Université du Québec en Outaouais

Maîtrise en sciences et technologies de l'information (profil mémoire)

Détection en des Deepfakes Audiovisuels dans les Flux Vidéo à l'aide de Représentations d'Apprentissage Profond

Mémoire présenté par

Dia Elhak Temmar

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

Codirecteur de recherche : Prof. Anderson Raymundo Avila (INRS)

Août 2026

Ce mémoire est évalué par un jury composé de :

- **Président du jury :** Prof. Raphael Khoury (UQO)
- **Membre du jury :** Prof. Zakaria Abou El Houda (INRS)
- **Directeur de recherche :** Prof. Mohand Saïd Allili (UQO)
- **Codirecteur de recherche :** Prof. Anderson Raymundo Avila (INRS)

Liste des abréviations, sigles et acronymes

Abréviation	Signification
AI	intelligence artificielle
AUC	aire sous la courbe ROC
AV	audiovisuel
AV-HuBERT	extension audiovisuelle de HuBERT
BLSTM	mémoire longue à court terme bidirectionnelle
CNN	réseau de neurones convolutif
CVF	Computer Vision Foundation
DFDC [1])	Défi de détection des deepfakes
EER	taux d'erreur égal
FAR	taux de fausse acceptation
FRR	taux de faux rejet
GAN	réseau antagoniste génératif
GRU	unité récurrente à portes
HuBERT	modèle de parole auto-supervisé BERT à unités cachées
IEEE	Institute of Electrical and Electronics Engineers
KLD	divergence de Kullback-Leibler
LRS3	jeu de données Lip Reading Sentences 3
LRW	jeu de données Lip Reading in the Wild
LSTM	mémoire longue à court terme
MAE	autoencodeur masqué (préentraînement auto-supervisé)
MFA	Montreal Forced Aligner
MSE	erreur quadratique moyenne
ResNet	réseau résiduel
RNN	réseau de neurones récurrent
ROC	caractéristique de fonctionnement du récepteur
ROI	région d'intérêt
t-SNE	plongement stochastique de voisins distribué selon t
TCN	réseau convolutif temporel
TPR	taux de vrais positifs
TTS	synthèse vocale à partir du texte
UCF	University of Central Florida (jeu de données UCF-Crime)
URL	localisateur uniforme de ressource

Abréviation	**Signification**
VAD	détection d'anomalies vidéo
VC	conversion vocale
VGG	réseau convolutif Visual Geometry Group
ViT	transformer de vision
VITS	modèle neuronal de synthèse vocale de bout en bout
wav2vec 2.0	modèle auto-supervisé de représentation de la parole à partir d'ondes audio brutes
XD	jeu de données d'anomalies vidéo XD-Violence

Dédicace

À la mémoire de mon cher père ***Ahmed****,*
dont la sagesse, les sacrifices et la présence constante dans mon cœur continuent de guider chacun de mes pas.

À ma chère mère ***Alia****,*
pour son amour inconditionnel, ses prières, sa patience et son soutien à chaque instant.

À ma sœur et à mon frère,
pour leur présence, leurs encouragements et la confiance qu'ils m'ont toujours accordée.

À mes amis et à mes collègues,
pour leur gentillesse, leur aide au quotidien et leur soutien discret mais précieux.

À mon directeur de recherche ainsi qu'à tous mes enseignants,
pour leurs conseils, leur exigence et tout ce qu'ils m'ont transmis,
tant sur le plan scientifique qu'humain.

Je vous dédie ce mémoire avec ma plus profonde gratitude et mon plus grand respect.

Remerciements

Je tiens tout d'abord à exprimer ma sincère gratitude à mon directeur de recherche, le professeur **Mohand Saïd Allili**, pour son encadrement, sa disponibilité et la confiance qu'il m'a accordée tout au long de ce travail. Ses conseils, ses remarques constructives et son expertise scientifique ont grandement contribué à l'avancement de ce mémoire et à ma formation en recherche.

Je souhaite également remercier chaleureusement mon codirecteur de recherche, **Pro. Anderson Raymundo Avila**, pour son accompagnement, ses conseils et les échanges scientifiques qui ont enrichi ce travail. Son regard complémentaire et ses suggestions ont été particulièrement précieux au cours de la réalisation de ce projet.

Mes remerciements vont également à l'ensemble des membres et collègues du laboratoire **LARIVA** pour leur accueil, leur disponibilité et l'environnement de travail stimulant qu'ils ont contribué à créer. Les discussions, les échanges d'idées et les moments partagés au laboratoire ont constitué une expérience enrichissante, tant sur le plan scientifique que humain.

Enfin, je remercie toutes les personnes qui, directement ou indirectement, ont contribué à la réalisation de ce mémoire et m'ont accompagné au cours de cette étape de mon parcours académique.

Résumé

Ce mémoire aborde le défi croissant de la détection des deepfakes audiovisuels, stimulé par les avancées rapides des technologies d'IA générative. Il propose une étude centrée sur les incohérences fines qui apparaissent dans la parole et dans la dynamique visuelle de la bouche, souvent ignorées par les approches traditionnelles basées sur des caractéristiques globales. Le travail met l'accent sur une modélisation fine des dynamiques de la parole et des expressions faciales, en particulier sur les différences entre parole réelle et synthétique ainsi que sur les variations de géométrie et de texture des lèvres.

Pour cela, le mémoire présente deux contributions principales : (1) une analyse au niveau des phonèmes basée sur des représentations auto-supervisées de la parole afin d'identifier des motifs linguistiques sensibles aux manipulations et d'améliorer l'interprétabilité ; et (2) une stratégie de fusion précoce combinant la géométrie des lèvres et les caractéristiques de texture pour modéliser finement les variations de mouvement et d'apparence faciale. Une extension de fusion audio–vidéo avec localisation temporelle des segments manipulés est conservée uniquement comme perspective de recherche et comme base possible pour un futur article scientifique.

Abstract

This thesis addresses the growing challenge of detecting audio-visual deepfakes, driven by rapid advances in generative AI technologies. It studies subtle inconsistencies in speech and visual mouth dynamics that are often overlooked by traditional approaches based on global features. The work focuses on fine-grained modeling of real versus synthetic speech, as well as on geometric and texture variations around the lips.

To achieve this, the thesis presents two main contributions : (1) a phoneme-level analysis based on self-supervised speech representations to identify manipulation-sensitive linguistic patterns and improve interpretability ; and (2) an early-fusion strategy that combines lip geometry and texture features to finely model variations in facial motion and appearance. A future audio-visual fusion extension with temporal localization of manipulated segments is kept only as future research and as a possible basis for a future scientific paper.

Table des matières

Liste des figures

Liste des tableaux

CHAPITRE 1

Introduction générale

1.1 Contexte

Le paysage numérique connaît une transformation profonde sous l'effet des avancées récentes en intelligence artificielle. L'une des évolutions les plus perturbatrices est la montée des *deepfakes*, c'est-à-dire des vidéos et des pistes audio synthétiques très réalistes capables de représenter de manière convaincante des personnes en train de dire ou de faire des choses qu'elles n'ont jamais dites ni faites. Ces falsifications sont rendues possibles par des modèles génératifs puissants, notamment les réseaux antagonistes génératifs (GAN) [14] et les modèles de diffusion [15], qui peuvent synthétiser un visage parlant complet ou cloner une voix à partir de quelques images ou d'une courte consigne textuelle [16]. Dans des systèmes récents tels que SORA, les modèles de diffusion sont conditionnés par des descriptions en langage naturel et produisent des séquences vidéo cohérentes où les mouvements des lèvres sont alignés avec les phrases prononcées.

Les implications de cette technologie sont considérables. Si les avatars réalistes permettent de nouvelles formes d'interaction humain-machine et de narration virtuelle, des acteurs malveillants peuvent aussi exploiter les deepfakes à des fins de fraude, de chantage ou de désinformation [17]. Une fausse vidéo montrant une personnalité publique tenant des propos incendiaires, ou une voix synthétique autorisant une transaction bancaire frauduleuse, peut avoir des conséquences graves pour les individus et pour la société [18]. À mesure que les deepfakes se diffusent sur les réseaux sociaux et les plateformes d'information, le maintien de l'intégrité des médias numériques devient un enjeu critique.

En parallèle, l'ensemble des dispositifs de captation du monde réel, caméras de surveillance, assistants intelligents et caméras embarquées, génère de grandes quantités de données audiovisuelles qui doivent être surveillées en temps réel pour des raisons de sûreté et de sécurité. Les systèmes doivent non seulement filtrer les contenus fabriqués, mais aussi repérer des anomalies réelles, comme des collisions, des bagarres ou d'autres événements dangereux. Les opérateurs humains ne peuvent pas examiner l'ensemble de ces données, et les solutions entièrement automatisées doivent composer avec les occlusions, les variations d'éclairage et le bruit de fond [19, 20]. Une simple moyenne de l'information sur toute la durée d'une séquence est

insuffisante, car les indices révélateurs d'une manipulation ou d'une anomalie réelle résident souvent dans des motifs subtils et transitoires.

Ce mémoire aborde cette lacune en explorant des *représentations audiovisuelles* capables de capter ces incohérences fines. Nous partons de l'idée que les erreurs de synchronisation, c'est-à-dire les décalages entre les mouvements des lèvres et les mots prononcés, trahissent souvent les deepfakes [16], tandis que des événements acoustiques anormaux, comme des cris ou des explosions, peuvent signaler des anomalies dans des séquences visuellement peu informatives [20]. En portant une attention particulière à la région de la bouche et au contenu phonétique de l'audio, notre objectif est de détecter où et comment une vidéo s'écarte de la réalité.

1.1.1 Deepfakes Audiovisuels en temps réel

De nos jours, les deepfakes sont générés au moyen d'une combinaison de techniques de vision par ordinateur et de synthèse de la parole [21]. Les modèles de réanimation faciale projettent un visage source sur un visage cible tout en préservant les expressions et les mouvements de la tête [22]. Les systèmes de clonage vocal, fondés sur des architectures de synthèse vocale à partir du texte (TTS) et de conversion vocale (VC), reproduisent l'intonation et la prononciation d'une personne à partir de courts enregistrements [13]. Combinées, ces technologies peuvent produire en temps réel une entrevue entièrement fabriquée, avec des mouvements de lèvres qui semblent correspondre à la parole synthétique. La facilité de création de tels contenus abaisse considérablement la barrière d'entrée à la désinformation.

1.1.2 Limites des détecteurs existants

Malgré l'intérêt croissant de la recherche, de nombreux détecteurs de deepfakes reposent encore sur des caractéristiques globales au niveau de la séquence, comme des statistiques agrégées des valeurs de pixels ou des plongements audio moyennés [23]. Ces approches résistent difficilement aux falsifications sophistiquées qui maintiennent une cohérence globale tout en introduisant des incohérences subtiles. Les micro-expressions autour de la bouche, des paupières ou des joues peuvent paraître légèrement non naturelles ; les motifs temporels, comme l'accélération et la décélération des mouvements des lèvres, peuvent ne pas suivre les normes physiologiques ; et le minutage des phonèmes prononcés peut être légèrement désynchronisé par rapport à l'articulation visuelle. Par conception, les descripteurs globaux lissent ces détails.

Un autre défi tient au couplage entre les modalités audio et vidéo. Une vidéo peut sembler authentique lorsqu'elle est examinée seule, et une piste audio peut paraître naturelle lorsqu'elle

est écoutée isolément ; toutefois, lorsqu'elles sont alignées image par image, les mouvements des lèvres et les segments de parole peuvent ne pas correspondre parfaitement [16]. À l'inverse, un événement réellement anormal, comme un cri soudain hors champ suivi d'une agitation, peut ne pas se manifester clairement dans le canal visuel seul. Les détecteurs qui ignorent les relations intermodales risquent donc de manquer des indices critiques.

1.2 Problématique

Les deepfakes modernes produisent des contenus audio et vidéo de plus en plus réalistes, rendant leur détection particulièrement difficile. Dans une vidéo authentique, les mouvements de la bouche et des lèvres sont naturellement synchronisés avec la parole produite. Cette relation étroite entre les modalités audio et visuelle constitue une source d'information importante pour l'évaluation de l'authenticité d'un contenu. Cependant, lors de la génération d'un deepfake, cette cohérence n'est pas toujours parfaitement préservée. Des incohérences subtiles peuvent apparaître entre les phonèmes prononcés et les mouvements observés dans la région buccale, sous la forme d'erreurs de synchronisation, de transitions articulatoires non naturelles ou de déformations visuelles localisées.

La problématique étudiée dans ce mémoire consiste à exploiter conjointement les informations audio et visuelles afin d'identifier ces incohérences et de distinguer un contenu authentique d'un contenu manipulé. L'hypothèse centrale est que les deepfakes introduisent des ruptures de cohérence entre la parole et les mouvements faciaux associés, particulièrement au niveau de la bouche. Ainsi, l'analyse de la correspondance entre les caractéristiques phonétiques du signal audio et les indices visuels de l'articulation constitue une piste prometteuse pour la détection robuste des manipulations audiovisuelles.

1.3 Objectifs

Pour traiter ce problème, nous poursuivons les objectifs suivants :

1. **Développer des représentations audio** capables de capter les différences entre parole réelle et parole synthétique au niveau des phonèmes, des voyelles, des consonnes et de certains mots courts.
2. **Concevoir un détecteur visuel efficace** fondé sur la fusion précoce des descripteurs géométriques et texturaux de la bouche, afin d'améliorer la détection des manipulations faciales subtiles.
3. **Préparer l'extension audiovisuelle** en identifiant comment les résultats audio et

visuels pourront être intégrés ultérieurement dans un détecteur en temps réel avec localisation temporelle, qui fera l'objet de travaux futurs et d'une publication scientifique dédiée.

1.3.1 Contributions

À la lumièe de ces objectifs, nous proposons une approche située à l'intersection de la phonétique et de la dynamique faciale. Dans cette version du mémoire, les contributions retenues sont les suivantes :

1. **Analyse de divergence au niveau des phonèmes :** en exploitant le modèle de parole auto-supervisé HuBERT, nous extrayons des plongements pour des phonèmes et des mots individuels. En projetant les plongements d'énoncés réels et synthétiques dans un espace commun et en calculant les divergences de Kullback-Leibler, nous identifions les sons de parole les plus sensibles à la manipulation [13]. Cette méthode met en évidence des indices linguistiquement significatifs et améliore l'interprétabilité : lorsqu'un détecteur signale une anomalie, il peut pointer vers des voyelles ou des consonnes précises où la voix synthétique s'écarte de la distribution naturelle.
2. **Fusion précoce de la géométrie et de la texture des lèvres :** du côté visuel, nous nous concentrons sur la région de la bouche, où les articulateurs de la parole sont les plus visibles. Nous suivons 40 points de repère des lèvres par image et les normalisons afin de réduire les effets de translation, de rotation et d'échelle. Nous recadrons également une région d'intérêt serrée autour des lèvres et extrayons des caractéristiques de texture à l'aide d'un réseau de neurones convolutif léger. En concaténant tôt les descripteurs géométriques et texturaux dans le réseau, nous permettons au modèle d'apprendre des représentations conjointes qui capturent à la fois la forme des lèvres et leur apparence de surface au fil du temps [23]. Cette fusion précoce facilite la détection d'incohérences subtiles d'une image à l'autre et fournit une base solide pour une extension audiovisuelle ultérieure.

L'intégration complète de ces deux axes dans une extension de fusion audio–vidéo n'est pas présentée comme une contribution expérimentale dans cette version. Elle est plutôt conservée comme perspective de recherche et comme sujet potentiel d'un futur article scientifique.

1.4 Organisation du mémoire

Afin de guider la lecture, ce mémoire est organisé comme suit. Le chapitre 2 présente les travaux existants sur la génération et la détection des deepfakes, la synchronisation

audiovisuelle et la détection d'anomalies, tout en mettant en évidence les tendances récentes. Le chapitre 3 décrit la méthodologie proposée, notamment l'extraction des caractéristiques, l'architecture du réseau, les stratégies d'entraînement et les protocoles d'évaluation. Le chapitre 4 présente les jeux de données utilisés dans nos expériences, les protocoles d'acquisition, les étapes de prétraitement nécessaires et les résultats expérimentaux. Enfin, le chapitre 5 résume les principales contributions, discute les limites de l'approche proposée et esquisse des pistes de recherche futures.

CHAPITRE 2

État de l'art

2.1 Introduction

Au cours des dernières années, la technologie des deepfakes a progressé à une vitesse préoccupante, réduisant l'écart entre fabrication et réalité et soulevant des enjeux majeurs en matière de sécurité, de vie privée et d'analyse forensique numérique [21]. Par *deepfake*, nous désignons des images, des vidéos ou des signaux audio synthétiques ou manipulés qui imitent de manière crédible des locuteurs et des identités réels [24]. Les premiers travaux se concentraient principalement sur des indices visuels dans de courts clips vidéo, alors que les techniques récentes sont de plus en plus multimodales et de haute qualité, ce qui rend les falsifications plus difficiles à détecter et plus faciles à exploiter dans des campagnes de désinformation. Plusieurs synthèses récentes présentent un panorama complet de la génération et de la détection des deepfakes selon les modalités image, vidéo et audio [25–31]. Ces travaux mettent en évidence les progrès spectaculaires réalisés par les modèles génératifs modernes, notamment les GANs et les modèles de diffusion, qui produisent désormais des contenus de plus en plus réalistes et difficiles à distinguer des données authentiques. Ils soulignent également le besoin croissant de méthodes forensiques robustes, généralisables et interprétables, capables de résister à l'évolution rapide des techniques de manipulation et aux diverses transformations subies par les contenus multimédias dans des environnements réels.

Dans ce chapitre, nous commençons par les deepfakes *fondés sur l'image*, puisqu'ils restent largement diffusés sur les réseaux sociaux et constituent la base de nombreuses falsifications vidéo. Nous passons d'abord en revue les fondements classiques et profonds de la vision par ordinateur qui soutiennent les détecteurs modernes de deepfakes (CNN, réseaux récurrents et modèles fondés sur les Transformers). Nous étendons ensuite la discussion aux deepfakes *vidéo* et *audio*, où les indices temporels et intermodaux deviennent centraux. Enfin, nous mettons en évidence le rôle d'outils d'interprétabilité tels que *Grad-CAM*, qui permettent de visualiser l'attention du modèle et de mieux comprendre quelles régions ou quels instants contribuent à la décision de détection.

2.2 Fondements de la vision par ordinateur et de l'apprentissage profond

Avant d'aborder les détecteurs spécialisés de deepfakes, nous rappelons brièvement les principales architectures d'apprentissage profond utilisées dans ce mémoire : les réseaux de neurones convolutifs (CNN) pour l'extraction spatiale de caractéristiques, les réseaux récurrents (LSTM/BiLSTM) pour la modélisation temporelle, ainsi que les modèles fondés sur les Transformers pour les interactions à longue portée et multimodales. L'objectif n'est pas de proposer un tutoriel exhaustif, mais de mettre en évidence les éléments réutilisés dans la chaîne de détection proposée.

2.2.1 Réseaux de neurones convolutifs (CNN)

Les CNN apprennent des caractéristiques visuelles hiérarchiques de bout en bout : les premières couches capturent des motifs locaux tels que les contours et les textures, tandis que les couches profondes encodent des structures plus abstraites comme des parties d'objets ou des visages [11, 32–34]. Un CNN typique comprend deux parties : un extracteur de caractéristiques composé de convolutions, de non-linéarités (p. ex. ReLU) et de couches de sous-échantillonnage, suivi de têtes propres à la tâche (p. ex. couches entièrement connectées pour la classification ou blocs de suréchantillonnage pour la segmentation).

Des architectures établies comme VGGNet [32], Inception [33], ResNet [11] et EfficientNet [35] demeurent de solides références pour la détection visuelle des deepfakes. Elles offrent différents compromis entre exactitude, rapidité et efficacité paramétrique, ce qui est important lorsque les modèles doivent fonctionner sur du contenu compressé provenant des réseaux sociaux ou sous des contraintes de temps réel.

Pour concrétiser cette discussion, la figure 2.1 illustre la chaîne de traitement canonique d'un CNN. Elle montre comment une image d'entrée est progressivement transformée par des filtres convolutifs, des activations non linéaires et des opérations de sous-échantillonnage avant d'être transmise à une tête de prédiction propre à la tâche.

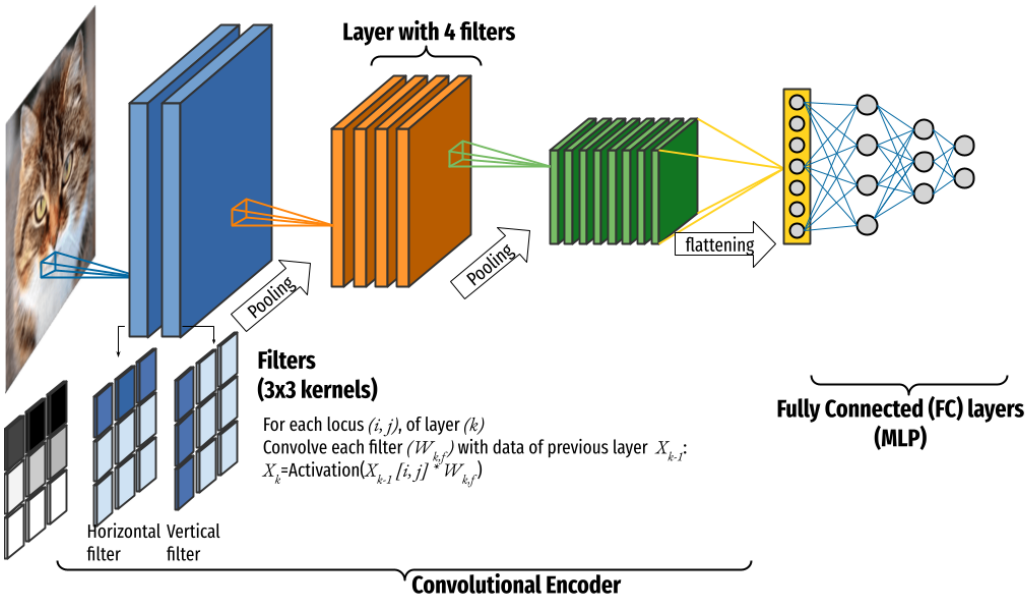


FIGURE 2.1 – Structure générale d'un réseau de neurones convolutif, montrant l'étape d'extraction des caractéristiques suivie de la tête de classification (d'après [2]).

Comme le montre la figure 2.1, les premières couches convolutives capturent généralement des informations de bas niveau, comme les contours et les textures, tandis que les couches profondes encodent des structures plus abstraites. Dans le contexte de la détection des deepfakes, cette hiérarchie est précieuse, car les manipulations peuvent laisser à la fois des artefacts texturaux locaux et des incohérences faciales de plus haut niveau.

Plus précisément, les CNN utilisés pour la détection des deepfakes fonctionnent souvent au niveau de l'image : chaque image faciale (ou région d'intérêt) est transmise à un réseau dorsal, puis les plongements obtenus sont soit classés directement (cadre fondé sur l'image), soit traités par des modules temporels (cadres vidéo et audiovisuels). Les CNN servent aussi de briques de base dans des architectures plus complexes combinant des indices spatiaux et temporels [25, 27].

2.2.2 Réseaux récurrents : LSTM et BiLSTM

Alors que les CNN conviennent bien à la modélisation de la structure spatiale, les vidéos deepfake et les clips audio sont intrinsèquement temporels. Les réseaux de neurones récurrents (RNN) constituent un choix naturel pour ce type de données, mais les RNN standards souffrent de gradients qui disparaissent ou explosent lorsqu'ils modélisent de longues séquences. Les réseaux à mémoire longue à court terme (LSTM) répondent à cette limite au moyen de mécanismes de portes qui contrôlent le flux d'information dans le temps, permettant au réseau de conserver les informations pertinentes sur des horizons plus longs [3].

Dans ce mémoire, les LSTM et leur variante bidirectionnelle (BiLSTM, aussi notée BLSTM [4]) sont utilisés pour agréger :

— des *caractéristiques visuelles image par image* extraites par un réseau dorsal CNN sous forme de descripteurs temporellement cohérents, capables de capter les micro-expressions, la dynamique des lèvres, les clignements et des indices de mouvement plus globaux.

— des *caractéristiques audio au niveau de la séquence* obtenues à partir de spectrogrammes Mel ou de plongements de parole auto-supervisés, afin de modéliser la prosodie, la coarticulation et les artefacts éventuels de vocodeur propres à la parole synthétique.

Cellule LSTM : Pour clarifier la manière dont l'information temporelle est préservée, les équations suivantes décrivent les opérations internes d'une cellule LSTM à chaque pas de temps. Étant donné l'entrée $\mathbf{x}_t \in \mathbb{R}^d$, l'état caché précédent $\mathbf{h}_{t-1} \in \mathbb{R}^h$ et l'état de cellule précédent $\mathbf{c}_{t-1} \in \mathbb{R}^h$, une cellule LSTM calcule :

Dans les équations suivantes, $\sigma(\cdot)$ désigne la fonction sigmoïde, $\tanh(\cdot)$ la fonction tangente hyperbolique et $\odot$ la multiplication élément par élément. Les matrices $W_\bullet$ et $U_\bullet$ contiennent les poids appris par le modèle, tandis que $b_\bullet$ désigne les vecteurs de biais. Les vecteurs i_t, f_t et o_t correspondent respectivement aux portes d'entrée, d'oubli et de sortie, alors que $\widetilde{c}_t$ représente l'état candidat de la cellule.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i) \quad \text{(porte d'entrée)} \tag{2.1}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f) \quad \text{(porte d'oubli)} \tag{2.2}$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c) \quad \text{(état candidat)} \tag{2.3}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad \text{(mise à jour de la cellule)} \tag{2.4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o) \quad \text{(porte de sortie)} \tag{2.5}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad \text{(sortie cachée).} \tag{2.6}$$

Pris ensemble, les équations (2.1)–(2.6) montrent que la porte d'oubli contrôle l'information supprimée, que la porte d'entrée détermine les nouvelles informations stockées et que la porte de sortie régule ce qui est exposé à l'état caché suivant. Ce mécanisme explique pourquoi les LSTM sont mieux adaptés que les RNN standards à la modélisation de longues dépendances temporelles dans les vidéos deepfake et les signaux de parole.

La figure 2.2 offre une vue schématique de ce mécanisme. Elle met en évidence l'interaction entre les portes et la cellule mémoire, ce qui rend la formulation mathématique plus facile à

interpréter.

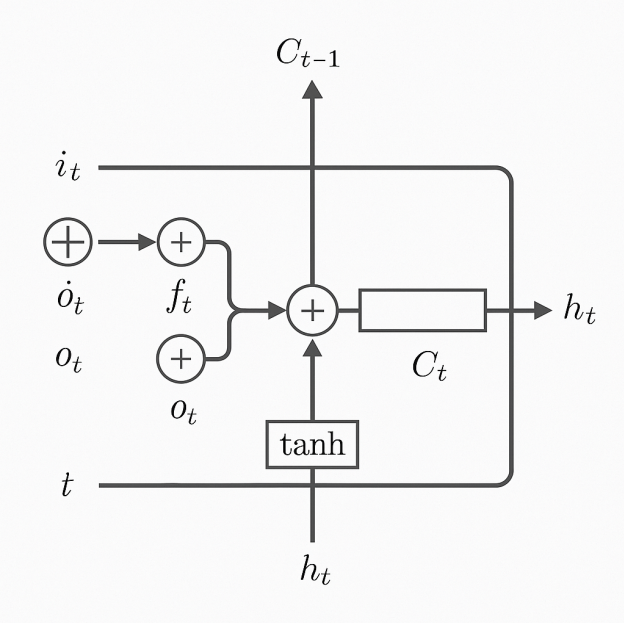


FIGURE 2.2 – Schéma d'une cellule LSTM avec les portes d'entrée ($\mathbf{i}_t$), d'oubli ($\mathbf{f}_t$) et de sortie ($\mathbf{o}_t$) contrôlant la cellule mémoire $\mathbf{c}_t$ (adapté de [3]).

Comme l'illustre la figure 2.2, l'état interne de la cellule agit comme une mémoire persistante capable de transporter l'information sur de nombreux pas de temps. Cette propriété est particulièrement importante en détection des deepfakes, où des indices comme les irrégularités du mouvement des lèvres ou les incohérences de synchronisation de la parole peuvent n'apparaître qu'à l'échelle d'une séquence plutôt que dans une seule image.

LSTM bidirectionnel (BiLSTM) : Un BiLSTM exécute deux LSTM en parallèle : l'un traite la séquence dans le sens direct et l'autre dans le sens inverse [4]. À chaque pas de temps, les états cachés sont combinés comme suit :

$$\overrightarrow{\mathbf{h}}_t = \mathrm{LSTM}_{\rightarrow}(\mathbf{x}_t), \qquad \overleftarrow{\mathbf{h}}_t = \mathrm{LSTM}_{\leftarrow}(\mathbf{x}_t), \qquad \mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right]. \tag{2.7}$$

Dans cette notation, $\overrightarrow{\mathbf{h}}_t$ représente l'état caché produit par le LSTM qui traite la séquence dans le sens chronologique, tandis que $\overleftarrow{\mathbf{h}}_t$ représente l'état caché produit dans le sens inverse. La notation $[\mathbf{a}; \mathbf{b}]$ désigne la concaténation des vecteurs $\mathbf{a}$ et $\mathbf{b}$. Ainsi, $\mathbf{h}_t$ contient à la fois l'information provenant du contexte passé et celle provenant du contexte futur.

La représentation obtenue $\mathbf{h}_t$ exploite à la fois le contexte passé et futur, ce qui améliore la sensibilité aux dérives inter-images, comme un désalignement lèvres-phonèmes s'étendant sur plusieurs images. Comme le passage inverse est non causal, les BiLSTM sont surtout adaptés

à l'analyse hors ligne ou à l'inférence par fenêtres avec un léger regard vers l'avant, tandis que les applications strictement en temps réel privilégient des LSTM causaux unidirectionnels.

La figure 2.3 présente le principe général de la modélisation bidirectionnelle des séquences. Elle clarifie la manière dont les états cachés direct et inverse sont combinés avant l'étape finale de prédiction.

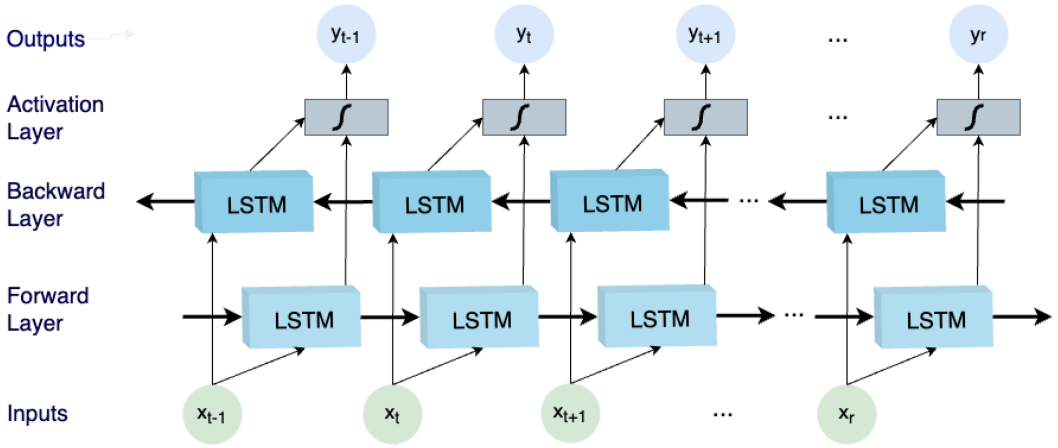


FIGURE 2.3 – Architecture BiLSTM : un LSTM direct et un LSTM inverse traitent la même séquence, puis leurs états cachés sont concaténés avant les couches de classification ou de fusion (adapté de [4]).

Dans les applications pratiques de détection des deepfakes, cette conception bidirectionnelle peut améliorer les performances lorsque la séquence complète est disponible, car les manipulations subtiles sont souvent plus faciles à détecter lorsque les contextes antérieur et postérieur sont pris en compte.

2.3 Transformers de vision et modèles fondés sur les Transformers

Les Transformers sont devenus un paradigme central de l'analyse forensique visuelle et audiovisuelle. Initialement proposés pour le traitement automatique des langues [36], ils reposent sur un mécanisme d'auto-attention qui modélise les interactions par paires entre des jetons (par exemple des patchs d'image, des tubes vidéo ou des trames audio). Contrairement aux CNN, qui agrègent l'information à travers des champs récepteurs locaux et du sous-échantillonnage, les Transformers offrent une manière plus flexible de capter des dépendances spatiales et temporelles à longue portée. Cette propriété est particulièrement intéressante pour la détection des deepfakes, où les artefacts peuvent être subtils, non locaux ou répartis entre plusieurs modalités.

2.3.1 Fondements de l'auto-attention

Soit $\mathbf{X} \in \mathbb{R}^{N \times d}$ une séquence de N jetons d'entrée, chacun de dimension d (par exemple des patchs d'image ou des représentations intermédiaires extraites par un réseau neuronal). Contrairement aux réseaux convolutifs qui exploitent principalement des voisinages locaux, le mécanisme d'auto-attention permet à chaque jeton d'interagir directement avec tous les autres jetons de la séquence. Cette propriété favorise la modélisation de dépendances à longue portée et la capture de relations globales entre différentes régions d'une image.

Pour formaliser ce mécanisme, le Transformer projette d'abord les jetons d'entrée dans trois espaces distincts correspondant aux requêtes (*queries*), aux clés (*keys*) et aux valeurs (*values*) :

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \qquad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \qquad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{2.8}$$

où $\mathbf{W}_Q$, $\mathbf{W}_K$ et $\mathbf{W}_V$ sont des matrices de projection apprises. Les matrices $\mathbf{Q}$, $\mathbf{K}$ et $\mathbf{V}$ représentent respectivement les requêtes, les clés et les valeurs. Les requêtes et les clés servent à mesurer la similarité entre les jetons, tandis que les valeurs contiennent l'information qui sera agrégée.

Les requêtes et les clés servent à mesurer la similarité entre les jetons, tandis que les valeurs contiennent l'information qui sera agrégée. L'attention à produit scalaire mis à l'échelle avec une seule tête est définie par :

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{2.9}$$

Dans cette équation, le symbole $\top$ désigne la transposition d'une matrice et d_k la dimension des vecteurs de clés et de requêtes. Le produit $\mathbf{Q}\mathbf{K}^\top$ produit une matrice de scores de similarité entre les jetons. La division par $\sqrt{d_k}$ limite l'amplitude de ces scores, tandis que la fonction softmax les transforme en poids d'attention normalisés. La multiplication finale par $\mathbf{V}$ produit, pour chaque jeton, une combinaison pondérée des informations disponibles.

Le terme $\mathbf{Q}\mathbf{K}^\top$ calcule une matrice de similarité entre tous les couples de jetons, tandis que la fonction *softmax* transforme ces scores en poids d'attention normalisés. Chaque jeton est ainsi mis à jour comme une combinaison pondérée de l'ensemble des autres jetons de la séquence. Cette capacité à agréger de l'information provenant de régions éloignées constitue l'un des principaux avantages des Transformers pour l'analyse d'images complexes et la détection d'artefacts subtils de deepfake, souvent dispersés dans différentes parties du visage.

En pratique, l'auto-attention multi-têtes (*Multi-Head Self-Attention*, MSA) applique plusieurs mécanismes d'attention en parallèle à l'aide de projections différentes. Chaque tête apprend ainsi des relations complémentaires, certaines pouvant se concentrer sur la

structure globale du visage tandis que d'autres modélisent des détails plus fins. Les sorties des différentes têtes sont ensuite concaténées et traitées par un perceptron multicouche (MLP). Des connexions résiduelles et des opérations de normalisation de couche (*Layer Normalization*) sont également utilisées afin d'améliorer la stabilité et la convergence de l'apprentissage [36].

Une limite importante de l'auto-attention standard réside toutefois dans sa complexité quadratique $\mathcal{O}(N^2)$ par rapport au nombre de jetons N. Lorsque la résolution des images augmente ou que plusieurs images vidéo doivent être analysées simultanément, le coût mémoire et computationnel peut devenir prohibitif. Pour répondre à cette problématique, plusieurs variantes efficaces ont été proposées, notamment l'attention locale, l'attention fenêtrée et les mécanismes de réduction progressive du nombre de jetons.

Afin de mieux illustrer cette architecture, la figure 2.4 présente la chaîne de traitement standard du Transformer de vision (ViT). L'image est d'abord divisée en patchs de taille fixe, puis chaque patch est converti en un vecteur de caractéristiques appelé *embedding*. Après l'ajout d'un codage positionnel, les jetons sont transmis à une série de blocs encodeurs composés de couches d'auto-attention et de propagation avant.

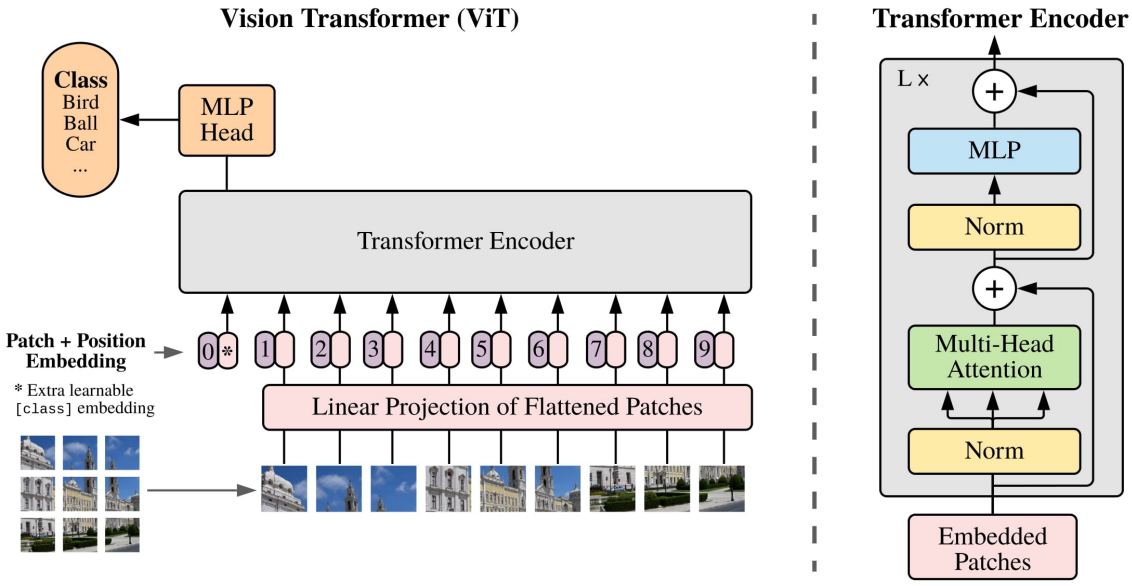


FIGURE 2.4 – Chaîne de traitement standard du Transformer de vision (ViT) et bloc encodeur. L'image est divisée en patchs, convertie en plongements de jetons, puis traitée par des couches empilées d'auto-attention et de propagation avant (d'après [5]).

Comme le montre la figure 2.4, l'auto-attention globale permet à chaque patch d'interagir avec l'ensemble des autres patchs de l'image. Cette capacité est particulièrement pertinente dans le contexte de la détection de deepfakes, où les incohérences générées artificiellement peuvent apparaître simultanément dans plusieurs régions du visage et nécessitent donc une compréhension globale de la scène plutôt qu'une simple analyse locale.

2.3.2 Transformers pour l'image

Transformer de vision (ViT) [5] : ViT divise une image en patchs de taille fixe (p. ex. 16×16), aplatit chaque patch et le projette dans un plongement de jeton. Un jeton de classe apprenable est ajouté, puis la séquence est traitée par des blocs encodeurs Transformer avec auto-attention globale. ViT s'est montré efficace sur plusieurs bancs d'essai de vision et demeure compétitif pour la détection des deepfakes lorsque des incohérences globales et non locales doivent être captées.

Transformers efficaces en données (DeiT) [37] : DeiT introduit des stratégies de distillation des connaissances et d'entraînement qui permettent aux modèles de type ViT de converger sur des jeux de données plus petits. Cela est précieux en analyse forensique, où la collecte de grands jeux de données annotés de deepfakes est coûteuse.

Swin Transformer [6] : Swin adopte une architecture hiérarchique dans laquelle les jetons sont progressivement fusionnés, formant une pyramide de caractéristiques proche de celle des CNN. L'auto-attention est limitée à des fenêtres locales décalées, ce qui conduit à une complexité linéaire par rapport à la taille de l'image et à une bonne latence. Cela rend Swin attractif pour des chaînes de détection des deepfakes proches du temps réel.

Auto-supervisé préentraînement (BEiT, MAE) : Des modèles tels que BEiT [38] et MAE [39] utilisent la modélisation d'images masquées pour le préentraînement auto-supervisé. Les représentations obtenues se transfèrent bien aux tâches en aval, notamment la détection de manipulations, et tendent à être plus robustes face à la compression, au réencodage et aux changements de domaine que les modèles purement supervisés.

La figure 2.5 complète la discussion précédente en montrant comment le Swin Transformer introduit un apprentissage hiérarchique des représentations tout en maintenant un coût computationnel maîtrisable. Au lieu d'appliquer une attention globale sur tous les jetons, il restreint l'attention à des fenêtres locales et décale ces fenêtres d'une couche à l'autre, ce qui préserve à la fois l'efficacité et les interactions entre fenêtres.

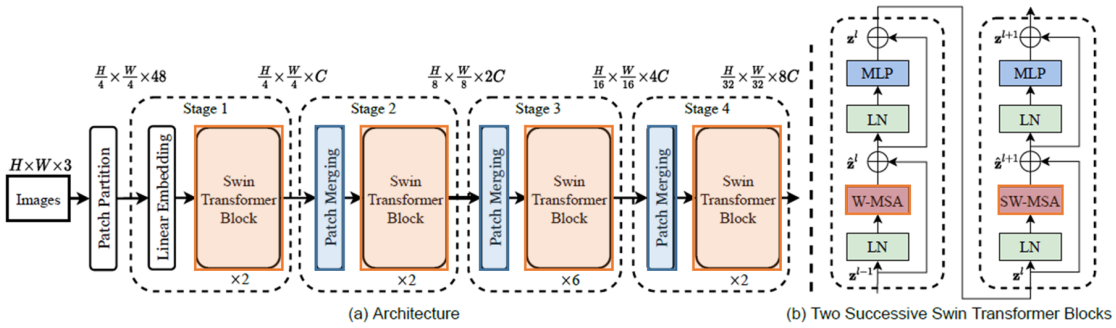


FIGURE 2.5 – Vue d'ensemble de l'architecture Swin Transformer. Le modèle construit une pyramide hiérarchique de caractéristiques au moyen du plongement de patchs, de la fusion de patchs et de l'auto-attention à fenêtres décalées, ce qui améliore l'efficacité sur des images à haute résolution (d'après [6]).

Cette conception hiérarchique explique pourquoi le Swin Transformer est souvent privilégié dans les contextes pratiques de détection des deepfakes, où la représentation multi-échelle et l'efficacité computationnelle sont toutes deux importantes.

2.3.3 Transformers pour la vidéo

Étendre les Transformers à la vidéo exige de modéliser à la fois les dimensions spatiale et temporelle. Les architectures récentes se distinguent surtout par la manière dont elles factorisent l'attention et contrôlent le coût computationnel.

TimeSformer [40] : factorise l'attention en composantes spatiale et temporelle séparées. Cela réduit la complexité par rapport à une attention spatio-temporelle complète, tout en captant des indices de mouvement tels que les taux de clignement, les mouvements de tête et les trajectoires des lèvres, essentiels à la détection des deepfakes.

ViViT [41] : ViViT étend ViT à la vidéo en opérant sur des tubes spatio-temporels ou sur d'autres schémas de tokenisation. Il offre plusieurs variantes qui échangent l'exactitude contre la vitesse, ce qui le rend adaptable à différentes contraintes de déploiement.

Transformers de vision multi-échelle (MViT) [42] et Video Swin [43] : Ces architectures construisent des hiérarchies pyramidales de jetons et appliquent l'attention dans des fenêtres spatio-temporelles locales. Elles offrent un débit élevé sur les GPU modernes et soutiennent l'analyse forensique fine au niveau de l'image ainsi que la notation d'anomalies.

2.3.4 Transformers audiovisuels

Au-delà de la vision seule, les travaux récents se sont intéressés aux Transformers audiovisuels (AV) conjoints qui exploitent la synchronie naturelle entre la parole et les mouvements des lèvres. Cette synchronie est fréquemment perturbée dans les deepfakes audiovisuels, ce qui en fait un indice fort de détection.

Fusion audiovisuelle de type CLIP : Inspirés par CLIP [44], les modèles de fusion audiovisuelle projettent les jetons visuels (p. ex. régions de la bouche ou du visage) et les jetons audio (p. ex. trames Mel ou plongements de parole auto-supervisés) dans un espace de plongement partagé. Des couches d'attention croisée alignent ensuite les formes des lèvres avec le minutage des phonèmes, tandis que des objectifs contrastifs rapprochent les paires audio-vidéo correspondantes et éloignent les paires discordantes. Ces architectures sont efficaces pour révéler les artefacts de désynchronisation caractéristiques de nombreux deepfakes.

AV-HuBERT et hybrides de lecture labiale : AV-HuBERT [45] et les modèles apparentés effectuent un préentraînement auto-supervisé sur de grands corpus audiovisuels, apprenant des représentations de la parole qui encodent conjointement l'acoustique et les mouvements des lèvres. L'ajustement fin de ces réseaux dorsaux pour la détection des deepfakes produit des plongements sensibles à la synchronie et plus robustes face à des générateurs inconnus et à des dégradations de canal.

Têtes propres aux tâches : Au-dessus d'un réseau dorsal Transformer audiovisuel partagé, des têtes légères propres aux tâches peuvent être ajoutées pour prendre en charge :

- la classification des deepfakes au niveau du clip ou du segment,
- la localisation de falsifications image par image,
- et, optionnellement, des scores d'authenticité propres à chaque modalité (audio seul, vidéo seule ou audiovisuel conjoint).

Cette conception modulaire permet au même réseau dorsal de servir plusieurs tâches forensiques tout en maintenant l'inférence et l'entraînement à un coût raisonnable.

Pour résumer les principales familles de Transformers discutées ci-dessus, le tableau 2.1 compare leurs domaines, leurs idées clés et leurs usages pratiques en détection des deepfakes ou d'anomalies.

TABLEAU 2.1 – Familles de Transformers utilisées pour la détection des deepfakes et des anomalies. « Domaine » indique image (I), vidéo (V) ou audiovisuel (AV).

Modèle (année)	**Domaine**	**Idée clé**	**Contrôle de la complexité**	**Usage typique**
ViT (2021)	I	Auto-attention globale sur les jetons de patchs	Nombre de jetons (taille des patchs), distillation	Indices de falsification d'images
DeiT (2021)	I	ViT efficace en données grâce à la distillation	Recette d'entraînement, petits réseaux dorsaux	Contextes avec peu de données
Swin (2021)	I	Attention hiérarchique à fenêtres décalées	Fenêtres locales, pyramide	Réseaux visuels compatibles avec le temps réel
TimeSformer (2021)	V	Attention factorisée espace/temps	Attention séparable	Artefacts temporels (clignement, lèvres)
ViViT (2021)	V	Extensions de ViT à la vidéo (plusieurs variantes)	Choix de tokenisation	Compromis exactitude/vitesse
MViT / Video Swin (2021–2022)	V	Espace-temps multi-échelle/fenêtré	Pyramides et fenêtres	Localisation rapide en vidéo
CLIP-style AV (depuis 2021)	AV	Plongement partagé et contraste audiovisuel	Projection de jetons et attention croisée	Synchronisation audiovisuelle et discordance modale
AV-HuBERT (depuis 2022)	AV	Apprentissage auto-supervisé sur l'audio et la vidéo labiale	Tâches audiovisuelles masquées	Plongements audiovisuels robustes

2.4 Explicabilité et saillance visuelle

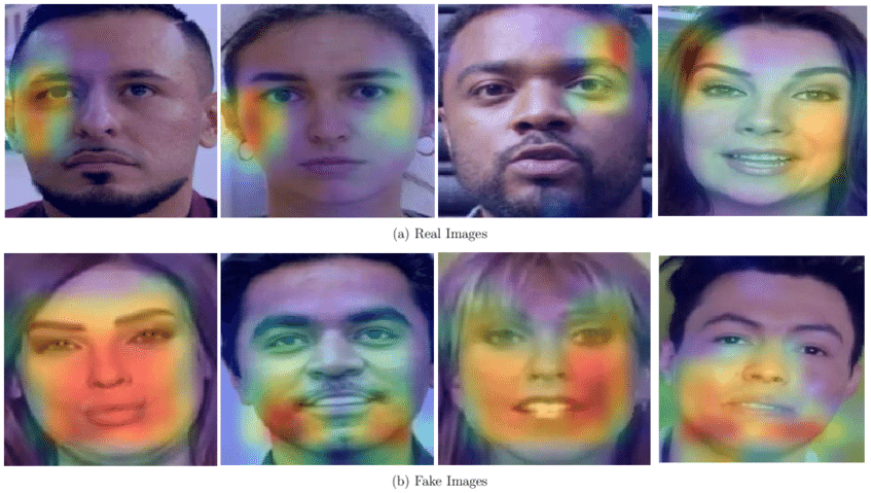
Les méthodes d'explicabilité comme Grad-CAM aident à identifier les régions de l'image qui influencent le plus la décision d'un modèle. En détection des deepfakes, cela est particulièrement utile, car les contenus manipulés contiennent souvent des artefacts subtils difficiles à reconnaître par une simple inspection visuelle.

Pour mieux comprendre la décision du détecteur, la figure 2.6 présente des visualisations Grad-CAM représentatives pour des images de visages réelles et synthétiques. La rangée supérieure correspond aux échantillons réels, tandis que la rangée inférieure montre les échantillons synthétiques. Les couleurs plus chaudes indiquent les régions qui contribuent le plus fortement à la prédiction du modèle.

La comparaison entre les deux rangées suggère que les images synthétiques tendent à produire des activations plus fortes et moins régulières sur des régions faciales sensibles, en

particulier autour de la bouche, des joues, du nez et des yeux. Ces motifs peuvent refléter des artefacts de fusion, des incohérences de texture ou des erreurs de synthèse introduites lors de la manipulation. À l'inverse, les activations observées dans les images réelles paraissent plus stables et anatomiquement cohérentes.

Ces visualisations sont précieuses dans un contexte forensique, car elles fournissent plus qu'une simple prédiction. Elles indiquent aussi les régions faciales qui soutiennent la décision du modèle, ce qui rend le processus de détection plus interprétable et renforce la crédibilité du résultat.



(a) Real Images

(b) Fake Images

FIGURE 2.6 – Visualisations Grad-CAM pour des images de visages réelles et synthétiques. adapté de [7]).

2.5 Génération de deepfakes : mécanismes et techniques

La génération de deepfakes repose sur quelques grandes familles de modèles génératifs — autoencodeurs, GAN et, plus récemment, modèles de diffusion — qui ont progressivement fait évoluer les médias synthétiques d'images floues et rudimentaires vers des contenus photoréalistes. Dans cette section, nous rappelons brièvement les principes de chaque famille, en soulignant les propriétés les plus importantes pour la détection en aval.

2.5.1 Autoencodeurs et variantes

Les autoencodeurs sont des architectures composées d'un encodeur et d'un décodeur. L'encodeur transforme l'entrée en une représentation latente compacte, tandis que le décodeur cherche à reconstruire les données originales à partir de cette représentation. L'ensemble du modèle est entraîné en minimisant une fonction de perte de reconstruction, comme l'erreur

quadratique moyenne (*Mean Squared Error*, MSE) [46].

Dans le contexte de la manipulation faciale, des autoencodeurs appariés, également appelés autoencodeurs doubles, peuvent être utilisés pour modéliser un visage source et un visage cible. En partageant une représentation latente commune tout en utilisant des décodeurs propres à chaque identité, ils permettent de transférer les expressions et les mouvements d'un visage vers un autre [47].

Les autoencodeurs variationnels (VAE) ajoutent une modélisation probabiliste de l'espace latent. Cette contrainte favorise un espace de représentation plus régulier et permet de générer des variations faciales plausibles par échantillonnage, plutôt que de simplement mémoriser les images utilisées pendant l'entraînement [48].

2.5.2 Réseaux antagonistes génératifs (GAN)

Les réseaux antagonistes génératifs (GAN) opposent un générateur et un discriminateur dans un jeu minimax, ce qui produit des détails nets et de haute fréquence que les autoencodeurs classiques reproduisaient difficilement [14]. Les variantes modernes, notamment ProgressiveGAN, StyleGAN, BigGAN et CycleGAN, ont rapproché le photoréalisme de la perception humaine en améliorant les textures, l'éclairage et la cohérence des expressions [49–52]. Pour la falsification faciale, les chaînes fondées sur les GAN synthétisent ou remplacent généralement des régions du visage avec une grande fidélité tout en cherchant à supprimer des empreintes forensiques évidentes, comme les coutures de fusion ou les motifs de bruit de bas niveau.

2.5.3 Modèles de diffusion

Les modèles de diffusion inversent un processus d'ajout progressif de bruit en apprenant une séquence d'étapes de débruitage allant d'un bruit pur vers un échantillon propre [53]. Les approches fondées sur le score et les équations différentielles stochastiques (SDE) raffinent cette perspective, en permettant des trajectoires d'échantillonnage flexibles et une bonne couverture des modes [54]. Les modèles de diffusion latents déplacent le calcul dans un espace latent appris, préservant des détails fins (pores de la peau, cheveux, reflets dans les yeux) tout en réduisant les coûts mémoire et calcul par rapport à un traitement direct dans l'espace des pixels [55]. Leur échantillonnage itératif demeure coûteux, de sorte que les générateurs deepfake pratiques s'appuient souvent sur des variantes accélérées (p. ex. échantillonnage de type DDIM, distillation) pour réduire le nombre d'étapes sans perdre trop de fidélité.

2.6 Approches de détection

2.6.1 Analyse forensique classique des images

Avant l'ère de l'apprentissage profond, l'analyse forensique des images se concentrait sur des incohérences physiques et statistiques : direction de l'éclairage, ombres, fusion de contours, quantification JPEG, traces de double compression ou bruit propre au capteur [56–58]. Ces méthodes sont efficaces et interprétables, mais leur robustesse diminue à mesure que les modèles génératifs gagnent en réalisme et que le post-traitement (filtrage, réencodage, redimensionnement) détruit de nombreux indices de bas niveau.

2.6.2 Apprentissage profond : modalités visuelles seulement

Les détecteurs modernes de deepfakes ajustent généralement des CNN ou des Transformers sur de grands jeux de données de manipulations faciales comme FaceForensics++[8]. Les détecteurs fondés sur les CNN et les ViT apprennent des artefacts subtils : résidus de haute fréquence, statistiques de style, frontières de fusion, anomalies de clignement des yeux et incohérences temporelles entre images [59, 60]. Les architectures hybrides combinent des réseaux dorsaux CNN 2D avec des modules temporels (RNN, TCN, ConvLSTM), tandis que les CNN 3D et les Transformers vidéo modélisent directement les séquences.

2.6.3 Anti-usurpation fondée uniquement sur l'audio

Les détecteurs audio seuls visent les indices acoustiques de parole synthétique ou convertie : artefacts de vocodeur, manque de micro-prosodie, motifs spectraux anormaux et incohérences de phase. Les méthodes récentes utilisent des représentations de parole auto-supervisées, des spectrogrammes log-Mel ou des formes d'onde brutes afin de distinguer la parole naturelle de la parole générée.

2.6.4 Fusion audiovisuelle (AV)

La fusion audiovisuelle exploite l'idée qu'un visage parlant naturel doit présenter une cohérence entre le signal acoustique et l'articulation visuelle. Les approches de fusion comparent les mouvements des lèvres avec le contenu phonétique, agrègent des caractéristiques audio et visuelles ou utilisent l'attention croisée pour repérer les désalignements. Cette famille de méthodes est particulièrement pertinente pour les deepfakes de type tête parlante, où une modalité peut paraître crédible isolément tout en étant incohérente avec l'autre.

2.6.5 Considérations liées au temps réel

Les contraintes de temps réel constituent un enjeu majeur pour les systèmes modernes de détection de deepfakes. Contrairement aux scénarios d'analyse hors ligne, où plusieurs secondes ou minutes de calcul peuvent être tolérées, les applications de modération en direct, de visioconférence, de diffusion en continu ou de surveillance numérique exigent des décisions rapides avec une latence minimale. Dans ce contexte, les architectures doivent maintenir un équilibre entre précision, coût computationnel, consommation mémoire et vitesse d'inférence. Les modèles très profonds ou reposant sur des mécanismes d'attention globale appliqués à de longues séquences peuvent atteindre d'excellentes performances expérimentales, mais leur complexité limite souvent leur déploiement pratique [5, 6].

Plusieurs travaux ont ainsi souligné l'importance de concevoir des détecteurs efficaces capables de fonctionner sur des ressources limitées tout en conservant un niveau élevé de fiabilité [21]. Les contraintes de temps réel imposent généralement de réduire la profondeur des réseaux, de limiter la résolution des images traitées, de raccourcir les fenêtres temporelles utilisées pour l'analyse vidéo ou encore d'employer des mécanismes d'attention locale plutôt que globale. Dans les architectures vidéo, le traitement de séquences plus courtes permet de diminuer significativement le coût mémoire et le temps d'inférence, au prix d'une réduction potentielle du contexte temporel disponible pour la détection.

2.7 Détection d'anomalies vidéo (VAD)

2.7.1 Formulation du problème et supervision

Du point de vue de la modélisation, la détection des deepfakes peut, dans certains contextes, être rapprochée de la détection d'anomalies vidéo (*Video Anomaly Detection*, VAD). Les segments manipulés sont souvent localisés dans le temps et peuvent présenter des caractéristiques différentes de celles observées dans les séquences authentiques.

Cette relation explique l'intérêt porté aux méthodes non supervisées et faiblement supervisées, qui peuvent limiter le besoin de disposer d'un grand nombre d'exemples falsifiés précisément annotés. Dans les approches fondées sur les autoencodeurs, le modèle est entraîné à reconstruire des séquences considérées comme normales. Lorsqu'une région ou un segment diffère fortement des données authentiques observées pendant l'entraînement, sa reconstruction devient généralement moins précise. Une erreur de reconstruction élevée peut alors être utilisée comme indice d'anomalie [61].

Les approches d'apprentissage multi-instances (*Multiple Instance Learning*, MIL) utilisent,

quant à elles, des étiquettes globales attribuées aux vidéos afin d'identifier les segments les plus susceptibles de contenir une anomalie, sans exiger une annotation temporelle précise pour chaque événement [62].

Plus récemment, les Transformers, l'apprentissage contrastif et les modèles prédictifs ont permis de mieux représenter les dépendances temporelles complexes et de localiser plus précisément les événements anormaux. Ces avancées sont pertinentes pour la détection des deepfakes, car les manipulations peuvent être subtiles, intermittentes et limitées à quelques images ou à de courts intervalles temporels. Dans ce cas, leur détection présente certaines similitudes avec le repérage d'anomalies rares dans de longues séquences vidéo.

2.7.2 Anomalies audiovisuelles

Les indices audio, tels que coups de feu, alarmes, cris ou silences soudains, permettent souvent de lever l'ambiguïté de scènes visuellement incertaines affectées par l'occlusion, le flou de mouvement ou un faible éclairage. Les travaux récents de localisation audiovisuelle d'événements montrent que la modélisation conjointe de l'audio et de la vidéo réduit les faux positifs et produit des frontières temporelles plus nettes pour les événements anormaux [63]. Dans notre contexte, le même principe s'applique : les anomalies sont détectées plus fiablement lorsque les deux modalités s'écartent des motifs normaux, ou lorsque leur synchronie se rompt, que lorsque l'on s'appuie sur une seule modalité.

2.8 Jeux de données et bancs d'essai

Nous résumons dans cette section les principaux jeux de données utilisés pour l'évaluation des méthodes de détection de deepfakes (visuels et audiovisuels) ainsi que de détection d'anomalies vidéo (Video Anomaly Detection, VAD). Ces corpus couvrent des scénarios variés allant des manipulations faciales synthétiques aux événements anormaux observés dans des séquences de surveillance réelles. Ils diffèrent par leur taille, leurs modalités (image, vidéo, audio ou audiovisuelle), leur niveau d'annotation et la complexité des situations représentées. Le tableau **??** présente un aperçu des ensembles de données les plus fréquemment utilisés dans la littérature récente, en mettant en évidence leurs principales caractéristiques ainsi que les protocoles d'évaluation associés. Ces jeux de données constituent aujourd'hui des références essentielles pour mesurer la robustesse, la capacité de généralisation et les performances en conditions réelles des systèmes de détection.

Les jeux de données présentés dans le tableau 2.2 couvrent deux tâches principales : la détection des deepfakes et la détection d'anomalies vidéo. La colonne **Modalité** précise le

type de données disponible dans chaque corpus. La lettre **V** désigne la modalité vidéo, tandis que **AV** indique que le jeu de données contient à la fois des informations audio et vidéo.

TABLEAU 2.2 – Jeux de données représentatifs pour la détection des deepfakes visuels et audiovisuels, ainsi que pour la détection d'anomalies vidéo.

Jeu de données	**Année**	**Modalité**	**Portée et caractéristiques principales**	**Usage typique**
FaceForensics++ (FF++) [8]	2019	V	Environ 1 000 vidéos réelles et 4 000 vidéos synthétiques ; plusieurs méthodes de manipulation et différents niveaux de compression.	Évaluation de référence et étude de la robustesse à la compression.
DFDC [1]	2020	V/AV	Plus de 100 000 vidéos présentant des acteurs, des identités et des contextes variés ; corpus associé à un défi public.	Évaluation comparative à grande échelle.
Celeb-DF [9]	2020	V	Vidéos falsifiées de célébrités présentant un niveau de réalisme visuel élevé.	Évaluation de la généralisation entre différents jeux de données.
DeeperForensics-1.0 [64]	2020	V	Variations réalistes de pose, d'éclairage, d'expression et de compression.	Évaluation de la robustesse dans des conditions proches du réel.
FakeAVCeleb [65]	2021	AV	Falsifications audio seules, vidéo seules et audiovisuelles.	Étude de la synchronisation audiovisuelle et des stratégies de fusion.
AV-deepfake1M [66]	2023	AV	Environ un million de clips, avec des annotations temporelles et plusieurs formes de manipulation.	Détection et localisation temporelle des manipulations audiovisuelles.
UCF-Crime [67]	2018	V	Longues vidéos de surveillance non segmentées, annotations faibles et nombreuses catégories d'anomalies.	Détection d'anomalies vidéo avec supervision faible.
XD-Violence [68]	2020	AV	Longues vidéos non segmentées contenant une piste audio et des annotations liées à des événements violents.	Détection audiovisuelle d'anomalies et évaluation au niveau des événements.
ShanghaiTech Campus [69]	2017	V	Séquences de surveillance provenant de plusieurs scènes et caméras, avec différents types d'événements anormaux.	Évaluation de la détection d'anomalies au niveau de l'image.

Le tableau 2.2 montre que les corpus de deepfakes visuels, tels que FaceForensics++, Celeb-DF et DeeperForensics-1.0, sont principalement utilisés pour évaluer la détection d'artefacts faciaux et la robustesse aux variations de qualité. Les jeux de données audiovisuels, comme FakeAVCeleb et AV-deepfake1M, permettent en plus d'étudier la cohérence entre la parole et les mouvements du visage, ainsi que la localisation temporelle des segments manipulés.

Les corpus UCF-Crime, XD-Violence et ShanghaiTech Campus relèvent davantage de la détection d'anomalies dans des vidéos de surveillance. Ils sont inclus afin de montrer les similitudes méthodologiques entre le repérage de manipulations localisées et la détection d'événements rares dans de longues séquences. Toutefois, ces corpus ne constituent pas des jeux de données de deepfakes à proprement parler.

2.9 Métriques d'évaluation

Une évaluation rigoureuse des systèmes de détection des deepfakes doit aller au-delà d'un seul score agrégé. Dans un contexte forensique réaliste, il ne suffit pas de savoir si un clip est correctement classé : il faut aussi vérifier si le détecteur reste fiable en présence d'un déséquilibre des classes, s'il peut localiser les intervalles manipulés et s'il peut fonctionner sous des contraintes pratiques de latence. Pour cette raison, nous rapportons un ensemble de métriques complémentaires couvrant la qualité de classification, la qualité du classement, la localisation temporelle et l'efficacité de déploiement.

Le tableau 2.3 résume les métriques retenues dans ce mémoire. L'exactitude demeure utile comme indicateur global initial, mais elle peut être trop optimiste sur des jeux de données déséquilibrés ; elle doit donc être interprétée avec la précision, le rappel et le score F1. L'AUC-ROC est particulièrement importante, car elle évalue la qualité du classement indépendamment d'un seuil fixe. Pour l'anti-usurpation audio, le taux d'erreur égal (EER) demeure une mesure synthétique standard, puisqu'il correspond au point de fonctionnement où les fausses acceptations et les faux rejets sont égaux. Lorsque la tâche consiste à identifier les intervalles falsifiés plutôt qu'à attribuer une seule étiquette à tout le clip, les métriques de localisation temporelle, telles que l'IoU temporelle et la précision moyenne à des seuils d'IoU donnés, deviennent essentielles. Enfin, comme l'objectif ultime de ce mémoire est la détection audiovisuelle proche du temps réel, le débit et la latence doivent être rapportés avec les performances de détection.

Autrement dit, le choix de la métrique doit refléter le cas d'utilisation visé. Un modèle peut obtenir une forte AUC au niveau du clip tout en ayant une valeur forensique limitée s'il n'indique pas *où* la manipulation se produit ou si son coût d'inférence empêche un déploiement

pratique. Le protocole d'évaluation adopté ici est donc aligné avec la formulation du problème de ce mémoire, qui combine détection, localisation et faisabilité en temps réel.

Dans le tableau 2.3, TP désigne le nombre de vrais positifs, TN le nombre de vrais négatifs, FP le nombre de faux positifs et FN le nombre de faux négatifs. Le taux de fausse acceptation est noté FAR, tandis que le taux de faux rejet est noté FRR. L'IoU, ou intersection sur union, mesure le recouvrement entre un intervalle manipulé prédit et l'intervalle de vérité terrain.

TABLEAU 2.3 – Métriques courantes pour la détection des deepfakes (visuelle/AV), l'anti-usurpation audio et la localisation temporelle.

Métrique	**Définition**	**Interprétation**
Exactitude	$\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$	Justesse globale, mais potentiellement trompeuse en cas de déséquilibre des classes
Précision	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$	Fiabilité des prédictions positives ; importante lorsque les fausses alertes sont coûteuses
Rappel (TPR)	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	Capacité à retrouver les échantillons falsifiés ; importante lorsque les omissions sont coûteuses
Score F1	$2 \cdot \frac{\mathrm{Prec \cdot Rec}}{\mathrm{Prec+Rec}}$	Résumé équilibré de la précision et du rappel
AUC-ROC	Aire sous la courbe ROC	Mesure indépendante du seuil de la séparabilité et de la qualité du classement
EER	Point de fonctionnement où FAR = FRR	Mesure standard pour l'anti-usurpation audio et la détection de parole synthétique
AP / mAP à l'IoU	Précision moyenne sous contraintes de recouvrement temporel	Mesure la qualité de localisation des segments manipulés
IoU temporelle	IoU entre intervalles falsifiés prédits et réels	Mesure directe de l'exactitude de localisation dans le temps
Latence / débit	ms par fenêtre ou clips par seconde	Indique si la méthode convient au temps réel ou au quasi-temps réel
Robustesse aux perturbations	Performance sous compression, bruit, occlusion ou réencodage	Reflète la fiabilité du déploiement en conditions réalistes

2.10 Synthèse des méthodes représentatives

Cette section ne se limite pas à énumérer des détecteurs représentatifs ; elle les utilise pour clarifier l'évolution du domaine et situer la contribution de ce mémoire. La comparaison porte en particulier sur quatre critères récurrents : la généralisation à des manipulations inconnues, la sensibilité aux artefacts locaux et temporels, la capacité à exploiter les indices intermodaux et l'adéquation à une inférence en temps réel ou quasi temps réel.

Détecteurs visuels de deepfakes

Le tableau 2.4 résume des détecteurs visuels représentatifs, en mettant l'accent sur des méthodes récentes qui abordent explicitement la généralisation, la localisation ou le raisonnement temporel. L'évolution est instructive : les premiers détecteurs étaient souvent efficaces parce que les images et vidéos synthétiques contenaient des traces de fusion ou de compression visibles. À l'inverse, les méthodes récentes supposent de plus en plus que les manipulations sont visuellement subtiles et qu'une détection robuste exige une localisation fine, une modélisation temporelle plus forte ou des représentations transférables.

TABLEAU 2.4 – Détecteurs **visuels** représentatifs de deepfakes. « Loc. » indique la prise en charge de la localisation temporelle ou au niveau des pixels ; « Gen. » renvoie à la robustesse ou à la généralisation inter-domaines.

Méthode (année)	**Type**	**Idée principale**	**Jeux de données**	**Notes (Loc./Gen.)**
LipForensics (2021) [19]	V	Plongements centrés sur la bouche et indices temporels d'articulation labiale	FF++, Celeb-DF	A priori fort sur les visages parlants ; région faciale restreinte
LAA-Net (2024) [70]	V	Attention explicite aux régions d'artefacts vulnérables avec supervision fine	Celeb-DF, DFD, DFDC,[1, 9]) Wilddeepfake	Forte généralisation inter-jeux de données ; localisation centrée sur les artefacts
Style latent flows / StyleGRU (2024) [71]	V	Modélise la variation temporelle des vecteurs latents de style afin de détecter des comportements temporels propres aux générateurs	FF++, Celeb-DF, DFDC [1, 9])	Bon comportement inter-manipulations ; met l'accent sur les dynamiques temporelles
Classement multi-étiquettes pour classification + localisation (2024) [72]	V	Objectif unifié pour la classification au niveau du clip et la localisation des régions falsifiées	FF++, Celeb-DF, DFDC [1, 9]), Wilddeepfake	Localisation explicite ; forte interprétabilité forensique
Fusion vidéo Plug-and-Play + adaptateurs spatio-temporels (2025) [73]	V	Adaptateurs spatio-temporels légers insérés dans des détecteurs existants	FF++, Celeb-DF, DFDC [1, 9])	Améliore la généralisation au niveau vidéo avec un coût supplémentaire modéré
Apprentissage spatio-temporel sensible aux vulnérabilités (2025) [74]	V	Apprentissage multi-branches adapté aux vulnérabilités spatiales et temporelles subtiles	FF++[8], Celeb-DF, DFDC [1, 9])	Vue fine de la détection ; robustesse accrue aux manipulations inconnues
Détecteur de fréquence temporelle au niveau du pixel (2025) [75]	V	Analyse de Fourier temporelle 1D par pixel + fusion Transformer des indices temporels et contextuels	FF++, KoDF, cross-jeu de données évaluation	Particulièrement fort sur les artefacts temporels et les types de synthèse inconnus

Plusieurs observations découlent du tableau 2.4. Premièrement, le domaine s'est clairement éloigné de la classification binaire purement image par image vers des formulations plus structurées où la localisation et la cohérence temporelle sont traitées explicitement. Des méthodes comme LAA-Net et le cadre de classement multi-étiquettes cherchent à identifier les zones où les preuves forensiques se concentrent, plutôt que de s'appuyer uniquement sur des descripteurs faciaux globaux. Cette tendance est importante du point de vue forensique, car elle soutient un processus de décision plus interprétable.

Deuxièmement, les détecteurs visuels récents ciblent de plus en plus la *généralisation*, et non seulement l'exactitude dans le même domaine. L'approche temporelle fondée sur le style de Choi et al. [71], la stratégie d'apprentissage sensible aux vulnérabilités de Nguyen et al. [74] et la stratégie d'adaptateurs légers de Yan et al. [73] reflètent la même préoccupation : les deepfakes modernes paraissent souvent visuellement plausibles, de sorte qu'un détecteur surajusté à une famille étroite d'artefacts risque d'être peu fiable en pratique.

Troisièmement, les méthodes récentes les plus fortes montrent aussi que la modélisation temporelle est devenue indispensable. Cela se voit particulièrement dans l'analyse de fréquence temporelle au niveau du pixel [75], où la cible n'est plus seulement l'apparence spatiale d'une image, mais l'évolution subtile des preuves visuelles dans le temps. Ce déplacement est directement pertinent pour ce mémoire, puisque la branche visuelle proposée suppose également que la région de la bouche doit être analysée comme un signal spatio-temporel plutôt que comme un ensemble d'images isolées.

Ces approches présentent toutefois encore des limites. Une modélisation temporelle plus forte augmente souvent le coût computationnel, tandis que la généralisation demeure difficile lorsque les manipulations de test diffèrent substantiellement des procédés de synthèse vus à l'entraînement. De plus, les détecteurs plein visage répartissent leur capacité de représentation sur l'ensemble du visage, ce qui peut être sous-optimal lorsque les indices les plus discriminants se concentrent dans des régions liées à la parole, comme les lèvres, les dents et l'intérieur de la bouche.

Détecteurs audio et audiovisuels de deepfakes

Les méthodes audio et audiovisuelles sont résumées dans le tableau 2.5. Ici encore, la littérature récente révèle un déplacement méthodologique clair. Les systèmes uniquement audio demeurent très compétitifs pour la détection de la parole synthétique, surtout sur les bancs d'essai d'anti-usurpation, mais ils ne peuvent pas vérifier si le contenu parlé est cohérent avec l'articulation visible. À l'inverse, les méthodes AV fondées sur la synchronie peuvent révéler des discordances entre modalités, mais elles peuvent être moins efficaces lorsque la

chaîne de manipulation préserve une synchronisation labiale apparente. Les détecteurs AV récents cherchent donc à apprendre des représentations intermodales plus riches plutôt qu'à s'appuyer uniquement sur des indices de synchronie.

TABLEAU 2.5 – Détecteurs **audio** et **audiovisuels** (AV) représentatifs. « Sync » désigne la modélisation explicite de la synchronie ou l'apprentissage de la cohérence intermodale.

Méthode (année)	**Type**	**Idée principale**	**Jeux de données**	**Notes (Sync)**
AASIST / RawNet2 / RawFormer (2021 2022) [76–78]	A	Modélisation de la forme d'onde brute ou spectro-temporelle pour les artefacts d'usurpation	ASVspoof 2019 / 2021	Références audio fortes ; aucune preuve visuelle
Temporel-channel MHSA for parole synthétique détection (2024) [79]	A	Auto-attention multi-têtes adaptée aux interactions temporelles et fréquentielles de la parole usurpée	ASVspoof 2021	Détecteur moderne compétitif uniquement audio ; toujours unimodal
Antideepfake post-entraînement (2025) [80]	A	Post-entraîne des modèles fondamentaux de parole SSL sur des données multilingues diverses riches en artefacts	ASVspoof 5, deepfake-Eval-2024, multi-corpus setup	Forte robustesse et transfert ; coût d'entraînement élevé
SyncNet / AV-sync scoring (2016) [81]	AV	Apprend des plongements d'alignement audiovisuel entre la parole et le mouvement des lèvres	LRS2, LRS3, LRW	Référence de base fondatrice pour la synchronie
AVFF (2024) [82]	AV	Apprentissage de représentations AV auto-supervisé avec masquage complémentaire et fusion de caractéristiques	FakeAVCeleb, DFDC[8]	Apprentissage intermodal fort ; pas principalement une méthode de localisation
MMMS-BA (2024) [83]	AV	Attention croisée contextuelle sur des caractéristiques multi-séquences multimodales pour la détection et la	AV-deepfake1M, FakeAVCeleb, LAV-DF, TVIL	Détection/localisation conjointe ; bon comportement inter-jeux de données

La comparaison du tableau 2.5 met en évidence trois points. Premièrement, les détecteurs uniquement audio sont indispensables lorsque l'attaque affecte principalement le signal de parole. Les systèmes construits autour d'AASIST, d'architectures de type RawNet ou de stratégies récentes fondées sur SSL demeurent des références solides, car ils capturent les traces de vocodeur, les anomalies spectrales et les artefacts résiduels de génération qui peuvent ne pas être visibles sur le visage. Toutefois, leur portée probante est fondamentalement limitée : un signal naturel à l'écoute peut tout de même être incohérent avec le flux visuel.

Deuxièmement, la synchronie audiovisuelle est un indice puissant, mais insuffisant à elle seule. Des méthodes fondatrices comme SyncNet demeurent conceptuellement importantes, car elles rendent le problème opérationnel : un visage parlant doit présenter un alignement mesurable entre les visèmes et le contenu phonétique. Toutefois, les deepfakes modernes cherchent de plus en plus à préserver cette cohérence apparente. Les détecteurs récents dépassent donc le simple score de synchronie et modélisent plutôt des représentations conjointes plus riches, comme dans AVFF [82], MMMS-BA [83] et le cadre de préentraînement fondé sur la synchronisation de Anshul et al. [84].

Troisièmement, la localisation est devenue un critère décisif. Pour les manipulations pilotées par le contenu, surtout celles qui impliquent insertion, suppression ou remplacement de mots, la question forensique n'est souvent pas seulement de savoir *si* une vidéo est manipulée, mais *quel segment* l'est. Des méthodes comme MMMS-BA et le cadre de synchronisation ICCV 2025 répondent à cette exigence plus directement que les premiers modèles de classification AV. Cette évolution est fortement cohérente avec ce mémoire, dont le cadre cible n'est pas seulement l'authenticité au niveau du clip, mais la détection audiovisuelle en temps réel avec segments suspects localisés dans le temps.

Malgré cela, les méthodes AV récentes demeurent plus exigeantes que les détecteurs unimodaux. Leurs chaînes d'entraînement sont généralement plus lourdes, leur comportement peut encore dépendre de la conception des jeux de données, et leurs gains intermodaux peuvent s'effondrer si le modèle apprend des raccourcis sans lien avec de véritables preuves forensiques. Cette limite n'est pas marginale ; elle est désormais reconnue comme un enjeu central dans l'évaluation comparative de la détection audiovisuelle des deepfakes.

2.11 Synthèse et problèmes ouverts

Discussion comparative

Pris ensemble, les tableaux 2.4 et 2.5 montrent que le domaine a atteint un stade où les seuls scores bruts sur bancs d'essai ne suffisent plus à comparer les méthodes de manière

significative. La question centrale n'est plus simplement de savoir quel modèle atteint la meilleure AUC dans le même domaine, mais quel modèle généralise, localise les manipulations et fournit des preuves interprétables sous des contraintes réalistes de déploiement.

Les détecteurs visuels restent indispensables, car la falsification laisse souvent des traces spatiales ou temporelles locales. C'est pourquoi les travaux récents continuent de raffiner l'attention centrée sur les artefacts, la modélisation temporelle du style et l'analyse fréquentielle temporelle. Cependant, le raisonnement purement visuel présente une limite structurelle dans les scénarios de visage parlant : il ne peut pas vérifier si l'articulation observée est cohérente avec le signal de parole associé. Cette limite devient plus importante à mesure que la génération visuelle s'améliore.

Les détecteurs uniquement audio complètent cette faiblesse en ciblant plus directement la parole synthétique. En particulier, le succès récent des modèles de parole fondés sur SSL et le post-entraînement suggère que la détection audio des deepfakes évolue vers des représentations plus transférables et plus robustes. Toutefois, les systèmes uniquement audio restent incomplets pour le problème traité dans ce mémoire, car des manipulations réalistes peuvent préserver une parole plausible tout en modifiant l'articulation visuelle, ou inversement conserver le flux visuel tout en modifiant le contenu audio.

Les détecteurs audiovisuels sont donc les candidats les plus naturels pour l'analyse forensique réaliste des visages parlants. Leur principal avantage est d'exploiter à la fois les preuves intramodales et les incohérences intermodales. Néanmoins, ils constituent aussi la famille la plus fragile du point de vue de l'évaluation : leurs gains peuvent dépendre fortement de la qualité de synchronisation, de la préparation du jeu de données et de biais cachés dans le banc d'essai. Ainsi, un détecteur AV convaincant doit être analysé non seulement en termes d'AUC ou d'AP, mais aussi selon ses modes d'échec, sa robustesse aux perturbations et le type de preuve qu'il utilise réellement.

Tendances récentes (2024 à 2026)

Une première tendance claire de la littérature récente est le passage vers une *détection orientée vers la généralisation*. Dans le domaine visuel, les méthodes récentes traitent de plus en plus la détection des deepfakes comme un problème fin ou orienté transfert plutôt que comme une classification binaire fermée. LAA-Net [70], la modélisation temporelle du style latent [71], l'apprentissage spatio-temporel sensible aux vulnérabilités [74] et l'adaptation efficace en paramètres fondée sur CLIP [85] reflètent tous le même déplacement stratégique : les détecteurs doivent apprendre des représentations utiles au-delà d'un seul banc d'essai ou d'une seule famille de générateurs. Une tendance similaire apparaît en détection audio,

où le post-entraînement sur des corpus divers riches en artefacts améliore la robustesse plus efficacement qu'un ajustement fin étroit seul [80].

Une deuxième grande tendance est le renforcement des formulations *temporelles et sensibles à la localisation*. Les détecteurs visuels récents ciblent de plus en plus directement les incohérences temporelles, plutôt que de traiter le temps comme un simple axe d'agrégation secondaire. En parallèle, les travaux AV se déplacent vers une détection et une localisation conjointes, comme l'illustrent MMMS-BA [83] et les cadres récents de localisation fondés sur la synchronisation [84]. Cette évolution est particulièrement importante pour les falsifications pilotées par le contenu, où seule une courte phrase ou un court segment est manipulé tandis que le reste du clip demeure authentique.

Une troisième tendance est la montée du *réalisme des bancs d'essai*. Les jeux de données et plateformes d'évaluation récents ne supposent plus qu'un deepfake correspond nécessairement à un clip entier présentant de forts artefacts visibles. ASVspoof 5 élargit l'évaluation audio à des conditions acoustiques et adversariales plus diverses [86], tandis qu'AV-deepfake1M++ [12] étend l'évaluation AV à plus grande échelle avec des perturbations proches du monde réel [12]. Du côté visuel et des visages parlants, TalkingHeadBench soutient explicitement que de nombreux anciens bancs d'essai ne représentent pas entièrement les menaces modernes liées aux visages parlants [87]. Cela signifie qu'un état de l'art contemporain doit discuter non seulement les détecteurs, mais aussi le réalisme et la validité des jeux de données sur lesquels ils sont mesurés.

Une quatrième tendance est l'importance croissante de l'*interprétabilité*. Des travaux récents commencent à combiner la détection avec des explications textuelles ou au niveau des régions, par exemple au moyen de modèles forensiques multimodaux et vision-langage [88, 89]. Cette ligne de recherche demeure relativement jeune, mais répond à un besoin forensique réel : dans de nombreux contextes sensibles, une sortie binaire sans preuve d'appui a une valeur pratique limitée.

En même temps, des études récentes ont rendu la communauté plus prudente. En particulier, l'apprentissage par raccourcis est devenu une préoccupation sérieuse en détection audiovisuelle des deepfakes. Smeu et al. [90] montrent que des jeux de données AV largement utilisés peuvent contenir des indices parasites, comme un court préfixe silencieux dans des clips synthétiques, à partir desquels des classificateurs triviaux peuvent obtenir des performances artificiellement élevées. Cette observation est cruciale, car elle rappelle que des scores élevés ne signifient pas nécessairement un raisonnement forensique solide.

Problèmes ouverts et positionnement de ce mémoire

Dans ce contexte, plusieurs problèmes ouverts demeurent non résolus.

Le premier est la *généralisation inter-jeux de données et inter-manipulations*. Un détecteur peut très bien fonctionner sur FF++, Celeb-DF, FakeAVCeleb ou même AV-deepfake1M, tout en échouant devant une autre chaîne de génération, un autre profil de compression ou une autre distribution de styles de parole. Ce problème est visible dans les contextes visuels et AV, et constitue l'un des arguments les plus forts en faveur de conceptions orientées transfert et d'une évaluation prudente.

Le deuxième problème ouvert est la *localisation de manipulations courtes et sémantiquement ciblées*. Beaucoup de manipulations AV récentes ne remplacent pas une vidéo entière ; elles altèrent un mot, une phrase ou un court intervalle. La classification globale au niveau du clip devient alors insuffisante. Un détecteur doit donc préserver l'information locale dans le temps et éviter de moyenner les indices mêmes qu'il cherche à détecter.

Le troisième problème est la *faisabilité en temps réel*. Certaines des méthodes récentes les plus performantes reposent sur un préentraînement lourd, de grands réseaux dorsaux ou des chaînes d'inférence en plusieurs étapes. Ces avancées sont utiles, mais elles ne résolvent pas automatiquement le problème pratique traité dans ce mémoire, soit le traitement en ligne ou quasi en ligne avec ressources computationnelles limitées. Pour la modération, l'analyse de réunions ou la surveillance, la conception du détecteur doit équilibrer la richesse représentationnelle, la latence et la simplicité de mise en œuvre.

Le quatrième problème est l'*interprétabilité forensique*. Un détecteur pratiquement utile ne devrait pas seulement classifier un échantillon comme falsifié ; il devrait aussi indiquer quelle modalité, quelle région ou quel intervalle a contribué à la décision. À cet égard, le défi est autant méthodologique que technique : le détecteur doit encourager l'inspection, et non simplement produire un score.

Ces points non résolus permettent de situer plus précisément la présente thèse. Plutôt que de concurrencer les plus grands détecteurs plein visage ou modèles de fondation entièrement généraux, ce travail cible volontairement un régime plus étroit mais très important : les *manipulations liées à la parole, localisées et temporellement structurées*. La méthodologie proposée repose sur trois hypothèses issues de la littérature analysée. Premièrement, la région de la bouche est un site forensique privilégié dans les scénarios de visage parlant, car elle concentre l'interaction entre la géométrie articulatoire et la texture visuelle locale. Deuxièmement, la parole synthétique n'est pas uniformément informative sur toutes les unités : l'analyse au niveau des phonèmes ou des mots peut révéler où le flux audio s'écarte le plus de la production naturelle. Troisièmement, la preuve intermodale est la plus utile

lorsqu'elle est alignée temporellement et conservée à un niveau local, plutôt que réduite prématurément à un descripteur global.

C'est pourquoi cette thèse combine une représentation visuelle centrée sur la bouche, une analyse phonétique de l'audio fondée sur HuBERT et une stratégie légère de fusion audiovisuelle visant la localisation temporelle. Autrement dit, la contribution se situe à l'intersection de quatre besoins identifiés par la littérature récente : la preuve locale, le raisonnement intermodal, l'interprétabilité et la faisabilité en temps réel. Ce positionnement est cohérent avec les objectifs de la thèse, mais répond aussi à une lacune réelle de l'état de l'art : de nombreux détecteurs existants sont soit trop globaux, soit trop spécifiques à un banc d'essai, soit trop coûteux computationnellement pour le cadre de déploiement visé.

Pour rendre ces défis concrets, les figures 2.7 et 2.8 montrent des exemples représentatifs issus de jeux de données largement utilisés. Ces figures demeurent utiles dans le chapitre, car elles rappellent visuellement que certaines manipulations sont encore associées à des traces de fusion ou de compression visibles, tandis que d'autres sont beaucoup plus difficiles à distinguer par inspection seule. Leur rôle n'est donc pas seulement illustratif : elles soutiennent l'argument méthodologique selon lequel une détection robuste ne peut pas reposer sur un seul indice visuel ou une seule condition de jeu de données.

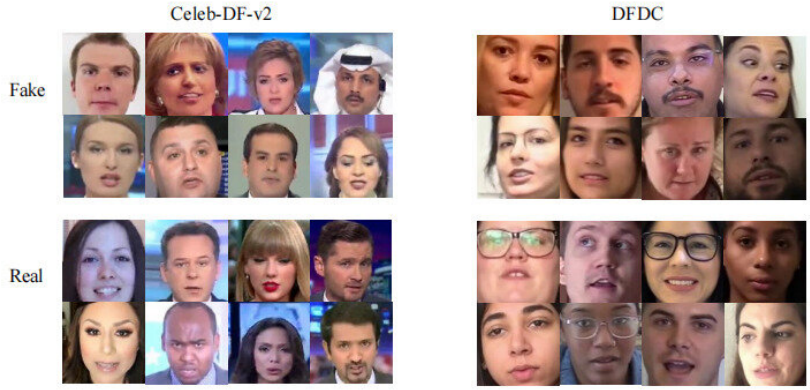


FIGURE 2.7 – Exemples d'images réelles et manipulées provenant de Celeb-DF V2 et DFDC [8]. La figure illustre la diversité visuelle des artefacts de deepfake entre les jeux de données et les chaînes de génération (adapté de [1, 9]).

Comme le montre la figure 2.8, le réalisme des manipulations varie considérablement selon la méthode de synthèse. Dans certains cas, des artefacts visuels restent clairement perceptibles, tandis que, dans d'autres, les images manipulées paraissent suffisamment réalistes pour rendre leur détection difficile par une simple inspection visuelle. Cette diversité explique en partie pourquoi la généralisation entre différents jeux de données et différentes méthodes de manipulation demeure complexe.

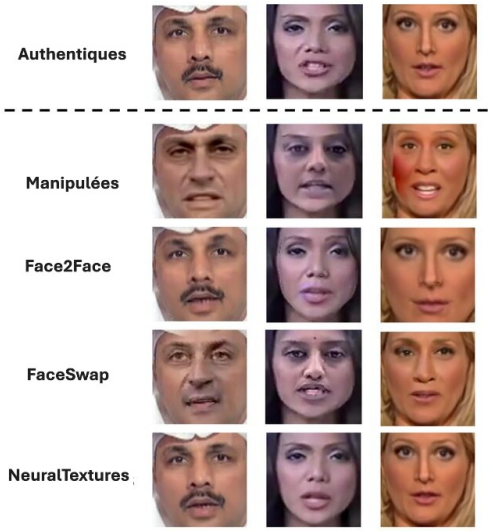


FIGURE 2.8 – Exemples d'images authentiques et manipulées provenant de FaceForensics++. La première ligne présente les images authentiques, tandis que les lignes suivantes montrent les résultats produits par différentes méthodes de manipulation faciale, notamment Deepfakes, Face2Face, FaceSwap et NeuralTextures. Ces exemples illustrent des artefacts visuels tels que les erreurs de fusion, les incohérences de texture et les déformations subtiles des frontières faciales. Adapté de [8].

La figure 2.8 montre que la nature et l'intensité des artefacts varient selon la méthode de manipulation. Lorsque ces artefacts restent localisés et visibles, un détecteur spatial ou spatio-temporel peut les identifier efficacement. Cependant, le réencodage, la compression et l'amélioration des techniques de synthèse peuvent atténuer ces traces. Dans de telles conditions, l'exploitation des informations temporelles et audiovisuelles devient particulièrement importante.

2.12 Conclusion

Ce chapitre a passé en revue les principales familles méthodologiques utilisées dans la détection contemporaine des deepfakes et a montré que le domaine traverse une transition décisive. Les premiers succès reposaient souvent sur des artefacts visibles et des indices propres aux bancs d'essai ; les travaux récents recherchent de plus en plus la robustesse aux manipulations inconnues, la localisation temporelle explicite, un raisonnement intermodal plus fort et une meilleure interprétabilité. En ce sens, l'état de l'art n'est plus défini uniquement par une meilleure exactitude dans le même domaine, mais par un ensemble plus exigeant de critères incluant la généralisation, la qualité de localisation et le réalisme du déploiement.

L'analyse comparative montre également qu'aucune famille de méthodes n'est suffisante dans tous les contextes. Les détecteurs visuels restent indispensables, surtout lorsque les

anomalies de texture ou de mouvement sont marquées, mais ils peuvent manquer des incohérences qui n'apparaissent que lorsque la parole et la dynamique faciale sont analysées conjointement. Les systèmes uniquement audio sont puissants contre la parole synthétique, mais ils ne peuvent pas vérifier la cohérence visuelle. Les modèles audiovisuels sont donc la direction la plus prometteuse pour l'analyse forensique réaliste des visages parlants, même s'ils soulèvent aussi les plus grands défis en matière de biais des jeux de données, de complexité d'entraînement et de déploiement en temps réel.

Pour la présente thèse, cette revue mène à une lacune de recherche précise. Il existe encore un besoin pour un détecteur qui ne se contente pas de classifier des clips entiers, mais qui combine des preuves visuelles locales, une analyse informée par la parole et un raisonnement AV aligné temporellement, tout en restant faisable sur le plan computationnel et interprétable sur le plan forensique. C'est précisément dans cet espace que se situe la méthodologie proposée. Le chapitre suivant passe donc de l'analyse de la littérature à la conception méthodologique, en introduisant les composantes phonétique, visuelle et audiovisuelle développées dans ce mémoire.

CHAPITRE 3

Méthodologie

Ce chapitre présente la trajectoire méthodologique suivie dans ce mémoire pour étudier la détection des deepfakes à partir de deux sources complémentaires : la parole et la dynamique visuelle de la bouche. L'idée directrice consiste à partir de représentations contrôlées et interprétables au niveau des unités, puis à construire un détecteur visuel centré sur la bouche. Concrètement, nous procédons en deux étapes. D'abord (section 3.1), nous introduisons une analyse phonétique fondée sur des plongements HuBERT auto-supervisés. Ensuite (section 3.2), nous proposons un détecteur visuel à fusion précoce centré sur la région de la bouche. L'intégration complète de ces deux composantes dans une future approche de fusion audio–vidéo est discutée comme perspective de recherche, et non comme une section méthodologique principale dans cette version. Le chapitre se termine par une discussion (section 3.3) sur les gains obtenus, les fragilités restantes et la manière dont ces deux blocs préparent une future fusion audiovisuelle.

3.1 Analyse phonétique de la parole réelle et synthétique à partir des plongements HuBERT : perspectives pour la détection des deepfakes

Cette section introduit le bloc d'analyse audio qui complète la méthodologie visuelle développée plus loin dans ce chapitre. L'idée centrale est que la parole synthétique moderne est souvent convaincante au niveau de l'énoncé, mais demeure imparfaite à une échelle linguistique plus fine. Plutôt que de traiter l'ensemble de la forme d'onde comme uniformément informatif, nous cherchons à identifier les unités phonétiques les plus touchées par la synthèse et donc les plus utiles pour la détection des deepfakes. Pour ce faire, nous nous appuyons sur HuBERT, un modèle de parole auto-supervisé qui convertit la forme d'onde en plongements contextuels au niveau des trames, puis nous analysons ces plongements au niveau des phonèmes.

3.1.1 Justification

La preuve visuelle est souvent puissante, mais elle n'est pas toujours entièrement observable. Dans des vidéos réalistes, la bouche peut être partiellement occultée, le visage peut apparaître à faible résolution ou le locuteur peut détourner la tête de la caméra. Dans de telles situations, le flux audio devient particulièrement précieux. Toutefois, les systèmes récents de synthèse vocale (TTS) et de conversion vocale (VC) ont progressé au point que la parole synthétique peut sembler globalement naturelle à l'écoute humaine. Cela signifie que les descripteurs grossiers au niveau de l'énoncé sont souvent trop peu sensibles : ils moyennent l'ensemble du signal et peuvent masquer les unités locales où la synthèse échoue réellement.

Notre hypothèse de travail est donc que la parole synthétique n'est pas uniformément inexacte. Certains phonèmes, certaines diphtongues et certains mots fonctionnels courts sont plus difficiles à synthétiser fidèlement que d'autres. Si cette hypothèse est correcte, un détecteur efficace ne devrait pas traiter toutes les parties de l'audio de la même manière. Il devrait plutôt identifier les unités qui divergent le plus de la parole naturelle et utiliser ensuite cette information comme guide pour la détection. L'objectif de cette section est précisément de rendre cette intuition explicite, mesurable et réutilisable dans le cadre audiovisuel plus large.

3.1.2 Corpus parallèle de parole réelle et synthétique

Pour étudier ce phénomène dans des conditions contrôlées, nous supposons un corpus *parallèle*. Pour chaque phrase prononcée, nous disposons d'un énoncé réel et de plusieurs versions synthétiques générées à partir du même contenu linguistique. En pratique, cela signifie qu'un même texte est réalisé par un locuteur naturel et par un ou plusieurs systèmes de synthèse, tels que Coqui TTS, VITS TTS ou des modèles de conversion vocale. Comme le texte est fixe, toute différence systématique observée ensuite dans l'espace des plongements peut être attribuée au processus de synthèse plutôt qu'à une variation lexicale.

Cette construction parallèle est essentielle. Si les énoncés réels et synthétiques avaient des transcriptions différentes, une séparation dans l'espace des caractéristiques pourrait simplement refléter une variation linguistique. Ici, au contraire, la comparaison est beaucoup plus interprétable : les signaux réel et synthétique sont alignés au niveau du sens, et l'écart restant reflète la manière dont la chaîne de synthèse transforme la réalisation acoustique du même contenu. Autrement dit, le corpus isole l'effet de la génération de parole elle-même.

3.1.3 Plongements HuBERT et agrégation au niveau des unités

HuBERT traite la forme d'onde et produit une séquence de représentations cachées à un pas temporel régulier. Dans notre configuration, le modèle produit un plongement toutes les 20 ms. Soit

$$h_t \in \mathbb{R}^d$$

le plongement HuBERT à la trame t, où d est la dimension du plongement. Intuitivement, h_t résume le contenu acoustique local autour de cet instant, mais dans un espace de représentation appris à partir d'un entraînement de parole auto-supervisé à grande échelle plutôt qu'au moyen de caractéristiques conçues manuellement.

Une séquence au niveau de la trame n'est toutefois pas encore l'objet le plus approprié pour une comparaison phonétique. Un phonème s'étend généralement sur plusieurs trames consécutives, et le nombre exact de trames dépend du débit de parole, de la coarticulation et des détails d'alignement. Pour comparer la parole réelle et la parole synthétique au niveau des phonèmes, il faut donc d'abord connaître les frontières temporelles de chaque unité linguistique. Nous obtenons ces frontières au moyen d'un aligneur forcé, qui fournit pour chaque phonème ou mot un intervalle temporel

$$[t_s, t_e],$$

où t_s et t_e sont les indices de début et de fin associés à cette unité.

Une fois ces frontières disponibles, nous convertissons la séquence de trames HuBERT de longueur variable correspondant à un phonème en un vecteur de taille fixe au moyen d'un moyennage :

$$\phi(p) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} h_t. \tag{3.1}$$

Dans cette expression, p désigne une occurrence de phonème, t_s et t_e sont respectivement les indices de sa première et de sa dernière trame, et $h_t \in \mathbb{R}^d$ est le plongement HuBERT de la trame t. Le vecteur $\phi(p)$ est donc la moyenne des plongements HuBERT associés à toutes les trames appartenant à cette occurrence du phonème.

Le rôle de (3.1) est direct, au lieu de conserver une description fluctuante image par image du phonème, nous résumons tout le phonème par sa représentation moyenne. Cette opération réduit le bruit local des trames, atténue les petites imprécisions d'alignement et permet de comparer des phonèmes de durées différentes dans un espace commun. Le résultat, noté $\phi(p)$, peut être interprété comme un plongement compact d'une occurrence du phonème p.

La répétition de cette procédure sur tous les énoncés produit, pour chaque phonème p,

deux collections empiriques de plongements :

$$\left\{\phi_i^{\text{rel}}(p)\right\}_{i=1}^{N_{\text{rel}}(p)}, \qquad \left\{\phi_j^{\text{synthtique}}(p)\right\}_{j=1}^{N_{\text{synthtique}}(p)}. \tag{3.2}$$

Dans cette notation, $N_{\text{rel}}(p)$ désigne le nombre d'occurrences réelles disponibles pour le phonème p, tandis que $N_{\text{synthtique}}(p)$ désigne le nombre d'occurrences synthétiques. Les accolades représentent les deux ensembles de plongements utilisés pour estimer les distributions réelle et synthétique du phonème.

Ces deux ensembles peuvent être vus comme deux nuages dans l'espace de plongement HuBERT : un nuage pour la parole réelle et un nuage pour la parole synthétique. Si un système de synthèse reproduit bien le phonème p, les deux nuages devraient se recouvrir largement. Si, au contraire, ce phonème est systématiquement déformé, le nuage synthétique devrait s'éloigner du nuage réel. La sous-section suivante formalise cet écart.

3.1.4 Classement fondé sur la divergence et principaux constats

La question suivante consiste à mesurer l'écart entre les nuages de plongements réels et synthétiques associés à chaque phonème. L'objectif n'est pas seulement d'établir que certains phonèmes diffèrent, mais également de les classer selon leur pouvoir informatif pour la détection. Pour rendre cette comparaison traitable, nous estimons, pour chaque phonème, une distribution à partir des plongements réels et une autre à partir des plongements synthétiques, puis nous quantifions leur divergence.

Pour un phonème p, nous notons P_p^{rel} et $P_p^{\text{synthtique}}$ les distributions estimées respectivement à partir de ses plongements réels et synthétiques. Nous définissons alors son score de divergence par :

$$\text{KLD}(p) = D_{\text{KL}}\left(P_p^{\text{rel}} \quad P_p^{\text{synthtique}}\right). \tag{3.3}$$

Le symbole $D_{\text{KL}}(P\|Q)$ désigne la divergence de Kullback–Leibler de la distribution P par rapport à la distribution Q. Dans l'équation (3.3), la distribution de la parole réelle est utilisée comme distribution de référence. Une valeur proche de zéro indique que les distributions réelle et synthétique sont similaires dans l'espace des plongements HuBERT. À l'inverse, une valeur élevée indique que les réalisations synthétiques du phonème s'écartent de manière systématique de ses réalisations naturelles. La divergence agit ainsi comme un *score diagnostique*, permettant d'identifier les unités phonétiques les plus affectées par la synthèse et les plus informatives pour la détection des deepfakes.

Il convient de noter que la divergence de Kullback–Leibler n'est généralement pas symé-

trique :

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P).$$

L'ordre des distributions dans l'équation doit donc être conservé lors de toutes les comparaisons.

Une fois la divergence estimée pour l'ensemble des phonèmes, nous classons les unités par ordre décroissant de KLD(p). Le classement obtenu s'étend ainsi des phonèmes les plus sensibles à la synthèse aux phonèmes les moins sensibles. Empiriquement, les unités situées en haut de ce classement comprennent souvent des voyelles, des diphtongues et certains mots fonctionnels courts. Ce résultat suggère que les systèmes de synthèse peuvent reproduire de manière convaincante la structure prosodique globale, tout en échouant à reproduire certains détails articulatoires et spectraux fins concentrés dans des unités linguistiques particulières.

La conséquence pratique de ce classement est importante pour la suite de la méthodologie. Elle suggère qu'un détecteur audio ne devrait pas ignorer la structure linguistique. Au contraire, il peut accorder plus d'attention aux fenêtres contenant les unités phonétiques les plus divergentes. En ce sens, la présente analyse joue un rôle exploratoire et orienté vers l'interprétabilité : avant de construire le détecteur complet, elle identifie où la parole synthétique est la plus vulnérable.

Les principaux constats sont donc de deux ordres :

1. La détection audio des deepfakes ne nécessite pas de traiter uniformément tous les segments de parole, certaines unités phonétiques étant empiriquement plus informatives que d'autres.
2. Le classement des phonèmes selon leur niveau de divergence pourra ensuite être exploité dans une future approche de fusion audio–vidéo, notamment pour mettre en évidence, pondérer ou prioriser les fenêtres temporelles contenant des phonèmes à forte divergence.

Cette analyse phonétique n'a pas vocation à remplacer la détection des deepfakes sur séquence complète. Elle fournit plutôt une manière fondée de comprendre où la preuve acoustique est la plus forte, et elle complète la méthodologie visuelle développée dans la section suivante. Dans la logique générale de ce mémoire, elle joue le rôle de contrepartie audio à l'analyse locale de la bouche : les deux visent à passer d'un jugement global grossier à des indices locaux, interprétables et pertinents pour le diagnostic.

3.2 Détecteur visuel à fusion précoce géométrie-texture

3.2.1 Positionnement méthodologique

La plupart des détecteurs visuels récents de deepfakes appartiennent à deux familles : (i) ceux qui reposent sur des *artefacts faciaux globaux* et des traces de compression ; (ii) ceux qui reposent sur un *seul* indice physiologique ou géométrique (clignement, rPPG, points de repère). Ces deux approches sont utiles, mais chacune possède ses fragilités. Pendant la parole, la région de la bouche est toutefois particulière : deux processus physiques différents y sont couplés, soit la cinématique des lèvres et de la mâchoire, et l'apparence de la zone intra-orale (bords des dents, langue, reflets spéculaires). Un manipulateur peut réussir l'un de ces aspects et échouer l'autre. La fusion précoce, effectuée *avant* l'intégration temporelle, n'est donc pas un choix esthétique, mais une manière de placer les deux signaux devant la même tête temporelle afin que les incohérences brèves ne soient pas lissées.

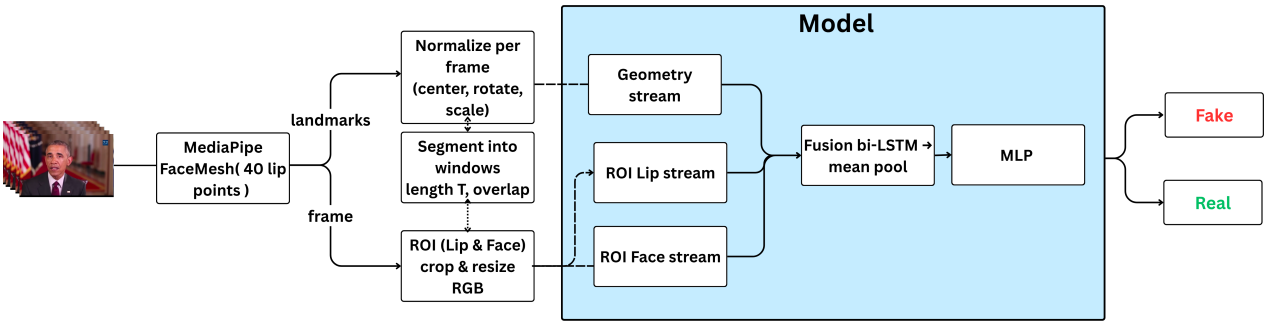


FIGURE 3.1 – Chaîne de traitement globale : la vidéo est découpée en fenêtres temporelles, analysée par le modèle, puis utilisée pour produire une décision finale.

3.2.2 Formulation du problème

L'objectif est de décider, pour chaque court segment d'une séquence visuelle, s'il contient un contenu *manipulé* ou un segment *authentique*, et de prendre cette décision d'une manière compatible avec un traitement en temps réel ou quasi temps réel. Nous ciblons la classe de deepfakes où la région de la bouche porte l'essentiel de la preuve, comme le montrent des jeux de données AV récents tels que FakeAVCeleb et AV-deepfake1M, où les désalignements labiaux et les artefacts de fusion autour de la bouche sont fréquents [16, 65, 91]. Nous voulons aussi que la même formulation puisse soutenir les cas d'anomalies (cris, collisions, mouvements inhabituels) dans des contextes proches de la surveillance [20].

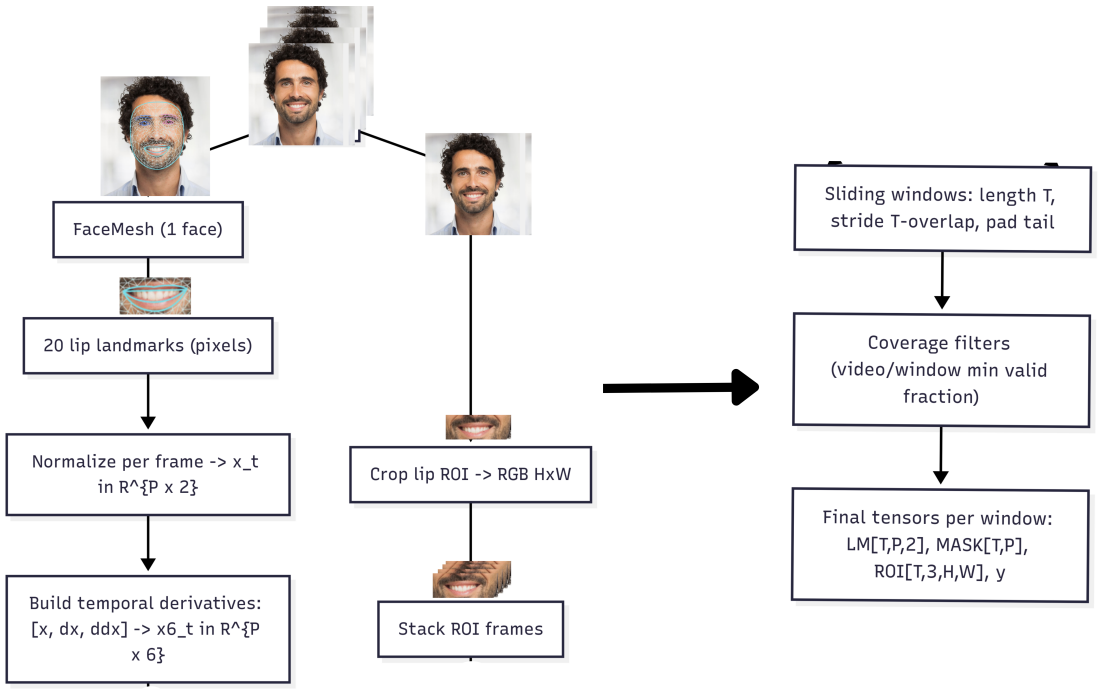


FIGURE 3.2 – Étapes de prétraitement : MediaPipe FaceMesh [10], recadrage de la ROI de la bouche, dérivées temporelles et fenêtres glissantes avec filtrage selon la couverture des points de repère.

Flux vidéo brut : Nous notons

$$V = \{I_t\}_{t=1}^{L}$$

une séquence vidéo composée de L images, échantillonnée à 25 images par seconde (ips), ou rééchantillonnée à cette fréquence afin d'assurer la cohérence du traitement. Chaque image

$$I_t \in [0,1]^{H_0 \times W_0 \times 3}$$

est une image RGB dont les valeurs de pixels sont normalisées dans l'intervalle $[0,1]$.

Pour chaque image, nous appliquons **MediaPipe Face Landmarker v2 [10]** en mode visage unique afin d'extraire $P = 40$ points de repère associés aux lèvres. Ces points sont regroupés dans la matrice

$$u_t \in \mathbb{R}^{P \times 2},$$

où chaque ligne contient les coordonnées bidimensionnelles d'un point de repère dans l'image.

En parallèle, nous extrayons une région d'intérêt visuelle

$$R_t \in [0,1]^{H \times W \times 3}$$

à partir de la région de la bouche. Dans la configuration par défaut, ce recadrage est obtenu au moyen d'une boîte englobante serrée autour des lèvres. La boîte est agrandie par un facteur d'environ 1,3, puis redimensionnée à une résolution de 96×96 pixels par interpolation bilinéaire déterministe.

Dans certaines variantes, nous utilisons un recadrage plus large incluant les lèvres ainsi qu'une partie du bas du visage. Cette représentation permet au flux de texture de capter à la fois les indices intra-oraux, comme les dents, la langue et les frontières internes des lèvres, et les indices faciaux voisins susceptibles de révéler des artefacts de fusion ou d'autres incohérences locales.

Structure temporelle par fenêtres glissantes : La vidéo est analysée au moyen de fenêtres glissantes de longueur fixe T et de pas $s < T$. Pour la k-ième fenêtre, nous définissons l'ensemble de ses indices temporels par

$$\mathcal{I}_k = \{t_k, t_k + 1, \ldots, t_k + T - 1\}, \tag{3.4}$$

où t_k désigne l'indice de la première image de la fenêtre.

Les observations associées à cette fenêtre sont regroupées sous la forme

$$U_k = \{u_t\}_{t \in \mathcal{I}_k}, \qquad M_k = \{m_t\}_{t \in \mathcal{I}_k}, \qquad \mathcal{R}_k = \{R_t\}_{t \in \mathcal{I}_k}. \tag{3.5}$$

Le terme

$$m_t \in \{0, 1\}$$

est un masque de validité associé à l'image t. Il vaut 1 lorsque les points de repère des lèvres sont détectés avec une qualité suffisante et 0 lorsque le suivi est considéré comme invalide. Le masque M_k indique ainsi quelles images de la fenêtre peuvent contribuer au calcul des représentations et à leur agrégation temporelle.

Les fenêtres constituent les unités élémentaires utilisées pour l'entraînement, l'évaluation et l'inférence en flux continu. Cette organisation permet de traiter une longue vidéo par blocs temporels de taille fixe, d'ignorer les images dont le suivi des lèvres est invalide et de rejeter les fenêtres dont la couverture des points de repère est insuffisante. Elle permet également de ramener la décision au niveau de la vidéo à une agrégation des prédictions obtenues sur les différentes fenêtres.

Normalisation géométrique : Les coordonnées brutes u_t dépendent de la position du visage dans l'image, de son échelle et de sa rotation dans le plan. Afin de réduire l'influence

de ces variations, nous appliquons une normalisation par translation, rotation dans le plan et mise à l'échelle.

Soient c_ℓ et c_r les indices des commissures gauche et droite de la bouche. Le vecteur reliant ces deux commissures est défini par

$$\mathbf{v}_t = u_t[c_r, :] - u_t[c_\ell, :] \in \mathbb{R}^2. \tag{3.6}$$

Nous calculons ensuite le centre moyen des points de repère :

$$\overline{\mathbf{u}}_t = \frac{1}{P} \sum_{j=1}^{P} u_t[j, :] \in \mathbb{R}^2. \tag{3.7}$$

Les points centrés sont alors donnés par

$$\widetilde{u}_t[j, :] = u_t[j, :] - \overline{\mathbf{u}}_t, \qquad j = 1, \ldots, P. \tag{3.8}$$

L'orientation de la bouche et la distance entre les commissures sont calculées comme suit :

$$\theta_t = \operatorname{atan2}(v_{t,y}, v_{t,x}), \qquad d_t = \max\left(\|\mathbf{v}_t\|_2, \varepsilon\right). \tag{3.9}$$

Dans cette expression, θ_t est l'angle formé par l'axe des commissures avec l'axe horizontal de l'image. La quantité d_t correspond à la distance entre les deux commissures et sert à normaliser l'échelle. La constante $\varepsilon > 0$ est une petite valeur numérique utilisée pour empêcher une division par zéro lorsque cette distance devient très faible.

Pour un vecteur $\mathbf{z} = (z_1, \ldots, z_n)$, la norme euclidienne et son carré sont définis par

$$\|\mathbf{z}\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2}, \qquad \|\mathbf{z}\|_2^2 = \sum_{i=1}^{n} z_i^2. \tag{3.10}$$

La norme au carré est notamment utilisée dans les fonctions de perte, car elle pénalise davantage les grandes différences tout en évitant le calcul d'une racine carrée.

La matrice appliquant une rotation de $-\theta_t$ est définie par

$$\mathbf{Q}_t = \begin{bmatrix} \cos(-\theta_t) & -\sin(-\theta_t) \\ \sin(-\theta_t) & \cos(-\theta_t) \end{bmatrix}. \tag{3.11}$$

Comme chaque point est représenté sous la forme d'un vecteur-ligne, les points de repère

normalisés sont obtenus par

$$x_t = \frac{1}{d_t}\widetilde{u}_t \mathbf{Q}_t^\top \in \mathbb{R}^{P\times 2}. \tag{3.12}$$

Le symbole $\top$ désigne ici la transposition de la matrice de rotation. Cette transformation centre les points de repère, aligne l'axe des commissures avec l'axe horizontal et normalise la distance entre les deux commissures. Le détecteur devient ainsi moins sensible aux translations, aux changements d'échelle et aux petites rotations du visage dans le plan de l'image.

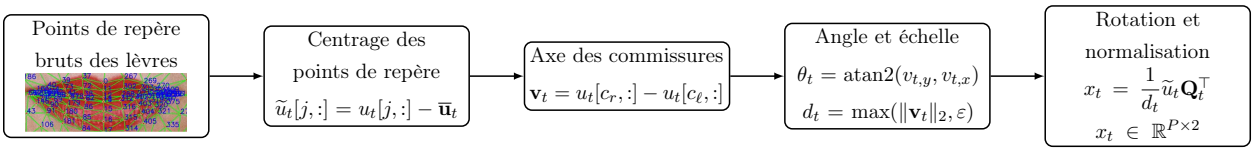


FIGURE 3.3 – Normalisation des points de repère des lèvres par centrage, rotation dans le plan et mise à l'échelle.

Descripteurs géométriques temporels : L'observation de la forme des lèvres dans une image isolée ne suffit pas toujours à révéler une manipulation. Certaines irrégularités apparaissent principalement dans la manière dont les lèvres évoluent d'une image à l'autre. Nous calculons donc des différences discrètes du premier et du second ordre à partir des points normalisés :

$$x_t^{(0)} = x_t, \qquad x_t^{(1)} = x_t - x_{t-1}, \qquad x_t^{(2)} = x_t^{(1)} - x_{t-1}^{(1)}. \tag{3.13}$$

Le terme $x_t^{(1)}$ constitue une approximation discrète de la vitesse des points de repère entre les images $t-1$ et t. Il mesure donc la variation image par image de la configuration normalisée des lèvres.

Le terme $x_t^{(2)}$ constitue une approximation discrète de l'accélération, puisqu'il mesure la variation de cette vitesse entre deux pas de temps consécutifs.

Pour la première image de chaque fenêtre, aucune image précédente n'est disponible. Nous utilisons donc la convention

$$x_{t_k}^{(1)} = \mathbf{0}, \qquad x_{t_k}^{(2)} = \mathbf{0}. \tag{3.14}$$

Ces descripteurs temporels permettent de révéler des transitions articulatoires irrégulières qui peuvent être difficiles à observer dans une image isolée. Une vidéo manipulée peut en effet préserver une forme de lèvres visuellement plausible tout en échouant à reproduire la continuité naturelle de son mouvement.

Pour chaque image, les positions, les vitesses et les accélérations sont concaténées le long

de la dimension des caractéristiques :

$$x_t^{(6)} = \text{concat}\left(x_t^{(0)}, x_t^{(1)}, x_t^{(2)}\right) \in \mathbb{R}^{P\times 6}. \tag{3.15}$$

La dernière dimension contient ainsi les six composantes

$$(x, y, \Delta x, \Delta y, \Delta^2 x, \Delta^2 y)$$

associées à chaque point de repère.

Sur la k-ième fenêtre, les représentations de ses T images sont empilées dans le tenseur

$$X_k^{(6)} = \text{stack}\left(x_t^{(6)}\right)_{t\in\mathcal{I}_k} \in \mathbb{R}^{T\times P\times 6}. \tag{3.16}$$

Les composantes de vitesse et d'accélération sont particulièrement utiles, car les falsifications subtiles peuvent préserver la forme générale des lèvres tout en introduisant des irrégularités dans les micro-transitions temporelles [16, 91].

Modèle d'étiquetage : Chaque fenêtre k est associée à une étiquette binaire

$$y_k \in \{0, 1\},$$

où $y_k = 1$ désigne une fenêtre manipulée et $y_k = 0$ une fenêtre authentique.

Dans une extension audiovisuelle future, cette représentation peut être généralisée à un vecteur multi-étiquettes :

$$\mathbf{y}_k = \left(y_k^{(V)}, y_k^{(A)}, y_k^{(\text{sync})}\right) \in \{0, 1\}^3. \tag{3.17}$$

Les trois composantes représentent respectivement la présence d'une manipulation visuelle, d'une manipulation audio et d'une désynchronisation entre les mouvements des lèvres et la parole [91, 92].

Dans un contexte plus général de détection d'anomalies, le même schéma peut être adapté en remplaçant y_k par une étiquette $y_k^{(\text{anom})}$ indiquant si la fenêtre contient un événement anormal dans la scène [20].

Objectif au niveau de la vidéo : Au moment de l'inférence, le réseau f_Θ produit, pour chaque fenêtre valide k, un logit scalaire $\widehat{\ell}_k \in \mathbb{R}$. Ce logit est transformé en probabilité par la fonction sigmoïde :

$$p_k = \sigma(\widehat{\ell}_k) \in [0, 1]. \tag{3.18}$$

Une valeur de p_k proche de 1 indique que la fenêtre est considérée comme synthétique, tandis qu'une valeur proche de 0 indique qu'elle est considérée comme authentique.

Le score global de la vidéo, noté

$$s(V) \in [0, 1],$$

est ensuite obtenu en agrégeant les probabilités des fenêtres valides. Cette agrégation peut être réalisée au moyen d'une moyenne, d'une médiane ou d'une règle de type top-K, selon le protocole retenu.

La décision finale est obtenue en comparant le score $s(V)$ à un seuil calibré τ :

$$\widehat{y}(V) = \begin{cases} 1, & \text{si } s(V) \geq \tau, \\ 0, & \text{sinon.} \end{cases} \tag{3.19}$$

La valeur $\widehat{y}(V) = 1$ indique que la vidéo est déclarée synthétique, tandis que $\widehat{y}(V) = 0$ indique qu'elle est déclarée authentique.

Justification de cette formulation : Cette formulation place la région de la bouche au centre du processus de détection, car elle contient des indices directement liés à l'articulation de la parole. Elle réduit également l'influence de l'arrière-plan et d'autres régions du visage qui peuvent varier fortement d'une identité ou d'une vidéo à l'autre.

Le traitement par courtes fenêtres permet par ailleurs de maintenir une consommation mémoire limitée et facilite l'inférence en flux continu. Enfin, la même organisation temporelle pourra être utilisée ultérieurement pour aligner les représentations visuelles avec les représentations audio dans une architecture audiovisuelle.

Résumé de la notation

TABLEAU 3.1 – Principaux symboles utilisés dans la formulation du problème.

Symbole	Description
$V = \{I_t\}_{t=1}^{L}$	Séquence vidéo d'entrée composée de L images.
$I_t \in [0,1]^{H_0 \times W_0 \times 3}$	Image RGB observée au temps t.
$u_t \in \mathbb{R}^{P \times 2}$	Matrice contenant les coordonnées des $P = 40$ points de repère des lèvres à l'image t.
$R_t \in [0,1]^{H \times W \times 3}$	Région d'intérêt visuelle extraite à l'image t.
T	Nombre d'images contenues dans une fenêtre temporelle.
s	Pas temporel séparant deux fenêtres glissantes consécutives.
$\mathcal{I}_k$	Ensemble des indices temporels appartenant à la fenêtre k.
$m_t \in \{0,1\}$	Masque indiquant si le suivi des lèvres est valide à l'image t.
M_k	Ensemble des valeurs du masque associées à la fenêtre k.
$\mathbf{v}_t \in \mathbb{R}^2$	Vecteur reliant les commissures gauche et droite de la bouche.
θ_t	Angle de l'axe des commissures par rapport à l'axe horizontal.
d_t	Distance entre les commissures utilisée pour normaliser l'échelle.
ε	Petite constante positive empêchant une division par zéro.
$\mathbf{Q}_t$	Matrice appliquant une rotation de $-\theta_t$.
$x_t \in \mathbb{R}^{P \times 2}$	Points de repère centrés, alignés et normalisés en échelle.
$x_t^{(1)}$	Différence temporelle du premier ordre, interprétée comme une vitesse discrète.
$x_t^{(2)}$	Différence temporelle du second ordre, interprétée comme une accélération discrète.
$x_t^{(6)} \in \mathbb{R}^{P \times 6}$	Représentation combinant la position, la vitesse et l'accélération de chaque point de repère.
$X_k^{(6)} \in \mathbb{R}^{T \times P \times 6}$	Tenseur géométrique associé à la fenêtre k.
$y_k \in \{0,1\}$	Étiquette binaire authentique ou synthétique de la fenêtre k.
$p_k \in [0,1]$	Probabilité prédite pour la classe synthétique à la fenêtre k.
$s(V) \in [0,1]$	Score agrégé au niveau de la vidéo complète.
τ	Seuil utilisé pour produire la décision binaire finale.

3.2.3 Encodeurs à trois branches et fusion précoce

Dans la suite de ce chapitre, le terme *géométrie* désigne la représentation du mouvement obtenue à partir des points de repère des lèvres après leur normalisation et le calcul des différences temporelles.

La *texture de la bouche* désigne la représentation d'apparence extraite d'une région d'intérêt serrée autour des lèvres. Cette région peut contenir les dents, la langue, les frontières des lèvres, l'ombrage local et d'éventuels artefacts introduits par la manipulation.

La *texture du visage* désigne la représentation extraite d'un recadrage plus large centré sur le visage. Elle conserve un contexte supplémentaire autour de la bouche, notamment au niveau du menton et des joues.

La *fusion précoce* consiste à concaténer, pour chaque image, les descripteurs géométriques, les descripteurs de texture de la bouche et les descripteurs de texture du visage avant l'étape finale d'intégration temporelle.

Enfin, la *tête temporelle* désigne le modèle séquentiel chargé d'agréger les représentations fusionnées sur l'ensemble de la fenêtre afin de produire la décision de classification.

L'encodeur visuel proposé comporte ainsi trois branches complémentaires : une branche géométrique et deux branches d'apparence. Cette organisation permet de combiner des informations articulatoires locales avec des informations de texture provenant de la bouche et de son contexte facial.

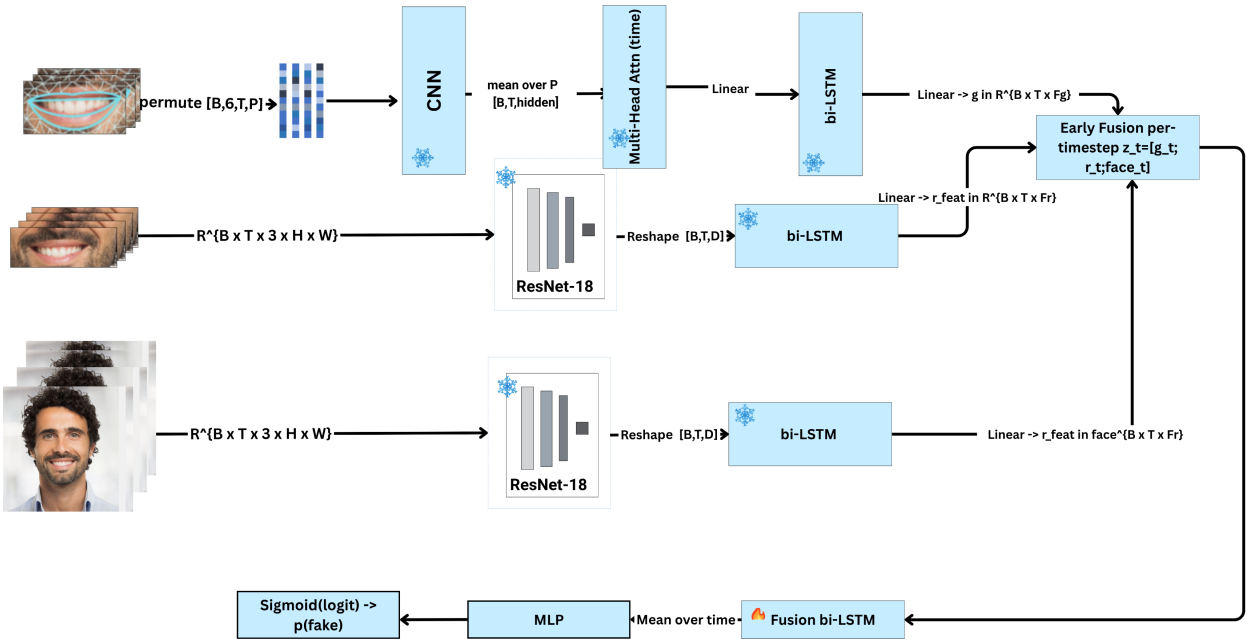


FIGURE 3.4 – Architecture de fusion précoce à trois branches. La branche géométrique traite les points de repère des lèvres normalisés ainsi que leurs dérivées temporelles. La deuxième branche traite une ROI serrée de la bouche avec un ResNet-18 [11] suivi d'un BiLSTM, tandis que la troisième traite un recadrage plus large du visage avec la même chaîne ResNet-18–BiLSTM. Les trois descripteurs sont alignés dans le temps et fusionnés avant l'étape finale d'intégration temporelle.

Nous notons B la taille du lot, T le nombre d'images par fenêtre et P le nombre de points de repère des lèvres, avec $P = 40$ dans notre configuration.

Un lot de tenseurs géométriques est représenté par

$$\mathcal{X}^{(6)} \in \mathbb{R}^{B\times T\times P\times 6}. \tag{3.20}$$

La dernière dimension contient les coordonnées et leurs différences temporelles :

$$(x, y, \Delta x, \Delta y, \Delta^2 x, \Delta^2 y).$$

Les deux entrées d'apparence sont des séquences de recadrages :

$$\mathcal{R}^{\text{mouth}} \in \mathbb{R}^{B\times T\times 3\times H_m\times W_m}, \qquad \mathcal{R}^{\text{face}} \in \mathbb{R}^{B\times T\times 3\times H_f\times W_f}. \tag{3.21}$$

La quantité $\mathcal{R}^{\text{mouth}}$ désigne les recadrages serrés de la bouche, tandis que $\mathcal{R}^{\text{face}}$ désigne les recadrages plus larges du visage. Dans notre implémentation, la ROI de la bouche est redimensionnée à environ 96×96 pixels. Le recadrage du visage est également redimensionné à une résolution fixe compatible avec le réseau dorsal de texture.

Comme l'illustre la figure 3.4, la branche géométrique reçoit $\mathcal{X}^{(6)}$ et applique une courte pile de convolutions sur le plan temporel et spatial ($T \times P$). La dimension correspondant aux points de repère est ensuite agrégée par moyennage. La séquence obtenue est raffinée au moyen d'un bloc d'auto-attention multi-têtes.

Dans la configuration géométrique seule, cet encodeur fournit directement une représentation temporelle structurée du mouvement des lèvres. Dans l'architecture complète, la séquence est projetée puis transmise à un LSTM bidirectionnel afin de l'aligner avec les deux branches d'apparence avant la fusion.

Flux géométrique Le flux géométrique repose sur un module convolutionnel muni d'un mécanisme d'auto-attention. Il reçoit en entrée la représentation géométrique

$$\mathcal{X}^{(6)} \in \mathbb{R}^{B \times T \times P \times 6},$$

qui regroupe les positions normalisées des points de repère des lèvres, ainsi que leurs différences temporelles du premier et du second ordre.

Pour le traitement convolutif, l'ordre des dimensions est modifié :

$$\widetilde{\mathcal{X}}^{(6)} = \text{permute}\left(\mathcal{X}^{(6)}\right) \in \mathbb{R}^{B \times 6 \times T \times P}. \tag{3.22}$$

Les six composantes géométriques jouent ainsi le rôle de canaux d'entrée pour les couches convolutives.

Une pile de deux couches convolutives est ensuite appliquée sur le plan ($T \times P$). Elle permet d'extraire des motifs spatio-temporels locaux à partir de la séquence de points de repère, comme les déformations du contour des lèvres, les mouvements articulatoires courts et les transitions susceptibles de révéler une manipulation.

La dimension associée aux P points de repère est ensuite agrégée par moyennage, ce qui produit une séquence de caractéristiques temporelles :

$$F \in \mathbb{R}^{B \times T \times D}, \tag{3.23}$$

où D désigne la dimension intermédiaire des caractéristiques géométriques.

Pour modéliser des dépendances temporelles plus longues, cette séquence est transmise à un bloc d'auto-attention multi-têtes comportant une connexion résiduelle et une normalisation de couche.

Dans la configuration complète à trois branches, la séquence issue du bloc d'attention est ensuite projetée et transmise à un LSTM bidirectionnel. Ce dernier raffine la représentation en

exploitant les contextes temporels antérieur et postérieur. Cette étape améliore la continuité temporelle de la représentation géométrique et facilite son alignement avec les représentations de texture.

Une dernière normalisation de couche produit la sortie

$$G = (g_1, \ldots, g_T) \in \mathbb{R}^{B \times T \times F_g}, \tag{3.24}$$

où F_g désigne la dimension du descripteur géométrique.

Pour alléger la notation, l'indice du lot est omis. Pour chaque élément du lot, le vecteur

$$g_t \in \mathbb{R}^{F_g}$$

représente donc le descripteur géométrique associé à l'image t.

Dans le modèle complet, ces descripteurs sont fusionnés à chaque pas de temps avec les descripteurs de texture de la bouche et du visage avant d'être transmis à la tête temporelle finale. La branche géométrique combine ainsi une extraction convolutive de motifs locaux, un raisonnement global par attention et une modélisation récurrente de la dynamique des lèvres.

Flux de texture : La branche de texture de la bouche prend la ROI de la bouche R_t^{mouth} à chaque image, extrait un descripteur spatial avec un réseau dorsal ResNet-18 [11], puis transmet la séquence obtenue à un LSTM bidirectionnel afin de capter la cohérence temporelle à court terme de l'apparence de la bouche. Cela produit un descripteur r_t^{mouth} par image.

En parallèle, la branche de texture du visage prend le recadrage plus large R_t^{face}, extrait des caractéristiques d'apparence au niveau de l'image avec un autre réseau dorsal ResNet-18 [11], puis traite la séquence obtenue avec un LSTM bidirectionnel. Elle produit ainsi un descripteur r_t^{face} par image. L'objectif de cette branche est de conserver un contexte facial supplémentaire pouvant contenir des indices de manipulation utiles hors de la ROI stricte de la bouche.

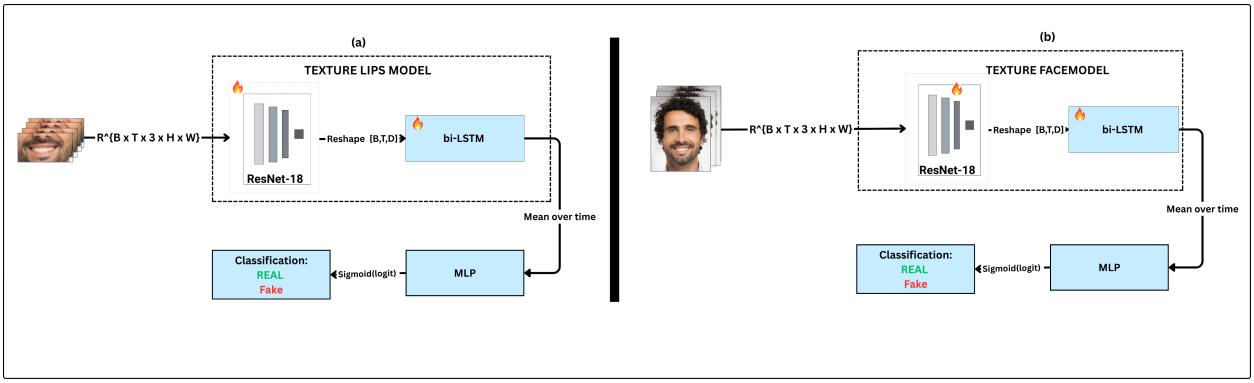


FIGURE 3.5 – Variantes d'entrée de texture considérées dans ce travail. (a) ROI par défaut centrée sur la bouche utilisée dans le flux de texture proposé. (b) Recadrage plus large centré sur le visage, utilisé comme configuration de comparaison afin d'évaluer si un contexte facial supplémentaire améliore la performance de détection.

La figure 3.5 résume la chaîne du flux de texture. Le CNN extrait une représentation spatiale de chaque recadrage d'entrée, puis le BiLSTM modélise l'évolution temporelle de ces descripteurs au niveau de l'image sur la fenêtre. Cela permet aux branches de texture de capter à la fois des indices d'apparence statiques et la cohérence temporelle à court terme dans la région parlante.

Régime d'entraînement des flux visuels : En pratique, nous ne nous sommes pas appuyés exclusivement sur une optimisation conjointe de bout en bout à partir de zéro. Les branches de géométrie, de texture de la bouche et de texture du visage ont d'abord été étudiées dans des configurations séparées ou partiellement découplées afin de vérifier la contribution individuelle de chaque représentation. En particulier, l'encodeur géométrique a d'abord été entraîné dans une configuration géométrique seule. Dans le modèle complet à trois branches, cet encodeur géométrique préentraîné peut ensuite être réutilisé comme extracteur de caractéristiques gelé, tandis que la projection ajoutée, le raffinement BiLSTM et les couches de fusion sont entraînés conjointement avec les branches d'apparence. Cette stratégie rend la comparaison entre géométrie seule, texture seule et fusion plus interprétable, puisqu'elle isole la contribution discriminante de chaque branche avant d'analyser leur interaction par fusion précoce.

Fusion précoce et tête temporelle : La fusion précoce se fait au niveau des images individuelles. Pour chaque indice temporel t, le modèle concatène les trois descripteurs en une représentation conjointe

$$z_t = [g_t; r_t^{\text{mouth}}; r_t^{\text{face}}] \in \mathbb{R}^{F_g + F_r^{\text{mouth}} + F_r^{\text{face}}}, \tag{3.25}$$

de sorte que la tête temporelle observe toujours la géométrie, la texture de la bouche et la texture du visage *ensemble*. La séquence

$$\{z_t\}_{t=1}^{T}$$

est ensuite traitée par un bi-LSTM de fusion, qui agrège les preuves sur la fenêtre. Un moyennage tenant compte du masque sur les images valides produit un vecteur unique résumant la fenêtre, puis un petit MLP à deux couches projette ce vecteur vers un logit $\hat{\ell}$ et une probabilité $p = \sigma(\hat{\ell})$ pour la classe « synthétique ». Ce score est utilisé dans les objectifs au niveau de la fenêtre et de la vidéo décrits précédemment.

3.2.4 Objectif d'apprentissage, déséquilibre et calibration

Pour chaque fenêtre, nous observons une séquence de T images. La branche géométrique reçoit le tenseur de points de repère normalisés par SE(2)

$$X_{1:T}^{(6)} \in \mathbb{R}^{T\times P\times 6},$$

où P est le nombre de points de repère de la bouche et où la dernière dimension empile position, vitesse et accélération,

$$X_{t,p,:}^{(6)} = \big[x_{t,p}, y_{t,p}, \Delta x_{t,p}, \Delta y_{t,p}, \Delta^2 x_{t,p}, \Delta^2 y_{t,p}\big].$$

Après les étapes de convolution, de moyennage, d'attention, de projection et de raffinement BiLSTM décrites ci-dessus, chaque image t est représentée par un descripteur géométrique $g_t \in \mathbb{R}^{F_g}$.

En parallèle, les deux branches d'apparence traitent les séquences de recadrages

$$R_{1:T}^{\text{mouth}} \in \mathbb{R}^{T\times 3\times H_m\times W_m}, \qquad R_{1:T}^{\text{face}} \in \mathbb{R}^{T\times 3\times H_f\times W_f}.$$

Un réseau dorsal ResNet-18 [11] extrait d'abord un descripteur spatial de chaque image, puis les séquences de descripteurs obtenues sont transmises à des LSTM bidirectionnels, qui modélisent la cohérence temporelle à court terme et produisent des descripteurs de texture par image

$$r_t^{\text{mouth}} \in \mathbb{R}^{F_r^{\text{mouth}}}, \qquad r_t^{\text{face}} \in \mathbb{R}^{F_r^{\text{face}}}.$$

La fusion précoce est effectuée au niveau des caractéristiques par concaténation,

$$z_t = [g_t; r_t^{\text{mouth}}; r_t^{\text{face}}],$$

puis la séquence fusionnée $(z_1, \ldots, z_T)$ est transmise à un LSTM bidirectionnel final et à une agrégation temporelle afin de produire une représentation unique au niveau de la fenêtre. Un petit MLP projette cette représentation vers un logit scalaire $\hat{\ell} \in \mathbb{R}$, interprété comme le score non normalisé de la classe « synthétique ».

Étant donné l'étiquette binaire $y \in \{0, 1\}$ associée à la fenêtre, la perte d'entraînement combine un terme standard de classification avec un régularisateur temporel propre à la géométrie :

$$\mathcal{L} \;=\; \mathrm{BCEWithLogits}(\hat{\ell}, y) \;+\; \lambda \frac{1}{T-1} \sum_{t=2}^{T} \; g_t - g_{t-1} \; {}_2^2, \qquad \lambda \in [10^{-4}, 10^{-3}]. \tag{3.26}$$

Terme de classification Le premier terme correspond à l'entropie croisée binaire calculée à partir du logit $\hat{\ell}$ produit par le modèle et de l'étiquette de vérité terrain $y \in \{0, 1\}$. Il vise à séparer les fenêtres réelles des fenêtres synthétiques en favorisant des logits élevés pour les séquences synthétiques ($y = 1$) et faibles pour les séquences réelles ($y = 0$). Dans notre implémentation, ce terme est calculé à l'aide de la fonction `BCEWithLogitsLoss` de PyTorch, qui combine l'application de la fonction sigmoïde et le calcul de l'entropie croisée binaire de manière numériquement stable.

Terme de lissage géométrique Le second terme est appliqué à la séquence de descripteurs géométriques $\{g_t\}_{t=1}^{T}$ transmise au module de fusion, où $g_t \in \mathbb{R}^{D_g}$ représente le descripteur géométrique associé au pas de temps t. Pour chaque paire de pas de temps consécutifs $(t-1, t)$, la distance euclidienne au carré

$$\|g_t - g_{t-1}\|_2^2 \tag{3.27}$$

mesure la variation de la représentation géométrique entre deux images successives. Le terme de lissage est alors défini par

$$\mathcal{L}_{\text{liss}} = \frac{1}{T-1} \sum_{t=2}^{T} \|g_t - g_{t-1}\|_2^2 \,. \tag{3.28}$$

Ce coût reste faible lorsque les descripteurs géométriques évoluent de manière continue au cours du temps. À l'inverse, il augmente lorsque la représentation présente des variations brusques ou des discontinuités entre deux images consécutives.

Ce régularisateur tient compte de la nature temporelle des mouvements des lèvres. Bien que ces mouvements soient généralement continus, les trajectoires des points de repère peuvent être perturbées par des erreurs de suivi, des occlusions partielles ou le flou de mouvement.

Sans régularisation, le modèle pourrait accorder une importance excessive à ces variations instables, ce qui risquerait de réduire sa capacité de généralisation.

Le terme de lissage encourage ainsi une évolution temporelle cohérente des descripteurs géométriques, tout en conservant les variations utiles à la classification. Il est appliqué uniquement aux descripteurs géométriques g_t. Les descripteurs de texture r_t ne sont pas soumis à cette contrainte, afin de permettre au modèle d'exploiter des indices visuels locaux, tels que les contours des dents, les frontières de fusion et les artefacts de compression.

Rôle du facteur de pondération λ : Le coefficient λ contrôle le compromis entre l'exactitude de classification et la régularité temporelle. Si λ était trop grand, l'optimiseur serait fortement encouragé à rendre tous les g_t presque constants, ce qui nuirait à la discrimination. Pour éviter ce comportement dégénéré, nous limitons λ au petit intervalle $[10^{-4}, 10^{-3}]$ et choisissons une valeur fixe sur l'ensemble de validation. Dans ce régime, le terme BCE domine l'optimisation et pilote la séparation entre réel et synthétique, tandis que le terme de lissage réduit surtout les fluctuations non informatives dans l'espace géométrique. Autrement dit, λ met en œuvre une contrainte *souple* : il décourage les changements violents d'une image à l'autre, mais ne force pas les différentes séquences (ou classes) à partager le même plongement. Dans toutes les expériences, λ est traité comme un hyperparamètre fixe une fois sélectionné.

Extracteurs de caractéristiques gelés : Une fois l'entraînement terminé, les branches de géométrie, de texture de la bouche et de texture du visage peuvent être réutilisées comme extracteurs de caractéristiques gelés. Pour l'ablation géométrique seule, nous conservons l'encodeur `GeometryCNNWithAttention` préentraîné gelé et n'entraînons qu'un MLP peu profond au-dessus, en utilisant le terme de classification sans régularisateur de lissage lorsque les descripteurs géométriques eux-mêmes ne sont plus mis à jour. Dans le modèle complet à trois branches, la séquence géométrique préentraînée peut encore être réutilisée comme représentation d'entrée fixe, tandis que la projection ajoutée, le raffinement BiLSTM, les branches d'apparence et les couches de fusion sont entraînés conjointement. De même, les branches de texture de la bouche et du visage peuvent être utilisées soit comme extracteurs de caractéristiques gelés indépendants, soit comme composantes optimisées conjointement dans le modèle complet. Dans tous les régimes d'inférence, aucune perte n'est appliquée : le réseau produit simplement des plongements géométriques g_t, des plongements de texture de la bouche r_t^{mouth}, des plongements de texture du visage r_t^{face}, des caractéristiques fusionnées z_t et des probabilités synthétiques calibrées $p_{\text{synthétique}} = \sigma(\hat{\ell})$.

Déséquilibre des classes : Le déséquilibre des classes dans nos données d'entraînement apparaît principalement au *niveau des fenêtres*, plutôt qu'au seul niveau de la vidéo. Dans les jeux de données deepfake pilotés par le contenu, les manipulations sont généralement confinées à de courts intervalles temporels plutôt qu'à l'ensemble du clip. Par conséquent, même lorsqu'une vidéo est étiquetée synthétique, de nombreuses fenêtres glissantes extraites de celle-ci demeurent entièrement authentiques ou ne chevauchent que marginalement le segment manipulé, et se voient donc attribuer l'étiquette réelle. Ce déséquilibre découle directement du cadre de localisation temporelle : la supervision est appliquée à de courtes fenêtres, tandis que le contenu manipulé n'occupe généralement qu'une fraction limitée de la durée totale de la vidéo. En pratique, cela produit beaucoup plus de fenêtres négatives que positives. Ce phénomène est particulièrement attendu dans des jeux de données comme AV-deepfake1M++, où les intervalles falsifiés représentent en moyenne une faible proportion de la durée vidéo totale. Pour atténuer ce biais pendant l'entraînement, nous sous-échantillonnons la classe majoritaire ou utilisons un échantillonneur équilibré par classe lors de la construction des mini-lots. Ces stratégies ne modifient que la distribution d'échantillonnage des fenêtres ; elles n'altèrent pas la forme analytique de la perte dans (3.26).

Calibration des probabilités : Pour le déploiement, nous calibrons les scores de sortie par mise à l'échelle de température. Étant donné un ensemble de validation, nous apprenons une température scalaire $T_{\text{cal}} > 0$ qui minimise la log-vraisemblance négative des étiquettes lorsque les logits $\hat{\ell}/T_{\text{cal}}$ sont utilisés. Au moment du test, la probabilité synthétique finale pour une fenêtre est $p_{\text{synthétique}} = \sigma(\hat{\ell}/T_{\text{cal}})$. Selon l'application, le seuil de décision sur $p_{\text{synthétique}}$ est choisi soit au point de taux d'erreur égal (EER), afin d'équilibrer les faux positifs et les faux négatifs, soit à la valeur qui maximise le score F1 lorsqu'une plus grande sensibilité de détection est requise.

3.2.5 Considérations de complexité

Du point de vue computationnel, le coût par fenêtre est dominé par le réseau dorsal ResNet-18 [11] appliqué à de petites ROI d'environ 96×96 pixels. Les convolutions géométriques et la couche d'attention temporelle évoluent quadratiquement avec la longueur de fenêtre T, mais dans notre configuration T reste modeste (quelques dizaines d'images), de sorte que ces composantes sont relativement légères. Les bi-LSTM des têtes de texture et de fusion évoluent linéairement avec T et avec la taille cachée.

Dans l'ensemble, l'architecture tient confortablement sur un seul GPU récent en précision mixte et peut traiter les fenêtres en temps réel ou quasi temps réel. Cela la rend compatible

avec des scénarios de modération en ligne ou d'analyse de réunions, où la latence et les contraintes matérielles sont importantes.

3.3 Discussion

Les deux blocs méthodologiques précédents forment une progression volontaire. Le premier bloc montre empiriquement que l'audio n'est pas uniformément fiable et que certains phonèmes sont plus révélateurs que d'autres. Le deuxième bloc fournit un détecteur *visuel* compact, bien normalisé et sensible aux masques, dont le comportement peut être expliqué et reproduit. Ensemble, ces deux composantes définissent une base solide pour une future fusion audiovisuelle, sans présenter dans cette version une architecture audio–vidéo complète comme résultat méthodologique.

CHAPITRE 4

Expérimentations et discussion

4.1 Jeux de données et protocole d'évaluation

La détection des deepfakes est, au final, un problème guidé par les données. Un détecteur ne devient utile en pratique que s'il est entraîné et évalué sur du matériel ressemblant à ce qui circule dans des conditions réalistes de déploiement : vidéos compressées en ligne, modifications locales subtiles insérées dans des enregistrements autrement authentiques, et falsifications audiovisuelles de plus en plus réalistes. Pour cette raison, le principal banc d'essai utilisé dans ce mémoire est **AV-deepfake1M++**, un banc d'essai audiovisuel à grande échelle conçu à la fois pour la détection au niveau vidéo et la localisation temporelle.

AV-deepfake1M++ peut être compris comme une extension plus difficile du banc d'essai AV-deepfake1M. Il conserve la même idée générale de manipulation *pilotée par le contenu*, où le sens d'un énoncé parlé est modifié par des changements locaux au niveau des mots, comme le *remplacement*, la *suppression* et l'*insertion*. Cependant, AV-deepfake1M++ est nettement plus difficile pour la détection : il est plus grand, repose sur plusieurs sources de vidéos réelles, inclut un ensemble plus large de générateurs audio et visuels, et introduit surtout des perturbations réalistes de redistribution et de diffusion susceptibles de supprimer ou d'imiter des indices de manipulation.

Cette section se concentre donc sur AV-deepfake1M++ comme banc d'essai principal pour les expériences rapportées dans ce chapitre. Nous résumons d'abord son échelle et ses caractéristiques générales, puis nous décrivons sa chaîne de génération, ses partitions, son protocole d'évaluation et la raison pour laquelle il est particulièrement aligné avec les objectifs de ce mémoire.

4.1.1 Vue d'ensemble du banc d'essai AV-deepfake1M++

AV-deepfake1M++ est un banc d'essai audiovisuel de deepfakes à grande échelle, introduit comme extension d'AV-deepfake1M, avec un accent beaucoup plus fort sur le réalisme. Selon sa description officielle, il contient **2,051,154** clips vidéo, dont **627,936** échantillons réels et **1,423,218** échantillons synthétiques, pour un total d'environ **4,655.9** heures de vidéo

provenant de **7,109** sujets.

Une force importante d'AV-deepfake1M++ est qu'il ne se limite pas à une seule source de vidéos réelles. Il combine plutôt du matériel provenant de **VoxCeleb2**, **LRS3** et **EngageNet**, ce qui augmente la diversité des conditions d'enregistrement, des styles de parole et des contextes d'interaction. Cet aspect est important, car les détecteurs entraînés sur une seule source tendent souvent à se surajuster aux caractéristiques propres à cette source plutôt qu'à apprendre des indices robustes de manipulation.

Une autre différence importante est qu'AV-deepfake1M++ est explicitement conçu pour être plus difficile qu'AV-deepfake1M. Le banc d'essai augmente la diversité des sources, la diversité des générateurs et la difficulté d'évaluation, tout en introduisant des perturbations audio et visuelles réalistes. Il constitue donc un cadre plus exigeant pour évaluer la robustesse de détecteurs en conditions proches du déploiement.

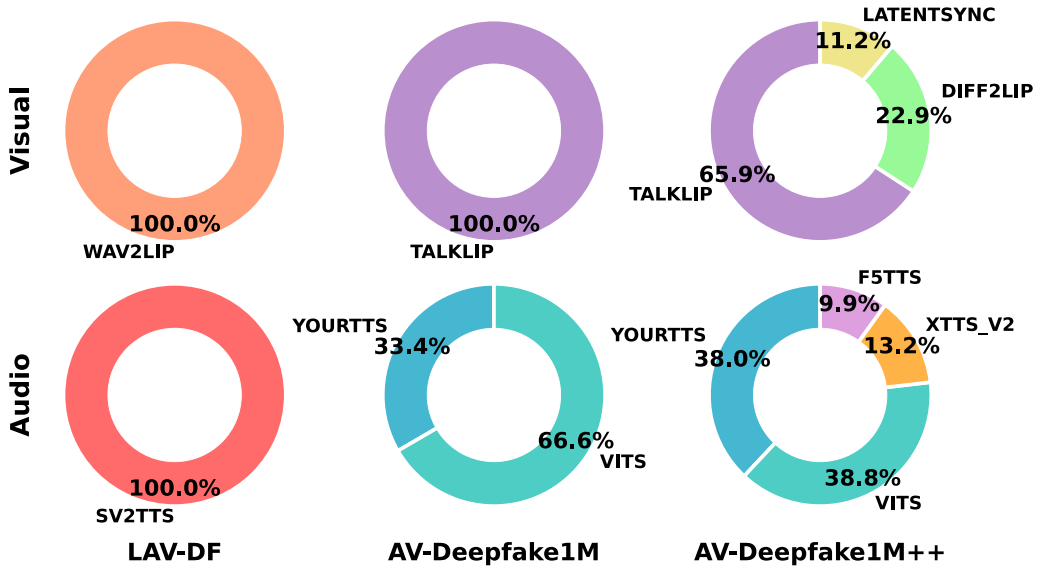


FIGURE 4.1 – Comparaison de LAV-DF, AV-deepfake1M et AV-deepfake1M++ en termes de diversité des méthodes de génération de deepfakes. La première rangée montre les proportions des méthodes de génération visuelle et la deuxième rangée montre les proportions des méthodes de génération audio (adapté de [12]).

La figure 4.1 illustre l'une des principales raisons pour lesquelles AV-deepfake1M++ est plus difficile que les bancs d'essai antérieurs. Comparativement à LAV-DF et à la version originale d'AV-deepfake1M, il repose sur un ensemble plus large de générateurs, à la fois pour l'audio et pour le visuel, ce qui réduit le risque que les détecteurs mémorisent simplement les artefacts d'une seule chaîne de synthèse.

4.1.2 Chaîne de génération pilotée par le contenu

AV-deepfake1M++ suit une chaîne en plusieurs étapes pour générer des falsifications audiovisuelles pilotées par le contenu. La structure générale reprend la conception originale d'AV-deepfake1M, mais l'étend par des sources supplémentaires de vidéos réelles, davantage de systèmes de synthèse et une étape de perturbation appliquée après la génération de la falsification.

Étape 1 : planification de la manipulation : Pour chaque vidéo réelle, la transcription de la parole est d'abord obtenue par reconnaissance automatique de la parole. Un grand modèle de langage est ensuite utilisé pour générer des modifications sémantiques au niveau des jetons, destinées à inverser ou à altérer le sens de l'énoncé. Ces modifications prennent la forme d'opérations de **remplacement**, de **suppression** et d'**insertion**, renvoyées dans un format structuré précisant les mots modifiés et leurs positions.

Étape 2 : génération audio : La transcription manipulée est reconvertie en parole après séparation de la parole et du bruit de fond. AV-deepfake1M++ [12] utilise un ensemble plus large de systèmes de synthèse vocale, notamment **VITS**, **YourTTS**, **F5TTS** et **XTTSv2**. Pour le remplacement et l'insertion, la chaîne peut générer la phrase modifiée complète et découper le segment pertinent, ou synthétiser uniquement l'intervalle du mot inséré ou remplacé.

Étape 3 : génération visuelle : Le flux visuel est ensuite manipulé au moyen de plusieurs modèles de synchronisation labiale plutôt que d'un seul générateur visuel. Dans AV-deepfake1M++ [12], les générateurs visuels incluent **TalkLip**, **LatentSync** et **Diff2Lip**. L'audio manipulé guide la génération visuelle synchronisée avec les lèvres, tout en préservant la pose générale et l'apparence du locuteur.

Étape 4 : post-traitement par perturbations : Un ajout majeur d'AV-deepfake1M++ est l'étape explicite de perturbation appliquée après la génération de la falsification. Cette étape simule des distorsions réalistes couramment observées dans les vidéos redistribuées en ligne, notamment le flou, le bruit, la réduction du débit binaire, la variation de la fréquence d'images, les pertes d'images, l'écrêtage audio, la réverbération, les interférences et les effets de bégaiement.

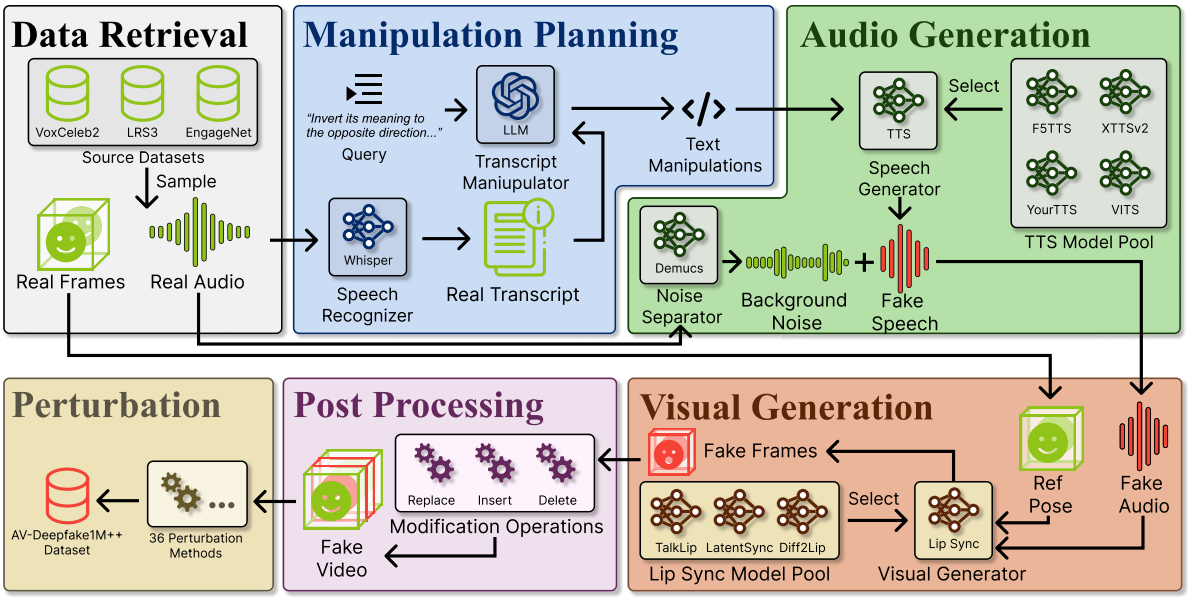


FIGURE 4.2 – Chaîne de génération des données d'AV-deepfake1M++ [12]. Le banc d'essai étend la conception originale d'AV-deepfake1M en combinant plusieurs sources de vidéos réelles, plusieurs générateurs audio et visuels ainsi qu'une étape de post-traitement par perturbations (adapté de [12]).

4.1.3 Structure du jeu de données, partitions et statistiques

AV-deepfake1M++ est divisé en sous-ensembles **entraînement**, **validation**, **TestA** et **TestB**. Cette division est plus exigeante que la structure entraînement/validation/test plus simple d'AV-deepfake1M, car elle vise à évaluer à la fois l'apprentissage dans le même domaine et la robustesse inter-domaines sous différentes identités, sources de données réelles, chaînes de génération et configurations de perturbation.

TABLEAU 4.1 – Nombre officiel de sujets et de vidéos dans AV-deepfake1M++ [12].

Sous-ensemble	#Vidéos	#Réels	#Synthétiques	#Images	Durée (heures)	#Sujets
Entraînement	1,099,217	297,389	801,828	264,053,153	2,444.9	2,606
Validation	77,326	20,220	57,106	18,488,518	171.2	
TestA	828,318	287,517	540,801	208,290,429	1,928.6	4,503
TestB	46,293	22,810	23,483	12,012,810	111.2	
Total	2,051,154	627,936	1,423,218	502,844,910	4,655.9	7,109

Au-delà de son échelle, AV-deepfake1M++ est aussi structuré pour augmenter la difficulté de plusieurs façons. Il utilise plusieurs sources de vidéos réelles, mélange plusieurs générateurs

audio et visuels et applique différents calendriers de perturbation selon les sous-ensembles. Par conséquent, de bonnes performances sur des jeux de données antérieurs ne se transfèrent pas automatiquement à ce banc d'essai.

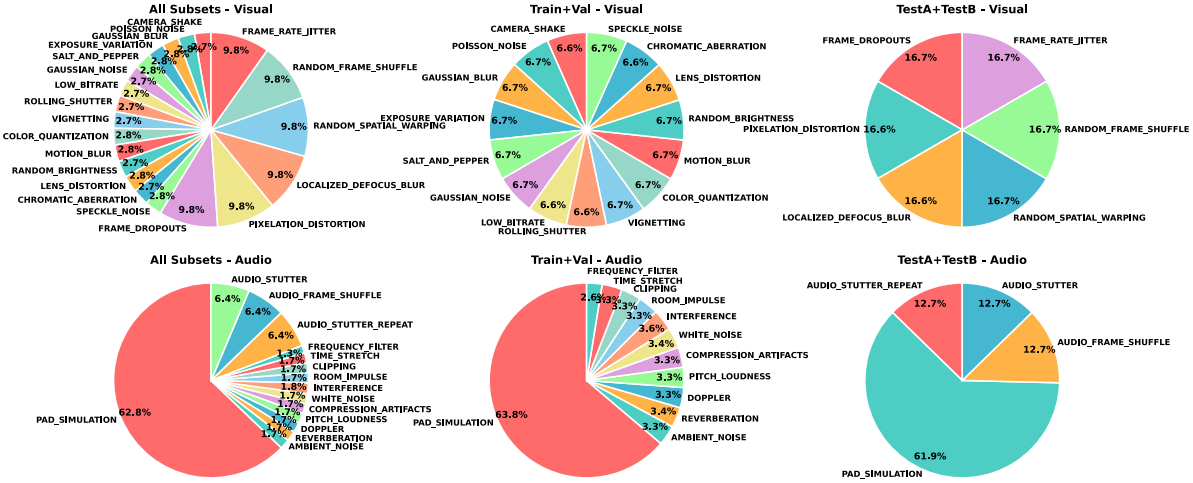


FIGURE 4.3 – Distribution des perturbations audio et visuelles dans AV-deepfake1M++. La première rangée montre les perturbations visuelles et la deuxième rangée montre les perturbations audio. Le banc d'essai applique des perturbations réalistes à l'ensemble du jeu de données et de ses sous-ensembles, ce qui augmente la difficulté forensique dans des conditions de redistribution plus réalistes (adapté de [12]).

4.1.4 Protocole de référence

AV-Deepfake1M++ constitue le jeu de données de référence du *1M-Deepfakes Detection Challenge 2025*. Dans ce cadre, les méthodes sont entraînées sur la partition officielle d'entraînement, puis évaluées sur les partitions **TestA** et **TestB**. Le protocole comprend deux tâches principales.

1. **Détection au niveau de la vidéo :** à partir d'un clip audiovisuel, le modèle doit déterminer si la vidéo contient au moins un segment manipulé. La performance est mesurée au moyen de l'aire sous la courbe ROC (AUC).
2. **Localisation temporelle :** le modèle doit identifier les intervalles temporels contenant une manipulation. L'évaluation repose sur la précision moyenne calculée à plusieurs seuils d'intersection sur union (IoU) :

$$\text{AP@}\{0{,}50;\ 0{,}75;\ 0{,}90;\ 0{,}95\},$$

ainsi que sur le rappel moyen calculé pour plusieurs nombres de propositions :

AR@{5; 10; 20; 30; 50}.

Ce protocole est particulièrement pertinent pour notre travail, car il distingue la détection globale d'une vidéo de la localisation temporelle précise des segments manipulés.

4.1.5 Pertinence pour notre plan expérimental

AV-Deepfake1M++ est particulièrement adapté aux objectifs de ce mémoire pour plusieurs raisons.

— Ses manipulations sont **guidées par le contenu** : la modification de quelques mots peut changer le sens d'un énoncé tout en laissant la majeure partie du clip inchangée.

— Le jeu de données fournit des **annotations temporelles**, nécessaires à l'entraînement par fenêtres et à la localisation précise des segments manipulés.

— Il regroupe, dans un même cadre, des manipulations **audio**, **visuelles** et **audiovisuelles**.

— La diversité de ses sources réelles, de ses méthodes de synthèse et de ses perturbations permet une évaluation plus réaliste de la robustesse et de la capacité de généralisation des détecteurs.

Dans la suite de ce chapitre, ce jeu de données constitue le principal support des expériences visuelles. Les résultats de fusion précoce sont rapportés sur des sous-ensembles correspondant aux trois opérations considérées dans AV-Deepfake1M++ : l'insertion (*insert*), la suppression (*delete*) et le remplacement (*replace*).

4.2 Vue d'ensemble des expérimentations

Ce chapitre présente l'entraînement et l'évaluation des deux composantes introduites au chapitre 3. Les expériences sont organisées sous la forme de deux études complémentaires :

1. une étude audio fondée sur HuBERT, consacrée à l'analyse des différences entre la parole réelle et la parole synthétique aux niveaux phonétique et lexical ;
2. une étude visuelle fondée sur la fusion précoce des informations géométriques et texturales autour de la bouche, dans laquelle plusieurs configurations de détection des deepfakes sont évaluées.

La fusion audiovisuelle complète n'est pas évaluée expérimentalement dans la présente version du mémoire. Elle est toutefois abordée dans les perspectives de recherche et pourra faire l'objet d'un travail scientifique ultérieur.

4.3 Expérience 1 : analyse phonétique avec HuBERT

4.3.1 Objectif

L'objectif de cette expérience est de vérifier, sur des données réelles, l'hypothèse introduite au chapitre 3 : *les deepfakes audio ne sont pas uniformément imparfaits*. Certains phonèmes, certains mots fonctionnels courts et certaines chaînes de synthèse (TTS vs VC) produisent des représentations qui s'écartent davantage de la parole naturelle que d'autres. Si nous pouvons *mesurer* cet écart au niveau de l'unité, nous pouvons alors prioriser les segments audio les plus informatifs dans une future extension audiovisuelle et justifier pourquoi un détecteur pourrait accorder davantage d'attention à certains horodatages. L'expérience reprend donc l'analyse de l'étude fondée sur HuBERT menée sur une parole réelle/synthétique parallèle construite à partir de LibriSpeech et de trois systèmes de synthèse [13].

4.3.2 Données

Cette expérience repose sur un corpus *parallèle* de parole réelle et synthétique. Pour chaque énoncé provenant d'un sous-ensemble propre de LibriSpeech (16 kHz, texte aligné), nous avons généré plusieurs équivalents synthétiques à l'aide de systèmes représentatifs de synthèse vocale et de conversion de voix. L'objectif n'est pas seulement de comparer différents corpus, mais de comparer différentes *réalisations d'un même contenu linguistique*. Cela est important, car l'analyse phonétique subséquente devient interprétable : toute séparation systématique observée dans l'espace HuBERT peut être attribuée au processus de synthèse plutôt qu'à une différence de texte [13].

Plus précisément, pour chaque énoncé réel, nous avons généré trois ou quatre versions synthétiques à l'aide de systèmes représentatifs des attaques actuelles :

- **Coqui TTS** (single-speaker, end-to-end),
- **VITS TTS** (multi-speaker, high quality),
- **StarGANv2-VC** avec ParallelWaveGAN pour la conversion de voix,
- et, dans certains cas, une piste supplémentaire **VC-multi** afin de diversifier les conditions de locuteur.

Toutes les formes d'onde ont été rééchantillonnées à 16 kHz, converties en mono 16 bits et alignées de force au moyen de MFA (modèle anglais ARPA), de sorte que chaque énoncé réel et chaque équivalent synthétique soit associé à des intervalles temporels au niveau des phonèmes et des mots au format TextGrid. Cette étape d'alignement est essentielle, car les

étapes suivantes de l'expérience opèrent au niveau des phonèmes et des mots plutôt qu'au seul niveau de l'énoncé.

Le tableau 4.2 résume les statistiques de base de ce corpus. Il indique, pour chaque sous-ensemble, le nombre de fichiers audio ainsi que la durée minimale, moyenne et maximale des énoncés. Ces valeurs remplissent deux fonctions. Premièrement, elles confirment que le corpus est assez grand pour soutenir des comparaisons distributionnelles entre phonèmes, plutôt que des comparaisons anecdotiques fondées sur quelques exemples. Deuxièmement, elles montrent que les systèmes synthétiques ne préservent pas exactement la temporalité de la même manière, ce qui est en soi pertinent pour l'analyse des deepfakes.

TABLEAU 4.2 – Statistiques de la parole réelle/synthétique parallèle utilisée dans l'expérience 1.

Jeu de données	#Fichiers	Min. (s)	Moy. (s)	Max. (s)
LibriSpeech (réel)	28,539	1.41	12.69	24.52
VITS TTS	28,539	0.77	11.24	24.90
VC Multi	28,538	1.40	12.68	24.50
Coqui TTS	28,539	1.21	12.18	116.67

Plusieurs observations découlent du tableau 4.2. Premièrement, le nombre de fichiers est presque identique entre les sous-ensembles, ce qui confirme la conception parallèle du corpus et garantit une comparaison équitable entre la parole réelle et la parole synthétique. Deuxièmement, les durées moyennes restent dans le même ordre de grandeur, ce qui indique que les systèmes produisent des énoncés de longueur globale comparable. En même temps, les statistiques de durée ne correspondent pas parfaitement : certains systèmes synthétiques génèrent des énoncés légèrement plus courts ou plus longs que les enregistrements réels correspondants, et Coqui TTS présente une durée maximale particulièrement élevée. Cela indique que la synthèse peut introduire des irrégularités temporelles ou une dérive occasionnelle de durée, ce qui est pertinent, car ces effets peuvent influencer la coarticulation, la prosodie et les frontières phonémiques.

Du point de vue de cette expérience, ces statistiques justifient l'utilisation de l'alignement forcé et du regroupement au niveau des phonèmes. Comme la durée n'est pas strictement préservée entre les systèmes, une comparaison directe image par image serait peu fiable. En alignant linguistiquement les signaux, puis en regroupant les plongements à l'intérieur des intervalles phonémiques, nous obtenons une représentation mieux adaptée à la comparaison de la *réalisation* d'un phonème, plutôt que de sa simple *durée*. En ce sens, le tableau 4.2

n'est pas seulement descriptif ; il motive aussi les choix méthodologiques adoptés dans la sous-section suivante [13].

4.3.3 Extraction de caractéristiques

Chaque forme d'onde réelle et synthétique a été transmise à un modèle HuBERT Base gelé afin d'obtenir des plongements SSL au pas de 20 ms. En parallèle, les frontières MFA nous fournissent les intervalles phonémiques $[t_s, t_e]$. Nous avons ensuite associé les caractéristiques par trame aux caractéristiques phonémiques par un simple regroupement moyen :

$$\phi(p) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} h_t, \tag{4.1}$$

Ainsi, chaque occurrence de phonème dans chaque énoncé, et pour chaque système, devient un vecteur de taille fixe $\phi(p) \in \mathbb{R}^D$.

Nous obtenons ainsi, pour *chaque type de phonème* c, deux nuages empiriques :

$$\mathcal{R}_c = \{\phi_1^{\text{réel}}, \ldots, \phi_{N_c}^{\text{réel}}\}, \qquad \mathcal{S}_c = \{\phi_1^{\text{synthétique}}, \ldots, \phi_{M_c}^{\text{synthétique}}\}, \tag{4.2}$$

L'un pour la parole réelle et l'autre pour la parole synthétique. Il s'agit de la même chaîne d'alignement que celle illustrée à la figure 4.4 [13].

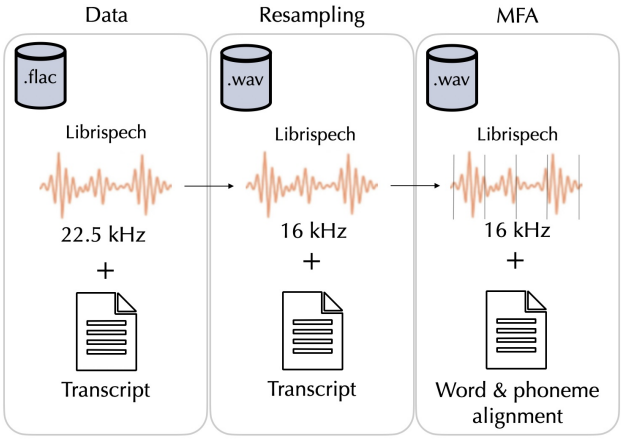


FIGURE 4.4 – Chaîne d'alignement et de regroupement pour l'expérience 1 : plongements HuBERT (20 ms) + intervalles phonémiques MFA → vecteurs de phonèmes $\phi(p)$ pour la parole réelle et synthétique.

4.3.4 Estimation de la divergence et classement

Pour savoir *quels* phonèmes changent le plus lorsqu'ils sont synthétisés, nous avons suivi l'analyse de divergence de l'étude originale et ajusté une distribution gaussienne à chaque nuage :

$$\mathcal{R}_c \sim \mathcal{N}(\mu_c^{\text{réel}}, \Sigma_c^{\text{réel}}), \qquad \mathcal{S}_c \sim \mathcal{N}(\mu_c^{\text{synthétique}}, \Sigma_c^{\text{synthétique}}), \tag{4.3}$$

Nous avons ensuite calculé la divergence KL

$$D_c = D_{\text{KL}}\big(\mathcal{N}(\mu_c^{\text{réel}}, \Sigma_c^{\text{réel}}) \,\|\, \mathcal{N}(\mu_c^{\text{synthétique}}, \Sigma_c^{\text{synthétique}})\big). \tag{4.4}$$

Enfin, nous avons classé les phonèmes selon D_c séparément pour chaque synthétiseur (VITS TTS, VC Multi, Coqui), car chaque système possède ses propres faiblesses. Les tableaux 4.3 et 4.4 ci-dessous correspondent directement aux analyses des voyelles et des consonnes rapportées dans l'article SMC 2025 ; nous les reproduisons ici afin que la personne lectrice de ce mémoire n'ait pas à consulter un autre document [13].

Tableau 4.3 – Divergence KL (KLD) et exactitude de classification (ACC) pour les voyelles dans les trois systèmes de synthèse. Une KLD élevée indique une voyelle plus discriminante.

	VITS TTS		VC Multi		Coqui	
Phonème	KLD	ACC	KLD	ACC	KLD	ACC
/OY/	4.22	89.2%	1.20	76.3%	2.68	76.2%
/AW/	3.40	92.0%	0.81	79.1%	1.85	75.1%
/EY/	2.61	85.0%	0.60	71.1%	1.52	79.5%
/AY/	2.54	92.5%	0.86	70.5%	1.22	77.2%
...	...	...	...	...	...	...

TABLEAU 4.4 – Divergence KL (KLD) et exactitude de classification (ACC) pour les consonnes dans les trois systèmes de synthèse. Les consonnes sont globalement moins séparables que les voyelles.

	VITS TTS		VC Multi		Coqui	
Phonème	KLD	ACC	KLD	ACC	KLD	ACC
ZH	2.79	50.0%	0.63	60.0%	1.61	46.7%
SH	2.69	84.6%	0.26	61.6%	1.59	75.0%
S	2.27	89.4%	0.23	65.2%	2.10	87.3%
...	...	...	...	...	...	...

Deux observations apparaissent immédiatement : (i) les *voyelles* sont systématiquement plus affectées que les consonnes, en particulier pour VITS et VC Multi, et (ii) le *classement n'est pas universel*. Coqui est plus difficile à modéliser ; par conséquent, les détecteurs entraînés uniquement sur des faux de type VITS surestimeront leur confiance sur des faux de type Coqui. C'est pourquoi nous insistons plus loin sur l'entraînement avec *plusieurs* générateurs audio [13].

4.3.5 Expérience de classification

Pour vérifier si ces divergences ont réellement un effet en pratique, nous avons entraîné un petit MLP par phonème. Chaque MLP reçoit en entrée (1) la moyenne de *tous* les phonèmes de l'énoncé et (2) la moyenne des k phonèmes les plus divergents pour le modèle de synthèse considéré. L'utilisation de (1) seulement donne un taux d'erreur égal (EER) d'environ 10% ; l'ajout de (2) réduit l'EER à environ 6% sur la partition de test parallèle, ce qui confirme l'intuition suivante : *tous les phones ne se valent pas*, et l'on gagne plusieurs points en se concentrant sur les unités les plus problématiques.

Nous avons également répété l'analyse au niveau des *mots* pour les mots-outils et les pronoms courts. Des mots comme *himself*, *again* et *themselves* présentent une KLD et une exactitude élevées, ce qui signifie que notre détecteur AV peut accorder davantage de confiance à l'audio dans ces régions [13].

TABLEAU 4.5 – KLD et ACC au niveau des mots-outils (extrait). Des valeurs élevées indiquent que la synthèse éprouve des difficultés avec ces mots.

Mot-outil	KLD (VITS)	ACC (VITS)	KLD (VC-Multi)	ACC (VC-Multi)
himself	6.69	96.1%	3.81	95.7%
again	5.92	97.7%	2.37	83.8%
themselves	6.04	93.0%	2.63	86.0%
same	5.54	94.0%	2.77	75.8%
...	...	...	...	...

4.3.6 Visualisation

Afin de rendre cette analyse lisible pour des personnes non spécialistes de la parole, nous avons projeté les plongements phonémiques HuBERT en 2D avec t-SNE, comme dans les figures 4.5 et 4.6 [13] : un panneau pour une voyelle (/OY/) et un autre pour un mot-outil discriminant (*himself*). Les deux nuages (réel vs synthétique) se séparent clairement pour les unités à forte KLD, mais se superposent pour les unités à faible KLD.

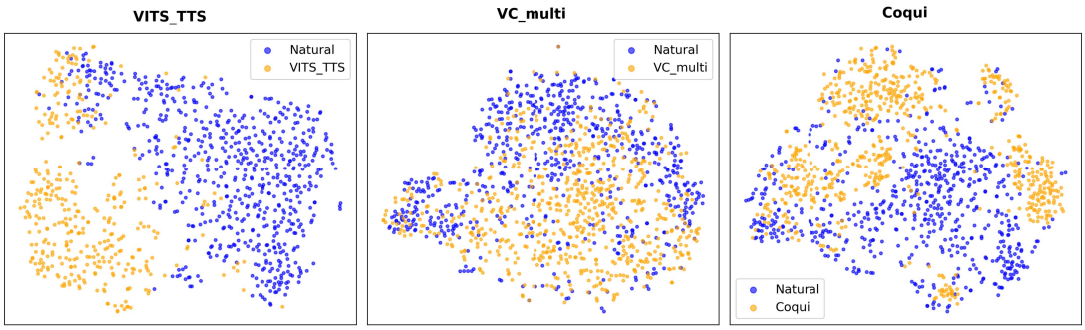


FIGURE 4.5 – t-SNE des plongements HuBERT pour la voyelle /OY/ (réel vs synthétique). Une séparation nette indique une forte discriminabilité [13].

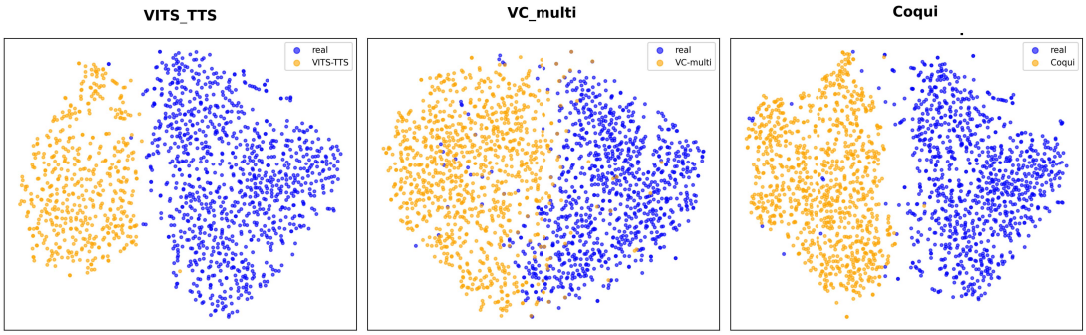


FIGURE 4.6 – t-SNE des plongements HuBERT pour le mot-outil *himself*. Les instances synthétiques se regroupent à distance des instances réelles pour les systèmes VITS et VC [13].

4.3.7 Discussion

Cette expérience est plus limitée que le détecteur visuel de l'expérience 2, mais elle fournit deux résultats exploitables :

1. **La sélectivité au niveau des phonèmes est réelle :** les voyelles et certains mots fonctionnels expressifs présentent une KLD beaucoup plus élevée et une exactitude par unité nettement meilleure que les consonnes (tableaux 4.3 et 4.5). Il est donc raisonnable de transmettre cette information à une future fusion audio–vidéo sous forme de poids par fenêtre.
2. **Le système de synthèse compte :** Coqui est constamment le plus difficile à séparer (plus faible R^2 entre KLD et ACC), ce qui signifie qu'une évaluation inter-jeux de données contenant de l'audio de type Coqui sera moins performante si l'entraînement s'est fait uniquement sur de l'audio de type VITS. Cela correspond exactement au problème de changement de distribution mentionné pour les défis ASVspoof et ADD.[13]

Au chapitre 3, nous avons indiqué que le détecteur A/V « écoute plus attentivement » lorsqu'un phonème à forte divergence est présent. Après cette expérience, nous pouvons justifier ce choix de conception : sur ce corpus parallèle, l'utilisation des k phonèmes les plus divergents réduit l'EER d'environ 10% à 6%, ce qui constitue un gain non négligeable pour un composant qui sera ensuite intégré à un détecteur AV fonctionnant par fenêtres glissantes.

4.4 Expérience 2 : fusion précoce visuelle géométrie-texture

4.4.1 Données

Les données visuelles proviennent du jeu de données AV-Deepfake1M++ [12], conçu pour l'étude des deepfakes audiovisuels à grande échelle. Bien que ce corpus contienne des manipulations audio et vidéo, seule sa composante visuelle a été exploitée dans cette expérience. La partition officielle d'entraînement a été utilisée pour l'apprentissage du modèle, tandis que la partition officielle de validation, également utilisée dans le cadre de la compétition, a servi à la sélection et à l'évaluation des configurations.

Aucune sous-sélection manuelle fondée sur l'identité des locuteurs ou sur le type de manipulation n'a été effectuée. Cette stratégie permet de conserver la diversité des sujets et des conditions présentes dans les partitions officielles, et ainsi de favoriser l'évaluation de la capacité de généralisation du modèle.

Lors du prétraitement, les vidéos ont été analysées automatiquement à l'aide de MediaPipe FaceMesh [10]. Le filtrage a été effectué au niveau des fenêtres temporelles et non au niveau des vidéos ou des sujets. Une fenêtre était considérée comme valide lorsque le visage du locuteur était suffisamment visible et qu'au moins $P = 40$ points de repère des lèvres étaient détectés sur la majorité de ses images. Les fenêtres ne satisfaisant pas ces conditions ont été écartées afin de garantir la fiabilité des informations géométriques transmises au modèle.

Après ce filtrage, les séquences ont été traitées à une fréquence de 25 images par seconde afin d'assurer la cohérence temporelle entre les images vidéo et les points de repère extraits. Pour chaque image au temps t, deux représentations visuelles ont été conservées :

— l'image faciale complète

$$I_t \in [0,1]^{H_0 \times W_0 \times 3},$$

où H_0 et W_0 désignent respectivement la hauteur et la largeur de l'image ;

— une région d'intérêt centrée sur la bouche

$$R_t \in [0,1]^{H \times W \times 3},$$

utilisée pour l'analyse locale de la texture. Dans certaines configurations, cette région a été élargie par un facteur de 1,3 afin d'inclure davantage de contexte facial, puis redimensionnée à une résolution de 96×96 pixels.

Ces deux représentations permettent d'étudier conjointement les artefacts visuels locaux autour de la bouche et les incohérences plus larges susceptibles d'apparaître dans le visage.

Afin d'améliorer la robustesse du modèle aux dégradations couramment introduites par les plateformes numériques, plusieurs augmentations visuelles ont été appliquées aux régions d'intérêt pendant l'entraînement. Elles comprennent un réencodage JPEG aléatoire avec une qualité comprise entre 40 et 90, un léger flou gaussien, des variations aléatoires de couleur ainsi que des retournements horizontaux occasionnels lorsque leur application ne modifiait pas une information textuelle pertinente.

4.4.2 Configurations des modèles et implémentation

Les expériences visuelles ont été conduites de manière progressive, en passant d'une configuration de référence initiale à des configurations de fusion de plus en plus riches. Dans le carnet final, trois étapes expérimentales pratiques peuvent être distinguées.

Premièrement, une série d'expériences préliminaires à exécution unique a été réalisée sur une version plus ancienne et plus petite du jeu de données. Ces essais ont surtout servi de référence exploratoire et ont permis de valider la faisabilité de la chaîne visuelle proposée avant l'introduction d'une validation croisée plus systématique.

Deuxièmement, le modèle visuel principal a été implémenté comme une **architecture de fusion précoce à deux branches**, combinant :

— une **branche géométrique** fondée sur des points de repère des lèvres normalisés par SE(2) et leurs dérivées temporelles, notés $X_{1:T}^{(6)} \in \mathbb{R}^{T\times P\times 6}$.

— une **branche de texture** fondée sur une ROI serrée des lèvres, centrée sur la bouche et extraite de chaque image.

Dans cette configuration, la séquence de points de repère est traitée par l'encodeur `GeometryCNNWithAttenti` décrit au chapitre 3, tandis que la séquence de ROI est traitée par un réseau dorsal ResNet-18 [11] suivi d'un LSTM bidirectionnel à une couche. Les deux descripteurs sont ensuite fusionnés à chaque pas de temps et transmis à un module de fusion temporelle pour la classification.

Troisièmement, un **modèle de fusion visuelle à trois branches** a été introduit en ajoutant une **branche ROI du visage** plus large à la configuration précédente à deux branches. Cette extension vise à capter des artefacts contextuels qui peuvent ne pas être entièrement visibles dans un recadrage serré de la bouche, en particulier pour des types de manipulation plus subtils comme *delete* et *replace*. La configuration finale à trois branches combine donc :

$$x6 \ + \ \text{lips ROI} \ + \ \text{ROI du visage.}$$

L'implémentation a été réalisée en PyTorch 2.x avec prise en charge CUDA. L'entraînement en précision mixte (`autocast` et `GradScaler`) a été utilisé pour réduire la consommation

mémoire, et toutes les entrées ont été organisées sous forme de fenêtres $[B, T, \cdot]$ compatibles avec l'inférence en flux continu.

4.4.3 Configuration de l'entraînement

Nous avons optimisé l'objectif suivant :

$$\mathcal{L} = \mathrm{BCEWithLogits}(\hat{\ell}, y) + \lambda \frac{1}{T-1} \sum_{t=2}^{T} \|g_t - g_{t-1}\|_2^2, \tag{4.5}$$

avec $\lambda = 5 \times 10^{-4}$. La régularisation de lissage temporel a été appliquée uniquement aux descripteurs géométriques g_t, afin de ne pas supprimer les artefacts de texture de haute fréquence, comme les frontières des dents, la fusion des lèvres ou les incohérences à l'intérieur de la bouche.

L'entraînement a utilisé AdamW avec une décroissance de poids de 10^{-4}, une phase de réchauffement de 5 époques et une décroissance cosinus du taux d'apprentissage. La taille de lot était typiquement $B = 48$ fenêtres sur un GPU de 24 Go et $B = 32$ sur un GPU de 12 Go. Un écrêtage du gradient à 1.0 a été utilisé pour stabiliser l'optimisation des couches récurrentes.

La sélection du modèle reposait sur la conservation du meilleur point de contrôle selon l'AUC de validation. Après chaque époque, l'AUC-ROC était calculée sur l'ensemble de validation, et le point de contrôle ayant l'AUC de validation la plus élevée était retenu.

4.4.4 Protocole d'évaluation

L'évaluation a été effectuée au niveau des fenêtres. Nous rapportons :

— **AUC**,

— **EER**,

calculés uniquement sur les fenêtres valides. De plus, lorsque cela était possible, les prédictions au niveau de la vidéo ont été obtenues en agrégeant les scores de fenêtres par une moyenne pondérée par le masque. La robustesse a aussi été examinée sous des conditions visuelles dégradées, telles que la recompression JPEG et la suppression simulée de points de repère.

4.4.5 Résultats

Les résultats de l'expérience 4.4 sont présentés comme une progression entre plusieurs variantes du modèle visuel. L'objectif n'est pas seulement de rapporter une valeur d'AUC

TABLEAU 4.6 – Comparaison globale des variantes du modèle visuel en termes d'AUC. La référence géométrique seule est évaluée sur une seule partition de validation, tandis que les modèles de fusion sont rapportés avec une validation croisée à 5 plis.

Modèle	**Flux / caractéristiques**	**Protocole**	**AUC (%)**
Géométrie seule	$x^{(6)}$: points de repère des lèvres normalisés et dérivées temporelles	Single split	92.59
Fusion précoce à deux branches	$x^{(6)}$ + ROI des lèvres	5-fold CV	94.34
Fusion précoce à trois branches (proposé)	$\mathbf{x^{(6)}}$ **+ ROI des lèvres + ROI du visage**	**5-fold CV**	**97.85 ± 0.20**

finale, mais aussi de comprendre l'effet de chaque source d'information ajoutée au modèle. Nous comparons donc une référence fondée uniquement sur la géométrie des lèvres, une fusion précoce à deux branches combinant la géométrie et la ROI des lèvres, puis notre variante finale à trois branches, qui ajoute une ROI faciale plus large.

Comparaison globale des variantes du modèle visuel Le tableau 4.6 résume les résultats principaux en termes d'AUC. La référence géométrique seule obtient déjà une performance solide avec une AUC de 92.59%. Ce résultat montre que les trajectoires des points de repère des lèvres, après normalisation et enrichissement temporel, contiennent une information discriminante importante pour la détection des manipulations faciales.

L'ajout d'une branche de texture centrée sur la ROI des lèvres améliore la performance et permet d'atteindre une AUC de 94.34%. Cette amélioration indique que la géométrie seule ne suffit pas toujours à capturer les artefacts visuels introduits par les deepfakes. L'apparence locale de la bouche apporte donc un complément utile au mouvement des lèvres.

La meilleure performance est obtenue avec le modèle à trois branches proposé. En ajoutant une ROI faciale plus large à la géométrie des lèvres et à la ROI de la bouche, le modèle atteint une AUC moyenne de 97.85% ± 0.20%. Ce résultat suggère que certains artefacts de manipulation ne sont pas strictement limités à la région des lèvres, mais peuvent également apparaître dans le contexte facial proche, notamment autour des joues, du menton ou des frontières de la bouche.

Comparaison avec des références externes Afin de situer notre résultat par rapport à des travaux récents, le tableau 4.7 présente une comparaison indicative avec des méthodes

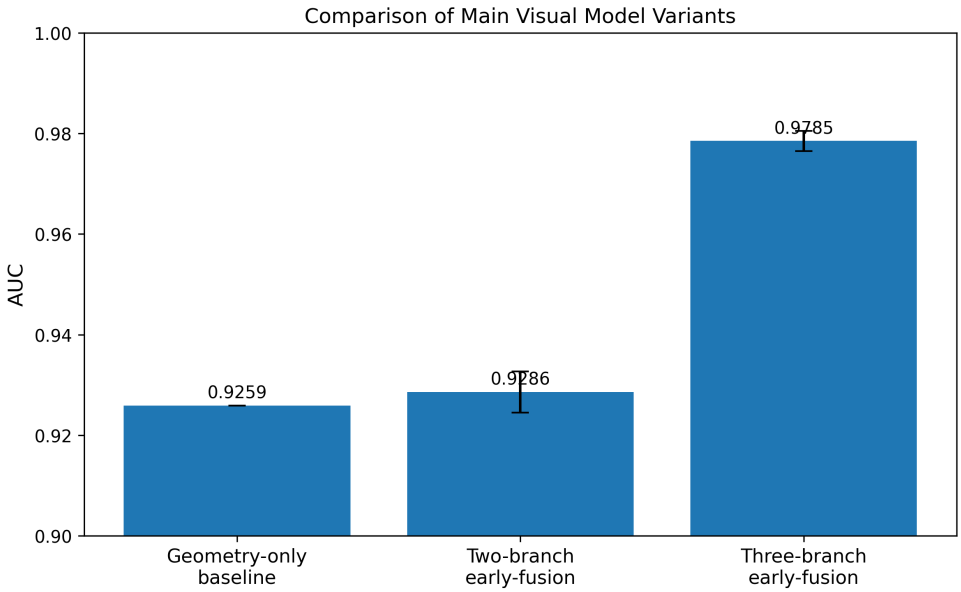


FIGURE 4.7 – Comparaison des principales variantes du modèle visuel en termes d'AUC. La figure illustre l'amélioration progressive obtenue par l'ajout de la ROI des lèvres puis de la ROI du visage.

rapportées dans la littérature. Les variantes KLASSify reposent sur un TCN alimenté par des caractéristiques visuelles artisanales et atteignent des AUC de 84.29%, 85.48% et 88.41%. Notre modèle à trois branches atteint 97.85%, ce qui représente un gain de 9.44 points d'AUC par rapport à la meilleure variante KLASSify rapportée.

Nous comparons également notre résultat avec le travail de Pindrop Labs sur AV-Deepfake1M++. Leur système rapporte une AUC de 92.49% sur le split TestA pour la tâche de classification globale, tandis que leur fusion évaluée sur les étiquettes visuelles atteint 91.98% sur l'ensemble de validation. Notre résultat de 97.85% est supérieur à ces deux valeurs. Cependant, cette comparaison doit rester prudente, car les protocoles expérimentaux, les partitions et les modalités évaluées ne sont pas identiques. De plus, les modèles visuels individuels de Pindrop fondés sur LipForensics atteignent des AUC supérieures à 99% sur les étiquettes visuelles de validation. Ainsi, notre contribution doit être présentée comme une approche visuelle compétitive et plus performante que certaines références, mais pas comme une supériorité générale sur toutes les méthodes existantes.

Comparaison par type de manipulation Pour mieux comprendre le comportement du modèle, nous avons également évalué les principales variantes selon les trois types de manipulation considérés : *insert*, *delete* et *replace*. Les résultats du tableau 4.8 montrent que les performances restent élevées dans les trois cas, avec des différences selon la nature de la manipulation.

TABLEAU 4.7 – Comparaison indicative avec des références externes en termes d'AUC. Les résultats externes sont rapportés selon leurs protocoles originaux ; la comparaison doit donc être interprétée comme un positionnement expérimental et non comme une évaluation strictement contrôlée.

Méthode	**Description**	**Protocole rapporté**	**AUC (%)**
KLASSify v1 [93]	TCN avec caractéristiques visuelles 1 et 2	Validation	84.29
KLASSify v2 [93]	TCN avec caractéristiques visuelles 1, 2 et 3	Validation	85.48
KLASSify v3 [93]	TCN avec caractéristiques visuelles 1, 2, 3 et 4	Validation	88.41
Pindrop Labs [94]	Fusion évaluée sur les étiquettes visuelles	Validation	91.98
Notre modèle	**Fusion précoce à trois branches : géométrie + ROI des lèvres + ROI du visage**	Validation	**93.85**

TABLEAU 4.8 – AUC par type de manipulation pour les principales variantes de fusion.

Type de manipulation	**Deux branches**	**Deux branches, texture gelée**	**Trois branches**
Insert	98.18	94.72	98.11
Delete	97.09	92.43	96.41
Replace	96.79	92.31	97.09

Les manipulations de type *insert* sont les plus faciles à détecter, car elles introduisent souvent des incohérences locales plus visibles autour de la bouche. Les manipulations *delete* et *replace* sont plus difficiles, car elles peuvent produire des modifications plus subtiles ou plus distribuées. Dans ce contexte, l'ajout de la ROI du visage reste utile, en particulier pour le cas *replace*, où le modèle à trois branches obtient la meilleure AUC.

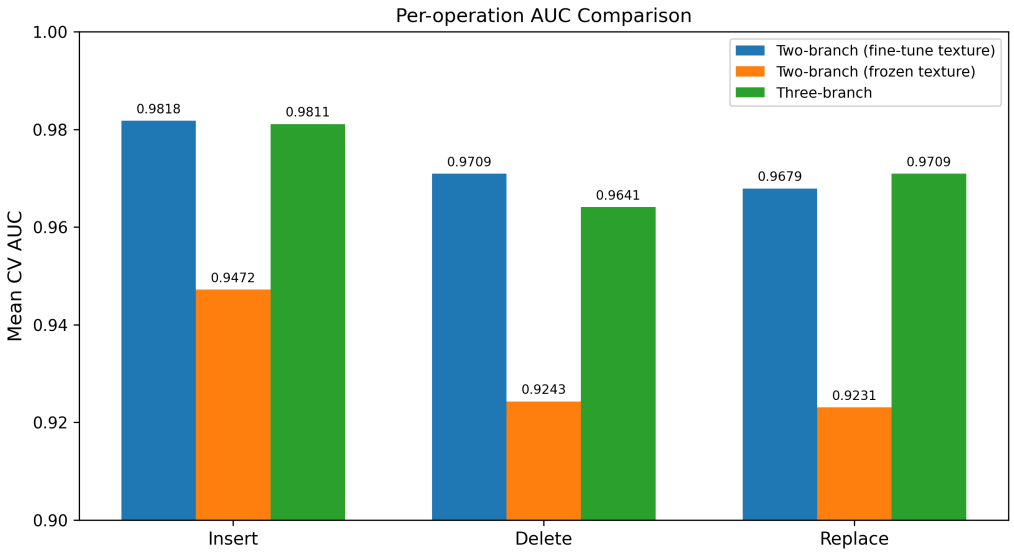


FIGURE 4.8 – Comparaison de l'AUC par type de manipulation. Le graphique montre que les modèles de fusion restent performants sur les trois opérations, avec un avantage notable de la variante à trois branches pour les manipulations de type *replace*.

Interprétation générale Pris ensemble, ces résultats montrent que la progression de la géométrie seule vers la fusion à deux branches, puis vers la fusion à trois branches, apporte une amélioration réelle et mesurable. Le flux géométrique capture la dynamique des lèvres, mais il ne décrit pas entièrement les artefacts d'apparence. La ROI des lèvres ajoute une information locale directement liée à la parole visuelle. Enfin, la ROI du visage apporte un contexte complémentaire, utile lorsque la manipulation affecte des régions voisines de la bouche ou produit des incohérences plus globales.

Ainsi, le modèle à trois branches constitue la configuration la plus complète et la plus performante de cette étude visuelle. Il reste centré sur la bouche, mais il ne limite pas l'analyse à cette seule région. Cette conception offre un bon compromis entre interprétabilité, richesse des représentations et performance de détection.

4.4.6 Courbes AUC de validation par pli

Par souci d'exhaustivité, les courbes AUC de validation par pli au fil des époques pour les principales variantes visuelles sont fournies aux figures 4.9–4.14.

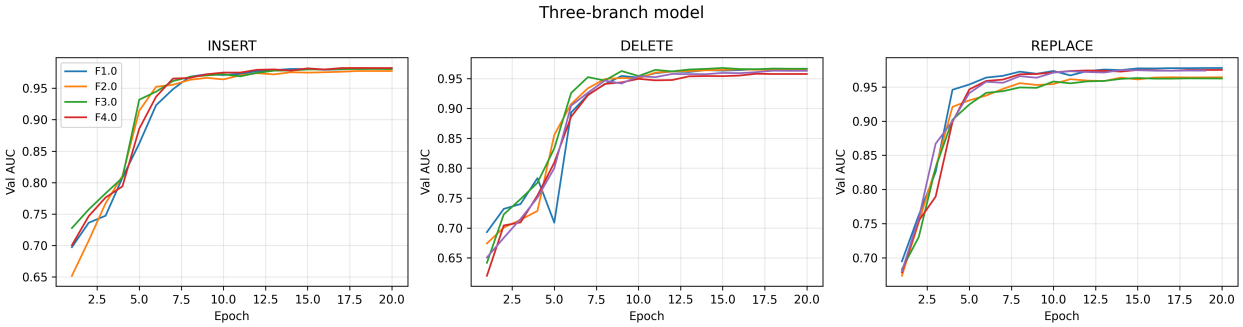


FIGURE 4.9 – AUC de validation par pli au fil des époques pour le modèle à trois branches sur les sous-ensembles insertion, suppression et remplacement.

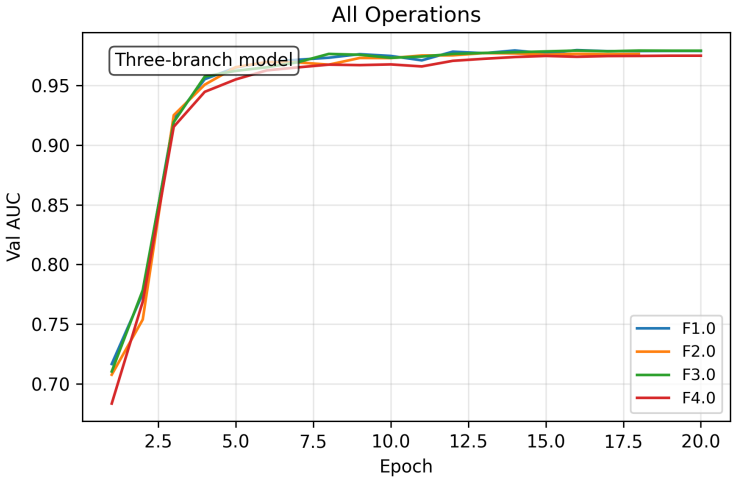


FIGURE 4.10 – AUC de validation par pli au fil des époques pour le modèle à trois branches dans la configuration toutes opérations.

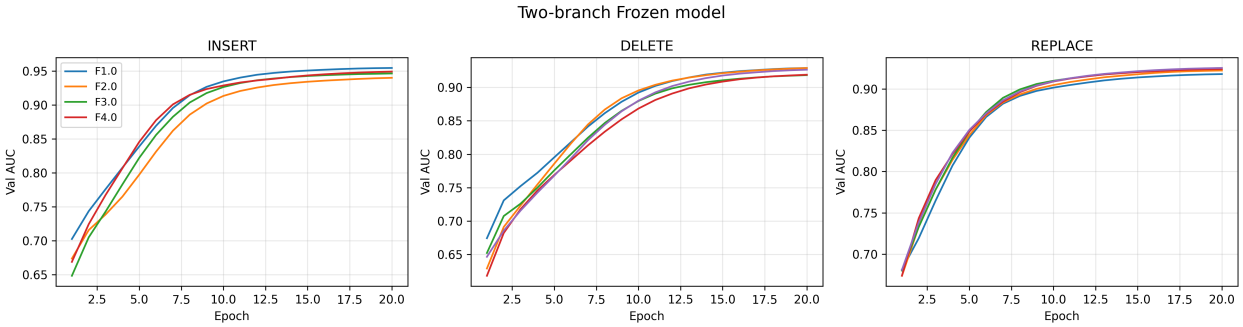


FIGURE 4.11 – AUC de validation par pli au fil des époques pour le modèle à deux branches avec texture gelée sur les sous-ensembles insertion, suppression et remplacement.

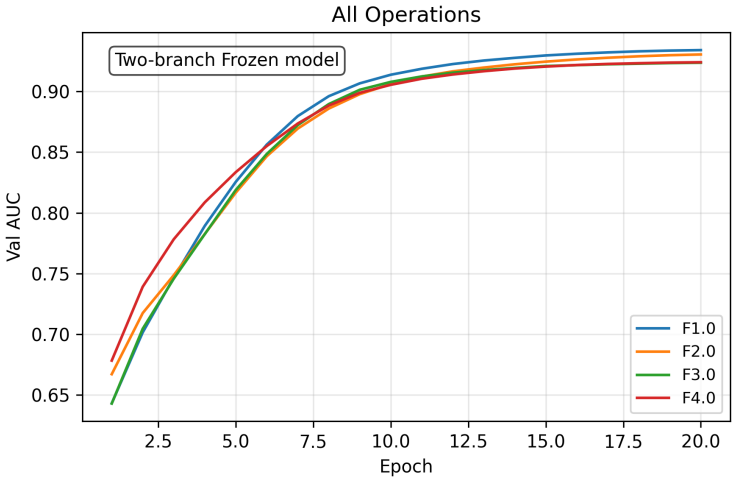


FIGURE 4.12 – AUC de validation par pli au fil des époques pour le modèle à deux branches avec texture gelée dans la configuration toutes opérations.

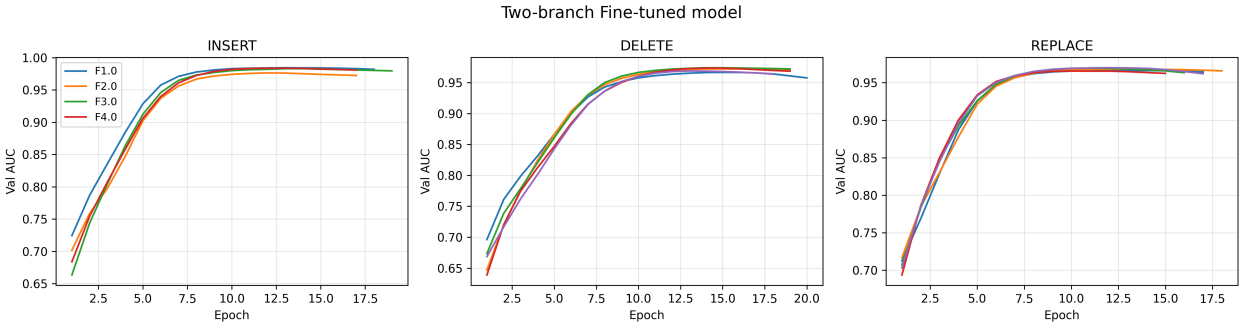


FIGURE 4.13 – AUC de validation par pli au fil des époques pour le modèle à deux branches ajusté finement sur les sous-ensembles insertion, suppression et remplacement.

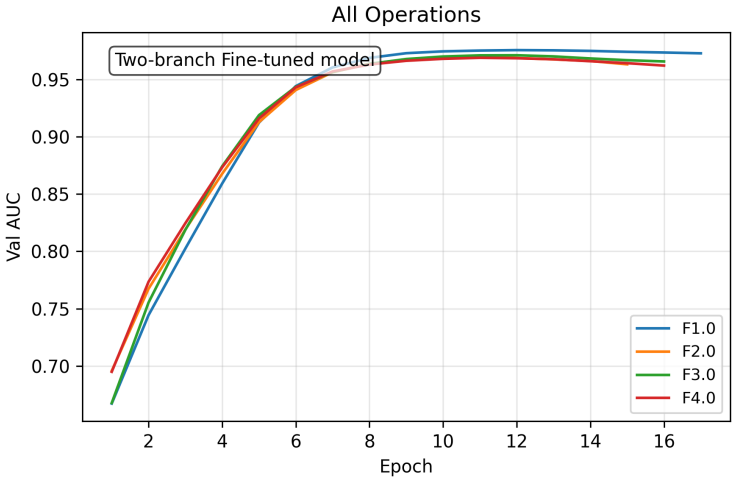


FIGURE 4.14 – AUC de validation par pli au fil des époques pour le modèle à deux branches ajusté finement dans la configuration toutes opérations.

4.4.7 Discussion

L'expérience 4.4 conduit à trois observations principales. Premièrement, la géométrie de la bouche seule fournit déjà une référence compétitive, ce qui confirme que la dynamique temporelle des lèvres reste informative même sans l'apparence brute des pixels. Deuxièmement, la combinaison de la géométrie et de la texture des lèvres par fusion précoce est bénéfique, mais sa capacité devient limitée lorsque le modèle doit généraliser à des types de manipulation hétérogènes en utilisant seulement un recadrage serré centré sur la bouche. Troisièmement, les meilleurs résultats sont obtenus en ajoutant un contexte facial plus large par un flux ROI du visage dédié.

L'écart entre les modèles à deux et à trois branches est particulièrement informatif. Dans la configuration toutes opérations, l'AUC moyenne passe de 0.9434 à 0.9785, ce qui constitue une amélioration substantielle pour un détecteur purement visuel. Au niveau par opération, le gain le plus important apparaît sur le sous-ensemble *replace*, ce qui suggère que cette manipulation laisse souvent des artefacts hors de la région stricte des lèvres, par exemple dans la fusion faciale, la cohérence d'identité ou les transitions locales de texture autour du bas du visage. Cette observation est cohérente avec l'intuition selon laquelle un détecteur limité à la bouche peut être trop restrictif pour certaines classes de falsification audiovisuelle.

L'ablation avec texture gelée clarifie aussi le rôle de l'optimisation conjointe. Pour les trois types d'opérations, geler la branche d'apparence conduit à une AUC plus faible que lorsque l'encodeur de texture peut s'adapter aux données cibles. Cela indique que le gain de performance ne provient pas seulement de l'ajout de caractéristiques, mais aussi de l'apprentissage de représentations visuelles propres à la tâche et alignées avec le flux géométrique pendant l'entraînement.

Dans l'ensemble, les résultats soutiennent l'hypothèse centrale de ce chapitre : la détection visuelle des deepfakes bénéficie de la combinaison du mouvement géométrique local et d'indices de texture complémentaires, et ce bénéfice devient plus fort lorsque le modèle peut intégrer à la fois le détail fin des lèvres et le contexte facial plus large. Le modèle de fusion précoce à trois branches fournit donc la configuration uniquement visuelle la plus fiable parmi les variantes testées, et constitue le réseau dorsal visuel approprié pour les extensions multimodales discutées plus loin dans le manuscrit.

4.5 Discussion générale

À travers les deux expériences conservées dans cette version, quelques constats se dégagent :

— **La conception centrée sur la bouche fonctionne :** se concentrer sur les lèvres

MediaPipe [10] et une ROI de 96×96 suffit à atteindre une AUC élevée, ce qui confirme les observations de Koutlis et al. [16] et Cai et al. [91].

— **MediaPipe est suffisant :** nous avons volontairement remplacé RetinaFace par MediaPipe FaceMesh pour le suivi des lèvres ; avec le filtre de couverture et la normalisation SE(2), les performances sont restées élevées et la chaîne est devenue plus facile à reproduire sur Colab ou sur des dispositifs à faible coût.

— **La fusion précoce est importante :** chaque fois que nous avons retardé la fusion, le détecteur est devenu légèrement moins sensible aux mouvements de bouche courts et incohérents. C'est précisément ce que nous voulions vérifier.

— **L'audio n'est pas uniforme :** l'analyse HuBERT l'a rendu explicite et nous a fourni une manière fondée de surpondérer certaines fenêtres audio.

Des limites demeurent : les vidéos très peu éclairées, les fortes rotations de tête et les locuteurs avec une pilosité faciale importante peuvent encore compromettre les points de repère des lèvres ; dans ces cas, un détecteur sensible à l'identité ou fondé sur le visage complet devrait prendre le relais.

4.6 Conclusion

Ce chapitre a documenté, de manière claire et reproductible, les deux expériences conservées dans cette version du mémoire. L'expérience 1 a montré que les modèles de parole auto-supervisés, en particulier HuBERT, peuvent révéler quels phonèmes sont les plus affectés par la synthèse, et que cette information est directement exploitable. L'expérience 2 a établi qu'un détecteur compact, centré sur la bouche et à fusion précoce, utilisant ResNet-18 [11] et MediaPipe [10], peut déjà égaler ou approcher l'état de l'art en détection visuelle des deepfakes. La fusion audio–vidéo complète est donc replacée dans les perspectives de recherche et pourra faire l'objet d'un article scientifique futur.

CHAPITRE 5

Conclusion générale et perspectives

5.1 Conclusion générale

Ce mémoire s'inscrit dans le domaine de la détection des deepfakes audio et visuels, dont l'importance s'est renforcée avec les progrès rapides des modèles génératifs. L'objectif principal était d'étudier des représentations capables de mettre en évidence des indices fins de manipulation dans la parole synthétique et dans la région visuelle de la bouche.

Pour répondre à cette problématique, le travail a été organisé autour de deux composantes complémentaires. La première porte sur l'analyse audio de la parole réelle et synthétique à partir de représentations auto-supervisées extraites par HuBERT. Cette étude a permis d'examiner les divergences observées au niveau des phonèmes et de déterminer les unités linguistiques les plus sensibles aux systèmes de synthèse vocale.

La seconde composante concerne la détection visuelle des manipulations faciales. Une architecture de fusion précoce a été développée afin de combiner les informations géométriques issues des points de repère des lèvres avec les caractéristiques de texture extraites autour de la bouche et du visage. Cette combinaison permet de représenter conjointement la forme, le mouvement et l'apparence locale des régions susceptibles de contenir des artefacts de manipulation.

Les résultats obtenus montrent l'intérêt de ces deux directions. L'analyse audio indique que certaines unités phonétiques présentent des écarts plus marqués entre la parole réelle et la parole synthétique. Du côté visuel, la combinaison de la géométrie des lèvres et des informations de texture améliore la détection des manipulations, notamment lorsque les artefacts apparaissent autour de la bouche ou dans le contexte facial proche.

Ces travaux constituent ainsi deux bases complémentaires pour le développement ultérieur d'un système audiovisuel unifié. La fusion audio–vidéo complète n'est toutefois pas présentée comme une contribution expérimentale finalisée dans ce mémoire. Elle est considérée comme une perspective de recherche nécessitant une validation spécifique.

5.2 Synthèse des contributions

Contribution 1 : analyse phonétique de la parole synthétique avec HuBERT

La première contribution repose sur l'utilisation de représentations auto-supervisées de la parole pour analyser les différences entre des énoncés réels et synthétiques. À partir de plongements HuBERT et d'un alignement phonétique, les distributions des phonèmes issus de la parole naturelle ont été comparées à celles produites par différents systèmes de synthèse vocale.

Cette analyse montre que les phonèmes ne sont pas tous affectés de la même manière par la synthèse. Certaines voyelles et certains mots courts présentent des divergences plus importantes et peuvent ainsi constituer des indices pertinents pour la détection de la parole synthétique. Cette approche apporte également une dimension interprétable, puisqu'elle relie la décision du modèle à des unités linguistiques précises.

Cette contribution a donné lieu à une publication dans les actes de l'*IEEE International Conference on Systems, Man, and Cybernetics 2025* [13].

Contribution 2 : détecteur visuel à fusion précoce géométrie–texture

La deuxième contribution concerne le développement d'un détecteur visuel centré sur la région de la bouche. Le modèle combine les descripteurs géométriques des lèvres, obtenus à partir de points de repère normalisés et de leurs variations temporelles, avec des descripteurs de texture extraits de régions d'intérêt centrées sur la bouche et le visage.

La fusion de ces sources d'information permet au modèle d'analyser simultanément la dynamique des lèvres, leur apparence locale et le contexte facial. Les résultats expérimentaux montrent que cette complémentarité améliore la détection des manipulations visuelles par rapport à une représentation fondée uniquement sur la géométrie.

Les résultats de cette seconde contribution constituent la base d'un manuscrit scientifique actuellement en préparation, consacré à la détection visuelle des deepfakes par fusion des informations géométriques et texturales.

5.3 Limites du travail

Malgré les résultats obtenus, plusieurs limites doivent être considérées.

Premièrement, la composante visuelle dépend de la qualité du suivi des points de repère

des lèvres. Les variations importantes de pose, les occlusions, le flou de mouvement et les conditions d'éclairage difficiles peuvent réduire la fiabilité des caractéristiques géométriques.

Deuxièmement, l'analyse audio dépend de la qualité du signal et de la précision de l'alignement phonétique. Le bruit de fond, la compression audio et les erreurs d'alignement peuvent masquer certaines différences entre la parole réelle et la parole synthétique.

Troisièmement, les expériences ont été réalisées sur un ensemble déterminé de systèmes de synthèse et de conditions expérimentales. La généralisation à de nouveaux générateurs, à d'autres langues et à des conditions d'enregistrement différentes devra être étudiée plus largement.

Enfin, les composantes audio et visuelle ont principalement été évaluées séparément. La complémentarité entre les indices phonétiques et visuels reste donc à valider dans une architecture audiovisuelle complète.

5.4 Perspectives

Intégration audiovisuelle

La principale perspective consiste à développer une architecture de fusion audio–vidéo combinant les représentations phonétiques issues de HuBERT avec les caractéristiques géométriques et texturales de la bouche. Une telle architecture permettrait d'étudier la cohérence entre la parole et les mouvements des lèvres, ainsi que de produire des prédictions temporelles pour localiser les segments suspects.

Cette extension nécessitera un protocole expérimental dédié, comprenant une comparaison avec des méthodes audiovisuelles récentes et une évaluation de la détection, de la localisation temporelle et du coût d'inférence.

Extension des données et des systèmes de synthèse

Une deuxième perspective concerne l'élargissement des données expérimentales. Il serait pertinent d'intégrer davantage de locuteurs, d'accents, de langues et de conditions d'enregistrement. L'évaluation sur des manipulations produites par des générateurs plus récents permettrait également de mieux mesurer la robustesse du modèle face à des méthodes de synthèse inconnues pendant l'entraînement.

Robustesse et efficacité

La robustesse du système pourrait être améliorée face aux dégradations réelles, notamment la compression vidéo, le bruit audio, les occlusions, le flou de mouvement et les variations d'éclairage. Des architectures plus légères pourraient également être étudiées afin de réduire le temps d'inférence et de faciliter le déploiement sur des équipements disposant de ressources limitées.

Explicabilité des décisions

Enfin, l'explicabilité constitue une perspective importante. Du côté visuel, des méthodes telles que Grad-CAM pourraient aider à identifier les régions ayant le plus influencé la décision du détecteur. Du côté audio, la visualisation des phonèmes et des mots présentant les divergences les plus élevées permettrait de mieux comprendre les prédictions produites par le modèle.

En définitive, ce mémoire propose deux contributions complémentaires : une analyse phonétique interprétable de la parole synthétique et un détecteur visuel fondé sur la fusion de la géométrie et de la texture. Leur intégration future dans un cadre audiovisuel unifié représente la continuité naturelle de ces travaux.

Bibliographie

[1] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. URL `https://openaccess.thecvf.com/content_CVPRW_2020/papers/w40/Dolhansky_The_Deepfake_Detection_Challenge_Dataset_CVPRW_2020_paper.pdf`.

[2] Bill Kromydas. Convolutional neural network (cnn) : A complete guide. `https://learnopencv.com/understanding-convolutional-neural-networks-cnn/`, jan 2023. LearnOpenCV, accessed March 29, 2026.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8) :1735–1780, 1997.

[4] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11) :2673–2681, 1997.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://arxiv.org/abs/2010.11929`.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer : Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[7] Aakash Varma Nadimpalli and Ajita Rattani. *GBDF : Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection*, pages 320–337. 08 2023. ISBN 978-3-031-37741-9. doi:10.1007/978-3-031-37742-6_25.

[8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++ : Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*,

2019. URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf.

[9] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df : A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf.

[10] Google Research, Google Health, and Google AI. Mediapipe face mesh : A 3d facial landmark detector with 468 landmarks. White-paper / Model card, 2021. URL https://storage.googleapis.com/mediapipe-assets/Model%20Card%20MediaPipe%20Face%20Mesh%20V2.pdf.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. Av-deepfake1m++ : A large-scale audio-visual deepfake benchmark with real-world perturbations. *arXiv preprint arXiv :2507.20579*, 2025.

[13] Dia Elhak Temmar, Assia Hamadene, Vamshi Nallaguntla, Aishwarya Fursule, Mohand Saïd Allili, Shruti Kshirsagar, and Anderson R. Avila. Phonetic analysis of real and synthetic speech using HuBERT embeddings : Perspectives for deepfake detection. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 86–91. IEEE, 2025. doi:10.1109/SMC58881.2025.11343334. URL https://doi.org/10.1109/SMC58881.2025.11343334.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

[15] Ziyi Chang, George Alex Koulieris, Hyung Jin Chang, and Hubert P. H. Shum. On the design fundamentals of diffusion models : A survey. *Pattern Recognition*, 169 :111934, 2026.

[16] Christos Koutlis, Ioannis Chronopoulos, Alexandros Stergiou, Dimitrios Zarpalas, and Petros Daras. AV-Deepfake1M : A large-scale audio–visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024. doi:10.1145/3746027.3755563.

[17] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. Deepfakes : Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154 :113368, 2023.

[18] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, 11 :144497–144529, 2023.

[19] Anastasios Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie : A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL `https://openaccess.thecvf.com/content/CVPR2021/papers/Haliassos_Lips_Dont_Lie_A_Generalisable_and_Robust_Approach_to_Face_CVPR_2021_paper.pdf`.

[20] Peng Wu, Wanshun Su, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, and Yanning Zhang. AVadCLIP : Audio–visual collaboration for robust video anomaly detection. *arXiv preprint arXiv :2504.04495*, 2025. URL `https://arxiv.org/abs/2504.04495`.

[21] Luisa Verdoliva. Media forensics and deepfakes : An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5) :910–932, 2020.

[22] Yifei Fan, Modan Xie, Peihan Wu, and Gang Yang. Real-time deepfake system for live streaming. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, page 202–205, 2022.

[23] Chenfeng Feng, Zhiwei Chen, and Andrew Owens. Self-supervised video forensics by audio–visual anomaly detection, 2023. Conference venue not specified ; convert to @inproceedings if known.

[24] Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Zhe Peng, Qian Wang, and Chao Shen. A survey of defenses against ai-generated visual media : Detection, disruption, and authentication. *ACM Computing Surveys*, 58(5) :1–35, 2025.

[25] AA Bekheet and A Ghoneim. A comprehensive comparative analysis of deepfake detection techniques in visual, audio, and audio-visual domains. *IEEE*, 2024. URL https://ieeexplore.ieee.org/abstract/document/10652683.

[26] WA Jbara and NAHK Hussein. Deepfake detection in video and audio clips : a comprehensive survey and analysis. *Mesopotamian Journal of Cybersecurity*, 2024. URL https://mesopotamian.press/journals/index.php/CyberSecurity/article/view/675.

[27] D Tan, Y Yang, C Niu, et al. A review of deep learning based multimodal forgery detection for video and audio. *Discover Applied Sciences*, 2025. URL https://link.springer.com/article/10.1007/s42452-025-07629-3.

[28] HH Nguyen-Le, VT Tran, DT Nguyen, and NA Le-Khac. Deepfake detection across image, video, and audio : A comprehensive survey with empirical evaluation of generalization and robustness. *ResearchSquare*, 2025. URL https://www.researchsquare.com/article/rs-7753326/latest.

[29] AA Bekheet, AS Ghoneim, and G Khoriba. Unmasking the digital deception : A comprehensive survey of large vision models (lvms) for deepfake detection. *EKB*, 2025. URL https://fcihib.journals.ekb.eg/article_428110_9735fd0ae51e9f03f17c250fe9a794fe.pdf.

[30] A Sar, S Roy, T Choudhury, and A Abraham. Zero-shot visual deepfake detection : Can ai predict and prevent fake content before it's created ? *arXiv preprint arXiv :2509.18461*, 2025. URL https://arxiv.org/abs/2509.18461.

[31] Sami Belguesmia, Mohand Saïd Allili, and Assia Hamadene. Unmasking facial deepfakes : A robust multiview detection framework for natural images. *CoRR*, abs/2510.15576, 2025. doi:10.48550/arXiv.2510.15576. URL https://arxiv.org/abs/2510.15576.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[34] Assia Hamadene and Mohand Saïd Allili. Cross-model deepfake detection through contourlet-based inter-channel spectral analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2026. doi:10.1109/TBIOM.2026.3687469.

[35] Mingxing Tan and Quoc V. Le. Efficientnet : Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[38] Hangbo Bao, Li Dong, and Furu Wei. Beit : Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[39] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[40] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding ? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL `https://openaccess.thecvf.com/content/CVPR2021/papers/Bertasius_Is_Space-Time_Attention_All_You_Need_for_Video_Understanding_CVPR_2021_paper.pdf`.

[41] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit : A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL `https://openaccess.thecvf.com/content/ICCV2021/papers/Arnab_ViViT_A_Video_Vision_Transformer_ICCV_2021_paper.pdf`.

[42] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[43] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[45] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022.

[46] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786) :504–507, 2006.

[47] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face : Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016.

[48] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 2014.

[49] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[50] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[51] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[53] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[54] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[56] Hany Farid. Image forgery detection. *IEEE Signal Processing Magazine*, 26(2) :16–25, 2009.

[57] Hany Farid. Exposing digital forgeries from JPEG ghosts. *IEEE Transactions on Information Forensics and Security*, 4(1) :154–160, 2009.

[58] Jan Lukás, Jessica Fridrich, and Miroslav Goljan. Detecting digital image forgeries using sensor pattern noise. In *Proceedings of SPIE, Electronic Imaging : Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, pages 362–372, 2006.

[59] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet : a compact facial video forgery detection network. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. doi:10.1109/WIFS.2018.8630761.

[60] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[61] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016.

[62] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, 2018.

[63] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[64] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0 : A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL `https://openaccess.thecvf.com/content_CVPR_2020/papers/Jiang_DeeperForensics-1.0_A_Large-Scale_Dataset_for_Real-World_Face_Forgery_Detection_CVPR_2020_paper.pdf`.

[65] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb : A novel audio-video multimodal deepfake dataset. In *NeurIPS 2021 Datasets and Benchmarks Track*, 2021. arXiv :2108.05080.

[66] Christos Koutlis, Ioannis Chronopoulos, Alexandros Stergiou, Dimitrios Zarpalas, and Petros Daras. Av-deepfake1m : A large-scale audio-visual deepfake dataset. In *CVPR Workshops*, 2023.

[67] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, 2018. doi:10.1109/CVPR.2018.00677.

[68] Yuchen Wu, Peng Wang, Yinjie Lei, Haifeng Sun, and Ming Liu. XD-Violence : Cross-dataset transfer learning for violence detection using audio–visual streams. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2020. URL `https://www.ecva.net/papers/eccv_2020/workshops/w30/papers/08.pdf`.

[69] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[70] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net : Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17395–17405, June 2024.

[71] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1133–1143, June 2024.

[72] Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17627–17637, 2024. doi:10.1109/CVPR52733.2024.01669.

[73] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play : Video-level blending and spatiotemporal adapter tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12615–12625, 2025. doi:10.1109/CVPR52734.2025.01177.

[74] Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10786–10796, October 2025.

[75] Taehoon Kim, Jongwook Choi, Yonghyun Jeong, Haeun Noh, Jaejun Yoo, Seungryul Baek, and Jongwon Choi. Beyond spatial frequency : Pixel-wise temporal frequency-based deepfake video detection, 2025. URL `https://arxiv.org/abs/2507.02398`.

[76] Jee-weon Jung, Hee-Soo Kim, Hyung-il Shim, and Ha-Jin Yu. AASIST : Audio anti-spoofing using integrated spectro–temporal graph attention networks. In *INTERSPEECH*, 2021. URL `https://arxiv.org/abs/2104.01279`.

[77] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2, 2021. URL `https://arxiv.org/abs/2011.01108`.

[78] Xiaohui Liu, Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, and Jianwu Dang. Leveraging positional-related local-global dependency for synthetic speech detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[79] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pages 537–541. ISCA, September 2024.

doi:10.21437/interspeech.2024-659. URL http://dx.doi.org/10.21437/Interspeech.2024-659.

[80] Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. Post-training for deepfake speech detection. *arXiv preprint arXiv :2506.21090*, 2025.

[81] Joon Son Chung and Andrew Zisserman. Out of time : Automated lip sync in the wild. In *Asian Conference on Computer Vision (ACCV)*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16/chung16.pdf.

[82] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff : Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27102–27112, June 2024.

[83] Vinaya Sree Katamneni and Ajita Rattani. Contextual cross-modal attention for audio-visual deepfake detection and localization. *arXiv preprint arXiv :2408.01532*, 2024.

[84] Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13826–13836, October 2025.

[85] Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of clip in generalizable deepfake detection. *arXiv preprint arXiv :2503.19683*, 2025.

[86] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. Asvspoof 5 : crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8, 2024. doi:10.21437/ASVspoof.2024-1.

[87] Xinqi Xiong, Prakrut Patel, Qingyuan Fan, Amisha Wadhwa, Sarathy Selvam, Xiao Guo, Luchao Qi, Xiaoming Liu, and Roni Sengupta. Talkingheadbench : A multi-modal benchmark & analysis of talking-head deepfake detection. *arXiv preprint arXiv :2505.24866*, 2025.

[88] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics : Multi-modal interpretable forged face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–116, June 2025.

[89] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19576–19586, 2025. doi:10.1109/CVPR52734.2025.01823.

[90] Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18815–18825, June 2025.

[91] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m : A large-scale llm-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7414–7423, 2024. doi:10.1145/3664647.3680795.

[92] Hassan Tariq Khalid, Sanghoon Woo, Suha Kwak, and In So Kweon Kim. FakeAVceleb : A novel audio–video multimodal deepfake dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. URL `https://bmvc2021-virtualconference.com/assets/papers/0153.pdf`.

[93] Ivan Kukanov and Jun Wah Ng. Klassify to verify : Audio-visual deepfake detection using ssl-based audio and handcrafted visual features, 2025. URL `https://arxiv.org/abs/2508.07337`.

[94] Nicholas Klein, Hemlata Tak, James Fullwood, Krishna Regmi, Leonidas Spinoulas, Ganesh Sivaraman, Tianxiang Chen, and Elie Khoury. Pindrop it ! audio and visual deepfake countermeasures for robust detection and fine grained-localization, 2025. URL `https://arxiv.org/abs/2508.08141`.