UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

SEGMENTATION SÉMANTIQUE GÉNÉRALISABLE DE SCÈNES DE DÉSASTRES ACQUISES PAR DRONES SOUS DÉSÉQUILIBRES DE CLASSES ET HÉTÉROGÉNÉITÉ INTER-DOMAINES

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

PAR

AHCENE ZETOUT

OCTOBRE 2026

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

Prof. Halim Boutayeb (UQO) ..... Président du jury

Prof. Etienne Gael Tajeuna (UQO) ..... Membre du jury interne

Prof. Mohamed Dahmane (UQTR) ..... Membre du jury externe

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

Thèse acceptée le : 21 octobre 2026

Remerciements

Avant toute chose, je remercie Dieu de m'avoir accordé la force et la patience nécessaires pour mener à bien ce parcours doctoral.

J'adresse mes sincères remerciements à mon directeur de recherche, le Pr. Mohand Said Allili, pour son encadrement, ses conseils, sa disponibilité et son soutien tout au long de cette thèse.

Je remercie profondément mon épouse pour son soutien inconditionnel et sa patience, ainsi que mes parents, mon frère et ma sœur pour leurs encouragements et leur présence constante.

Je remercie également les membres du jury pour le temps et l'expertise consacrés à l'évaluation de cette thèse.

Enfin, mes remerciements vont à l'Université du Québec en Outaouais, ainsi qu'à ma famille, mes collègues, mes amis et toutes les personnes qui m'ont soutenu, de près ou de loin, dans la réalisation de ce travail.

Table des matières

Liste des figures **v**

Liste des tableaux **viii**

Liste des abréviations et acronymes **x**

Résumé **xii**

Abstract **xiii**

1 Introduction Générale **1**

1.1 Contexte et problématique de recherche 1

1.2 Objectifs de la thèse 4

1.2.1 Objectif principal 4

1.2.2 Objectif secondaire 5

1.3 Contributions 5

1.3.1 Première contribution : segmentation guidée par la détection pour les classes rares et les petits objets 5

1.3.2 Seconde contribution : segmentation hiérarchique et adaptation inter-domaine avec taxonomies hétérogènes 7

1.4 Publications associées 8

1.5 Organisation de la thèse 9

2 Les fondements **10**

2.1 Segmentation sémantique 10

2.1.1 Architectures de segmentation 10

2.1.2 Fonctions de perte 12

2.1.3 Défis de la segmentation sémantique 12
2.2 Réseaux de neurones convolutifs (CNN) 13
2.2.1 Convolution 14
2.2.2 Pooling 16
2.2.3 Couches entièrement connectées 18
2.3 Vision Transformers (ViTs) 19
2.3.1 Patch Embedding 20
2.3.2 Self-Attention Mechanism 21
2.3.3 Transformer Encoder 21
2.4 Détection d'objets 22
2.4.1 Détecteurs en deux étapes 22
2.4.2 Détecteurs en une étape 23
2.5 Généralisation inter-domaines 25
2.5.1 Apprentissage de représentations invariantes 26
2.6 Représentations par prototypes et structuration hiérarchique 27
2.6.1 Alignement des représentations 28
2.6.2 Structuration hiérarchique 28
2.6.3 Avantages de l'approche hiérarchique 29

3 Etat de l'art **30**
3.1 Segmentation sémantique pour la compréhension des scènes de catastrophe . 30
3.1.1 Architectures encodeur–décodeur 30
3.1.2 Agrégation du contexte par convolutions dilatées et modules pyramidaux 33
3.2 Cadres de détection d'objets pour l'analyse de scènes de catastrophes 35
3.2.1 Architectures haute résolution et modèles hybrides basés sur les Transformers 37
3.3 Segmentation sémantique inter-domaine 41
3.3.1 Apprentissage avec taxonomies et étiquettes hétérogènes 42
3.3.2 Apprentissage de représentations sémantiques fondé sur les prototypes 43
3.3.3 Segmentation sémantique hiérarchique et conditionnelle 44
3.3.4 Travaux liés à l'adaptation inter-domaines et aux taxonomies hétérogènes 45

4 Première contribution **47**
4.1 Motivation et objectifs 47
4.2 Méthodologie proposée 49

4.2.1 Module d'amélioration des classes ciblées 49
4.2.2 Module d'attention contextuelle profonde 54
4.2.3 Module de fusion multi-échelle des caractéristiques 55
4.3 Entraînement du modèle, implémentation et fonction de perte 56
4.3.1 Stratégie d'entraînement et implémentation 56
4.3.2 Fonction de perte 57
4.4 Expérimentations et résultats 58
4.4.1 Jeux de données 59
4.4.2 Configuration d'entraînement 60
4.4.3 Métriques d'évaluation 61
4.4.4 Étude d'ablation 61
4.4.5 Résultats quantitatifs et analyse 64
4.4.6 Analyse computationnelle 68
4.5 Résultats qualitatifs 70
4.6 Discussion, limites et perspectives de la première contribution 77

5 Deuxième Contribution **80**
5.1 Méthodologie proposée 81
5.1.1 Vue d'ensemble de l'approche 81
5.1.2 Formulation hiérarchique de la segmentation 82
5.1.3 Têtes hiérarchiques et raffinement conditionnel 85
5.1.4 Transfert inter-domaine guidé par prototypes 87
5.1.5 Interaction des modules 90
5.2 Fonction de perte et stratégie d'entraînement 91
5.3 Expérimentation et résultats 94
5.3.1 Jeux de données 94
5.3.2 Configuration expérimentale 96
5.3.3 Métriques d'évaluation 97
5.3.4 Résultats quantitatifs 98
5.3.5 Étude d'ablation 100
5.3.6 Résultats qualitatifs 102
5.3.7 Discussion, limites et perspectives de la seconde contribution 108

6 Conclusion générale et perspectives **112**
6.1 Perspectives à long terme 112
6.2 Synthèse des contributions 113
6.3 Cohérence des travaux et apports scientifiques 113
6.4 Remarques finales 114

Bibliographie **114**

Liste des figures

2.1 Architecture U-Net 11
2.2 Architecture typique d'un réseau de neurones convolutif (CNN). 14
2.3 Illustration d'une opération de convolution 15
2.4 Illustration des opérations de pooling. 17
2.5 Illustration des convolutions dilatées (*atrous*). 18
2.6 Vue d'ensemble de l'architecture Vision Transformer (ViT). 20
2.7 Vue d'ensemble d'un pipeline typique de détection d'objets. 22
2.8 Vue d'ensemble de la famille des détecteurs R-CNN. 23
2.9 Illustration des détecteurs d'objets en une étape. 24
2.10 Vue d'ensemble de l'architecture RetinaNet. 25

4.1 Distributions des classes dans les jeux de données FloodNet (**a**) et RescueNet (**b**) utilisés dans ce travail. Elles illustrent le fort déséquilibre entre les catégories, certaines classes telles que les véhicules et les piscines étant fortement sous-représentées. 51
4.2 Architecture globale du modèle proposé : (1) module d'amélioration des classes ciblées, (2) module d'attention contextuelle profonde, et (3) module de fusion multi-échelle des caractéristiques. 52
4.3 Illustration du module Atrous Spatial Pyramid Pooling (ASPP). 53
4.4 Comparaison qualitative des résultats de segmentation obtenus avec différentes configurations d'ablation du modèle. Chaque colonne illustre un échantillon de test issu du jeu de données FloodNet. 65
4.5 Progression de l'IoU sur l'ensemble de validation au cours de 100 époques pour les classes "vehicle" et "pool" sur le jeu de données FloodNet. 70

4.6 Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données FloodNet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 1). 72
4.7 Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données FloodNet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 2). 73
4.8 Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données Rescuenet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 1). 74
4.9 Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données Rescuenet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 2). 75
4.10 Défis liés aux annotations. **(A)** correspond à un échantillon du jeu de données FloodNet ; **(B)** et **(C)** proviennent du jeu de données RescueNet. Les cercles rouges indiquent des erreurs d'annotation telles que des objets manquants ou des régions mal classifiées. 76

5.1 Vue d'ensemble du cadre proposé combinant une segmentation hiérarchique et un transfert inter-domaines guidé sur des prototypes. Le modèle prédit d'abord les super-classes partagées, puis affine conditionnellement les prédictions des sous-classes, tout en exploitant des prototypes guidés par le domaine source pour générer des pseudo-labels et assurer un alignement inter-domaines. 83
5.2 Comparaison qualitative de la segmentation des super-classes partagées dans le scénario de transfert RescueNet→FloodNet. Le cadre proposé produit des cartes de segmentation plus cohérentes spatialement et réduit le bruit de prédiction dans les régions complexes de bâtiments, routes et eau. 103

5.3 Comparaison qualitative de la segmentation des super-classes partagées dans le scénario de transfert FloodNet→RescueNet. Le cadre proposé améliore la cohérence des prédictions dans des scènes présentant des variations visuelles importantes et des taxonomies d'annotation différentes. 104
5.4 Comparaison qualitative verticale entre DeepLabV3-ResNet50 et la méthode proposée dans le scénario RescueNet→FloodNet. Chaque colonne correspond à un exemple distinct. L'ajout du cadre hiérarchique et des mécanismes de transfert permet de produire des cartes plus cohérentes spatialement, avec une meilleure délimitation des bâtiments, des routes et des zones d'eau. 106
5.5 Comparaison qualitative verticale entre DeepLabV3-ResNet50 et la méthode proposée dans le scénario FloodNet→RescueNet. Malgré une variabilité visuelle plus importante du domaine cible, la méthode proposée génère des prédictions plus régulières, réduit certaines régions incohérentes et améliore la localisation des structures principales. 107

Liste des tableaux

3.1 Positionnement comparatif de modèles représentatifs de segmentation sémantique et de travaux pertinents pour la segmentation d'images UAV, en lien avec la première contribution. CMS : contexte multi-échelle ; DS : détails spatiaux ; AR : attention ou raisonnement contextuel ; PCO : prise en compte des petits objets ; CR : classes rares ou sous-représentées ; GD : guidance par détection. ✓: explicitement traité ; △: partiellement traité ou indirectement lié ; ✗: non explicitement traité. 40

3.2 Positionnement comparatif des travaux représentatifs. DVS : décalage visuel entre domaines ou changement d'apparence ; DTS : décalage de taxonomie sémantique entre domaines source et cible ; SCP : modélisation de superclasses partagées ; AP : alignement fondé sur les prototypes ; PH : prédiction hiérarchique ; RC : raffinement conditionnel ; CPS : correspondance partielle entre sous-classes hétérogènes. ✓: explicitement traité ; △: partiellement traité ou indirectement lié ; ✗: non explicitement traité. 46

4.1 Étude d'ablation : impact de la suppression des composants clés sur l'IoU par classe (%) et la mIoU (%). 62

4.2 Comparaison de l'IoU par classe (%) entre différents modèles et classes thématiques sur l'ensemble de test FloodNet(BF: flooded building, BNF: non-flooded building, RF: flooded road, and RNF: non-flooded road). 66

4.3 Comparaison de l'IoU par classe (%) entre différents modèles et classes thématiques sur l'ensemble de test Rescuenet (BG: background, BND: building no damage, BMND: building minor damage, BMJD: building major damage, BTD: building total destruction, BR: blocked road and CR: clear road). . . . 67

4.4 Comparaison de l'efficacité des modèles : nombre de paramètres et temps d'inférence par image (sur un seul GPU NVIDIA A100). 69

5.1 Class-wise IoU (%) for RescueNet→FloodNet transfer. 98
5.2 Class-wise IoU (%) for FloodNet→RescueNet transfer. 99
5.3 Étude d'ablation sur les sous-classes (mIoU %). 101

Liste des abréviations et acronymes

AI Artificial Intelligence

ASPP Atrous Spatial Pyramid Pooling

CMTFNet CNN and Multiscale Transformer Fusion Network

CNN Convolutional Neural Network

CSDNet Context-Aware Segmentation with Detection-Guided Features and Lightweight Transformers

DL Deep Learning

ENet Efficient Neural Network for Real-Time Segmentation

Fast R-CNN Improved Region-based Convolutional Neural Network

Faster R-CNN Region-based Convolutional Neural Network with Region Proposal Network

FCN Fully Convolutional Network

GAN Generative Adversarial Network

GPU Graphics Processing Unit

HRNet High-Resolution Network

IoU Intersection over Union

LLM Large Language Model

Mask R-CNN Region-based Convolutional Neural Network extended to segmentation

mIoU mean Intersection over Union

ML Machine Learning

PSPNet Pyramid Scene Parsing Network

R-CNN Region-based Convolutional Neural Network

ReLU Rectified Linear Unit

RetinaNet RetinaNet with Focal Loss

RNN Recurrent Neural Network

RS Remote Sensing

SE Squeeze-and-Excitation

SegFormer Segmentation Transformer

SSD Single Shot MultiBox Detector

Swin-Unet Swin Transformer-based U-Net

TransUNet Transformer-based U-Net

ViT Vision Transformer

VLM Vision-Language Model

YOLO You Only Look Once

Résumé

La segmentation sémantique d'images aériennes à haute résolution constitue une tâche fondamentale en vision par ordinateur, avec des applications importantes en cartographie, surveillance environnementale et gestion des catastrophes naturelles. Dans les contextes post-catastrophe, une compréhension précise des scènes est essentielle pour identifier les zones affectées, évaluer les dommages aux infrastructures et soutenir les interventions d'urgence. Malgré les avancées récentes de l'apprentissage profond, cette tâche demeure difficile en raison du déséquilibre des classes, de la présence d'objets de petite taille, de la similarité visuelle entre certaines catégories et des variations importantes entre jeux de données.

Cette thèse s'intéresse à la segmentation sémantique d'images aériennes dans des conditions réelles en proposant des approches robustes adaptées à ces défis. La première contribution concerne un cadre supervisé mono-domaine, où une architecture exploitant des informations issues de la détection d'objets est développée afin d'améliorer la reconnaissance des classes sous-représentées. Cette approche combine signaux guidés par la détection, fusion multi-échelle et mécanismes d'attention légers pour renforcer la segmentation des catégories difficiles. La seconde contribution aborde le problème plus général du transfert inter-domaine sous taxonomies hétérogènes. Une reformulation hiérarchique est proposée, fondée sur des super-classes sémantiques partagées et des sous-classes spécifiques à chaque domaine, complétée par un raffinement conditionnel et un transfert sémantique guidé par des prototypes. Les résultats expérimentaux montrent que cette approche demeure compétitive face aux méthodes de référence en robustesse visuelle et adaptation inter-domaine, tout en traitant un cadre plus réaliste caractérisé par une supervision asymétrique et des taxonomies partiellement incompatibles. Au-delà des performances obtenues, cette thèse met en évidence que la robustesse d'un système de segmentation dépend non seulement de l'architecture utilisée, mais aussi de la manière dont les connaissances sémantiques sont structurées et transférées. Elle ouvre aussi la voie à des méthodes de segmentation plus adaptées à la diversité des données rencontrées dans les applications critiques de télédétection.

Abstract

Semantic segmentation of high-resolution aerial imagery is a fundamental task in computer vision, with important applications in mapping, environmental monitoring, and natural disaster management. In post-disaster contexts, accurate scene understanding is essential for identifying affected areas, assessing infrastructure damage, and supporting emergency response operations. Despite recent advances in deep learning, this task remains challenging because of class imbalance, the presence of small objects, visual similarities between certain categories, and substantial variations across datasets.

This thesis investigates semantic segmentation of aerial imagery under real-world conditions by proposing robust approaches designed to address these challenges. The first contribution focuses on a supervised single-domain setting, in which an architecture leveraging information from object detection is developed to improve the recognition of under-represented classes. This approach combines detection-guided signals, multi-scale feature fusion, and lightweight attention mechanisms to strengthen the segmentation of challenging categories. The second contribution addresses the more general problem of cross-domain transfer under heterogeneous taxonomies. A hierarchical reformulation is proposed based on shared semantic superclasses and domain-specific subclasses, complemented by conditional refinement and prototype-guided semantic transfer. Experimental results show that the proposed approach remains competitive with reference methods for visual robustness and domain adaptation while addressing a more realistic setting characterized by asymmetric supervision and partially incompatible taxonomies. Beyond the achieved performance, this thesis highlights that the robustness of a segmentation system depends not only on the architecture employed, but also on how semantic knowledge is structured and transferred. It therefore opens the way toward segmentation methods that are better suited to the diversity of data encountered in critical remote-sensing applications.

Chapitre 1

Introduction Générale

1.1 Contexte et problématique de recherche

L'interprétation des scènes visuelles constitue aujourd'hui l'un des objectifs centraux de la vision par ordinateur moderne. Parmi les différentes tâches associées à cette problématique, la segmentation sémantique occupe une place particulièrement importante puisqu'elle vise à produire une cartographie dense de l'image en attribuant une étiquette à chaque pixel [1, 2, 3]. Contrairement à la classification d'images, qui fournit uniquement une description globale de la scène, ou à la détection d'objets, qui se limite généralement à la localisation approximative d'instances, la segmentation sémantique cherche à reconstruire une représentation structurée et détaillée de l'environnement observé.

Cette capacité de compréhension fine joue désormais un rôle essentiel dans de nombreux domaines nécessitant une interprétation précise des structures visuelles, notamment la conduite autonome, l'analyse urbaine, la surveillance environnementale, la cartographie intelligente et l'évaluation des dommages après catastrophe. Dans ce contexte, les images acquises par drones (*Unmanned Aerial Vehicles*, UAV) ont suscité un intérêt croissant au cours des dernières années en raison de leur flexibilité d'acquisition, de leur faible coût opérationnel et de leur très haute résolution spatiale [4, 5]. Contrairement aux images satellitaires traditionnelles, souvent contraintes par leur résolution ou leur fréquence de revisite, les plateformes UAV permettent une observation rapide, localisée et détaillée des scènes complexes. Cette richesse visuelle ouvre la voie à des applications critiques dans les scénarios réels, en particulier dans les contextes post-catastrophe où une analyse rapide des infrastructures, des routes, des bâtiments endommagés ou des zones inondées peut soutenir les opérations

d'urgence et améliorer la prise de décision. Toutefois, cette même richesse visuelle introduit également des difficultés majeures pour les systèmes de segmentation automatique.

Les scènes UAV présentent généralement une forte variabilité d'échelle, de perspective, d'illumination et de texture [6]. Les objets peuvent apparaître sous des angles variés, être partiellement occlus, fortement déformés ou n'occuper qu'une portion extrêmement réduite de l'image. Dans les environnements post-catastrophe, ces difficultés sont encore amplifiées par la présence de débris, de structures dégradées et de modifications importantes de l'apparence des scènes. Ainsi, malgré les progrès récents de l'apprentissage profond, la segmentation sémantique d'images UAV réelles demeure un problème particulièrement difficile lorsque plusieurs contraintes visuelles et sémantiques apparaissent simultanément [5].

Les avancées récentes en apprentissage profond ont transformé le domaine de la segmentation sémantique. Les architectures entièrement convolutives (*Fully Convolutional Networks*, FCN) ont démontré la possibilité d'effectuer une segmentation dense de bout en bout [1]. Les architectures encodeur–décodeur, telles que U-Net, ont ensuite amélioré la préservation des détails spatiaux grâce aux connexions de saut [2]. Les modèles de la famille DeepLab ont introduit les convolutions dilatées ainsi que les modules d'*Atrous Spatial Pyramid Pooling* (ASPP) afin de mieux capturer le contexte multi-échelle [3]. Plus récemment, les architectures basées sur les Transformers, telles que ViT et SegFormer, ont permis une modélisation plus efficace des dépendances globales et des relations spatiales à longue portée [7, 8].

Malgré ces avancées, de bonnes performances globales sur les benchmarks standards ne garantissent pas des performances homogènes pour toutes les catégories, en particulier dans les scènes UAV réelles et complexes. En pratique, les difficultés de segmentation n'apparaissent pas de manière isolée, mais résultent souvent de l'interaction simultanée de plusieurs contraintes visuelles, spatiales et sémantiques. Une première difficulté majeure concerne le déséquilibre des classes ainsi que la segmentation des objets de petite taille [9, 10]. Dans les images UAV haute résolution, certaines catégories critiques n'occupent qu'une portion extrêmement réduite de l'image comparativement aux classes dominantes. Les modèles profonds tendent alors naturellement à privilégier les catégories majoritaires, ce qui réduit fortement la qualité des représentations apprises pour les classes rares. Cette difficulté est accentuée par les mécanismes de sous-échantillonnage successifs utilisés dans les architectures profondes. Bien qu'ils permettent d'augmenter le champ réceptif et d'extraire des représentations sémantiques plus riches, ils entraînent également une perte progressive d'informations spatiales fines [10]. Les petits objets, les structures fragmentées et les frontières complexes deviennent alors particulièrement difficiles à localiser avec précision.

Au-delà du déséquilibre des classes, les scènes post-catastrophe introduisent également de fortes ambiguïtés visuelles entre catégories proches [11, 12]. Une route inondée peut partager des caractéristiques visuelles similaires à une surface d'eau naturelle ; un bâtiment partiellement endommagé peut conserver l'apparence générale d'une structure intacte ; certaines régions de débris peuvent présenter des textures proches du sol ou de la végétation. Dans ce contexte, les modèles de segmentation classiques peinent souvent à maintenir des prédictions cohérentes dans les régions ambiguës, particulièrement lorsque les catégories concernées sont faiblement représentées dans les données d'apprentissage [5].

Ces limitations montrent que la robustesse d'un système de segmentation ne dépend pas uniquement de sa capacité à apprendre des représentations globales performantes, mais également de sa capacité à préserver et renforcer les informations locales critiques associées aux structures rares, ambiguës ou spatialement limitées. Toutefois, même lorsque ces difficultés intra-domaines sont partiellement maîtrisées, une limitation plus fondamentale demeure : la faible capacité de généralisation des modèles face aux changements de distribution des données.

Dans les applications réelles de télédétection, les modèles sont rarement déployés dans des conditions strictement identiques à celles rencontrées durant l'apprentissage [13, 14]. Les images peuvent provenir de régions géographiques différentes, être acquises par des capteurs distincts, présenter des conditions climatiques variées ou correspondre à des scénarios de catastrophe différents. Cette différence entre les distributions d'apprentissage et de test, généralement désignée sous le terme de *domain shift*, entraîne souvent une dégradation importante des performances des modèles [14, 15].

Les approches récentes de segmentation inter-domaine, incluant l'adaptation de domaine et la généralisation de domaine, cherchent précisément à améliorer cette robustesse en réduisant l'effet des décalages de distribution entre domaines [16, 17, 14]. Plusieurs stratégies ont été proposées dans la littérature, notamment les mécanismes de normalisation adaptés aux domaines, les perturbations fréquentielles, les techniques de diversification de styles, l'alignement de caractéristiques ou encore l'utilisation de pseudo-labels [18, 19, 20, 21]. Bien que ces approches améliorent souvent la robustesse face aux variations d'apparence, elles reposent généralement sur une hypothèse simplificatrice importante : les domaines source et cible partagent un espace sémantique identique ou directement compatible [22, 23].

1.2 Objectifs de la thèse

1.2.1 Objectif principal

L'objectif principal de cette thèse est de développer des approches de segmentation sémantique robustes pour les images aériennes UAV acquises dans des conditions réelles, en particulier dans les scénarios de catastrophe naturelle caractérisés par une forte variabilité visuelle, des classes déséquilibrées et des distributions de données hétérogènes. Plus précisément, cette recherche cherche à répondre à la question suivante :

> *Comment concevoir des architectures de segmentation capables de maintenir des performances robustes face aux classes rares, aux petits objets, aux ambiguïtés visuelles et aux variations inter-domaines dans les images UAV réelles ?*

Afin de répondre à cette problématique, cette thèse suit une progression méthodologique structurée en deux niveaux complémentaires. Dans un premier temps, nous nous intéressons à la robustesse intra-domaine des modèles de segmentation. L'objectif est d'améliorer la représentation des catégories difficiles à segmenter dans un même domaine de données, notamment les classes rares, les petits objets et les structures ambiguës présentes dans les scènes post-catastrophe. Cette première étape conduit à explorer des stratégies de fusion entre segmentation et détection afin d'enrichir les représentations utilisées par le réseau de segmentation et de mieux préserver les informations locales importantes.

Dans un second temps, nous étendons cette réflexion vers la robustesse inter-domaines. Plus précisément, nous cherchons à développer un cadre d'adaptation inter-domaine capable de transférer les connaissances entre plusieurs jeux de données UAV présentant non seulement des différences visuelles importantes, mais également des taxonomies d'annotation partiellement incompatibles. Cet objectif nécessite de dépasser les approches classiques centrées uniquement sur l'invariance visuelle des représentations, afin de prendre également en compte les relations sémantiques entre classes partagées et attributs spécifiques aux domaines.

Dans cette perspective, cette thèse explore une structuration hiérarchique de l'espace sémantique reposant sur des super-classes partagées entre domaines et des sous-classes spécifiques à chaque dataset. Nous étudions également l'utilisation de prototypes sémantiques et de mécanismes de raffinement conditionnel afin de mieux transférer les connaissances entre domaines tout en préservant les spécificités propres à chaque jeu de données. Ainsi, cette thèse poursuit une progression allant de l'amélioration locale des représentations de segmentation vers le développement d'un cadre plus général de segmentation hiérarchique

et d'adaptation inter-domaine adapté aux contraintes réelles des données UAV et aux taxonomies hétérogènes.

1.2.2 Objectif secondaire

Au-delà de l'amélioration des performances de segmentation, cette thèse cherche également à maintenir un équilibre entre robustesse, complexité architecturale et transférabilité des modèles. Les images UAV haute résolution introduisent des contraintes importantes en mémoire et en temps de calcul, particulièrement lorsque plusieurs mécanismes multi-échelles, modules attentionnels ou stratégies de transfert inter-domaines sont combinés. Dans ce contexte, l'objectif n'est pas uniquement d'augmenter les performances, mais également de développer des architectures capables d'exploiter efficacement les informations pertinentes sans introduire une complexité computationnelle excessive. Cette recherche vise ainsi à proposer des cadres de segmentation suffisamment robustes pour traiter des scénarios UAV complexes, tout en conservant une faisabilité pratique pour des applications réelles de télédétection et d'analyse post-catastrophe.

1.3 Contributions

Cette thèse propose deux contributions principales s'inscrivant dans une progression méthodologique cohérente visant à améliorer progressivement la robustesse des systèmes de segmentation sémantique pour les images aériennes UAV. La première contribution traite principalement les limitations intra-domaines liées aux classes rares, aux petits objets et aux ambiguïtés locales, tandis que la seconde étend cette réflexion vers l'adaptation inter-domaine et la gestion de taxonomies d'annotation hétérogènes.

Contrairement à de nombreux travaux qui abordent séparément les problèmes de segmentation des petits objets, de déséquilibre des classes ou de robustesse inter-domaine, les contributions proposées dans cette thèse cherchent à traiter conjointement plusieurs contraintes réelles de la segmentation UAV.

1.3.1 Première contribution : segmentation guidée par la détection pour les classes rares et les petits objets

La première contribution de cette thèse porte sur l'amélioration de la robustesse intra-domaine de la segmentation sémantique dans les images aériennes UAV post-catastrophe.

Les architectures modernes de segmentation, telles que DeepLabV3+ [24], PSPNet [25], OCRNet [26] ou SegFormer [8], ont démontré des performances remarquables grâce à leur capacité à modéliser efficacement le contexte spatial et multi-échelle [3, 25, 26, 8]. Les architectures basées sur les Transformers ont également amélioré la modélisation des dépendances globales dans les images [7]. Malgré ces avancées, plusieurs travaux montrent que les petits objets et les classes rares demeurent particulièrement difficiles à segmenter dans les images aériennes haute résolution [4, 27].

Dans les scènes UAV post-catastrophe naturelles, certaines catégories critiques telles que les véhicules, piscines, routes fragmentées ou structures endommagées occupent souvent une portion très réduite de l'image. Les représentations apprises par les architectures classiques restent alors dominées par les classes majoritaires, ce qui réduit fortement la qualité des prédictions sur les catégories minoritaires [10]. De plus, les mécanismes de sous-échantillonnage successifs utilisés dans les architectures profondes peuvent entraîner une perte importante d'information spatiale, affectant directement la localisation des petits objets et des frontières fines.

Afin de répondre à ces limitations, nous proposons une architecture de segmentation guidée par la détection appelée *CSDNet* (*Context-Aware Segmentation of Disaster Aerial Imagery Using Detection-Guided Features and Lightweight Transformers*). L'idée principale de cette contribution est d'exploiter les capacités de localisation des modèles de détection afin d'enrichir les représentations utilisées par le réseau de segmentation. Contrairement aux approches de segmentation classiques reposant uniquement sur une supervision dense pixel par pixel, notre méthode introduit des indices guidés par la détection afin de renforcer les régions associées aux catégories difficiles. Cette stratégie permet d'orienter plus efficacement l'apprentissage vers les classes sous-représentées et les structures de petite taille. Le cadre proposé combine plusieurs composantes complémentaires : un backbone EfficientNet-B5 pour l'extraction de représentations multi-échelles riches, des caractéristiques guidées par la détection afin de renforcer les catégories rares, des mécanismes de fusion multi-échelle pour préserver simultanément les détails locaux et le contexte global, ainsi que des modules attentionnels légers inspirés des Transformers afin d'améliorer la modélisation des dépendances spatiales longues portées.

Cette contribution se distingue des travaux existants à plusieurs niveaux. Premièrement, elle introduit une interaction explicite entre segmentation et détection dans un contexte UAV post-catastrophe où les classes rares jouent un rôle opérationnel important. Deuxièmement, elle combine efficacement des représentations locales guidées par la détection avec des mé-

canismes de contexte global inspirés des Transformers. Enfin, elle conserve une complexité relativement modérée comparativement à certaines architectures hybrides plus coûteuses. Les résultats expérimentaux obtenus sur FloodNet montrent que cette approche améliore significativement la segmentation des catégories minoritaires tout en maintenant de bonnes performances globales sur les autres classes. Cette contribution a conduit à la publication du modèle CSDNet dans la revue *Remote Sensing*.

1.3.2 Seconde contribution : segmentation hiérarchique et adaptation inter-domaine avec taxonomies hétérogènes

La seconde contribution de cette thèse étend la réflexion précédente vers un problème plus général : la robustesse inter-domaines des modèles de segmentation sémantique.

Les approches récentes de segmentation inter-domaine cherchent à améliorer la robustesse des modèles face aux changements de distribution en apprenant des représentations moins sensibles aux variations visuelles [15, 14]. Plusieurs stratégies ont été proposées dans la littérature, notamment les mécanismes de normalisation adaptés aux domaines, les perturbations fréquentielles, les stratégies de diversification de styles, l'alignement de caractéristiques ou encore le pseudo-étiquetage [28, 29, 30]. Bien que ces approches aient démontré leur efficacité dans différents contextes de vision par ordinateur, elles supposent généralement que les domaines source et cible partagent un espace de classes identique ou directement compatible [31, 23, 32]. Cette hypothèse n'est cependant pas toujours vérifiée dans les scénarios réels de segmentation UAV. Dans les jeux de données de désastres, les bâtiments peuvent être annotés selon leur état d'inondation ou leur niveau de dommage, tandis que les routes sont décrites par leur accessibilité, leur obstruction ou leur état d'inondation. Cette hétérogénéité sémantique, encore peu étudiée, complique fortement l'apprentissage inter-domaine. De plus, restreindre l'apprentissage aux seules classes communes entraîne une perte importante d'information sémantique.

Afin de répondre à cette problématique, nous proposons un cadre hiérarchique d'adaptation inter-domaine guidé par prototypes, reposant sur la distinction entre super-classes partagées et sous-classes spécifiques à chaque domaine. L'idée centrale de cette contribution est que différents datasets peuvent partager des concepts sémantiques généraux communs tout en possédant des attributs ou états spécifiques propres à chaque domaine. Des classes telles que *building*, *road*, *water*, *tree* ou *vehicle* peuvent ainsi être considérées comme des super-classes

communes, tandis que les états d'inondation ou les niveaux de dommages représentent des sous-classes spécifiques.

Cette structuration hiérarchique reformule la segmentation comme une suite de décisions sémantiques cohérentes : le modèle identifie d'abord une catégorie générale partagée, puis affine sa prédiction vers des attributs spécifiques au domaine. Le cadre proposé combine un espace de super-classes partagé, des têtes hiérarchiques dédiées aux sous-classes propres à chaque jeu de données, des prototypes sémantiques pour faciliter le transfert inter-domaine et un mécanisme de raffinement conditionnel assurant la cohérence entre prédictions fines et catégories parentes. Les prototypes sont utilisés à deux niveaux : toutes les super-classes partagées participent au transfert, tandis que seules les sous-classes présentant une correspondance sémantique pertinente sont alignées.

Cette contribution se distingue des approches classiques de robustesse visuelle et d'adaptation inter-domaine à plusieurs niveaux. Premièrement, elle traite explicitement les différences de taxonomies d'annotation et non uniquement les variations visuelles entre domaines. Deuxièmement, elle introduit une structuration hiérarchique de l'espace sémantique afin de préserver simultanément les connaissances partagées et les informations spécifiques à chaque dataset. Troisièmement, elle montre qu'il est possible d'exploiter des annotations fines hétérogènes sans imposer une correspondance exacte entre toutes les classes des différents domaines.

Les expérimentations réalisées sur les jeux de données FloodNet [4] et RescueNet [5] montrent que cette approche demeure compétitive face aux approches classiques de robustesse visuelle et de transfert inter-domaine, tout en traitant un problème plus général et plus réaliste caractérisé par des taxonomies partiellement incompatibles et une supervision asymétrique. Plus largement, cette contribution montre que la robustesse inter-domaine ne dépend pas uniquement de l'invariance visuelle des représentations, mais également de la manière dont les relations sémantiques entre super-classes partagées et sous-classes spécifiques aux domaines sont structurées et exploitées pendant l'apprentissage.

1.4 Publications associées

Les travaux réalisés dans le cadre de cette thèse ont conduit aux publications scientifiques suivantes :

- **Ahcene Zetout**, Mohand Said Allili. *CSDNet: Context-Aware Segmentation of Disaster Aerial Imagery Using Detection-Guided Features and Lightweight Transformers. Remote Sensing*, 2025.

– **Ahcene Zetout**, Mohand Said Allili. *HPDA : Hierarchical Prototype-Based Domain Adaptation for Semantic Segmentation under Heterogeneous UAV Taxonomies.* Manuscrit en préparation / soumis.

La première publication correspond à la contribution portant sur l'amélioration de la segmentation intra-domaine des classes rares et des petits objets à travers une architecture guidée par la détection et enrichie par des mécanismes de contexte multi-échelle.

La seconde publication étend cette réflexion vers l'adaptation inter-domaine dans des scénarios caractérisés par des taxonomies d'annotation hétérogènes, en introduisant une structuration hiérarchique des classes, un raffinement conditionnel ainsi que des mécanismes de transfert guidés par prototypes sémantiques.

1.5 Organisation de la thèse

Le reste de cette thèse est organisé comme suit :

– Le Chapitre 2 présente l'état de l'art relatif à la segmentation sémantique d'images aériennes, aux architectures profondes de segmentation, aux approches guidées par l'attention ainsi qu'aux méthodes récentes d'adaptation inter-domaine, de généralisation de domaine et de segmentation hiérarchique.

– Le Chapitre 3 présente la première contribution de cette thèse, consacrée à l'architecture CSDNet pour la segmentation guidée par la détection dans les images UAV post-catastrophe.

– Le Chapitre 4 présente la seconde contribution, centrée sur l'adaptation inter-domaine hiérarchique avec taxonomies hétérogènes et mécanismes de transfert guidés par prototypes sémantiques.

– Le Chapitre 5 présente une discussion générale des contributions, des limites et des perspectives ouvertes par les travaux réalisés dans cette thèse.

– Enfin, le Chapitre 6 conclut cette thèse en résumant les principales contributions et en présentant plusieurs perspectives de recherche futures.

Chapitre 2

Les fondements

2.1 Segmentation sémantique

La segmentation sémantique est une tâche qui consiste à attribuer une étiquette de classe à chaque pixel d'une image. Contrairement aux tâches de classification, qui attribuent une seule étiquette à l'image entière, ou à la détection d'objets, qui localise des instances à l'aide de boîtes englobantes, la segmentation sémantique permet une compréhension dense et détaillée de la scène.

Formellement, étant donnée une image d'entrée $X \in R^{H \times W \times C}$, l'objectif est de prédire une carte de segmentation $Y \in \{1, \ldots, K\}^{H \times W}$, où chaque pixel (i, j) est associé à une classe parmi K catégories :

$$Y(i,j) = \arg \max_{k \in \{1,\ldots,K\}} P(y_{i,j} = k \mid X) \tag{2.1}$$

où $P(y_{i,j} = k \mid X)$ représente la probabilité que le pixel (i, j) appartienne à la classe k.

Cette tâche peut être vue comme un problème de classification multi-classe au niveau du pixel, nécessitant à la fois l'extraction de caractéristiques locales fines et la prise en compte du contexte global de l'image [1, 3].

2.1.1 Architectures de segmentation

Les architectures modernes de segmentation sémantique reposent généralement sur des structures de type encodeur-décodeur. L'encodeur est chargé d'extraire des caractéristiques

hiérarchiques en réduisant progressivement la résolution spatiale, tandis que le décodeur reconstruit une représentation dense afin de produire une prédiction pixel par pixel.

Parmi les modèles les plus influents, on peut citer :

- **Fully Convolutional Networks (FCN)** [1] : remplacent les couches entièrement connectées par des convolutions afin de produire des cartes de sortie de même taille que l'entrée.
- **U-Net** [2] : introduit des connexions de saut (*skip connections*) entre l'encodeur et le décodeur afin de préserver les détails spatiaux.
- **DeepLab** [3] : utilise des convolutions dilatées (*atrous*) et des modules ASPP pour capturer le contexte multi-échelle.

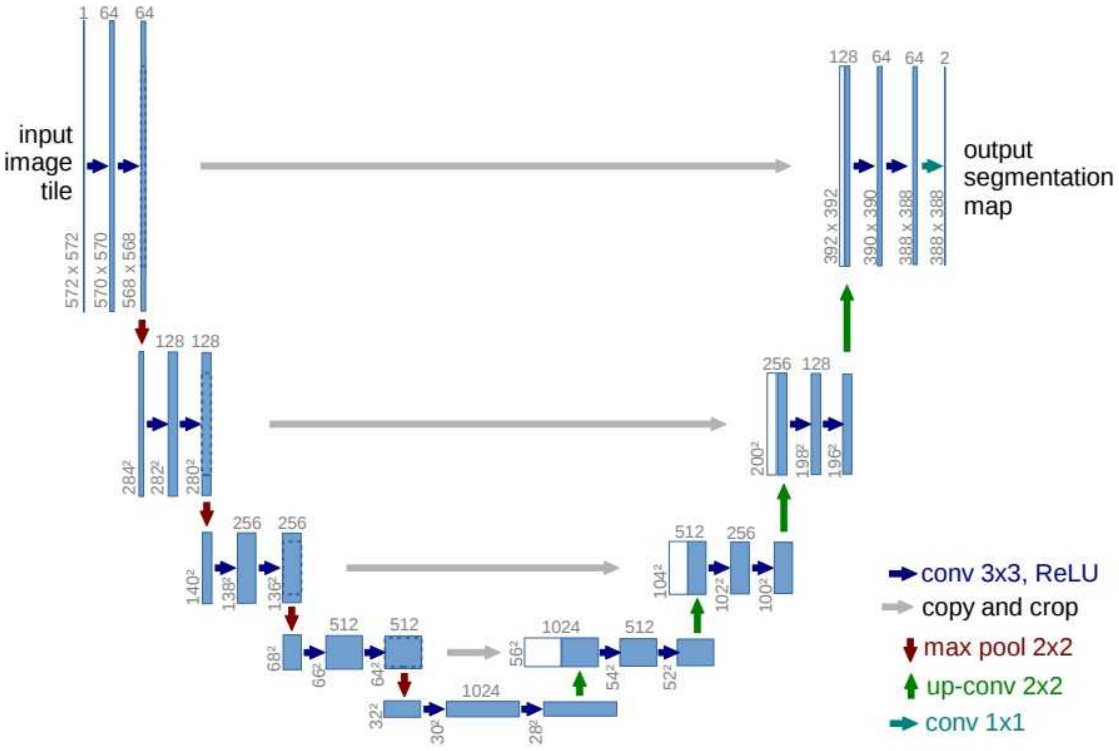


Figure 2.1: Architecture de type encodeur-décodeur illustrée par U-Net. Les connexions de saut permettent de combiner des caractéristiques de bas niveau et de haut niveau afin d'améliorer la précision de la segmentation (tiré de [2]).

2.1.2 Fonctions de perte

L'apprentissage des modèles de segmentation repose généralement sur des fonctions de perte adaptées à la classification dense. La plus courante est l'entropie croisée :

$$\mathcal{L}_{CE} = -\sum_{i,j}\sum_{k=1}^{K} I\left[y_{i,j} = k\right] \log p_{i,j}^{(k)} \tag{2.2}$$

où $y_{i,j} \in \{1, \ldots, K\}$ désigne l'étiquette réelle du pixel (i, j), $I[y_{i,j} = k]$ vaut 1 si le pixel appartient à la classe k et 0 sinon, et $p_{i,j}^{(k)}$ représente la probabilité prédite pour la classe k.

D'autres fonctions de perte ont été proposées pour mieux gérer le déséquilibre des classes, comme la Dice loss ou la Jaccard loss, qui mesurent directement le recouvrement entre la prédiction et la vérité terrain :

$$\mathcal{L}_{Dice} = 1 - \frac{1}{K}\sum_{k=1}^{K} \frac{2\sum_{i,j} p_{i,j}^{(k)} y_{i,j}^{(k)} + \epsilon}{\sum_{i,j} p_{i,j}^{(k)} + \sum_{i,j} y_{i,j}^{(k)} + \epsilon} \tag{2.3}$$

où K désigne le nombre de classes, $p_{i,j}^{(k)}$ la probabilité prédite pour la classe k, $y_{i,j}^{(k)} \in \{0, 1\}$ l'indicateur de vérité terrain correspondant, et ϵ une constante utilisée pour assurer la stabilité numérique. La Dice loss est ainsi calculée séparément pour chaque classe puis moyennée sur l'ensemble des classes.

Ces fonctions sont particulièrement utiles dans les scénarios où certaines classes occupent une faible proportion de l'image.

2.1.3 Défis de la segmentation sémantique

Malgré les progrès récents, la segmentation sémantique reste un problème difficile, en particulier dans les images aériennes. Les principaux défis incluent :

- **Déséquilibre des classes :** certaines classes rares sont sous-représentées.
- **Objets de petite taille :** difficiles à détecter et segmenter précisément.
- **Variabilité des scènes :** changements d'échelle, d'illumination et de contexte.
- **Ambiguïté visuelle :** classes similaires difficiles à distinguer.

Ces défis motivent le développement de nouvelles approches capables de combiner des informations locales fines et un contexte global riche, tout en étant robustes aux variations de distribution des données.

2.2 Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN) constituent le cœur de l'apprentissage profond. Ils sont composés de couches de neurones, organisées sous la forme d'une couche d'entrée suivie d'une ou plusieurs couches intermédiaires, et enfin d'une couche de sortie. Les neurones sont connectés entre eux par des liens, et chaque neurone possède un poids.

Lors de l'utilisation de ces réseaux sur des images, lorsqu'on atteint une couche convolutionnelle, un ensemble de neurones interconnectés forme ce que l'on appelle un filtre sous forme de matrice, où chaque valeur représente un poids qui sera appliqué à l'image afin d'extraire des caractéristiques. Ces caractéristiques sont appelées cartes de caractéristiques (*feature maps*). Elles sont extraites à différents niveaux correspondant aux couches du réseau. Les premières cartes de caractéristiques, issues des premières couches, sont appelées caractéristiques de bas niveau, car elles capturent généralement des éléments simples tels que les contours, les angles et les textures. À l'inverse, celles des couches profondes sont appelées caractéristiques de haut niveau, car elles permettent d'extraire des informations plus complexes, telles que des parties d'objets ou des objets complets [33]. Selon la tâche à accomplir, certaines caractéristiques peuvent être plus importantes que d'autres. Par exemple, dans le cas de la classification, les caractéristiques de haut niveau peuvent être plus représentatives que les caractéristiques de bas niveau. En revanche, pour la segmentation, comme démontré par U-Net, les premières couches sont aussi importantes que les dernières.

Bien que les réseaux convolutionnels soient largement utilisés pour la classification et les tâches de vision par ordinateur, il existe différents types de réseaux neuronaux adaptés à divers cas d'usage et types de données. Par exemple, les réseaux de neurones récurrents sont couramment utilisés pour le traitement du langage naturel et la reconnaissance vocale. Cependant, les réseaux de neurones convolutifs constituent aujourd'hui une approche efficace pour la classification d'images, la détection d'objets, la segmentation et d'autres tâches de vision, en s'appuyant notamment sur des opérations d'algèbre linéaire pour extraire des motifs à partir des images. Néanmoins, ces modèles peuvent être coûteux en calcul, nécessitant souvent l'utilisation de processeurs graphiques (GPU) pour leur entraînement. Les CNN se distinguent des autres réseaux neuronaux par leurs performances supérieures sur les données visuelles. Toutefois, les convolutions ne sont pas les seuls éléments du réseau : d'autres types de couches contribuent également à affiner les cartes de caractéristiques.

La structure globale d'un CNN est illustrée dans la Figure 2.2. Un CNN est généralement composé de trois types principaux de couches :

– **Couche convolutionnelle**

– **Couche de pooling**

– **Couche entièrement connectée (Fully Connected)**

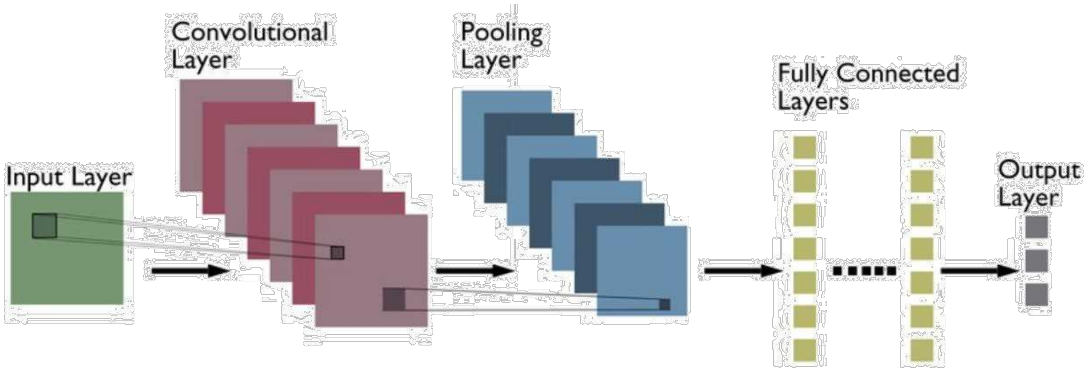


Figure 2.2: Architecture typique d'un réseau de neurones convolutif (CNN). Elle consiste en une succession de couches convolutionnelles et de pooling pour l'extraction de caractéristiques, suivies de couches entièrement connectées pour la classification ou la prédiction (tiré de [33]).

2.2.1 Convolution

Les couches convolutionnelles sont les composants principaux des CNN, où s'effectue la majorité des calculs. Elles sont conçues pour extraire des informations spatiales à partir des données d'entrée. L'entrée est généralement une image 2D, considérée comme un volume 3D en raison de la profondeur de couleur RGB.

Le processus consiste à appliquer des filtres entraînables, appelés noyaux (*kernels*), sur des régions locales de l'image afin d'extraire des caractéristiques, qui sont des représentations de motifs tels que les contours, les textures et des structures plus complexes. Chaque filtre est une petite matrice 2D de poids (généralement de taille 3×3) qui calcule un produit scalaire avec la région d'entrée correspondante, produisant ainsi une carte de caractéristiques (ou carte d'activation). Ces cartes forment des représentations hiérarchiques : à mesure que l'on progresse dans le réseau, les cartes extraites à une couche donnée deviennent l'entrée de la couche suivante. C'est pourquoi les premières couches capturent des caractéristiques de bas niveau, tandis que les couches plus profondes, dites de haut niveau, identifient des structures plus abstraites. L'opération de convolution entre une entrée X et un noyau W peut être

exprimée comme suit :

$$Y(i,j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X(i+m, j+n) \cdot W(m,n) + b \tag{2.4}$$

où k est la taille du filtre (généralement un carré $k \times k$) et b est le terme de biais (Figure 2.3. Cette opération est appliquée à l'aide d'un mécanisme de fenêtre glissante, où chaque région locale de l'entrée est multipliée élément par élément avec les poids du filtre, puis sommée afin de produire la valeur de sortie correspondante. Ce processus permet de préserver les relations spatiales dans les données tout en réduisant considérablement le nombre de paramètres par rapport aux couches entièrement connectées.

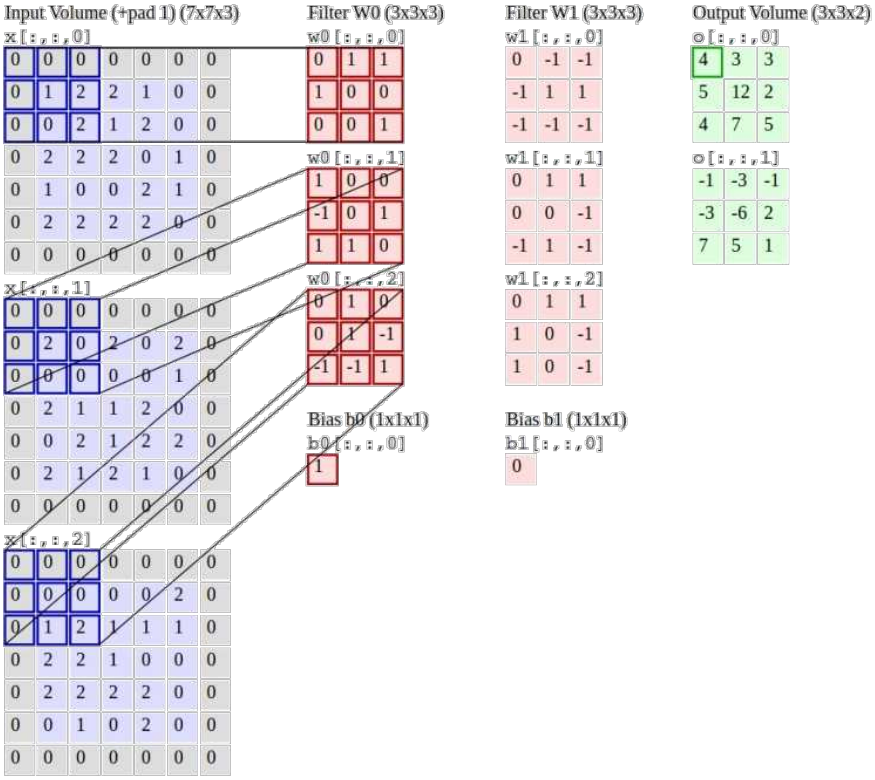


Figure 2.3: Illustration d'une opération de convolution : un filtre (noyau) se déplace sur l'image d'entrée pour produire une carte de caractéristiques (tiré de [34]).

La convolution repose sur le principe du partage de paramètres, ce qui signifie que les mêmes poids du filtre sont appliqués à l'ensemble de l'entrée. Ces poids sont optimisés lors de l'entraînement par propagation en arrière, en utilisant une fonction de perte qui

mesure l'écart entre les prédictions du modèle et les valeurs réelles. La conception d'un bloc convolutionnel nécessite un réglage soigneux de plusieurs hyperparamètres, notamment :

- **Nombre de filtres :** détermine la profondeur des cartes de caractéristiques en sortie.
- **Stride :** définit le pas de déplacement du noyau sur l'entrée.
- **Padding :** permet de gérer les effets de bord :
  - *Valid padding :* aucune extension, la sortie est plus petite que l'entrée.
  - *Same padding :* la taille de sortie est identique à celle de l'entrée.
  - *Full padding :* la taille de sortie augmente en ajoutant des zéros autour de l'entrée.

Après la convolution, une fonction de non-linéarité est appliquée à la carte de caractéristiques afin de capturer des motifs complexes. La fonction la plus couramment utilisée est la *Rectified Linear Unit* (ReLU) [35], définie par :

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

qui met à zéro les valeurs négatives tout en conservant les valeurs positives. Une autre fonction largement utilisée est la fonction Softmax [36], qui normalise les valeurs à travers les canaux en les transformant en probabilités :

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \quad i = 1, \ldots, K \tag{2.6}$$

Cette normalisation garantit que les probabilités somment à 1 dans chaque canal, ce qui rend Softmax particulièrement adaptée aux tâches de classification multi-classe. La Figure 2.3 illustre une opération de convolution dans laquelle le filtre (noyau) se déplace sur l'image afin de produire une carte de caractéristiques.

2.2.2 Pooling

Le pooling constitue un autre composant fondamental des CNN. Comme illustré dans la Figure 2.4, ces couches visent à sous-échantillonner (*downsample*) les cartes de caractéristiques en réduisant progressivement leurs dimensions spatiales (hauteur et largeur), tout en conservant les informations les plus importantes, afin de diminuer le nombre de paramètres et la complexité computationnelle du modèle [37].

Ces couches opèrent sur chaque carte d'activation en entrée, en réduisant ses dimensions à l'aide d'une matrice de taille fixe (par exemple 2×2) appliquée de manière glissante, en utilisant une fonction de regroupement sur une région de pixels. Dans la plupart des CNN, les stratégies de pooling les plus courantes sont :

– **Max Pooling :** sélectionne la valeur maximale dans chaque région, préservant ainsi la caractéristique la plus saillante :

$$y_{i,j} = \max_{(m,n)\in R_{i,j}} x_{m,n} \tag{2.7}$$

– **Average Pooling :** calcule la moyenne des valeurs dans la région, produisant des cartes de caractéristiques plus lisses :

$$y_{i,j} = \frac{1}{|R_{i,j}|} \sum_{(m,n)\in R_{i,j}} x_{m,n} \tag{2.8}$$

– **Global Average Pooling (GAP) :** réduit chaque carte de caractéristiques à un scalaire unique en moyennant toutes ses valeurs, souvent utilisé avant la couche de classification finale afin de réduire drastiquement le nombre de paramètres :

$$y = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j} \tag{2.9}$$

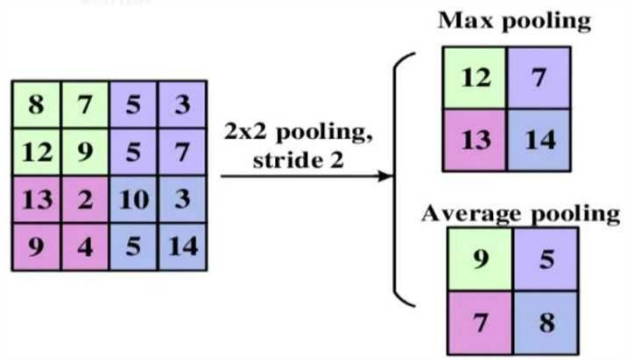


Figure 2.4: Illustration des opérations de pooling. Le max pooling sélectionne la valeur maximale dans chaque région, préservant la caractéristique la plus saillante, tandis que le average pooling calcule la moyenne, produisant des cartes de caractéristiques plus lisses (tiré de [37]).

Le pooling rend le réseau moins sensible aux petits déplacements ou aux distorsions dans l'image. Par exemple, si un objet se déplace légèrement ou est capturé sous un angle légèrement différent, la représentation après pooling sera toujours capable de le reconnaître de manière similaire.

Cependant, un pooling excessif peut entraîner une perte de détails spatiaux fins, ce qui est particulièrement problématique dans les tâches de prédiction dense telles que la segmentation sémantique. Pour cette raison, de nombreuses architectures modernes de segmentation (par exemple U-Net [2], DeepLab [38]) remplacent ou minimisent l'utilisation du pooling, en s'appuyant plutôt sur des convolutions dilatées (*atrous*), voir Fig 2.5 ou des connexions de saut (*skip connections*) afin de préserver la résolution spatiale.

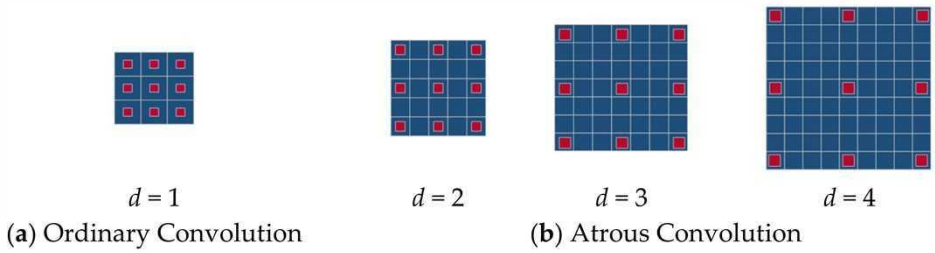


Figure 2.5: Illustration des convolutions dilatées (*atrous*). Contrairement aux convolutions standards, les convolutions dilatées introduisent des espaces entre les éléments du noyau, permettant d'élargir le champ réceptif sans augmenter le nombre de paramètres ni réduire la résolution spatiale (tiré de [39]).

2.2.3 Couches entièrement connectées

La couche entièrement connectée (Fully Connected, FC) est généralement la dernière couche du réseau de neurones. Elle se distingue par la manière dont elle est connectée à la couche précédente : chaque neurone de cette couche est connecté à tous les neurones de la couche précédente. Son rôle principal est de combiner les caractéristiques extraites par les couches convolutionnelles et de pooling afin de produire une décision finale. Les cartes d'activation issues des couches précédentes sont d'abord aplaties en un vecteur unidimensionnel, puis transmises à travers les couches entièrement connectées. Ces couches fonctionnent de manière similaire à celles des perceptrons multicouches (MLP).

Les motifs spatiaux (contours, textures, formes) capturés par les convolutions sont ainsi transmis aux couches entièrement connectées, qui transforment ces représentations en décisions de haut niveau, telles que des scores de classification ou des résultats de régression.

Enfin, la dernière couche applique généralement une fonction d'activation, comme la fonction Softmax pour la classification multi-classe ou une activation linéaire pour les tâches de régression [33]. Par exemple, après que les couches convolutionnelles et de pooling ont détecté des caractéristiques locales telles que des contours ou des fenêtres dans une image, la couche FC les combine afin de déterminer, avec une certaine probabilité, si l'objet global correspond à une voiture ou à un bâtiment, selon les classes définies.

Globalement, la couche FC effectue une combinaison linéaire pondérée de ses entrées, suivie d'une fonction d'activation (par exemple ReLU ou Softmax). Alors que les couches convolutionnelles et de pooling sont responsables de l'extraction des caractéristiques, la couche FC est responsable de la classification ou de la prédiction en projetant les caractéristiques apprises dans l'espace de sortie. Dans les tâches de segmentation, au lieu de terminer par des couches entièrement connectées, les modèles les adaptent souvent en couches entièrement convolutionnelles (comme dans les réseaux FCN [1]) afin de préserver la structure spatiale et de produire des prédictions au niveau du pixel.

2.3 Vision Transformers (ViTs)

Les Vision Transformers (ViTs) sont un type d'architecture de réseau de neurones qui a émergé comme une alternative puissante aux réseaux de neurones convolutifs (CNN) pour les tâches de compréhension d'images [7]. Ils constituent une extension de l'architecture Transformer, initialement développée pour les tâches de traitement du langage naturel [40]. Les architectures Transformer utilisent un mécanisme d'attention pour apprendre les relations entre les éléments d'une séquence d'entrée.

Pour illustrer ce principe, considérons une phrase dans laquelle chaque mot entretient des relations avec les autres afin de construire un sens cohérent. Le mécanisme d'attention permet au modèle d'identifier et de pondérer ces relations en évaluant l'importance de chaque mot dans le contexte global. Cette capacité à modéliser simultanément les dépendances entre tous les éléments d'une séquence explique en grande partie le succès des Transformers par rapport aux approches séquentielles traditionnelles, telles que les réseaux de neurones récurrents (RNN) [41]. Contrairement aux RNN, qui traitent les mots un à un et peuvent avoir des difficultés à capturer les dépendances lointaines, les Transformers analysent l'ensemble de la séquence en parallèle, ce qui favorise une meilleure préservation du contexte et de la cohérence sémantique.

Sur ce principe, les Vision Transformers (ViT) appliquent le mécanisme d'auto-attention à l'analyse d'images [7]. Comme illustré à la Figure 2.6, l'image est d'abord divisée en patches de taille fixe, puis chaque patch est converti en un jeton (*token*) enrichi par un encodage positionnel afin de préserver l'information spatiale [42]. Ces jetons sont ensuite traités par une succession de blocs Transformer composés d'auto-attention multi-têtes et de couches feedforward. Cette architecture permet de capturer efficacement les relations entre régions éloignées de l'image et de modéliser le contexte global, un avantage majeur pour de nombreuses tâches de vision par ordinateur.

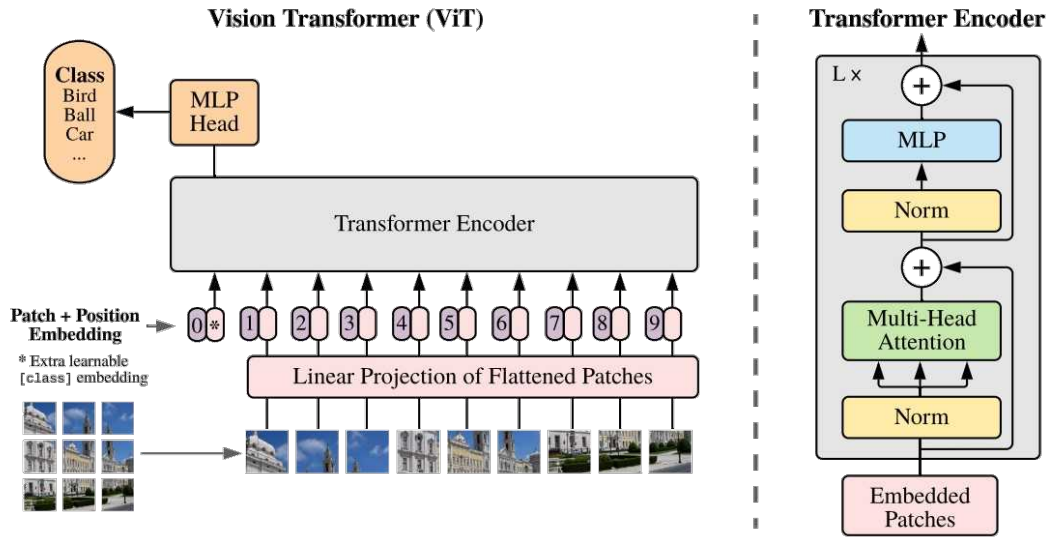


Figure 2.6: Vue d'ensemble de l'architecture Vision Transformer (ViT). L'image est divisée en patches, transformée en une séquence de tokens avec encodage positionnel, puis traitée par des couches encodeur Transformer (tiré de [7]).

2.3.1 Patch Embedding

Étant donné que les architectures Transformer attendent des données séquentielles en entrée, une image $X \in R^{H \times W \times C}$ est d'abord divisée en un ensemble de patches non chevauchants de taille $P \times P$. Chaque patch est ensuite aplati en un vecteur et projeté dans un espace d'embedding latent à l'aide d'une transformation linéaire. Ce processus produit une séquence de tokens :

$$Z_0 = [x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos} \tag{2.10}$$

où E est une matrice d'embedding apprenable, $N = \frac{HW}{P^2}$ est le nombre de patches, et E_{pos} est un encodage positionnel ajouté afin de conserver l'information spatiale perdue lors de l'aplatissement. Ces embeddings servent d'entrée à l'encodeur Transformer.

2.3.2 Self-Attention Mechanism

Au cœur des ViT se trouve le mécanisme de **multi-head self-attention (MSA)**, qui permet au modèle de calculer des relations contextuelles entre tous les tokens. Pour chaque token, des vecteurs requête (Q), clé (K) et valeur (V) sont calculés comme des projections linéaires des embeddings d'entrée. L'opération d'attention est définie comme suit :

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

où d_k est la dimension des vecteurs clés. Cette formulation permet à chaque patch de prêter attention à tous les autres, en pondérant leur contribution en fonction de leur pertinence apprise. L'attention multi-têtes étend ce principe en utilisant plusieurs ensembles de Q, K et V, afin de capturer des informations complémentaires issues de différents sous-espaces de représentation.

2.3.3 Transformer Encoder

Les embeddings des patches sont ensuite traités par une série de couches encodeur Transformer, chacune composée d'un bloc d'auto-attention multi-têtes et d'un réseau feed-forward (FFN), tous deux entourés de connexions résiduelles et de normalisation de couche :

Chaque bloc Transformer comprend un mécanisme d'auto-attention multi-têtes (*Multi-Head Self-Attention*, MSA), une normalisation de couche (*Layer Normalization*, LN) et un réseau *feed-forward* (FFN).

$$Z'_\ell = \text{MSA}(\text{LN}(Z_{\ell-1})) + Z_{\ell-1} \tag{2.12}$$

$$Z_\ell = \text{FFN}(\text{LN}(Z'_\ell)) + Z'_\ell \tag{2.13}$$

où ℓ représente l'indice de la couche encodeur. À travers ces couches successives, le modèle construit des représentations de plus en plus riches et globales de l'image.

2.4 Détection d'objets

La détection d'objets consiste à identifier et localiser des objets individuels dans une image [43]. Cette tâche implique la prédiction à la fois de la classe de chaque objet et des coordonnées d'une boîte englobante qui l'entoure précisément, comme illustré dans la Figure 2.7. Ce double objectif fait de la détection d'objets l'un des problèmes les plus fondamentaux et les plus complexes en vision par ordinateur, puisqu'un modèle doit apprendre simultanément ce qu'est un objet et où il se trouve.

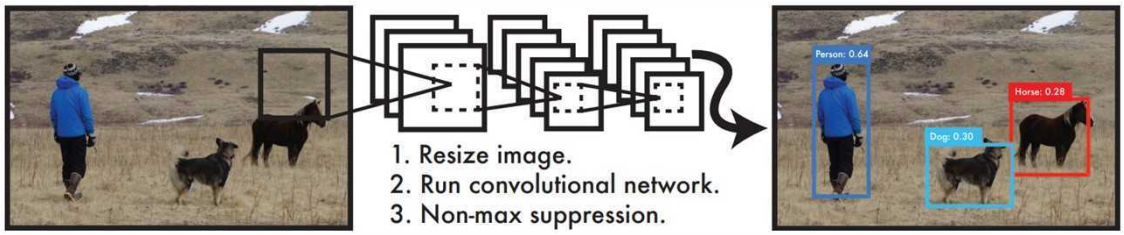


Figure 2.7: Vue d'ensemble d'un pipeline typique de détection d'objets. Le réseau prédit simultanément des boîtes englobantes et des étiquettes de classes pour tous les objets présents dans l'image (tiré de [44]).

2.4.1 Détecteurs en deux étapes

Les premières approches basées sur l'apprentissage profond pour la détection d'objets reposaient principalement sur un pipeline en deux étapes, dans lequel le processus de détection est divisé en deux phases successives : la génération de régions candidates (*region proposal*) et la classification. La famille de méthodes la plus influente dans ce cadre est celle des réseaux R-CNN (*Region-based Convolutional Neural Networks*) :

- **R-CNN** [43] : génère des propositions de régions à l'aide de la méthode *selective search*, puis classe chaque région à l'aide d'un CNN.
- **Fast R-CNN** [45] : améliore l'efficacité en calculant les caractéristiques convolutionnelles une seule fois par image et en appliquant une couche de *Region of Interest* (RoI) pooling pour classifier les régions proposées.
- **Faster R-CNN** [46] : introduit un *Region Proposal Network* (RPN) qui apprend à proposer directement des régions candidates à partir des cartes de caractéristiques, permettant un entraînement de bout en bout et des performances proches du temps réel.

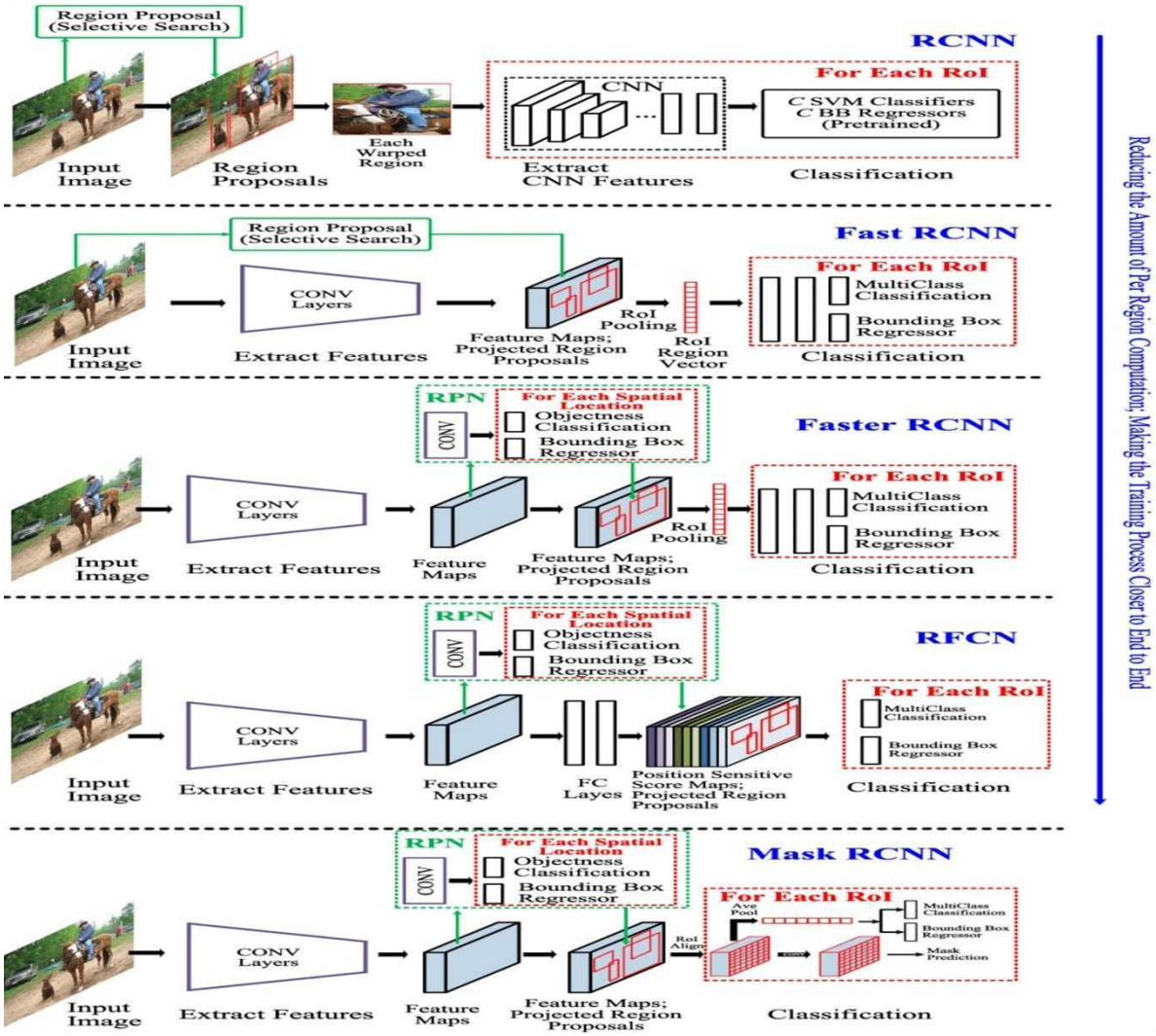


Figure 2.8: Vue d'ensemble de la famille des détecteurs R-CNN. La figure illustre l'évolution de R-CNN vers Fast R-CNN, Faster R-CNN, R-FCN et Mask R-CNN, montrant l'intégration progressive des étapes de génération de régions, d'extraction de caractéristiques et de classification (tiré de [47]).

Ces détecteurs en deux étapes offrent généralement une grande précision et de bonnes performances de localisation, ce qui les rend adaptés aux tâches nécessitant une grande précision.

2.4.2 Détecteurs en une étape

Afin d'améliorer la vitesse et de permettre des applications en temps réel, les détecteurs en une seule étape ont été introduits. Ces modèles suppriment l'étape de génération de

propositions de régions et prédisent directement les probabilités de classes et les coordonnées des boîtes englobantes à partir de cartes de caractéristiques denses :

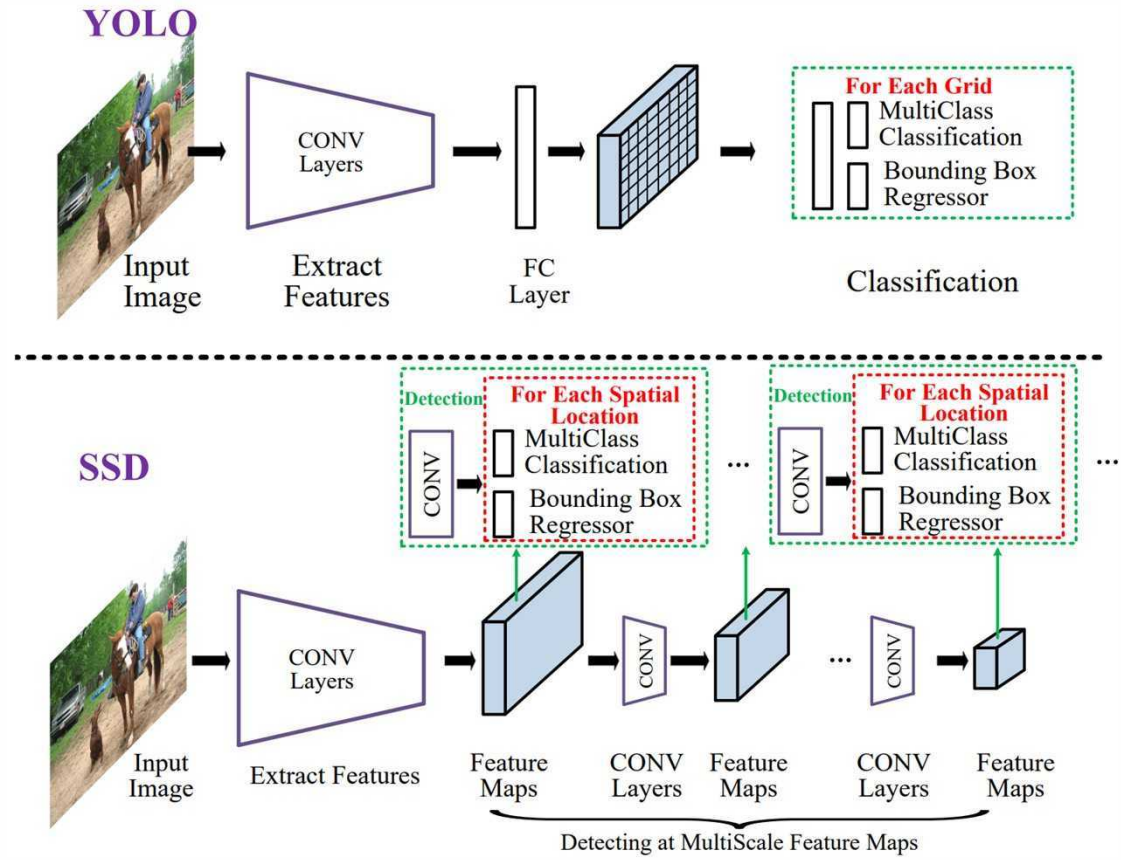


Figure 2.9: Illustration des détecteurs d'objets en une étape, notamment YOLO et SSD. Ces modèles formulent la détection comme un problème de régression unique en divisant l'image en cellules de grille ou en boîtes d'ancrage, et en prédisant simultanément les coordonnées des boîtes et les probabilités de classes directement à partir des cartes de caractéristiques, permettant une inférence en temps réel (tiré de [47]).

- **YOLO (You Only Look Once)** [44] : traite la détection d'objets comme un problème de régression unique, en divisant l'image en une grille et en prédisant des boîtes englobantes et des classes pour chaque cellule.
- **SSD (Single Shot MultiBox Detector)** [48] : utilise des cartes de caractéristiques multi-échelles pour détecter des objets de tailles variées, améliorant la précision sans compromettre la vitesse.

– **RetinaNet** [49] : introduit la fonction de perte *focal loss* pour traiter le problème du déséquilibre des classes dans la détection dense, améliorant significativement les performances pour les objets rares et de petite taille.

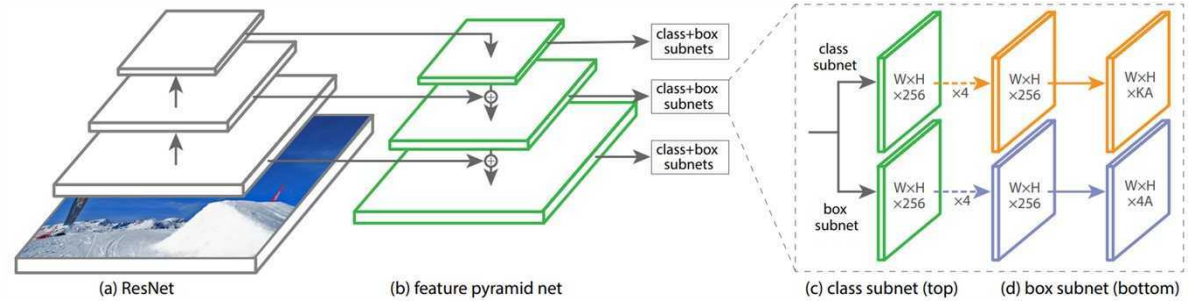


Figure 2.10: Vue d'ensemble de l'architecture RetinaNet. Un réseau de pyramide de caractéristiques (FPN) [49] est construit sur un backbone ResNet [50] (a) afin de générer une représentation multi-échelle riche (b). Deux sous-réseaux sont ensuite utilisés : l'un pour la classification des boîtes d'ancrage (c), et l'autre pour la régression des boîtes vers les objets réels (d). Cette architecture permet d'exploiter la focal loss afin de réduire l'écart de performance entre les détecteurs en une étape et ceux en deux étapes, tout en conservant une inférence rapide (tiré de [49]).

Les détecteurs en une étape sont généralement plus rapides et plus efficaces en termes de calcul, bien qu'ils puissent présenter une légère baisse de précision par rapport aux méthodes en deux étapes.

2.5 Généralisation inter-domaines

La généralisation inter-domaines (*Domain Generalization*, DG) vise à apprendre des modèles capables de maintenir de bonnes performances lorsqu'ils sont appliqués à des données provenant de distributions différentes de celles observées durant l'entraînement. Contrairement à l'adaptation de domaine (*Domain Adaptation*), qui suppose l'accès à des données du domaine cible, la DG cherche à apprendre des représentations robustes sans accès préalable aux données cibles [51].

Formellement, soit un ensemble de domaines sources $\mathcal{D}_s = \{D_1, D_2, \ldots, D_n\}$, où chaque domaine D_i est associé à une distribution de données $P_i(X, Y)$. L'objectif de la DG est d'apprendre une fonction f_θ telle que :

$$\min_{\theta} \sum_{i=1}^{n} E_{(x,y)\sim P_i} \left[\mathcal{L}(f_\theta(x), y) \right], \tag{2.14}$$

tout en assurant de bonnes performances sur un domaine cible inconnu D_t de distribution $P_t(X, Y)$, non observée durant l'entraînement. Le principal défi en DG réside dans le *domain shift*, qui correspond aux différences de distribution entre les domaines sources et le domaine cible. Ce phénomène peut être exprimé comme :

$$P_s(X, Y) \neq P_t(X, Y) \tag{2.15}$$

où P_s représente la distribution agrégée des domaines sources et P_t celle du domaine cible. Dans le contexte des images aériennes, ce décalage peut être causé par des variations de capteurs, de conditions d'acquisition, d'illumination, de résolution ou encore de contexte géographique. Ces variations rendent difficile la généralisation des modèles appris sur un domaine vers un autre.

Afin de traiter ce problème, les méthodes de transfert inter-domaine cherchent à apprendre des représentations invariantes aux variations de domaine tout en conservant les informations discriminantes nécessaires à la tâche. Cela peut être formulé comme la recherche d'une transformation Φ telle que :

$$\Phi(X_s) \approx \Phi(X_t) \tag{2.16}$$

tout en garantissant que les labels restent séparables dans l'espace transformé.

2.5.1 Apprentissage de représentations invariantes

L'un des principes fondamentaux de la DG est l'apprentissage de représentations invariantes aux domaines. L'objectif est de projeter les données issues de différents domaines dans un espace commun où les distributions sont alignées.

Une manière de formaliser cet objectif consiste à minimiser une divergence entre distributions, par exemple la divergence de Jensen-Shannon ou la distance de Wasserstein :

$$\min_{\Phi} \ \mathcal{D}(P_s(\Phi(X)), P_t(\Phi(X))) \tag{2.17}$$

où $\mathcal{D}$ mesure la distance entre les distributions dans l'espace des features.

Cependant, en l'absence de données du domaine cible, les approches DG reposent souvent sur la simulation de variations de domaine ou sur l'apprentissage à partir de multiples domaines sources afin de capturer des invariances robustes [52]. Malgré les avancées récentes, la DG reste un problème difficile en raison de plusieurs facteurs :

- **Variabilité des distributions :** les différences entre domaines peuvent être complexes et non linéaires.
- **Compromis invariance-discrimination :** rendre les représentations invariantes peut entraîner une perte d'information discriminante.
- **Absence de données cibles :** rend l'apprentissage plus difficile que dans les approches supervisées classiques.

Ces défis sont particulièrement critiques dans les tâches de segmentation sémantique, où une prédiction précise au niveau du pixel est requise, et où les erreurs locales peuvent fortement dégrader les performances globales.

2.6 Représentations par prototypes et structuration hiérarchique

Les approches basées sur les prototypes constituent une stratégie efficace pour représenter les classes dans l'espace des caractéristiques. L'idée principale consiste à associer à chaque classe un vecteur représentatif, appelé prototype, qui correspond généralement au centre de gravité des caractéristiques extraites pour cette classe.

Formellement, étant donné un ensemble de caractéristiques $\{f_i\}_{i=1}^{N_k}$ appartenant à une classe k, le prototype $\mathbf{c}_k$ peut être défini comme la moyenne des caractéristiques :

$$\mathbf{c}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} f_i \tag{2.18}$$

où N_k est le nombre d'échantillons de la classe k.

Dans ce cadre, la classification peut être formulée comme une mesure de similarité entre les caractéristiques d'un pixel ou d'un patch et les prototypes des différentes classes. Par exemple, en utilisant la similarité cosinus :

$$s_k(x) = \frac{f(x) \cdot \mathbf{c}_k}{\|f(x)\|\|\mathbf{c}_k\|} \tag{2.19}$$

où $f(x)$ représente la caractéristique extraite pour l'entrée x.

Les prototypes offrent une représentation compacte et interprétable des classes, facilitant la comparaison entre différentes distributions de données et améliorant la robustesse des modèles face aux variations inter-domaines [53].

2.6.1 Alignement des représentations

Dans un contexte de généralisation inter-domaines, les prototypes peuvent être utilisés pour aligner les représentations issues de différents domaines. L'objectif est de rapprocher les prototypes correspondants dans l'espace des caractéristiques :

$$\min \sum_{k=1}^{K} \|\mathbf{c}_k^{(s)} - \mathbf{c}_k^{(t)}\|^2 \tag{2.20}$$

où $\mathbf{c}_k^{(s)}$ et $\mathbf{c}_k^{(t)}$ représentent respectivement les prototypes de la classe k dans les domaines source et cible.

Cet alignement permet de réduire l'effet du *domain shift* en assurant que les représentations sémantiques restent cohérentes entre les différents domaines.

2.6.2 Structuration hiérarchique

La structuration hiérarchique consiste à organiser les classes en plusieurs niveaux de granularité, typiquement sous la forme de super-classes et de sous-classes. Cette approche permet de décomposer un problème complexe en plusieurs sous-problèmes plus simples.

Soit une hiérarchie définie par une fonction $h : \mathcal{Y} \to \mathcal{S}$, où $\mathcal{Y}$ représente l'ensemble des classes fines et $\mathcal{S}$ l'ensemble des super-classes. Chaque classe fine y est associée à une super-classe $s = h(y)$.

Dans ce cadre, la prédiction peut être décomposée en deux étapes :

$$P(y|x) = P(y|s, x) \cdot P(s|x) \tag{2.21}$$

où $P(s|x)$ correspond à la prédiction au niveau des super-classes et $P(y|s, x)$ à la discrimination fine au sein de chaque groupe.

Cette décomposition permet de réduire la complexité du problème en contraignant l'espace des décisions et en améliorant la stabilité de l'apprentissage, notamment lorsque les classes sont déséquilibrées ou visuellement similaires.

2.6.3 Avantages de l'approche hiérarchique

L'introduction d'une structure hiérarchique présente plusieurs avantages :

- **Réduction de la complexité :** la classification est décomposée en plusieurs étapes plus simples.
- **Amélioration de la robustesse :** les erreurs grossières sont limitées grâce à une première décision au niveau des super-classes.
- **Meilleure généralisation :** les connaissances partagées entre classes similaires sont exploitées.

Cette approche est particulièrement pertinente dans les scénarios où les classes présentent des relations sémantiques fortes ou des similarités visuelles importantes.

Chapitre 3

Etat de l'art

3.1 Segmentation sémantique pour la compréhension des scènes de catastrophe

3.1.1 Architectures encodeur–décodeur

Avec l'avènement de l'apprentissage profond, des tâches de classification à la segmentation, les CNN ont considérablement renforcé les capacités de l'intelligence artificielle. Une avancée majeure a été introduite avec les Fully Convolutional Networks (FCN) [1], qui ont remplacé les couches entièrement connectées traditionnelles par des couches convolutionnelles afin de permettre des prédictions denses de bout en bout. Ce changement architectural a permis aux modèles de prédire des probabilités de classe pour chaque pixel. Cependant, malgré son succès initial, FCN présentait des limitations dans la capture des détails fins et dans la délimitation précise des contours des objets.

L'architecture U-Net [2] a répondu à ce problème en introduisant une structure encodeur–décodeur avec des connexions de saut (*skip connections*), permettant de préserver les informations spatiales. Des extensions telles que U-Net++ [54] ont ajouté des connexions de saut imbriquées et denses afin d'améliorer la fusion multi-échelle. Ces architectures restent fondamentales pour la segmentation dans l'imagerie aérienne et médicale en raison de leur capacité à préserver les contours et les structures spatiales fines.

En s'appuyant sur cette base, U-Net++ [54] a étendu la conception initiale en introduisant des chemins de connexions imbriqués et denses, permettant une fusion multi-échelle plus efficace. Cette architecture affine la représentation sémantique à plusieurs échelles et

améliore la segmentation des objets de petite taille ou de forme irrégulière, fréquents dans les images post-catastrophe. U-Net++ réduit également l'écart sémantique entre les cartes de caractéristiques de l'encodeur et du décodeur, améliorant ainsi la précision des contours et la délimitation des objets.

U-Net et ses variantes ont inspiré de nombreuses études, notamment dans le domaine des catastrophes. Wenmei Li et al. [55] ont proposé U-Net-CBAM, une variante de U-Net enrichie par un mécanisme d'attention pour la cartographie des inondations à partir d'images radar à synthèse d'ouverture (SAR). L'idée consiste à intégrer le module CBAM (*Convolutional Block Attention Module*), un module d'attention léger et modulaire conçu pour améliorer la capacité du CNN à se concentrer sur les caractéristiques importantes.

Ce module fonctionne en deux étapes. La première est le *Channel Attention Module* (CAM), dont le rôle est d'identifier quelles caractéristiques (canaux) sont les plus importantes pour la tâche. Cela est réalisé en appliquant des opérations de *global average pooling* et de *max pooling* sur chaque canal, suivies d'un petit réseau qui génère des poids d'attention. Les canaux contenant des informations discriminantes (par exemple eau vs sol sec) sont ainsi renforcés, tandis que les autres sont atténués.

La seconde étape est le *Spatial Attention Module* (SAM), qui apprend où concentrer l'attention dans l'espace. Il applique des opérations de moyenne et de maximum sur les canaux afin de générer une carte d'attention spatiale. Ainsi, l'attention est focalisée sur les régions pertinentes de l'image (par exemple zones inondées ou bâtiments détruits), ce qui améliore la continuité des contours et la détection des petites régions.

Bien que l'adaptation de modèles puissants au domaine des catastrophes soit une solution naturelle, les paramètres et hyperparamètres peuvent également être adaptés. Nugraha et al. [56] ont appliqué U-Net et PSPNet à des images aériennes de catastrophes et utilisé une optimisation des hyperparamètres par *grid search* pour améliorer le mIoU, montrant que le réglage précis des architectures encodeur–décodeur peut significativement améliorer les performances dans des environnements complexes.

Il existe également plusieurs pipelines multi-étapes spécialisés qui ont amélioré les performances de segmentation dans les scénarios de catastrophes. Deng et al. [57] ont proposé une approche en deux étapes dans laquelle un modèle basé sur U-Net segmente d'abord les empreintes de bâtiments à partir d'images aériennes haute résolution, produisant un masque binaire indiquant la localisation et la forme des bâtiments. Dans une seconde étape, les bâtiments segmentés sont extraits et transmis à un classificateur chargé d'évaluer le niveau de dommage (par exemple : aucun dommage, mineur, majeur, détruit). Cette décomposition

permet un apprentissage plus spécialisé à chaque étape, réduisant la propagation des erreurs et améliorant la précision ainsi que l'interprétabilité par rapport aux approches en une seule étape.

De manière similaire, DeepDamageNet [12] est un modèle en deux étapes conçu pour la détection et la classification des dommages aux bâtiments après des catastrophes, principalement à partir d'images satellites (dataset xBD). Le processus combine segmentation et classification dans un pipeline modulaire. Contrairement à Deng et al. [57], où les deux étapes sont découplées et peuvent être entraînées séparément, DeepDamageNet adopte une architecture conjointe dans laquelle les deux modules sont optimisés ensemble. Cette approche montre une meilleure généralisation entre différents types de catastrophes.

Des stratégies de fusion multimodale et multi-temporelle ont également été explorées. Multi3Net [58], basé sur une architecture U-Net modifiée, combine des données provenant de trois types de capteurs satellites et de différentes résolutions : multi-résolution, multi-source (optique et radar), et multi-temporelle (avant et après catastrophe). Ces données sont fusionnées dans un pipeline unique afin d'améliorer la précision de la segmentation des bâtiments inondés. Cependant, cette approche est plus complexe et coûteuse en ressources, car elle nécessite plusieurs sources de données et des étapes de prétraitement supplémentaires.

De manière similaire, Rahnemoonfar et al. [59] ont montré que la fusion de données multi-capteurs (optique + radar) dans une architecture U-Net permet d'améliorer la robustesse du modèle face aux variations atmosphériques, aux nuages et aux conditions d'éclairage difficiles. Shen et al. [60] ont proposé BDANet (*Building Damage Assessment Network*), un réseau convolutionnel multi-échelle basé sur U-Net enrichi par un mécanisme d'attention croisée bidirectionnelle. Le modèle capture les relations spatiales et contextuelles entre différentes zones d'un bâtiment, améliorant la détection des petits dommages et des structures fines. Cependant, ces mécanismes d'attention augmentent la complexité computationnelle.

D'autres travaux ont exploré des modifications architecturales spécifiques. TreeUNet [61] introduit une structure hiérarchique en combinant un bloc Tree-CNN avec un module de segmentation, améliorant l'extraction multi-échelle. ResUNet-a [62] intègre des connexions résiduelles pour améliorer la propagation des caractéristiques et réduire le problème de disparition du gradient (*vanishing gradient problem*). Le modèle utilise également des convolutions dilatées parallèles pour capturer le contexte multi-échelle. Priyanka et al. [63] ont proposé DIResUNet, qui intègre un module de pooling global dense pour extraire des caractéristiques locales et globales, améliorant la segmentation dans des scènes complexes.

Certaines approches ont également intégré des informations spectrales spécifiques à la télédétection. Liu et al. [64] ont combiné des images RGB avec des données LiDAR, améliorant la segmentation des structures urbaines. Chen et al. [65] ont proposé SDFCNv2, utilisant une fusion spatiale et des blocs convolutionnels hybrides pour optimiser la segmentation multi-spectrale. Enfin, Gupta et al. [66] ont montré que l'exploitation de données ouvertes issues d'OpenStreetMap permet de réduire le besoin d'annotations manuelles et de soutenir la segmentation d'images aériennes pour l'analyse post-catastrophe.

Dans l'ensemble, les architectures encodeur–décodeur et les variantes de U-Net constituent la base de la segmentation sémantique dans les scènes de catastrophe. Leur capacité à capturer des représentations hiérarchiques, à préserver les détails spatiaux et à intégrer des données variées les rend indispensables. Cependant, des défis tels que le *domain shift*, le manque de données annotées et la segmentation d'objets de petite taille persistent, motivant le développement d'approches hybrides et basées sur l'attention.

3.1.2 Agrégation du contexte par convolutions dilatées et modules pyramidaux

Bien que les architectures encodeur–décodeur telles que U-Net excellent dans la préservation des détails spatiaux, notamment grâce aux connexions de saut, leurs couches convolutionnelles opèrent avec des champs réceptifs fixes, ce qui limite leur capacité à capturer des informations contextuelles à longue portée. Cela pose un problème, car dans les environnements complexes liés aux catastrophes, l'interprétation sémantique dépend fortement du contexte global. Pour remédier à cela, une importante direction de recherche s'est concentrée sur l'extension du champ réceptif et l'agrégation du contexte à différentes échelles, en utilisant des convolutions dilatées (*atrous*) et des modules pyramidaux.

Ces approches permettent d'améliorer la compréhension sémantique sans compromettre la résolution spatiale, ce qui améliore significativement les performances de segmentation dans des environnements encombrés et hétérogènes, tels que les zones urbaines après une catastrophe. La série de modèles DeepLab a été pionnière dans cette direction. DeepLab v1 [67] a introduit les convolutions dilatées afin d'élargir le champ réceptif tout en conservant la résolution des cartes de caractéristiques. Cette innovation permet au réseau d'intégrer un contexte plus large, ce qui est particulièrement utile pour segmenter des phénomènes de grande échelle tels que les zones inondées ou les champs de débris.

De plus, DeepLab v1 intègre un modèle CRF (*Conditional Random Field*) comme étape de post-traitement pour affiner les contours des objets, ce qui est essentiel pour la délimitation de structures fines telles que les routes. Toutefois, son coût computationnel et sa dépendance au CRF limitent son utilisation en temps réel. DeepLab v2 [24] a amélioré ce concept en introduisant le module ASPP (*Atrous Spatial Pyramid Pooling*), qui applique des convolutions dilatées parallèles avec différents taux de dilation afin de capturer des caractéristiques à plusieurs échelles. ASPP améliore significativement la segmentation d'objets présentant une forte variabilité d'échelle, comme les bâtiments endommagés ou les véhicules dispersés. DeepLab v3 [68] a ensuite raffiné ASPP en intégrant des couches convolutionnelles supplémentaires et de la normalisation de lot (*batch normalization*), améliorant la robustesse et la stabilité de l'entraînement. Cette version supprime également la nécessité d'un post-traitement CRF, simplifiant ainsi le pipeline. DeepLab v3+ [3] étend encore ce cadre en introduisant un décodeur inspiré de U-Net, permettant une meilleure reconstruction des contours et une segmentation plus précise des petits objets, ce qui est essentiel dans les scénarios post-catastrophe.

Plusieurs travaux ont appliqué directement les architectures DeepLab à l'analyse de scènes de catastrophe. Venugopal [69] a proposé un CNN dilaté basé sur DeepLab pour la segmentation des zones inondées, montrant de meilleures performances dans la capture des frontières de l'eau. Wang et al. [70] ont introduit un modèle DeepLab régularisé pour atténuer le surapprentissage dans des contextes de données limitées.

D'autres travaux ont ciblé des défis spécifiques. Andrade et al. [71] ont adapté DeepLab v3+ pour la segmentation de zones forestières affectées par des incendies, en traitant le déséquilibre de classes et le bruit de fond. Wang et al. [72] ont proposé CFAMNet, une variante de DeepLab v3+ intégrant un module d'attention sensible aux classes, améliorant la discrimination entre classes visuellement similaires. Wang et al. [73] ont également exploré l'intégration de la normalisation Instance–Batch (IBN) dans le backbone DeepLab, améliorant la généralisation à différents domaines et types de capteurs.

En parallèle, les techniques de *pyramid pooling* sont devenues essentielles pour capturer le contexte à différentes échelles. PSPNet [25] introduit un module de pooling pyramidal qui agrège des informations globales et locales, ce qui est particulièrement efficace pour des tâches de grande échelle comme la cartographie des inondations. Nugraha et al. [56] ont combiné PSPNet avec U-Net pour améliorer les performances dans les zones affectées par des catastrophes.

L'importance des représentations haute résolution a également été reconnue. HRNet [74] maintient des représentations haute résolution tout au long du réseau, contrairement aux architectures classiques qui réduisent puis restaurent la résolution. Cette approche est particulièrement adaptée à la segmentation d'objets fins ou de petite taille, tels que les routes ou les lignes électriques.

En conservant plusieurs résolutions en parallèle et en échangeant continuellement des informations entre elles, HRNet permet d'obtenir une meilleure précision spatiale et une meilleure compréhension contextuelle. Au-delà des innovations architecturales, plusieurs travaux ont démontré l'intérêt de ces techniques dans les applications liées aux catastrophes. DeepLab v3+ a été utilisé pour la segmentation des zones inondées dans des images UAV [75], tandis que Jo et al. [76] ont combiné YOLOv7 et DeepLabv2 dans un pipeline multi-étapes, utilisant la détection pour localiser les régions d'intérêt avant d'y appliquer une segmentation fine.

Dans l'ensemble, les convolutions dilatées, le pyramid pooling et les représentations multi-résolutions ont considérablement amélioré la capacité des réseaux à capturer le contexte dans les scènes de catastrophe. Ces approches permettent de mieux gérer la variabilité d'échelle et d'améliorer la détection des structures fines. Cependant, elles présentent encore certaines limites : les convolutions dilatées augmentent la complexité computationnelle, le pyramid pooling peut introduire de la redondance, et les architectures haute résolution nécessitent davantage de mémoire. Les recherches récentes s'orientent vers la combinaison de ces techniques avec des mécanismes d'attention et des architectures basées sur les Transformers afin d'améliorer simultanément précision et efficacité dans des scénarios réels.

3.2 Cadres de détection d'objets pour l'analyse de scènes de catastrophes

Alors que la segmentation sémantique permet une compréhension dense au niveau du pixel des scènes de catastrophe, la détection d'objets joue un rôle complémentaire en localisant des entités clés telles que les bâtiments endommagés, les zones inondées, les véhicules ou les débris. L'intégration de la détection dans les pipelines d'analyse de catastrophes a évolué des approches classiques basées sur des régions vers des modèles légers en temps réel et des cadres hybrides combinant détection et segmentation. Une tendance croissante dans les recherches récentes vise à améliorer la reconnaissance des classes de petite taille ou sous-représentées.

Zhu et al. [77] et Zou et al. [78] ont exploité des architectures basées sur YOLO et Mask R-CNN pour l'analyse de catastrophes, améliorant la localisation des bâtiments endommagés et des zones inondées. De même, YOLO-Damage [79] a démontré des performances élevées pour la détection de dommages à partir d'images UAV. Les détecteurs basés sur les Transformers, tels que DETR [80], ont également simplifié la détection grâce à une prédiction directe d'ensembles. Cependant, la plupart des approches centrées sur la détection ne fournissent que des annotations au niveau des objets et ne produisent pas de segmentation dense au niveau des pixels. En revanche, le modèle proposé intègre directement des indices guidés par la détection dans le pipeline de segmentation, améliorant la discrimination au niveau des pixels pour les classes rares et de petite taille.

Les premiers travaux ont adapté des détecteurs classiques tels que Faster R-CNN et Mask R-CNN [46, 81] pour la détection de dommages et la cartographie post-catastrophe. Ces méthodes ont montré une excellente précision de localisation pour les empreintes de bâtiments et les débris, mais souffraient de vitesses d'inférence faibles et d'une forte complexité computationnelle, limitant leur utilisation pour la surveillance aérienne en temps réel. Les détecteurs en une étape basés sur des ancres, tels que YOLO et RetinaNet [44, 49], ont permis de surmonter ces limitations en offrant une détection rapide, particulièrement efficace pour identifier des objets petits et dispersés tels que les véhicules, les routes ou les débris dans les images UAV.

Des travaux récents ont exploré des pipelines multi-étapes combinant détection et segmentation afin d'améliorer l'interprétation des scènes. Par exemple, la combinaison de YOLOv7 et DeepLabv2 [76] utilise la détection d'objets comme étape préalable à la segmentation au niveau du pixel, améliorant significativement la qualité de la segmentation dans des environnements de catastrophe complexes et encombrés. De manière similaire, des cadres basés sur UAV intégrant détection et segmentation [82] ont été proposés pour détecter la présence humaine, les zones inondées et les dommages structurels en temps réel, facilitant une prise de décision rapide lors des interventions d'urgence.

Les approches en deux étapes étendent davantage ce paradigme en utilisant la détection ou la segmentation initiale pour localiser les régions d'intérêt, suivie d'une classification ou d'une évaluation plus fine de la sévérité. DeepDamageNet [12], présenté précédemment, constitue un exemple de cette stratégie. De même, un pipeline en deux étapes basé sur U-Net appliqué au dataset xBD [83] segmente les régions de bâtiments avant d'effectuer une classification fine des dommages. Ces méthodes hiérarchiques exploitent les indices de détection pour réduire les faux positifs et améliorer la granularité de l'analyse des impacts.

Des adaptations spécifiques au domaine ont également émergé pour répondre à des contraintes opérationnelles particulières. DIS-Mine [84] se concentre sur la segmentation d'instances dans des environnements souterrains à faible luminosité, en utilisant des étapes de prétraitement et des mécanismes d'attention pour améliorer l'extraction des caractéristiques dans des conditions de faible visibilité. Une approche améliorée de segmentation d'instances pour les scénarios post-sismiques [85] exploite YOLOv5-Seg afin de fournir des évaluations rapides et précises des structures endommagées. Par ailleurs, des cadres basés sur des données ouvertes [66] intègrent la détection avec des mises à jour des réseaux routiers pour soutenir l'analyse des infrastructures à grande échelle.

Dans l'ensemble, la détection d'objets est devenue un composant essentiel des pipelines de compréhension des scènes de catastrophe. Elle permet non seulement la localisation rapide des objets clés, mais améliore également la précision de la segmentation via des indices régionaux, et facilite des tâches d'analyse ultérieures telles que l'évaluation de la sévérité des dommages et l'analyse des infrastructures. La transition des détecteurs classiques vers des cadres hybrides détection–segmentation reflète une tendance plus large vers des systèmes intelligents multi-tâches et sensibles au contexte pour l'analyse des catastrophes.

3.2.1 Architectures haute résolution et modèles hybrides basés sur les Transformers

Bien que les architectures encodeur–décodeur et les modules d'agrégation de contexte aient considérablement amélioré les performances de la segmentation sémantique, elles restent limitées dans leur capacité à capturer des relations globales et à modéliser des dépendances à longue portée, pourtant essentielles pour interpréter des images de catastrophe à grande échelle. Les CNN traditionnels reposent intrinsèquement sur des champs réceptifs locaux, ce qui limite leur compréhension des indices contextuels répartis sur de larges zones géographiques, tels que l'étendue des inondations, les motifs de destruction urbaine ou la perte de végétation.

De plus, les opérations répétées de sous-échantillonnage et de sur-échantillonnage peuvent entraîner une perte de précision spatiale, conduisant à une segmentation imprécise des structures fines ou de petite taille. Pour surmonter ces limitations, les recherches récentes se sont orientées vers des architectures haute résolution et des modèles hybrides basés sur les Transformers, capables de préserver les détails spatiaux tout en intégrant un contexte sémantique global.

Les mécanismes d'attention ont d'abord été introduits dans les RNN [86], avant de révolutionner les tâches de vision avec l'auto-attention et l'attention multi-têtes dans l'architecture Transformer [40]. Les Vision Transformers (ViT) [7] appliquent ces mécanismes aux patches d'image, permettant de capturer des dépendances à longue distance au-delà des capacités des CNN. Des modules d'attention plus récents, tels que Squeeze-and-Excitation (SE) [87], Criss-Cross Attention [88] et Dual Attention [89], ont amélioré le raisonnement contextuel et la précision de segmentation, en particulier pour les structures petites ou ambiguës. Des études récentes [90, 91, 92, 93] confirment que l'intégration d'une attention spatiale et canal améliore les performances sur les classes rares et permet de mieux distinguer des régions visuellement similaires (par exemple routes inondées vs rivières). Cependant, ces mécanismes restent coûteux en calcul et nécessitent de grandes quantités de données pour atteindre leur plein potentiel.

Une avancée notable dans la segmentation haute résolution est HRNet [74], qui rompt avec le paradigme encodeur–décodeur en maintenant des représentations haute résolution tout au long du réseau. Plutôt que de réduire puis reconstruire la résolution, HRNet traite des cartes de caractéristiques multi-résolutions en parallèle et échange continuellement des informations entre elles. Cette conception permet de capturer simultanément le contexte global et les détails fins. Dans le cadre des catastrophes, HRNet a été appliqué à la segmentation d'objets fins et allongés tels que les lignes électriques, les routes ou les canaux d'eau, souvent manqués par les CNN classiques. Toutefois, le coût computationnel élevé de ces représentations limite son utilisation dans des scénarios temps réel ou à ressources contraintes comme les drones.

L'introduction des architectures basées sur les Transformers a marqué un changement de paradigme en segmentation sémantique, en permettant une modélisation explicite des relations globales. Le modèle SETR [94] a remplacé l'encodeur CNN par un Vision Transformer pur, traitant l'image comme une séquence de patches et utilisant l'auto-attention pour capturer des dépendances à longue portée. Cette approche a amélioré la segmentation de structures à grande échelle telles que les zones inondées. Cependant, la complexité quadratique de SETR par rapport à la taille de l'image limite son utilisation avec des images haute résolution. Des travaux ultérieurs ont cherché à réduire cette complexité. Le Swin Transformer [95] introduit un mécanisme de fenêtres décalées (*shifted windows*), permettant de calculer l'attention localement tout en conservant une communication globale. Cette innovation réduit la complexité de quadratique à linéaire, tout en conservant une bonne modélisation du contexte global.

Les architectures hybrides CNN–Transformer ont émergé comme une solution efficace combinant les avantages des deux paradigmes : les biais inductifs locaux des CNN et la capacité de modélisation globale des Transformers. SegFormer [8] illustre cette approche en combinant un encodeur Transformer hiérarchique léger avec un décodeur MLP simple mais efficace, atteignant un excellent compromis entre précision et efficacité. De même, Swin-UNet [96] intègre un encodeur Swin Transformer avec un décodeur de type U-Net, améliorant la segmentation dans des scènes complexes caractérisées par des occlusions ou des variations de couverture du sol. Plusieurs travaux ont exploré ces approches hybrides dans le domaine de la télédétection et des catastrophes. Zhong et al. [97] ont proposé un réseau hybride combinant CNN et Transformers pour améliorer la délimitation des contours et réduire les erreurs de sur-segmentation. Zhao et al. [98] ont également combiné ces deux approches pour mieux capturer les changements de paysage induits par les catastrophes. Gao et al. [99] ont proposé STransFuse, un cadre de fusion combinant extraction locale par CNN et modélisation globale par Transformers, permettant d'améliorer significativement les performances, notamment dans les scénarios multi-capteurs.

Des travaux récents se sont également concentrés sur l'optimisation pour des applications réelles. Xu et al. [100] ont proposé une architecture Transformer légère adaptée aux environnements contraints, tandis que Liu et al. [101] ont combiné Swin Transformer avec UPerNet pour améliorer la segmentation de couverture terrestre. Li et al. [102] ont étendu ces approches en intégrant des données multimodales (optique et SAR), améliorant la robustesse face au bruit et aux occlusions. La fusion multi-niveaux et multi-modale constitue également une direction prometteuse. Ma et al. [103] ont montré que les Transformers multi-niveaux permettent de capturer efficacement des structures complexes à différentes échelles. De plus, la fusion de données multi-capteurs améliore les performances dans des conditions difficiles telles que la couverture nuageuse ou les images nocturnes. Les approches hybrides ont également été étendues à la segmentation d'instances. Zhou et al. [85] ont proposé un modèle basé sur YOLOv5-Seg combiné avec des mécanismes d'attention Transformer pour l'évaluation post-sismique.

En résumé, les architectures haute résolution et les modèles hybrides CNN–Transformer représentent l'état de l'art en segmentation sémantique pour les scènes de catastrophe. En combinant précision spatiale fine et compréhension globale du contexte, ces modèles surpassent les approches purement convolutionnelles. Toutefois, des défis persistent, notamment le coût computationnel élevé, les besoins importants en données et les contraintes liées

Table 3.1: Positionnement comparatif de modèles représentatifs de segmentation sémantique et de travaux pertinents pour la segmentation d'images UAV, en lien avec la première contribution. CMS : contexte multi-échelle ; DS : détails spatiaux ; AR : attention ou raisonnement contextuel ; PCO : prise en compte des petits objets ; CR : classes rares ou sous-représentées ; GD : guidance par détection. ✓: explicitement traité ; △: partiellement traité ou indirectement lié ; ✗: non explicitement traité.

Travail	**CMS**	**DS**	**AR**	**PCO**	**CR**	**GD**	**Positionnement principal**
FCN [1]	✗	△	✗	✗	✗	✗	Introduit la segmentation dense entièrement convolutionnelle, mais reste limitée pour les contours fins, les petits objets et les scènes complexes.
U-Net [2]	△	✓	✗	△	✗	✗	Exploite une architecture encodeur–décodeur avec connexions de saut, efficace pour préserver les détails spatiaux, mais sans mécanisme spécifique pour les classes rares.
U-Net++ [54]	△	✓	✗	△	✗	✗	Améliore les connexions de saut par une structure imbriquée, mais reste principalement centrée sur la fusion spatiale plutôt que sur les régions difficiles.
DeepLabV3+ [3]	✓	△	△	△	✗	✗	Utilise l'ASPP et les convolutions dilatées pour capturer le contexte multi-échelle, mais les petits objets et les classes sous-représentées restent difficiles.
PSPNet [25]	✓	△	△	✗	✗	✗	Agrège le contexte global par pooling pyramidal, mais ne cible pas explicitement les classes rares ou les objets de petite taille.
SegFormer [8]	✓	△	✓	△	✗	✗	Combine une architecture Transformer hiérarchique et un décodeur léger, offrant un bon contexte global, mais sans guidance explicite pour les catégories difficiles.
YOLO + DeepLab [98]	△	△	✗	✓	△	✓	Combine détection et segmentation pour améliorer la localisation des objets d'intérêt, mais la détection reste souvent utilisée comme module externe ou pipeline découplé.
CSDNet (notre contribution)	✓	✓	✓	✓	✓	✓	Propose une segmentation guidée par la détection, combinant fusion multi-échelle, attention légère et intégration directe de caractéristiques de détection pour renforcer les classes rares, les petits objets et les régions ambiguës.

au déploiement en temps réel. Les recherches actuelles s'orientent vers des architectures plus légères, multimodales et optimisées pour les systèmes embarqués.

Positionnement Le Tableau 3.1 positionne les principales familles de méthodes de segmentation sémantique par rapport à la première contribution de cette thèse. Il met en évidence que les architectures existantes abordent certains aspects importants, tels que le contexte multi-échelle, la préservation spatiale, l'attention ou l'efficacité computationnelle. Cependant, peu d'approches traitent conjointement les classes rares, les petits objets, les régions ambiguës et l'intégration explicite de signaux guidés par la détection dans un cadre de segmentation dense.

3.3 Segmentation sémantique inter-domaine

La segmentation sémantique inter-domaine vise à améliorer la robustesse des modèles lorsque les données d'entraînement et de test proviennent de distributions visuelles différentes. Dans le cadre de l'adaptation de domaine, ce problème est généralement abordé en transférant les connaissances d'un domaine source annoté vers un domaine cible non annoté ou faiblement annoté. À l'inverse, la généralisation de domaine cherche à améliorer la robustesse à des domaines non observés durant l'entraînement, sans utiliser d'échantillons du domaine cible. Ces deux cadres sont particulièrement importants pour l'analyse d'images UAV en contexte de catastrophe, où les variations de capteurs, d'altitude de vol, d'angle de vue, d'illumination, de conditions météorologiques et d'apparence environnementale peuvent fortement affecter les performances des modèles [51, 104, 105].

Les premières méthodes d'adaptation de domaine pour la segmentation sémantique se sont principalement concentrées sur la réduction des écarts visuels entre les domaines source et cible, notamment par l'alignement adversarial des caractéristiques, la traduction d'images, la minimisation d'entropie, l'auto-apprentissage et le raffinement de pseudo-étiquettes. Des approches plus récentes exploitent également l'alignement des caractéristiques, l'adaptation fondée sur les prototypes et l'apprentissage multi-résolution du domaine cible afin d'améliorer la cohérence sémantique entre domaines [20, 106]. En parallèle, les méthodes de généralisation de domaine cherchent à améliorer la robustesse sans utiliser d'échantillons du domaine cible, en s'appuyant sur l'apprentissage de représentations invariantes, la normalisation, la diversification de style ou les perturbations fréquentielles [18, 52, 19, 107]. Dans le contexte de la segmentation sémantique, les modèles de fondation visuels, tels que DINOv2, peuvent être utilisés comme extracteurs de caractéristiques pré-entraînés afin de fournir des représentations visuelles génériques à une tête de segmentation. Ces représentations, apprises à grande échelle, peuvent ensuite être adaptées à des tâches de prédiction dense telles que la segmentation sémantique [108]. Des travaux récents tels que VDD [109] soulignent la grande diversité visuelle présente dans les images UAV, tandis que les approches récentes de segmentation avec généralisation de domaine exploitent davantage les modèles de fondation, le pré-entraînement géospatial et les priors sémantiques, comme GeoSA-BaSA [110], CrossEarth [111] et DPMFormer [112].

Malgré leur efficacité face aux décalages visuels et aux variations liées aux capteurs, la plupart des méthodes de segmentation inter-domaine supposent que les domaines source et cible partagent un espace d'étiquettes identique ou directement compatible. Cette hy-

pothèse est souvent violée dans les images UAV de catastrophes, où des éléments similaires peuvent être annotés selon des critères et des niveaux de granularité différents, tels que l'état d'inondation, le degré de dommage structurel, l'accessibilité des routes ou la présence d'obstacles. Dans ce contexte de taxonomies hétérogènes, la correspondance entre classes devient ambiguë. Restreindre l'apprentissage aux seules catégories communes entraîne une perte d'information spécifique aux datasets, tandis que fusionner toutes les classes dans une taxonomie plate unique peut introduire des conflits sémantiques et une supervision instable. Ainsi, une segmentation UAV inter-domaine robuste doit traiter non seulement le décalage de distribution visuelle, mais aussi le décalage de taxonomie sémantique.

3.3.1 Apprentissage avec taxonomies et étiquettes hétérogènes

Au-delà des décalages visuels, les modèles de segmentation entraînés sur plusieurs jeux de données doivent également faire face à des taxonomies d'annotation hétérogènes, issues de protocoles de labellisation, de granularités sémantiques et d'objectifs applicatifs différents. Ce problème est particulièrement important dans l'imagerie UAV de catastrophes, où les mêmes éléments de scène peuvent être annotés selon l'état d'inondation, le niveau de dommage, l'accessibilité ou le blocage. Une solution courante consiste à harmoniser les datasets dans une taxonomie partagée, comme dans MSeg [113], mais ces approches sont difficiles à étendre et peuvent sacrifier des informations spécifiques aux datasets. Des méthodes plus récentes, telles que UniSeg [22], apprennent directement à partir d'espaces d'étiquettes hétérogènes en modélisant les relations et les conflits entre catégories spécifiques aux datasets, soulignant l'importance d'exploiter la structure sémantique plutôt que d'imposer une unification stricte des labels.

Des problématiques similaires ont également été étudiées dans l'adaptation universelle de domaine et la segmentation sensible aux taxonomies. UniDA-SS [114], DynAlign [32], Automatic Universal Taxonomies [115] et Automated Label Unification [116] cherchent à réduire les incohérences entre espaces d'étiquettes à travers l'apprentissage par prototypes, l'alignement de taxonomies ou l'unification automatique des labels. Bien que ces méthodes montrent l'intérêt de modéliser les relations sémantiques, elles se concentrent principalement sur la réconciliation taxonomique et ne traitent pas explicitement le défi conjoint du décalage visuel entre domaines et du décalage de taxonomie sémantique rencontré dans les datasets UAV de catastrophes.

Malgré ces avancées, la plupart des méthodes sensibles aux taxonomies se concentrent sur la réconciliation des espaces d'étiquettes plutôt que sur l'adaptation inter-domaine sous décalages visuels et sémantiques simultanés. Dans l'imagerie UAV de catastrophes, il est pourtant essentiel de préserver les attributs spécifiques aux datasets, tels que l'état d'inondation, le niveau de dommage ou l'accessibilité des routes. C'est pourquoi nous séparons les superclasses partagées des sous-classes spécifiques aux domaines, afin d'apprendre des concepts communs tout en conservant les informations spécialisées liées aux catastrophes. Cette formulation traite conjointement le décalage visuel entre domaines et le décalage de taxonomie sémantique.

3.3.2 Apprentissage de représentations sémantiques fondé sur les prototypes

L'apprentissage fondé sur les prototypes structure l'espace des caractéristiques autour d'ancres sémantiques de classes, généralement définies comme la moyenne des représentations associées à une catégorie. Initialement introduit dans le contexte de l'apprentissage few-shot avec les Prototypical Networks [53], ce principe a ensuite été étendu à la segmentation sémantique [117] et à l'adaptation de domaine [20]. En associant des pixels ou des régions à des prototypes de classes, ces méthodes fournissent des représentations sémantiques explicites qui améliorent la compacité des caractéristiques, réduisent l'effet des pseudo-étiquettes bruitées et renforcent la cohérence sémantique entre domaines. Les prototypes constituent ainsi des ancres efficaces pour transférer les connaissances et stabiliser l'apprentissage sous décalage de domaine.

L'apprentissage guidé par prototypes a également montré son intérêt en segmentation d'images de télédétection, où les prototypes peuvent soutenir la génération de pseudo-étiquettes et l'alignement inter-domaine des caractéristiques [29]. En parallèle, l'apprentissage contrastif au niveau pixel améliore l'organisation de l'espace sémantique en favorisant la compacité intra-classe et la séparation inter-classe [118]. Cependant, la plupart des méthodes fondées sur les prototypes ou l'apprentissage contrastif supposent un espace de labels fixe et partagé entre domaines. Cette hypothèse devient restrictive pour les datasets UAV de catastrophes, où les taxonomies sémantiques sont souvent hétérogènes et comportent à la fois des concepts partagés et des attributs spécifiques aux domaines.

Ainsi, les prototypes ne doivent pas uniquement servir à aligner des caractéristiques sous un espace de labels identique, mais aussi à transférer des connaissances sémantiques entre

taxonomies partiellement compatibles. Dans les images UAV de catastrophes, des classes partagées telles que *building* ou *road* peuvent correspondre à différents attributs selon les datasets, comme l'état d'inondation, le degré de dommage ou l'accessibilité. Un prototype unique peut alors être insuffisant pour représenter l'ensemble de ces variations sémantiques. Contrairement aux méthodes existantes fondées sur les prototypes, notre approche utilise les prototypes non seulement pour la compacité et l'alignement des caractéristiques, mais aussi pour relier les super-classes partagées aux sous-classes spécifiques aux domaines, permettant un transfert sémantique à travers des taxonomies hétérogènes tout en préservant les informations fines liées aux catastrophes.

3.3.3 Segmentation sémantique hiérarchique et conditionnelle

La segmentation sémantique conventionnelle formule la prédiction dense comme une classification pixel par pixel dans un espace de labels plat, ignorant souvent les relations hiérarchiques entre concepts généraux et attributs plus fins. Pour mieux exploiter cette structure sémantique, plusieurs travaux ont intégré des mécanismes de raisonnement contextuel, de raffinement progressif ou de supervision hiérarchique. Par exemple, EncNet [119] exploite le contexte sémantique global pour guider l'apprentissage des caractéristiques, Label Refinement Network [120] améliore progressivement les prédictions par un raffinement grossier-vers-fin, et les formulations de pertes hiérarchiques [121] exploitent les relations entre classes à plusieurs niveaux d'abstraction. Ces travaux montrent que la modélisation de la structure sémantique peut améliorer la segmentation au-delà des formulations classiques fondées sur un espace de labels plat.

Des méthodes hiérarchiques plus récentes montrent également l'intérêt de modéliser explicitement les relations entre classes. HSSN [122] exploite les hiérarchies de classes pour régulariser les prédictions, tandis que les pertes hiérarchiques taxonomiques [123] utilisent les relations sémantiques pour apprendre à partir d'annotations hétérogènes ou partiellement spécifiées. Toutefois, ces approches supposent généralement une taxonomie unique et cohérente entre domaines, ce qui est restrictif pour les images UAV de catastrophes, où différents datasets fournissent souvent des annotations fines différentes pour les mêmes concepts de haut niveau.

Pour répondre à cette limite, nous adoptons une formulation hiérarchique conditionnelle qui prédit d'abord des super-classes partagées, puis raffine les sous-classes spécifiques aux domaines uniquement dans les régions sémantiquement valides. Contrairement aux ap-

proches hiérarchiques existantes, notre cadre combine prédiction hiérarchique, adaptation inter-domaine et alignement sémantique guidé par prototypes afin d'apprendre à partir de taxonomies UAV hétérogènes tout en préservant les informations propres à chaque domaine.

3.3.4 Travaux liés à l'adaptation inter-domaines et aux taxonomies hétérogènes

Le Tableau 3.2 synthétise le positionnement de méthodes représentatives par rapport au cadre proposé. Les approches existantes d'adaptation de domaine et de généralisation de domaine traitent principalement les décalages visuels entre domaines à travers l'alignement adversarial [16, 124], l'auto-apprentissage et les pseudo-étiquettes [17, 106], la normalisation et la diversification de style [18, 52], les perturbations fréquentielles [19, 107], ou encore les représentations fondées sur des modèles de fondation [110, 111, 112]. Cependant, ces méthodes supposent généralement des taxonomies sémantiques cohérentes entre les domaines. En parallèle, les méthodes sensibles aux taxonomies abordent les incohérences entre espaces d'étiquettes par la réconciliation taxonomique, l'unification des labels ou le regroupement sémantique [113, 22, 115, 116, 32], tandis que les approches fondées sur les prototypes et la segmentation hiérarchique mettent en évidence l'intérêt des ancres sémantiques et des relations structurées entre classes [20, 29, 122, 123]. Contrairement à ces travaux, notre cadre unifie ces directions dans un contexte de segmentation UAV inter-domaine, en traitant conjointement le décalage visuel entre domaines et le décalage de taxonomie sémantique à travers la modélisation de super-classes partagées, le raffinement conditionnel, le transfert sémantique guidé par prototypes et les correspondances partielles entre sous-classes.

Table 3.2: Positionnement comparatif des travaux représentatifs. DVS : décalage visuel entre domaines ou changement d'apparence ; DTS : décalage de taxonomie sémantique entre domaines source et cible ; SCP : modélisation de super-classes partagées ; AP : alignement fondé sur les prototypes ; PH : prédiction hiérarchique ; RC : raffinement conditionnel ; CPS : correspondance partielle entre sous-classes hétérogènes. ✓: explicitement traité ; △: partiellement traité ou indirectement lié ; ✗: non explicitement traité.

Travail	**DVS**	**DTS**	**SCP**	**AP**	**PH**	**RC**	**CPS**	**Positionnement principal**
Adaptation/généralisation de domaine et robustesse visuelle								
AdaptSegNet (2018) [16]	✓	✗	✗	✗	✗	✗	✗	Réalise un alignement adversarial dans l'espace de sortie pour l'adaptation non supervisée de domaine en segmentation sémantique.
ADVENT (2019) [124]	✓	✗	✗	✗	✗	✗	✗	Utilise la minimisation d'entropie et l'apprentissage adversarial pour adapter les modèles de segmentation au domaine cible.
FDA (2020) [19]	✓	✗	✗	✗	✗	✗	✗	Réduit les décalages visuels de bas niveau par transfert de style dans le domaine fréquentiel.
ProDA (2021) [20]	✓	✗	✗	✓	✗	✗	✗	Utilise des prototypes pour guider l'apprentissage de la structure du domaine cible et le raffinement des pseudo-étiquettes en adaptation de domaine.
DAFormer (2022) [17]	✓	✗	✗	✗	✗	✗	✗	Introduit un cadre UDA fondé sur les Transformers avec une stratégie robuste d'auto-apprentissage pour la segmentation sémantique.
HRDA (2022) [106]	✓	✗	✗	✗	✗	✗	✗	Traite l'adaptation non supervisée de domaine à haute résolution pour la segmentation sémantique sous décalage visuel.
MIC (2023) [21]	✓	✗	✗	✗	✗	✗	✗	Améliore la robustesse en UDA grâce à une régularisation par cohérence d'images masquées.
IBN-Net (2018) [18]	✓	✗	✗	✗	✗	✗	✗	Améliore la robustesse aux variations d'apparence grâce à la combinaison de normalisation par instance et par lot.
MixStyle (2021) [52]	✓	✗	✗	✗	✗	✗	✗	Simule des domaines visuels non vus en mélangeant les statistiques de style au niveau des caractéristiques.
GeoSA-BaSA (2025) [110]	✓	✗	✗	✗	✗	✗	✗	Exploite des représentations fondées sur des modèles de fondation pour la segmentation en télédétection avec généralisation de domaine.
DPMFormer (2025) [112]	✓	✗	✗	✗	✗	✗	✗	Exploite le guidage vision-langage pour la segmentation sémantique avec généralisation de domaine sous espace d'étiquettes partagé.
Espaces d'étiquettes hétérogènes et réconciliation taxonomique								
MSeg (2020) [113]	△	✓	△	✗	△	✗	△	Réalise une harmonisation manuelle des taxonomies pour la segmentation sémantique multi-dataset.
UniSeg (2022) [22]	△	✓	△	✗	△	✗	△	Apprend à partir d'espaces d'étiquettes hétérogènes et d'annotations conflictuelles entre datasets.
Automatic Universal Taxonomies (2022) [115]	✗	✓	✓	✗	✓	✗	△	Construit des taxonomies universelles à partir de relations de type sous-ensemble–sur-ensemble entre labels.
Automated Label Unification (2024) [116]	✗	✓	△	✗	△	✗	△	Utilise une unification des labels fondée sur des graphes pour réconcilier des taxonomies hétérogènes.
UniDA-SS (2025) [114]	△	△	△	✓	✗	✗	△	Aborde l'adaptation universelle de domaine avec classes communes et privées à l'aide d'une distinction par prototypes.
DynAlign (2025) [32]	△	✓	△	✗	△	✗	△	Aligne dynamiquement les taxonomies afin de gérer de nouveaux espaces d'étiquettes dans le domaine cible.
Apprentissage de représentations sémantiques fondé sur les prototypes								
Prototypical Networks (2017) [53]	✗	✗	✗	✓	✗	✗	✗	Introduit des prototypes de classes pour des représentations métriques compactes.
PANet (2019) [117]	✗	✗	✗	✓	✗	✗	✗	Utilise une correspondance prototype-pixel pour la segmentation sémantique en few-shot.
Pixel Contrast (2021) [118]	△	✗	✗	△	✗	✗	✗	Organise les caractéristiques sémantiques par apprentissage contrastif au niveau pixel.
Prototype-Guided RS Adaptation (2024) [29]	✓	✗	✗	✓	✗	✗	✗	Utilise une adaptation guidée par prototypes pour la segmentation sémantique en télédétection sous décalage visuel entre domaines.
Segmentation sémantique hiérarchique et conditionnelle								
Label Refinement Network (2017) [120]	✗	✗	✗	✗	△	△	✗	Réalise un raffinement spatial grossier-vers-fin pour la segmentation sémantique.
Hierarchical Loss (2020) [121]	✗	✗	△	✗	✓	✗	✗	Introduit une supervision hiérarchique utilisant plusieurs niveaux d'abstraction des classes.
HSSN (2022) [122]	✗	✗	△	✗	✓	△	✗	Exploite une segmentation sensible à la hiérarchie sous une taxonomie unique et cohérente.
Taxonomic Hierarchical Loss (2025) [123]	✗	△	△	✗	✓	✗	△	Utilise des pertes hiérarchiques pour apprendre à partir d'annotations taxonomiques hétérogènes ou partiellement spécifiées.
Notre cadre	✓	✓	✓	✓	✓	✓	✓	Cadre inter-domaine hiérarchique guidé par prototypes, traitant le décalage visuel entre domaines, les taxonomies UAV hétérogènes, la modélisation de super-classes partagées, l'alignement guidé par prototypes, la prédiction hiérarchique, le raffinement conditionnel et les correspondances partielles entre sous-classes.

Chapitre 4

Première contribution

4.1 Motivation et objectifs

Malgré les progrès récents en segmentation sémantique, l'analyse d'images aériennes dans des contextes réels demeure confrontée à plusieurs défis majeurs. Les images UAV présentent une résolution élevée et une forte variabilité d'échelle, nécessitant à la fois une compréhension globale du contexte et une capacité à capturer des détails fins [125, 126]. Cette contrainte augmente significativement la complexité computationnelle, limitant le déploiement sur des systèmes embarqués [127, 128]. Les scènes post-désastres naturels introduisent également des phénomènes d'occlusion et d'ambiguïté visuelle, où des structures importantes peuvent être partiellement masquées ou difficiles à distinguer [129, 12]. De plus, le déséquilibre des classes constitue un problème critique : les classes dominantes occupent la majorité des pixels, tandis que des objets essentiels mais rares, tels que les véhicules ou les bâtiments endommagés, sont sous-représentés. Enfin, la variabilité inter-domaines limite fortement la généralisation des modèles. Les différences entre datasets, capteurs et environnements entraînent une dégradation significative des performances lors du déploiement sur des données non vues.

Ces limitations motivent le développement de nouvelles approches capables de mieux structurer l'espace des représentations, de renforcer la discrimination des classes rares, et d'améliorer la robustesse face aux variations de domaine. Dans ce contexte, ce travail propose une approche progressive. Dans un premier temps, le modèle **CSDNet** est introduit afin d'améliorer la segmentation des classes sous-représentées en intégrant des informations guidées par la détection et des mécanismes légers d'attention. Dans un second temps, ce travail est étendu vers un cadre plus général intégrant des mécanismes de structuration

hiérarchique et de représentation par prototypes, visant à améliorer la généralisation inter-domaines et la cohérence sémantique dans des environnements complexes.

Dans le cadre de la segmentation de scènes de catastrophe, la détection précise des objets de petite taille et des classes sous-représentées demeure un défi majeur, souvent négligé. Des catégories essentielles telles que les véhicules, les débris, les piscines ou encore les structures inondées occupent une portion très limitée de l'image (voir Figure 4.1), tout en présentant une importance opérationnelle critique pour les interventions d'urgence [130]. Dans les contextes post-catastrophe naturelle, ces objets sont fréquemment occlus, visuellement dégradés ou dispersés dans des environnements complexes, ce qui complique davantage leur identification [131]. Par ailleurs, la dominance des classes majoritaires, telles que les routes, la végétation ou les bâtiments, introduit un biais dans les modèles de segmentation classiques. Ces derniers tendent à optimiser la précision globale au niveau des pixels, au détriment des classes rares mais cruciales [132]. Ce déséquilibre des classes, combiné à la forte variabilité d'échelle, limite l'efficacité des systèmes de segmentation dans des scénarios réels de gestion des catastrophes.

Afin de répondre à ces limitations, nous proposons **CSDNet**, un modèle de segmentation multi-classes conçu spécifiquement pour améliorer la représentation et la détection des classes sous-représentées, notamment les objets de petite taille ou peu fréquents. Dans de nombreux cas, les classes dominantes masquent les erreurs de segmentation sur ces catégories critiques, donnant une impression trompeuse de performance globale [9]. CSDNet intègre plusieurs modules spécialisés qui interagissent de manière complémentaire afin d'améliorer la précision de segmentation dans des scènes complexes et déséquilibrées. L'architecture globale, illustrée dans la Figure 4.2, repose sur trois composants principaux : (1) un module d'amélioration des classes ciblées, (2) un module d'attention contextuelle profonde, et (3) un module de fusion multi-échelle des caractéristiques. Ensemble, ces modules forment une architecture cohérente et synergique, adaptée à la segmentation de scènes de catastrophe UAV caractérisées par des ambiguïtés visuelles, des occlusions et un fort déséquilibre des classes. Les principales contributions de ce travail sont résumées comme suit :

- Nous proposons CSDNet, un modèle de segmentation innovant, spécifiquement conçu pour l'analyse des zones affectées par des catastrophes naturelles. CSDNet intègre trois contributions majeures : (i) un module Transformer léger permettant de capturer le contexte global, (ii) des convolutions séparables en profondeur (*Depthwise Separable Convolutions*, DWSC) pour améliorer l'efficacité computationnelle, et (iii) un mécanisme de fusion guidée par la détection injectant des informations issues d'un module de détection auxiliaire dans le pipeline de segmentation. Cette combinaison permet de

traiter efficacement les problèmes de déséquilibre des classes, de segmentation d'objets de petite taille et de discrimination entre classes visuellement similaires.

- Nous introduisons une stratégie de fusion multi-échelle qui combine les caractéristiques spatiales de bas niveau avec des caractéristiques sémantiques de niveau intermédiaire dans le décodeur. Cette approche améliore la précision des contours et la détection de structures fines et sous-représentées telles que les débris, les véhicules, les routes étroites ou les infrastructures endommagées. Afin de renforcer la séparation entre classes similaires (par exemple "toit intact" vs "toit effondré"), nous intégrons des caractéristiques guidées par la détection comme signaux d'attention spécifiques aux classes. Ces informations orientent le modèle vers des régions discriminantes et améliorent la robustesse de la segmentation dans des scènes UAV complexes.
- Nous évaluons CSDNet sur les datasets FloodNet [4] et RescueNet [5], deux ensembles de données UAV particulièrement exigeants, caractérisés par des conditions environnementales variées, des structures de scène complexes et un fort déséquilibre des classes (voir Figure 4.1). Les résultats expérimentaux montrent que notre approche surpasse systématiquement les modèles de référence, tant pour la segmentation des zones étendues que pour la détection d'objets critiques de petite taille. Ces résultats confirment l'efficacité de notre architecture en termes de robustesse, de précision des contours et de cohérence sémantique, faisant de CSDNet un outil pertinent pour l'analyse rapide de situations de catastrophe en conditions réelles.

Dans les sous-sections suivantes, nous détaillons l'implémentation de chacun des modules constituant l'architecture CSDNet.

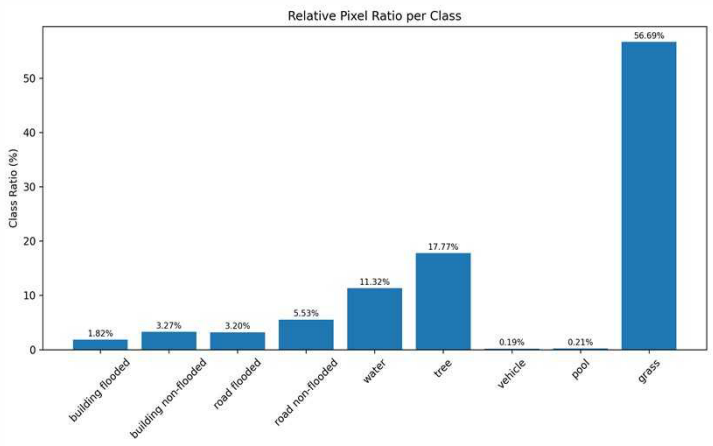
4.2 Méthodologie proposée

4.2.1 Module d'amélioration des classes ciblées

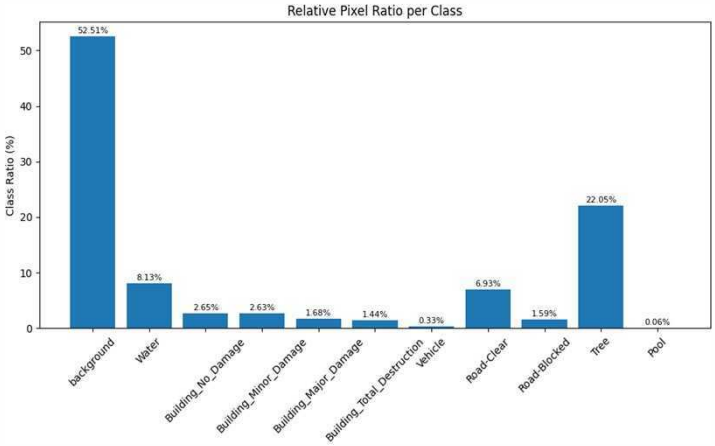
Dans les images aériennes à haute résolution, certaines classes sémantiques (par exemple bâtiments inondés vs non inondés) présentent une forte similarité visuelle et sont souvent sous-représentées dans les données d'entraînement. Ces facteurs contribuent à des performances de segmentation limitées pour ces classes critiques. Afin de répondre à ce problème, nous proposons une stratégie d'amélioration ciblée des classes, qui consiste à intégrer des

caractéristiques issues d'un modèle de détection spécifique aux classes dans le pipeline de segmentation sémantique.

À cette fin, nous exploitons le modèle RetinaNet [49], initialement conçu pour la détection d'objets, comme source de caractéristiques sémantiques riches et spécifiques aux classes, venant compléter le pipeline de segmentation. La détection fournit des indices à la fois globaux et locaux pour les objets sous-représentés, souvent négligés par les architectures encodeur–décodeur classiques. Plutôt que d'entraîner RetinaNet sur l'ensemble des classes de segmentation, nous limitons son apprentissage à un sous-ensemble ciblé de classes, noté $\mathcal{T}$, focalisé sur deux tâches critiques : la distinction entre routes inondées et non inondées, et entre bâtiments inondés et non inondés.



(**a**)



(**b**)

Figure 4.1: Distributions des classes dans les jeux de données FloodNet (**a**) et RescueNet (**b**) utilisés dans ce travail. Elles illustrent le fort déséquilibre entre les catégories, certaines classes telles que les véhicules et les piscines étant fortement sous-représentées.

Soit une image aérienne d'entrée $\mathcal{X} \in R^{H \times W \times 3}$, et le masque de segmentation associé $\mathcal{Y} \in \{1, ..., C\}^{H \times W}$, où C désigne le nombre total de classes. Un modèle de détection D_{θ_D} est entraîné uniquement sur les classes appartenant à $\mathcal{T}$, à partir d'annotations de type boîtes englobantes dérivées des masques de segmentation.

Après entraînement, les paramètres θ_D de D_{θ_D} sont gelés (mode évaluation) durant l'entraînement du modèle de segmentation. Le modèle de détection agit alors comme un extracteur de caractéristiques sensible aux classes, produisant des représentations sémantiques de haut niveau :

$$F_D = D_{\theta_D}^{\text{eval}}(\mathcal{X}) \in R^{H' \times W' \times d_D}, \tag{4.1}$$

où $H' \times W'$ représente la résolution spatiale de la carte de caractéristiques et d_D sa profondeur en canaux. Cette stratégie permet de réutiliser des caractéristiques stables et sémantiquement riches sans mettre à jour le modèle de détection.

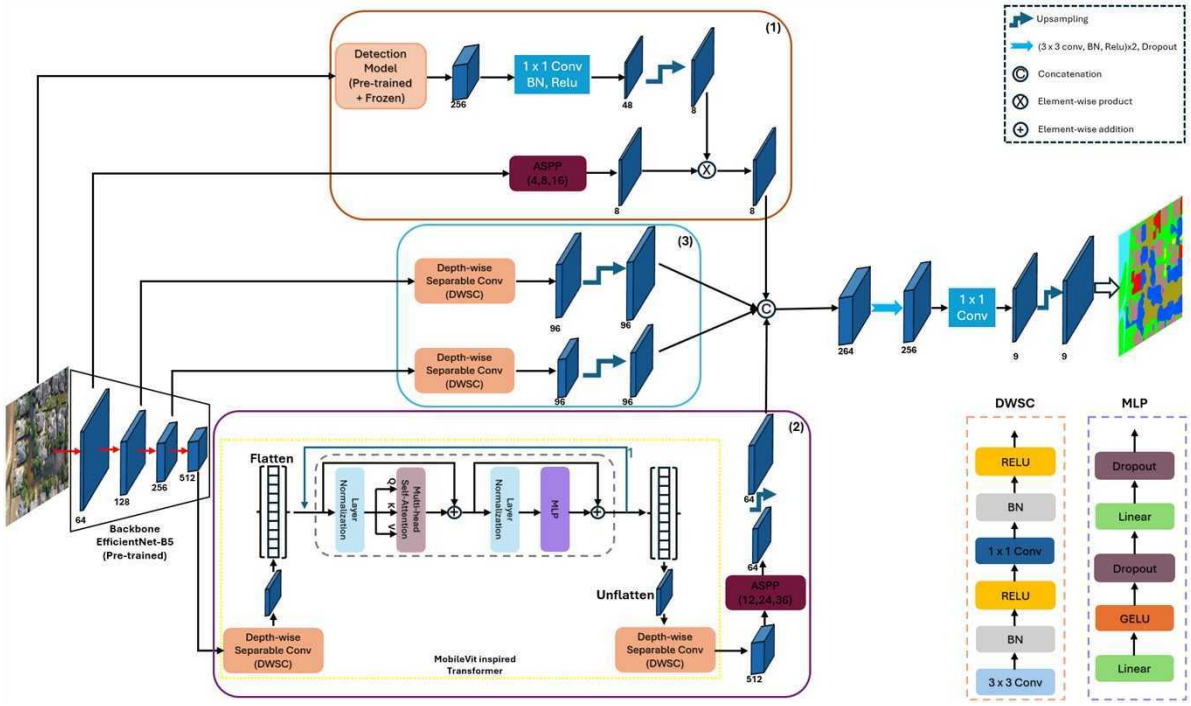


Figure 4.2: Architecture globale du modèle proposé : (1) module d'amélioration des classes ciblées, (2) module d'attention contextuelle profonde, et (3) module de fusion multi-échelle des caractéristiques.

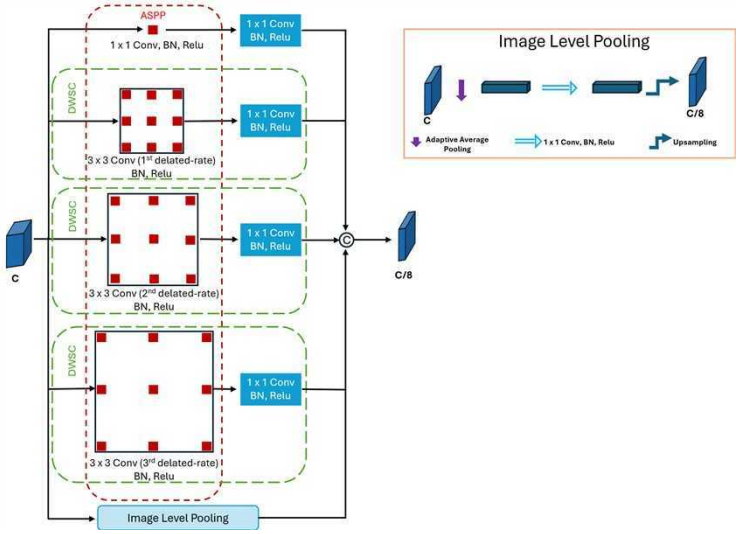


Figure 4.3: Illustration du module Atrous Spatial Pyramid Pooling (ASPP), introduit dans DeepLab v2 [24].

En parallèle du module de détection, l'image d'entrée est traitée par un encodeur de segmentation standard $\mathcal{B}$. Nous utilisons EfficientNet-B5, un backbone CNN de grande capacité pré-entraîné sur ImageNet, afin d'extraire des caractéristiques de bas et moyen niveaux. Ces caractéristiques sont notées :

$$F_1 = \mathcal{B}_1(\mathcal{X}) \in R^{\frac{H}{2}\times\frac{W}{2}\times d_1}, \quad F_2 = \mathcal{B}_2(F_1) \in R^{\frac{H}{4}\times\frac{W}{4}\times d_2},$$
$$F_3 = \mathcal{B}_3(F_2) \in R^{\frac{H}{8}\times\frac{W}{8}\times d_3}, \quad F_4 = \mathcal{B}_4(F_3) \in R^{\frac{H}{16}\times\frac{W}{16}\times d_4},$$

où $\mathcal{B}_l$ désigne le réseau tronqué jusqu'à la couche l, et d_l le nombre de canaux correspondants.

Afin d'améliorer la compréhension contextuelle, nous intégrons les caractéristiques issues de la détection F_D dans le pipeline de segmentation en alignant d'abord leurs dimensions via une transformation ϕ. Les caractéristiques résultantes sont ensuite fusionnées avec les cartes de caractéristiques de l'encodeur F_1 après application du module ASPP [3], selon :

$$F_{det} = \phi(F_D) \otimes \text{ASPP}(F_1), \tag{4.2}$$

où $\otimes$ désigne la multiplication élément par élément. Cette opération agit comme un mécanisme de filtrage (gating), amplifiant les activations spécifiques aux classes et guidant le modèle vers des régions discriminantes. Elle permet notamment de mieux capturer des distinctions subtiles, en particulier pour les classes rares ou ambiguës.

4.2.2 Module d'attention contextuelle profonde

Dans la segmentation sémantique multi-classes, un défi majeur provient du déséquilibre des classes, en particulier dans les jeux de données aériens à haute résolution, où certaines classes occupent des régions spatiales très limitées et apparaissent rarement dans l'ensemble des données [133]. Cela entraîne deux problèmes complémentaires : (1) le problème des petits objets, où certaines classes pertinentes ne couvrent qu'un faible nombre de pixels dans chaque image, et (2) la sous-représentation au niveau du dataset, où peu d'échantillons contiennent ces classes. Dans notre cas, ces deux difficultés sont présentes, rendant nécessaire l'amélioration de la compréhension du contexte global ainsi que le renforcement des représentations associées aux classes rares.

Pour répondre à ce problème, nous intégrons un mécanisme d'attention léger basé sur les Transformers, inspiré de MobileViT [134], correspondant au module d'attention contextuelle profonde illustré dans la Figure 4.2. Ce module est appliqué aux cartes de caractéristiques de haut niveau extraites du backbone de segmentation afin de capturer les relations entre des régions distantes de l'image appartenant potentiellement au même objet ou à la même classe. En intégrant ces interactions globales, le réseau acquiert une compréhension plus complète de la scène, ce qui est particulièrement important pour la segmentation de structures étendues ou de régions spatialement disjointes mais sémantiquement liées.

Dans un premier temps, nous appliquons un bloc de convolution séparable en profondeur (*Depthwise Separable Convolution*, DWSC) afin d'extraire des caractéristiques locales à partir de la dernière carte de caractéristiques du backbone :

$$F_{loc} = \mathrm{DWSC}(F_4) \in R^{\frac{H}{16} \times \frac{W}{16} \times d_4}. \tag{4.3}$$

Ensuite, les deux dimensions spatiales de ce tenseur sont réorganisées en une séquence de N tokens, tout en conservant la dimension des caractéristiques d_4. La séquence résultante est ensuite passée à travers un encodeur Transformer (TE) composé de blocs d'auto-attention multi-têtes et de réseaux feed-forward :

$$F_{\mathrm{attn}} = \mathrm{TE}\big(\mathrm{flatten}(F_{loc})\big) \in R^{N \times d_4}, \tag{4.4}$$

où $N = \frac{H}{16} \times \frac{W}{16}$ désigne le nombre de positions spatiales, chacune étant représentée par un vecteur de caractéristiques de dimension d_4. La sortie enrichie par l'attention est ensuite remise sous forme spatiale, puis traitée par un module ASPP afin de capturer des champs réceptifs multiples :

$$F_{global} = \mathrm{ASPP}\big(\mathrm{unflatten}(F_{\mathrm{attn}})\big). \tag{4.5}$$

La représentation résultante F_{global} fournit un encodage multi-échelle enrichi par l'attention, qui est ensuite fusionné avec les caractéristiques du décodeur. Cette intégration vise à renforcer la représentation des régions contenant des classes rares ou des objets de petite taille. En combinant l'attention globale et le raffinement contextuel multi-échelle, ce module favorise une meilleure modélisation des catégories fines et sous-représentées, dont les caractéristiques peuvent être difficiles à préserver en raison de leur faible présence spatiale ou du déséquilibre des données.

4.2.3 Module de fusion multi-échelle des caractéristiques

En complément des modules spécialisés — l'un dédié à la discrimination sémantique des classes difficiles et l'autre à l'amélioration des classes rares et des petits objets — nous introduisons un troisième composant agissant comme un module convolutionnel générique. Ce module est conçu pour traiter les classes bien représentées en exploitant les caractéristiques de niveau intermédiaire extraites du backbone. Son rôle est d'assurer une couverture globale et une cohérence sémantique pour les catégories fréquentes, tout en servant de pont entre les représentations de haut et de bas niveaux issues des autres modules.

Soient $F_2, F_3 \in R^{H_i \times W_i \times d_i}$ deux cartes de caractéristiques intermédiaires extraites du backbone EfficientNet-B5, avec $i = 2, 3$. Ces caractéristiques contiennent des informations sémantiques riches tout en conservant un niveau de détail spatial suffisant. Afin de permettre leur fusion, nous réduisons d'abord leur dimension à l'aide de convolutions ponctuelles (1×1), notées $\psi_i : R^{H_i \times W_i \times d_i} \to R^{H_i \times W_i \times d'}$. Les caractéristiques transformées sont alors :

$$\hat{F}_2 = \psi_2(F_2), \quad \hat{F}_3 = \psi_3(F_3), \quad \hat{F}_2, \hat{F}_3 \in R^{H_i \times W_i \times d'}. \tag{4.6}$$

Ces cartes de caractéristiques sont ensuite mises à l'échelle vers une résolution commune (upsampling), puis combinées — par concaténation ou addition — suivies d'un bloc convolutionnel $\mathcal{C}$ afin de produire une représentation intermédiaire raffinée :

$$F_{\mathrm{gen}} = \mathcal{C}\big[\uparrow(\hat{F}_2), \uparrow(\hat{F}_3)\big] \in R^{H_i \times W_i \times d''}, \tag{4.7}$$

où d'' correspond à la profondeur unifiée des caractéristiques, $\uparrow$ désigne l'opération d'upsampling, et $\mathcal{C}$ représente une opération de fusion suivie d'un traitement convolutionnel.

Cette représentation agit comme une composante sémantique générique, complémentaire aux sorties des modules spécialisés, tout en assurant une cohérence entre différentes échelles de représentation.

Enfin, les sorties des trois modules — les caractéristiques guidées par la détection F_{det}, les caractéristiques enrichies par l'attention globale F_{global}, et les caractéristiques convolutionnelles génériques F_{gen} — sont fusionnées pour constituer l'entrée finale du décodeur :

$$F_{\text{final}} = \mathcal{C}\big[F_{\text{det}}, F_{\text{global}}, F_{\text{gen}}\big], \tag{4.8}$$

où $\mathcal{C}$ correspond ici à une concaténation le long de la dimension des canaux. Cette représentation unifiée intègre des informations complémentaires issues des indices de détection, des dépendances globales à longue portée et des représentations convolutionnelles locales. Elle fournit ainsi au décodeur un contexte riche, multi-échelle et sémantiquement diversifié.

La représentation fusionnée est ensuite transmise au décodeur de segmentation afin de produire la prédiction dense finale $\hat{\mathcal{Y}}$. Cette stratégie permet d'assurer une segmentation précise à la fois des petits objets et des régions complexes ou visuellement ambiguës.

4.3 Entraînement du modèle, implémentation et fonction de perte

4.3.1 Stratégie d'entraînement et implémentation

Le cadre de segmentation proposé repose sur trois modules complémentaires : un module d'amélioration des classes basé sur la détection, un module d'attention dédié aux classes rares et aux petits objets, et un module convolutionnel générique pour la compréhension sémantique globale.

Afin de garantir la stabilité et la modularité du modèle, le module de détection est entraîné indépendamment sur un sous-ensemble de classes binaires ciblées (par exemple routes inondées vs non inondées, bâtiments inondés vs non inondés). Une fois entraîné, ce module est figé (mode évaluation) lors de l'apprentissage du modèle de segmentation, et est utilisé uniquement comme extracteur de caractéristiques sémantiques de haut niveau pour la distinction des classes ciblées.

Les autres composants, incluant le backbone (EfficientNet-B5), les modules d'attention, les couches de raffinement et le décodeur, sont entraînés de manière end-to-end à partir des annotations complètes de segmentation multi-classes. Le modèle est implémenté sous PyTorch (version 2.6.0), avec un backbone EfficientNet-B5 pré-entraîné sur ImageNet afin d'accélérer la convergence et d'exploiter des représentations génériques transférables.

Afin d'améliorer la robustesse et la capacité de généralisation, notamment dans des scénarios de catastrophe complexes, plusieurs techniques d'augmentation de données sont appliquées. Celles-ci incluent des retournements horizontaux et verticaux aléatoires, des rotations, des ajustements de luminosité et de contraste, ainsi que des variations d'échelle. Ces transformations permettent d'accroître la variabilité apparente des données d'entraînement en reproduisant certaines variations d'orientation, d'illumination et d'échelle susceptibles d'être observées dans les images UAV.

Par ailleurs, un équilibrage des lots d'entraînement est appliqué afin de garantir une représentation suffisante des classes sous-représentées, contribuant ainsi à atténuer le problème de déséquilibre des classes.

4.3.2 Fonction de perte

Afin d'optimiser le modèle de segmentation et de répondre aux défis liés au déséquilibre des classes et aux erreurs de localisation spatiale, nous adoptons une fonction de perte composite combinant la *Focal Loss* et la *Jaccard Loss* (ou perte IoU). Cette formulation permet de mettre l'accent sur les exemples difficiles tout en améliorant la cohérence régionale des prédictions.

La Focal Loss est utilisée pour atténuer l'impact du déséquilibre des classes en réduisant la contribution des exemples faciles et en focalisant l'apprentissage sur les pixels mal classifiés :

$$\mathcal{L}_{\text{Focal}}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{4.9}$$

où $p_t \in [0,1]$ représente la probabilité prédite pour la classe correcte, γ est le paramètre de focalisation (fixé à 2.0 dans nos expériences), et α_t est un facteur de pondération spécifique aux classes.

La Jaccard Loss (ou IoU loss) vise à améliorer le recouvrement spatial entre les prédictions et les annotations en pénalisant les erreurs au niveau des régions plutôt que pixel par pixel :

$$\mathcal{L}_{\text{Jaccard}} = 1 - \frac{|\hat{Y} \cap Y| + \varepsilon}{|\hat{Y} \cup Y| + \varepsilon}, \tag{4.10}$$

où $\hat{Y} \in \{0,1\}^{H \times W}$ représente le masque prédit et $Y \in \{0,1\}^{H \times W}$ le masque de vérité terrain. L'intersection $|\hat{Y} \cap Y|$ correspond aux pixels correctement prédits, tandis que l'union $|\hat{Y} \cup Y|$ représente l'ensemble des pixels présents dans l'un ou l'autre masque. Le terme ε (par exemple 10^{-6}) est introduit pour éviter toute division par zéro.

La fonction de perte globale est définie comme la somme des deux composantes :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{Jaccard}}. \tag{4.11}$$

Cette combinaison permet d'assurer simultanément une bonne discrimination entre classes et une cohérence spatiale des prédictions, améliorant ainsi les performances sur les classes dominantes comme sur les classes rares, en particulier dans des scènes de catastrophe caractérisées par des objets fragmentés et visuellement similaires.

L'Algorithme 1 résume le processus d'entraînement de CSDNet. Le module de détection est d'abord pré-entraîné puis figé, avant d'entraîner le reste du modèle pendant N_e époques.

Algorithm 1 Script d'entraînement de CSDNet

Entrée: Ensemble d'images $\mathcal{D}$ et classes ciblées $\mathcal{T}$;

Sortie: Modèle CSDNet entraîné

Étape 1 : Pré-entraîner le module de détection sur $\mathcal{T}$ et le figer;

Étape 2 : Pour epoch = 1 à N_e:

- Calculer F_{det}, using Equation (4.2) ;
- Calculer F_{global} using Equation (4.5) ;
- Calculer F_{gen} using Equation (4.7) ;
- Calculer F_{total} using Equation (4.8);
- Prédire $\hat{\mathcal{Y}}$;
- Rétropropager avec $\mathcal{L}_{\text{total}}$;

4.4 Expérimentations et résultats

Afin d'évaluer l'efficacité du modèle de segmentation proposé, nous avons mené une série d'expérimentations approfondies combinant des analyses quantitatives et qualitatives, en comparaison avec des méthodes de l'état de l'art. L'évaluation repose principalement sur la

métrique de l'Intersection over Union moyenne (mIoU), ainsi que sur les IoU par classe, en particulier pour les classes fréquentes et sous-représentées.

Les résultats qualitatifs permettent également d'illustrer la capacité du modèle à segmenter des objets de petite taille, des régions ambiguës et des frontières complexes dans des scènes de catastrophe. Dans l'ensemble, les résultats mettent en évidence les avantages de CSDNet pour améliorer la segmentation des classes critiques, en particulier celles peu représentées.

4.4.1 Jeux de données

Pour l'évaluation expérimentale, nous utilisons deux jeux de données de référence pour la segmentation de scènes de catastrophe : FloodNet [59, 4] et RescueNet [5].

FloodNet est un jeu de données UAV à haute résolution collecté après l'ouragan Harvey dans la région de Houston, conçu pour l'analyse post-catastrophe. Il fournit des annotations au niveau pixel pour plusieurs classes, incluant les routes et bâtiments (inondés et non inondés), les véhicules, les arbres, les piscines ainsi que le fond. Le dataset couvre des zones urbaines, suburbaines et semi-rurales, caractérisées par différents niveaux d'inondation.

Un défi majeur de FloodNet réside dans le déséquilibre des classes et la forte similarité visuelle entre certaines catégories, telles que les structures inondées et non inondées. De plus, la présence d'objets de petite taille, comme les véhicules et les piscines, en fait un benchmark particulièrement adapté pour évaluer les méthodes d'amélioration ciblée des classes et de fusion multi-échelle. FloodNet contient 2345 images, réparties en ensembles d'entraînement ($\sim$60%), de validation ($\sim$20%) et de test ($\sim$20%).

RescueNet est un autre jeu de données UAV à haute résolution, collecté dans des zones affectées par l'ouragan Michael. Il comprend dix classes thématiques : eau, bâtiment sans dommage, bâtiment légèrement endommagé, bâtiment fortement endommagé, bâtiment totalement détruit, véhicule, route praticable, route bloquée, arbre et piscine. Le dataset contient 4494 images, réparties en ensembles d'entraînement ($\sim$80%), de validation ($\sim$10%) et de test ($\sim$10%).

La tâche de segmentation dans ces deux jeux de données couvre à la fois des éléments structurels (par exemple bâtiments et routes inondés ou non) et des éléments environnementaux (eau, arbres, végétation, véhicules, piscines), essentiels pour l'analyse de l'impact des catastrophes. Comme dans de nombreux jeux de données réels, FloodNet et RescueNet présentent un fort déséquilibre des classes, certaines catégories minoritaires comme

les véhicules ou les piscines étant nettement moins représentées que des classes dominantes telles que les bâtiments ou les arbres. Ce déséquilibre, illustré dans la Figure 4.1, met en évidence la difficulté de segmenter avec précision les classes rares.

4.4.2 Configuration d'entraînement

Tous les modèles ont été implémentés sous PyTorch et entraînés pendant 100 époques sur un GPU NVIDIA RTX A6000. Les images d'entrée ont été redimensionnées de leur résolution originale 4000×3000 à 512×512 pixels afin de réduire la consommation mémoire et d'accélérer l'entraînement. Cette réduction permet de diminuer significativement la charge computationnelle, autorisant des tailles de batch plus élevées et des itérations plus rapides, ce qui est essentiel pour un développement efficace du modèle.

Cependant, cette réduction de résolution introduit un compromis, car certains détails spatiaux fins, en particulier les structures de petite taille ou fines, peuvent être perdus, affectant potentiellement la précision de segmentation. Afin d'atténuer cet effet, le modèle de détection RetinaNet est entraîné sur des images de plus haute résolution (1024×1024), à partir de boîtes englobantes dérivées des masques de segmentation des classes ciblées. Les cartes de caractéristiques produites sont ensuite redimensionnées à 512×512 pour être intégrées dans le pipeline de segmentation.

Dans l'architecture de segmentation, quatre niveaux de caractéristiques hiérarchiques (F_1 à F_4) sont extraits du backbone EfficientNet-B5. Les caractéristiques intermédiaires F_2 et F_3 sont traitées par le module convolutionnel générique, tandis que F_4 est passé à travers un module ASPP pour l'extraction du contexte global.

Le backbone EfficientNet-B5 est initialisé avec des poids pré-entraînés sur ImageNet et affiné de manière end-to-end. L'entraînement repose sur un optimiseur SGD avec un momentum de 0.9, un poids de régularisation (*weight decay*) de 1×10^{-4}, et un taux d'apprentissage initial de 0.01. Celui-ci est ajusté dynamiquement à l'aide d'un scheduler de type ReduceLROnPlateau (patience = 10, facteur = 0.1, taux minimal 1×10^{-4}).

Une stratégie d'échantillonnage pondéré (*WeightedRandomSampler*) est également utilisée pour atténuer le déséquilibre des classes, avec des poids calculés à partir de la distribution des pixels par classe. Des augmentations de données simples, incluant des retournements horizontaux et des rotations aléatoires, sont appliquées pour améliorer la robustesse du modèle.

4.4.3 Métriques d'évaluation

Afin d'évaluer les performances du modèle, nous utilisons les métriques standards de segmentation sémantique, en particulier l'Intersection over Union (IoU) par classe ainsi que la moyenne des IoU (mIoU).

L'IoU mesure le recouvrement entre les prédictions et les annotations de vérité terrain pour chaque classe. Pour une classe i, elle est définie comme suit :

$$\text{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{4.12}$$

où TP_i, FP_i et FN_i représentent respectivement les vrais positifs, faux positifs et faux négatifs pour la classe i.

La métrique mIoU correspond à la moyenne des IoU sur l'ensemble des classes :

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}_i, \tag{4.13}$$

où N est le nombre total de classes. Contrairement à la précision globale, qui favorise les classes dominantes, la mIoU fournit une évaluation équilibrée des performances sur l'ensemble des classes, y compris les catégories rares et sous-représentées.

4.4.4 Étude d'ablation

Afin d'évaluer l'impact de chaque composant architectural de notre modèle, nous avons mené une étude d'ablation sur le jeu de données FloodNet. Cette analyse consiste à supprimer successivement certains modules clés, notamment le module d'attention basé sur Transformer et le module de détection, afin d'évaluer leur contribution respective aux performances de segmentation.

Les résultats sont présentés dans le Tableau 4.1, où nous incluons également DeepLabV3+ comme modèle de référence.

Table 4.1: Étude d'ablation : impact de la suppression des composants clés sur l'IoU par classe (%) et la mIoU (%).

Classe	**Modèle complet**	**Sans Transformer**	**Sans RetinaNet**	**Sans Transformer + RetinaNet**	**DeepLabV3+**
Vehicle	58.62	34.62	47.91	31.32	44.25
Pool	66.12	48.12	60.22	47.67	51.80
Road Flooded	56.81	53.21	45.61	43.28	52.91
Building Flooded	75.77	73.14	65.87	48.42	47.24
mIoU	73.03	65.74	68.36	63.52	67.24

Effet de la suppression du module Transformer

La suppression du module Transformer entraîne une diminution d'environ 8% de la mIoU globale, avec un impact particulièrement marqué sur les classes de petite taille et sous-représentées. Par exemple, l'IoU de la classe "Vehicle" diminue d'environ 20%, tandis que celle de la classe "Pool" chute de 18%.

Ces résultats mettent en évidence le rôle essentiel de la modélisation du contexte global apportée par le Transformer. Le mécanisme d'auto-attention permet de capturer des dépendances à longue portée, souvent négligées par les architectures convolutionnelles classiques, qui se concentrent principalement sur des relations locales. Cette capacité est particulièrement importante pour la segmentation de structures de petite taille, souvent fragmentées ou visuellement ambiguës dans les scènes aériennes complexes.

En l'absence du module Transformer, la capacité du modèle à intégrer des informations provenant de régions distantes se dégrade, ce qui conduit à des erreurs de classification ou à l'omission d'objets pourtant critiques d'un point de vue opérationnel.

Effet de la suppression des caractéristiques issues de la détection

La suppression du module d'amélioration ciblée basé sur la détection entraîne une diminution d'environ 4% de la mIoU globale, avec un impact notable sur les classes visuellement similaires et spécifiques aux phénomènes d'inondation. En particulier, les classes "Road Flooded" et "Building Flooded" subissent une baisse d'environ 10% de leur IoU, tandis que leurs équivalents non inondés présentent une diminution d'environ 8%.

Ces résultats confirment l'intérêt d'intégrer des caractéristiques issues de la détection, qui agissent comme des *priors* sémantiques spécifiques aux classes. Ces informations guident le réseau de segmentation vers des indices visuels subtils permettant de distinguer des classes présentant une forte similarité spectrale et spatiale, comme les régions inondées et non inondées.

Dans des scénarios réels de catastrophe, ces distinctions sont cruciales, car les infrastructures inondées, les débris et les surfaces d'eau partagent souvent des caractéristiques visuelles proches, rendant leur différenciation particulièrement difficile sans guidance supplémentaire.

Effet de la suppression conjointe du Transformer et des caractéristiques de détection

La suppression simultanée du module Transformer et du module d'amélioration ciblée basé sur la détection entraîne une chute significative de la mIoU globale, atteignant 63.52%. Cette configuration dégradée affecte fortement la segmentation des classes critiques, notamment les objets de petite taille, les régions ambiguës et les catégories présentant des différences visuelles subtiles.

Ces résultats mettent en évidence le caractère complémentaire des deux mécanismes. Le Transformer apporte une compréhension globale et permet de modéliser des dépendances à longue portée, essentielles pour la segmentation d'objets fragmentés ou dispersés. En parallèle, les caractéristiques guidées par la détection introduisent des indices sémantiques spécifiques aux classes, améliorant la discrimination dans des régions complexes et fortement chevauchantes.

L'intégration conjointe de ces deux modules apparaît ainsi essentielle pour obtenir une segmentation robuste et précise dans des scènes aériennes de catastrophe, caractérisées par une forte variabilité, un déséquilibre des classes et des ambiguïtés visuelles importantes.

Afin d'illustrer l'impact de ces modules, la Figure 4.4 présente des exemples qualitatifs comparant différentes configurations d'ablation décrites dans le Tableau 4.1. Ces résultats visuels mettent en évidence la contribution de chaque module à la performance globale, en particulier dans des scénarios difficiles impliquant des objets de petite taille et des structures fines. L'absence de l'un ou l'autre module entraîne une dégradation visible de la précision des contours et de la localisation des objets, confirmant leur rôle complémentaire dans l'amélioration de la robustesse du modèle.

4.4.5 Résultats quantitatifs et analyse

Afin d'évaluer rigoureusement l'efficacité du modèle de segmentation proposé, nous avons mené une étude comparative approfondie avec huit modèles de segmentation sémantique de l'état de l'art. Ces modèles ont été sélectionnés en raison de leur diversité architecturale, de leurs performances reconnues et de leur pertinence pour des tâches de segmentation générales ainsi que pour des applications spécifiques au domaine.

Tous les modèles ont été entraînés et évalués dans des conditions expérimentales identiques sur les jeux de données FloodNet et RescueNet. L'évaluation repose sur la métrique de l'Intersection over Union moyenne (mIoU), ainsi que sur les IoU par classe, permettant d'obtenir à la fois une mesure globale des performances et une analyse détaillée au niveau des classes.

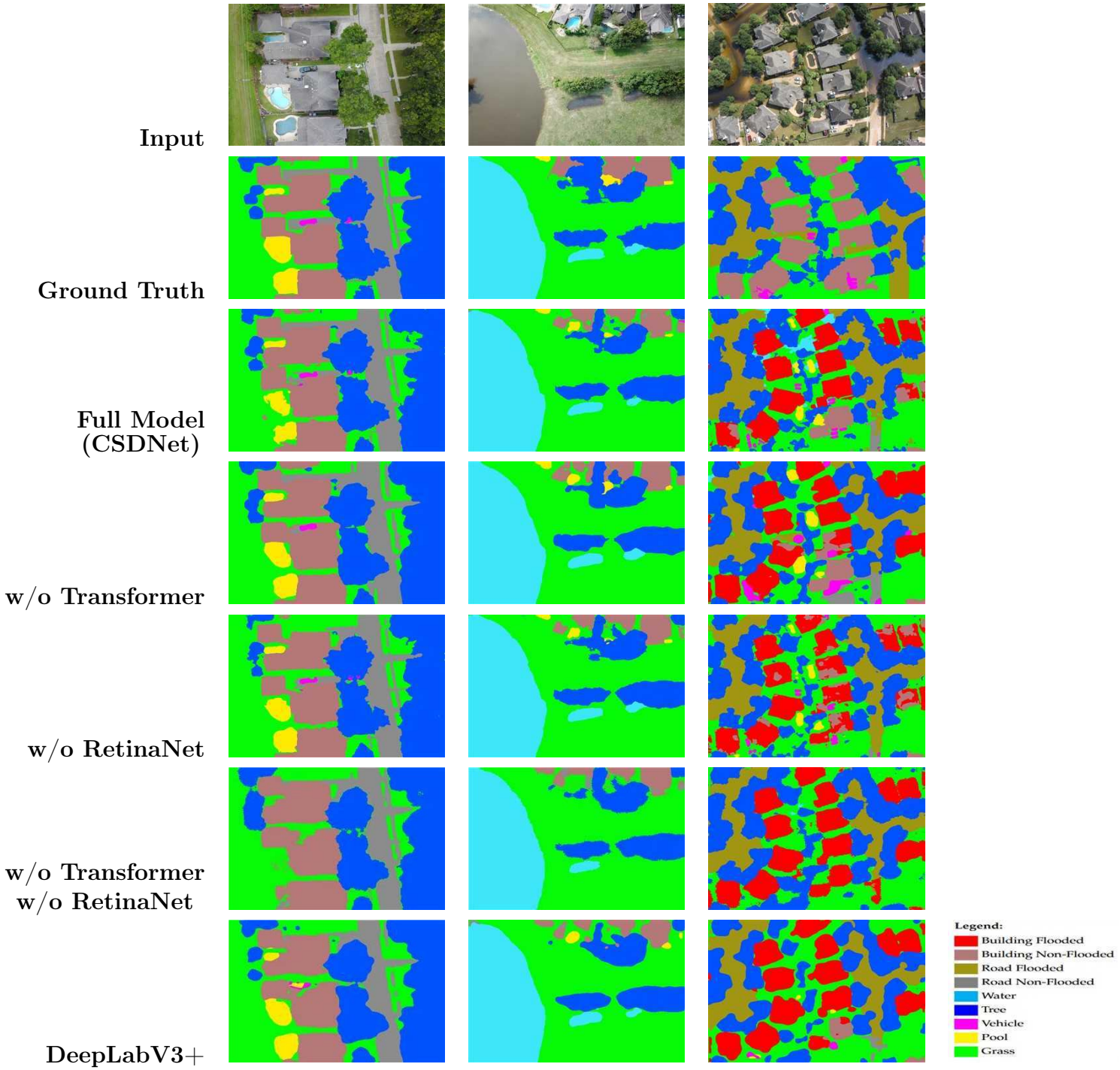


Figure 4.4: Comparaison qualitative des résultats de segmentation obtenus avec différentes configurations d'ablation du modèle. Chaque colonne illustre un échantillon de test issu du jeu de données FloodNet.

Les modèles inclus dans le benchmark sont les suivants : (1) SegFormer-B0 [8], une architecture légère basée sur les Transformers avec un encodage hiérarchique, reconnue pour sa bonne capacité de généralisation ; (2) DeepLabV3+ [3], un modèle convolutionnel robuste exploitant le module ASPP pour capturer le contexte multi-échelle, particulièrement efficace pour la segmentation de structures de grande taille ; (3) U-Net++ [54], une architecture encodeur–décodeur profondément imbriquée améliorant le U-Net classique grâce à des connexions de saut densément connectées ; (4) PSPNet [25], qui utilise un module de pooling pyramidal pour agréger des informations contextuelles à différentes échelles ; (5) ENet [135], un modèle de segmentation temps réel servant de référence pour des déploiements légers ; (6) TransUNet [136], combinant des représentations convolutionnelles et des mécanismes Transformer afin de capturer simultanément les dépendances locales et globales ; et (7) CMTFNet [137], intégrant des caractéristiques contextuelles et multi-échelles pour l'analyse de scènes de catastrophe.

Tous les modèles ont été ré-entraînés ou affinés à partir de poids pré-entraînés sur ImageNet, et configurés de manière homogène en termes de prétraitement des données, de résolution d'entrée (512×512) et de stratégies d'augmentation.

L'évaluation a été réalisée sur les ensembles de test, comprenant des scènes urbaines et semi-urbaines affectées par des inondations, présentant un niveau élevé de déséquilibre des classes ainsi que des distinctions fines entre objets.

Les résultats quantitatifs obtenus par notre modèle, ainsi que ceux des méthodes de référence, sont présentés dans les Tables 4.2 et 4.3 pour FloodNet et RescueNet, respectivement.

Table 4.2: Comparaison de l'IoU par classe (%) entre différents modèles et classes thématiques sur l'ensemble de test FloodNet(BF: flooded building, BNF: non-flooded building, RF: flooded road, and RNF: non-flooded road).

Model	**BF**	**BNF**	**RF**	**RNF**	**Water**	**Tree**	**Vehicle**	**Pool**	**Grass**	**mIoU**
SegFormer (B0)[8]	48.42	78.90	53.88	82.95	76.46	82.05	45.38	46.95	89.39	67.15
PSPNet[25]	44.18	74.53	49.60	80.14	75.05	78.55	28.40	39.89	87.54	61.99
ENet[135]	45.39	71.30	46.24	77.68	75.18	77.81	0.00	0.00	86.64	53.36
DeepLabV3+[3]	47.24	78.66	52.91	82.81	76.39	81.97	44.25	51.80	89.14	67.24
U-Net++[54]	57.31	76.18	45.73	80.13	65.58	79.28	26.45	47.37	86.59	62.74
TransUNet[136]	46.96	72.52	48.86	79.49	71.77	78.50	35.84	43.90	86.81	62.74
CMTFNet[137]	48.04	78.64	50.44	82.15	75.21	81.10	41.91	45.46	88.85	65.76
Our Model(CSDNet)	75.77	84.99	56.81	81.94	62.26	84.53	58.62	66.12	86.26	73.03

Table 4.3: Comparaison de l'IoU par classe (%) entre différents modèles et classes thématiques sur l'ensemble de test Rescuenet (BG: background, BND: building no damage, BMND: building minor damage, BMJD: building major damage, BTD: building total destruction, BR: blocked road and CR: clear road).

Model	**BG**	**Water**	**BND**	**BMND**	**BMJD**	**BTD**	**Vehicle**	**CR**	**BR**	**Tree**	**Pool**	**mIoU**
SegFormer(B0) [8]	82.76	83.86	62.95	53.33	51.48	53.41	50.65	74.50	41.22	81.12	63.02	61.55
PSPNet [25]	84.07	84.01	67.80	60.59	59.50	61.04	57.65	75.77	46.53	81.64	71.55	68.19
ENet [135]	76.35	74.84	45.12	36.02	31.10	41.67	0.00	52.43	16.98	74.12	0.00	40.78
DeepLabV3+ [3]	82.91	81.45	65.27	53.19	50.76	52.79	59.86	72.29	41.29	81.57	62.72	64.01
U-Net++ [54]	83.38	82.19	67.01	56.13	53.53	60.53	57.94	74.85	40.87	81.01	60.05	63.41
TransUNet [136]	78.81	77.93	47.68	38.11	30.12	40.59	41.40	67.22	19.02	76.71	31.91	49.95
CMTFNet [137]	83.84	82.70	66.00	56.69	55.74	60.40	54.49	74.29	40.50	81.96	56.70	64.84
Our Model(CSDNet)	84.60	84.71	67.98	59.17	58.61	62.04	60.35	76.77	48.73	82.57	74.07	69.05

Pour les deux jeux de données, CSDNet surpasse les architectures de référence, tant en termes de performance globale que sur les classes difficiles.

Pour FloodNet, notre modèle proposé atteint une mIoU de 73.03%, surpassant significativement des baselines solides telles que SegFormer (B0), qui obtient 67.15%, soit un gain absolu de près de 6%. Comparé à d'autres modèles largement utilisés, notre méthode montre des améliorations constantes, dépassant DeepLabV3+ (67.24%) d'environ 5.8%, et U-Net++ (62.74%) de plus de 10.2%.

La même tendance est observée sur le jeu de données RescueNet, où notre modèle atteint une mIoU de 69.05%, surpassant systématiquement les méthodes de référence. Plus précisément, il apporte un gain de 5.9% par rapport à DeepLabV3+ (63.12%), de 7.5% par rapport à SegFormer (B0) (61.55%), et dépasse U-Net++ (63.41%) de plus de 5.6%. Le modèle le plus proche, PSPNet, atteint une mIoU de 68.19%, grâce à son module de pooling pyramidal (PPM) qui capture efficacement le contexte multi-échelle, mais reste inférieur à notre approche d'environ 1%. Ces résultats confirment l'efficacité de notre architecture pour modéliser à la fois le contexte global et les ambiguïtés sémantiques dans des scènes de catastrophe complexes.

En se concentrant sur les classes ciblées, les améliorations sont encore plus marquées. Sur FloodNet, pour la classe *road flooded*, notre modèle atteint un IoU de 56.81%, dépassant le second meilleur score de 53.88% obtenu par SegFormer d'environ 3%. L'écart devient encore plus important pour la classe *building flooded*, où notre méthode atteint 75.77%, soit une amélioration de plus de 18% par rapport au meilleur concurrent, U-Net++ (57.31%). Une tendance similaire est observée sur RescueNet, où notre modèle obtient les meilleurs IoU pour des classes difficiles telles que *building-no-damage* et *building-total-destruction*. Ces

gains reflètent directement l'efficacité de notre stratégie d'amélioration ciblée des classes ainsi que du mécanisme de fusion guidée par la détection.

Notre modèle démontre également une grande robustesse sur les classes rares et de petite taille, particulièrement difficiles dans les scénarios de catastrophe. Notamment, des classes peu représentées comme *vehicle* (58.62% IoU) et *pool* (66.12% IoU) présentent des gains significatifs, attribués à l'intégration de caractéristiques issues de la détection qui guident la segmentation de manière spécifique aux classes. Le module d'auto-attention et l'ASPP renforcent également la modélisation du contexte global, améliorant la délimitation des contours dans des cas complexes, comme pour la classe *road flooded* atteignant 56.81% d'IoU, supérieure aux baselines. À titre d'exemple, pour la classe *vehicle*, des modèles comme ENet échouent totalement (0.00%) sur les deux jeux de données. Même des modèles plus performants comme SegFormer et DeepLabV3+ restent en dessous de notre modèle.

Il est important de noter que notre modèle maintient également des performances compétitives, voire supérieures, sur les classes fréquentes et bien représentées. Par exemple, sur FloodNet, notre modèle atteint 81.94% d'IoU pour *road non-flooded* et 84.53% pour la classe *tree*, montrant que les gains sur les classes rares ne se font pas au détriment des classes dominantes. La même observation s'applique à RescueNet, où notre modèle obtient de bonnes performances sur des classes dominantes telles que *water* et *tree*.

Dans l'ensemble, ces résultats montrent que notre modèle offre une performance de segmentation à la fois équilibrée et supérieure, excelling sur les classes fréquentes comme sur les classes rares. Il est particulièrement efficace pour traiter le déséquilibre des classes, les objets de petite taille et les régions visuellement ambiguës, caractéristiques des images de catastrophe en conditions réelles.

4.4.6 Analyse computationnelle

Le Tableau 4.4 compare notre modèle aux architectures existantes en termes de nombre de paramètres et de temps d'inférence. Malgré un nombre de paramètres significativement inférieur à celui de DeepLabV3+ et PSPNet, notre modèle offre le meilleur compromis entre précision et efficacité, atteignant la mIoU la plus élevée avec un temps d'inférence compétitif.

Les temps d'inférence rapportés correspondent au temps moyen de traitement (en ms) par image de taille 512×512 sur un GPU NVIDIA Tesla A100.

Table 4.4: Comparaison de l'efficacité des modèles : nombre de paramètres et temps d'inférence par image (sur un seul GPU NVIDIA A100).

Model	**Params (M)**	**Inference Time (ms)**	**mIoU (%)**
DeepLabV3+	39.64	9.14	67.24
SegFormer-B0	3.72	10.57	67.15
U-Net++	48.99	12.46	62.74
CMTFNet	30.07	19.69	65.76
ENet	0.35	16.32	53.36
TransUnet	107.68	36.47	62.74
PSPNet	53.58	12.25	61.99
Our Model(CSDNet)	33.52	22.84	73.03

Comme illustré dans la Figure 4.5, notre modèle proposé présente une augmentation plus rapide et plus stable de la métrique IoU pour les classes sous-représentées, telles que les classes *vehicle* et *pool*, comparativement aux architectures de référence. Ce gain de performance précoce, observable dès les premières époques d'entraînement, met en évidence l'efficacité de notre modèle à répondre à deux défis majeurs : (i) la segmentation fiable d'objets de petite taille et faiblement distribués, et (ii) la gestion du déséquilibre des classes.

En exploitant des caractéristiques issues de la détection, des mécanismes d'attention dynamiques et des interactions à longue portée entre caractéristiques, notre modèle est capable de capturer des contours fins et des indices sémantiques même dans des régions complexes ou à faible contraste. Par ailleurs, la convergence accélérée observée pour les classes minoritaires suggère que notre architecture n'améliore pas seulement les performances finales, mais facilite également un apprentissage plus efficace dès les premières phases d'optimisation — une propriété essentielle dans des contextes de déploiement contraints en temps ou en ressources, tels que les applications d'analyse de catastrophes.

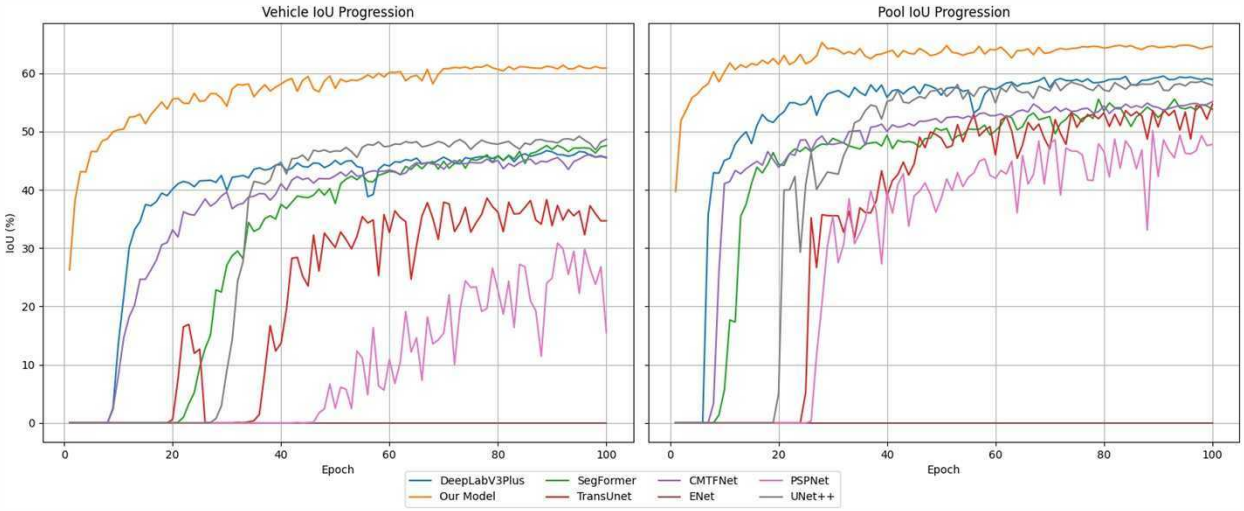


Figure 4.5: Progression de l'IoU sur l'ensemble de validation au cours de 100 époques pour les classes "vehicle" et "pool" sur le jeu de données FloodNet.

4.5 Résultats qualitatifs

Des comparaisons visuelles des masques de segmentation prédits sont présentées dans les Figures 4.6 et 4.8, respectivement. Ces figures montrent clairement que notre modèle proposé produit des contours plus précis, cohérents et spatialement plus fidèles, en particulier dans les régions contenant des objets de petite taille ou des structures complexes telles que les *vehicles*, les *pools* et les *flooded buildings*. Contrairement aux modèles de référence, qui génèrent souvent des prédictions fragmentées ou excessivement lissées dans ces zones, notre méthode parvient à capturer des détails fins tout en préservant la forme des objets.

Ces améliorations résultent de la conception synergique des différents modules de notre architecture. Le module d'attention favorise une meilleure localisation des objets de petite taille, tandis que les caractéristiques issues de la détection renforcent la capacité du modèle à distinguer des classes visuellement similaires ou chevauchantes — une propriété essentielle dans l'analyse de scènes de catastrophe, où les ambiguïtés visuelles sont fréquentes.

Enfin, il convient de noter que les jeux de données FloodNet et RescueNet présentent certaines imperfections d'annotation susceptibles d'induire en erreur les modèles de segmentation et d'affecter la fiabilité de l'évaluation. Dans de nombreux cas, nous observons que la qualité des annotations de vérité terrain peut avoir un impact significatif sur les résultats. La Figure 4.10 illustre des exemples où certaines régions importantes sont absentes des

annotations, notamment des objets de petite taille comme les véhicules ou les piscines (et, dans certains cas, des structures plus grandes telles que les bâtiments), ou encore des cas où certaines classes sont mal annotées, par exemple en représentant de manière incorrecte le niveau de dommage des bâtiments.

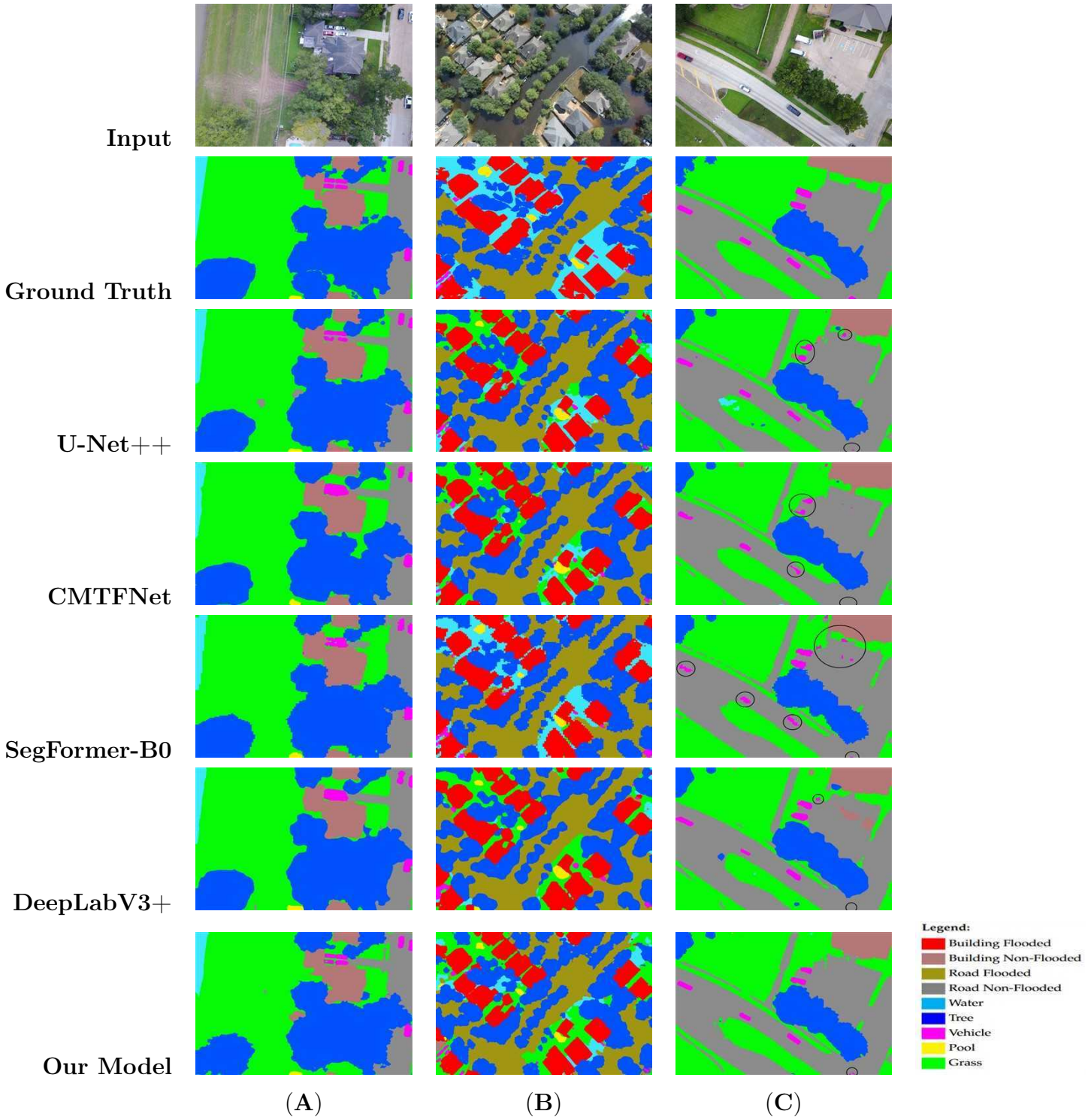


Figure 4.6: Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données FloodNet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 1).

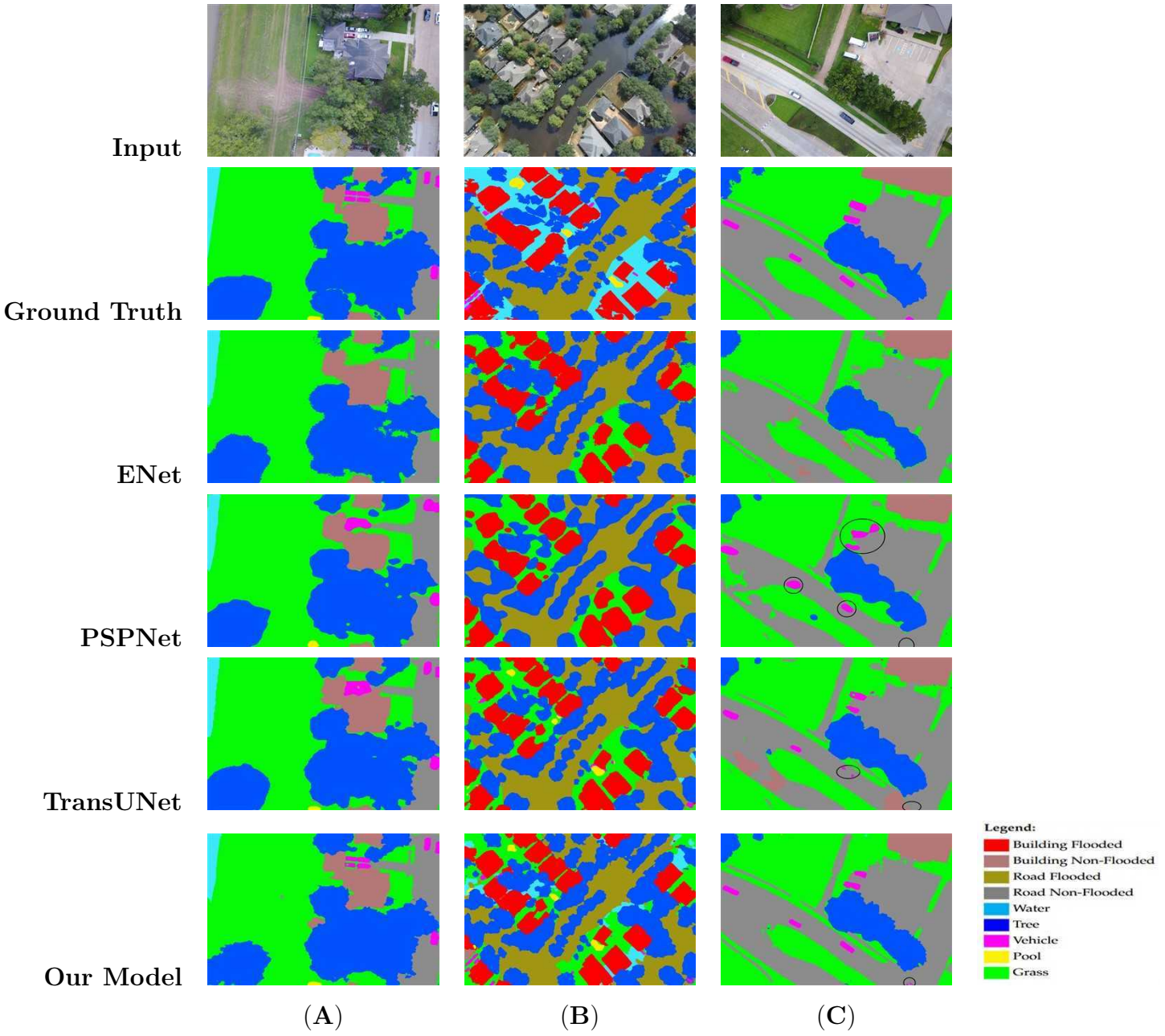


Figure 4.7: Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données FloodNet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 2).

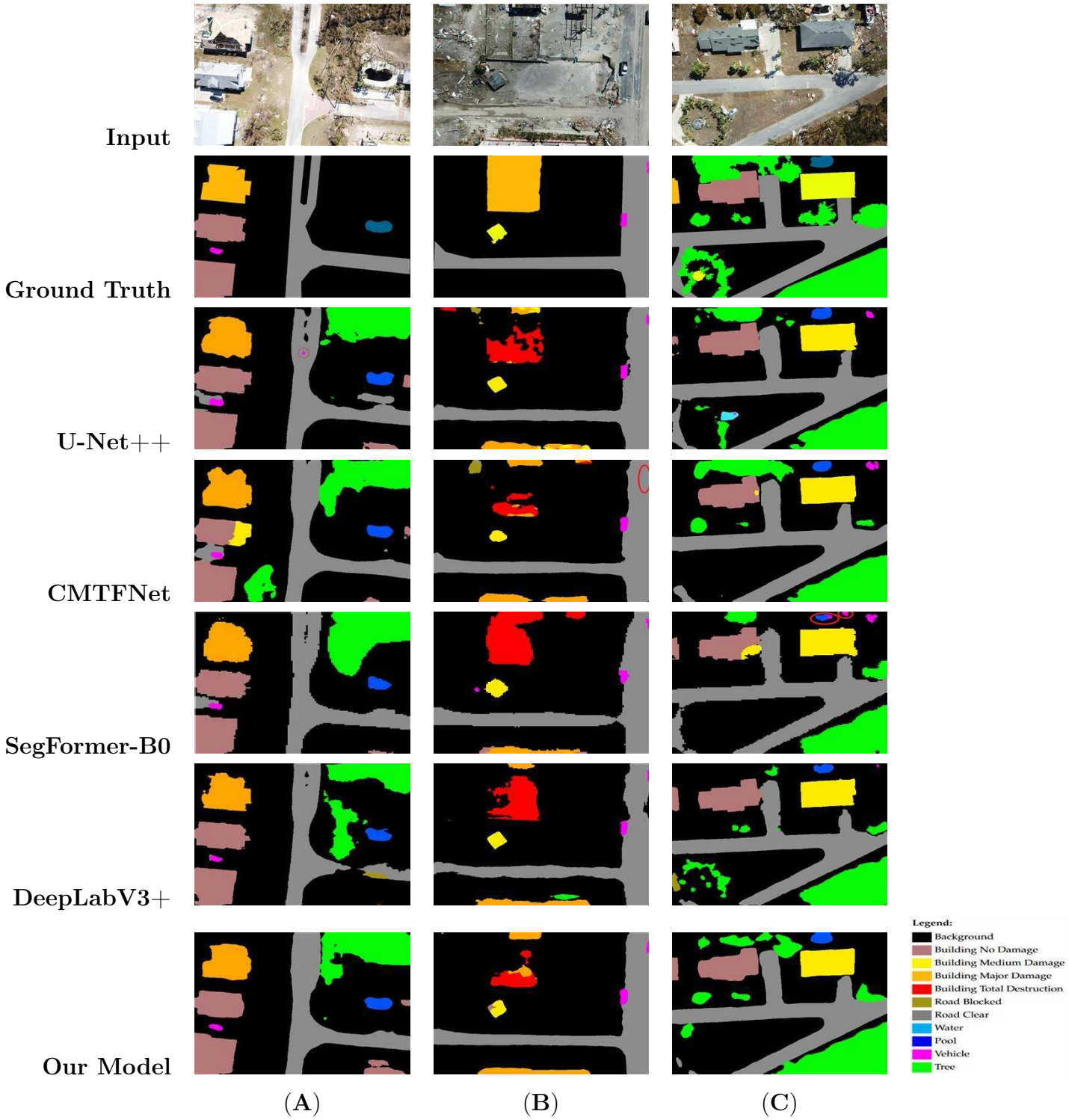


Figure 4.8: Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données Rescuenet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 1).

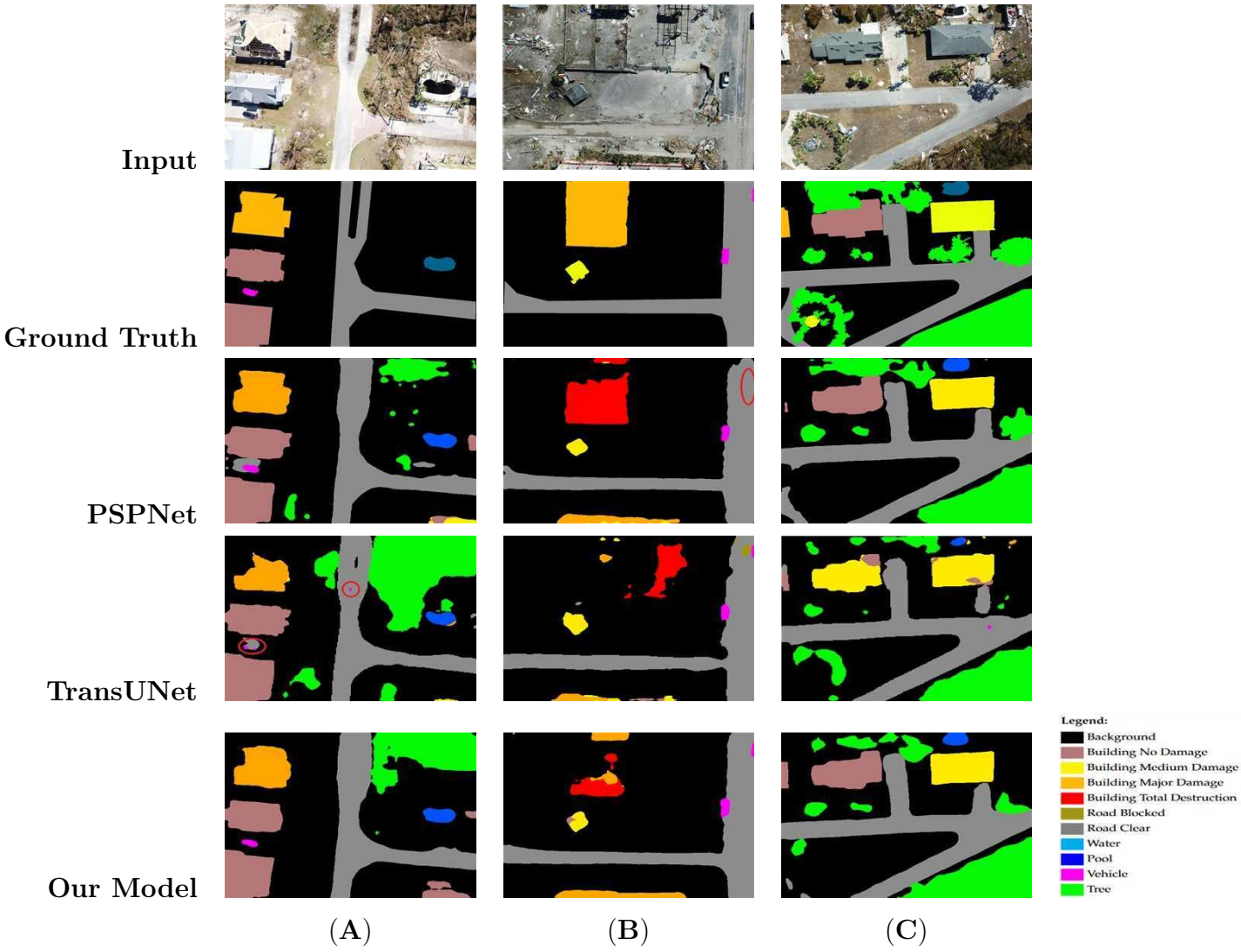


Figure 4.9: Exemples de segmentation obtenus avec les différentes méthodes comparées. Les colonnes (A), (B) et (C) correspondent à différents échantillons de test du jeu de données Rescuenet. Les différentes couleurs représentent les classes sémantiques, et des cercles noirs mettent en évidence les régions difficiles ou mal classifiées (Partie 2).

Ces imperfections introduisent du bruit lors de l'entraînement et peuvent induire les modèles en erreur, en particulier pour les classes rares ou visuellement similaires. Ces problèmes d'annotation peuvent conduire à une sous-estimation des performances du modèle pour les catégories concernées, lorsque des prédictions correctes s'écartent d'une vérité terrain imparfaite.

Cela met en évidence la nécessité de protocoles d'annotation plus fiables, ainsi que de modèles, tels que CSDNet, capables d'intégrer le contexte global et des indices spécifiques aux classes afin de compenser les imperfections des annotations et d'améliorer la segmentation dans des scénarios de catastrophe complexes.

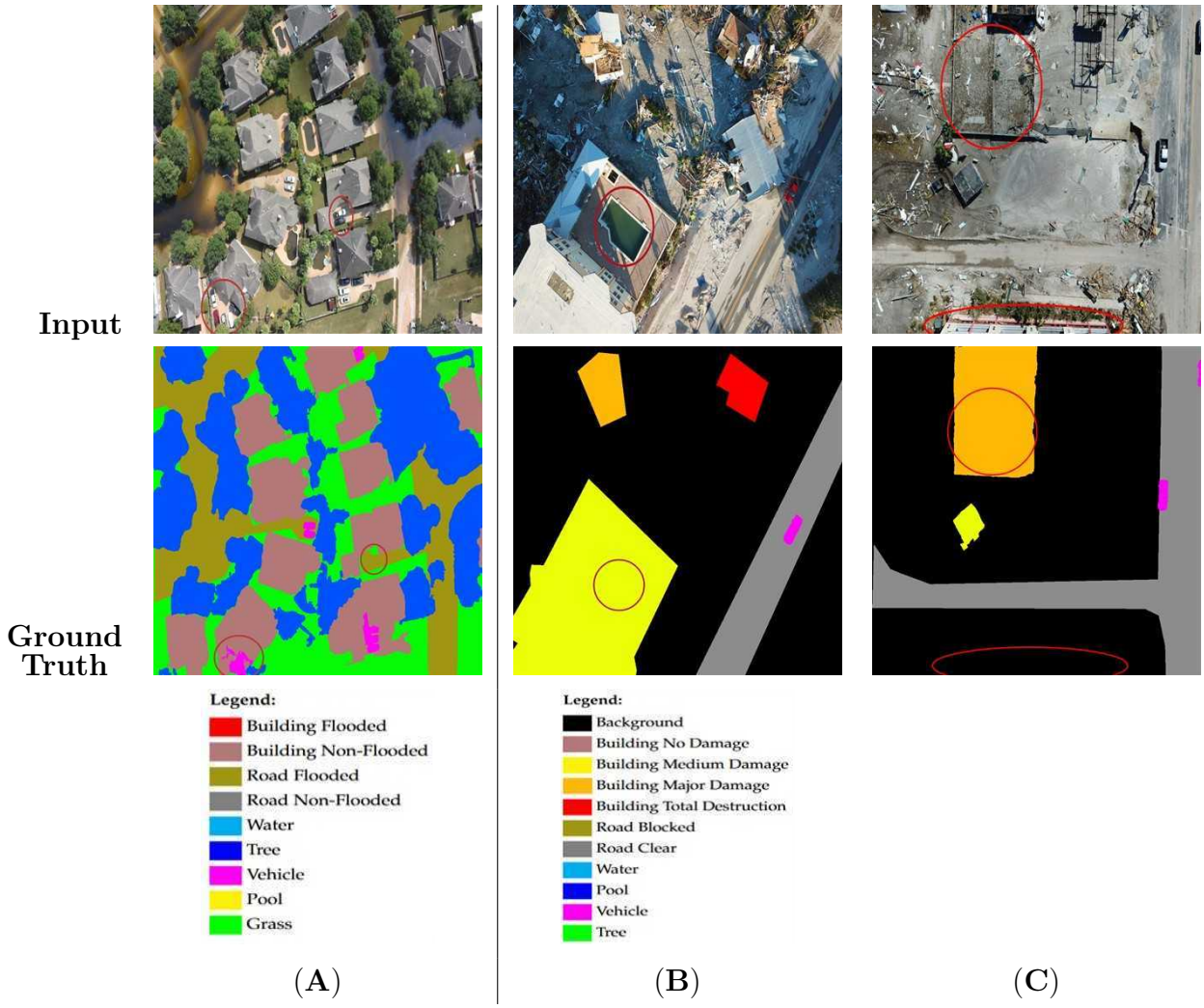


Figure 4.10: Défis liés aux annotations. **(A)** correspond à un échantillon du jeu de données FloodNet ; **(B)** et **(C)** proviennent du jeu de données RescueNet. Les cercles rouges indiquent des erreurs d'annotation telles que des objets manquants ou des régions mal classifiées.

4.6 Discussion, limites et perspectives de la première contribution

Les résultats obtenus montrent que CSDNet permet d'améliorer significativement la segmentation des images aériennes UAV dans des scénarios post-catastrophe complexes. Plus particulièrement, l'intégration d'indices guidés par la détection contribue à renforcer la représentation des classes rares et des objets de petite taille, qui demeurent souvent difficiles à segmenter pour les architectures classiques de segmentation dense [3, 8].

Forces de l'approche proposée. L'une des principales forces de CSDNet réside dans la complémentarité entre segmentation et détection. Alors que les approches classiques de segmentation apprennent principalement à partir d'une supervision dense pixel par pixel, notre méthode exploite également des informations de localisation issues d'un modèle de détection afin de guider plus efficacement les représentations associées aux catégories difficiles.

Cette interaction permet d'améliorer la localisation des petits objets ainsi que la cohérence des prédictions dans les régions complexes. Les résultats qualitatifs montrent notamment une meilleure délimitation des véhicules, piscines et structures fines, tandis que les résultats quantitatifs confirment une amélioration des performances sur plusieurs classes sous-représentées. Ces observations rejoignent plusieurs travaux récents soulignant l'importance des représentations multi-tâches et du contexte spatial dans les scènes aériennes complexes [26, 27].

Une autre force importante de cette contribution concerne l'équilibre entre contexte global et détails locaux. Grâce à la combinaison de fusion multi-échelle, de mécanismes attentionnels légers et de caractéristiques guidées par la détection, l'architecture proposée parvient à préserver des informations spatiales fines tout en maintenant une bonne compréhension contextuelle de la scène. Cette combinaison rejoint la tendance actuelle des architectures hybrides CNN–Transformer visant à exploiter simultanément les dépendances locales et globales dans les images [8].

Par ailleurs, contrairement à certaines architectures Transformer de grande taille nécessitant des ressources computationnelles importantes, CSDNet conserve une complexité relativement modérée tout en intégrant des mécanismes avancés de modélisation spatiale. Cette propriété constitue un avantage intéressant dans les applications UAV, où les contraintes mémoire et le traitement d'images haute résolution peuvent limiter l'utilisation de modèles très lourds [7].

Enfin, cette contribution montre plus largement l'intérêt d'intégrer des informations complémentaires issues d'autres tâches de vision pour améliorer la segmentation sémantique. Les résultats obtenus suggèrent que la détection peut servir non seulement à localiser des objets, mais également à enrichir les représentations sémantiques utilisées par les modèles de segmentation.

Limites théoriques. Malgré les améliorations obtenues, cette première contribution reste principalement formulée dans un cadre supervisé intra-domaine. Le modèle suppose que les données d'entraînement et de test proviennent de distributions relativement proches, ce qui limite naturellement sa capacité de généralisation lorsque les conditions d'acquisition changent fortement. Par ailleurs, bien que l'intégration des caractéristiques de détection améliore les performances sur certaines classes rares, cette stratégie dépend indirectement de la qualité des représentations apprises par le module de détection lui-même. Les erreurs de localisation ou les détections imprécises peuvent ainsi se propager vers les représentations de segmentation. Une autre limite théorique concerne la difficulté persistante de certaines catégories visuellement ambiguës. Dans les scènes post-catastrophe, plusieurs classes peuvent partager des textures ou des apparences proches, ce qui rend leur séparation difficile même avec des mécanismes de contexte avancés. Cette difficulté reste largement observée dans la littérature de segmentation aérienne [27, 4].

Limites pratiques. Sur le plan pratique, les performances du modèle restent fortement dépendantes de la qualité des annotations disponibles. Comme illustré dans la Figure 4.10, certains jeux de données présentent des annotations incomplètes, imprécises ou partiellement décalées par rapport aux images originales. Ces imperfections introduisent du bruit dans le processus d'apprentissage et affectent particulièrement les classes minoritaires et les petits objets. La collecte d'annotations denses de haute qualité représente également une contrainte importante dans les scénarios UAV post-catastrophe. La génération manuelle de masques précis est coûteuse, longue et nécessite souvent une expertise spécialisée, ce qui limite la disponibilité de grands jeux de données annotés [4]. De plus, les images UAV haute résolution impliquent des coûts computationnels élevés en mémoire et en temps d'entraînement. Bien que l'architecture proposée reste plus légère que certaines approches hybrides complexes, le traitement de scènes aériennes détaillées demeure exigeant, notamment lorsque plusieurs mécanismes multi-échelles et modules complémentaires sont combinés.

Perspectives. Ces limitations ouvrent plusieurs perspectives de recherche. Une première direction consiste à réduire la dépendance aux annotations denses par des approches semi-supervisées ou faiblement supervisées, exploitant le pseudo-étiquetage, l'apprentissage enseignant–élève et les méthodes contrastives. Une seconde perspective concerne le renforcement de la robustesse face aux variations de distribution, notamment par des architectures à mélange d'experts (Mixture of Experts) permettant d'adapter dynamiquement les représentations aux caractéristiques des scènes [138]. L'intégration de modèles statistiques de texture, notamment les mélanges de lois gaussiennes généralisées multivariées, constitue également une piste prometteuse pour mieux caractériser la diversité des surfaces et des structures visuelles [139]. Par ailleurs, le développement de modèles robustes aux conditions atmosphériques défavorables, telles que le brouillard et la faible visibilité, représente un enjeu important pour les applications UAV, comme l'illustrent les travaux sur la détection humaine en environnement brumeux [140]. Enfin, les limites de généralisation de CSDNet motivent directement la seconde contribution de cette thèse, consacrée à la généralisation inter-domaines et à l'apprentissage à partir de taxonomies sémantiques hétérogènes dans les scénarios UAV réels.

Chapitre 5

Deuxième Contribution

Malgré les avancées présentées dans le chapitre précédent, la segmentation sémantique d'images aériennes en conditions réelles demeure un problème complexe, notamment lorsque les modèles sont transférés entre des jeux de données acquis dans des conditions différentes [15, 141]. En pratique, un modèle entraîné sur un domaine source peut subir une dégradation significative lorsqu'il est appliqué à un domaine cible présentant des variations de capteurs, d'illumination, d'angle de vue, de résolution ou de structure des scènes.

L'adaptation de domaine vise à réduire cet écart en transférant les connaissances d'un domaine source annoté vers un domaine cible non annoté ou faiblement annoté [16, 17]. Ce cadre est particulièrement pertinent pour l'analyse UAV post-catastrophe, où les données disponibles proviennent souvent de catastrophes différentes, de plateformes d'acquisition différentes et de protocoles d'annotation distincts [4, 5]. Cependant, la difficulté ne se limite pas au décalage visuel entre domaines. Dans de nombreux scénarios réels, les jeux de données présentent également des taxonomies d'annotation hétérogènes [31, 23, 32]. Par exemple, FloodNet distingue les bâtiments et les routes selon leur état inondé ou non inondé, tandis que RescueNet décrit les bâtiments selon différents niveaux de dommages et les routes selon leur état d'accessibilité. Ainsi, deux jeux de données peuvent partager les mêmes concepts généraux, tels que *building* ou *road*, tout en fournissant des attributs fins différents.

Pour s'attaquer au problème, une stratégie simple consiste à restreindre l'apprentissage aux seules classes partagées entre les jeux de données [113, 22]. Bien que cette solution facilite l'entraînement inter-domaine, elle ignore une partie importante des annotations spécifiques à chaque domaine. À l'inverse, fusionner toutes les classes dans une taxonomie plate unique peut introduire des ambiguïtés sémantiques, car certaines sous-classes ne sont pas strictement équivalentes entre les jeux de données [142, 116].

Par ailleurs, la segmentation sémantique est généralement formulée comme un problème de classification plate, où chaque pixel est directement associé à une classe parmi un ensemble prédéfini. Pourtant, de nombreuses catégories réelles présentent une structure hiérarchique naturelle [122, 121, 123]. Par exemple, les classes *road flooded* et *road non-flooded* partagent la même super-classe *road*, tandis que différents niveaux de dommages aux bâtiments peuvent être considérés comme des sous-classes de la super-classe *building*. L'exploitation explicite de cette structure hiérarchique permet de simplifier le processus de prédiction et de mieux organiser les relations sémantiques entre catégories générales et attributs fins.

Dans cette contribution, nous proposons un cadre d'adaptation inter-domaine capable de traiter simultanément le décalage visuel entre domaines et le décalage de taxonomie sémantique. L'idée centrale consiste à organiser les annotations en deux niveaux : des super-classes partagées, communes aux différents jeux de données, et des sous-classes spécifiques à chaque domaine. Cette structuration permet d'apprendre des représentations transférables au niveau des concepts généraux tout en préservant les informations fines propres à chaque jeu de données. En complément, nous introduisons un mécanisme de transfert sémantique guidé par prototypes. Les prototypes de super-classes servent d'ancres communes entre domaines, tandis que les prototypes de sous-classes sont utilisés uniquement pour des paires de sous-classes sémantiquement reliées. Cette stratégie permet de transférer de l'information fine sans imposer une correspondance artificielle entre toutes les catégories spécifiques aux domaines. Ainsi, cette contribution se positionne comme un cadre hiérarchique d'adaptation inter-domaine sous supervision hétérogène. Elle vise à exploiter les annotations disponibles dans chaque domaine, à préserver les attributs spécifiques aux jeux de données, et à améliorer la robustesse du modèle face aux décalages visuels et sémantiques rencontrés dans l'imagerie UAV de catastrophes.

5.1 Méthodologie proposée

5.1.1 Vue d'ensemble de l'approche

Afin de répondre aux défis liés aux décalages visuels et aux taxonomies hétérogènes, nous proposons un cadre hiérarchique d'adaptation inter-domaine guidé par prototypes. L'idée principale consiste à organiser les connaissances sémantiques en plusieurs niveaux d'abstraction, afin de distinguer les concepts généraux transférables des attributs spécifiques à chaque domaine.

Le cadre proposé repose sur trois composantes complémentaires. Premièrement, les taxonomies propres à chaque jeu de données sont projetées vers un espace commun composé de super-classes partagées, telles que *building*, *road*, *water*, *tree*, *vehicle* et *pool*. Cette étape permet de définir une base sémantique commune entre les domaines source et cible, malgré leurs différences d'annotation. Deuxièmement, un mécanisme de raffinement conditionnel est introduit afin de prédire les sous-classes spécifiques aux domaines uniquement dans les régions appartenant à leur super-classe parente. Par exemple, les attributs liés à l'état d'inondation ou au niveau de dommage sont prédits seulement à l'intérieur des régions identifiées comme *building* ou *road*. Cette stratégie réduit les ambiguïtés entre classes fines et évite de prédire des sous-classes dans des régions sémantiquement non pertinentes. Troisièmement, nous intégrons un mécanisme de transfert sémantique guidé par prototypes afin d'exploiter les informations du domaine cible en l'absence d'annotations fines. Les prototypes source servent d'ancres sémantiques fiables pour guider la génération de pseudo-labels dans le domaine cible et favoriser l'alignement inter-domaine. Au niveau des super-classes, toutes les catégories partagées participent au transfert par prototypes. Au niveau des sous-classes, l'alignement est appliqué uniquement aux paires de sous-classes jugées sémantiquement reliées, afin d'éviter d'imposer des correspondances artificielles entre des attributs incompatibles.

Cette formulation permet d'exploiter les annotations spécifiques disponibles dans chaque domaine sans les réduire à un espace de labels plat et appauvri. Elle permet également de contrôler l'influence du domaine cible à travers des pseudo-labels de haute confiance et des interactions détachées du flux principal de gradient. Ainsi, le modèle bénéficie des informations du domaine cible tout en préservant la stabilité de l'apprentissage supervisé à partir du domaine source. La combinaison de la modélisation hiérarchique, du raffinement conditionnel et du transfert guidé par prototypes constitue le fondement de cette seconde contribution. La Figure 5.1 présente une vue d'ensemble du cadre proposé et illustre l'interaction entre la prédiction des super-classes, le raffinement des sous-classes et l'alignement sémantique inter-domaine.

5.1.2 Formulation hiérarchique de la segmentation

Nous reformulons la segmentation sémantique comme un problème de prédiction hiérarchique adapté à l'adaptation inter-domaine sous supervision hétérogène. Au lieu de prédire directement toutes les classes fines dans un espace de labels plat, le cadre proposé décompose

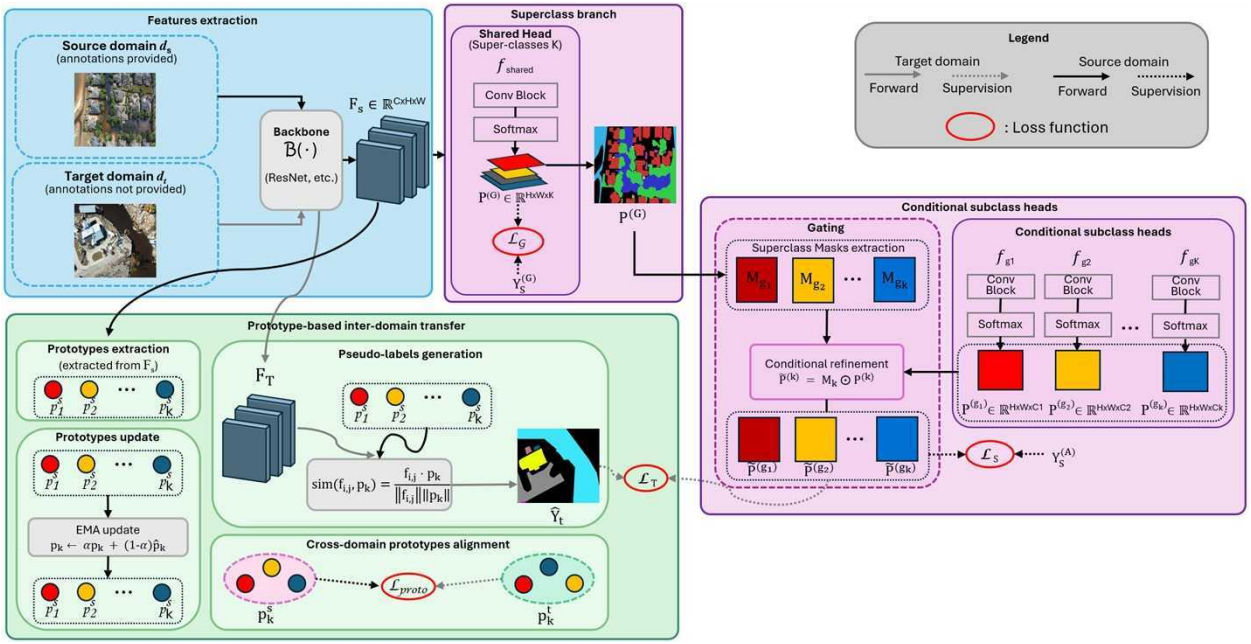


Figure 5.1: Vue d'ensemble du cadre proposé combinant une segmentation hiérarchique et un transfert inter-domaines guidé sur des prototypes. Le modèle prédit d'abord les super-classes partagées, puis affine conditionnellement les prédictions des sous-classes, tout en exploitant des prototypes guidés par le domaine source pour générer des pseudo-labels et assurer un alignement inter-domaines.

la tâche en deux niveaux : des super-classes partagées entre domaines et des sous-classes spécifiques à chaque domaine.

Soit $\mathcal{C}_d$ l'espace d'étiquettes original du domaine d, et soit $\mathcal{G} = \{g_1, \ldots, g_K\}$ l'ensemble des super-classes sémantiques partagées. Dans les scènes UAV de catastrophes, ces super-classes peuvent inclure des catégories génériques telles que *building*, *road*, *water*, *vegetation*, *vehicle* et *pool*. Afin d'établir une représentation sémantique commune entre des jeux de données hétérogènes, chaque taxonomie spécifique à un domaine est projetée vers l'espace partagé à l'aide d'une fonction de mapping :

$$\pi_d : \mathcal{C}_d \to \mathcal{G}. \tag{5.1}$$

Soit $I \in R^{H\times W\times 3}$ une image d'entrée et Y sa carte d'annotation pixel-wise dans le domaine d. Pour une étiquette de pixel $y \in \mathcal{C}_d$, l'étiquette correspondante dans l'espace des super-classes est donnée par :

$$y^{(G)} = \pi_d(y). \tag{5.2}$$

La carte d'annotation partagée est donc définie comme :

$$Y^{(G)} = \pi_d(Y), \tag{5.3}$$

où $Y^{(G)} \in \mathcal{G}^{H \times W}$ contient les annotations de super-classes utilisées pour apprendre des représentations sémantiques communes entre domaines.

Cette formulation permet de préserver les annotations disponibles tout en distinguant explicitement les concepts transférables des catégories spécifiques à chaque jeu de données. Par exemple, FloodNet distingue les bâtiments et les routes selon leur état inondé ou non inondé, tandis que RescueNet décrit les bâtiments selon plusieurs niveaux de dommages structurels et les routes selon leur état d'accessibilité. Ces annotations fines sont organisées comme des sous-classes conditionnées par leur super-classe parente.

Pour une super-classe donnée $g_k \in \mathcal{G}$, l'ensemble des sous-classes spécifiques au domaine d est défini par :

$$\mathcal{A}_d(g_k) = \{c \in \mathcal{C}_d \mid \pi_d(c) = g_k\}. \tag{5.4}$$

Les sous-classes provenant de différents domaines ne sont pas supposées être identiques. Certaines peuvent toutefois présenter une relation sémantique pertinente. Pour deux domaines d_1 et d_2, nous définissons alors un ensemble de paires de sous-classes associées à la même super-classe g_k :

$$\mathcal{P}(g_k) \subseteq \mathcal{A}_{d_1}(g_k) \times \mathcal{A}_{d_2}(g_k). \tag{5.5}$$

Chaque paire $(a_1, a_2) \in \mathcal{P}(g_k)$, avec $a_1 \in \mathcal{A}_{d_1}(g_k)$ et $a_2 \in \mathcal{A}_{d_2}(g_k)$, représente une relation sémantique entre deux attributs fins appartenant à la même super-classe parente. Ces paires ne constituent pas nécessairement des équivalences strictes entre labels ; elles expriment plutôt des similarités sémantiques utiles pour le transfert guidé par prototypes. Par exemple, un bâtiment non inondé dans FloodNet peut être rapproché d'un bâtiment sans dommage dans RescueNet.

Il est important de distinguer le transfert au niveau des super-classes et le transfert au niveau des sous-classes. Au niveau des super-classes, toutes les classes partagées $g_k \in \mathcal{G}$ sont prises en compte, car elles définissent l'espace sémantique commun entre domaines. Au niveau des sous-classes, en revanche, le transfert guidé par prototypes est appliqué uniquement aux paires sémantiquement reliées définies dans $\mathcal{P}(g_k)$. Cette stratégie évite d'imposer des correspondances artificielles entre toutes les sous-classes source et cible, tout en permettant un transfert fin lorsque des relations sémantiques pertinentes existent.

Le processus de segmentation est ainsi décomposé en deux étapes. Le modèle prédit d'abord les super-classes partagées afin d'apprendre des concepts généraux transférables entre domaines. Il raffine ensuite les prédictions vers des sous-classes spécifiques aux domaines uniquement dans les régions correspondant à leur super-classe parente. Cette formulation réduit l'ambiguïté sémantique, exploite les annotations hétérogènes disponibles et préserve les informations fines au-delà du simple espace des classes partagées.

5.1.3 Têtes hiérarchiques et raffinement conditionnel

À partir de la formulation hiérarchique introduite précédemment, nous concevons une architecture de prédiction composée de plusieurs têtes spécialisées afin de découpler la prédiction des super-classes de celle des sous-classes. Soit f_θ le backbone de segmentation paramétré par θ. Pour une image d'entrée I, le backbone extrait une représentation dense :

$$F = f_\theta(I), \tag{5.6}$$

avec $F \in R^{C\times H\times W}$, où C désigne la dimension des caractéristiques. Cette carte de caractéristiques sert de représentation sémantique partagée à partir de laquelle toutes les prédictions sont dérivées.

Prédiction des super-classes. Une tête de prédiction partagée f_G est d'abord utilisée pour estimer les super-classes sémantiques communes à tous les domaines :

$$P^{(G)} = \sigma\big(f_G(F)\big) \in R^{K\times H\times W}, \tag{5.7}$$

où $K = |\mathcal{G}|$ désigne le nombre de super-classes partagées, et $\sigma(\cdot)$ représente l'opérateur softmax appliqué sur les canaux. Cette tête produit donc, pour chaque pixel, une distribution de probabilités sur l'ensemble des super-classes. Elle est supervisée à l'aide des annotations projetées $Y^{(G)}$, ce qui permet à tous les domaines de contribuer à l'apprentissage d'une représentation sémantique commune.

Raffinement conditionnel des sous-classes. Pour chaque super-classe $g_k \in \mathcal{G}$, nous introduisons une tête de raffinement dédiée f_k, chargée de prédire les sous-classes spécifiques associées à cette super-classe. Au lieu de prédire toutes les classes fines globalement sur

l'image, le raffinement est appliqué de manière conditionnelle uniquement dans les régions sémantiquement pertinentes.

Un masque binaire $M_k \in \{0,1\}^{H\times W}$ est d'abord généré à partir de la prédiction des super-classes :

$$M_k = I\big(\arg\max_c P_c^{(G)} = k\big), \tag{5.8}$$

où $I(\cdot)$ désigne la fonction indicatrice. La tête de raffinement associée produit ensuite des prédictions locales de sous-classes :

$$P^{(k)} = \sigma\big(f_k(F)\big). \tag{5.9}$$

Ces prédictions sont restreintes aux régions correspondant à la super-classe concernée :

$$\tilde{P}^{(k)} = M_k \odot P^{(k)}, \tag{5.10}$$

où $\odot$ désigne la multiplication élément par élément.

Cette formulation conditionnelle contraint les prédictions fines à rester cohérentes avec les régions de super-classes correspondantes. Elle réduit ainsi les ambiguïtés entre catégories proches et évite que des sous-classes soient prédites dans des régions sémantiquement invalides.

Supervision asymétrique entre domaines. Dans le cadre d'adaptation inter-domaine considéré, la supervision est dite asymétrique car les deux domaines ne disposent pas du même niveau d'information pendant l'apprentissage. Le domaine source d_s est annoté et fournit des étiquettes au niveau des super-classes et des sous-classes, tandis que le domaine cible d_t est non annoté durant l'entraînement. Les informations relatives au domaine cible sont donc obtenues indirectement à partir des prédictions du modèle, des pseudo-labels de haute confiance et du transfert guidé par prototypes. Les annotations du domaine cible sont uniquement utilisées pour l'évaluation des performances.

Une supervision directe des sous-classes dans le domaine cible peut introduire du bruit en raison des décalages de domaine, des taxonomies hétérogènes et de l'incertitude des pseudo-labels. Pour cette raison, les têtes de sous-classes sont principalement supervisées à partir des annotations fiables du domaine source, tandis que les informations du domaine cible sont exploitées à travers le transfert guidé par prototypes, les pseudo-labels de haute confiance et la régularisation sémantique.

Plus précisément, soient F_s et F_t les représentations extraites respectivement des images source et cible. Afin d'éviter que des pseudo-labels incertains du domaine cible ne perturbent l'apprentissage du backbone, les interactions impliquant les caractéristiques du domaine cible sont réalisées avec un blocage du gradient :

$$\bar{F}_t = \mathrm{sg}(F_t), \tag{5.11}$$

où $\mathrm{sg}(\cdot)$ désigne un opérateur de blocage du gradient (*stop-gradient*), empêchant la rétro-propagation à travers son argument, indépendamment de toute implémentation logicielle particulière.

Ainsi, les caractéristiques du domaine cible peuvent contribuer à la génération de pseudo-labels, à l'estimation des prototypes et à l'alignement sémantique, sans corrompre directement l'apprentissage supervisé fondé sur les annotations source. Cette stratégie préserve la fiabilité de l'apprentissage source tout en exploitant de manière contrôlée les informations sémantiques du domaine cible.

L'architecture hiérarchique proposée décompose donc la segmentation en décisions sémantiques structurées. La prédiction des super-classes apprend des concepts transférables entre domaines, le raffinement conditionnel capture les attributs spécifiques aux domaines, et les interactions détachées avec le domaine cible améliorent la stabilité de l'optimisation sous supervision hétérogène. Ensemble, ces composantes facilitent le transfert inter-domaine tout en limitant les effets des divergences taxonomiques et de la propagation de pseudo-labels bruités.

5.1.4 Transfert inter-domaine guidé par prototypes

Afin de compléter le cadre hiérarchique et de compenser l'absence d'annotations fines dans le domaine cible, nous introduisons un mécanisme de transfert inter-domaine guidé par prototypes. L'objectif est d'apprendre des représentations sémantiques stables, capables de guider la prédiction des sous-classes dans un contexte de supervision asymétrique.

Définition des prototypes. Nous définissons des prototypes à deux niveaux sémantiques : les super-classes partagées et les sous-classes spécifiques aux domaines. Les prototypes agissent comme des ancres sémantiques dans l'espace des caractéristiques, en capturant la représentation centrale d'une catégorie indépendamment des variations d'apparence propres aux domaines.

Soit $F^d \in R^{C\times H\times W}$ la carte de caractéristiques extraite pour le domaine $d \in \{s, t\}$. Au niveau des super-classes, les prototypes sont calculés pour toutes les classes sémantiques partagées $g_k \in \mathcal{G}$. Le prototype de la super-classe g_k dans le domaine d est défini par :

$$\mathbf{p}^d_{g_k} = \frac{1}{|\Omega^d_{g_k}|} \sum_{(i,j)\in\Omega^d_{g_k}} F^d_{:,i,j}, \tag{5.12}$$

où $\Omega^d_{g_k}$ désigne l'ensemble des pixels assignés à la super-classe g_k dans le domaine d. Comme les super-classes sont partagées entre les domaines, toutes les classes $g_k \in \mathcal{G}$ sont utilisées pour l'alignement au niveau des super-classes.

Au niveau des sous-classes, le transfert par prototypes est restreint aux paires de sous-classes sémantiquement reliées $(a_s, a_t) \in \mathcal{P}(g_k)$. Le prototype de sous-classe source $\mathbf{p}^s_{a_s}$ est estimé à partir des annotations fiables du domaine source :

$$\mathbf{p}^s_{a_s} = \frac{1}{|\Omega^s_{a_s}|} \sum_{(i,j)\in\Omega^s_{a_s}} F^s_{:,i,j}. \tag{5.13}$$

En revanche, le prototype de sous-classe cible $\mathbf{p}^t_{a_t}$ est estimé uniquement à partir des pseudo-labels cible de haute confiance :

$$\mathbf{p}^t_{a_t} = \frac{1}{|\Omega^{t,\delta}_{a_t}|} \sum_{(i,j)\in\Omega^{t,\delta}_{a_t}} F^t_{:,i,j}, \tag{5.14}$$

où $\Omega^{t,\delta}_{a_t}$ désigne l'ensemble des pixels du domaine cible assignés à la sous-classe a_t avec une confiance supérieure au seuil δ. Cette stratégie sélective évite que des prédictions incertaines du domaine cible ou des sous-classes non reliées ne perturbent la mémoire des prototypes.

Comme les estimations obtenues sur les mini-batchs peuvent être bruitées et dépendantes du domaine, les prototypes sont maintenus dans une mémoire et mis à jour par moyenne exponentielle mobile (*Exponential Moving Average*, EMA). Pour les prototypes source de super-classes, la mise à jour est définie par :

$$\mathbf{p}^s_{g_k} \leftarrow \alpha\mathbf{p}^s_{g_k} + (1-\alpha)\hat{\mathbf{p}}^s_{g_k}, \tag{5.15}$$

où $\hat{\mathbf{p}}^s_{g_k}$ désigne le prototype estimé à partir du mini-batch courant. De manière similaire, les prototypes source de sous-classes sont mis à jour selon :

$$\mathbf{p}^s_{a_s} \leftarrow \alpha \mathbf{p}^s_{a_s} + (1-\alpha)\hat{\mathbf{p}}^s_{a_s}. \tag{5.16}$$

Les prototypes cible peuvent également être mis à jour par EMA, mais uniquement lorsqu'ils sont estimés à partir de pseudo-labels de haute confiance. Le paramètre $\alpha \in [0,1]$ contrôle le momentum de mise à jour.

Génération de pseudo-labels guidée par prototypes. Dans le domaine cible, les annotations fines de sous-classes ne sont pas disponibles. Les prototypes sont donc utilisés comme références sémantiques transférables afin de générer des pseudo-labels fiables. L'intuition est que les prototypes source encodent des concepts sémantiques stables appris à partir d'annotations fiables, et peuvent ainsi guider la prédiction dans le domaine cible au-delà des seules similarités d'apparence.

Soit $\mathbf{f}^t_{i,j} \in R^C$ la représentation du pixel (i,j) dans une image cible. Pour la génération de pseudo-labels au niveau des super-classes, chaque caractéristique cible $\mathbf{f}^t_{i,j}$ est comparée aux prototypes source de super-classes $\mathbf{p}^s_{g_k}$, qui servent d'ancres sémantiques fiables :

$$s^{(k)}_{i,j} = \cos(\mathbf{f}^t_{i,j}, \mathbf{p}^s_{g_k}) = \frac{(\mathbf{f}^t_{i,j})^\top \mathbf{p}^s_{g_k}}{\|\mathbf{f}^t_{i,j}\|_2 \, \|\mathbf{p}^s_{g_k}\|_2}. \tag{5.17}$$

Les scores de similarité sont ensuite convertis en distribution de pseudo-labels :

$$q^{(k)}_{i,j} = \frac{\exp(s^{(k)}_{i,j}/\tau)}{\sum_l \exp(s^{(l)}_{i,j}/\tau)}, \tag{5.18}$$

où τ est un paramètre de température contrôlant la netteté des prédictions. Le pseudo-label assigné au pixel (i,j) est donné par :

$$\hat{y}_{i,j} = \arg\max_k q^{(k)}_{i,j}. \tag{5.19}$$

Afin de réduire la propagation d'erreurs, seuls les pseudo-labels de haute confiance sont conservés :

$$\max_k q^{(k)}_{i,j} > \delta, \tag{5.20}$$

où δ est un seuil de confiance. Ces pseudo-labels guidés par prototypes fournissent une supervision faible pour les modules de raffinement conditionnel dans le domaine cible, sans nécessiter d'annotations fines explicites.

Rôle du transfert sémantique par prototypes. Au-delà de la génération de pseudo-labels, les prototypes constituent le principal mécanisme de transfert sémantique entre domaines sous supervision asymétrique. Les prototypes de super-classes permettent d'aligner les concepts partagés, tandis que les prototypes de sous-classes permettent de transférer des informations fines uniquement lorsque des correspondances sémantiques partielles sont définies.

Ce mécanisme évite de forcer une correspondance complète entre toutes les taxonomies source et cible. Il permet au modèle de bénéficier des relations sémantiques existantes, telles que le lien entre un bâtiment non inondé et un bâtiment sans dommage, tout en préservant les spécificités propres à chaque dataset. Ainsi, même en l'absence d'annotations fines dans le domaine cible, le modèle peut améliorer la cohérence des prédictions et renforcer la discrimination des sous-classes grâce à la structure induite par les prototypes.

5.1.5 Interaction des modules

La méthode proposée repose sur l'interaction de trois composantes complémentaires : la formulation hiérarchique, le raffinement conditionnel et le transfert inter-domaine guidé par prototypes. Ces composantes ne sont pas indépendantes ; elles forment un cadre cohérent dans lequel chaque module contribue à réduire une difficulté spécifique liée à l'adaptation inter-domaine sous taxonomies hétérogènes.

La formulation hiérarchique constitue la base du cadre proposé. Elle sépare les concepts généraux transférables, représentés par les super-classes partagées, des attributs fins propres à chaque domaine, représentés par les sous-classes spécifiques. Cette décomposition permet d'éviter une fusion naïve de toutes les catégories dans un espace de labels plat, ce qui pourrait introduire des conflits sémantiques entre des annotations non équivalentes.

Les têtes de raffinement conditionnel exploitent ensuite cette structure hiérarchique pour effectuer une prédiction plus ciblée. Au lieu de prédire toutes les sous-classes sur l'ensemble de l'image, chaque tête spécialisée est activée uniquement dans les régions correspondant à sa super-classe parente. Par exemple, les sous-classes associées aux bâtiments ne sont raffinées que dans les régions prédites comme *building*. Ce mécanisme réduit la confusion entre catégories fines et améliore la cohérence sémantique des prédictions.

Le transfert guidé par prototypes complète cette architecture en introduisant des ancres sémantiques entre les domaines source et cible. Les prototypes de super-classes sont utilisés pour toutes les catégories partagées, car elles définissent l'espace commun entre les domaines. En revanche, les prototypes de sous-classes ne sont utilisés que pour les paires sémantiquement reliées définies dans $\mathcal{P}(g_k)$. Cette distinction permet d'exploiter les similarités fines entre certains attributs sans imposer une correspondance artificielle entre toutes les sous-classes source et cible.

Dans le domaine cible, où les annotations fines sont absentes ou incomplètes, les prototypes source servent d'ancres fiables pour générer des pseudo-labels de haute confiance. Ces pseudo-labels guident les têtes de raffinement conditionnel, tandis que les interactions avec les caractéristiques cible sont contrôlées par un blocage du gradient afin de limiter la propagation d'erreurs vers le backbone. Cette stratégie permet d'exploiter l'information cible sans compromettre la stabilité de l'apprentissage supervisé à partir du domaine source.

Ainsi, la hiérarchie structure l'espace sémantique, les têtes conditionnelles spécialisent la prédiction fine, et les prototypes assurent le transfert sémantique entre domaines. L'interaction de ces modules permet de traiter simultanément le décalage visuel, l'hétérogénéité des taxonomies et l'asymétrie de supervision entre domaines.

5.2 Fonction de perte et stratégie d'entraînement

Dans cette section, nous décrivons l'objectif d'apprentissage global ainsi que la stratégie d'optimisation progressive adoptée dans le cadre d'une adaptation inter-domaine sous supervision asymétrique.

Perte sur les super-classes partagées. Les super-classes partagées fournissent un signal de supervision commun entre les domaines. Soit $P^{(G)} \in [0,1]^{K\times H\times W}$ la carte de probabilités produite par la tête de prédiction des super-classes, où $K = |\mathcal{G}|$ désigne le nombre de super-classes partagées. Soit $Y^{(G)} \in \{1,\dots,K\}^{H\times W}$ la carte d'annotation correspondante dans l'espace partagé. La perte de segmentation au niveau des super-classes est définie par :

$$\mathcal{L}_G = -\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\sum_{k=1}^{K} I\Big[Y_{i,j}^{(G)} = k\Big]\log P_{k,i,j}^{(G)}, \tag{5.21}$$

où Ω désigne l'ensemble des pixels annotés valides, et $I(\cdot)$ est la fonction indicatrice. Cette perte permet d'apprendre des représentations transférables au niveau des concepts généraux partagés entre domaines.

Perte sur les sous-classes du domaine source. Pour le domaine source d_s, les sous-classes fines sont disponibles et peuvent être utilisées pour superviser les têtes de raffinement conditionnel. Comme chaque tête de sous-classe est associée à une super-classe particulière, la supervision est appliquée uniquement aux pixels appartenant à la région de cette super-classe.

Soit $P^{(S,k)} \in [0,1]^{K_k \times H \times W}$ la carte de probabilités produite par la tête de sous-classes associée à la super-classe g_k, où K_k désigne le nombre de sous-classes rattachées à g_k. Soit $Y^{(S,k)}$ la carte d'annotations fines correspondante, et Ω_k l'ensemble des pixels dont l'étiquette de super-classe est g_k. La perte de supervision des sous-classes source est définie par :

$$\mathcal{L}_S = -\sum_{k=1}^{K} \frac{1}{|\Omega_k|} \sum_{(i,j)\in\Omega_k} \ell_{CE}\big(P_{i,j}^{(S,k)}, Y_{i,j}^{(S,k)}\big). \tag{5.22}$$

En restreignant l'optimisation aux régions sémantiquement valides, chaque tête de raffinement apprend à discriminer uniquement les sous-classes pertinentes pour sa super-classe parente. Cette supervision conditionnelle réduit l'ambiguïté entre catégories fines et encourage la spécialisation des têtes de prédiction.

Perte pseudo-supervisée dans le domaine cible. Dans le domaine cible d_t, les annotations ne sont pas disponibles. Nous utilisons donc les pseudo-labels générés par les prototypes pour fournir une supervision faible aux têtes de raffinement conditionnel. Soit $\hat{Y}_{i,j}^{(T,k)}$ le pseudo-label de haute confiance assigné au pixel (i,j) pour la tête associée à la super-classe g_k. La perte de raffinement dans le domaine cible est définie par :

$$\mathcal{L}_T = \sum_{k} \frac{1}{|\Omega_k|} \sum_{(i,j)\in\Omega_k} w_{i,j}\, \ell_{CE}\Big(P_{i,j}^{(T,k)}, \hat{Y}_{i,j}^{(T,k)}\Big), \tag{5.23}$$

où $w_{i,j} = \max_c q_{i,j}^{(c)}$ correspond à la confiance du pseudo-label généré par les prototypes. Ce poids réduit l'influence des prédictions incertaines.

Afin d'éviter que des pseudo-labels bruités ne perturbent l'apprentissage du backbone, cette perte est appliquée à partir des caractéristiques cibles détachées $\bar{F}_t$ définies précédem-

ment. Ainsi, la pseudo-supervision cible influence principalement les modules de raffinement et le processus d'adaptation par prototypes, tandis que le backbone reste principalement guidé par les annotations fiables du domaine source.

Perte d'alignement des prototypes. Afin de promouvoir la cohérence sémantique entre domaines, nous introduisons une perte d'alignement des prototypes composée de deux termes : un alignement au niveau des super-classes et un alignement sélectif au niveau des sous-classes.

Au niveau des super-classes, toutes les classes partagées sont alignées. Pour chaque super-classe $g_k \in \mathcal{G}$, le prototype source $\mathbf{p}^s_{g_k}$ agit comme une ancre sémantique fiable, tandis que le prototype cible $\mathbf{p}^t_{g_k}$ est estimé à partir des prédictions du domaine cible :

$$\mathcal{L}^G_{proto} = \sum_{g_k \in \mathcal{G}} \left\| \mathrm{sg}(\mathbf{p}^s_{g_k}) - \mathbf{p}^t_{g_k} \right\|_2^2 . \tag{5.24}$$

Au niveau des sous-classes, l'alignement est appliqué uniquement aux paires sémantiquement reliées définies dans $\mathcal{P}(g_k)$:

$$\mathcal{L}^A_{proto} = \sum_{g_k \in \mathcal{G}} \sum_{(a_s, a_t) \in \mathcal{P}(g_k)} \left\| \mathrm{sg}(\mathbf{p}^s_{a_s}) - \mathbf{p}^t_{a_t} \right\|_2^2 . \tag{5.25}$$

La perte finale d'alignement des prototypes est définie par :

$$\mathcal{L}_{proto} = \mathcal{L}^G_{proto} + \lambda_A \mathcal{L}^A_{proto}, \tag{5.26}$$

où λ_A contrôle la contribution de l'alignement sélectif au niveau des sous-classes. Cette formulation rapproche les prototypes cibles des ancres source fiables, sans imposer de correspondance artificielle entre toutes les sous-classes source et cible.

Objectif global. L'objectif total combine les différentes composantes de perte :

$$\mathcal{L}_{total} = \mathcal{L}_G + \lambda_S \mathcal{L}_S + \lambda_T \mathcal{L}_T + \lambda_P \mathcal{L}_{proto}, \tag{5.27}$$

où λ_S, λ_T et λ_P sont des coefficients de pondération.

Stratégie d'entraînement progressive. Afin de stabiliser l'optimisation, les pertes associées aux sous-classes, aux pseudo-labels et aux prototypes sont introduites progressivement

au cours de l'entraînement à l'aide d'un facteur de montée :

$$\lambda_t(t) = \min\left(1, \frac{t}{T}\right), \tag{5.28}$$

où t désigne l'itération courante et T contrôle la durée de la phase de montée.

Cette stratégie évite une dépendance précoce aux pseudo-labels, souvent bruités au début de l'apprentissage, et permet au modèle d'apprendre d'abord des représentations robustes au niveau des super-classes avant d'introduire les objectifs plus difficiles de raffinement et d'alignement inter-domaine.

L'objectif proposé permet de combiner la supervision fiable du domaine source, l'exploitation contrôlée du domaine cible et le transfert sémantique guidé par prototypes. La séparation des flux de gradient, associée aux prototypes sémantiques, assure un apprentissage plus stable sous supervision hétérogène et sous décalage de domaine.

5.3 Expérimentation et résultats

Dans cette section, nous évaluons la méthode proposée sur deux jeux de données UAV de scènes de catastrophes, FloodNet et RescueNet, dans un cadre d'adaptation inter-domaine sous supervision hétérogène. Les expériences sont menées dans les deux directions de transfert, FloodNet→RescueNet et RescueNet→FloodNet, afin d'évaluer la robustesse du modèle face à des décalages visuels et sémantiques importants.

L'évaluation repose sur trois aspects complémentaires : (i) la performance de segmentation au niveau des super-classes partagées, (ii) la capacité à exploiter les sous-classes spécifiques aux domaines sous supervision asymétrique, et (iii) la robustesse face à des taxonomies partiellement incompatibles. Les résultats sont rapportés à l'aide du mIoU et de l'IoU par classe, puis comparés à plusieurs méthodes de référence.

5.3.1 Jeux de données

FloodNet Le jeu de données FloodNet [4] contient des images UAV haute résolution capturées après l'ouragan Harvey à Houston. Il fournit des annotations pixel-wise pour les états inondé et non inondé des routes et des bâtiments, ainsi que pour les classes eau, arbres, véhicules, piscines, herbe et arrière-plan. Ce jeu de données présente un fort déséquilibre de classes, des variations d'illumination importantes et une grande variabilité intra-classe

due aux différents niveaux d'inondation. Il contient 2 345 images, réparties en 1 445 images d'entraînement (61,62%), 450 images de validation (19,19%) et 450 images de test (19,19%).

RescueNet

RescueNet Le jeu de données RescueNet [5] contient des images UAV collectées après l'ouragan Michael. Il comprend dix catégories, notamment l'eau, les véhicules, les piscines, les arbres, les routes (dégagées ou bloquées) et quatre niveaux de dommages aux bâtiments (aucun, mineur, majeur, destruction totale). Ce jeu de données se distingue visuellement de FloodNet en raison de différences dans le type de catastrophe, les conditions atmosphériques, les perspectives et la nature des dommages. Il contient 4 494 images, réparties en 3 595 images d'entraînement (80,00 %), 449 images de validation (9,99 %) et 450 images de test (10,01 %).

Alignement des classes partagées et attributs spécifiques. FloodNet et RescueNet adoptent des taxonomies d'annotation différentes. Afin de permettre l'apprentissage inter-domaine, nous construisons un espace sémantique partagé en consolidant et en alignant les catégories compatibles entre les deux jeux de données. Les labels fins sont d'abord regroupés en super-classes communes, telles que *building*, *road*, *tree*, *water*, *vehicle* et *pool*. Ces super-classes constituent la tâche principale de segmentation partagée entre domaines.

Les attributs spécifiques à chaque jeu de données sont cependant conservés sous forme de sous-classes propres au domaine. Dans FloodNet, ces attributs correspondent notamment aux états inondé et non inondé des bâtiments et des routes. Dans RescueNet, ils correspondent aux niveaux de dommages des bâtiments ainsi qu'à l'état des routes, dégagées ou bloquées. Ces sous-classes ne sont pas fusionnées dans une taxonomie plate unique ; elles sont supervisées dans leur domaine d'origine et exploitées à travers le raffinement conditionnel et le transfert guidé par prototypes.

Ce protocole bidirectionnel permet d'évaluer la méthode dans deux directions de transfert, FloodNet→RescueNet et RescueNet→FloodNet. Il permet ainsi d'analyser la capacité du modèle à apprendre des concepts partagés tout en préservant les attributs spécifiques aux domaines sous des taxonomies partiellement incompatibles. Par ailleurs, le déséquilibre des classes demeure un défi important, notamment pour les catégories sous-représentées telles que *vehicle* et *pool*.

5.3.2 Configuration expérimentale

Dans cette section, nous décrivons les paramètres d'entraînement ainsi que les choix d'implémentation adoptés pour évaluer la méthode proposée dans un cadre d'adaptation inter-domaine sous supervision hétérogène.

Prétraitement des données. Les images sont redimensionnées à une résolution fixe de 512×512 pixels afin de garantir une cohérence spatiale et de maîtriser la complexité computationnelle. Les masques de segmentation sont redimensionnés à l'aide d'une interpolation au plus proche voisin afin de préserver l'intégrité des étiquettes discrètes. Une normalisation standard de type ImageNet est appliquée aux images en entrée.

Architecture du modèle. Le backbone utilisé pour l'extraction des caractéristiques est DeepLabV3 avec un encodeur ResNet-50 pré-entraîné sur ImageNet. Le backbone produit une carte de caractéristiques dense $\mathbf{F} \in R^{C \times H \times W}$, qui est ensuite utilisée conjointement par une tête de prédiction des super-classes (*shared head*) et par des têtes auxiliaires dédiées aux sous-classes. Les branches de sous-classes sont alimentées par des caractéristiques détachées afin d'éviter toute interférence avec la branche principale supervisée par le domaine source.

Paramètres d'entraînement. Le modèle est entraîné à l'aide d'un optimiseur SGD avec un momentum de 0.9 et un facteur de décroissance des poids fixé à 10^{-4}. Le taux d'apprentissage initial est fixé à 10^{-2} et le modèle est entraîné pendant 100 époques avec une taille de batch de 8.

Stratégie d'apprentissage inter-domaines. À chaque itération, des mini-batchs sont échantillonnés simultanément à partir des domaines source et cible. La branche principale du modèle est supervisée à partir des annotations fiables du domaine source, tandis que le domaine cible est exploité de manière contrôlée via des mécanismes de pseudo-labeling et d'alignement guidés par prototypes. Les caractéristiques issues du domaine cible sont détachées du graphe de calcul lors des interactions sensibles avec les prototypes afin de limiter la propagation d'erreurs dues aux pseudo-labels bruités.

Mise à jour des prototypes. Les prototypes source sont estimés à partir des annotations fiables du domaine source et mis à jour dynamiquement au cours de l'entraînement à l'aide d'une moyenne exponentielle (EMA). Les prototypes cible, lorsqu'ils sont utilisés, sont

estimés uniquement à partir de pseudo-labels de haute confiance. Cette stratégie permet de stabiliser les représentations et de limiter l'influence des prédictions incertaines du domaine cible.

Activation progressive des pertes. Afin d'assurer une convergence stable, les différentes composantes de la fonction de perte sont introduites progressivement au cours de l'entraînement via un mécanisme de montée en puissance (*ramp-up*). Les pertes associées aux sous-classes, aux pseudo-labels et à l'alignement par prototypes sont introduites progressivement après les premières phases d'apprentissage, permettant au modèle d'acquérir d'abord des représentations robustes au niveau des super-classes. Cette stratégie réduit l'impact des pseudo-labels bruités en début d'entraînement et améliore la stabilité globale du processus d'optimisation.

Implémentation. Le modèle est implémenté en `PyTorch` et entraîné sur GPU. Les mêmes hyperparamètres sont utilisés pour les deux directions de transfert (FloodNet→RescueNet et RescueNet→FloodNet) afin d'assurer une comparaison équitable.

5.3.3 Métriques d'évaluation

Afin d'évaluer les performances du modèle proposé, nous utilisons des métriques standard en segmentation sémantique, en particulier l'Intersection over Union (IoU) et la moyenne des IoU (mIoU).

Intersection over Union (IoU). Pour chaque classe k, l'IoU est définie comme le rapport entre l'intersection et l'union des prédictions et des annotations de référence :

$$\mathrm{IoU}_k = \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{5.29}$$

où TP_k, FP_k et FN_k représentent respectivement les vrais positifs, faux positifs et faux négatifs pour la classe k.

Mean Intersection over Union (mIoU). La performance globale est mesurée à l'aide de la moyenne des IoU sur l'ensemble des classes considérées :

$$\mathrm{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{IoU}_k, \tag{5.30}$$

Table 5.1: Class-wise IoU (%) for RescueNet→FloodNet transfer.

Method	Eval	mIoU	Bld	Road	Water	Tree	Veh	Pool
Ours	RN (source)	90.03	95.45	94.90	93.60	96.44	76.47	80.34
	FN (target)	55.05	67.92	56.97	65.26	80.99	39.66	19.48
FDA	RN (source)	83.98	93.89	92.44	90.63	94.98	72.45	59.51
	FN (target)	54.25	68.66	58.20	60.33	77.62	39.77	20.93
IBN	RN (source)	87.99	93.84	93.76	90.38	94.99	76.56	78.43
	FN (target)	52.68	65.68	59.02	69.09	80.94	34.48	06.89
Deeplabv3-resnet50	RN (source)	77.12	81.38	79.12	87.73	82.85	63.22	68.45
	FN (target)	34.72	56.99	45.15	31.49	41.48	30.01	3.23

où K correspond au nombre de classes.

Évaluation des super-classes et sous-classes. Nous évaluons séparément les performances au niveau des super-classes partagées et des sous-classes spécifiques à chaque domaine. Pour les super-classes, le mIoU est calculé sur l'ensemble des catégories partagées entre les deux domaines. Pour les sous-classes, l'évaluation est réalisée uniquement sur les classes définies dans le domaine correspondant (par exemple, les états d'inondation pour FloodNet et les niveaux de dommages pour RescueNet).

Gestion des pixels ignorés. Certains pixels correspondant à des classes non pertinentes (telles que l'arrière-plan ou certaines classes non alignées) sont exclus de l'évaluation à l'aide d'un indice d'ignorance (*ignore index*). Ces pixels ne contribuent ni au calcul des IoU ni à celui du mIoU.

Analyse par classe. En complément du mIoU global, nous rapportons également l'IoU par classe afin d'analyser finement le comportement du modèle.

5.3.4 Résultats quantitatifs

Nous comparons la méthode proposée à plusieurs approches de référence dans un cadre de transfert inter-domaine bidirectionnel, en considérant les deux directions RescueNet→FloodNet et FloodNet→RescueNet. Les méthodes évaluées incluent le modèle de base DeepLabV3-ResNet50, ainsi que deux stratégies représentatives de robustesse visuelle et de transfert inter-domaine : FDA [19] et IBN [18].

Table 5.2: Class-wise IoU (%) for FloodNet→RescueNet transfer.

Method	Eval	mIoU	Bld	Road	Water	Tree	Veh	Pool
Ours	FN (source)	84.80	90.58	85.00	86.93	94.14	70.55	81.58
	RN (target)	53.04	60.82	62.36	63.58	82.85	35.23	10.40
FDA	FN (source)	84.00	90.45	80.54	83.16	94.09	73.71	82.07
	RN (target)	54.02	61.20	63.94	67.04	85.84	36.04	10.04
IBN	FN (source)	87.37	92.57	86.88	88.91	95.08	76.43	84.35
	RN (target)	49.15	57.39	58.32	63.85	82.30	24.87	08.16
Deeplabv3-resnet50	FN (source)	72.65	83.96	77.17	76.83	84.61	49.20	64.12
	RN (target)	24.65	38.41	29.77	34.88	24.52	16.28	4.02

Comparaison globale inter-domaines. Le Tableau 5.1 présente les performances pour le transfert RescueNet→FloodNet. Dans ce scénario, la méthode proposée obtient un mIoU de 55.05% sur le domaine cible FloodNet, ce qui la place à un niveau très compétitif face à FDA (54.25%) et IBN (52.68%), tout en dépassant largement le baseline DeepLabV3-ResNet50 (34.72%).

Le Tableau 5.2 présente les résultats du transfert inverse FloodNet→RescueNet. Dans ce cas plus difficile, notre méthode atteint 53.04% de mIoU, soit une performance comparable à FDA (54.02%), supérieure à IBN (49.15%) et nettement au-dessus du baseline DeepLabV3-ResNet50 (24.65%).

Ces résultats montrent que la méthode proposée demeure compétitive face aux approches classiques de robustesse visuelle, tout en traitant un cadre plus général intégrant l'adaptation inter-domaine, la segmentation hiérarchique et des taxonomies partiellement hétérogènes.

Apport par rapport au backbone de base: Nous incluons également DeepLabV3-ResNet50 comme baseline de référence. Bien qu'il ne s'agisse pas d'une méthode explicitement conçue pour l'adaptation inter-domaine, ce choix est particulièrement pertinent puisqu'il correspond au backbone central utilisé dans notre architecture.

Cette comparaison permet d'isoler l'apport réel des composantes proposées : apprentissage hiérarchique, prototypes sémantiques, raffinement conditionnel et stratégie d'entraînement progressive. Les gains observés par rapport à ce baseline montrent que les améliorations ne proviennent pas uniquement du réseau de base, mais surtout de la manière dont celui-ci est structuré et exploité dans le cadre proposé.

Analyse par direction de transfert. Le transfert RescueNet→FloodNet apparaît globalement favorable à notre approche. Cela peut s'expliquer par la richesse sémantique de

RescueNet, qui fournit des annotations détaillées sur les niveaux de dommages des bâtiments et l'état des routes. Ces informations fines permettent d'apprendre des représentations plus structurées au niveau source, qui peuvent ensuite être transférées vers FloodNet à travers les super-classes partagées et les prototypes sémantiques.

Dans cette direction, notre méthode obtient un mIoU cible de 55.05%, légèrement supérieur à IBN et proche de FDA. Les gains sont particulièrement visibles par rapport au baseline DeepLabV3-ResNet50, ce qui montre que la structuration hiérarchique et le transfert guidé par prototypes apportent une amélioration significative par rapport à une segmentation plate classique.

Le transfert inverse FloodNet→RescueNet est plus difficile. RescueNet présente une plus grande diversité visuelle et des annotations fines plus complexes, notamment pour les niveaux de dommages des bâtiments. Dans ce scénario, notre méthode reste compétitive avec FDA et dépasse IBN ainsi que DeepLabV3-ResNet50. Cette observation suggère que le cadre proposé permet de maintenir une robustesse inter-domaine malgré des différences importantes d'apparence et de taxonomie.

Globalement, les deux directions montrent que la méthode proposée ne se limite pas à réduire le décalage visuel entre domaines. Elle permet également d'exploiter une structure sémantique hiérarchique afin de préserver les attributs spécifiques aux datasets tout en apprenant des représentations transférables au niveau des super-classes partagées.

Analyse par classe. L'analyse détaillée des IoU par classe montre que les catégories structurales telles que *building* et *road* sont segmentées avec une grande précision dans les deux directions. Les classes naturelles comme *water* et *tree* présentent également des performances stables.

Les classes plus rares telles que *vehicle* et *pool* demeurent plus difficiles à prédire en raison de leur faible fréquence d'apparition et de leur petite taille spatiale. Néanmoins, la méthode proposée reste compétitive sur ces catégories, ce qui suggère que les prototypes et le raffinement hiérarchique contribuent à préserver l'information liée aux objets peu représentés.

5.3.5 Étude d'ablation

Afin de mieux quantifier l'apport de chaque composante du cadre proposé, nous réalisons une étude d'ablation en retirant successivement certains modules clés de l'architecture. Plus précisément, nous évaluons les variantes suivantes : (i) suppression du mécanisme de proto-

types (*w/o prototypes*), (ii) suppression des têtes hiérarchiques spécialisées (*w/o hierarchical heads*), et (iii) suppression du raffinement conditionnel (*w/o conditional refinement*).

L'objectif de cette analyse est de mesurer l'impact de chaque composant sur la prédiction des sous-classes, qui constitue la partie la plus difficile du problème en raison de l'hétérogénéité sémantique entre FloodNet et RescueNet. Contrairement aux super-classes partagées, les sous-classes diffèrent fortement entre les deux jeux de données : FloodNet distingue principalement les états d'inondation, tandis que RescueNet décrit différents niveaux de dommages structurels. Une comparaison directe entre domaines n'est donc pas pertinente ; nous évaluons plutôt la contribution relative de chaque module au sein de notre méthode.

Résultats globaux sur les sous-classes. Le Tableau 5.3 présente la moyenne des IoU sur les sous-classes pour les deux directions de transfert. On observe que la version complète du modèle obtient systématiquement les meilleures performances, confirmant la complémentarité des différents modules proposés.

Table 5.3: Étude d'ablation sur les sous-classes (mIoU %).

Méthode	RN→FN	FN→RN
Full model	**51.90**	**25.12**
w/o prototypes	24.85	14.83
w/o hierarchical heads	18.48	8.58
w/o conditional refinement	22.73	16.62

Ces résultats montrent une chute importante des performances lorsque l'un des modules est retiré. En particulier, la suppression des têtes hiérarchiques entraîne la plus forte dégradation, ce qui confirme l'importance de structurer explicitement la prédiction en super-classes puis sous-classes.

Impact des prototypes Le retrait des prototypes entraîne une baisse marquée dans les deux directions de transfert, particulièrement sur le domaine cible non annoté. Ce comportement confirme que les prototypes jouent un rôle central dans le transfert d'information entre domaines. Ils permettent de fournir un signal discriminant utile pour les sous-classes, malgré l'absence de supervision directe.

Par exemple, dans le scénario RescueNet→FloodNet, certaines classes telles que *road non-flooded* chutent drastiquement lorsque les prototypes sont retirés, ce qui souligne leur importance dans la génération de pseudo-labels cohérents.

Impact des têtes hiérarchiques La suppression des têtes hiérarchiques conduit à la plus forte baisse globale des performances. Sans cette décomposition, le modèle doit prédire directement des classes fines hétérogènes dans un espace global unique, ce qui augmente fortement l'ambiguïté entre catégories proches.

Ces résultats valident l'hypothèse centrale de ce travail : reformuler la segmentation comme une suite de décisions hiérarchiques simplifie significativement l'apprentissage.

Impact du raffinement conditionnel Le retrait du raffinement conditionnel dégrade également les performances, bien que de manière moins sévère que la suppression des têtes hiérarchiques. Cela indique que la structure hiérarchique constitue le facteur principal, tandis que le raffinement conditionnel agit comme un mécanisme complémentaire permettant d'améliorer la précision locale.

En restreignant la prédiction des sous-classes aux régions sémantiquement cohérentes, ce module réduit les confusions et améliore la robustesse du système.

Discussion Dans l'ensemble, cette étude d'ablation confirme que les gains observés ne proviennent pas uniquement du backbone utilisé, mais bien de l'interaction entre les composantes proposées. Les têtes hiérarchiques assurent la structuration du problème, les prototypes facilitent le transfert inter-domaines, et le raffinement conditionnel améliore la précision des prédictions fines.

Ces résultats renforcent ainsi la validité du cadre proposé pour traiter conjointement la supervision hétérogène, la segmentation hiérarchique et la généralisation inter-domaines.

5.3.6 Résultats qualitatifs

Au-delà des métriques quantitatives, une analyse qualitative permet d'évaluer plus finement le comportement du modèle face aux variations visuelles entre domaines ainsi que la cohérence spatiale des cartes produites.

Nous distinguons ici deux niveaux d'analyse complémentaires :

- une comparaison avec des méthodes reconnues de transfert inter-domaine (FDA et IBN), afin d'évaluer le positionnement global de notre approche ;
- une comparaison directe avec DeepLabV3-ResNet50, afin de visualiser l'apport concret des modules ajoutés au backbone de base.

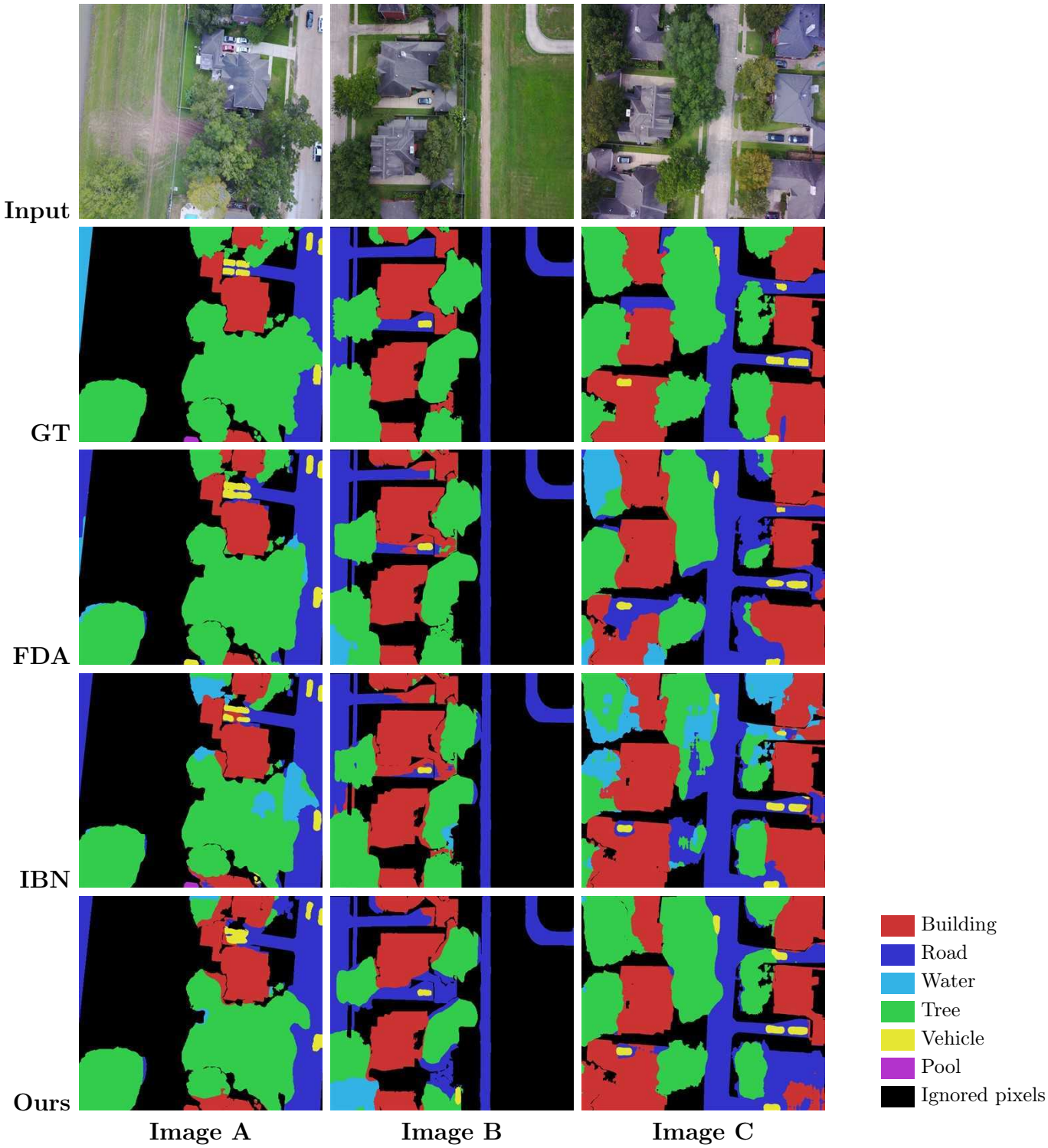


Figure 5.2: Comparaison qualitative de la segmentation des super-classes partagées dans le scénario de transfert RescueNet→FloodNet. Le cadre proposé produit des cartes de segmentation plus cohérentes spatialement et réduit le bruit de prédiction dans les régions complexes de bâtiments, routes et eau.

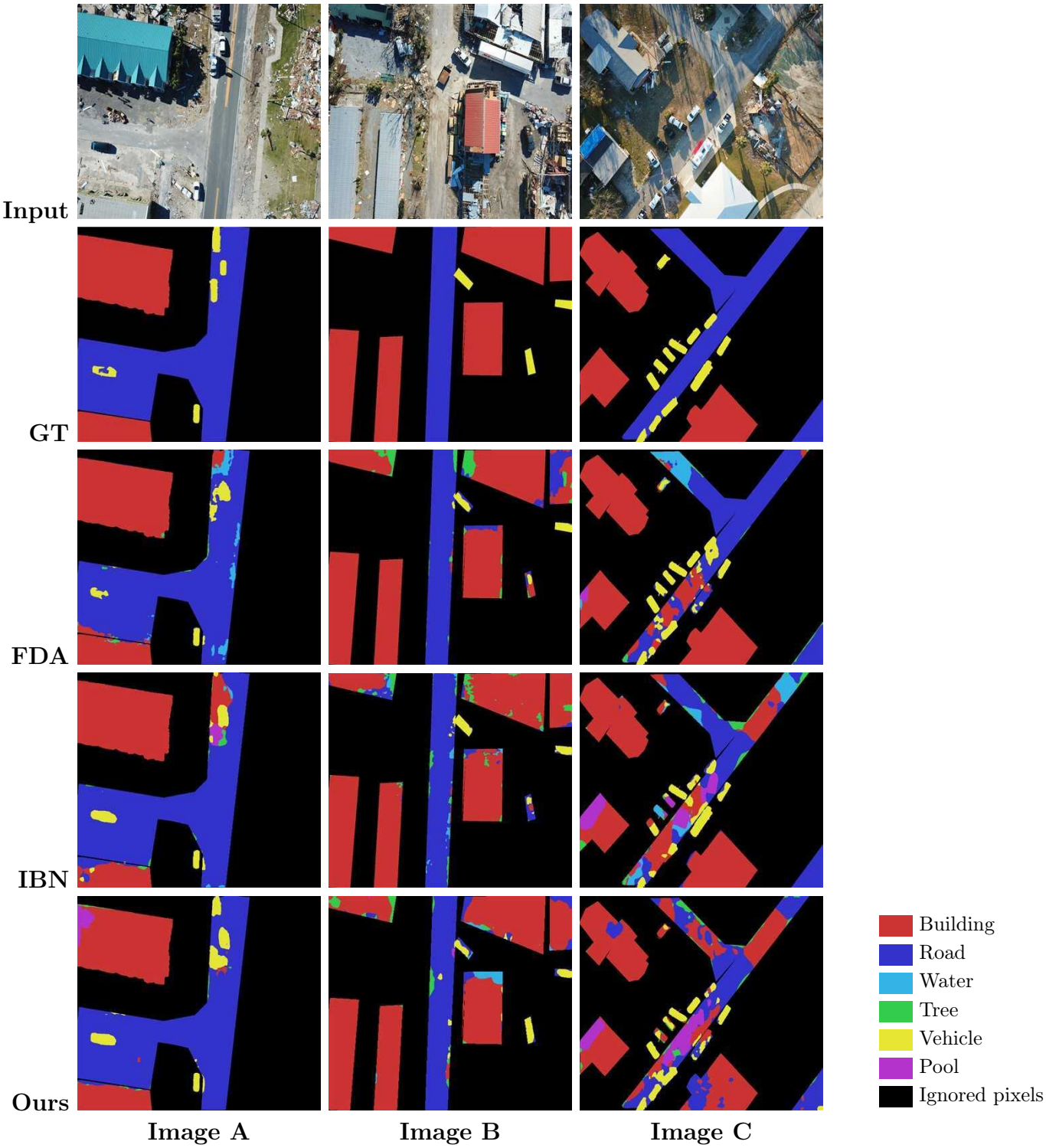


Figure 5.3: Comparaison qualitative de la segmentation des super-classes partagées dans le scénario de transfert FloodNet→RescueNet. Le cadre proposé améliore la cohérence des prédictions dans des scènes présentant des variations visuelles importantes et des taxonomies d'annotation différentes.

Analyse au niveau des super-classes. Les Figures 5.2 et 5.3 montrent que la méthode proposée présente des performances qualitatives concurrentielles face à FDA et IBN, tout en produisant dans plusieurs cas des cartes plus régulières et spatialement plus cohérentes.

Dans le scénario RescueNet→FloodNet, les bâtiments, routes et zones d'eau sont généralement mieux délimités, avec une réduction visible du bruit de prédiction dans certaines régions ambiguës.

Dans le scénario FloodNet→RescueNet, malgré une plus grande variabilité du domaine cible, la méthode proposée conserve de bonnes frontières d'objets et préserve mieux certaines structures complexes.

Comparaison avec DeepLabV3-ResNet50. Les Figures 5.4 et 5.5 comparent directement notre approche au backbone de base DeepLabV3-ResNet50.

Les différences observées mettent en évidence l'apport des composantes introduites : les cartes générées par la méthode proposée apparaissent plus stables, présentent moins de régions incohérentes et montrent une meilleure localisation des structures principales. Les frontières entre classes sont également plus nettes dans plusieurs cas.

Ces observations confirment que les gains obtenus ne résultent pas uniquement d'un backbone performant, mais bien de la structuration hiérarchique et du mécanisme de transfert proposés.

Analyse au niveau des sous-classes L'évaluation des sous-classes a été menée principalement de manière quantitative à travers les tableaux détaillés et l'étude d'ablation. Contrairement aux super-classes partagées, les sous-classes sont plus fines, plus nombreuses et fortement hétérogènes entre FloodNet et RescueNet, ce qui rend leur interprétation visuelle plus délicate dans un cadre inter-domaines.

Les résultats numériques montrent néanmoins que la méthode proposée améliore la distinction entre états inondé / non inondé dans FloodNet et permet une meilleure différenciation des niveaux de dommages dans RescueNet.

Globalement, les résultats qualitatifs confirment que la méthode proposée ne se limite pas à des gains numériques. Elle permet également de produire des cartes de segmentation plus cohérentes spatialement et plus robustes face aux changements de domaine.

Ces observations soutiennent l'idée centrale de cette seconde contribution : concevoir un cadre hiérarchique d'adaptation inter-domaine capable de rester compétitif face aux approches classiques de robustesse visuelle, tout en traitant un problème plus général in-

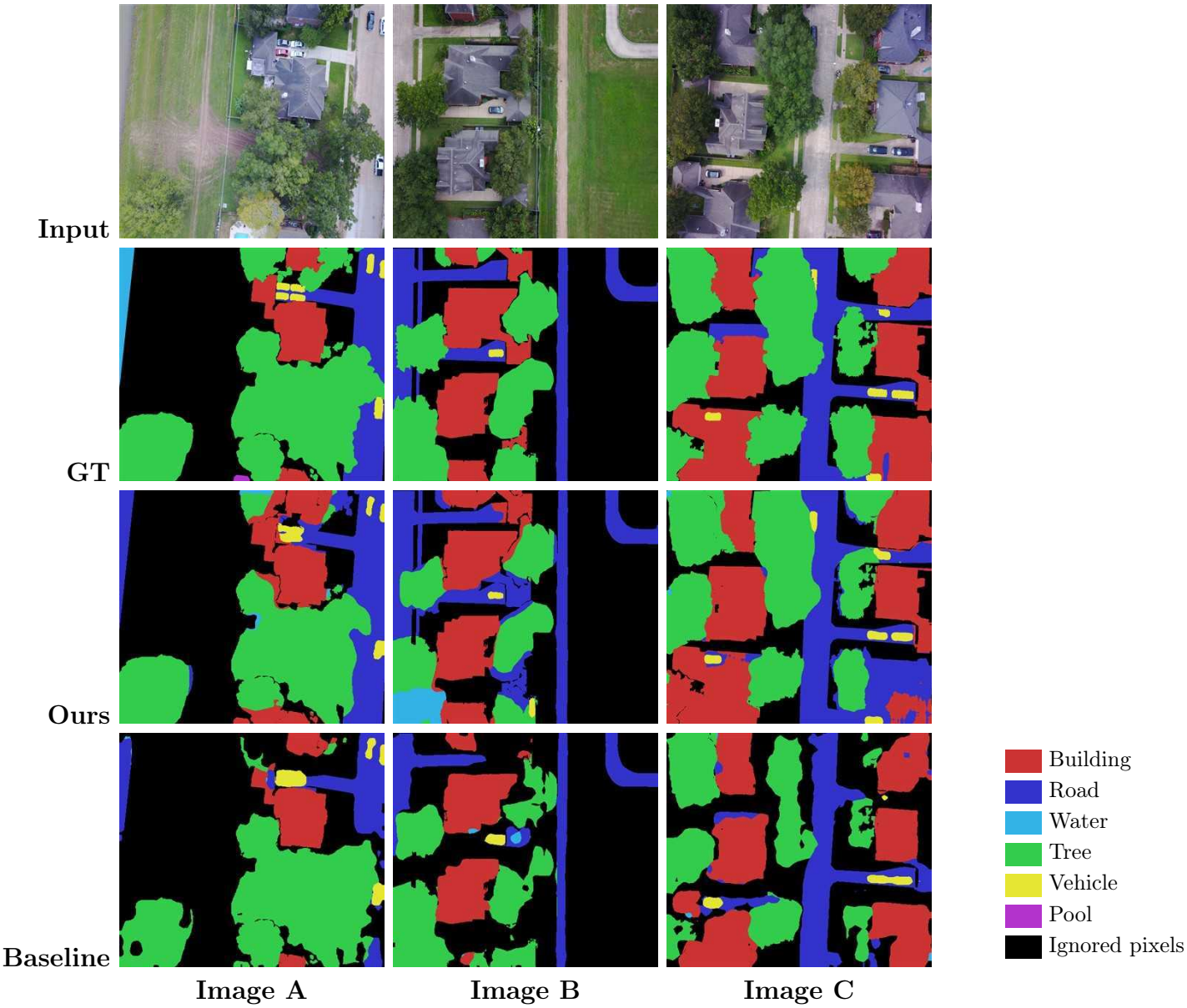


Figure 5.4: Comparaison qualitative verticale entre DeepLabV3-ResNet50 et la méthode proposée dans le scénario RescueNet→FloodNet. Chaque colonne correspond à un exemple distinct. L'ajout du cadre hiérarchique et des mécanismes de transfert permet de produire des cartes plus cohérentes spatialement, avec une meilleure délimitation des bâtiments, des routes et des zones d'eau.

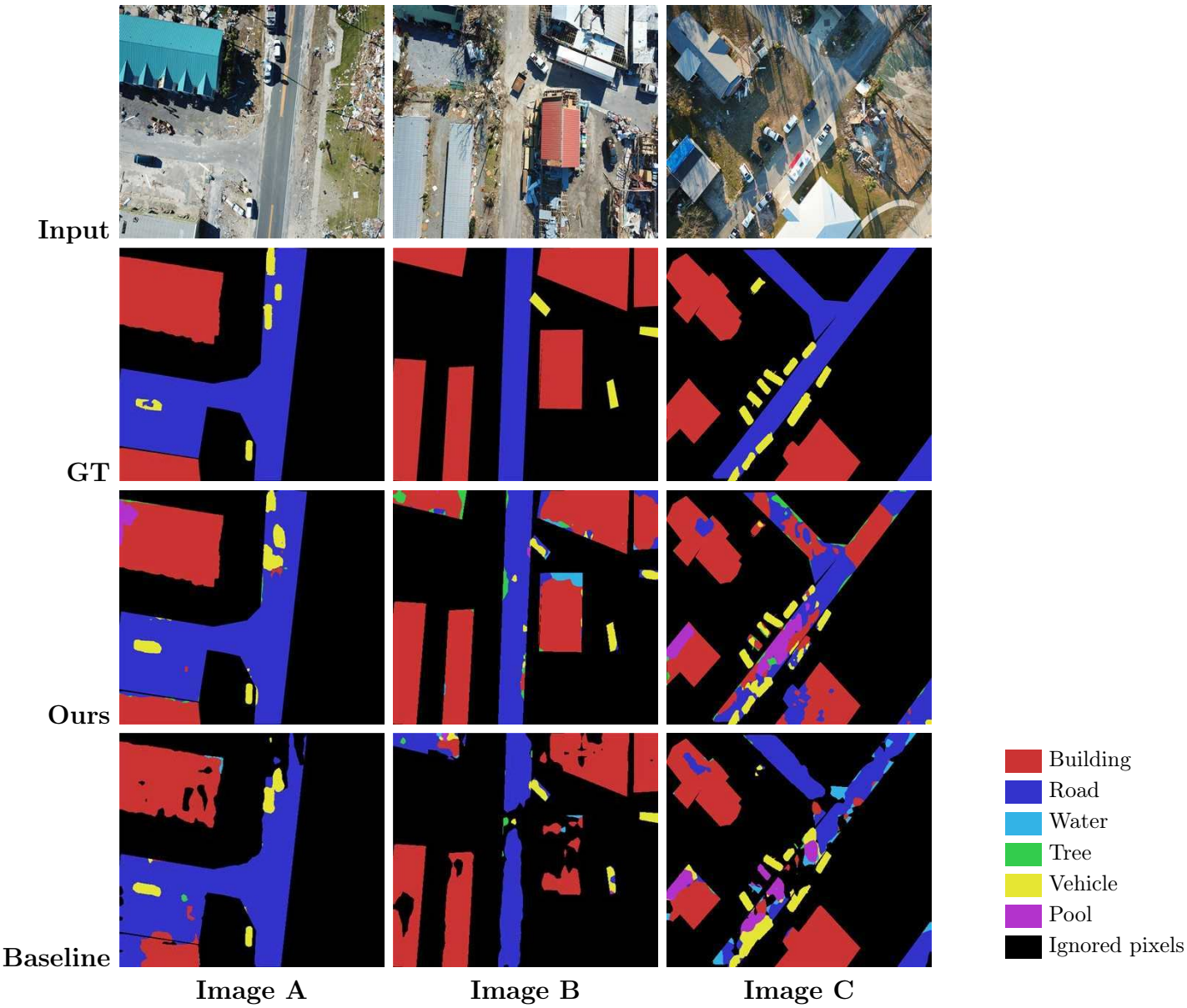


Figure 5.5: Comparaison qualitative verticale entre DeepLabV3-ResNet50 et la méthode proposée dans le scénario FloodNet→RescueNet. Malgré une variabilité visuelle plus importante du domaine cible, la méthode proposée génère des prédictions plus régulières, réduit certaines régions incohérentes et améliore la localisation des structures principales.

tégrant supervision hétérogène, taxonomies distinctes et correspondances partielles entre sous-classes.

5.3.7 Discussion, limites et perspectives de la seconde contribution

Les résultats obtenus montrent que le cadre proposé permet d'améliorer la robustesse de la segmentation dans des scénarios inter-domaines caractérisés non seulement par des variations visuelles importantes, mais également par des taxonomies d'annotation partiellement hétérogènes. Cette contribution étend ainsi les approches classiques d'adaptation inter-domaine en introduisant une structuration hiérarchique explicite de l'espace sémantique et un transfert guidé par prototypes sous taxonomies hétérogènes.

Forces de l'approche proposée L'une des principales forces de cette contribution réside dans la distinction explicite entre super-classes partagées et sous-classes spécifiques à chaque domaine. Contrairement à la majorité des approches d'adaptation inter-domaine, qui supposent un espace de classes identique ou directement compatible entre source et cible, notre méthode permet d'exploiter simultanément des connaissances communes et des annotations spécifiques à chaque jeu de données.

Cette structuration hiérarchique présente plusieurs avantages. Premièrement, elle stabilise le transfert inter-domaines en apprenant des représentations communes à un niveau sémantique plus général. Deuxièmement, elle permet de préserver des informations fines qui seraient perdues si l'apprentissage était limité uniquement à l'intersection des classes communes. Enfin, elle réduit certaines ambiguïtés entre catégories proches en reformulant la segmentation comme une suite de décisions hiérarchiques plus cohérentes.

Une autre force importante de cette contribution concerne l'utilisation de prototypes sémantiques pour représenter les catégories dans l'espace des caractéristiques. Les prototypes permettent de structurer le transfert des connaissances entre domaines de manière plus explicite et plus interprétable. Cette idée rejoint plusieurs travaux récents montrant l'intérêt des représentations compactes de classes pour l'apprentissage métrique, l'adaptation de domaine et les approches few-shot [53]. Dans notre cadre, ces prototypes contribuent à améliorer la cohérence des prédictions fines et facilitent l'exploitation indirecte d'attributs spécifiques à certains datasets.

Le mécanisme de raffinement conditionnel constitue également un apport important de cette contribution. En contraignant les sous-classes à rester cohérentes avec leurs catégories

parentes, il réduit l'apparition de prédictions sémantiquement incohérentes et améliore la stabilité globale des cartes de segmentation.

Par ailleurs, contrairement à plusieurs approches de robustesse visuelle centrées principalement sur les variations de style et d'apparence [18, 52, 19], notre méthode traite explicitement une difficulté rarement abordée dans la littérature : l'hétérogénéité des taxonomies d'annotation. Cette propriété rend le cadre proposé particulièrement adapté aux scénarios réels de télédétection, où les jeux de données sont souvent collectés indépendamment et annotés selon des conventions différentes.

Enfin, les résultats expérimentaux montrent que la méthode proposée demeure compétitive face aux approches classiques de robustesse visuelle et de transfert inter-domaine, tout en traitant un problème plus général intégrant à la fois le décalage visuel et le décalage de taxonomie sémantique. Cela suggère que la structuration sémantique du problème peut jouer un rôle aussi important que l'invariance visuelle des représentations dans la robustesse inter-domaines.

Limites théoriques Malgré les résultats obtenus, cette seconde contribution présente certaines limites théoriques.

La première concerne la dépendance à une hiérarchie définie a priori. Dans ce travail, les relations entre super-classes et sous-classes ont été construites à partir d'une analyse sémantique des annotations disponibles dans les différents jeux de données. Bien que cette hiérarchie soit naturelle dans notre cas d'étude, cette étape peut devenir plus complexe lorsque les catégories présentent des relations ambiguës, partiellement recouvrantes ou difficilement comparables.

Une autre limite concerne la dépendance aux prototypes sémantiques et aux pseudo-labels de haute confiance générés pour certaines sous-classes du domaine cible. Cette limite est particulièrement importante pour les sous-classes appariées dans $\mathcal{P}(g_k)$, car l'alignement fin dépend de la qualité des correspondances sémantiques définies entre les domaines. Lorsque les distributions source et cible deviennent très éloignées, ou lorsque certaines catégories sont fortement sous-représentées, les pseudo-labels peuvent contenir du bruit et limiter la qualité du transfert.

Par ailleurs, bien que la structuration hiérarchique améliore la cohérence des prédictions, certaines sous-classes restent intrinsèquement difficiles à distinguer. Les catégories présentant des différences visuelles très faibles, comme certains niveaux de dommages structurels ou

certains états intermédiaires, continuent de générer des confusions même avec des mécanismes de raffinement conditionnel.

Limites pratiques Sur le plan pratique, le cadre proposé reste dépendant de la disponibilité de plusieurs jeux de données annotés permettant de construire un espace hiérarchique partagé. La définition des correspondances sémantiques entre datasets nécessite encore une intervention humaine et une connaissance experte du domaine applicatif.

De plus, la présence de plusieurs niveaux hiérarchiques, de têtes spécialisées et de mécanismes de transfert supplémentaires augmente la complexité globale de l'entraînement. La stabilisation simultanée des super-classes, des sous-classes et des prototypes peut nécessiter des stratégies d'apprentissage progressives et un réglage attentif des hyperparamètres.

Enfin, cette étude se concentre principalement sur deux jeux de données UAV liés aux catastrophes naturelles. Bien que ces scénarios présentent déjà des écarts visuels et taxonomiques importants, des validations supplémentaires sur un plus grand nombre de domaines permettraient de mieux évaluer la robustesse et l'adaptabilité du cadre proposé.

Perspectives Ces limitations ouvrent plusieurs perspectives de recherche intéressantes.

Une première direction consiste à apprendre automatiquement les relations hiérarchiques entre classes à partir des données elles-mêmes, par exemple via des approches auto-supervisées, des mécanismes de clustering sémantique ou des représentations vision-langage.

Une seconde perspective concerne l'amélioration de la robustesse des pseudo-labels et des prototypes. Des mécanismes fondés sur l'estimation d'incertitude, l'apprentissage enseignant–étudiant ou l'auto-entraînement progressif pourraient permettre de mieux contrôler le bruit introduit pendant l'adaptation inter-domaine et l'estimation des prototypes cible.

Une autre direction importante concerne l'extension vers des scénarios multi-domaines ou multi-sources. Dans des contextes réels, plusieurs jeux de données peuvent être combinés simultanément, avec des distributions et taxonomies encore plus variées. L'apprentissage d'un espace hiérarchique partagé à plus grande échelle constitue ainsi une perspective particulièrement prometteuse.

Enfin, les avancées récentes des modèles fondation en vision et des approches multimodales ouvrent de nouvelles possibilités pour la segmentation inter-domaines. L'intégration de représentations pré-entraînées à grande échelle, de données multispectrales, radar ou textuelles pourrait enrichir la structuration sémantique et améliorer davantage la robustesse des modèles face aux variations réelles des scènes UAV.

Synthèse Dans l'ensemble, cette seconde contribution montre qu'il est possible de traiter conjointement segmentation hiérarchique, supervision hétérogène et adaptation inter-domaine au sein d'un cadre unifié. Les résultats obtenus suggèrent que la robustesse des systèmes de segmentation ne dépend pas uniquement de l'invariance visuelle des représentations, mais également de la manière dont les relations sémantiques entre super-classes partagées et sous-classes spécifiques sont structurées et exploitées pendant l'apprentissage.

Chapitre 6

Conclusion générale et perspectives

6.1 Perspectives à long terme

Au-delà des contributions développées dans cette thèse, les problématiques étudiées s'inscrivent dans une évolution plus large de la vision artificielle appliquée à la télédétection. Les besoins opérationnels actuels dépassent désormais la simple segmentation supervisée sur un jeu de données unique. Les systèmes futurs devront être capables de généraliser à de nouveaux environnements, de s'adapter à des capteurs variés et d'exploiter des annotations hétérogènes ou incomplètes.

Dans cette perspective, une première direction naturelle concerne les approches multi-domaines et multi-sources. De nombreux scénarios réels combinent plusieurs bases de données acquises dans des régions différentes, avec des conditions d'illumination, de résolution ou de saisonnalité variables. Le développement de modèles capables d'apprendre à partir de ces sources hétérogènes constitue un enjeu central pour la robustesse future des systèmes de segmentation.

Une seconde perspective réside dans l'intégration de données multimodales. Les images RGB peuvent être complétées par des informations multispectrales, radar ou LiDAR, permettant d'accéder à des indices complémentaires sur la structure des scènes, la végétation, l'humidité ou la géométrie des objets. La fusion cohérente de ces modalités représente un axe de recherche particulièrement prometteur.

Par ailleurs, l'émergence récente des modèles fondation en vision ouvre de nouvelles possibilités. Des architectures pré-entraînées à très grande échelle pourraient permettre de réduire la dépendance aux annotations spécialisées, d'améliorer la transférabilité entre domaines et de faciliter l'adaptation à de nouvelles tâches avec peu de données supervisées.

Enfin, l'interaction entre vision et langage constitue une perspective de long terme particulièrement intéressante. Des systèmes capables d'exploiter des descriptions textuelles, des consignes expertes ou des connaissances externes pourraient assister plus efficacement l'interprétation automatique des scènes aériennes complexes, notamment en contexte de gestion de crise.

6.2 Synthèse des contributions

Cette thèse avait pour objectif général d'améliorer la robustesse de la segmentation sémantique d'images aériennes à haute résolution, en particulier dans des environnements complexes liés aux catastrophes naturelles.

La première contribution s'est concentrée sur le cadre supervisé classique, en étudiant les difficultés liées au déséquilibre des classes, aux objets de petite taille et aux structures difficiles à segmenter. Pour répondre à ces limites, une architecture combinant indices de détection, fusion multi-échelle et mécanismes attentionnels légers a été proposée. Les résultats obtenus ont montré une amélioration notable de la segmentation des classes minoritaires tout en maintenant de bonnes performances globales.

La seconde contribution a étendu cette réflexion vers un cadre plus général d'adaptation inter-domaine sous taxonomies hétérogènes. Une formulation hiérarchique reposant sur des super-classes communes et des sous-classes spécifiques à chaque domaine a été introduite, accompagnée de mécanismes fondés sur des prototypes sémantiques et un raffinement conditionnel. Les expériences menées ont confirmé l'intérêt de cette approche pour traiter simultanément le décalage visuel entre domaines, la supervision asymétrique et l'hétérogénéité des taxonomies d'annotation.

Ces deux contributions répondent ainsi à deux niveaux complémentaires d'un même problème : améliorer la segmentation lorsque le domaine est connu, puis préserver cette performance lorsque le domaine change.

6.3 Cohérence des travaux et apports scientifiques

Bien que reposant sur des formulations différentes, les travaux présentés dans cette thèse suivent une trajectoire scientifique cohérente. La première contribution s'intéresse à la robustesse locale de la segmentation, c'est-à-dire à la capacité d'un modèle à mieux reconnaître les catégories difficiles au sein d'un domaine donné. La seconde contribution élargit ensuite

cette problématique à la robustesse inter-domaine, en considérant la capacité d'adaptation à des données cibles issues de distributions différentes et à des espaces de classes partiellement incompatibles. Cette progression traduit un passage naturel d'une amélioration centrée sur la qualité de prédiction intra-domaine vers une réflexion plus générale sur l'adaptabilité des systèmes de segmentation. Sur le plan scientifique, plusieurs apports peuvent être retenus :

- l'intérêt d'intégrer des indices issus d'autres tâches de vision, telles que la détection, pour renforcer la segmentation de classes rares ;
- la pertinence d'une structuration hiérarchique des classes pour simplifier l'apprentissage et améliorer la cohérence des prédictions fines ;
- le potentiel des représentations compactes de classes, telles que les prototypes, pour faciliter l'adaptation inter-domaine et le transfert sémantique entre taxonomies partiellement compatibles ;
- l'importance de la formulation du problème, au-delà du seul choix d'une architecture profonde.

Ces éléments montrent qu'une segmentation robuste ne repose pas uniquement sur des réseaux plus profonds ou plus lourds, mais également sur des stratégies d'apprentissage mieux adaptées à la structure réelle des données.

6.4 Remarques finales

Comme tout travail de recherche, cette thèse présente certaines limites, notamment le périmètre expérimental principalement centré sur des jeux de données UAV liés aux catastrophes, ainsi que la dépendance partielle à des structures hiérarchiques définies manuellement. Ces limites constituent toutefois des prolongements naturels plutôt que des remises en cause des approches proposées.

En définitive, cette thèse a cherché à contribuer au développement de méthodes de segmentation plus robustes, plus flexibles et mieux adaptées aux contraintes des environnements réels. Les résultats obtenus montrent qu'il est possible d'améliorer à la fois la reconnaissance des classes difficiles et l'adaptation entre jeux de données hétérogènes. Nous espérons que ces travaux pourront servir de base à de futures recherches visant à concevoir des systèmes de compréhension visuelle plus généraux, capables d'assister efficacement les applications critiques de télédétection, d'analyse environnementale et de gestion des catastrophes.

Bibliographie

[1] Jonathan Long et al. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[2] Olaf Ronneberger et al. 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*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[3] L. C. Chen et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[4] M. Rahnemoonfar et al. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

[5] M. Rahnemoonfar et al. Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment. *Scientific Data*, 10(913), 2023.

[6] L. P. Osco et al. A review on deep learning in uav remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 102:102456, 2021.

[7] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[8] E. Xie et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

[9] P. Chen et al. A deep learning-based solution to the class imbalance problem in high-resolution land cover classification. *Remote Sensing*, 17:1845, 2025.

[10] X. Li et al. Semantic segmentation of uav remote sensing images based on edge feature fusing and multi-level upsampling integrated with deeplabv3+. *PLOS ONE*, 18(1):e0279097, 2023.

[11] N. S. Hewawiththi et al. Damage assessment after natural disasters with uavs: Semantic feature extraction using deep learning. *arXiv preprint arXiv:2412.10756*, 2024.

[12] I. Alisjahbana et al. Deepdamagenet: A two-step deep-learning model for multi-disaster building damage segmentation and classification using satellite imagery. *Remote Sensing of Environment*, 312:113456, 2024.

[13] S. Verma et al. Segdesicnet: Lightweight semantic segmentation in remote sensing with geo-coordinate embeddings for domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[14] S. Lyu et al. Deep learning based domain adaptation methods in remote sensing: A comprehensive survey. *arXiv preprint arXiv:2510.15615*, 2025.

[15] L. Hoyer et al. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):220–235, 2024.

[16] Y.-H. Tsai et al. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7472–7481, 2018.

[17] L. Hoyer et al. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9924–9935, 2022.

[18] X. Pan et al. Two at once: Enhancing learning and generalization capacities via ibn-net. In *European Conference on Computer Vision (ECCV)*, 2018.

[19] Y. Yang and S. Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] P. Zhang et al. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12414–12423, 2021.

[21] L. Hoyer et al. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11721–11732, 2023.

[22] D. Kim et al. Learning semantic segmentation from multiple datasets with label shifts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2022.

[23] J. Lim and Y. Kim. Cross-domain semantic segmentation on inconsistent taxonomy using vision-language models. *arXiv preprint arXiv:2408.02261*, 2024.

[24] L.-C. Chen et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pages 834–848, 2017.

[25] H. Zhao et al. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[26] Yuhui Yuan et al. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2020.

[27] Lei Ma et al. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, 2019.

[28] L. Ran et al. Ddf: A novel dual-domain image fusion strategy for remote sensing image semantic segmentation with unsupervised domain adaptation. *arXiv preprint arXiv:2403.02784*, 2024.

[29] W. Zeng et al. Domain adaptive remote sensing image semantic segmentation with prototype guidance. *Neurocomputing*, 580:127484, 2024.

[30] W. Liu et al. Source-free domain adaptive semantic segmentation of remote sensing images with diffusion-guided label enrichment. *arXiv preprint arXiv:2509.18502*, 2025.

[31] R. Gong et al. Tacs: Taxonomy adaptive cross-domain semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–35, 2022.

[32] H. Sun et al. Dynalign: Unsupervised dynamic taxonomy alignment for cross-domain segmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[33] Ian Goodfellow et al. *Deep Learning*. MIT Press, 2016.

[34] Andrej Karpathy and Fei-Fei Li. Cs231n convolutional neural networks for visual recognition: Convolutional networks. `http://cs231n.github.io/convolutional-networks/`. Accessed: 2025-09-11.

[35] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[36] John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing*, 68:227–236, 1990.

[37] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[38] Liang-Chieh Chen et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*, 2014.

[39] Q. Zou et al. Multi-scale aggregation stereo matching network based on dense grouping atrous convolution. *Applied Sciences*, 13(12):7033, 2023.

[40] Ashish Vaswani et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[41] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.

[42] Salman Khan et al. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022.

[43] Ross Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[44] J. Redmon et al. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[45] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[46] S. Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[47] Li Liu et al. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318, February 2020.

[48] Wei Liu et al. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[49] T.-Y. Lin et al. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[50] Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[51] Jindong Wang et al. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[52] K. Zhou et al. Domain generalization with mixstyle. In *International Conference on Learning Representations (ICLR)*, 2021.

[53] J. Snell et al. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[54] Z. Zhou et al. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA + ML-CDS 2018), MICCAI Workshops*, pages 3–11, Granada, Spain, 2018. Springer.

[55] W. Li et al. Unet-cbam: Attention-augmented u-net for flood mapping from sar imagery. *Remote Sensing*, 14(18):4527, 2022.

[56] D. W. Nugraha et al. Performance improvement of deep convolutional networks for aerial imagery segmentation of natural disaster-affected areas. *International Journal on Informatics Visualization*, 2023.

[57] S. Deng et al. A two-stage deep learning approach for post-disaster building damage assessment from remote sensing imagery. *Remote Sensing*, 14(21):5311, 2022.

[58] T. G. J. Rudner et al. Multi3net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 702–709, 2019.

[59] M. Rahnemoonfar et al. Flood mapping using multi-sensor satellite imagery and deep learning. *Remote Sensing*, 10(11):1683, 2018.

[60] Y. Shen et al. Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 60(5402114), 2022.

[61] J. Yue et al. Treeunet: A tree-structured u-net for high-resolution remote sensing image segmentation. In *Proceedings of IGARSS*, pages 4235–4238, 2019.

[62] A. Foivos et al. Resunet-a: A deep learning framework for semantic segmentation of high-resolution satellite imagery. *Remote Sensing*, 12(3):402, 2020.

[63] R. Priyanka et al. Diresunet: A dense inception residual u-net for disaster scene segmentation. *Remote Sensing Letters*, 13(4):321–330, 2022.

[64] X. Liu et al. Integrating lidar and optical imagery for urban structure segmentation using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156:108–119, 2019.

[65] G. Chen et al. Sdfcnv2: Spatial-channel fusion network for multi-spectral semantic segmentation. *Remote Sensing*, 13(15):2935, 2021.

[66] A. Gupta et al. Deep learning-based aerial image segmentation with open data for disaster impact assessment. *Neurocomputing*, 439:22–33, 2021.

[67] L.-C. Chen et al. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[68] L.-C. Chen et al. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[69] S. Venugopal. A dilated cnn for flood extent segmentation using deeplab. *Remote Sensing Letters*, 11(5):451–460, 2020.

[70] Y. Wang et al. Feature-regularized deeplab for robust disaster segmentation under limited data. *Remote Sensing*, 13(18):3572, 2021.

[71] P. Andrade et al. Improved deeplabv3+ for wildfire-affected forest segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 112:102895, 2022.

[72] X. Wang et al. Cfamnet: Class-aware feature attention module for disaster segmentation. *Remote Sensing*, 14(14):3298, 2022.

[73] M. Wang et al. Improving domain generalization with instance-batch normalization in deeplabv3+. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:125–137, 2022.

[74] J. Wang et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.

[75] A. A. Sundaresan and A. A. Solomon. Post-disaster flooded region segmentation using deeplabv3+ and unmanned aerial system imagery. *International Journal of Applied Earth Observation and Geoinformation*, 126:103671, 2024.

[76] So-Hyeon Jo et al. Damage detection and segmentation in disaster environments using combined yolo and deeplab. *Remote Sensing*, 16(22):4267, 2024.

[77] X. Zhu et al. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2022–2031, Virtual, 2021.

[78] R.Z.J. Liu et al. An improved instance segmentation method for fast assessment of damaged buildings based on post-earthquake uav images. *Sensors*, 24:4371, 2024.

[79] Y. Zou et al. Yolo-based building damage assessment in uav imagery for post-earthquake disaster management. preprint, 2024.

[80] Nicolas Carion et al. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[81] K. He et al. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[82] J. Santos et al. Uav-based detection and segmentation framework for real-time disaster response. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[83] Zongze Zhao et al. Deep object segmentation and classification networks for building damage detection using the xbd dataset. *International Journal of Digital Earth*, 17(1):1–18, 2024.

[84] M. R. Jewel et al. Dis-mine: Instance segmentation for disaster-awareness in poor-light condition in underground mines. In *Proceedings of the IEEE International Conference on Big Data (BigData)*, pages 456–465, Los Angeles, CA, USA, 2024.

[85] X. Zhou et al. Yolov5-seg with transformer attention for post-earthquake building assessment. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2024.

[86] Dzmitry Bahdanau et al. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[87] J. Hu et al. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, Salt Lake City, UT, USA, 2018.

[88] Z. Huang et al. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 603–612, Seoul, Korea, 2019.

[89] J. Fu et al. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3141–3149, Long Beach, CA, USA, 2019.

[90] Q. Zhao et al. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:5403913, 2022.

[91] Y. Chen et al. Hybrid attention fusion embedded in transformer for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4421–4435, 2024.

[92] Y. Xu et al. Remote sensing image semantic segmentation sample generation using a decoupled latent diffusion framework. *Remote Sensing*, 17:2143, 2025.

[93] R. Zhang et al. Lsrformer: Efficient transformer supply convolutional neural networks with global information for aerial image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5610713, 2024.

[94] Sixiao Zheng et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021.

[95] Z. Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[96] H. Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the IEEE International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022.

[97] Y. Zhong et al. Cnn-transformer hybrid network for flood and wildfire mapping. *Remote Sensing*, 2022.

[98] L. Zhao et al. Hybrid cnn–transformer model for disaster-induced landscape change detection. *Remote Sensing Letters*, 2021.

[99] Y. Gao et al. Stransfuse: Fusing transformers and cnns for remote sensing image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[100] Y. Xu et al. Edgetrans: Lightweight transformer with edge detection for disaster segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[101] Y. Liu et al. Land-cover segmentation with swin transformer and upernet decoder. *Remote Sensing*, 2022.

[102] H. Li et al. Cross-modal fusion of optical and sar data using transformers for disaster mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022.

[103] X. Ma et al. Multi-level transformer fusion for segmentation of complex disaster scenes. *Remote Sensing of Environment*, 2022.

[104] T. H. Rafi et al. Domain generalization for semantic segmentation: A survey. *Artificial Intelligence Review*, 57(9):247, 2024.

[105] M. T. Shaheen et al. Swinsegformer: Advancing aerial image semantic segmentation for flood detection. *IEEE Open Journal of the Computer Society*, 6:645–657, 2025.

[106] L. Hoyer et al. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 372–391, 2022.

[107] J. Huang et al. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[108] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[109] W. Cai et al. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025.

[110] M. Luo et al. Domain generalization for semantic segmentation of remote sensing images via vision foundation model fine-tuning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:126–146, 2025.

[111] Z. Gong et al. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[112] S. Jeon et al. Exploiting domain properties in language-driven domain generalization for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[113] J. Lambert et al. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2888, 2020.

[114] S.-A. Choe et al. Universal domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4607–4617, 2025.

[115] P. Bevandić and S. Šegvić. Automatic universal taxonomies for multi-domain semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022.

[116] R. Ma et al. Automated label unification for multi-dataset semantic segmentation with gnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 80141–80164, 2024.

[117] K. Wang et al. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019.

[118] X. Wang et al. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021.

[119] H. Zhang et al. Context encoding for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[120] M. A. Islam et al. Label refinement network for coarse-to-fine semantic segmentation. *arXiv preprint arXiv:1703.00551*, 2017.

[121] B. R. Muller and W. A. P. Smith. A hierarchical loss for semantic segmentation. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP/VISAPP)*, 2020.

[122] Liulei Li et al. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[123] A. Picon et al. Taxonomic hierarchical loss function for enhanced crop and weed phenotyping in multi-task semantic segmentation. *Smart Agricultural Technology*, 10:100761, 2025.

[124] T.-H. Vu et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2517–2526, 2019.

[125] Y. Lyu et al. Uavid: A semantic segmentation dataset for uav imagery. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-2-2020, pages 1–8, 2020.

[126] D. Zhang et al. Prototype-based low altitude uav semantic segmentation. *arXiv preprint arXiv:2604.01550*, 2026.

[127] S. Bultmann et al. Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation. *arXiv preprint arXiv:2210.09739*, 2022.

[128] L. Zhang et al. Efficientsegnet: Lightweight semantic segmentation with feature enhancement for real-time applications. *Sensors*, 2025.

[129] A. Zetout and M. S. Allili. Csdnet: Context-aware segmentation of disaster aerial imagery using detection-guided features and lightweight transformers. *Remote Sensing*, 17(14):2337, 2025.

[130] A. Khan et al. Emerging uav technology for disaster detection, mitigation, response, and preparedness. *Journal of Field Robotics*, 39(6):905–955, 2022.

[131] N. Le et al. 3daerorelief: The first 3d benchmark uav dataset for post-disaster assessment. *arXiv preprint arXiv:2509.11097*, 2025.

[132] S. Chu et al. Learning debiased and disentangled representations for semantic segmentation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 8355–8366, Virtual, 2021.

[133] H. Zhang et al. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, Salt Lake City, UT, USA, 2018.

[134] S. Mehta and M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 2022.

[135] A. Paszke et al. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv*, arXiv:1606.02147, 2016.

[136] X. Chen et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

[137] H. Wu et al. Cmtfnet: Cnn and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:2004612, 2023.

[138] M.L. Allaoui and M.S. Allili. MixLVMM: A mixture of lightweight vision mamba model for enhancing skin lesion segmentation across high tone variability. *IEEE Access*, 13:121234–121249, 2025.

[139] D. Yapi, A. Nouboukpo, and M.S. Allili. Mixture of multivariate generalized gaussians for multi-band texture modeling and representation. *Signal Processing*, 209:109011, 2023.

[140] Y. Bouafia, M.S. Allili, L. Hebbache, and L. Guezouli. SES-ReNet: Lightweight deep learning model for human detection in hazy weather conditions. *Signal Processing: Image Communication*, 130:117223, 2025.

[141] A. Yaghmour et al. A sensor agnostic domain generalization framework for leveraging geospatial foundation models: Enhancing semantic segmentation via synergistic pseudo-labeling and generative learning. *arXiv preprint arXiv:2505.01558*, 2025.

[142] P. Bevandić et al. Multi-domain semantic segmentation with overlapping labels. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2422–2431, 2022.