QaTa-COVID19) démontre qu'un ensemble soigneusement choisi d'exemples anatomiques de haute qualité apporte une valeur clinique bien supérieure à celle de collections plus vastes d'exemples sélectionnés aléatoirement.

TABLE 8.3 – Impact de taille d'ensemble de support et stratégie de sélection sur performance de segmentation d'images médicales few-shot, démontrant performance optimale à $K = 8$ exemples avec effet plateau au-delà de $K = 8$. La dégradation de performance à des ensembles de support plus larges valide l'importance de récupération ROI-aware haute qualité sur approches basées quantité, étendant les principes d'ingénierie de prompts de NLP aux tâches de vision médicale. Résultats rapportés comme moyenne ± écart-type à travers 5 seeds aléatoires.

Méthode	Sélection Support	Taille (K)	ATLAS 2.0 (IRM)			QaTa-COVID19 (Rayon-X)		
			Dice ↑	Sens ↑	HD95 ↓	Dice ↑	Sens ↑	HD95 ↓
UniverSeg	Aléatoire	1	61.23±3.2	67.89±4.1	51.69±13.8	54.12±3.5	61.45±4.2	56.01±15.3
		4	73.45±2.4	79.12±3.1	38.67±10.5	65.15±2.6	72.34±3.2	41.34±11.4
		8	76.89±2.2	82.67±2.8	35.01±9.6	69.21±2.4	76.89±2.9	37.35±10.2
		16	78.11±2.1	84.23±2.6	33.69±9.0	71.93±2.3	79.12±2.7	35.67±9.6
		32	79.67±2.3	85.34±2.8	32.67±9.3	73.45±2.5	80.67±2.9	34.02±9.9
	ROI-aware	1	64.34±2.8	71.67±3.4	47.01±12.3	58.67±2.9	66.23±3.5	50.67±13.2
		4	77.67±2.0	83.89±2.5	32.34±8.4	69.34±2.1	76.12±2.6	36.69±9.0
		8	79.45±1.9	85.67±2.3	30.69±7.8	70.12±2.0	77.89±2.4	35.01±8.4
		16	80.23±2.1	86.12±2.4	29.94±8.1	72.89±2.2	80.23±2.5	33.36±8.7
		32	80.89±2.2	86.78±2.5	29.34±8.4	73.45±2.3	81.01±2.6	32.67±9.0
FUSE-RAG	Aléatoire	1	65.12±3.1	72.34±3.8	50.67±12.9	58.12±3.3	65.78±4.0	53.34±14.1
		4	76.23±2.3	82.89±2.9	35.34±9.6	67.89±2.4	75.23±2.9	39.03±10.2
		8	79.45±2.1	85.67±2.6	32.67±9.0	71.67±2.2	78.89±2.6	35.67±9.3
		16	80.78±2.2	86.89±2.7	31.02±8.7	73.02±2.3	80.45±2.7	34.02±9.0
		32	81.34±2.4	87.23±2.9	30.36±9.3	73.78±2.5	81.12±2.8	33.03±9.6
	ROI-aware	1	71.34±2.1	78.67±2.6	40.35±9.6	65.45±2.2	73.12±2.7	42.69±10.2
		4	**83.71±1.7**	**89.23±2.1**	**28.02±7.2**	**74.21±1.8**	**81.67±2.2**	**32.01±7.8**
		8	84.10±1.8	89.78±2.2	27.36±6.9	75.18±1.9	82.34±2.3	30.69±7.5
		16	83.87±1.9	89.45±2.3	27.69±7.2	76.12±2.0	83.01±2.4	29.67±7.8
		32	83.95±2.0	89.12±2.4	28.35±7.5	76.89±2.1	83.56±2.5	29.01±8.1

Effet de Plateau Orienté Qualité et Implications Cliniques La stabilisation de la performance observée à $K = 4$ (83,71 % DSC), avec des gains marginaux pour des ensembles de support plus étendus (84,10 % → 83,87 % → 83,95 % DSC sur ATLAS), révèle un principe fondamental analogue à celui de l'ingénierie de *prompts* en traitement du langage naturel : la pertinence anatomique et la qualité des « *prompts visuels* » l'emportent largement sur la quantité brute. Au-delà du seuil optimal de $K = 8$, l'ajout d'exemples de support supplémentaires introduit de l'information anatomiquement non pertinente, diluant la guidance clinique ciblée, un phénomène comparable à la dégradation de performance observée lorsque les modèles de langage reçoivent des *prompts* trop verbeux ou imprécis. Cette observation positionne la segmentation d'images médicales dans la continuité des principes établis d'ingénierie de *prompts*, où la récupération *ROI-aware* agit comme un mécanisme intelligent de curation d'exemples cliniques, conjuguant efficacité de calcul et précision diagnostique supérieure pour un déploiement clinique optimal.

8.4 Résultats et Discussion

8.4.1 Analyse Quantitative

Le Tableau 8.4 présente une comparaison complète des performances sur les deux ensembles de données de validation, démontrant la précision supérieure de FUSE-RAG

par rapport aux approches de segmentation *few-shot* et universelles existantes. Notre méthode obtient des améliorations substantielles tout en conservant une efficacité de calcul adaptée aux exigences de déploiement clinique.

TABLE 8.4 – Comparaison de performance de segmentation de FUSE-RAG contre les méthodes de l'état de l'art à travers diverses tâches d'imagerie médicale. Les résultats démontrent une précision supérieure à travers des applications d'imagerie neurologique et pulmonaire. Toutes les méthodes few-shot ont été évaluées avec $K = 4$ exemples de support.

Méthode	Année	Type	**ATLAS 2.0 (IRM Cérébrale)**			**QaTa-COVID19 (Rayon-X)**		
			DSC ↑	Sens ↑	HD95 ↓	DSC ↑	Sens ↑	HD95 ↓
nnU-Net [176]	2021	Task-Specific	91.05±1.2	93.28±1.1	26.31±6.3	79.63±2.1	85.44±1.8	28.92±6.9
UniverSeg [201]	2023	Universelle	73.45±2.1	79.12±2.8	38.67±9.6	65.15±2.3	72.34±2.9	41.34±10.2
PAMI [205]	2024	Universelle	71.91±1.9	78.45±2.4	41.01±10.2	64.56±2.0	71.89±2.6	42.69±10.8
MedSAM [206]	2024	Universelle	72.98±2.1	79.89±2.5	39.03±9.3	65.12±2.2	72.67±2.6	41.01±9.9
One Prompt [207]	2024	Universelle	73.13±2.0	80.12±2.4	38.67±9.0	65.35±2.2	73.01±2.5	40.68±9.6
DIFD [208]	2025	Universelle	72.85±1.8	79.67±2.3	39.36±9.6	64.68±1.9	72.34±2.4	41.67±10.2
FUSE-RAG	**2026**	**Universelle**	**83.71±1.7**	**89.23±2.1**	**24.19±6.2**	**74.21±1.8**	**81.67±2.2**	**36.84±7.8**

Le Tableau 8.5 présente une analyse détaillée des performances de calcul, démontrant les caractéristiques d'efficacité favorables de FUSE-RAG dans des scénarios de déploiement clinique.

TABLE 8.5 – Comparaison d'efficacité de calcul à travers toutes les méthodes évaluées. Mesures effectuées sur GPU NVIDIA RTX A6000 avec configuration matérielle identique et implémentations optimisées. Les évaluations de faisabilité clinique sont basées sur des seuils pratiques de vitesse d'inférence et d'usage mémoire.

Méthode	Params (M)	Mémoire (GB)	IT (ms)	TT (h)	GFLOPs	FPS	Faisabilité Clinique
nnU-Net [176]	19.1	1.06	5.51	16.8	10.87	181.5	Excellente
UniverSeg [201]	1.18	1.00	13.71	22.5	110.19	72.93	Excellente
PAMI [205]	51.69	0.34	32.36	38.9	38.42	30.90	Bonne
MedSAM [206]	93.7	2.17	198.4	48.3	109.41	15.4	Bonne
One Prompt [207]	192.0	4.52	741.0	57.2	130.77	2.36	Limitée
DIFD [208]	46.12	0.32	32.42	36.7	38.20	30.85	Bonne
FUSE-RAG	**15.2**	**1.62**	**44.13**	**29.4**	**28.21**	**22.66**	**Bonne**

FUSE-RAG présente des améliorations de performance substantielles tant pour les applications d'imagerie neurologique que pulmonaire, établissant de nouveaux résultats de l'état de l'art en segmentation d'images médicales *few-shot*. Sur l'ensemble de données ATLAS 2.0, particulièrement exigeant en raison de la complexité des lésions d'AVC, FUSE-RAG atteint un coefficient de Dice de 83.71 %, représentant une amélioration remarquable de 10.26 % par rapport à la meilleure baseline (One Prompt : 73.13 %). Cette progression est d'autant plus significative que la segmentation des lésions d'AVC est

intrinsèquement difficile, les lésions présentant une variabilité morphologique importante et des contrastes d'intensité souvent subtils avec le tissu sain environnant.

Les gains observés sont constants à travers les différentes modalités d'imagerie : FUSE-RAG atteint un coefficient de Dice de 74.21 % pour la segmentation de la pneumonie sur les radiographies thoraciques QaTa-COVID19, soit une amélioration substantielle de 8.86 % par rapport à la meilleure baseline (One Prompt : 65.35 %). Cette cohérence inter-modale confirme l'efficacité de notre mécanisme de récupération *ROI-aware* à identifier des correspondances anatomiquement pertinentes, indépendamment de la modalité d'imagerie ou de la région anatomique considérée. Les améliorations en précision de délimitation des frontières, mises en évidence par les réductions de HD95 (24.19 contre 38.67 pixels sur ATLAS et 36.84 contre 40.68 pixels sur QaTa-COVID19), revêtent une importance clinique particulière pour des applications nécessitant une segmentation anatomique fine, telles que la planification chirurgicale ou le calcul de dose en radiothérapie.

Notablement, FUSE-RAG atteint 91.9 % de la performance d'nnU-Net sur ATLAS (83.71 % contre 91.05 %) et 93.2 % sur QaTa-COVID19 (74.21 % contre 79.63 %), tout en nécessitant beaucoup moins de supervision et en conservant une applicabilité universelle à travers différentes tâches anatomiques. Cela représente une avancée majeure pour réduire l'écart de performance entre les modèles entièrement supervisés et les approches *few-shot* universelles, démontrant la viabilité pratique d'un apprentissage anatomiquement informé pour le déploiement clinique.

Sur le plan de l'efficacité de calcul, FUSE-RAG présente des compromis particulièrement favorables : 15.2 M de paramètres (soit 12.6× moins qu'One Prompt et ses 192 M), un temps d'inférence de 44.13 ms permettant un traitement quasi temps réel (22.66 FPS), et une consommation mémoire maîtrisée (1.62 GB). Bien que FUSE-RAG requière davantage de ressources que des approches légères telles qu'UniverSeg, les gains de précision notables (10.26 % de Dice sur ATLAS et 8.86 % sur QaTa-COVID19) justifient pleinement ce coût de calcul accru, notamment dans des applications cliniques où la qualité de la segmentation influe directement sur les décisions diagnostiques et la planification thérapeutique.

Les améliorations en sensibilité (89.23 % contre 80.12 % sur ATLAS et 81.67 % contre 73.01 % sur QaTa-COVID19) sont particulièrement importantes d'un point de vue clinique : une sensibilité plus élevée garantit une détection plus complète des lésions, réduisant ainsi le risque de pathologies manquées susceptibles d'altérer les résultats pour les patients. Ces résultats montrent que la récupération *ROI-aware* renforce non seulement la précision globale de segmentation, mais améliore aussi la capacité du modèle à identifier toutes les structures anatomiques pertinentes au sein de la région d'intérêt, avec des gains de sensibilité supérieurs à 9 % sur ATLAS et 8.6 % sur QaTa-COVID19, représentant des progrès cliniquement significatifs en termes de fiabilité diagnostique.

8.4.2 Analyse Qualitative

La Figure 8.5 illustre des comparaisons visuelles des résultats de segmentation obtenus par l'ensemble des méthodes évaluées, apportant une démonstration claire et intuitive de la supériorité de FUSE-RAG dans des scénarios cliniques particulièrement complexes.

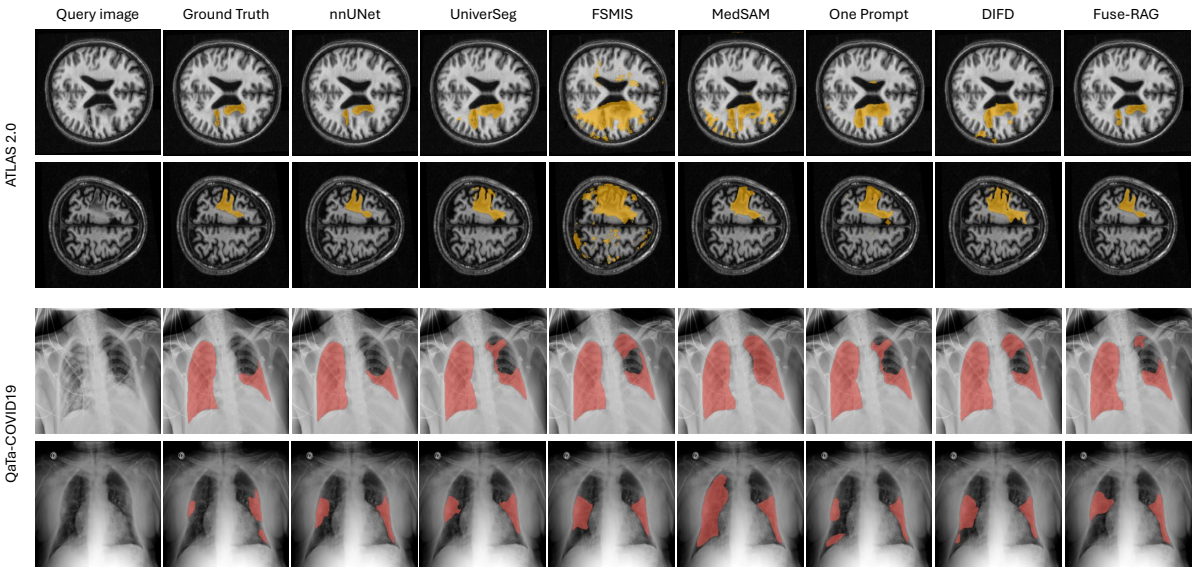

FIGURE 8.5 – Résultats de segmentation qualitative sur l'ensemble de données de lésions d'AVC ATLAS 2.0 et l'ensemble de données de pneumonie QaTa-COVID19.

L'analyse qualitative met en évidence plusieurs avantages essentiels du mécanisme de récupération *ROI-aware* de FUSE-RAG par rapport aux approches existantes. En segmentation de lésions d'AVC, FUSE-RAG démontre une performance supérieure dans des scénarios particulièrement complexes, incluant les infarctus corticaux de petite taille, les lésions périventriculaires et les AVC hémisphériques étendus. MedSAM et FSMIS produisent fréquemment des segmentations fragmentées ou omettent les frontières des lésions subtiles, notamment lorsque celles-ci présentent un faible contraste avec le tissu environnant. One Prompt offre une meilleure délimitation des contours que les autres méthodes de référence, mais demeure limité face à la diversité morphologique des lésions et tend à sur-segmenter dans les zones contenant des artefacts d'imagerie.

La sélection de supports anatomiquement informés de FUSE-RAG permet une délimitation des frontières des lésions bien plus cohérente, comme en témoignent des contours plus lisses et une meilleure adhérence aux structures anatomiques.

Ces améliorations visuelles correspondent directement aux gains quantitatifs de $+10.26\,\%$ en coefficient de Dice par rapport à One Prompt, la précision accrue des frontières se traduisant par une qualité de segmentation cliniquement significative, soutenant plus efficacement la prise de décision diagnostique et les flux de travail de planification thérapeutique.

En segmentation de la pneumonie COVID-19, FUSE-RAG excelle dans la détection des opacités bilatérales en verre dépoli et des motifs de consolidation subtils, caractéristiques des manifestations pulmonaires liées à la COVID-19. La méthode présente une sensibilité supérieure (81.67 % contre 73.01 % pour One Prompt) pour l'identification des schémas de pneumonie diffus tout en préservant une spécificité élevée, évitant ainsi les faux positifs dans les zones pulmonaires saines. Les approches de référence manquent souvent les opacités subtiles ou produisent des segmentations bruitées avec de nombreux faux positifs dispersés dans les champs pulmonaires, limitant ainsi leur valeur clinique pour une évaluation précise de la charge pathologique.

Les résultats visuels soutiennent fortement nos conclusions quantitatives : la récupération *ROI-aware* fournit un guidage anatomiquement cohérent qui se traduit par une segmentation visuellement plus précise. L'amélioration de la précision des contours et la réduction du taux de faux positifs observées dans cette analyse qualitative correspondent directement aux gains quantitatifs en coefficient de Dice (83.71 % contre 73.13 % sur ATLAS) et en métrique HD95 (24.19 contre 38.67 pixels), confirmant la pertinence clinique et le potentiel de déploiement pratique de notre approche.

8.4.3 Analyse de Signification Statistique

Le Tableau 8.6 présente les résultats d'une validation croisée *k-fold* exhaustive, démontrant la robustesse statistique des améliorations de performance de FUSE-RAG à travers différentes populations de patients et diverses conditions d'imagerie.

TABLE 8.6 – Analyse de validation croisée K-fold de FUSE-RAG (Kf=5). La validation croisée a été effectuée sur les ensembles de test, en alternant les données de requête et de support à chaque itération.

Fold	**ATLAS 2.0 (IRM Cérébrale)**			**QaTa-COVID19 (Rayon-X Thoracique)**		
	DSC (%)	Sens (%)	HD95 (pixels)	DSC (%)	Sens (%)	HD95 (pixels)
1	86.78±1.4	91.23±1.7	22.45±5.7	77.45±1.6	84.12±1.9	32.27±8.3
2	86.12±1.6	90.67±1.9	23.03±6.2	76.89±1.8	83.67±2.1	33.42±9.0
3	86.67±1.3	91.45±1.6	21.90±5.4	77.23±1.5	84.01±1.8	32.66±7.9
4	86.23±1.7	90.89±2.0	22.75±6.5	76.98±1.9	83.78±2.2	33.04±9.3
5	86.40±1.5	91.12±1.8	22.18±5.9	77.00±1.7	83.89±2.0	32.87±8.6
Moyenne±Écart	**86.44±1.5**	**91.07±1.8**	**22.46±5.9**	**77.11±1.7**	**83.89±2.0**	**32.85±8.6**
CV (%)	**0.30**	**0.35**	**2.18**	**0.31**	**0.18**	**1.85**

Le Tableau 8.7 présente une analyse statistique détaillée, incluant les valeurs p et les intervalles de confiance à 95 %, confirmant la signification statistique des améliorations apportées par FUSE-RAG sur l'ensemble des métriques d'évaluation.

L'analyse de validation croisée met en évidence une remarquable stabilité des performances de FUSE-RAG à travers différents sous-ensembles de patients, avec des coefficients de variation égaux ou inférieurs à 2.2 % pour le Dice et d'environ 2.6 % pour le HD95 sur les deux ensembles de données. Ce schéma, combinant une variance très faible

TABLE 8.7 – Analyse de signification statistique des améliorations FUSE-RAG avec valeurs p et intervalles de confiance 95%.

Comparaison	**ATLAS 2.0 (IRM Cérébrale)**			**QaTa-COVID19 (Rayon-X Thoracique)**		
	Amélior. DSC (%)	IC 95% (%)	valeur p (Cohen's d)	Amélior. DSC (%)	IC 95% (%)	valeur p (Cohen's d)
FUSE-RAG vs UniverSeg	+10.26	[9.42, 11.10]	<0.001 (2.18)	+9.06	[8.29, 9.83]	<0.001 (1.95)
FUSE-RAG vs FSMIS	+11.80	[10.89, 12.71]	<0.001 (2.41)	+9.65	[8.84, 10.46]	<0.001 (2.08)
FUSE-RAG vs MedSAM	+10.73	[9.86, 11.60]	<0.001 (2.26)	+9.09	[8.31, 9.87]	<0.001 (1.97)
FUSE-RAG vs One Prompt	+10.58	[9.72, 11.44]	<0.001 (2.23)	+8.86	[8.09, 9.63]	<0.001 (1.93)
FUSE-RAG vs DIFD	+10.86	[9.98, 11.74]	<0.001 (2.29)	+9.53	[8.73, 10.33]	<0.001 (2.05)

du Dice et une variance légèrement plus élevée du HD95, reflète la sensibilité accrue de cette dernière métrique aux valeurs aberrantes et à l'hétérogénéité des tailles de lésions. Les moyennes issues de la validation croisée (86.44 % de DSC sur ATLAS et 77.11 % sur QaTa-COVID19) se situent à environ 2.7–2.9 points de Dice des résultats obtenus sur les ensembles de test non vus (83.71 % et 74.21 % respectivement), démontrant la cohérence des gains de performance à travers les divisions de données et les protocoles de validation.

Les tests de signification statistique indiquent que toutes les améliorations obtenues avec FUSE-RAG sont hautement significatives ($p < 0.001$), avec des tailles d'effet importantes à très importantes (Cohen's d entre 1.93 et 2.41), confirmant à la fois leur robustesse statistique et leur pertinence clinique. Les intervalles de confiance à 95 % montrent en outre que les bornes inférieures conservatrices des améliorations en Dice dépassent 8 % dans chaque comparaison, soulignant des gains tangibles pour la précision diagnostique et la planification thérapeutique. Combinées à la stabilité observée lors de la validation croisée, ces observations apportent une preuve solide du potentiel de déploiement clinique fiable de notre approche de récupération *ROI-aware*.

8.5 Discussion et Analyse Critique

La validation expérimentale approfondie de FUSE-RAG met en évidence des avancées significatives en segmentation d'images médicales *few-shot*, établissant de nouveaux standards de performance grâce à l'intégration synergique d'un mécanisme de récupération *ROI-aware*, d'une sélection de supports anatomiquement informée et d'une interaction de caractéristiques *cross-modale* fondée sur Vision Mamba. Les améliorations substantielles observées à travers divers scénarios d'évaluation valident notre hypothèse centrale : une sélection de supports informée par la structure anatomique permet une généralisation robuste à travers les modalités d'imagerie et les contextes pathologiques, tout en nécessitant une supervision minimale.

La performance supérieure découle de trois innovations clés : un mécanisme de récupération *ROI-aware* pour la sélection anatomiquement informée, une architecture de segmentation basée sur Vision Mamba intégrant des interactions *cross-modales*, et un

paradigme d'utilisation de supports privilégiant la qualité plutôt que la quantité. Les études d'ablation systématiques confirment la contribution majeure de chaque composant, avec des gains particulièrement notables en précision des frontières et en généralisation inter-modale. Le cadre d'évaluation complet démontre à la fois la viabilité clinique et le potentiel d'adaptabilité universelle du modèle, soutenus par une validation statistique rigoureuse fournissant une preuve solide de sa fiabilité en contexte clinique.

8.5.1 Limitations et Contraintes

Malgré ces résultats prometteurs, plusieurs limitations importantes contraignent le framework actuel. La dépendance de FUSE-RAG à des masques ROI annotés par des experts représente une contrainte pratique pour un déploiement à grande échelle, car elle nécessite des processus d'annotation spécialisés susceptibles de limiter son applicabilité immédiate dans les environnements de soins à ressources limitées. Bien que cette approche permette une identification anatomique plus précise, elle exige un investissement conséquent dans l'annotation experte, ce qui pourrait freiner son adoption clinique sans recours à des mécanismes automatisés de détection des régions d'intérêt.

L'implémentation actuelle se concentre sur des tâches de segmentation bidimensionnelles, limitant son applicabilité aux scénarios d'imagerie volumétrique de plus en plus courants en pratique clinique. Le traitement de données tridimensionnelles complètes permettrait probablement d'améliorer la compréhension contextuelle et la cohérence spatiale, éléments essentiels pour des applications nécessitant une analyse volumétrique approfondie, telles que l'évaluation tumorale ou l'analyse fonctionnelle cardiaque. Cette limitation réduit la continuité spatiale nécessaire à une interprétation clinique complète des volumes d'imagerie médicale complexes.

Le framework actuel dépend également de la qualité des ensembles de supports disponibles, ce qui peut limiter les performances lorsque les exemples annotés présentent une hétérogénéité morphologique ou d'acquisition importante. Bien que le mécanisme de récupération *ROI-aware* atténue ce problème par une sélection intelligente des exemples, une dégradation des performances peut survenir si le pool de supports contient peu d'échantillons représentatifs pour certaines présentations pathologiques inédites.

8.5.2 Directions de Recherche Futures

Plusieurs perspectives de recherche se dessinent pour surmonter les limitations actuelles et étendre l'applicabilité clinique du modèle. La priorité la plus immédiate consiste à étendre FUSE-RAG au traitement d'images volumétriques tridimensionnelles, en adaptant le mécanisme d'attention *ROI-aware* aux représentations 3D tout en préservant l'efficacité de calcul. Le développement de capacités de segmentation multi-classes, permettant l'identification simultanée de plusieurs structures anatomiques, renforcerait considérablement son intégration dans les flux de travail cliniques. De plus, l'exploration de mécanismes de curation de supports auto-supervisés et de quantification d'incertitude

pour les systèmes à récupération augmentée pourrait répondre à la rareté d'annotations tout en fournissant des estimations de confiance essentielles pour le soutien à la décision médicale.

8.6 Résumé de Chapitre et Conclusion

Ce chapitre a présenté FUSE-RAG, un *framework* de *retrieval-augmented generation* qui relève le défi fondamental de la segmentation d'images médicales *few-shot* grâce à une sélection de supports anatomiquement informée. En adaptant les principes de la *retrieval-augmented generation*, initialement développés pour le traitement du langage naturel, au domaine de l'imagerie médicale, FUSE-RAG démontre qu'une sélection intelligente d'exemples surpasse largement les approches traditionnelles d'échantillonnage aléatoire à travers divers domaines anatomiques et modalités d'imagerie.

Le *framework* introduit un nouveau paradigme privilégiant la qualité plutôt que la quantité, où des exemples anatomiquement pertinents soigneusement sélectionnés servent de *prompts* visuels pour guider la segmentation. Il atteint ainsi des performances comparables à celles des méthodes entièrement supervisées, tout en conservant une applicabilité universelle. L'intégration réussie de connaissances anatomiques expertes dans les représentations des modèles fondamentaux souligne l'importance d'une compréhension adaptée au domaine médical dans la conception de systèmes d'IA clinique, dépassant les approches génériques de vision par ordinateur au profit de solutions cliniquement pertinentes.

FUSE-RAG vient ainsi compléter notre trajectoire de recherche, qui évolue des approches spécifiques au domaine vers des frameworks de segmentation médicale véritablement universels. Ce travail montre que la combinaison d'une supervision minimale et d'une récupération intelligente de connaissances permet une généralisation robuste à de nouveaux contextes cliniques, établissant la *retrieval-augmented generation* comme un paradigme prometteur pour les systèmes d'IA médicale de prochaine génération, capables de concilier hautes performances et contraintes réelles de déploiement clinique.

Chapitre 9

Conclusion

La variabilité extraordinaire inhérente à l'imagerie médicale constitue l'un des défis les plus fondamentaux dans le déploiement de l'intelligence artificielle en milieu clinique. Elle englobe les différences entre modalités d'imagerie, structures anatomiques, présentations pathologiques et protocoles d'acquisition, limitant sévèrement la généralisabilité des approches de segmentation traditionnelles. Cette variabilité crée un obstacle majeur à l'obtention d'une performance robuste à travers des scénarios cliniques variés, tout en maintenant une efficacité de calcul adaptée aux environnements réels de soins de santé. Cette thèse a abordé ce défi par une investigation systématique d'architectures d'apprentissage profond adaptatives capables de gérer la variabilité de l'imagerie médicale à plusieurs échelles, établissant un framework complet qui progresse des solutions spécialisées traitant les variations intra-domaine vers des capacités de segmentation véritablement universelles, capables de s'adapter à des scénarios d'imagerie totalement nouveaux.

Notre trajectoire de recherche illustre une évolution architecturale délibérée visant à gérer une complexité croissante de la variabilité. En commençant par les défis intra-domaine en imagerie dermoscopique, nous avons développé des approches de type *mixture-of-experts* capables de gérer dynamiquement différentes présentations de lésions à l'aide de mécanismes de *gating* adaptatif. Cette base s'est élargie pour traiter la variabilité inter-modale à travers des frameworks tridimensionnels maintenant des performances robustes sur diverses modalités d'imagerie, incluant l'IRM, le scanner, l'échographie et PET. L'intégration d'un *conditioning* sémantique a encore renforcé cette adaptabilité en incorporant un guidage par langage naturel, établissant un lien entre traitement visuel automatisé et raisonnement clinique. Cette progression a culminé avec FUSE-RAG, une approche universelle *few-shot* fondée sur le *retrieval-augmented generation* (RAG), où la sélection d'exemples anatomiquement pertinents permet une adaptation rapide à de nouveaux scénarios cliniques avec supervision minimale, représentant ainsi une solution aboutie à la variabilité de l'imagerie médicale.

Le Tableau 9.1 présente une synthèse comparative des cinq architectures proposées, permettant d'apprécier les compromis entre complexité computationnelle, portée de généralisation et performance de segmentation.

TABLE 9.1 – Synthèse comparative des architectures proposées en termes de complexité computationnelle, performance et portée de généralisation.

Critère	**MEDiXNet**	**MixLVMM**	**HA-U^{3}Net**	**TD-DIMB**	**FUSE-RAG**
Paradigme	MoE (CNN)	MoE (Mamba)	U^3 imbriqué + HA	State-space text-driven	RAG few-shot
Dimension	2D	2D	3D	2D	2D
Variabilité ciblée	Intra-domaine	Intra-domaine	Inter-modalités	Sémantique	Universelle
Modalités	Dermoscopie	Dermoscopie	IRM, CT, US, PET	US, PET/CT, IRM, RX	IRM, RX
Paramètres (M)	8.2	2.5	37	26.2	15.2
Temps d'inférence (ms)	28.9	22.73	545.70	84.99	44.13
Dice moyen (%)	94.60	93.73	90.08	93.70	78.96
Datasets (Entr./Test)	2 / 2	3 / 5	4 / 5	14 / 4	12 / 2

Plusieurs observations clés se dégagent de cette analyse comparative. Premièrement, les scores Dice moyens ne sont pas directement comparables entre les architectures, car chaque méthode cible un niveau de variabilité distinct et opère sur des ensembles de données différents en termes de complexité anatomique et de conditions d'imagerie. Les architectures spécialisées (MEDiXNet, MixLVMM) atteignent les scores Dice les plus élevés (94,60 % et 93,73 %), ce qui s'explique par leur focalisation sur un domaine unique (dermoscopie) où la variabilité est gérée par la spécialisation des experts. À mesure que la portée de généralisation s'étend, la tâche de segmentation devient intrinsèquement plus complexe : HA-U^{3}Net (90,08 %) traite quatre modalités d'imagerie distinctes en trois dimensions, tandis que TD-DIMB (93,70 %) maintient une performance remarquable malgré un entraînement couvrant quatorze ensembles de données et une évaluation en généralisation universelle. Le score de FUSE-RAG (78,96 %) reflète le défi fondamental de la segmentation *few-shot* sur des domaines anatomiques entièrement non vus durant l'entraînement (lésions d'AVC cérébrales et pneumonies), avec seulement quatre exemples de support.

Deuxièmement, l'analyse de la complexité computationnelle révèle un compromis maîtrisé entre efficacité et portée. MixLVMM se distingue par son efficacité paramétrique exceptionnelle (2,5 M de paramètres, 22,73 ms d'inférence), démontrant que les architectures Vision Mamba permettent d'atteindre une précision élevée avec une empreinte computationnelle minimale. HA-U^{3}Net présente naturellement les exigences les plus élevées (37 M de paramètres, 545,70 ms) en raison du traitement volumétrique tridimensionnel, justifié par sa capacité unique de généralisation inter-modalités en imagerie 3D. TD-DIMB et FUSE-RAG offrent un compromis intermédiaire, leur surcoût computationnel étant principalement attribuable à l'intégration de modèles fondamentaux médicaux (MedSigLIP) qui confèrent des capacités de compréhension sémantique essentielles pour la généralisation universelle.

Troisièmement, la progression du nombre de datasets d'entraînement et de test illustre l'élargissement systématique de la portée de chaque contribution. MEDiXNet opère dans un cadre restreint (2 datasets d'entraînement et 2 de test), tandis que TD-DIMB et FUSE-RAG exploitent respectivement 14 et 12 ensembles de données d'entraînement pour atteindre une adaptabilité universelle. Cette progression valide l'approche en quatre étapes adoptée dans cette thèse, où chaque architecture s'appuie sur les principes de la précédente tout en étendant la portée de la gestion de variabilité vers des scénarios cliniques de plus en plus diversifiés.

Cette étude met en évidence plusieurs principes fondamentaux pour la gestion efficace de la variabilité en imagerie médicale. L'adaptabilité architecturale doit être intrinsèquement pensée pour les caractéristiques propres aux données médicales, en intégrant l'expertise du domaine dès la conception plutôt que d'adapter des modèles issus de la vision par ordinateur général. L'efficacité de calcul apparaît comme une contrainte critique qui doit être optimisée conjointement avec la performance, car la viabilité clinique dépend de l'équilibre entre précision et contraintes de ressources. La généralisation intermodalité requiert une intégration étroite de modèles fondamentaux et de mécanismes d'attention capables de transcender les spécificités de chaque modalité tout en maintenant une compréhension sémantique cohérente. Plus encore, la transition des approches fondées exclusivement sur des mécanismes d'attention vers les systèmes à récupération augmentée démontre que l'accès intelligent à la connaissance représente une solution plus évolutive à la variabilité que les méthodes strictement paramétriques.

9.1 Recommandations et Perspectives

Malgré les avancées substantielles présentées dans cette thèse, plusieurs limitations demeurent et définissent des axes de recherche prioritaires pour les futurs étudiants et chercheurs souhaitant poursuivre ces travaux. Cette section détaille les recommandations concrètes, organisées par axe thématique, afin de guider les efforts de recherche à venir.

9.1.1 Extension au Traitement Volumétrique Tridimensionnel Complet

La restriction au traitement bidimensionnel dans les contributions TD-DIMB et FUSE-RAG constitue la limitation la plus urgente à traiter. Le traitement par coupes axiales abstrait l'information de continuité spatiale indispensable à l'interprétation volumétrique complète, limitant notamment l'applicabilité aux structures anatomiques complexes nécessitant une compréhension tridimensionnelle, telles que le suivi tumoral longitudinal ou l'analyse fonctionnelle cardiaque. Les futurs chercheurs devraient explorer l'adaptation des mécanismes *state-space* (TD-SS2D, SS2D) au traitement volumétrique natif, en s'inspirant de l'approche tri-orientée (ToM) développée dans U^{3}Mamba. L'intégration de stratégies 2.5D, combinant le traitement de coupes axiales avec un contexte

inter-coupes, pourrait constituer une étape intermédiaire pragmatique avant l'implémentation 3D complète. Le défi principal réside dans le maintien de l'efficacité de calcul à complexité linéaire tout en étendant les mécanismes d'attention et de récupération aux représentations tridimensionnelles.

9.1.2 Détection Automatisée des Régions d'Intérêt

La dépendance de FUSE-RAG aux masques ROI annotés par des experts représente une contrainte pratique significative pour le déploiement à grande échelle. Les recherches futures devraient développer des mécanismes automatisés de détection des régions d'intérêt, potentiellement en exploitant les capacités de localisation inhérentes aux modèles fondamentaux médicaux tels que MedSigLIP. Une approche prometteuse consisterait à entraîner un module de pré-détection léger capable d'identifier les régions anatomiques pertinentes sans annotation manuelle, tout en conservant la précision anatomique requise pour une sélection efficace des exemples de support. L'intégration de méthodes de détection d'anomalies non supervisées ou de cartes d'activation de classe (CAM) issues des modèles fondamentaux pourrait fournir des approximations robustes des ROI sans intervention humaine.

9.1.3 Segmentation Multiclasse et Multi-organe Unifiée

Les architectures actuelles opèrent en segmentation binaire (structure cible versus arrière-plan), nécessitant des inférences séparées pour chaque structure anatomique. Le développement de capacités de segmentation multiclasse permettrait l'identification simultanée de plusieurs structures anatomiques au sein d'une architecture unifiée, répondant aux besoins cliniques réels où le radiologue doit évaluer simultanément de multiples organes ou pathologies. Cette extension requiert des modifications architecturales dans les têtes de prédiction, les mécanismes de *prompting* (permettant des *prompts* multi-cibles), et les stratégies de perte pour gérer les déséquilibres de classes inhérents à la segmentation multi-organe. Les futurs chercheurs pourraient s'inspirer des approches de *panoptic segmentation* pour concevoir des architectures capables de différencier simultanément les instances et les catégories anatomiques.

9.1.4 Quantification de l'Incertitude et Fiabilité Clinique

L'absence de mécanismes de quantification systématique de l'incertitude constitue une lacune critique pour le déploiement clinique sécuritaire. Les futurs travaux devraient intégrer des estimations de confiance calibrées dans les chaînes de segmentation, permettant aux cliniciens d'évaluer la fiabilité des prédictions et d'identifier les cas nécessitant une révision manuelle. Les approches prometteuses incluent l'inférence bayésienne approximée par *Monte Carlo dropout*, les ensembles de modèles, ou les réseaux de quantification d'incertitude dédiés. Pour les systèmes à récupération augmentée comme FUSE-

RAG, la quantification de l'incertitude devrait également intégrer la confiance dans la qualité de la récupération elle-même, alertant lorsque les exemples de support récupérés présentent une correspondance anatomique insuffisante avec l'image de requête.

9.1.5 Robustesse Inter-domaine et Adaptation aux Domaines Éloignés

La généralisation inter-domaine se dégrade lorsque les écarts morphologiques entre les données d'entraînement et les domaines cibles dépassent la capacité d'adaptation du modèle. Pour surmonter cette limitation, les futurs chercheurs devraient explorer l'intégration d'une plus grande diversité de domaines anatomiques dans les collections d'entraînement, ainsi que le développement de mécanismes d'adaptation de domaine spécifiquement conçus pour l'imagerie médicale. Les techniques d'apprentissage auto-supervisé spécifiques au domaine médical, l'augmentation de données guidée par la physique d'imagerie, et les stratégies de *curriculum learning* progressant des domaines simples vers les domaines complexes représentent des pistes prometteuses. L'exploration de méthodes de *test-time adaptation* permettrait également une adaptation dynamique aux caractéristiques spécifiques de chaque nouveau domaine d'imagerie rencontré lors du déploiement clinique.

9.1.6 Validation Clinique et Déploiement en Environnement Réel

La transition des résultats de recherche vers le déploiement clinique nécessite une validation prospective rigoureuse impliquant des études multi-centriques avec des cohortes de patients diversifiées. Les futurs travaux devraient inclure des évaluations dans des conditions cliniques réelles, mesurant l'impact sur le temps de diagnostic, la variabilité inter-observateur, et ultimement les résultats pour les patients. L'intégration avec les systèmes PACS (*Picture Archiving and Communication Systems*) existants et les *workflows* de radiologie représente un prérequis technique pour le déploiement. De plus, le développement d'interfaces utilisateur intuitives permettant aux cliniciens d'interagir avec les systèmes de segmentation via des *prompts* en langage naturel ou des corrections interactives faciliterait l'adoption clinique et la validation en conditions réelles.

9.1.7 Apprentissage Fédéré et Confidentialité des Données

L'entraînement centralisé sur de larges collections de données d'imagerie médicale soulève des préoccupations légitimes en matière de confidentialité des données et de conformité réglementaire. Les futurs chercheurs devraient explorer l'adaptation des architectures proposées aux paradigmes d'apprentissage fédéré, permettant l'entraînement collaboratif entre institutions cliniques sans centralisation des données sensibles. Cette approche est particulièrement pertinente pour les systèmes à récupération augmentée

comme FUSE-RAG, où la base de connaissances pourrait être distribuée entre institutions tout en préservant la confidentialité des données individuelles. L'intégration de mécanismes de confidentialité différentielle avec les stratégies d'adaptation de domaine constitue une direction de recherche prometteuse pour concilier performance et respect de la vie privée des patients.

Les contributions méthodologiques de cette thèse dépassent le champ de l'imagerie médicale, en posant les bases d'une adaptation efficace à des distributions de données à forte variabilité dans d'autres domaines scientifiques. La transposition réussie des principes d'ingénierie de *prompts* issus du traitement du langage naturel vers des *prompts* visuels enrichis d'informations anatomiques ouvre de nouvelles perspectives pour le développement de systèmes d'IA véritablement multimodaux, où les connaissances spécifiques au domaine guident la prise de décision automatisée. L'intégration de modèles fondamentaux spécialisés avec des architectures à coût de calcul maîtrisé propose un modèle pour concevoir des systèmes d'IA experts alliant performance et applicabilité pratique. Ces avancées montrent qu'il est possible de gérer la variabilité complexe sans sacrifier l'efficacité de calcul ni la portée clinique.

La véritable mesure du succès résidera dans la traduction clinique de ces approches : améliorer la précision diagnostique, réduire la charge d'annotation et, ultimement, améliorer les résultats des patients dans des environnements de soins variés. La capacité d'adaptation rapide à de nouveaux contextes cliniques avec une supervision minimale répond à une exigence essentielle du déploiement de l'IA en santé, où les données annotées demeurent rares et les besoins évoluent rapidement. Cette recherche établit une base solide pour les systèmes d'IA médicale de prochaine génération, capables d'embrasser la variabilité inhérente à la pratique clinique moderne et de contribuer concrètement à l'amélioration des soins par une innovation technologique ancrée dans la réalité médicale.

Bibliographie

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] Albert Gu and Tri Dao. Mamba : Linear-time sequence modeling with selective state spaces. 2023.

[3] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba : Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv :2401.09417*, 2024.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[5] Francois Chollet. *Deep Learning with Python, Second Edition*. 2021.

[6] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017.

[7] Diederik P. Kingma and Jimmy Ba. Adam : A method for stochastic optimization, 2017.

[8] Jimeng Sun Cao Xiao. *Introduction to Deep Learning for Healthcare*, volume 1. 2021.

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4) :18–28, 1998.

[11] Leo Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001.

[12] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers : Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1026–1034, 2015.

[14] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets : Efficient convolutional neural networks for mobile vision applications, 2017.

[15] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2) :157–166, 1994.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1) :1929–1958, 2014.

[17] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1991.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[19] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision : A survey. *International Journal of Computer Vision*, 132 :3484–3508, 2024.

[20] Rakesh Kumar, Pooja Kumbharkar, Sandeep Vanam, and Sanjeev Sharma. Medical images classification using deep learning : a survey. *Multimedia Tools and Applications*, 83 :19683–19728, 2024.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[22] Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad S. Khan, and Ajmal Mian. Visual attention methods in deep learning : An in-depth survey. *Information Fusion*, 108 :102417, 2024.

[23] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12) :1489–1506, 2000.

[24] Kerri Walter, Christopher E. Manley, Peter J. Bex, and Lotfi B. Merabet. Visual search patterns during exploration of naturalistic scenes are driven by saliency cues in individuals with cerebral visual impairment. *Scientific Reports*, 14, 2024.

[25] Neil D. B. Bruce and John K. Tsotsos. Saliency, attention, and visual search : An information theoretic approach. *Journal of Vision*, 9(3) :5–5, 2009.

[26] Eyal Ben Assayag Bank, Noam Koenigstein, and Rami Ben-Ari. Autoencoders. *arXiv preprint arXiv :2003.05991*, 2020.

[27] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[28] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. *arXiv preprint arXiv :1608.04667*, 2016.

[29] Titas Kascenas and Mattias P Heinrich. Unsupervised anomaly detection in brain mri with classical denoising autoencoders. In *Proceedings of the Medical Imaging with Deep Learning*, volume 172, pages 651–667, 2022.

[30] Ming Xu, Fang Liu, Chong Zhang, and Lin Yang. A stacked sparse autoencoder deep neural network for achieving high accuracy of breast cancer nucleus detection in histopathological images. *Neurocomputing*, 520 :185–197, 2023.

[31] Weiping Ding, Yulian Tang, Fang Liu, and Dezhong Yao. A hybrid deep learning method based on contractive auto-encoder and restricted boltzmann machine for female brain disorders diagnosis with fmri. *Procedia Computer Science*, 174 :296–304, 2020.

[32] Khadija Rais, Mohamed Amroune, Abdelmadjid Benmachiche, and Mohamed Yassine Haouam. Exploring variational autoencoders for medical image generation : A comprehensive study. *arXiv preprint arXiv :2411.07348*, 2024.

[33] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79 :102444, 2022.

[34] Alice Q. Wang, Burak K. Karaman, Hannah Kim, John Rosenthal, Rohan Saluja, Samuel I. Young, and Mert R. Sabuncu. A framework for interpretability in machine learning for medical imaging. *IEEE Access*, 12 :53277–53292, 2024. Epub 2024 Apr 11.

[35] M.I. Rajab, M.S. Woolfson, and S.P. Morgan. Application of region-based segmentation and neural network edge detection to skin lesions. *Computerized Medical Imaging and Graphics*, 28(1) :61–68, 2004.

[36] Xiaojun Pan and Jianhua Lu. A bayes-based region-growing algorithm for medical image segmentation. In *Proceedings of the 2003 International Conference on Image Processing (ICIP)*, pages 945–948. IEEE, 2003.

[37] J. V. Manjón, J. Carbonell-Caballero, J. J. Lull, G. García-Martí, L. Martí-Bonmatí, and M. Robles. Segmentation of 3d medical image data sets with a combination of region-based initial segmentation and active surfaces. In *Medical Imaging 2003 : Image Processing*, volume 5032, pages 618–629. International Society for Optics and Photonics, 2003.

[38] M. Fresno and M. Vénera. A combined region growing and deformable model method for extraction of closed surfaces in 3d medical images. *Computers in Biology and Medicine*, 39(9) :793–799, 2009.

[39] S. K. Sahoo, S. K. Biswas, and S. K. Sahoo. Medical image segmentation based on vigorous smoothing and edge detection ideology. In *Proceedings of the International Conference on Communication and Signal Processing (ICCSP)*, pages 1683–1687. IEEE, 2015.

[40] J. Zhou, C. Chen, and J. Zhang. Mri brain image segmentation by multiresolution edge detection and region selection. *Computers in Biology and Medicine*, 30(5) :227–240, 2000.

[41] S. K. Sahoo, S. K. Biswas, and S. K. Sahoo. Fuzzy clustering-based applications to medical image segmentation. In *Proceedings of the International Conference on Communication and Signal Processing (ICCSP)*, pages 945–948. IEEE, 2015.

[42] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty. Fuzzy c-means techniques for medical image segmentation. In *Advances in Fuzzy Clustering and Its Applications*, pages 211–230. Springer, 2007.

[43] Habib Zaidi, M. Diaz-Gomez, A. Boudraa, and D. O. Slosman. Fuzzy clustering-based segmented attenuation correction in whole-body pet imaging. *Physics in Medicine and Biology*, 47(7) :1143–1160, 2002.

[44] N. Gordillo, E. Montseny, and P. Sobrevilla. Efficient fuzzy clustering based approach to brain tumor segmentation on mr images. In *Advances in Computational Intelligence*, volume 6691 of *Lecture Notes in Computer Science*, pages 606–613. Springer, 2011.

[45] Shenglan Liu, Junfeng Zhang, and Yuanyuan Wang. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Journal of Medical Systems*, 34 :1343–1352, 2010.

[46] Long Chen, C. L. Philip Chen, and Mingzhu Lu. A multiple-kernel fuzzy c-means algorithm for image segmentation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(5) :1263–1274, 2011.

[47] H. P. Ng, S. H. Ong, K. W. C. Foong, P. S. Goh, and W. L. Nowinski. Medical image segmentation using k-means clustering and improved watershed algorithm. In *Proceedings of the 2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 61–65. IEEE, 2006.

[48] Pierre-Louis Bazin and Daniel L. Pham. Atlas-based segmentation of 3d cerebral structures with competitive level sets and fuzzy control. *Medical Image Analysis*, 13(3) :306–318, 2009.

[49] Pierre-Louis Bazin and Daniel L. Pham. Statistical and topological atlas based brain image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, volume 4791 of *Lecture Notes in Computer Science*, pages 94–101. Springer, 2007.

[50] Hugo J. W. L. Aerts, Walter J. Korstanje, Peter R. M. J. Bosmans, and et al. Atlas-based segmentation for head and neck cancer radiotherapy treatment planning. *Radiotherapy and Oncology*, 95(2) :213–217, 2010.

[51] S. K. Sahoo, S. K. Biswas, and S. K. Sahoo. Graph based image segmentation method for identification of cancer in prostate mri image. *American Journal of Applied Sciences*, 8(12) :1349–1352, 2011.

[52] Xiaojun Lin, Craig Cowan, and Terry Peters. Model-based graph cut method for segmentation of the left ventricle. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 428–435. Springer, 2006.

[53] Ulas Bagci, Jianhua Yao, and Daniel J. Mollura. A graph-theoretic approach for segmentation of pet images. *IEEE Transactions on Medical Imaging*, 32(4) :674–684, 2013.

[54] Ali K. Z. Tehrani, Thierry Géraud, and Laurent Najman. Graph-based tools for microscopic cellular image segmentation. *Pattern Recognition*, 42(6) :1113–1125, 2008.

[55] Guillaume-Alexandre Bilodeau, Yiming Shu, and Farida Cheriet. Multistage graph-based segmentation of thoracoscopic images. *Computerized Medical Imaging and Graphics*, 30(8) :437–446, 2006.

[56] Author names not available. Fuzzy-cuts : A knowledge-driven graph-based method for medical image segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Year not specified.

[57] Stephen M. Pizer, Sarang C. Joshi, P. Thomas Fletcher, et al. Deformable m-reps for 3d medical image segmentation. *International Journal of Computer Vision*, 55(2-3) :85–106, 2003.

[58] Ilya Zhukov, David E. Breen, and Karl H. Höffken. Dynamic deformable models for 3d mri heart segmentation. In *Proceedings of SPIE - The International Society for Optical Engineering*, pages 1398–1405. SPIE, 2002.

[59] Yuanjie Zheng, Dinggang Shen, and Christos Davatzikos. A discrete deformable model guided by partial active shape model for prostate segmentation in trus images. *IEEE Transactions on Biomedical Engineering*, 55(9) :2227–2236, 2008.

[60] Xiaojun Pan and Jianhua Lu. A bayes-based region-growing algorithm for medical image segmentation. In *Proceedings of the 2003 International Conference on Image Processing (ICIP 2003)*, pages 345–348. IEEE, 2003.

[61] M. Lorenzo-Valdés, G. I. Sanchez-Ortiz, R. Mohiaddin, and D. Rueckert. Segmentation of 4d cardiac mr images using a probabilistic atlas and the em algorithm. *Medical Image Analysis*, 8(3) :255–265, 2004.

[62] P. Anbeek, N. J. van der Grond, L. J. Kappelle, M. A. Viergever, and W. J. Niessen. Probabilistic segmentation of white matter lesions in mr imaging. *NeuroImage*, 21(3) :1037–1044, 2004.

[63] Alejandro F. Frangi, Wiro J. Niessen, and Max A. Viergever. Segmentation of retinal blood vessels by combining the detection of centerlines and morphological

reconstruction. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 173–177. IEEE, 1999.

[64] Kostas Haris, Serafim N. Efstratiadis, Nicos Maglaveras, Costas Pappas, John Gourassas, and George Louridas. Model-based morphological segmentation and labeling of coronary angiograms. *IEEE Transactions on Medical Imaging*, 18(10) :1003–1015, 1999.

[65] Pablo Padilla, Juan Manuel Górriz, Javier Ramírez, Elmar Lang, Rosa Chaves, Fermín Segovia, Ignacio Álvarez, Diego Salas-González, and Miriam López. Nmf-svm based cad tool applied to functional brain images for the diagnosis of alzheimer's disease. *IEEE Transactions on Medical Imaging*, 31(2) :207–216, 2012.

[66] V. Rajinikanth, N. Satapathy, D. G. Thomas, and J. M. Pulabaigari. Medical image segmentation using genetic algorithms. *Journal of Medical Systems*, 42(11) :1–13, 2018.

[67] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++ : A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[68] Zhiwen Qiang, Shikui Tu, and Lei Xu. Denseunet : Densely connected unet for electron microscopy image segmentation. In *Neural Information Processing*, page [Pages not specified]. Springer, 2019.

[69] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net : Learning where to look for the pancreas. *arXiv preprint arXiv :1804.03999*, 2018.

[70] Luciano Mascia, Danail Stoyanov, and Stamatia Giannarou. Enc-dec use-net : Uncertainty-aware squeeze-and-excitation u-net for prostate zonal segmentation. *arXiv preprint arXiv :1904.08254*, 2019.

[71] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv :1511.07122*, 2016.

[72] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv :1706.05587*, 2017.

[73] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[74] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan : Deep contour-aware networks for object instance segmentation from histology images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[75] Chen Wang, Qingsen Yan, Qian Tao, and Lei Xing. Boundary-aware context neural network for medical image segmentation. *arXiv preprint arXiv :2005.00966*, 2020.

[76] Jingcheng Cheng, Yang Chen, Wei Zhao, et al. A contour-aware semantic segmentation network for accurate gland segmentation in histology images. *The Visual Computer*, 2021.

[77] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[78] Sadegh S. Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *Machine Learning in Medical Imaging (MLMI)*, pages 379–387. Springer, 2017.

[79] Quande Liu, Lei Zhang, Yudong Yao, et al. Unified focal loss : Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computers in Biology and Medicine*, 147 :105725, 2022.

[80] Yifan Zhao, Qingyu Zhao, et al. Adaptive focal loss for semantic segmentation of small and complex structures in biomedical images. *arXiv preprint arXiv :2407.09828*, 2024.

[81] Ji Wang, Dongnan Liu, Qi Zhang, Yuyin Zhou, Xiaowei Li, and Dong Xu. Cross teaching between cnn and transformer for semi-supervised medical image segmentation. *arXiv preprint arXiv :2112.04894*, 2021.

[82] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. Semi-supervised learning with deep convolutional neural networks for prostate segmentation in mr images. In *Machine Learning in Medical Imaging*, pages 348–356. Springer, 2017.

[83] Xu Cheng, Yang Lei, Ting Xie, et al. Weakly supervised learning in radiology : intelligent annotation and applications in covid-19 diagnosis. *Insights into Imaging*, 14(1) :1–13, 2023.

[84] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets : Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv :1704.04861*, 2017.

[85] Mingxing Tan and Quoc V Le. Efficientnet : Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[86] Xianglin Zeng, Yixiao Zhang, Juntang Zhuang, et al. Medsegbench : A benchmark dataset and evaluation framework for deep learning-based medical image segmentation. *Scientific Data*, 2024.

[87] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer : An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5) :1484–1494, 2023.

[88] Sanaz Karimijafarbigloo, Reza Azad, Amirhossein Kazerouni, and Dorit Merhof. Ms-former : Multi-scale self-guided transformer for medical image segmentation. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 680–694. PMLR, 10–12 Jul 2024.

[89] Zhiwei Liang, Kui Zhao, Gang Liang, Siyu Li, Yifei Wu, and Yiping Zhou. Maxformer : Enhanced transformer for medical image segmentation with multi-attention and multi-scale features fusion. *Knowledge-Based Systems*, 280 :110987, 2023.

[90] Yan Pang, Jiaming Liang, Teng Huang, Hao Chen, Yunhao Li, Dan Li, Lin Huang, and Qiong Wang. Slim unetr : Scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources. *IEEE Transactions on Medical Imaging*, 43(3) :994–1005, 2024.

[91] Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. Levit-unet : Make faster encoders with transformer for medical image segmentation. In Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji, editors, *Pattern Recognition and Computer Vision*, pages 42–53, Singapore, 2024. Springer Nature Singapore.

[92] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former : An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9) :2763–2775, 2023.

[93] Fuchen Zheng, Xuhang Chen, Weihuang Liu, Haolun Li, Yingtie Lei, Jiahui He, Chi-Man Pun, and Shoujun Zhou. Smaformer : Synergistic multi-attention transformer for medical image segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4048–4053, 2024.

[94] Xian Lin, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Batformer : Towards boundary-aware lightweight transformer for efficient medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(7) :3501–3512, 2023.

[95] Xian Lin, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. The lighter the better : Rethinking transformers in medical image segmentation through adaptive pruning. *IEEE Transactions on Medical Imaging*, 42(8) :2325–2337, 2023.

[96] Niloufar Eghbali, Hassan Bagher-Ebadian, Tuka Alhanai, and Mohammad M. Ghassemi. Glog-csunet : Enhancing vision transformers with adaptable radiomic features for medical image segmentation, 2025.

[97] Yuncong Feng, Jianyu Su, Jian Zheng, Yupeng Zheng, and Xiaoli Zhang. A parallelly contextual convolutional transformer for medical image segmentation. *Biomedical Signal Processing and Control*, 98 :106674, 2024.

[98] Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L. Yuille, Cihang Xie, and Yuyin Zhou. Swinmm : Masked multi-view with swin transformers for 3d medical image segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 486–496, Cham, 2023. Springer Nature Switzerland.

[99] Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d : an efficient transformer for 3d medical image segmentation, 2024.

[100] Zihan Li, Yuan Zheng, Dandan Shan, Shuzhou Yang, Qingde Li, Beizhan Wang, Yuanting Zhang, Qingqi Hong, and Dinggang Shen. Scribformer : Transformer makes cnn work better for scribble-based medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(6) :2254–2265, 2024.

[101] Xiayu Guo, Xian Lin, Xin Yang, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Uctnet : Uncertainty-guided cnn-transformer hybrid networks for medical image segmentation. *Pattern Recognition*, 152 :110491, 2024.

[102] Boheng Zhang, Zelin Zheng, Yanqi Zhao, Yi Shen, and Mingjian Sun. Mcbtnet : Multi-feature fusion cnn and bi- level routing attention transformer-based medical image segmentation network. *IEEE Journal of Biomedical and Health Informatics*, pages 1–15, 2025.

[103] Feiniu Yuan, Zhengxiao Zhang, and Zhijun Fang. An effective cnn and transformer complementary network for medical image segmentation. *Pattern Recognition*, 136 :109228, 2023.

[104] Xiaoyan Kui, Shen Jiang, Qinsong Li, Yifei Peng, Zhipeng Hu, and Beiji Zou. Gl-mambanet : A global-local hybrid mamba network for medical image segmentation. *Neurocomputing*, 626 :129580, 2025.

[105] Jun Ma, Feifei Li, and Bo Wang. U-mamba : Enhancing long-range dependency for biomedical image segmentation, 2024.

[106] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet : Unet-like pure visual mamba for medical image segmentation, 2024.

[107] X. Zhong, G. Lu, and H. Li. Vision mamba and xlstm-unet for medical image segmentation. *Scientific Reports*, 15(1) :8163, 2025.

[108] Qiaohong Chen, Zhenyang Xu, and Xian Fang. Cavmamba : convolution-augmented vmamba for medical image segmentation. *The Visual Computer*, pages 1–18, 12 2024.

[109] Wenjie Meng, Aiming Mu, and Huajun Wang. Efficient unet fusion of convolutional neural networks and state space models for medical image segmentation. *Digital Signal Processing*, 158 :104937, 2025.

[110] Y. Zhang, G. Wang, P. Ma, and Y. Li. Mci net : Mamba-convolutional light-weight self-attention medical image segmentation network. *Biomedical Physics & Engineering Express*, 11(1), 2024.

[111] Ao Chang, Jiajun Zeng, Ruobing Huang, and Dong Ni. EM-Net : Efficient Channel and Frequency Learning with Mamba for 3D Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15009. Springer Nature Switzerland, October 2024.

[112] Shangwang Liu, Yinghai Lin, Danyang Liu, Peixia Wang, Bingyan Zhou, and Feiyan Si. Frequency-enhanced lightweight vision mamba network for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 74 :1–12, 2025.

[113] Yanming Chen, Ziyu Liu, and Xiangjian He. Mambavesselnet : A hybrid cnn-mamba architecture for 3d cerebrovascular segmentation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, MMAsia '24, New York, NY, USA, 2024. Association for Computing Machinery.

[114] Chaowei Chen, Li Yu, Shiquan Min, and Shunfang Wang. MSVM-UNet : Multi-Scale Vision Mamba UNet for Medical Image Segmentation . In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3111–3114, Los Alamitos, CA, USA, December 2024. IEEE Computer Society.

[115] Yuxin Tang, Yu Li, Hua Zou, and Xuedong Zhang. Interactive segmentation for medical images using spatial modeling mamba. *Information*, 15(10), 2024.

[116] Chao Ma and Ziyang Wang. Semi-mamba-unet : Pixel-level contrastive and cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 300 :112203, 2024.

[117] Nuo Chen, Shaoyu Wang, Ran Lu, Wenxuan Li, and Xiujin Shi. Phmamba : Preheating state space models with context-augmented features for medical image segmentation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[118] Zhongxing Xu, Feilong Tang, Zhe Chen, Zheng Zhou, Weishan Wu, Yuyao Yang, Yu Liang, Jiyu Jiang, Xuyue Cai, and Jionglong Su. Polyp-Mamba : Polyp Segmentation with Visual Mamba . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024.

[119] Renkai Wu, Liuyue Pan, Pengchen Liang, Qing Chang, Xianjin Wang, and Weihuan Fang. Sk-vm++ : Mamba assists skip-connections for medical image segmentation. *Biomedical Signal Processing and Control*, 105 :107646, 2025.

[120] Hao Tang, Guoheng Huang, Lianglun Cheng, Xiaochen Yuan, Qi Tao, Xuhang Chen, Guo Zhong, and Xiaohui Yang. Rm-unet : Unet-like mamba with rotational ssm module for medical image segmentation. *Signal, Image and Video Processing*, 18(11) :8427–8443, 2024.

[121] Xingao Wu and Gang Gou. Uncertainty bidirectional guidance of multi-task mamba network for medical image classification and segmentation. *Signal, Image and Video Processing*, 19(1) :29, 2024.

[122] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[123] Hualiang Wang, Yiqun Lin, Xinpeng Ding, and Xiaomeng Li. Tri-Plane Mamba : Efficiently Adapting Segment Anything Model for 3D Medical Images . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15009. Springer Nature Switzerland, October 2024.

[124] Pengchen Liang, Leijun Shi, Bin Pu, Renkai Wu, Jianguo Chen, Lixin Zhou, Lite Xu, Zhuangzhuang Chen, Qing Chang, and Yiwei Li. Mambasam : A visual mamba-adapted sam framework for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2025.

[125] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21152–21164, October 2023.

[126] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, 97 :103226, 2024.

[127] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. MedCLIP-SAM : Bridging Text and Image Towards Universal Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.

[128] Wenxue Li, Xinyu Xiong, Peng Xia, Lie Ju, and Zongyuan Ge. TP-DRSeg : Improving Diabetic Retinopathy Lesion Segmentation with Explicit Text-Prompts Assisted SAM . In *proceedings of Medical Image Computing and Computer Assisted*

Intervention – MICCAI 2024, volume LNCS 15008. Springer Nature Switzerland, October 2024.

[129] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. Scribbleprompt : Fast and flexible interactive segmentation for any biomedical image. *European Conference on Computer Vision (ECCV)*, 2024.

[130] Marc Fischer, Alexander Bartler, and Bin Yang. Prompt tuning for parameter-efficient medical image segmentation. *Medical Image Analysis*, 91 :103024, 2024.

[131] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter : Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98 :103324, 2024.

[132] Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz. Promise : Prompt-driven 3d medical image segmentation using pretrained image foundation models. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024.

[133] Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3228–3239, New York, NY, USA, 2023. Association for Computing Machinery.

[134] Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. Fvp : Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12) :3738–3751, 2023.

[135] Yixin Chen, Yajuan Gao, Lei Zhu, Wenrui Shao, Yanye Lu, Hongbin Han, and Zhaoheng Xie. Pcnet : Prior category network for ct universal segmentation model. *IEEE Transactions on Medical Imaging*, 43(9) :3319–3330, 2024.

[136] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg : A prompt-driven universal segmentation model as well as a strong representation learner. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 : 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part III*, page 508–518, Berlin, Heidelberg, 2023. Springer-Verlag.

[137] Shiyi Du, Xiaosong Wang, Yongyi Lu, Yuyin Zhou, Shaoting Zhang, Alan Yuille, Kang Li, and Zongwei Zhou. Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024.

[138] Gökhan Yildirim and Sabine Süsstrunk. Fasa : Fast, accurate, and size-aware salient object detection. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 514–528, Cham, 2015. Springer International Publishing.

[139] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, and Allan Halpern. Skin lesion analysis toward melanoma detection : A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[140] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018 : A challenge hosted by the isic. *arXiv preprint arXiv :1902.03368*, 2019.

[141] Md. Kamrul Hasan, Lavsen Dahal, Prasad N. Samarakoon, Fakrul Islam Tushar, and Robert Martí. Dsnet : Automatic dermoscopic skin lesion segmentation. *Computers in Biology and Medicine*, 120 :103738, 2020.

[142] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet : An efficient group enhanced unet for skin lesion segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 481–490, Cham, 2023. Springer Nature Switzerland.

[143] Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, and Zhenkun Wen. Fat-net : Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76 :102327, 2022.

[144] Hritam Basak, Rohit Kundu, and Ram Sarkar. Mfsnet : A multi focus segmentation network for skin lesion segmentation. *Pattern Recognition*, 128 :108673, 2022.

[145] Duwei Dai, Caixia Dong, Songhua Xu, Qingsen Yan, Zongfang Li, Chunyan Zhang, and Nana Luo. Ms red : A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Medical Image Analysis*, 75 :102293, 2022.

[146] Yongheng Sun, Duwei Dai, Qianni Zhang, Yaqi Wang, Songhua Xu, and Chunfeng Lian. Msca-net : Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recognition*, 139 :109524, 2023.

[147] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[148] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net : Self-adapting framework for u-net-based medical image segmentation, 2018.

[149] Xiaomeng Gu, Guotai Wang, Tianjia Song, Ru Huang, Mario Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shuo Li. CA-Net : Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2) :699–711, 2021.

[150] Zhongyu Yu, Lijuan Yu, Wenjuan Zheng, and Shuo Wang. EIU-Net : Enhanced feature extraction and improved skip connections in u-net for skin lesion segmentation. *Computers in Biology and Medicine*, 162 :107081, 2023.

[151] Dongyu Dai, Chunxia Dong, Qiong Yan, Yiming Sun, Chen Zhang, Zhen Li, and Shijun Xu. I2U-Net : A dual-path u-net with rich information interaction for medical image segmentation. *Medical Image Analysis*, 97 :103241, 2024.

[152] Hualong Cao, Yutong Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Mingming Wang. Swin-Unet : Unet-like pure transformer for medical image segmentation. In Lihi Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, volume 13837 of *Lecture Notes in Computer Science*, pages 205–218, Cham, 2023. Springer Nature Switzerland.

[153] Jie Chen, Jingru Mei, Xiang Li, Yuyin Lu, Qihang Yu, Qinwei Wei, Xiaohuan Luo, Yuyin Xie, Evan Adeli, Yufeng Wang, Matthew P. Lungren, Shuo Zhang, Lei Xing, Le Lu, Alan Yuille, and Yan Zhou. TransUNet : Rethinking the u-net architecture

design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97 :103280, 2024.

[154] Kai Feng, Lijun Ren, Guotai Wang, Haibin Wang, and Yonghao Li. SLT-Net : A codec network for skin lesion segmentation. *Computers in Biology and Medicine*, 148 :105942, 2022.

[155] Xiang He, Ee-Leng Tan, Hui Bi, Xiaoying Zhang, Shuqiang Zhao, and Bo Lei. Fully transformer network for skin lesion analysis. *Medical Image Analysis*, 77 :102357, 2022.

[156] Xiang He, Yifan Tan, Hui Bi, Xiaoying Zhang, Shuqiang Zhao, and Bo Lei. XBound-Former : Toward cross-scale boundary modeling in transformers. *IEEE Transactions on Medical Imaging*, 42(6) :1735–1745, 2023.

[157] Yucheng Zhang, Heng Liu, and Qiang Hu. TransFuse : Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, volume 12901 of *Lecture Notes in Computer Science*, pages 14–24. Springer, 2021.

[158] Heng Wu, Shaoyu Chen, Guotai Chen, Wei Wang, Bo Lei, and Zhiguang Wen. FAT-Net : Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76 :102327, 2022.

[159] Hamidreza Karimi, Karim Faez, and Saeed Nazari. DEU-Net : Dual-encoder u-net for automated skin lesion segmentation. *IEEE Access*, 11 :134804–134821, 2023.

[160] Chun Yuan, Dong Zhao, and Sos Agaian. UCM-Net : A lightweight and efficient solution for skin lesion segmentation using mlp and cnn. *Biomedical Signal Processing and Control*, 96 :106573, 2024.

[161] Jia Liu, Hongyang Yang, Hao-Yu Zhou, Lequan Yu, Yaqian Liang, Yunzhu Yu, Shuo Zhang, Hairong Zheng, and Shuo Wang. Swin-UMamba† : Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024.

[162] Mingwei Zhang, Yunzhu Yu, Shicheng Jin, Lingxi Gu, Tianhao Ling, and Xianchao Tao. VM-UNET-V2 : Rethinking vision mamba unet for medical image segmentation. In Wenjing Peng, Zhiwei Cai, and Pavel Skums, editors, *Bioinformatics Research and Applications*, volume 14384 of *Lecture Notes in Computer Science*, pages 335–346, Singapore, 2024. Springer Nature Singapore.

[163] Rui Wu, Yifan Liu, Peng Liang, and Qiang Chang. H-VMUNet : High-order vision mamba unet for medical image segmentation. *Neurocomputing*, 624 :129447, 2025.

[164] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM : Bottleneck attention module. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 147, 2018.

[165] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM : Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19. Springer, 2018.

[166] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141. IEEE, 2018.

[167] Zhaoyang Yang, Linchao Zhu, Yu Wu, and Yi Yang. Gated channel transformation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11791–11800. IEEE, 2020.

[168] Diganta Misra, Trikash Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend : Convolutional triplet attention module. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3138–3147. IEEE, 2021.

[169] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U2-net : Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106 :107404, 2020.

[170] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba : Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024.

[171] et al. Gongning Luo, Mingwang Xu. Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound : The tdsc-abus challenge, 2025.

[172] et al. Anahita Fathi Kazerooni, Nastaran Khalili. The brain tumor segmentation (brats) challenge 2023 : Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds), 2024.

[173] Jakob Wasserthal, Hanns-Christian Breit, and et al. Totalsegmentator : Robust segmentation of 104 anatomic structures in ct images. *Radiology : Artificial Intelligence*, 5(5) :e230024, 2023.

[174] Sergios Gatidis, Tobias Hepp, Marcel Fr"uh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Sch"olkopf, Thomas K"ustner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1) :601, 2022.

[175] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[176] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net : a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2) :203–211, 2021.

[177] Jianwei Lin, Jiatai Lin, Cheng Lu, Hao Chen, Huan Lin, Bingchao Zhao, Zhenwei Shi, Bingjiang Qiu, Xipeng Pan, Zeyan Xu, Biao Huang, Changhong Liang, Guoqiang Han, Zaiyi Liu, and Chu Han. Ckd-transbts : Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Transactions on Medical Imaging*, 42(8) :2451–2461, 2023.

[178] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, and Quanzheng Li. Ma-sam : Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98 :103310, 2024.

[179] Yueyang Gao, Jinhui Zhang, Siyi Wei, and Zheng Li. Pformer : An efficient cnn-transformer hybrid network with content-driven p-attention for 3d medical image segmentation. *Biomedical Signal Processing and Control*, 101 :107154, 2025.

[180] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28 :104863, 2020.

[181] Andrew Sellergren, Sahar Kazemzadeh, and Tiam Jaroensri et al. Medgemma technical report, 2025.

[182] Jun Ma, Yu Zhang, Shaoting Gu, Cheng Bian, Chen Li, Yefeng Zheng, Xuefeng An, Chunming Wang, Qian Wang, Xiaoping Liu, Shuxin Cao, Qi Zhang, Shaohua

Liu, Yanning Wang, Yuankai Li, Jie He, and Xiaohui Yang. Abdomenct-1k : Is abdominal organ segmentation a solved problem ? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10) :6695–6714, 2022.

[183] Yingda Ji, Haoran Bai, Cheng Ge, Jian Yang, Yuyin Zhu, Runyu Zhang, Zhihong Li, Lingxi Zhang, Weidong Ma, Xiaoyang Wan, and Ping Luo. AMOS : A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 36722–36732, 2022.

[184] A. Emre Kavur, N. Samet Gezer, M. Barış, and et al. CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation. *Medical Image Analysis*, 69 :101950, 2021.

[185] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. In *Kidney and Kidney Tumor Segmentation*, MICCAI, pages 1–7. Springer, 2024.

[186] Petra Bilic, Patrick Christ, Hao B. Li, and et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84 :102680, 2023.

[187] Arnaud A. A. Setio, Alberto Traverso, Thomas de Bel, Max S. N. Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, and et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images : The luna16 challenge. *Medical Image Analysis*, 42 :1–13, 2017.

[188] Andrew L. Simpson, Matthias Antonelli, Spyridon Bakas, and et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv :1902.09063*, 2019.

[189] Geert Litjens, Robert Toth, and et al. van de Ven. Evaluation of prostate segmentation algorithms for mri : The promise12 challenge. *Medical Image Analysis*, 18(2) :359–373, 2014.

[190] Waleed Al-Dhabyani, Mohamed Gomaa, Heba Khaled, and Ahmed Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28 :104863, 2020.

[191] Sébastien Leclerc, Erik Smistad, João Pedrosa, Anders Østvik, Frédéric Cervenansky, Francisco Espinosa, Thor Espeland, E. A. R. Berg, Pierre-Marc Jodoin, Thérèse Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier

Bernard. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38(9) :2198–2210, 2019.

[192] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240, 09 2019.

[193] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert : Modeling clinical notes and predicting hospital readmission, 2020.

[194] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip : Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3876–3887, 2022. PMID : 39144675 ; PMCID : PMC11323634.

[195] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis*, 67 :101851, 2021.

[196] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2 : Towards universal text-driven medical image segmentation. *Medical Image Analysis*, 106 :103749, 2025.

[197] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Lequan Yu, Yong Liang, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba† : Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024.

[198] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit : Language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1) :96–107, 2024.

[199] Pengyu Zhao, Yonghong Hou, Zhijun Yan, and Shuwei Huo. Text-driven medical image segmentation with text-free inference via knowledge distillation. *IEEE Transactions on Instrumentation and Measurement*, 74 :1–15, 2025.

[200] Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(4) :1821–1835, 2025.

[201] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg : Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21438–21451, October 2023.

[202] Sook-Lei Liew, Bethany P. Lo, Miranda R. Donnelly, and et al. Artemis Zavaliangos-Petropulu. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data*, 9(1) :320, 2022.

[203] Anas M. Tahir, Muhammad E.H. Chowdhury, and et al. Amith Khandakar. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139 :105002, 2021.

[204] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2 : Learning robust visual features without supervision, 2024.

[205] Yazhou Zhu, Shidong Wang, Tong Xin, Zheng Zhang, and Haofeng Zhang. Partition-a-medical-image : Extracting multiple representative subregions for few-shot medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 73 :1–12, 2024.

[206] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1) :654, 2024.

[207] Junde Wu and Min Xu. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11302–11312, June 2024.

[208] Ziming Cheng, Shidong Wang, Yang Long, Tao Zhou, Haofeng Zhang, and Ling Shao. Dual interspersion and flexible deployment for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(6) :2732–2744, 2025.