UQAC
Université du Québec
à Chicoutimi

AV-ID LITE : UNE APPROCHE MULTIMODALE LÉGÈRE POUR LA DÉTECTION D'INCIDENTS DANS LES MAGASINS SANS PERSONNEL

PAR MAMADOU THIAW

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE ÈS SCIENCES (M.SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

Les approches multimodales audio-visuelles connaissent un essor important en raison de leur capacité à analyser simultanément des scènes complexes en exploitant à la fois les informations visuelles et audio. Dans le secteur du commerce de détail, en particulier dans les magasins sans personnel, la détection automatique des comportements anormaux constitue un défi majeur pour garantir la sécurité et prévenir des incidents tels que les bagarres ou les évanouissements dans des environnements fonctionnant en continu.

Dans ce travail, nous proposons AV-ID Lite, une approche multimodale légère et efficace pour la détection d'incidents dans les magasins sans personnel. L'approche proposée repose sur l'extraction de caractéristiques visuelles à partir de squelettes humains à l'aide d'AlphaPose et de caractéristiques audio à l'aide de VGGish, qui sont fusionnées selon une stratégie de fusion précoce strictement synchronisée. Les représentations fusionnées sont ensuite analysées par une architecture Transformer légère, combinée à un mécanisme de pooling Log-Sum-Exp afin de mettre en évidence les segments temporels les plus discriminants.

Nous développons également un nouveau jeu de données multimodal dédié aux magasins sans personnel, composé de vidéos annotées en trois classes : activités normales, bagarres et évanouissements. Les expériences montrent que l'approche proposée surpasse systématiquement plusieurs méthodes de l'état de l'art, atteignant un score F1 de 95.91%, un rappel de 98.22% et un mAP de 99.23%, ce qui correspond à des améliorations d'au moins 6.42%, 10.22% et 4.08% respectivement, confirmant ainsi l'efficacité et la pertinence de l'architecture proposée pour la détection multimodale d'incidents dans les environnements commerciaux sans personnel.

ABSTRACT

Multimodal audio-visual approaches are experiencing significant growth due to their ability to simultaneously analyze complex scenes by exploiting both visual and audio information. In the retail sector, particularly in unmanned stores, automatic detection of abnormal behavior is a major challenge in ensuring safety and preventing incidents such as fights or fainting in environments that operate continuously.

In this work, we propose AV-ID Lite, a lightweight and efficient multimodal approach for incident detection in unmanned retail stores. The proposed approach relies on the extraction of visual features from skeletons using AlphaPose and audio features using VGGish, which are fused by a strictly synchronized early fusion strategy. The fused representations are then analyzed by a lightweight Transformer architecture, combined with a Log-Sum-Exp pooling mechanism to highlight the most discriminative temporal segments.

We also develop a new multimodal dataset dedicated to unmanned stores, consisting of videos annotated in three classes : normal activities, fights, and fainting. Experiments show that the proposed approach consistently outperforms several state-of-the-art methods, achieving an F1-score of 95.91%, a recall of 98.22%, and a mAP of 99.23%, representing improvements of at least 6.42%, 10.22%, and 4.08%, respectively, and confirming the effectiveness and relevance of the proposed architecture for multimodal incident detection in unmanned commercial environments.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

ADL Activités normales ou activités normales de la vie quotidienne
AP Average Precision
AUC Area Under the ROC Curve
AVE Average Pooling
AV-ID Lite Audio-Visual Incident Detection Lite
CLIP Contrastive Language-Image Pretraining
CMU Carnegie Mellon University
CNN Convolutional Neural Networks
DL Deep Learning
FN Faux négatifs
FP Faux positifs
GCN Graph Convolutional Network
GPU Graphics Processing Unit
GRU Gated Recurrent Unit
HRNet High Resolution Network
I3D Inflated 3D ConvNet
KHZ Kilohertz
KNN k-Nearest Neighbors
Le2i Laboratoire Électronique, Informatique et Image
LSE Log-Sum-Exp
LSTM Long Short-Term Memory
mAP mean Average Precision
MCC Matthews Correlation Coefficient
MLP Multi-Layer Perceptron
PAF Part Affinity Fields
ResNet Residual Network
RF Random Forest
RGB Red Green Blue
RNN Recurrent Neural Network
RWF Real World Fighting Dataset
STARSS23 Sony TAu Realistic Spatial Soundscapes 2023
SVM Support Vector Machine
TN Vrais négatifs
TP Vrais positifs

ViT-B Vision Transformer Base
ViViT Video Vision Transformer
XD-Violence eXtreme-Domain Violence Dataset
YOLO You Only Look Once

DÉDICACE

Je dédie ce mémoire à la mémoire de ma chère maman, dont les prières continuent de m'accompagner, et à mon père, pour son amour, son soutien et ses sacrifices constants.

Je le dédie également à ma famille, en particulier au Professeur Clément Roger Kouly Tine et à son épouse, le Docteur Sigapale Ndong, qui m'ont accueilli comme leur propre fils et sans qui ce parcours n'aurait pas été possible.

Je dédie également ce mémoire à Maman Dorothée Diop, pour ses prières et son soutien. Que Dieu vous garde.

À Justin Tine et son épouse Isabelle, ainsi qu'à leurs enfants, pour leur soutien moral et leurs encouragements sincères.

À ma grande sœur Marie Monique Codou Tine, pour son soutien et ses encouragements constants.

À Raphaël Tine et son épouse et leurs enfants, à Bernard, Pascal et Isidore, sans oublier ma grande sœur Meïté et toute la famille Ndong.

À toute ma famille, à mes proches, à mes amis et à toutes les personnes de mon village Keur WACK, ce travail est aussi le vôtre.

Dédié au Professeur Clément Roger Kouly Tine.

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude à ma codirectrice de recherche, la **Professeure Haïfa Nakouri**, pour la confiance qu'elle m'a accordée en acceptant de diriger ce travail. Ses conseils avisés, sa disponibilité constante et son encadrement rigoureux ont été déterminants pour la réussite de ce projet. Tout au long de ce travail, elle a toujours été présente pour répondre à mes questions et m'orienter. Elle m'a énormément appris dans le domaine des données et de l'apprentissage automatique, et sans elle, ce travail n'aurait jamais pu voir le jour. Je la remercie également pour le soutien financier accordé, qui a grandement facilité la réalisation de ce parcours. Grâce à elle, j'ai renforcé ma rigueur, ma discipline et mon sens du travail bien fait, des qualités fondamentales pour devenir un chercheur accompli. Merci de m'avoir aidé à découvrir mon potentiel et à dépasser mes propres limites.

J'exprime également ma sincère reconnaissance à mon directeur de recherche, le **Professeur Fehmi Jaafar**, pour son soutien constant et précieux. Son expertise, ses conseils judicieux et sa grande disponibilité ont été des atouts majeurs dans la réalisation de ce projet. Son accompagnement a joué un rôle essentiel dans la poursuite sereine de mes études. Je lui adresse mes plus sincères remerciements, car sans lui également, l'aboutissement de ce travail n'aurait pas été possible.

Je tiens à exprimer une reconnaissance toute particulière à mon mentor, mon tuteur et mon parent de cœur, le Professeur Clément Roger Kouly Tine. Il a été pour moi à la fois un père, un grand frère et un guide. Si tout cela a été possible, c'est en grande partie grâce à lui. Merci pour votre soutien inestimable, votre disponibilité constante dans les meilleurs comme dans les pires moments, et pour tout ce que vous avez fait pour moi. Les mots me manquent pour exprimer l'ampleur de ma gratitude. Je remercie Dieu de m'avoir permis de vous connaître. Que Dieu vous bénisse, vous protège, ainsi que votre épouse et toute votre famille.

Mes sincères remerciements vont également à l'ensemble des professeurs de l'UQAC, notamment Kevin Bouchard, Abdenour Bouzouane et Pascal Fortin. En arrivant au Canada, je n'avais que très peu de connaissances en apprentissage automatique et en méthodologie de recherche. Grâce à vous, je suis aujourd'hui devenu un étudiant-chercheur passionné dans ce domaine. Merci pour ces enseignements précieux.

Enfin, je remercie du fond du cœur ma famille, mes amis et toutes les personnes de mon village pour leur soutien moral et leurs encouragements constants.

AVANT-PROPOS

Ce mémoire représente l'aboutissement d'un long parcours académique, exigeant et parfois difficile, mais extrêmement riche en apprentissages. Il m'a permis d'approfondir un sujet qui me passionne profondément : l'apprentissage automatique. Grâce aux enseignements théoriques et pratiques reçus, j'ai pu développer des compétences solides dans ce domaine et mieux comprendre l'impact que ces technologies peuvent avoir sur le monde numérique.

Ce travail m'a offert l'opportunité de proposer une approche, AV-ID Lite, visant à contribuer au développement technologique des magasins sans personnel. L'objectif principal de ce mémoire est de concevoir un modèle capable de détecter à la fois les situations normales (ADL) et les situations anormales (évanouissement et bagarre) dans ces environnements, tout en garantissant le respect de la vie privée des personnes.

Mon parcours de recherche a été jalonné de défis, d'échecs, de doutes et de stress, mais aussi de découvertes, d'explorations et de réussites qui ont profondément enrichi mon expérience. Ce mémoire est le fruit de douze mois de travail intensif et de réflexion, guidés par une passion sincère pour la recherche.

Conscient que tout travail humain reste perfectible, j'accueillerai avec intérêt toute suggestion ou critique constructive, qui ne pourra que contribuer à approfondir mes connaissances et à élargir mes perspectives de recherche.

J'espère que cette contribution pourra avoir un impact positif, notamment dans le domaine du commerce de détail et des systèmes intelligents.

CHAPITRE I

INTRODUCTION

1.1 CONTEXTE ET MOTIVATION

L'essor des modèles multimodaux vision-audio constitue une avancée majeure dans le domaine de l'intelligence artificielle. En exploitant conjointement les informations visuelles et sonores, ces modèles permettent une compréhension plus fine et plus robuste des scènes complexes, ouvrant la voie à de nombreuses applications, notamment en sécurité publique, en surveillance médicale et dans les systèmes de vidéosurveillance des villes intelligentes (Jang *et al.*, 2025).

Parallèlement, le secteur du commerce de détail connaît une transformation profonde avec l'émergence rapide des magasins sans personnel (Japan, 2023). Ces environnements, conçus pour fonctionner en continu (24 h/24), nécessitent des systèmes intelligents capables de détecter automatiquement des comportements humains anormaux tels que les bagarres, les évanouissements ou d'autres incidents pouvant compromettre la sécurité des clients et des infrastructures (Liu *et al.*, 2025).

L'intégration de telles technologies dans les magasins autonomes présente de nombreux avantages, notamment l'amélioration de l'expérience d'achat, la réduction des coûts opérationnels et le renforcement de la sécurité. Plus précisément, l'intégration de systèmes intelligents de détection d'incidents permet de garantir un environnement d'achat plus sûr et plus rassurant pour les clients, tout en réduisant les coûts liés à la surveillance humaine continue (Ramachandra *et al.*, 2020). Elle contribue également à une détection rapide des incidents critiques, limitant ainsi les risques pour les personnes et les infrastructures.

Cette évolution s'inscrit également dans un contexte socio-économique marqué par des tensions persistantes sur le marché du travail, caractérisées notamment par une pénurie de main-d'œuvre, accentuée par le vieillissement de la population active et diverses contraintes structurelles. Au Canada, plusieurs entreprises signalent des difficultés de recrutement et un faible ratio chômeurs-postes vacants, révélant un resserrement important du marché du travail (Canada, 2025).

1.2 PROBLÉMATIQUE

Malgré les progrès récents réalisés en vision-audio, la détection automatique d'incidents dans les magasins sans personnel demeure un défi majeur. Les approches existantes reposent le plus souvent sur une seule modalité ou ne prennent pas suffisamment en compte les contraintes spécifiques de ces environnements, notamment la nécessité d'un traitement en temps réel et la préservation de la vie privée des individus (Liu *et al.*, 2025; Jang *et al.*, 2025).

Par ailleurs, l'intégration conjointe de l'extraction de poses (Human Pose Estimation) et de l'analyse acoustique pour la détection d'incidents dans les magasins autonomes reste encore peu explorée. La littérature actuelle manque de solutions multimodales frugales capables de concilier performance en temps réel et protection de la vie privée.

1.3 CONTRIBUTIONS

Pour pallier cette lacune et répondre à cette problématique, nous proposons AV-ID Lite, un modèle hybride multimodal léger et efficace dédié à la détection d'incidents dans les magasins sans personnel. En exploitant la complémentarité de la pose squelettique, non identifiante, et des signatures sonores, l'approche proposée garantit à la fois une faible latence et une robustesse accrue face aux occlusions visuelles classiques du commerce de détail.

L'approche repose sur l'extraction de caractéristiques visuelles squelettiques à l'aide d'AlphaPose et de caractéristiques audio à l'aide de VGGish. Les deux modalités sont ensuite combinées par une fusion précoce strictement synchronisée, puis analysées par un Transformer léger intégrant un mécanisme de pooling Log-Sum-Exp (LSE), conçu pour mettre en valeur les instants temporels les plus discriminants.

Le système vise à détecter automatiquement trois types d'événements pertinents pour les environnements commerciaux autonomes : les activités normales de la vie quotidienne (ADL), les bagarres et les évanouissements, tout en garantissant la protection de la vie privée des clients.

Les principales contributions de ce travail sont les suivantes :

1. Nous proposons, à notre connaissance, la première approche combinant AlphaPose et VGGish dans le contexte spécifique des magasins sans personnel, permettant une analyse multimodale robuste tout en assurant la protection de la vie privée grâce à l'utilisation exclusive de représentations squelettiques et de caractéristiques audio anonymisées.
2. Nous introduisons un pipeline de fusion multimodale précoce strictement synchronisé, associé à un modèle Transformer léger et à un mécanisme de pooling Log-Sum-Exp (LSE), permettant d'améliorer l'identification des événements critiques tout en maintenant une faible complexité computationnelle adaptée aux contraintes du temps réel.
3. Nous développons un nouveau jeu de données multimodal dédié aux magasins sans personnel, incluant des vidéos annotées dans trois catégories pertinentes pour les environnements commerciaux autonomes.

Ainsi, notre approche vise à répondre simultanément aux défis majeurs des magasins autonomes : la détection d'incidents en temps réel, la préservation de la vie privée des

utilisateurs et la réduction de la dépendance à la main-d'œuvre, tout en assurant un niveau de sécurité élevé.

1.4 ORGANISATION DU MANUSCRIT

Le présent document est structuré en cinq chapitres. Le chapitre 1 présente l'introduction générale. Le chapitre 2 expose les fondements théoriques et les concepts clés de l'apprentissage profond. Le chapitre 3 est consacré à l'étude des travaux connexes. Le chapitre 4 décrit en détail la méthodologie de l'approche proposée pour la détection d'incidents dans les magasins sans personnel. Enfin, le chapitre 5 se concentre sur la mise en œuvre du système et la présentation des résultats expérimentaux.

CHAPITRE II
FONDEMENTS THÉORIQUES ET CONCEPTS CLÉS

2.1 INTRODUCTION

Dans ce chapitre, nous présentons les principaux fondements théoriques et concepts clés qui sous-tendent ce travail. L'objectif est de fournir les bases nécessaires à la compréhension de l'utilisation de l'apprentissage profond dans l'analyse vidéo, ainsi que des outils et paradigmes constituant le socle de notre approche multimodale.

Nous commençons par rappeler les principes généraux de l'analyse vidéo par apprentissage profond, en mettant en évidence ses apports et ses principales limites. Nous présentons ensuite les principaux modèles de réseaux de neurones profonds, notamment les réseaux de neurones convolutifs, les réseaux récurrents, les réseaux LSTM et les Transformers. Nous abordons également les différentes stratégies de fusion multimodale, en soulignant leurs avantages et leurs limites respectives. Enfin, nous présentons en détail plusieurs méthodes d'estimation de pose humaine, telles que OpenPose, AlphaPose, BlazePose, OpenPifPaf et HRNet, qui occupent aujourd'hui une place centrale dans les approches respectueuses de la vie privée.

2.2 APPRENTISSAGE PROFOND

L'apprentissage profond (Deep Learning, DL) est une branche de l'apprentissage automatique reposant principalement sur les réseaux de neurones artificiels (Dif, 2020). Contrairement aux perceptrons multicouches (MLP) classiques, qui ne comportent généralement qu'une seule couche cachée, les réseaux de deep learning sont constitués de plusieurs couches cachées

successives, comme le montre la Figure 2.1. Ces couches sont exploitées pour l'extraction et la transformation hiérarchique des caractéristiques.

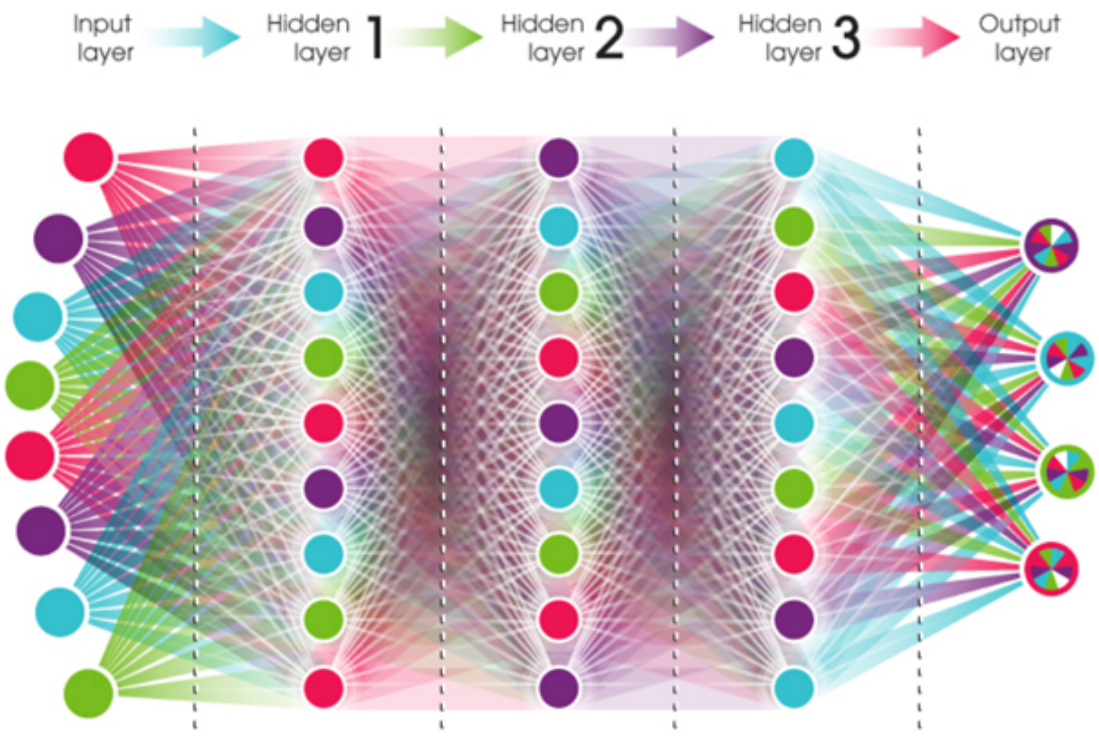

FIGURE 2.1 : Architecture d'un réseau de neurone profond entièrement connecté IRIC (2024).

Les premières couches apprennent des caractéristiques simples, telles que des contours ou des motifs élémentaires, tandis que les couches plus profondes capturent des structures de plus en plus complexes. Ce principe d'apprentissage multi-représentation permet aux réseaux profonds de résoudre efficacement des problèmes complexes. Cependant, le nombre élevé de paramètres dans les réseaux entièrement connectés peut conduire à des problèmes de surapprentissage ainsi qu'à une complexité de calcul importante.

Afin de réduire ces coûts et d'adapter les réseaux à des types de données spécifiques, plusieurs architectures spécialisées ont été proposées, notamment les réseaux de neurones convolutifs (CNN), les réseaux de neurones récurrents (RNN), les réseaux à mémoire à long et court terme (LSTM), les auto-encodeurs empilés (SAE), les réseaux antagonistes génératifs (GAN) et, plus récemment, les Transformeurs, etc.

2.2.1 RÉSEAUX DE NEURONES CONVOLUTIFS (CNN)

Les réseaux de neurones convolutifs (Convolutional Neural Networks, CNN) sont des architectures d'apprentissage profond inspirées du cortex visuel humain (Hubel & Wiesel, 1962; Dif, 2020). Le cortex visuel est la région du cerveau responsable du traitement des informations visuelles. Un CNN est généralement composé de trois types de couches : les couches de convolution, les couches de sous-échantillonnage (pooling) et les couches entièrement connectées, comme l'illustre la Figure 2.2.

La couche de convolution constitue le bloc fondamental du réseau. Elle permet d'extraire automatiquement les caractéristiques pertinentes à partir des données d'entrée en appliquant des filtres convolutifs appris durant l'entraînement. La couche de pooling est utilisée pour réduire la dimensionnalité des cartes de caractéristiques résultantes, ce qui permet de diminuer la complexité du modèle tout en conservant les informations les plus discriminantes. Quant aux couches entièrement connectées, elles permettent d'apprendre des combinaisons non linéaires entre les caractéristiques extraites par les couches de convolution afin de produire la prédiction finale.

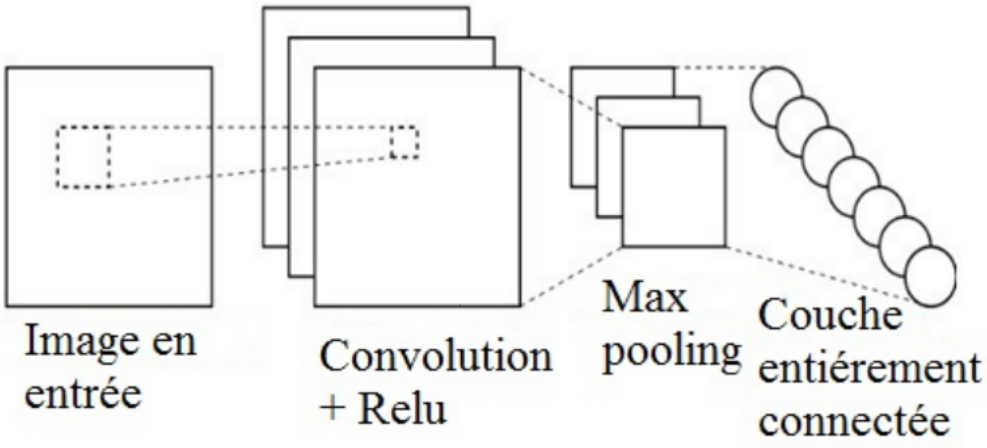

FIGURE 2.2 : Architecture d'un réseau de neurone profond convolutif Dif (2020).

Les CNN ont démontré leur efficacité dans de nombreux domaines, notamment dans les systèmes de recommandation, le traitement du langage naturel et la vision par ordinateur (Dif, 2020).

2.2.2 RÉSEAUX DE NEURONES RÉCURRENTS (RNN)

Les réseaux de neurones récurrents (Recurrent Neural Networks, RNN) sont conçus pour l'apprentissage à partir de données séquentielles (Pineda, 1987; Dif, 2020). Dans un RNN, les activations des couches cachées dépendent à la fois des entrées actuelles et des états précédents. Autrement dit, la sortie à l'instant *t-1* influence directement le calcul à l'instant *t*, ce qui permet de modéliser les dépendances temporelles.

L'architecture d'un RNN est caractérisée par la présence de connexions cycliques permettant de traiter des séquences de longueur variable (Dif, 2020). Les mêmes poids sont partagés à travers le temps, ce qui permet de limiter le nombre total de paramètres du modèle. La Figure 2.3 illustre la structure basique d'un RNN.

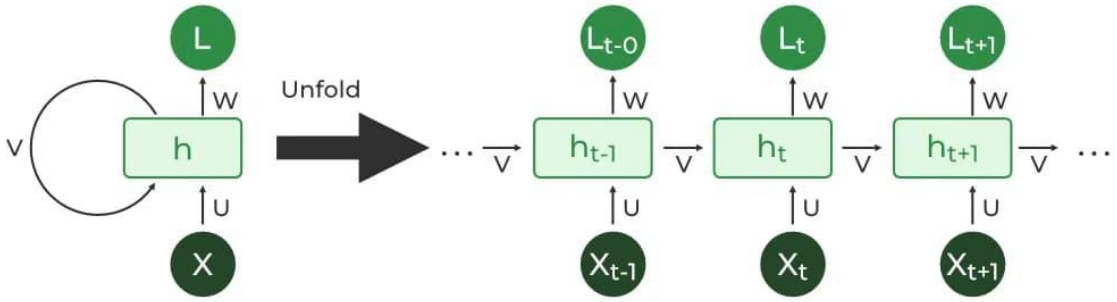

FIGURE 2.3 : Architecture d'un réseau de neurones récurrent geeksforgeeks (2025).

Les RNN ont été utilisés dans diverses applications, notamment en traitement du texte, en reconnaissance d'actions et en segmentation d'images. Cependant, malgré leurs avantages, les RNN présentent des limitations importantes liées à la dégradation du gradient lors de l'apprentissage de longues séquences. Ce phénomène empêche le modèle de capturer efficacement les dépendances à long terme. Pour pallier ce problème, les réseaux LSTM ont été proposés.

2.2.3 RÉSEAUX À MÉMOIRE À LONG ET COURT TERME (LSTM)

Les réseaux LSTM (Long Short-Term Memory) constituent une amélioration majeure des RNN classiques (Hochreiter & Schmidhuber, 1997). Contrairement à ces derniers, les LSTM sont capables de mémoriser des informations sur de longues périodes grâce à un mécanisme de mémoire interne contrôlé par trois portes : la porte d'entrée, la porte d'oubli et la porte de sortie. La Figure 2.4 illustre l'architecture d'un LSTM.

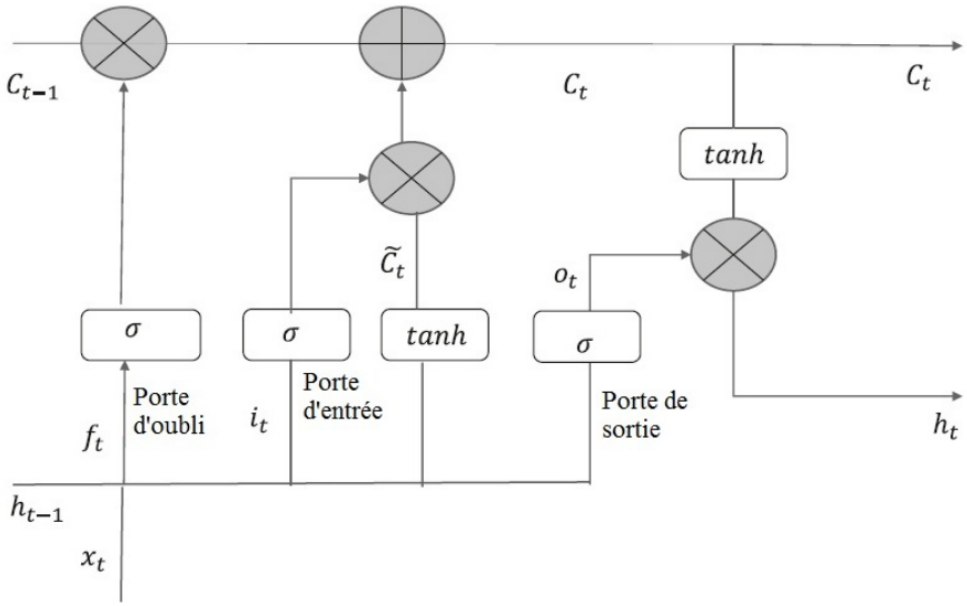

FIGURE 2.4 : Architecture d'une unité de mémoire à court et à long terme Dif (2020).

Chaque unité LSTM possède un état de cellule c_t, qui est mis à jour dynamiquement en fonction des informations pertinentes à conserver ou à oublier. La porte d'oubli contrôle quelles informations issues de l'état précédent $c_{(t-1)}$ doivent être supprimées ou conservées. La porte d'entrée décide quelles nouvelles informations doivent être ajoutées à l'état de cellule, tandis que la porte de sortie détermine quelles informations sont utilisées pour calculer l'état caché h_t. Grâce à ce mécanisme, les LSTM ont été utilisés avec succès dans de nombreux domaines, notamment l'amélioration de la parole, la détection d'activité vocale et le sous-titrage automatique d'images (Dif, 2020).

Cependant, malgré leurs performances, les LSTM restent basés sur un traitement séquentiel, ce qui limite leur capacité de parallélisation et complique l'extraction du contexte global (Islam *et al.*, 2024). Les variantes bidirectionnelles (Graves & Schmidhuber, 2005; Li *et al.*, 2020) atténuent partiellement ce problème, mais ne le résolvent pas complètement. Ces limitations ont conduit au développement des architectures de type Transformer.

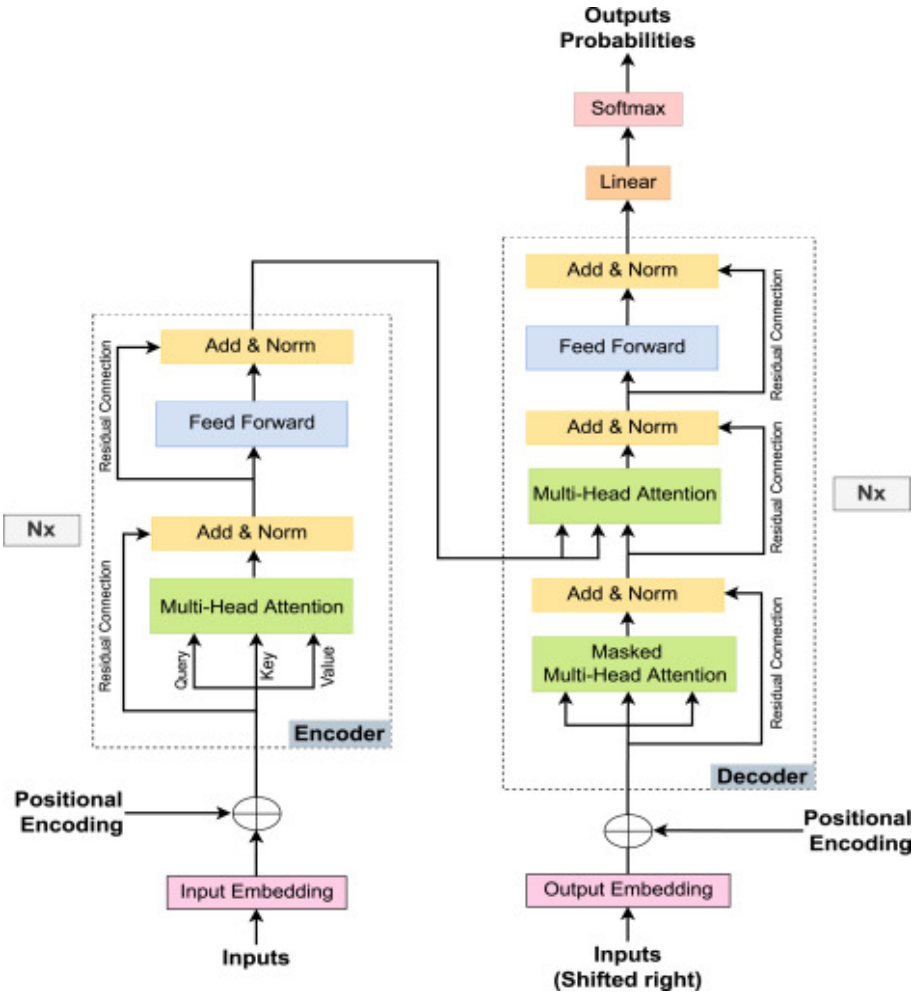

FIGURE 2.5 : Architecture d'un transformer Vaswani *et al.* (2017).

2.2.4 LES TRANSFORMERS

Les Transformers, introduits par Vaswani *et al.* (2017), constituent une classe de modèles d'apprentissage profond conçue pour surmonter les limitations des architectures séquence-à-séquence traditionnelles, notamment la dépendance au traitement strictement séquentiel et la difficulté à capturer les dépendances à long terme (Islam *et al.*, 2024; Ren *et al.*, 2023). La Figure 2.5 illustre l'architecture d'un transformer.

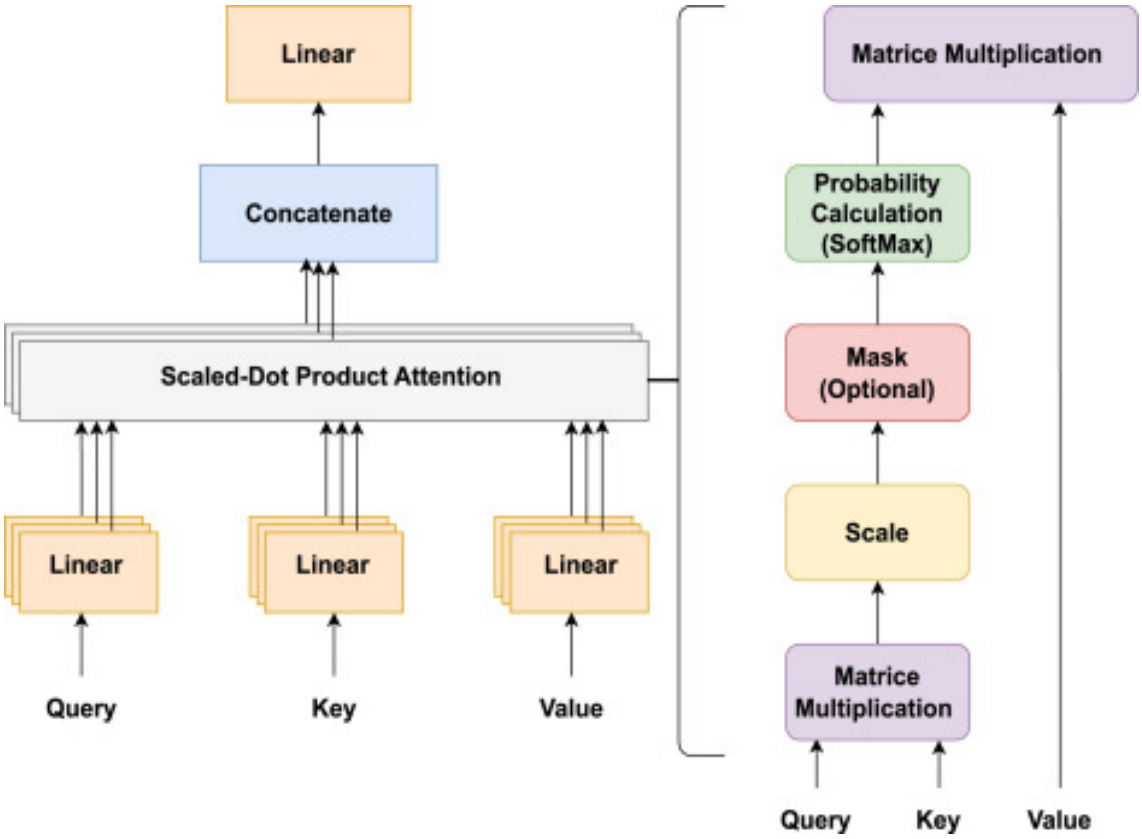

FIGURE 2.6 : Attention multi-têtes et attention produit scalaire à l'échelle Vaswani *et al.* (2017).

Les transformers reposent sur le mécanisme d'auto-attention multi-têtes, comme le montre la Figure 2.6, qui permet au modèle d'analyser l'ensemble de la séquence simultanément et de pondérer dynamiquement l'importance relative de chaque élément (Vaswani *et al.*, 2017).

Cette approche rend possible une parallélisation efficace sur GPU et accélère considérablement l'entraînement (Niu *et al.*, 2021; Zheng *et al.*, 2020). Grâce à leurs performances et à leur flexibilité, les Transformers ont été appliqués avec succès dans de nombreux domaines, notamment le traitement du langage naturel, la vision par ordinateur et l'analyse multimodale (Ren *et al.*, 2023; Reza *et al.*, 2022; Wang *et al.*, 2019; Yeh *et al.*, 2019).

2.3 MÉTHODES DE FUSION MULTIMODALE

Les données multimodales impliquent l'intégration de différents types d'informations, telles que les images, le texte et l'audio, qui peuvent être traitées par des modèles d'apprentissage automatique afin d'améliorer les performances dans de nombreuses tâches (Li & Tang, 2024; Baltrušaitis *et al.*, 2018; Barua *et al.*, 2023; Liang *et al.*, 2024; Tang *et al.*, 2019; Tao *et al.*, 2022; Duan *et al.*, 2021). En combinant plusieurs sources d'information, la fusion multimodale permet de tirer parti des atouts propres à chaque modalité tout en compensant les faiblesses ou les lacunes qui pourraient résulter de l'utilisation d'un seul type de données (Baltrušaitis *et al.*, 2018; Barua *et al.*, 2023; Binte Rashid *et al.*, 2024). Par exemple, chaque modalité peut contribuer différemment à la prédiction finale, l'une pouvant être plus informative ou moins bruitée que les autres à un instant donné. Les méthodes de fusion jouent ainsi un rôle essentiel pour combiner efficacement les informations provenant de différentes modalités.

Traditionnellement, les méthodes de fusion sont catégorisées en fonction de l'étape du pipeline de traitement des données à laquelle la fusion intervient. La fusion précoce intègre les données au niveau des caractéristiques, la fusion tardive au niveau de la décision, et la fusion hybride combine des aspects des deux approches (Li & Tang, 2024; Baltrušaitis *et al.*, 2018; Barua *et al.*, 2023).

La fusion précoce consiste à combiner les données issues de différentes modalités dès le stade de l'extraction des caractéristiques (Snoek *et al.*, 2005), ce qui permet de capturer précocement les interactions entre les modalités. Elle présente généralement un faible taux de faux positifs, ce qui est particulièrement avantageux pour les applications où la réduction des alertes inutiles est une priorité. En revanche, la fusion tardive combine les sorties de modèles spécifiques à chaque modalité au niveau de la décision. Cette approche est avantageuse lorsque l'une ou plusieurs modalités sont manquantes lors de la prédiction. Toutefois, le fait d'entraîner

séparément des modèles pour les données audio et visuelles avant la fusion augmente le temps de traitement global ainsi que les ressources de calcul requises, ce qui constitue son principal inconvénient. Enfin, la fusion hybride intègre à la fois des éléments de la fusion précoce et de la fusion tardive.

2.4 ESTIMATION DE LA POSE HUMAINE

Les systèmes de vision basés sur l'analyse du squelette humain, dérivé d'images classiques, gagnent en popularité en raison de leur respect de la vie privée et de leur robustesse face aux variations environnementales (Raza *et al.*, 2025). Contrairement aux méthodes basées directement sur les pixels, ces approches exploitent uniquement les points d'articulation du corps humain pour représenter les mouvements et les postures. Ces méthodes permettent d'identifier diverses activités, telles que la marche, la course, les sauts, les chutes ou les bagarres, et sont capables de gérer la présence de plusieurs individus dans une même scène. Plusieurs extracteurs de pose ont été proposés dans la littérature, parmi lesquels AlphaPose, OpenPose, BlazePose, OpenPifPaf et HRNet.

2.4.1 ALPHAPOSE

AlphaPose est un système d'estimation de pose multi-personnes en temps réel reposant sur un module de détection de personnes (par exemple YOLO), suivi d'un réseau spécialisé pour la localisation précise des articulations. Il a obtenu des performances élevées sur des jeux de données de référence, avec une précision moyenne d'environ 70% sur le jeu de données COCO et 80% sur le jeu de données MPII. De plus, AlphaPose peut être couplé à un module de suivi temporel, tel que Flow, permettant de relier les poses d'un même individu à travers le temps. Cette approche améliore la cohérence temporelle des squelettes estimés. Le modèle fournit généralement 17 points d'articulation par personne, comme illustré à la Figure 2.7.

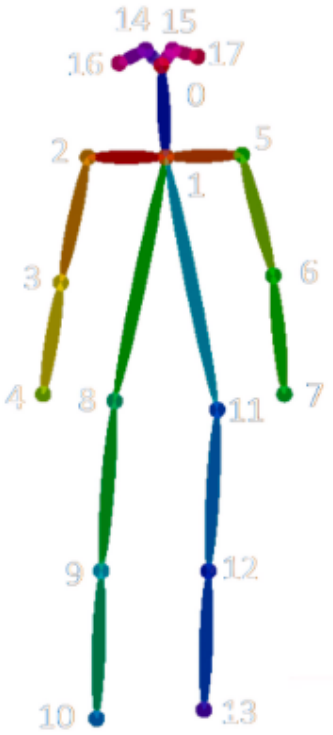

FIGURE 2.7 : Squelette d'AlphaPose Raza *et al.* (2025).

2.4.2 OPENPOSE

OpenPose, développé par l'Université Carnegie Mellon (CMU), est l'une des premières méthodes capables d'estimer en temps réel la pose de plusieurs personnes dans une même image. Cette méthode repose sur un réseau de neurones convolutif supervisé et utilise à la fois des cartes de confiance et des champs d'affinité des parties (Part Affinity Fields, PAF). Dans un premier temps, une carte de caractéristiques est générée à l'aide d'un réseau de type VGG-19. Cette carte est ensuite transmise à un réseau à deux branches et plusieurs étapes : l'une est chargée de prédire les cartes de confiance des articulations, tandis que l'autre prédit les PAF décrivant les connexions entre les différentes parties du corps. La combinaison de ces deux sorties permet de reconstruire le squelette humain en deux dimensions. OpenPose

peut estimer jusqu'à 25 points d'articulation par individu. Un exemple de squelette détecté est présenté à la Figure 2.8.

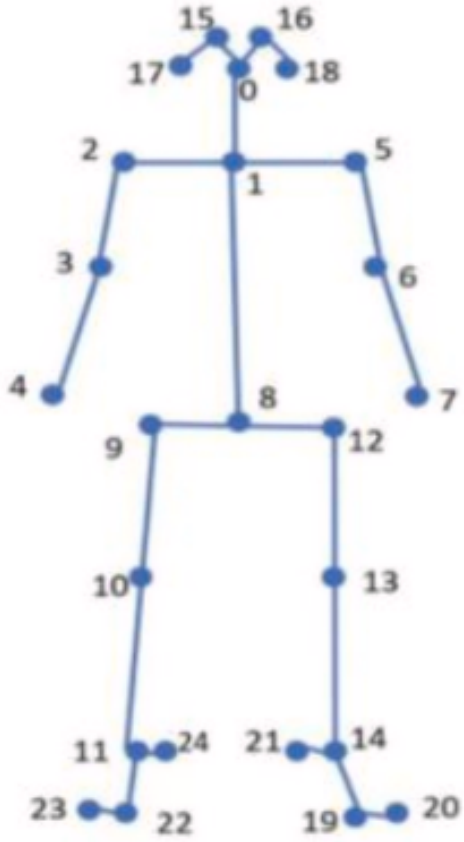

FIGURE 2.8 : Squelette OpenPose Raza *et al.* (2025).

2.4.3 BLAZEPOSE

BlazePose (Lugaresi *et al.*, 2019) est une méthode d'estimation de la pose humaine basée sur l'apprentissage automatique, conçue pour fonctionner efficacement en temps réel. Elle permet d'estimer 33 points d'articulation, comme le montre la Figure 2.9, à partir d'images RGB. Cette approche utilise un masque de segmentation afin de séparer le sujet de l'arrière-plan.

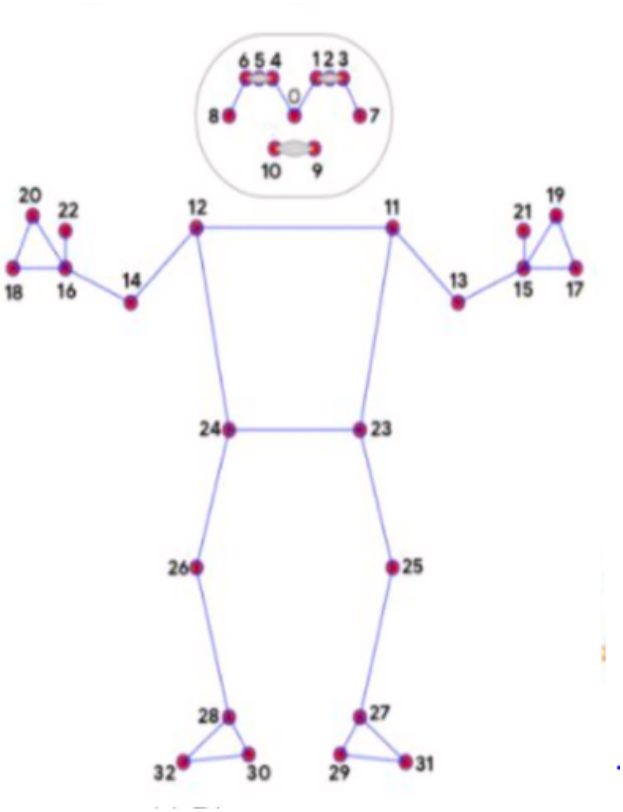

FIGURE 2.9 : Squelette BlazePose Raza *et al.* (2025).

Le fonctionnement de BlazePose repose sur une approximation géométrique du corps humain assimilé à un cercle. Le modèle prédit d'abord deux points clés virtuels, puis calcule l'orientation du corps ainsi que le rayon du cercle englobant. Le processus se déroule en deux étapes : la détection de la personne dans l'image, suivie de l'estimation précise de sa posture.

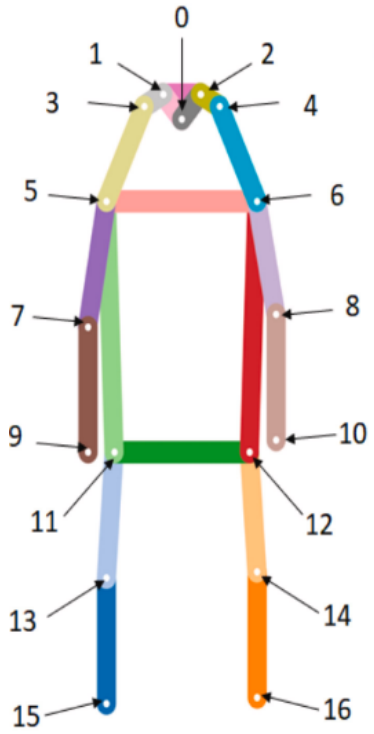

FIGURE 2.10 : Squelette OpenPifPaf Raza *et al.* (2025).

2.4.4 OPENPIFPAF

OpenPifPaf (Kreiss *et al.*, 2021) est un framework moderne destiné à l'estimation et au suivi spatio-temporel de la pose humaine. Il est conçu pour identifier simultanément les points clés et leurs relations à l'aide de champs composites. Dans cette architecture, les nœuds représentent des points clés sémantiques, tels que les articulations du corps, et sont reliés à travers plusieurs images, ce qui permet de construire un graphe unifié pour l'inférence de la posture dans le temps. Le squelette produit par OpenPifPaf est illustré à la Figure 2.10.

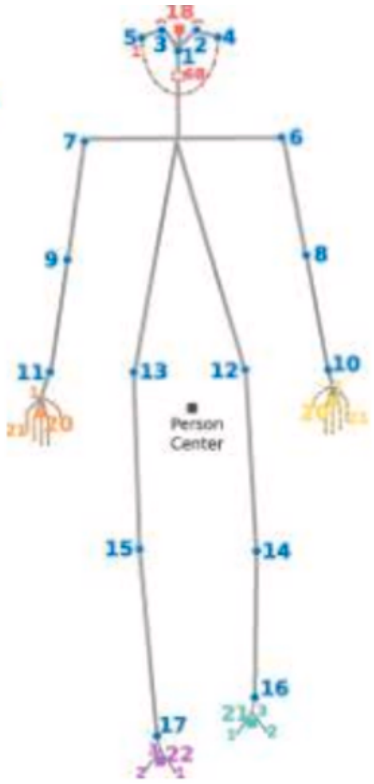

FIGURE 2.11 : Squelette HRNet Raza *et al.* (2025).

.

2.4.5 HRNET

HRNet (Cheng *et al.*, 2020) est un modèle avancé d'estimation de la pose humaine qui maintient des représentations de haute résolution tout au long du processus de traitement. Contrairement aux architectures classiques qui reconstruisent progressivement la résolution, HRNet conserve simultanément plusieurs branches de différentes résolutions et permet des échanges continus d'information entre elles.

Dans un premier temps, un réseau de haute résolution est construit. Ensuite, des sous-réseaux de plus basse résolution sont ajoutés et interconnectés. Cette architecture favorise

le développement de représentations riches et précises, ce qui améliore significativement la localisation des points d'articulation. Le squelette estimé par HRNet est illustré à la Figure 2.11

2.5 CONCLUSION

Dans ce chapitre, nous avons présenté les fondements théoriques de l'apprentissage profond, les principales architectures séquentielles modernes, les stratégies de fusion multimodale ainsi que plusieurs méthodes d'estimation de la pose humaine. Ces éléments constituent la base scientifique sur laquelle repose notre approche AV-ID Lite. Le chapitre suivant est consacré à l'étude des travaux connexes portant sur la détection de situations anormales.

CHAPITRE III

LES TRAVAUX CONNEXES

3.1 INTRODUCTION

De nombreuses approches ont été proposées dans la littérature pour la détection d'anomalies, en particulier pour l'analyse des comportements humains anormaux. Ces méthodes peuvent être classées en deux grandes catégories : les approches unimodales qui reposent exclusivement sur des informations visuelles telles que la pose humaine ou le flux optique et les approches multimodales, qui exploitent conjointement des données visuelles et audio afin de capturer une représentation plus complète des scènes analysées.

3.2 APPROCHES UNIMODALES BASÉES SUR LA POSE HUMAINE OU LE FLUX OPTIQUE POUR LA DÉTECTION D'ANOMALIE

Les méthodes unimodales utilisant l'estimation de la pose humaine ont récemment suscité un fort intérêt dans la recherche (Raza *et al.*, 2025). Ainsi, Ramirez *et al.* (2021) ont proposé une méthode de détection de chutes et de reconnaissance d'activités à partir des caractéristiques du squelette humain. L'approche utilise AlphaPose pour extraire le squelette humain à partir de vidéos issues de caméras. Les caractéristiques extraites sont ensuite évaluées à l'aide de quatre modèles d'apprentissage automatique (RF, SVM, MLP et KNN). Sur le jeu de données UP-FALL, leur approche a atteint une précision moyenne de 98,59%. Dans (Inturi *et al.*, 2023), AlphaPose est utilisé pour générer des squelettes humains, lesquels sont ensuite analysés par un CNN pour capturer les corrélations spatiales, puis un LSTM pour modéliser les dépendances temporelles. Les auteurs rapportent une précision moyenne remarquable de 99,95% pour cinq types de chutes (chute vers l'avant en s'appuyant sur les

mains, chute vers l'avant en s'appuyant sur les genoux, chute vers l'arrière, chute latérale et chute en position assise sur une chaise vide). Guan *et al.* (2021)) proposent un modèle spatio-temporel léger pour détecter les mouvements rapides liés aux chutes. Ils utilisent OpenPose pour détecter les personnes dans chaque image et extraire les caractéristiques des points articulaires, lesquelles sont ensuite transmises à un LSTM afin de détecter les chutes. Leur méthode atteint une précision de 98,46% sur le jeu de données multicaméras. Dans (Dentamaro *et al.*, 2021), OpenPose est utilisé pour extraire les caractéristiques cinématiques, qui sont ensuite transmises à un classifieur SVM afin de distinguer les situations de chute et de non-chute. Leur méthode atteint une précision de 98% sur le jeu de données Le2i. Wang *et al.* (2021) combinent YOLOv3, flux vidéo et squelette humain avec un réseau 3D-CNN, obtenant un mAP (mean Average Precision) de 91,32% sur le dataset PKU-MMD. Fatima *et al.* (2021) adoptent une approche non supervisée pour détecter les chutes chez les personnes âgées. Leur méthode utilise AlphaPose, qui détecte d'abord les personnes dans des images RGB, puis extrait les squelettes humains, lesquels sont ensuite transmis à un réseau de codage-décodage à passage de messages (MPED-RNN) afin de détecter les chutes. Ils rapportent une précision de 77,52% sur le jeu de données Le2i. Lin *et al.* (2020) proposent un cadre de détection de chutes chez les personnes âgées. Pour cela, ils utilisent le modèle d'estimation de pose OpenPose afin d'extraire les 25 points clés du squelette humain. Ensuite, différents modèles, tels que les LSTM et les GRU, sont testés afin de capturer la dynamique temporelle des mouvements et détecter les chutes. Leur méthode atteint une précision de 98,2%.

Concernant le flux optique, Chhetri *et al.* (2021) proposent un système de détection de chutes basé sur le flux optique. Leur approche repose sur une amélioration du flux optique dynamique, en utilisant une méthode de regroupement afin d'encoder les données temporelles du flux optique, réduisant ainsi le temps de traitement. Les auteurs rapportent une précision de 93,04%. Dans (Apicella & Snidaro, 2021), PoseNet est utilisé pour estimer la pose humaine,

puis un modèle LSTM est employé comme classificateur afin de détecter les chutes en temps réel. Les auteurs intègrent également un modèle CNN, qui sert de relais lorsque PoseNet ne parvient pas à détecter les personnes de manière fiable. Leur méthode atteint une précision de 81.8% sur un ensemble de jeux de données multicaméras et UR-Fall. Enfin, Fei *et al.* (2023) proposent une méthode à deux flux, appelée Flow-PoseNet, pour détecter les chutes chez les personnes âgées. L'approche combine le flux optique, afin de capturer les caractéristiques du mouvement, et OpenPose pour détecter les articulations du corps. Les caractéristiques du mouvement sont ensuite apprises par un modèle CNN afin d'extraire des représentations riches, tandis qu'un modèle GCN est utilisé pour capturer les caractéristiques d'apparence corporelle issues de la sortie d'OpenPose. Les sorties du CNN et du GCN sont ensuite fusionnées et transmises à un réseau de classification pour la détection des chutes. Leur méthode atteint une précision de 98.95% sur le jeux de données Le2i.

Dans cette section, nous avons présenté plusieurs approches unimodales, reposant exclusivement sur la modalité visuelle, notamment l'estimation de la pose humaine et le flux optique. Ces travaux montrent que la pose humaine constitue une modalité efficace et relativement robuste, peu sensible aux variations d'éclairage, aux ombres, aux flous ou aux changements d'arrière-plan, tout en permettant de préserver la vie privée des individus (Liu *et al.*, 2025). De plus, l'utilisation de l'estimation de la pose humaine permet de réduire le temps d'entraînement car les données squelettique sont plus légères que les images RGB. Ces caractéristiques sont particulièrement importantes dans des environnements commerciaux, en particulier dans le contexte des magasins sans personnel fonctionnant en continu (24 h/24).

3.3 APPROCHES MULTIMODALES VISION-AUDIO POUR LA DÉTECTION D'ANOMALIES

Les systèmes de surveillance traditionnels basés sur la vidéosurveillance reposent principalement sur une analyse unimodale des données vidéo (images RGB). Bien que cette approche soit efficace pour détecter les mouvements d'objets et analyser certains schémas comportementaux, ses performances peuvent se dégrader en raison de facteurs environnementaux tels que les variations d'éclairage, les occlusions ou la complexité de l'arrière-plan (Jang *et al.*, 2025). De plus, l'exploitation exclusive des données vidéo ne permet pas toujours de saisir pleinement le contexte des actions, ce qui complique l'analyse précise de comportements humains complexes. Autrement dit, certaines actions visuellement similaires peuvent correspondre à des situations très différentes, rendant leur interprétation difficile lorsque seule la modalité visuelle est exploitée.

Afin de pallier ces limitations, plusieurs travaux récents se sont orientés vers l'apprentissage multimodal vision-audio, permettant de combiner les informations visuelles et sonores pour une meilleure compréhension des scènes (Elharrouss *et al.*, 2021; Jang *et al.*, 2025).

Ainsi, Yu *et al.* (2022) proposent une stratégie d'apprentissage contrastif audio-visuelle faiblement supervisée, couplée à un module d'auto-distillation. Leur réseau à deux flux regroupe les instances en ensembles unimodaux et multimodaux, génère des paires positives et négatives contrastives, puis applique un transfert de connaissances du modèle visuel vers le modèle audio-visuel. Leur méthode obtient un AP (Average Precision) de 83,40% sur le jeu de données XD-Violence. Wu *et al.* (2025) introduisent une approche faiblement supervisée basée sur une fusion audio-visuelle adaptative et une adaptation légère du modèle CLIP. Leur méthode atteint un AP (Average Precision) de 86,04% sur le jeu de données XD-Violence. Dans (Zhou *et al.*, 2023), les auteurs présentent UR-DMU, un réseau à double mémoire régulé par

l'incertitude, capable d'apprendre à la fois les représentations normales et les caractéristiques discriminantes des anomalies. Leur approche combine une auto-attention globale et locale avec deux banques de mémoire afin de maximiser la séparation entre comportements normaux et anormaux. Les auteurs rapportent un AP (Average Precision) de 81,66% en vision seule et de 81,77% en configuration vision-audio. Enfin, Jang *et al.* (2025) proposent une approche de détection multimodale dans les magasins sans personnel basée sur deux configurations : une fusion précoce vidéo + audio analysée par un Transformer et une fusion vidéo + capteurs traitée par ViViT. Les auteurs rapportent un F1-score de 84% pour la configuration vision + audio sur le dataset STARSS23, et 86% pour la configuration vision + capteur sur le dataset Fall Accident Risk Motion Video-Sensor Pair. Cependant, cette approche présente plusieurs limitations, notamment l'absence d'informations sur le modèle d'extraction audio utilisé, le prétraitement des données vidéo, ainsi que des ambiguïtés concernant la stratégie de fusion employée (fusion précoce synchronisée ou fusion précoce aléatoire). Ces omissions rendent l'approche difficilement reproductible et limitent les possibilités de comparaison expérimentale rigoureuse.

Ces travaux confirment l'intérêt croissant des approches multimodales (vision-audio) pour la détection d'anomalies, car elles permettent de mieux saisir le contexte des actions, qu'elles soient normales ou anormales.

3.4 CONCLUSION

Dans ce chapitre, nous avons étudié plusieurs travaux liés à la détection d'anomalies unimodales et multimodales. Ces travaux montrent que les méthodes unimodales basées sur l'estimation de la pose humaine offrent des performances élevées tout en préservant la vie privée, car elles manipulent des données compactes, structurées et robustes face à des conditions environnementales difficiles. En parallèle, les approches multimodales vision-audio

renforcent la robustesse globale des systèmes grâce à la complémentarité des informations visuelles et sonores.

Cependant, à notre connaissance, les travaux exploitant conjointement la pose humaine et les informations audio dans les environnements commerciaux, en particulier dans les magasins sans personnel, demeurent limités.

C'est dans cette optique que nous proposons AV-ID Lite, une approche multimodale combinant AlphaPose pour la modalité visuelle et VGGish pour la modalité audio. Les deux modalités sont intégrées à l'aide d'une fusion précoce synchronisée, puis analysées par un Transformer léger associé à un mécanisme de pooling Log-Sum-Exp (LSE). Cette architecture est conçue pour détecter efficacement et rapidement les comportements anormaux, tels que les bagarres et les évanouissements, dans les magasins sans personnel, tout en garantissant la confidentialité des clients. Nous introduisons également un nouveau jeu de données multimodal dédié aux environnements sans personnel, comprenant trois catégories pertinentes : ADL (activités de la vie quotidienne), bagarres et évanouissements. Dans le chapitre suivant, nous présentons en détail l'approche proposée.

CHAPITRE IV
MÉTHODOLOGIE DE L'APPROCHE PROPOSÉE

4.1 INTRODUCTION

Dans le chapitre précédent, nous avons présenté plusieurs approches existantes liées à la détection d'anomalies à partir de caméras de vidéosurveillance. Ces approches incluent aussi bien des méthodes unimodales, reposant exclusivement sur la modalité visuelle pour détecter des chutes à l'aide de techniques d'estimation de pose humaine ou de flux optique, que des méthodes multimodales, combinant les modalités visuelle et audio pour la détection de scènes violentes telles que les bagarres, les fusillades ou les explosions. Cependant, l'analyse des comportements anormaux dans des vidéos issues de caméras de surveillance publiques, de supermarchés, de domiciles ou d'environnements routiers demeure une tâche complexe. En effet, ces contextes exigent des modèles à la fois efficaces et légers, capables de détecter en temps réel les incidents afin de permettre une intervention rapide des services de secours.

Bien que les approches unimodales exploitant la modalité visuelle, notamment à travers des modèles d'estimation de pose humaine tels qu'AlphaPose, OpenPose ou BlazePose, présentent l'avantage d'être robustes et respectueuses de la vie privée, elles reposent souvent sur des architectures lourdes, telles que les CNN, LSTM ou GCN, et sont généralement conçues pour analyser le comportement d'une seule personne à la fois. À l'inverse, les approches multimodales combinant les modalités visuelle et audio permettent une meilleure compréhension globale des scènes, mais elles s'appuient fréquemment sur des extracteurs de caractéristiques visuelles coûteux en calcul, tels que I3D ou ViT-B/16, peu adaptés aux contraintes de la détection en temps réel. De plus, ces dernières sont souvent sensibles aux

variations d'éclairage, aux ombres, au flou ou aux changements d'arrière-plan, et ne prennent pas toujours en compte les enjeux liés à la protection de la vie privée.

C'est dans ce contexte que, dans ce mémoire, nous proposons AV-ID Lite, une approche multimodale légère destinée à la détection d'incidents dans les magasins sans personnel. AV-ID Lite combine AlphaPose pour l'extraction de la pose humaine (modalité visuelle) et VGGish pour l'extraction des caractéristiques audio. Les deux modalités sont ensuite intégrées à l'aide d'une fusion précoce synchronisée, puis analysées par un Transformer léger associé à un mécanisme de pooling Log-Sum-Exp (LSE). Nous introduisons également un nouveau jeu de données multimodal dédié aux environnements de magasins sans personnel, construit à partir de jeux de données existants.

L'objectif principal de ce travail est de détecter automatiquement trois classes d'événements : les activités normales de la vie quotidienne (ADL), les évanouissements ou chutes, et les bagarres, tout en tenant compte de la vie privée des clients.

4.2 ARCHITECTURE DE NOTRE APPROCHE AV-ID LITE

L'approche que nous proposons, appelée AV-ID Lite, est une approche multimodale légère capable de détecter rapidement trois types d'événements dans les magasins sans personnel : les activités normales (ADL), les évanouissements et les bagarres, tout en préservant la vie privée des clients. La Figure 4.1 présente l'architecture globale du modèle proposé AV-ID Lite. Dans ce travail, le terme architecture désigne l'organisation structurelle des différents modules composant le système, tandis que le terme modèle fait référence à l'approche d'intelligence artificielle complète utilisée pour la détection des incidents.

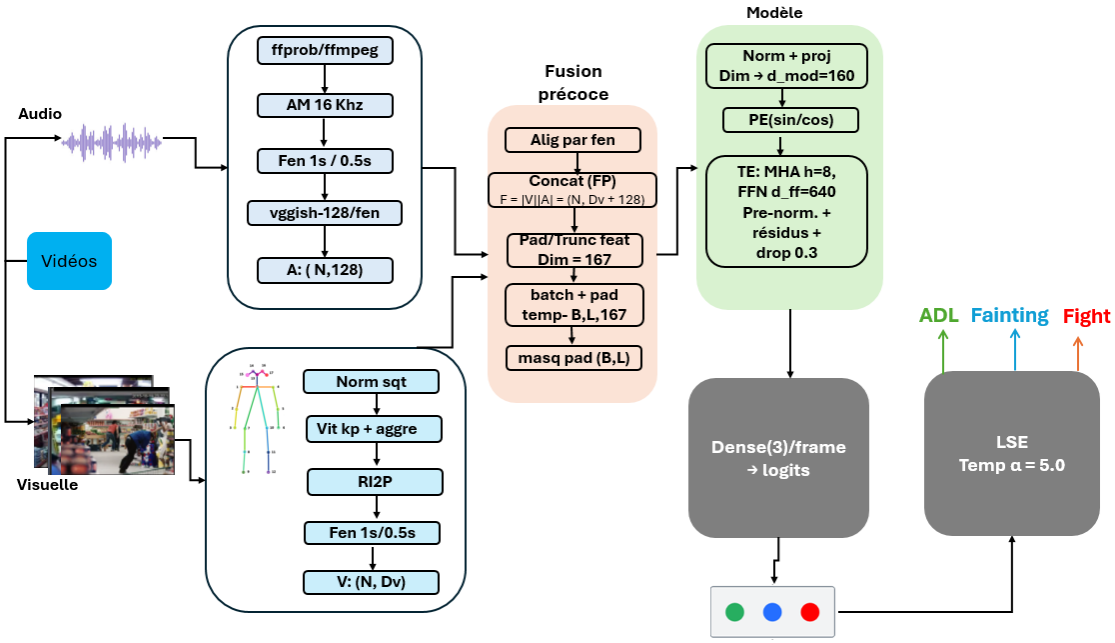

FIGURE 4.1 : Architecture de l'approche proposée AV-ID Lite.

Dans un premier temps, à partir des vidéos issues des caméras de surveillance, AlphaPose est utilisé pour extraire les points clés du squelette humain de chaque personne, constituant ainsi les caractéristiques visuelles. Ces points sont ensuite normalisés, puis organisés en fenêtres temporelles de 1 seconde avec un pas de 0,5 seconde. En parallèle, les caractéristiques audios sont extraites à l'aide du modèle VGGish, qui génère un vecteur de dimension 128 pour chaque fenêtre audio d'une seconde, avec le même pas temporel. Les caractéristiques visuelles et audio, issues respectivement d'AlphaPose et de VGGish, sont ensuite transmises au module de fusion précoce strictement synchronisée. Ce module aligne les fenêtres temporelles correspondantes des deux modalités, puis procède à leur fusion à chaque instant temporel. Les caractéristiques fusionnées sont ensuite normalisées et enrichies par un encodage positionnel, permettant de préserver l'ordre temporel des séquences. Les séquences ainsi obtenues sont ensuite traitées par un Transformer léger, dont le rôle est de capturer les dépendances temporelles et les patterns discriminants complexes présents dans les données multimodales. En sortie du Transformer, une couche dense est appliquée à chaque frame

afin de produire un score pour chacune des trois classes considérées (ADL, évanouissement, bagarre). Enfin, les scores frame-par-frame sont agrégés à l'aide d'un mécanisme de pooling Log-Sum-Exp (LSE), permettant de mettre en évidence les instants les plus discriminants de la séquence et de prédire la classe finale de l'événement observé.
L'architecture globale de AV-ID Lite repose ainsi sur quatre modules principaux :

4.2.1 MODULE D'EXTRACTION VISUELLE PAR ESTIMATION DE POSE HUMAINE : ALPHAPOSE

L'extraction des caractéristiques visuelles repose sur AlphaPose, un réseau performant d'estimation de pose humaine multi-personnes fonctionnant en temps réel (Fang *et al.*, 2022). AlphaPose utilise un réseau d'apprentissage profond pour identifier et localiser les 17 points articulaires du corps humain à partir de séquences vidéo, permettant ainsi de générer des représentations squelettiques en 2D ou 3D (Liu *et al.*, 2025). Cette capacité à capturer directement les schémas de mouvement humain rend l'estimation de pose particulièrement adaptée aux tâches d'analyse comportementale, notamment lorsque les anomalies sont liées à des mouvements corporels.

Dans ce travail, nous utilisons AlphaPose comme extracteur visuel en remplacement des images RGB et flux optique, car les données squelettiques offrent une meilleure robustesse face aux variations d'éclairage, aux ombres et aux arrière-plans complexes. De plus, Alphapose préserve la vie privée des personnes car les informations squelettiques ne révèlent pas l'apparence physique.

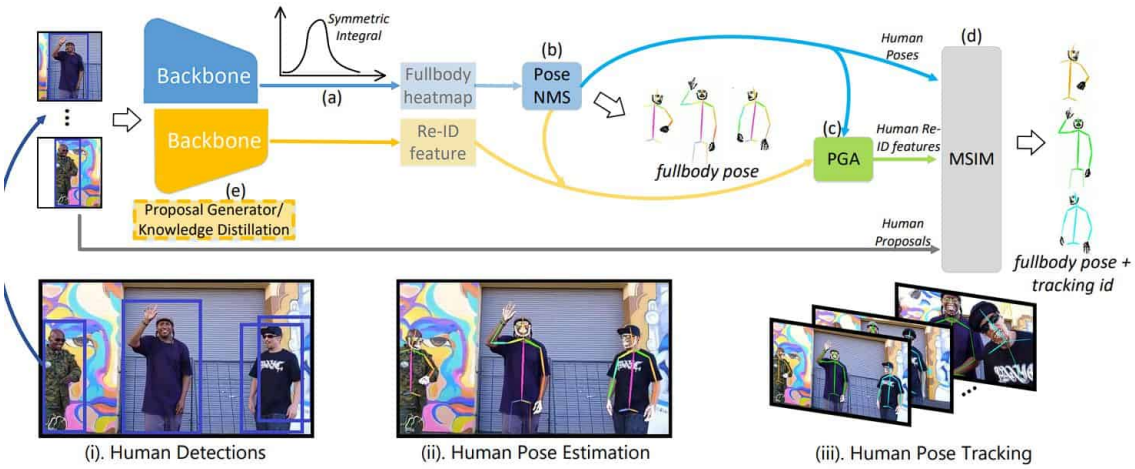

FIGURE 4.2 : Architecture de base d'AlphaPose Fang *et al.* (2022).

ARCHITECTURE DE BASE D'ALPHAPOSE

Basé sur un réseau neuronal profond, AlphaPose suit classiquement une architecture en deux étapes (Fang *et al.*, 2022; Yang *et al.*, 2025), comme illustrée à la Figure 4.2.

Dans un premier temps, un réseau de détection de personnes de type YOLOV3 localise les corps humains dans l'image d'entrée et génère des régions d'intérêt. Dans un second temps, chaque région détectée est transmise à un réseau de régression articulaire, tel que HRNet ou ResNet, chargé d'estimer précisément les principales articulations du corps humain (épaules, coudes, poignets, hanches, genoux et chevilles). Les points articulaires sont ensuite reliés afin de former une pose complète du squelette. Ce pipeline permet une estimation précise des poses humaines tout en conservant des performances compatibles avec une utilisation en temps réel.

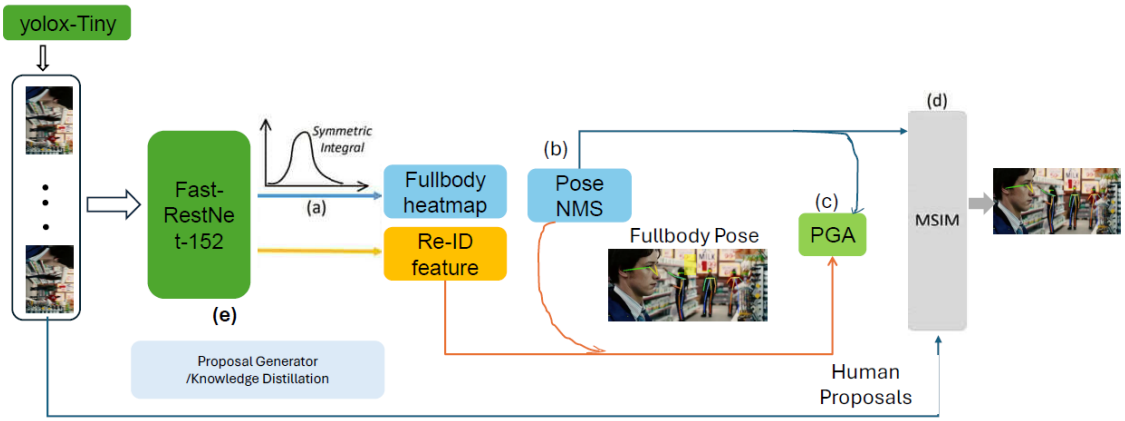

FIGURE 4.3 : Adaptation proposée d'AlphaPose

FIGURE 4.4 : Détection et estimation de la pose humaine en temps réel dans un environnement de magasin à l'aide du modèle AlphaPose amélioré.

ADAPTATION D'ALPHAPOSE POUR LES MAGASINS SANS PERSONNEL

Afin d'adapter ce pipeline aux contraintes spécifiques des magasins sans personnel et d'améliorer la détection des évanouissements, bagarres et activités normales, nous introduisons trois améliorations majeures. La Figure 4.3 présente l'architecture améliorée d'AlphaPose utilisée dans notre approche.

Premièrement, nous remplaçons le détecteur YOLOv3 par YOLOX-Tiny. Ce dernier est un modèle léger d'environ 5,06 millions de paramètres, offrant un excellent compromis entre rapidité et précision (Lin *et al.*, 2023; Wang *et al.*, 2024). Ce choix est motivé par la nécessité de détecter rapidement les personnes et d'alerter efficacement en cas d'incident dans les environnements commerciaux fonctionnant en continu.

Deuxièmement, nous utilisons le backbone FastPose-ResNet-152 (DUC) à la place du ResNet-50 par défaut. ResNet-152, basé sur le principe des connexions résiduelles, permet d'extraire des représentations plus profondes et plus discriminantes, tout en évitant les problèmes de dégradation des performances lors de l'apprentissage de réseaux très profonds (Hossain *et al.*, 2021; Krithika *et al.*, 2024). Ce choix améliore la précision de l'estimation des poses et permet de capturer des caractéristiques squelettiques plus complexes. La Figure 4.5 illustre l'architecture de ResNet-152.

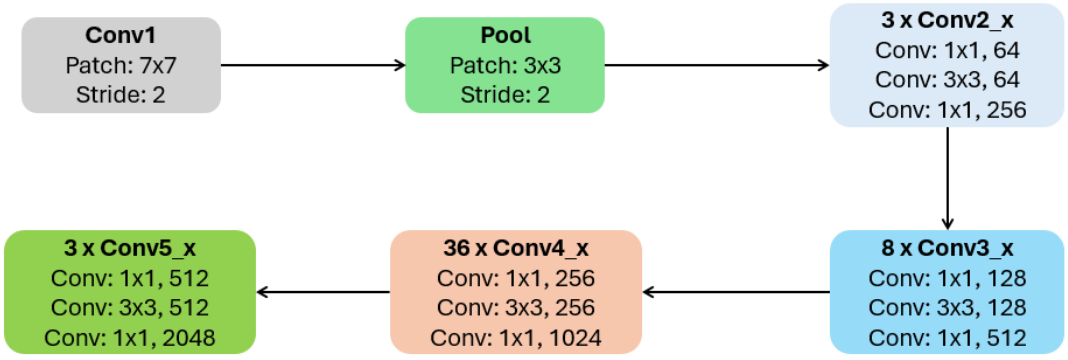

FIGURE 4.5 : Architecture de base du modèle ResNet-152 Hossain *et al.* (2021).

Troisièmement, nous enrichissons la représentation visuelle en calculant des caractéristiques cinématiques à partir des poses estimées, afin de mieux discriminer les différentes classes, en particulier les activités normales et les évanouissements. Après une normalisation des coordonnées autour du bassin, nous estimons la vitesse de chaque articulation ainsi que plusieurs agrégats robustes, notamment la vitesse moyenne globale, la vitesse maximale des mains, la vitesse maximale des pieds, le mouvement vertical du haut du corps, l'angle du tronc et la hauteur normalisée. Lorsque deux personnes sont présentes dans la scène, nous extrayons également des relations inter-personnelles, telles que la distance normalisée entre les centres de bassin, sa variation temporelle, ainsi que la différence d'angle de tronc.

Ces descripteurs permettent de caractériser efficacement les comportements associés aux trois classes étudiées. Les activités normales se traduisent par des mouvements modérés et une distance relativement stable entre individus. Les évanouissements se manifestent par une descente brutale du haut du corps suivie d'un ralentissement marqué des mouvements. À l'inverse, les bagarres se caractérisent par des vitesses élevées des mains et des pieds, des variations rapides de la distance inter-personnelle et des basculements importants du tronc. Dans notre approche, nous considérons qu'une bagarre implique au maximum deux personnes, tandis que pour les classes ADL (activités normales) et évanouissement, nous analysons une

seule personne à chaque instant. La Figure 4.4 montre un exemple de détection des personnes en temps réel réalisée avec notre version améliorée d'AlphaPose.

4.2.2 MODULE D'EXTRACTION DES CARACTÉRISTIQUES AUDIO À L'AIDE DE VGGISH

Pour la modalité audio, nous vérifions d'abord la présence d'une piste sonore à l'aide de ffprobe, avant d'extraire le signal mono 16 kHz via ffmpeg. Le signal est découpé en fenêtres de 1 seconde (pas de 0,5 s), alignées temporellement avec les fenêtres visuelles. Chaque segment audio est ensuite normalisé dans l'intervalle [-1,1] puis transmis au modèle VGGish, qui produit un embedding de dimension 128. Dans les cas où la vidéo ne contient pas d'audio, des vecteurs nuls strictement alignés sont générés.

4.2.3 MODULE DE FUSION PRÉCOCE STRICTEMENT SYNCHRONISÉE

Nous appliquons ensuite une fusion précoce synchronisée : pour chaque instants t, le vecteur visuel et le vecteur audio correspondants sont concaténés pour former un vecteur multimodal. La séquence finale $F \in \mathbb{R}^{N \times D_{\text{fused}}}$ capture ainsi explicitement les interactions entre les deux modalités. Ce choix de fusion frame-wise s'avère particulièrement pertinent pour la détection d'événements soudains (cris, chutes, impacts) et offre une précision supérieure à celle des approches de fusion tardive (Galanakis *et al.*, 2025). Lors du batching, les séquences fusionnées sont complétées par padding temporel et un masque est généré afin d'indiquer les frames artificieles. Ce masque est ensuite utilisé par le Transformer pour ignorer les positions invalides lors de l'attention et du pooling.

4.2.4 MODULE TRANSFORMER LÉGER ASSOCIÉ À UN POOLING LOG-SUM-EXP POUR LA CLASSIFICATION FINALE

Les séquences multimodales fusionnées sont ensuite analysées par un Transformer léger. Nous appliquons tout d'abord une normalisation, suivie d'une projection linéaire vers une dimension $d_{\text{model}} = 160$. Un encodeur positionnel sinusoïdal est ensuite ajouté afin d'injecter l'information d'ordre temporel dans la séquence. Le cœur du modèle est constitué d'un bloc Transformer Encoder comprenant un mécanisme d'attention multi-tête avec 8 têtes, un réseau feed-forward de dimension cachée $d_{\text{ff}} = 640$, ainsi que des connexions résiduelles, des normalisations de couche et du dropout. Cette architecture permet de capturer efficacement les dépendances temporelles tout en maintenant une faible complexité computationnelle, ce qui est essentiel pour une utilisation en quasi temps réel dans le contexte des magasins sans personnel (Jang *et al.*, 2025).

La sortie du Transformer est ensuite transmise à une couche Dense, appliquée indépendamment sur chaque frame, afin de produire un score pour chacune des trois classes considérées. Les scores frame-par-frame sont finalement agrégés à l'aide d'un mécanisme de pooling Log-Sum-Exp (LSE), paramétré par une température α. Ce mécanisme permet de mettre en évidence les instants les plus discriminants de la séquence et de prédire la classe finale de l'événement observé dans le magasin.

Nous avons choisi la fonction Log-Sum-Exp car elle permet de sélectionner les informations les plus pertinentes parmi les caractéristiques fusionnées. Autrement dit, le pooling LSE accorde davantage d'importance aux frames contenant effectivement une action significative, telles qu'une frame correspondant à un évanouissement ou à une bagarre, influençant ainsi de manière déterminante la prise de décision du modèle.
Selon Sun *et al.* (2016), le pooling Log-Sum-Exp est défini par l'équation 4.1 :

$$logits_{b,k} = \frac{1}{\alpha} \log \left(\sum_{t \in S_b} \exp(\alpha \, logits_{b,t,k}) \right) \tag{4.1}$$

où S_b désigne l'ensemble des indices temporels valides (non paddés) de la séquence b, $logits_{b,t,k}$ correspond au score de l'image t pour la classe k, et α est le paramètre de température.

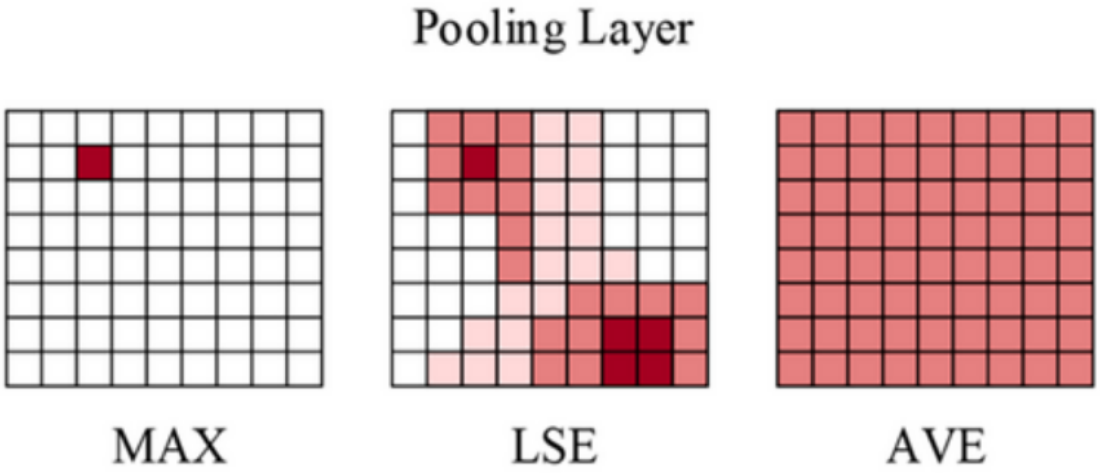

FIGURE 4.6 : Visualisation du max pooling, du LSE pooling et du average pooling Li *et al.* (2025).

Par ailleurs, comparé à l'average pooling (AVE) et au max pooling, le pooling LSE présente de meilleures performances en termes de mAP (mean Average Precision) (Sun *et al.*, 2016), comme illustré à la Figure 4.6. En effet, l'average pooling attribue une importance identique aux frames pertinentes et non pertinentes, c'est-à-dire que les frames contenant une action (par exemple un évanouissement ou une bagarre) ont le même poids que celles correspondant à l'arrière-plan. À l'inverse, le max pooling repose sur un seul frame pour produire la décision finale, ce qui le rend plus sensible au bruit et à l'initialisation du modèle lors de l'entraînement.

4.3 CONCLUSION

Dans ce chapitre, nous avons présenté la méthodologie de notre approche de détection audiovisuelle d'incidents dans les magasins sans personnel. Cette méthodologie repose sur plusieurs modules complémentaires : un module AlphaPose pour l'extraction des caractéristiques visuelles, un module VGGish pour l'extraction des caractéristiques audio, un module de fusion précoce strictement synchronisée permettant de combiner les informations visuelles et audio à chaque instant temporel, ainsi qu'un Transformer chargé de capturer les relations temporelles complexes entre les deux modalités. L'architecture est enfin complétée par une tête de classification intégrant un mécanisme de pooling Log-Sum-Exp (LSE).

Dans le chapitre suivant, nous présentons en détail l'implémentation de notre approche ainsi que les évaluations expérimentales réalisées.

CHAPITRE V
MISE EN ŒUVRE DE L'APPROCHE ET PRÉSENTATION DES RÉSULTATS EXPÉRIMENTAUX

5.1 INTRODUCTION

L'objectif de notre approche est de détecter automatiquement les situations normales et anormales (évanouissements et bagarres) dans les magasins sans personnel à l'aide d'une technologie multimodale audiovisuelle. L'approche AV-ID Lite repose sur la combinaison du modèle AlphaPose pour l'extraction des caractéristiques visuelles et du modèle VGGish pour l'extraction des caractéristiques audio. Ces deux modalités sont ensuite fusionnées à l'aide d'un module de fusion précoce strictement synchronisée, puis analysées par un Transformer léger chargé de capturer les relations temporelles complexes. La sortie du Transformer est enfin transmise à une couche de classification intégrant un mécanisme de pooling Log-Sum-Exp (LSE) afin de prédire l'une des trois classes considérées.

Dans ce chapitre, nous présentons tout d'abord la construction de notre ensemble de données, puis l'environnement de développement utilisé, ainsi que la méthode de recherche des meilleurs hyperparamètres pour l'entraînement de notre modèle. Nous décrivons ensuite les métriques d'évaluation employées pour mesurer les performances de notre approche. Nous présentons également une étude comparative avec plusieurs méthodes de l'état de l'art. Enfin, nous réalisons une étude d'ablation comparant le pooling LSE au mean pooling afin de mettre en évidence l'intérêt de l'utilisation du pooling LSE.

5.2 PRÉPARATION DU DATASET

L'un des défis majeurs rencontrés dans les magasins de détail sans personnel concerne l'absence de données adaptées à la détection automatique d'incidents. Les jeux de données existants restent souvent limités pour ce type de contexte et ne permettent pas toujours d'identifier simultanément les activités normales (ADL), les évanouissements et les bagarres. Afin de pallier ces limitations et de mieux représenter les situations rencontrées dans les magasins autonomes, nous avons constitué un nouvel ensemble de données multimodal dédié à la détection d'incidents dans ce type d'environnement. Ce dataset regroupe trois catégories d'événements pertinentes pour les environnements commerciaux autonomes : les activités normales de la vie quotidienne (ADL), les évanouissements et les bagarres. Il a été construit à partir de plusieurs jeux de données publics existants, notamment RWF-2000, XD-Violence, MERL Shopping, Le2i et le dataset multicaméra de Montréal.

5.2.1 LE JEU DE DONNÉES RWF-2000

Le jeu de données RWF-2000 [1], proposé par Cheng *et al.* (2021), est un jeu de données public à grande échelle dédié à la détection de la violence. Il est construit à partir de vidéos de vidéosurveillance réelles et comprend 1000 extraits violents et 1000 extraits non violents, d'une durée d'environ 5 secondes, comportant en moyenne 150 images chacun. Chaque vidéo présente diverses formes de violence, telles que des bagarres, des vols, des explosions, des fusillades ou des agressions.

Le jeu de données RWF-2000 est particulièrement complexe, car il intègre des conditions difficiles telles qu'une faible luminosité, du flou et des occlusions, ce qui le rend hautement

1. Jeu de données RWF-2000 : https://www.kaggle.com/datasets/vulamnguyen/rwf2000?resource=download

représentatif des scénarios du monde réel (Huang & Jiang, 2025). La Figure 5.1 présente un exemple issu du jeu de données RWF-2000.

FIGURE 5.1 : Exemple du jeu de données RWF-2000

5.2.2 LE JEU DE DONNÉES XD-VIOLENCE

Le jeu de données XD-Violence[2] (Wu *et al.*, 2020) est un ensemble de données audiovisuelles destiné à la détection de la violence. Il est constitué de films, de vidéos issues du web, de flux sportifs en streaming, ainsi que d'enregistrements provenant de caméras de sécurité et de systèmes de vidéosurveillance. Ce jeu de données comprend 4754 vidéos non découpées couvrant six types de violence courants (abus, accidents de la route, explosions, bagarres, émeutes et fusillades). Les annotations sont fournies au niveau de la vidéo pour l'ensemble d'entraînement et au niveau de l'image pour l'ensemble de test, pour une durée totale d'environ 217 heures (Zhou *et al.*, 2024). La Figure 5.2 présente un exemple issu du jeu de données XD-Violence.

2. Jeu de données XD-Violence : https://www.kaggle.com/datasets/nguhaduong/xd-violence-video-dataset

FIGURE 5.2 : Exemple du jeu de données XD-Violence

5.2.3 LE JEU DE DONNÉES MERL SHOPPING

Le jeu de données MERL Shopping [3] (Wen *et al.*, 2022) est un jeu de données public composé de 106 vidéos filmées en vue de dessus, avec une résolution de 920 x 680 pixels, une durée d'environ deux minutes chacune et une fréquence de 30 images par seconde. Les 41 participants ont été invités à effectuer des achats dans un magasin réel, ce qui permet de capturer des comportements naturels de clients. La Figure 5.3 présente un exemple d'image du jeu de données MERL Shopping.

3. Jeu de données MERL : https://www.merl.com/pub/tmarks/MERL_Shopping_Dataset/

FIGURE 5.3 : Exemple du jeu de données MERL Shopping

5.2.4 LE JEU DE DONNÉES LE2I

Le jeu de données Le2i[4] (Vishnu *et al.*, 2021) est un jeu de données standard dédié à la détection de chutes à partir de vidéos RGB. Il comprend 130 scènes de chutes et de non-chutes, enregistrées à 25 images par seconde par 17 acteurs à l'aide d'une seule caméra. La durée moyenne des vidéos est d'environ une minute, au cours de laquelle la personne chute ou poursuit une activité quotidienne. Les vidéos sont enregistrées dans différents environnements tels que des domiciles, des salles de pause, des bureaux et des salles de cours, avec une forte variabilité entre les événements de chute et de non-chute. Ces caractéristiques permettent de simuler des séquences vidéo réalistes susceptibles d'être rencontrées dans des environnements clos, comme les magasins. La Figure 5.4 présente un exemple issu du jeu de données Le2i.

4. Jeu de données Le2i : https://www.kaggle.com/datasets/tuyenldvn/falldataset-imvia

FIGURE 5.4 : Exemple du jeu de données Le2i Vishnu *et al.* (2021)

5.2.5 LE JEU DE DONNÉES MULTICAMÉRA DE MONTRÉAL

Enfin, le jeu de données multicaméra de Montréal[5] (Vishnu *et al.*, 2021) est un jeu de données dédié à la détection de chutes. Il comprend 24 scénarios enregistrés à l'aide de 8 caméras vidéo. Les 22 premiers scénarios incluent des chutes ainsi que des événements sans chute, tandis que les 2 derniers scénarios ne contiennent que des événements sans chute. Les vidéos sont enregistrées à 25 images par seconde, avec une résolution moyenne de 320 × 240 pixels. La Figure 5.5 présente un exemple d'images issues du jeu de données multicaméra.

5. Jeu de données multicaméra de Montréal : https://www.iro.umontreal.ca/~labimage/Dataset/.

FIGURE 5.5 : Exemple du jeu de données Multicaméra

5.2.6 CONSTRUCTION DE NOTRE JEU DE DONNÉES

Pour construire notre jeu de données final, nous avons d'abord extrait automatiquement, à partir des jeux de données RWF-2000 et XD-Violence, les vidéos correspondant à des scènes de supermarché. Pour ce faire, nous avons utilisé un script Python basé sur un ResNet-18 préentraîné sur Places365, un modèle capable de reconnaître 365 types de scènes (supermarché, plage, bureau, etc.). Les clips détectés comme appartenant à un environnement de magasin et contenant des activités normales ont été regroupés avec les vidéos du jeu de données MERL Shopping afin de former la classe ADL (activités normales de la vie quotidienne). Les vidéos contenant des bagarres, extraites de RWF-2000 et XD-Violence, constituent la classe bagarre. Enfin, les jeux de données Le2i et multicaméra de Montréal, tous deux enregistrés dans des environnements clos similaires à ceux d'un magasin, ont été utilisés pour composer la classe évanouissement, leurs scénarios de chute s'y prêtant naturellement.

Chaque vidéo a été segmentée en clips de 5 secondes. Au total, nous avons obtenu 1033 vidéos (740 ADL, 193 évanouissements et 100 bagarres), correspondant à environ 86,1 minutes de séquences. Chaque clip est ensuite divisé en fenêtres d'une seconde avec un pas de 0,5 seconde, produisant 9 fenêtres par clip et environ 9297 fenêtres multimodales synchronisées au total. Le jeu de données est accessible publiquement en ligne [6].

5.3 STRATÉGIE DE VALIDATION ET MÉTRIQUES D'ÉVALUATION

5.3.1 ENVIRONNEMENT DE DÉVELOPPEMENT

Pour implémenter notre approche, nous avons utilisé l'environnement Google Colab Pro afin de bénéficier de la puissance de calcul du GPU [7]. Nous avons choisi la version Pro principalement pour éviter les interruptions de session lors des entraînements prolongés, mais également pour accélérer le processus d'apprentissage, car l'entraînement d'un tel modèle sur CPU aurait nécessité un temps de calcul très important.

Notre approche a été implémentée à l'aide de la bibliothèque Keras sous le framework TensorFlow. Keras est une bibliothèque de haut niveau, à la fois simple d'utilisation et flexible, facilitant le prototypage rapide de modèles de deep learning. Par ailleurs, TensorFlow offre de bonnes garanties en termes de stabilité, de performances et de déploiement en production, ce qui a motivé notre choix par rapport à d'autres frameworks.
Les Tableaux 5.1 et 5.2 résument respectivement les principaux outils logiciels et leurs versions utilisés pour l'implémentation de notre approche.

---

6. https://drive.google.com/uc?export=download&id=1kVvVcEgFDDJTZOg9Mze2s158akjiIjPy
7. Code source : https://github.com/mamadou-thiaw/VLM_Retail.

TABLEAU 5.1 : Environnement logiciel utilisé pour l'implémentation de l'approche AV-ID Lite

Élément	Version
Python	3.12.12
Système d'exploitation	Linux 6.6.105+
Architecture	x86_64
TensorFlow	2.19.0
CUDA (TensorFlow)	12.5.1

TABLEAU 5.2 : Versions des principaux paquets logiciels utilisés

Bibliothèque	Version
optuna-integration	4.5.0
tensorflow-hub	0.16.1
keras	3.10.0
pydub	0.25.1
numpy	2.0.2
pandas	2.2.2
Cython	3.0.12
opencv-python	4.12.0.88
torchvision	0.24.0+cu126

5.3.2 MÉTHODE D'OPTIMISATION DES HYPERPARAMÈTRES

Avant l'entraînement, nous avons divisé le jeu de données en 90% pour l'entraînement et 10% pour le test. Afin d'obtenir une évaluation plus robuste et de limiter l'influence d'un découpage particulier, nous avons appliqué une validation croisée stratifiée à cinq plis sur l'ensemble d'entraînement. La validation croisée est une étape importante pour évaluer la fiabilité des modèles d'apprentissage automatique, car elle permet de mesurer la variabilité des performances selon différentes partitions des données (Mohammad *et al.*, 2025).

Dans notre protocole, l'ensemble d'entraînement (90%) est partitionné en cinq sous-ensembles de taille comparable. À chaque itération, un pli est utilisé comme ensemble de

validation, tandis que les quatre autres plis servent à l'entraînement. Ce processus est répété jusqu'à ce que chaque pli ait servi une fois pour la validation. L'ensemble de test (10%) est conservé séparément et n'est utilisé que pour l'évaluation finale.

Pour la recherche des meilleurs hyperparamètres, nous avons utilisé Optuna, un framework d'optimisation efficace pour l'exploration d'espaces de recherche complexes. Les hyperparamètres explorés concernent : le nombre de têtes d'attention, le nombre de couches du Transformer, la dimension du modèle (d_{model}), la taille du réseau feed-forward (via d_{ff}), le taux de dropout, le taux d'apprentissage, la taille des batchs, ainsi que la température α du pooling Log-Sum-Exp.

Pour chaque essai, le modèle est entraîné sur les cinq plis et la performance est évaluée à l'aide du F1-score macro, dont la moyenne sur les cinq plis constitue le critère d'optimisation. Les meilleures configurations identifiées par Optuna sont ensuite utilisées pour l'entraînement final du modèle AV-ID Lite sur l'ensemble d'entraînement complet. L'entraînement final est ensuite réalisé avec l'optimiseur Adam pendant 85 époques, ce choix étant motivé par la capacité d'Adam à ajuster dynamiquement le taux d'apprentissage et à favoriser une convergence stable (Kingma & Ba, 2014). Le Tableau 5.3 présente les meilleurs hyperparamètres obtenus avec Optuna.

TABLEAU 5.3 : Meilleurs hyperparamètres obtenus avec Optuna

Hyperparamètre	Valeur
d_{model}	160
Nombre de têtes (h)	8
Nombre de couches	1
d_{ff} (via $d_{\text{ff}} = d_{\text{model}} \times$ scale)	$160 \times 4 = 640$
Dropout	0.30
Taux d'apprentissage	0.00296
Température LSE (α)	5.0
Batch size	32
Meilleur F1-macro (moyenne 5 plis)	0.9450

5.3.3 MÉTRIQUES D'ÉVALUATIONS

Pour obtenir une évaluation complète et fiable des performances, plusieurs métriques ont été utilisées, notamment l'accuracy, la précision, le rappel, le F1-score, la précision moyenne (mAP), le coefficient de corrélation de Matthews (MCC), l'aire sous la courbe ROC (AUC-ROC), ainsi que le temps d'entraînement pour l'ensemble des 85 époques.

L'accuracy correspond à la proportion de classifications correctes, qu'elles soient positives ou négatives. Elle se définit par l'équation 5.1 :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Où *TP* désigne les vrais positifs, correctement identifiés ; *FP* représente les faux positifs, identifiés à tort comme positifs ; *FN* indique les faux négatifs, identifiés à tort comme négatifs ; et *TN* désigne les vrais négatifs, correctement identifiés.

La précision correspond à la proportion des prédictions positives du modèle qui sont effectivement positives. Elle est définie par l'équation 5.2 :

$$\text{Précision} = \frac{TP}{TP+FP} \tag{5.2}$$

Le rappel (recall) correspond à la proportion des exemples positifs réels correctement identifiés par le modèle. Il est défini par l'équation 5.3 :

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5.3}$$

Le F1-score représente la moyenne harmonique de la précision et du rappel, permettant d'équilibrer ces deux mesures. Il est défini par l'équation 5.4 :

$$\text{F1-score} = 2 \times \frac{\text{Précision} \times \text{Recall}}{\text{Précision} + \text{Recall}} \tag{5.4}$$

La précision moyenne (mean Average Precision, mAP) correspond à la moyenne des valeurs de précision calculées pour chaque classe. Elle est donnée par l'équation 5.5 :

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} AP_c \tag{5.5}$$

où C représente le nombre de classes et AP_{c} la précision moyenne pour la classe c.

Le coefficient de corrélation de Matthews (MCC) mesure la qualité globale des classifications en tenant compte des vrais et faux positifs et négatifs. Il est particulièrement pertinent lorsque les classes sont déséquilibrées. Il est défini par l'équation 5.6 :

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5.6}$$

L'aire sous la courbe ROC (AUC-ROC) représente la probabilité que le modèle classe un exemple positif plus haut qu'un exemple négatif, lorsque ces deux exemples sont choisis aléatoirement.

Ces métriques permettent ainsi d'évaluer à la fois la performance globale du modèle, sa robustesse face au déséquilibre des classes, ainsi que sa capacité à distinguer efficacement les comportements normaux et anormaux.

5.4 MÉTHODES COMPARATIVES ET RÉSULTATS

Afin de situer AV-ID Lite par rapport à l'état de l'art, nous avons réimplémenté les approches proposées par Yu *et al.* (2022), Zhou *et al.* (2023) et Wu *et al.* (2025). Ces méthodes sont initialement conçues pour la classification binaire dans le cadre de la détection d'anomalies. Pour permettre une comparaison équitable avec notre modèle multiclasse, nous avons modifié leur tête de classification afin de prédire les trois classes : ADL (activité normales), Évanouissement et Bagarre. De plus, toutes les méthodes ont été entraînées avec exactement le même jeu de données, la même procédure de validation croisée à cinq plis et le même ensemble de test, garantissant ainsi une comparaison juste et cohérente. Les Tableaux 5.4 et 5.5 présentent la comparaison quantitative entre AV-ID Lite et les méthodes de l'état de l'art.

TABLEAU 5.4 : Comparaison des performances.

Méthode	Accuracy	Précision	Rappel	F1-score	mAP	MCC	AUC
Yu *et al.* (2022)	95.19%	91.28%	89.11%	90.12%	95.34%	0.89	99.28%
Wu *et al.* (2025)	94.23%	92.33%	82.89%	86.20%	93.75%	0.87	97.54%
Zhou *et al.* (2023)	94.23%	88.67%	88.67%	88.67%	94.62%	0.8681	99.20%
AV-ID Lite	**96.15%**	**94.20%**	**98.22%**	**95.91%**	**99.23%**	**0.92**	**99.75%**
Amélioration d'AV-ID Lite par rapport à la méthode de référence(%)	**+1.01% ↗**	**+2.03% ↗**	**+10.22% ↗**	**+6.42% ↗**	**+4.08% ↗**	**+3.37% ↗**	**+0.47% ↗**

TABLEAU 5.5 : Comparaison des approches en termes de temps d'entraînement

Méthode	**Modalité (Visuelle + Audio)**	**Temps (min)**	**Gain relatif du temps d'AV-ID Lite (%)**
Yu *et al.* (2022)	I3D + VGGish	8.72	**36.92% ↘**
Wu *et al.* (2025)	ViT-B/16 + Wav2CLIP	8.45	**34.91% ↘**
Zhou *et al.* (2023)	I3D + VGGish	8.92	**38.34% ↘**
AV-ID Lite	**AlphaPose + VGGish**	**5.5**	

Nous constatons que notre approche AV-ID Lite obtient les meilleures performances sur l'ensemble des métriques, avec une accuracy globale de 96.15% (+1,01%), une précision de 94.20% (+2.03%), un rappel de 98.22% (+10.22%), un F1-score de 95.91% (+6.42%), un mAP de 99.23% (+4.08%), un MCC de 0.92 (+3.37%) et une AUC-ROC de 99,75% (+0.47%). De plus, grâce à son architecture légère et optimisée, le modèle présente un temps d'entraînement réduit de seulement 5.5 minutes pour 85 époques, ce qui constitue un avantage significatif pour une utilisation en conditions réelles.

5.4.1 ANALYSE DES COURBES ROC-AUC

La Figure 5.6 illustre les courbes ROC-AUC obtenues pour les trois classes.

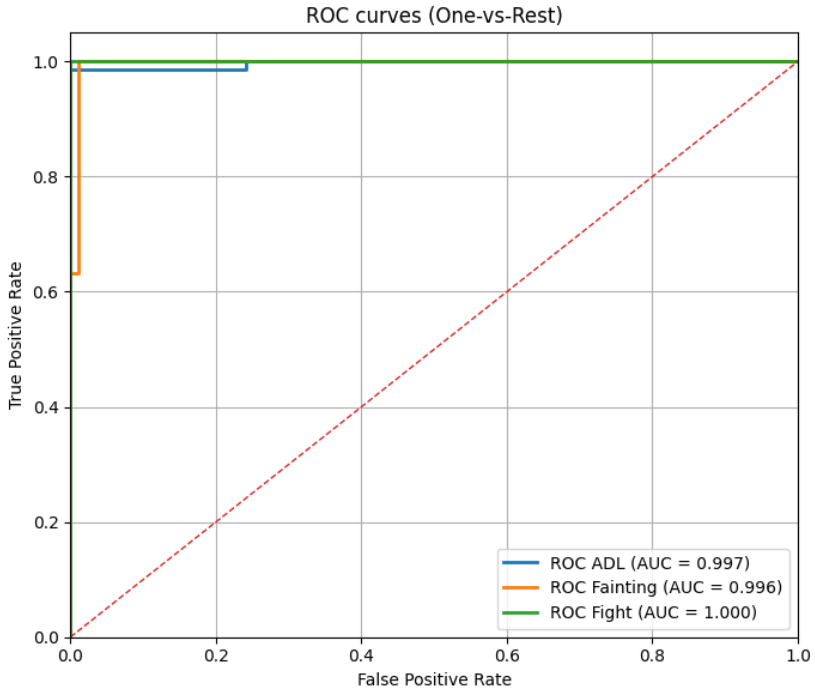

FIGURE 5.6 : Courbes ROC (one-vs-rest) et valeurs d'AUC pour les trois classes

Les résultats montrent une excellente séparabilité entre les comportements normaux et anormaux. Les valeurs d'AUC atteignent 0.997 pour la classe ADL (activités normales), 0.996 pour la classe Fainting (Évanouissement) et 1 pour la classe Fight (Bagarre). Le score parfait obtenu pour cette dernière classe confirme la capacité remarquable de notre modèle à détecter avec précision les scènes de bagarre.

De manière générale, la forme quasi rectangulaire des courbes ROC-AUC démontre l'efficacité d'AV-ID Lite à distinguer les comportements anormaux (évanouissement, bagarre) des comportements normaux, attestant ainsi de la robustesse et de la fiabilité de notre architecture multimodale.

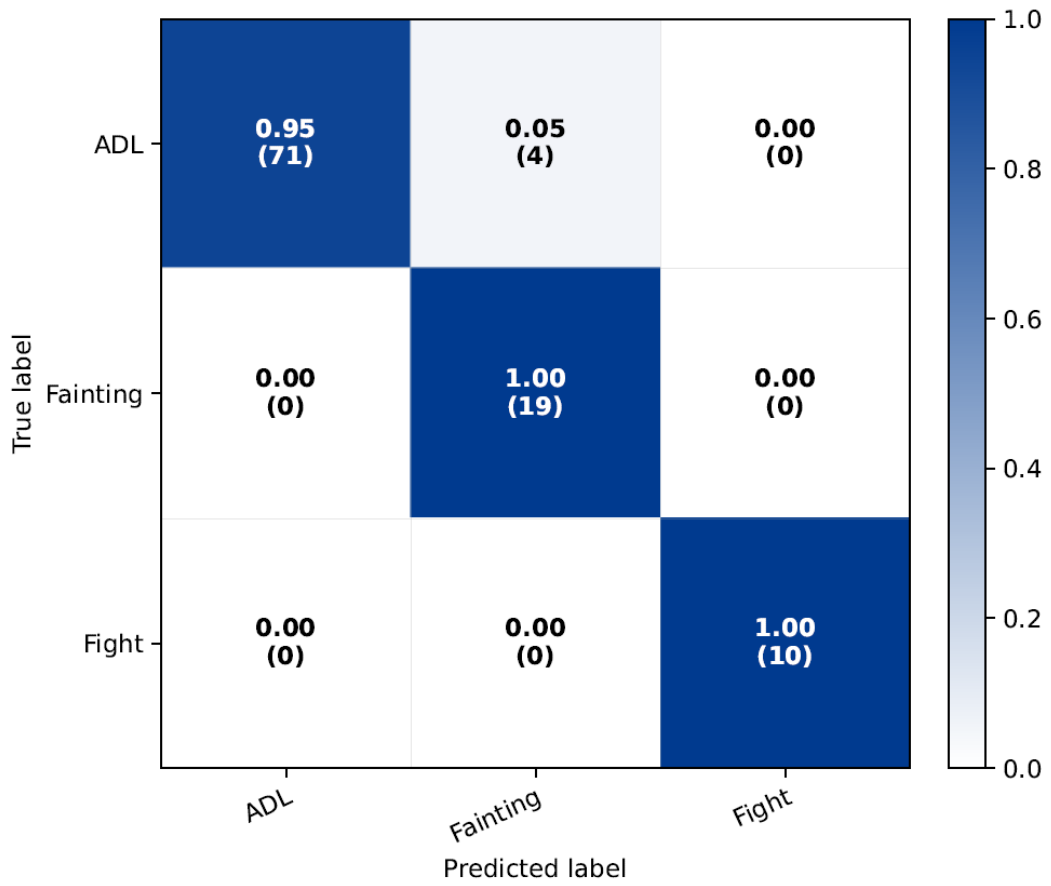

FIGURE 5.7 : Matrice de confusion AV-ID Lite

5.4.2 MATRICE DE CONFUSION

Nous avons également tracé la matrice de confusion, qui constitue un indicateur largement utilisé en apprentissage automatique. Cette matrice permet de comparer les étiquettes réelles de la variable cible avec celles prédites par le modèle. La Figure 5.7 présente la matrice de confusion obtenue par le modèle AV-ID Lite sur l'ensemble de données test. Les résultats montrent une excellente capacité de classification pour les trois classes considérées. Pour la classe ADL, 71 échantillons sur 75 sont correctement classés, tandis que 4 échantillons sont confondus avec la classe *Fainting (Évanouissement)*. Aucun échantillon ADL n'est confondu avec la classe *Fight (Bagarre)*. Cela indique que les rares erreurs concernent principalement

des cas ambigus entre activités normales et évanouissements, par exemple lorsqu'une personne trébuche, tombe puis se relève, ce qui est compréhensible dans des scénarios réels.

Pour la classe *Fainting*, l'ensemble des 19 échantillons est correctement classé, ce qui correspond à un taux de reconnaissance de 100%. Ce résultat confirme la capacité du modèle à détecter de manière fiable les événements de chute ou d'évanouissement.

De même, pour la classe *Fight*, les 10 échantillons sont tous correctement identifiés, sans aucune confusion avec les autres classes, ce qui témoigne d'une excellente discrimination des scènes de bagarre.

En résumé, cette matrice de confusion met en évidence une séparation très nette entre les classes, avec des erreurs de classification extrêmement limitées. Les confusions observées sont rares et se concentrent uniquement entre les classes ADL et *Évanouissement*, ce qui correspond aux cas les plus difficiles à distinguer visuellement. Ces résultats confirment la robustesse et la fiabilité de l'approche AV-ID Lite pour la détection d'incidents dans les magasins sans personnel.

5.5 ÉTUDE D'ABLATION : IMPACT DE LA STRATÉGIE DE POOLING

Afin d'évaluer l'impact du choix du mécanisme de pooling sur la performance globale de notre architecture, nous avons réalisé une étude d'ablation en remplaçant le pooling Log-Sum-Exp (LSE) par un mean pooling classique. Pour garantir une comparaison équitable, les deux modèles ont été entraînés en utilisant exactement les mêmes hyperparamètres ainsi que les mêmes ensembles de données d'entraînement et de test. La Figure 5.8 présente la comparaison des performances obtenues avec les deux stratégies de pooling.

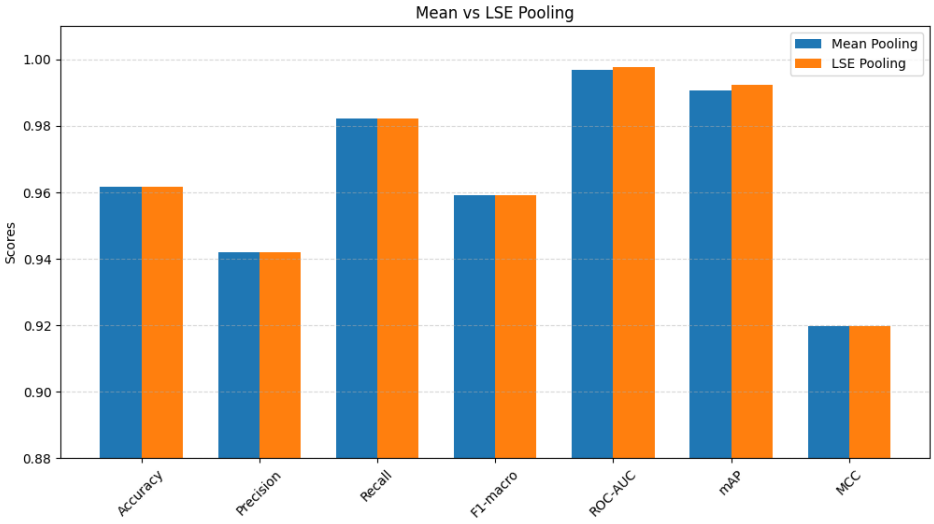

FIGURE 5.8 : Étude d'ablation entre Log-Sum-Exp (LSE) et mean pooling (AVE)

Les résultats montrent que les deux modèles obtiennent des performances globalement proches sur des métriques telles que l'accuracy, la précision, le rappel, le F1-score et le MCC. Toutefois, le pooling LSE surpasse systématiquement le mean pooling sur les métriques les plus discriminantes, en particulier le ROC-AUC, qui augmente de 0.0007, et le mAP, qui progresse de 0.0016.

Ces métriques étant particulièrement sensibles à la qualité de la séparation entre les classes et à la capacité du modèle à mettre en évidence des événements rares, cette amélioration indique que le pooling LSE permet de mieux valoriser les frames les plus informatives en amplifiant leur contribution dans la décision finale. À l'inverse, le mean pooling traite toutes les frames de manière uniforme, y compris celles qui sont peu pertinentes, ce qui peut diluer l'information cruciale présente dans les séquences contenant des comportements anormaux.

5.6 DISCUSSION ET LIMITES

Cette section discute les principaux facteurs ayant contribué aux performances du modèle AV-ID Lite, tout en mettant en évidence certaines limites de l'approche proposée. Les résultats obtenus peuvent être attribués à plusieurs choix de conception stratégiques intégrés dans le cadre proposé. Tout d'abord, l'utilisation de représentations squelettiques extraites par AlphaPose permet au modèle de se concentrer sur la posture humaine et la dynamique des mouvements tout en écartant les informations liées à l'apparence telles que la texture, les conditions d'éclairage ou l'encombrement de l'arrière-plan. Cette abstraction améliore la robustesse dans les environnements commerciaux complexes et garantit simultanément la préservation de la vie privée, ce qui contribue à l'amélioration observée du rappel (+10.22%) pour les tâches de détection d'incidents.

Deuxièmement, la stratégie de fusion précoce strictement synchronisée permet au modèle d'apprendre les corrélations conjointes entre la dynamique des poses et les signaux audio au niveau des images. Contrairement aux approches de fusion tardive (Snoek *et al.*, 2005), cet alignement temporel étroit fournit des représentations multimodales plus riches et aide à lever l'ambiguïté des situations visuellement similaires à l'aide d'indices acoustiques complémentaires, ce qui explique les gains constants en termes de précision (+2.03%), de F1-score (+6.42%) et de mAP (+4.08%).

Troisièmement, l'architecture Transformer légère améliore la modélisation temporelle tout en maintenant une faible complexité computationnelle (Jang *et al.*, 2025). En capturant les dépendances à long terme dans les séquences d'événements, le Transformer améliore la discrimination entre les activités normales et les incidents critiques sans introduire la lourde charge computationnelle typique des grandes structures visuelles, ce qui explique à la fois les performances élevées et la réduction du temps de formation.

Enfin, le mécanisme de pooling Log-Sum-Exp joue un rôle important en mettant en évidence les segments temporels les plus informatifs de chaque séquence. L'étude d'ablation comparant le pooling LSE au pooling moyen confirme que le LSE offre une meilleure séparation des classes et améliore les métriques telles que l'AUC (+0.47%) et le mAP, démontrant que la priorisation des images discriminantes est cruciale pour une reconnaissance fiable des incidents.

Malgré ces résultats très encourageants, notre approche présente certaines limites. Premièrement, le système dépend fortement de la qualité de l'estimation de pose : en cas d'occlusions sévères, de fortes densités de foule ou de détections erronées, les performances peuvent être dégradées. Deuxièmement, bien que le jeu de données construit soit varié, il reste composé de données issues de bases publiques et ne provient pas entièrement de magasins réels, ce qui peut limiter la capacité de généralisation du modèle dans certains scénarios pratiques. Troisièmement, le système est actuellement conçu pour un nombre limité de classes et ne prend pas encore en compte d'autres événements critiques tels que les vols, les comportements suspects ou les agressions verbales.

Plusieurs perspectives de recherche peuvent ainsi être envisagées. L'utilisation d'autres extracteurs de pose, tels qu'OpenPose ou BlazePose, permettrait d'évaluer la robustesse du système face à différentes architectures de détection corporelle. Des stratégies de fusion multimodale hybrides, combinant fusion précoce et fusion tardive, constituent également une piste intéressante pour capturer des interactions plus complexes entre l'audio et la vision. Enfin, la création d'un jeu de données à grande échelle entièrement dédié aux environnements commerciaux réels, incluant un large éventail d'événements (vols, agressions, chutes, comportements suspects), représenterait une avancée majeure pour la recherche et pour le déploiement industriel de ce type de systèmes.

En résumé, AV-ID Lite est proposé comme un modèle opérationnel pour des systèmes d'analyse audiovisuelle à la fois efficaces, rapides et respectueux de la vie privée dans les magasins sans personnel. Les perspectives envisagées ouvrent la voie à des applications plus complètes et à un déploiement à grande échelle, susceptibles d'avoir un impact important sur la sécurité, l'autonomie et la gestion intelligente des environnements commerciaux de demain.

5.7 CONCLUSION

Dans ce chapitre, nous avons évalué et discuté de la performance de l'approche proposée, AV-ID Lite, pour la détection d'incidents dans les magasins sans personnel. Les résultats expérimentaux ont montré une amélioration nette et cohérente des performances par rapport aux approches de l'état de l'art, tout en réduisant la complexité computationnelle et le temps d'entraînement.

L'étude d'ablation a également mis en évidence l'impact positif du mécanisme de pooling Log-Sum-Exp (LSE), qui permet de mieux exploiter les frames les plus informatives et d'améliorer la capacité du modèle à détecter des événements rares et discriminants.

Ces résultats confirment la pertinence de l'architecture proposée et démontrent qu'AV-ID Lite représente une solution efficace, rapide et adaptée aux contraintes des environnements de magasins autonomes.

CONCLUSION

Ce mémoire nous a permis d'étudier la détection de situations anormales dans les magasins sans personnel en exploitant des modèles multimodaux vision-audio. Après avoir présenté les fondements théoriques et les concepts clés de l'apprentissage profond, nous avons analysé les travaux existants liés à la détection d'anomalies. Nous nous sommes ensuite appuyés sur ces travaux pour proposer notre approche AV-ID Lite dédiée aux environnements de magasins autonomes.

L'objectif principal de ce travail est de détecter rapidement, tout en garantissant le respect de la vie privée des personnes, trois situations fréquentes dans les magasins sans personnel : les activités normales (ADL), les évanouissements ou chutes, et les bagarres. Afin d'atteindre cet objectif, nous avons apporté trois contributions majeures. Premièrement, nous proposons une approche combinant un extracteur de pose humaine squelettique pour la modalité visuelle et l'extracteur VGGish pour la modalité audio, ce qui permet de préserver la confidentialité des individus. Deuxièmement, nous avons introduit un pipeline de fusion précoce strictement synchronisée, associé à un Transformer léger équipé d'un mécanisme de pooling Log-Sum-Exp (LSE), permettant une analyse rapide et efficace des séquences multimodales, un aspect essentiel pour déclencher des alertes en temps réel lorsqu'un incident survient. Enfin, nous avons construit un nouveau jeu de données multimodal basé sur plusieurs jeux de données publics existants, couvrant trois types d'événements : ADL, évanouissement et bagarre.

Les expériences menées sur notre jeu de données démontrent que notre approche AV-ID Lite surpasse systématiquement les méthodes existantes sur l'ensemble des métriques d'évaluation, avec une exactitude (accuracy) de 96.15%, une précision de 94.20%, un rappel de 98.22%, un score F1 de 95.91%, un mAP de 99.23%, un MCC de 0.92 et un AUC-ROC

de 99.75%, tout en réduisant le temps d'entraînement du meilleur modèle à seulement 5.5 minutes.

Nous avons également réalisé une étude d'ablation comparant le pooling LSE au mean pooling (AVE), afin de mettre en évidence l'intérêt du pooling LSE dans ce contexte. Les résultats obtenus montrent que le pooling LSE est plus efficace que le mean pooling, notamment sur les métriques les plus sensibles telles que l'AUC-ROC et le mAP.

À travers ce mémoire, nous avons constaté que la détection de situations anormales dans les magasins sans personnel constitue une tâche particulièrement complexe. En effet, malgré les résultats encourageants obtenus, notre approche présente certaines limites. Tout d'abord, elle dépend fortement de la qualité de l'estimation de pose : des occlusions importantes, des scènes encombrées ou des erreurs de détection peuvent dégrader les performances. Ensuite, bien que le jeu de données construit soit diversifié, il repose principalement sur des bases de données publiques et non exclusivement sur des données collectées dans de véritables magasins, ce qui peut limiter la capacité de généralisation dans certains scénarios pratiques. Enfin, le système actuel cible un nombre limité de classes d'événements et ne couvre pas encore d'autres situations critiques telles que le vol, les comportements suspects ou les agressions verbales.

Les travaux futurs porteront sur l'exploration d'autres extracteurs de pose (tels que OpenPose ou BlazePose), l'étude de stratégies de fusion hybrides combinant fusion précoce et fusion tardive, ainsi que la création de jeux de données à grande échelle capturés dans des environnements commerciaux réels et couvrant un éventail plus large d'événements critiques.

BIBLIOGRAPHIE

Apicella, A. & Snidaro, L. (2021). Deep neural networks for real-time remote fall detection. Dans *International Conference on Pattern Recognition*, pp. 188–201. Springer.

Baltrušaitis, T., Ahuja, C. & Morency, L.-P. (2018). Multimodal machine learning : A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423–443.

Barua, A., Ahmed, M. U. & Begum, S. (2023). A systematic literature review on multimodal machine learning : Applications, challenges, gaps and future directions. *Ieee access*, *11*, 14804–14831.

Binte Rashid, M., Rahaman, M. S. & Rivas, P. (2024). Navigating the multimodal landscape : A review on integration of text and image data in machine learning architectures. *Machine Learning and Knowledge Extraction*, *6*(3), 1545–1563.

Canada, S. (2025). *Tendances relatives à la pénurie de main-d'œuvre au Canada*. https://www.statcan.gc.ca/fr/sujets-debut/travail_/tendances-penurie-main-oeuvre-canada. Accessed : 2026-02-17.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S. & Zhang, L. (2020). Higherhrnet : Scale-aware representation learning for bottom-up human pose estimation. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5386–5395.

Cheng, M., Cai, K. & Li, M. (2021). RWF-2000 : An open large scale video database for violence detection. Dans *2020 25th International conference on pattern recognition (ICPR)*, pp. 4183–4190. IEEE.

Chhetri, S., Alsadoon, A., Al-Dala'in, T., Prasad, P., Rashid, T. A. & Maag, A. (2021). Deep learning for vision-based fall detection system : Enhanced optical dynamic flow. *Computational Intelligence*, *37*(1), 578–595.

Dentamaro, V., Impedovo, D. & Pirlo, G. (2021). Fall detection by human pose estimation and kinematic theory. Dans *2020 25th international conference on pattern recognition (ICPR)*, pp. 2328–2335. IEEE.

Dif, N. (2020). *L'apprentissage profond pour le traitement des images*. *L'apprentissage profond pour le traitement des images*.

Duan, B., Tang, H., Wang, W., Zong, Z., Yang, G. & Yan, Y. (2021). Audio-visual event localization via recursive fusion by joint co-attention. Dans *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4013–4022.

Elharrouss, O., Almaadeed, N. & Al-Maadeed, S. (2021). A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, *77*, 103116.

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L. & Lu, C. (2022). Alphapose : Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, *45*(6), 7157–7173.

Fatima, M., Yousaf, M. H., Yasin, A. & Velastin, S. A. (2021). Unsupervised fall detection approach using human skeletons. Dans *2021 International Conference on Robotics and Automation in Industry (ICRAI)*, pp. 1–6. IEEE.

Fei, K., Wang, C., Zhang, J., Liu, Y., Xie, X. & Tu, Z. (2023). Flow-pose net : an effective two-stream network for fall detection. *The Visual Computer*, *39*(6), 2305–2320.

Galanakis, I., Soldatos, R. F., Karanikolas, N., Voulodimos, A., Voyiatzis, I. & Samarakou, M. (2025). Early and Late Fusion for Multimodal Aggression Prediction in Dementia Patients : A Comparative Analysis. *Applied Sciences*, *15*(11), 5823.

geeksforgeeks (2025). *Stacked RNNs in NLP*. https://www.geeksforgeeks.org/nlp/stacked-rnns-in-nlp/. Accessed : 2026-01-02.

Graves, A. & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, *18*(5-6), 602–610.

Guan, Z., Li, S., Cheng, Y., Man, C., Mao, W., Wong, N. & Yu, H. (2021). A video-based fall detection network by spatio-temporal joint-point model on edge devices. Dans *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 422–427. IEEE.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hossain, M., Sadik, K., Rahman, M. M., Ahmed, F., Bhuiyan, M. N. H. & Khan, M. M. (2021). Convolutional neural network based skin cancer detection (Malignant vs Benign). Dans *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0141–0147. IEEE.

Huang, H. & Jiang, Q. (2025). IDG-ViolenceNet : A Video Violence Detection Model Integrating Identity-Aware Graphs and 3D-CNN. *Sensors*, *25*(20), 6272.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Inturi, A. R., Manikandan, V. & Garrapally, V. (2023). A novel vision-based fall detection scheme using keypoints of human skeleton with long short-term memory network. *Arabian Journal for Science and Engineering*, *48*(2), 1143–1155.

IRIC (2024). *Une semaine d'apprentissage profond*. https://bioinfo.iric.ca/fr/une-semaine-dapprentissage-profond/#. Accessed : 2026-01-02.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G. & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666.

Jang, Y., Kim, N. & Chung, K. (2025). Enhancing Performance of Multimodal-Based Anomaly Detection using Early Fusion. Dans *2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 516–518. IEEE.

Japan, W. (2023). *Trends in Japan*. https://web-japan.org/trends/fr/tech-life/tec202309_unmanned-stores_fr.html. Accessed : 2026-02-17.

Kingma, D. P. & Ba, J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

Kreiss, S., Bertoni, L. & Alahi, A. (2021). Openpifpaf : Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 13498–13511.

Krithika, M., Gayathri, A. *et al.* (2024). Beyond Depth : Evaluating ResNet-50 and ResNet-152 Performance on Cifar-10 Dataset. Dans *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pp. 973–977. IEEE.

Li, G., Zhou, X., Huang, C., Lv, J., Zhang, H., Ji, D. & Wu, J. (2025). Diagnose Like a Doctor : A Vision-Guided Global–Local Fusion Network for Chest Disease Diagnosis. *International Journal of Imaging Systems and Technology*, *35*(5), e70203.

Li, P.-H., Fu, T.-J. & Ma, W.-Y. (2020). Why attention ? Analyze BiLSTM deficiency and its remedies in the case of NER. Dans *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, pp. 8236–8244.

Li, S. & Tang, H. (2024). Multimodal alignment and fusion : A survey. *arXiv preprint arXiv :2411.17040*.

Liang, P. P., Zadeh, A. & Morency, L.-P. (2024). Foundations & trends in multimodal machine learning : Principles, challenges, and open questions. *ACM Computing Surveys*, *56*(10), 1–42.

Lin, C.-B., Dong, Z., Kuan, W.-K. & Huang, Y.-F. (2020). A framework for fall detection based on OpenPose skeleton and LSTM/GRU models. *Applied Sciences*, *11*(1), 329.

Lin, J., Yu, D., Pan, R., Cai, J., Liu, J., Zhang, L., Wen, X., Peng, X., Cernava, T., Oufensou, S. *et al.* (2023). Improved YOLOX-Tiny network for detection of tobacco brown spot disease. *Frontiers in Plant Science*, *14*, 1135105.

Liu, Y., Liu, S., Zhu, X., Yang, H., Li, J., Guo, J., Teng, L., Yang, D., Wang, Y. & Liu, J. (2025). Privacy-preserving video anomaly detection : A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J. *et al.* (2019). Mediapipe : A framework for building perception

pipelines. *arXiv preprint arXiv :1906.08172*.

Mohammad, S. I., Owida, H. A., Vasudevan, A., Menon, S. V., Al-Hasnaawei, S., Ray, S., Talniya, N. C., Sinha, A., Jain, V. & Ranjbar, F. (2025). Thermophysical properties of used frying oil biodiesels blended with alcohols : Robust machine learning frameworks for density prediction. *Industrial Crops and Products*, *236*, 121916.

Niu, Z., Zhong, G. & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62.

Pineda, F. (1987). Generalization of back propagation to recurrent and higher order neural networks. Dans *Neural information processing systems*.

Ramachandra, B., Jones, M. J. & Vatsavai, R. R. (2020). A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, *44*(5), 2293–2312.

Ramirez, H., Velastin, S. A., Meza, I., Fabregas, E., Makris, D. & Farias, G. (2021). Fall detection and activity recognition using human skeleton features. *Ieee Access*, *9*, 33532–33542.

Raza, A., Yousaf, M. H., Ahmad, W., Velastin, S. A. & Viriri, S. (2025). Human fall detection using pose estimation : From traditional machine learning to vision transformers. *Engineering Applications of Artificial Intelligence*, *143*, 109809.

Ren, Q., Li, Y. & Liu, Y. (2023). Transformer-enhanced periodic temporal convolution network for long short-term traffic flow forecasting. *Expert Systems with Applications*, *227*, 120203.

Reza, S., Ferreira, M. C., Machado, J. J. & Tavares, J. M. R. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Systems with Applications*, *202*, 117275.

Snoek, C. G., Worring, M. & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. Dans *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402.

Sun, C., Paluri, M., Collobert, R., Nevatia, R. & Bourdev, L. (2016). Pronet : Learning to propose object-specific boxes for cascaded neural networks. Dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3485–3493.

Tang, H., Liu, H., Xiao, W. & Sebe, N. (2019). Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion. *Neurocomputing*, *331*, 424–433.

Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K. & Xu, C. (2022). Df-gan : A simple and effective baseline for text-to-image synthesis. Dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16515–16525.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vishnu, C., Datla, R., Roy, D., Babu, S. & Mohan, C. K. (2021). Human fall detection in surveillance videos using fall motion vector modeling. *IEEE Sensors Journal*, *21*(15), 17162–17170.

Wang, K., Li, X., Yang, J., Wu, J. & Li, R. (2021). Temporal action detection based on two-stream You Only Look Once network for elderly care service robot. *International Journal of Advanced Robotic Systems*, *18*(4), 17298814211038342.

Wang, Y., Dong, H., Bai, S., Yu, Y. & Duan, Q. (2024). Image recognition and classification of farmland pests based on improved Yolox-tiny algorithm. *Applied Sciences*, *14*(13), 5568.

Wang, Z., Ma, Y., Liu, Z. & Tang, J. (2019). R-transformer : Recurrent neural network enhanced transformer. *arXiv preprint arXiv :1907.05572*.

Wen, J., Abe, T. & Suganuma, T. (2022). Customer behavior recognition adaptable for changing targets in retail environments. Dans *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8. IEEE.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z. & Yang, Z. (2020). Not only look, but also listen : Learning multimodal violence detection under weak supervision. Dans *European conference on computer vision*, pp. 322–339. Springer.

Wu, P., Su, W., Pang, G., Sun, Y., Yan, Q., Wang, P. & Zhang, Y. (2025). AVad-CLIP : Audio-Visual Collaboration for Robust Video Anomaly Detection. *arXiv preprint arXiv :2504.04495*.

Yang, X., Wang, Y., Dong, Z., Li, J., Zhang, Q. & Qiang, S. (2025). Real-time fall detection algorithm based on FFD-AlphaPose and CTR–GCN. *Journal of Real-Time Image Processing*, *22*(3), 109.

Yeh, C.-F., Mahadeokar, J., Kalgaonkar, K., Wang, Y., Le, D., Jain, M., Schubert, K., Fuegen, C. & Seltzer, M. L. (2019). Transformer-transducer : End-to-end speech recognition with self-attention. *arXiv preprint arXiv :1910.12977*.

Yu, J., Liu, J., Cheng, Y., Feng, R. & Zhang, Y. (2022). Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. Dans *Proceedings of the 30th ACM international conference on multimedia*, pp. 6278–6287.

Zheng, Y., Li, X., Xie, F. & Lu, L. (2020). Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer. Dans *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6734–6738. IEEE.

Zhou, H., Yu, J. & Yang, W. (2023). Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 3769–3777.

Zhou, X., Peng, X., Wen, H., Luo, Y., Yu, K., Yang, P. & Wu, Z. (2024). Learning weakly supervised audio-visual violence detection in hyperbolic space. *Image and Vision Computing*, *151*, 105286.