UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

AVERTISSEMENT
DROIT D'AUTEUR

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document. Toute contrefaçon, plagiat ou reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR



Analyse conjointe des attributs et de l'identité en vision par ordinateur à l'aide d'un modèle vision-langage

Thèse

présentée et soutenue publiquement le 17 décembre 2025
pour l'obtention du

Doctorat de l'université de Lorraine
(mention informatique)

par
Mehran ADIBI SEDEH

Président **:**	El-hadi Djermoune
Rapporteurs **:**	Rémi Emonet
	Alexandre Benoit
Examinateurs **:**	Diarra Fall
	Gilles Gasso
Invité **:**	Frédéric Sur
Directeurs de thèse **:**	Cédric Pradalier et Marianne Clausel

---

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Table des matières

Table des figures v

Chapitre 1
Introduction générale

1.1 Contexte scientifique et industriel 5
1.2 Problématiques liées à l'analyse de personnes 7
1.3 Limites des approches classiques 8
1.4 Hypothèse de travail et intuition scientifique 9
1.5 Objectifs de la thèse 10
1.6 Plan du manuscrit 11
Publications associées à cette thèse 12

Chapitre 2
Fondamentaux de l'apprentissage pour la vision par ordinateur

2.1 Fondamentaux de l'apprentissage pour la vision par ordinateur 13
2.1.1 Apprentissage supervisé 13
2.1.2 Apprentissage non supervisé 16
2.1.3 Apprentissage auto-supervisé 17
2.1.4 Synthèse des paradigmes d'apprentissage 19
2.2 Architectures fondamentales 20
2.2.1 Réseaux de neurones convolutionnels (CNN) 20
2.2.2 Introduction aux transformeurs 25
2.2.3 Principe du self-attention 26
2.2.4 Vision Transformer (ViT) 30
2.3 Modèles vision-langage 31
2.3.1 CLIP et dérivés 31
2.3.2 Développements récents et adaptations à la PAR et à la ReID . . 33
2.3.3 Synthèse critique : CNNs, ViT et modèles vision-langage 34
2.4 Optimisation des modèles profonds 35

2.4.1 Fonctions de perte en classification 35
2.4.2 Stratégies d'optimisation des réseaux profonds 38
2.4.3 Améliorations pratiques et stabilité d'entraînement 41
2.4.4 Surapprentissage et régularisation 42
2.4.5 Métriques d'évaluation 45

Chapitre 3
AttriVision : La reconnaissance d'attributs avec CLIP

3.1 Introduction 49
3.2 Travaux connexes 51
3.2.1 Perte *Focal Cross-Entropy* (FCE) 52
3.3 Implémentation 54
3.3.1 Détails d'implémentation 54
3.3.2 Garantir la cohérence des données entre jeux de données 54
3.3.3 Augmentation des données et variabilité des phrases 55
3.3.4 Configuration du modèle 56
3.3.5 Processus d'évaluation 56
3.3.6 Résultats 57
3.3.7 Indépendance vis-à-vis du domaine 58
3.4 Étude d'ablation 59
3.4.1 Analyse 59
3.4.2 Représentation efficace des attributs 60
3.5 Conclusion 60

Chapitre 4
Re-Identification conjointe des personnes et de leurs attributs

4.1 Introduction 63
4.2 État de l'art 64
4.2.1 Introduction générale à la tâche de Re-Identification 64
4.2.2 Méthodes traditionnelles à base de descripteurs manuels 66
4.2.3 Apprentissage métrique et pertes contrastives 69
4.2.4 Attention et Transformers 71
4.2.5 Synthèse critique et limites des approches actuelles 77
4.2.6 Méthodes conjointes d'attributs et de Re-Identification 80
4.2.7 Intégration de modèles vision–langage dans la ReID 82
4.2.8 Conclusion : nécessité d'un apprentissage conjoint 84
4.3 Méthodologie proposée 85

4.3.1 Formulation du problème de ReID multi-attributs 85
4.3.2 Encodage partagé et structuration par centroïdes attributaires . . 86
4.3.3 Construction des triplets guidée par les attributs 88
4.3.4 Principes sous-jacents 89
4.3.5 Génération des centroïdes attributaires 89
4.3.6 Calcul de la Triplet Loss entre centroïde et images 90
4.3.7 Avantage d'un apprentissage en une seule étape 92
4.4 Configuration d'entraînement détaillée 92
4.4.1 Hyperparamètres d'entraînement 92
4.4.2 Optimiseur et stratégie de régularisation 92
4.4.3 Gestion de la précision mixte et du GPU 93
4.5 Jeux de données utilisés 93
4.5.1 Description des jeux de données 93
4.5.2 Protocole expérimental 94
4.5.3 Justification des choix dans un cadre multi-tâches 95
4.6 Résultats expérimentaux et analyses approfondies 95
4.6.1 Résultats quantitatifs 95
4.6.2 Recherche d'images par image témoin 97
4.6.3 Études d'ablation 98
4.6.4 Analyse de sensibilité aux hyperparamètres 99
4.6.5 Analyse qualitative 100
4.6.6 Robustesse hors-distribution 101
4.7 Discussion approfondie des résultats 101
4.7.1 Limitations identifiées 102
4.7.2 Perspectives méthodologiques 103
4.7.3 Synthèse 103
4.8 Conclusion du chapitre 104

Chapitre 5
Conclusion générale et perspectives

Récapitulatif des contributions 105
Modélisation conjointe des attributs et de l'identité 105
Centroïdes attributaires pour la structuration métrique 105
Standardisation inter-jeux et protocole d'évaluation 106
Implémentation expérimentale et validation 106
Extension à d'autres catégories d'objets (étude de cas : véhicules) 106
Discussion critique des résultats 107

Conclusion finale 109

Bibliographie **111**

Table des figures

2.1 Illustration d'une séparation linéaire entre deux classes de données dans un espace tridimensionnel. Chaque couleur représente une classe différente, et le plan coloré représente l'hyperplan de décision appris par un classifieur. 14

2.2 Visualisation des embeddings dans un espace 3D pour trois catégories distinctes. Chaque point représente une instance projetée dans l'espace latent, colorée selon sa classe. Les plans colorés représentent les hyperplans de séparation calculés par un classifieur linéaire (LinearSVC) en mode un-versus-le-reste (OvR), illustrant les régions de décision apprises. Cette figure met en évidence la séparabilité linéaire des classes dans l'espace projeté. 15

2.3 Évolution de la distribution des embeddings au cours de l'entraînement. On observe une structuration progressive de l'espace latent avec un regroupement par classe de plus en plus marqué. 17

2.4 Projection PCA des embeddings visuels extraits par CLIP pour 1000 images. Chaque point représente une image, colorée selon l'attribut "genre" (*man* en bleu, *woman* en rouge). 18

2.5 Architecture d'un auto-encodeur feed-forward. L'encodeur projette l'entrée $\mathbf{x}$ dans un espace latent $\mathbf{z}$, puis le décodeur reconstruit $\hat{\mathbf{x}}$. 18

2.6 Illustration du mécanisme de convolution 2D sur une image issue du jeu de données [1] : un filtre 3×3 (encadré bleu foncé) est appliqué à partir du coin supérieur gauche de l'image, puis se déplace de manière convolutive. 22

2.7 CNN inspiré de VGG sur deux lignes : convolution/pooling (ligne 1), puis Flatten, couches entièrement connectées et Softmax (ligne 2). Les sous-textes indiquent les dimensions ou tailles de sorties. 23

2.8 Architecture d'un transformeur déplié temporellement. Chaque bloc contient une attention multi-tête suivie d'un réseau feed-forward, avec normalisation et connexions résiduelles. 27

2.9 Schéma simplifié de l'attention multi-tête : l'entrée se divise en Q, K et V, qui sont combinés par la *Multi-Head Attention* pour produire une sortie contextualisée. 28

2.10 Effet du paramètre de focalisation γ sur la Focal Loss. Lorsque $\gamma = 0$, la Focal Loss correspond à l'entropie croisée classique. Des valeurs plus élevées de γ diminuent l'impact des exemples faciles (avec une probabilité prédite élevée pour la classe correcte) et concentrent l'apprentissage sur les exemples difficiles. 37

2.11 Visualisation des trajectoires d'optimisation des algorithmes *SGD*, *Adam* et *AdamW* sur une surface de perte non convexe. Ce graphique illustre les différences de comportement entre ces méthodes : SGD progresse avec des oscillations liées à l'absence d'adaptation du taux d'apprentissage, tandis qu'Adam et AdamW suivent des trajectoires plus lisses grâce à leur mécanisme de moyenne mobile. AdamW, en particulier, dissocie le terme de régularisation du gradient, ce qui lui confère une meilleure stabilité. . . . 40

2.12 Comparaison simulée des courbes de perte pour l'entraînement en précision **FP32** et **FP16**. La précision mixte (FP16) permet une convergence plus rapide dans les premières étapes, mais introduit une légère instabilité et un plateau de performance plus élevé, reflet des limitations numériques potentielles (par exemple : perte de précision dans le calcul de gradients de faible amplitude, entraînant un apprentissage moins efficace sur certaines couches du réseau). Ce phénomène est notamment documenté dans les travaux de [2]. 43

2.13 Exemple de surapprentissage : la perte d'entraînement continue de diminuer tandis que la perte de validation augmente après un certain nombre d'époques, signalant une perte de généralisation. 46

3.1 Comparaison entre notre approche et CLIP : pour une image donnée d'un lot, nous extrayons les plongements visuels de l'image ainsi que ceux des N attributs associés, puis nous faisons correspondre chaque paire *(images, attribute)*. 52

4.1 Architecture CLIP adaptée pour l'apprentissage conjoint attributs–ReID avec centroïdes et double supervision. 90

4.2 Impact des hyperparamètres sur les performances (mAP et Rank-1) sur le dataset Market1501. Trois paramètres sont analysés : λ (pondération de la perte Focal Cross-Entropy), α (marge de la triplet loss), et γ (paramètre de focalisation de la FCE). 99

4.3 Résultats qualitatifs sur le dataset Market1501. Chaque ligne montre une image requête (colonne "Query") accompagnée des cinq images les plus similaires selon deux branches du modèle : PAR et ReID. 100

4.4 Évolution des performances du modèle selon différentes variantes des données (résolution, luminosité, orientation). 101

Remerciements

Ce travail marque l'aboutissement d'un parcours aussi riche qu'imprévisible, jalonné de nombreux défis, parfois complexes, mais toujours porteurs d'apprentissage. Ces années ont été faites de hauts et de bas, souvent de ralentissements imposés plus que choisis, mais jamais d'abandon. Chaque étape, chaque obstacle a contribué à construire ce chemin jusqu'à son terme.

Une gratitude profonde va à Romain Martin, présent dès l'origine, lorsque l'idée de me lancer dans une thèse a émergé. Son écoute, sa confiance et sa disponibilité ont été un repère constant, en particulier dans les moments où la voie semblait incertaine.

Impossible également de ne pas évoquer Julien Trombini et Guillaume Cazenave, mes responsables dans mon ancienne entreprise, dont l'ouverture d'esprit et la vision ont rendu cette aventure possible. Défendre le lien entre la recherche académique et le monde industriel n'est pas chose aisée pour une structure de petite taille ; leur confiance et leur volonté d'encourager cette collaboration ont permis à ce projet de voir le jour à l'interface de deux univers souvent éloignés.

Reconnaissance également envers mes collègues Salim et Louis, dont les échanges ont toujours été stimulants et d'une grande richesse. Leur curiosité, leur rigueur et leur sens du partage ont largement contribué à nourrir la réflexion et à maintenir l'envie d'avancer.

Une pensée toute particulière pour Marianne Clausel et Cédric Pradalier, mes directeurs de thèse, dont la bienveillance, la compréhension et l'humanité ont profondément marqué ces années. Leur accompagnement, à la fois exigeant et attentif, a élargi ma vision scientifique et donné une cohérence durable à ce travail.

À ma famille, une reconnaissance infinie. Leur confiance, leur persévérance et leur humour ont été des moteurs essentiels. L'idée de devenir "docteur" leur appartenait bien avant qu'elle ne devienne la mienne — portée avec l'insistance affectueuse et la détermination légendaire des parents iraniens.

Enfin, à ma moitié, Marie, va toute ma gratitude et mon amour. Sa présence, sa patience et sa force tranquille ont donné sens à cette dernière ligne droite. Sans elle, rien de tout cela n'aurait eu la même portée ni la même lumière.

Résumé / Abstract

Titre de la thèse :

Analyse conjointe des attributs et de l'identité en vision par ordinateur à l'aide d'un modèle vision-langage

Résumé (français) :

L'identification et l'analyse automatique de personnes dans des environnements urbains surveillés constituent aujourd'hui un enjeu majeur en vision par ordinateur, notamment dans le cadre des systèmes de vidéoprotection intelligents. Deux tâches essentielles se distinguent : la reconnaissance d'attributs visuels (*Person Attribute Recognition*, PAR) et la re-identification de personnes (*Person Re-Identification*, ReID) à travers un réseau de caméras hétérogènes. Si ces deux problèmes ont historiquement été abordés de manière indépendante, la littérature récente suggère que leurs complémentarités peuvent être exploitées de manière conjointe pour améliorer la robustesse des modèles et leur capacité de généralisation.

Dans ce contexte, cette thèse propose une architecture d'apprentissage multimodale fondée sur le modèle vision-langage CLIP, permettant d'intégrer simultanément une supervision sémantique issue des attributs (sous forme de textes) et une supervision identitaire issue des identifiants de personne. L'approche repose sur un apprentissage unifié, sans pipeline séquentiel, dans lequel chaque image est associée à des descriptions attributaires textuelles encodées via un encodeur de texte, et à des représentations visuelles extraites par un encodeur d'image. À partir de cette correspondance croisée, nous introduisons le concept de *centroïdes attributaires*, obtenus comme moyenne des embeddings textuels associés à une image ancre. Ces centroïdes sont utilisés comme référence intermédiaire dans une fonction de perte de type Triplet Loss, permettant de rapprocher les images positives (même identité) et d'éloigner les images négatives (identité différente mais attributs visuellement proches), avec une métrique de similarité cosinus.

Notre pipeline d'entraînement unifié combine ainsi deux objectifs complémentaires : une perte Focal Cross-Entropy pour la classification attributaire supervisée (multi-label), et une Triplet Loss adaptée calculée entre le centroïde

et les images positives et négatives. Cette formulation permet d'optimiser un espace latent partagé, dans lequel les similarités sémantiques et visuelles sont alignées. Le modèle est entraîné sur plusieurs jeux de données publics majeurs en reconnaissance de personnes : **PETA**, **RAPv2** et **UPAR**. Les résultats expérimentaux montrent que l'intégration des attributs comme supervision intermédiaire améliore les performances en ReID (mAP, CMC@1), en particulier dans des situations de forte variabilité d'apparence (changement de vêtements, angles de vue, caméras). Une analyse d'ablation valide l'apport des centroïdes, ainsi que l'importance de la pondération des pertes pendant l'apprentissage.

Par ailleurs, ce travail intègre des considérations d'optimisation applicables en contexte industriel : gestion mémoire avec entraînement en précision mixte (FP16), stabilité par clipping de gradient, ajustement dynamique du poids des objectifs, et stratégie de validation continue. Le modèle proposé offre une meilleure explicabilité, grâce à la trace des attributs textuels et aux représentations visuelles interprétables.

En définitive, cette thèse démontre l'intérêt d'un cadre d'apprentissage couplé entre vision et langage pour la reconnaissance de personnes, et ouvre la voie à des extensions vers l'apprentissage faiblement supervisé, l'adaptation de domaine, et l'intégration dans des systèmes réels de vidéosurveillance urbaine.

Title of the thesis :

Joint Attribute and Identity Modeling in Computer Vision Using a Vision-Language Framework

Abstract (English) :

The automatic identification and analysis of individuals in urban surveillance environments has become a major challenge in computer vision, particularly within the context of intelligent video surveillance systems. Two key tasks are central to this domain : Person Attribute Recognition (PAR) and Person Re-Identification (ReID) across a network of heterogeneous cameras. Although traditionally studied independently, recent research suggests that the complementarity between these tasks can be jointly exploited to improve model robustness and generalization.

This thesis proposes a multimodal learning architecture based on the CLIP vision-language model, enabling simultaneous integration of semantic supervision through textual attribute descriptions and identity supervision through person identifiers. Our approach relies on a unified training framework — as opposed to sequential pipelines — where each image is associated with textual attribute descriptions encoded via a text encoder, and visual features extracted via an image encoder. From this cross-modal alignment, we introduce the concept of *attribute centroids*, computed as the mean of the textual embeddings associated with an anchor image. These centroids act as semantic anchors in a Triplet Loss formulation, bringing positive images (same identity) closer and pushing negative images (different identity but visually similar attributes) further apart, using a cosine similarity metric.

The proposed training pipeline combines two complementary objectives : a Focal Cross-Entropy loss for supervised multi-label attribute classification, and an adapted Triplet Loss between the centroid and the positive/negative images. This formulation optimizes a shared latent space where semantic and visual similarities are coherently aligned. The model is trained on major public datasets in pedestrian analysis : **PETA**, **RAPv2**, and **UPAR**. Experimental results demonstrate that incorporating attribute-level supervision improves ReID performance (in terms of mAP and CMC@1), particularly in challenging scenarios involving clothing changes, pose variations, or multi-camera viewpoints. An ablation study confirms the contribution of the centroids and the importance of dynamically balancing the losses during training.

In addition, the proposed framework incorporates practical optimizations relevant to industrial deployment : memory-efficient mixed-precision training (FP16), gradient clipping for stability, dynamic loss weighting, and conti-

nuous validation strategies. The joint vision-language architecture also enhances model interpretability, leveraging both textual traces and structured visual embeddings.

In summary, this thesis demonstrates the relevance of coupled vision-language learning for person recognition and opens the way for future extensions in weak supervision, domain adaptation, and deployment in real-world urban video surveillance systems.

Mots-clés : reconnaissance d'attributs, re-identification, vision par ordinateur, modèle vision-langage, CLIP, supervision sémantique.

Keywords : attribute recognition, re-identification, computer vision, vision-language model, CLIP, semantic supervision.

1

Introduction générale

1.1 Contexte scientifique et industriel

La vision par ordinateur est devenue un pilier technologique des systèmes intelligents dans des domaines variés tels que la robotique, la mobilité autonome, la santé ou encore la surveillance visuelle. En particulier, la *vidéoprotection* — c'est-à-dire l'analyse automatisée d'images issues de réseaux de caméras de surveillance — connaît une croissance significative, portée par les besoins croissants en sécurité urbaine, prévention d'incidents et gestion des flux de personnes dans des lieux sensibles (gares, aéroports, centres commerciaux, etc.) [3, 4, 5].

Cadre CIFRE et partenariat industriel. Cette thèse est conduite dans le cadre d'une convention CIFRE (*Conventions Industrielles de Formation par la REcherche*) en partenariat avec l'entreprise **Two-i**, financeur industriel du projet. Ce cadre confère un double ancrage académique–industriel : (i) il oriente les questions de recherche vers des cas d'usage réels (vidéoprotection urbaine, analyse de foules, gestion de flux multi-sites) ; (ii) il impose des contraintes opérationnelles et réglementaires fortes — latence et calcul embarqué, robustesse en environnements non coopératifs, scalabilité à large échelle, frugalité énergétique, ainsi que conformité RGPD/CNIL et protection de la vie privée (anonymisation, minimisation des données). Dans ce contexte, les contributions proposées sont conçues pour être transférables : elles sont évaluées sur des jeux de données publics et, lorsque cela est possible, sur des données industrielles anonymisées représentatives des scénarios cibles ; les développements logiciels visent une intégration progressive dans les chaînes d'analyse de **Two-i**, depuis la preuve de concept jusqu'à des briques réutilisables en production.

Face à la masse de données générées par ces systèmes (des centaines d'heures de vidéos par jour dans une grande ville), il devient indispensable de s'appuyer sur des solutions d'analyse automatique capables de détecter, caractériser, suivre, et parfois ré-identifier

les individus dans des contextes denses et non coopératifs. Deux tâches fondamentales émergent dans ce cadre :

— La **reconnaissance d'attributs visuels de personnes** (*Person Attribute Recognition*, PAR), qui consiste à prédire des caractéristiques comme le sexe, l'âge apparent, les types et couleurs de vêtements, la présence d'un sac ou d'accessoires. Cette tâche est souvent formulée comme une classification multi-label, chaque attribut correspondant à une variable binaire indépendante [6, 7].

— La **re-identification de personnes** (*Person Re-Identification*, ReID), qui vise à reconnaître une personne à travers plusieurs caméras non chevauchantes, sans supervision biométrique directe. Le système apprend un espace de représentation dans lequel les individus identiques sont rapprochés, malgré les variations de pose, d'apparence, de luminosité ou de point de vue [8, 9].

Dans un environnement de vidéoprotection, ces deux tâches ne sont pas dissociées. Une chaîne d'analyse robuste devra souvent recourir simultanément à une description attributaire pour la recherche ou le filtrage, et à une re-identification fine pour le suivi inter-caméras. Cependant, les approches traditionnelles les traitent de façon indépendante, en négligeant les redondances et les synergies potentielles entre les deux objectifs [10].

Les architectures classiques à base de réseaux de neurones convolutionnels (CNN) ont permis des progrès importants en reconnaissance visuelle. Toutefois, elles souffrent de deux limites majeures : (1) leur capacité à extraire des relations sémantiques complexes entre concepts visuels est limitée ; (2) les attributs sont souvent considérés comme des cibles indépendantes, sans hiérarchie ou co-occurrence modélisée.

Dans ce contexte, l'émergence de modèles **vision-langage** comme CLIP (*Contrastive Language–Image Pretraining*) [11] offre une nouvelle approche. CLIP apprend un espace latent commun entre images et textes, à partir d'un apprentissage contrastif sur des centaines de millions de paires image–texte. Il devient alors possible de formuler une supervision en langage naturel pour guider les représentations visuelles, et d'aligner directement des descriptions attributaires avec des images. Cette approche ouvre la voie à des systèmes plus robustes, plus explicables, et potentiellement plus généralisables à de nouveaux environnements visuels.

Dans cette thèse, nous défendons l'idée que l'apprentissage des attributs peut servir de pivot sémantique pour renforcer l'identité visuelle d'un individu. À ce titre, nous proposons un cadre d'apprentissage conjoint exploitant CLIP, combinant une supervision attributaire en langage naturel et une supervision par identifiants (PID) pour la ReID. Cette approche s'inscrit dans une logique d'intégration forte entre reconnaissance sémantique et identification visuelle, en phase avec les exigences opérationnelles de la vidéoprotection.

1.2 Problématiques liées à l'analyse de personnes

L'analyse automatisée de personnes dans des flux vidéo pose des défis multiples, qui se manifestent à la fois dans les tâches de reconnaissance d'attributs (PAR) et de re-identification (ReID). Ces défis sont amplifiés par les conditions spécifiques aux environnements de vidéoprotection : caméras fixes ou mobiles, angles variés, résolutions hétérogènes, conditions météo fluctuantes, et densité variable des scènes.

Dans la reconnaissance d'attributs, chaque image peut être annotée avec un ensemble de caractéristiques visuelles. Cette annotation est souvent bruitée, sujette à l'interprétation humaine, voire absente pour certaines classes. Les bases de données existantes (ex. PETA [12], RAPv2 [13], PA100k [14], UPAR [15]) sont déséquilibrées : certains attributs sont largement majoritaires (par ex. "porte un pantalon"), d'autres très minoritaires ("tient un parapluie"). Ce déséquilibre rend l'entraînement difficile, et requiert des stratégies de pondération ou d'échantillonnage spécifiques [6, 7].

La re-identification, quant à elle, est une tâche d'appariement inter-caméras où la difficulté provient de la variabilité intra-classe : une même personne peut apparaître sous des angles très différents, porter des objets différents, ou être partiellement occultée. La variabilité inter-classe est également problématique : des personnes distinctes peuvent avoir une apparence très similaire, notamment dans des environnements homogènes (uniformes, tenues professionnelles, etc.). Ces aspects rendent la tâche particulièrement sujette aux erreurs de confusion [7, 16].

Par ailleurs, les deux tâches sont souvent traitées de manière indépendante, malgré leur interdépendance sémantique. Les attributs peuvent pourtant constituer une information précieuse pour affiner la similarité identitaire. Par exemple, deux images présentant des individus de même gabarit, même couleur de cheveux et mêmes vêtements ont de fortes chances de représenter la même personne. Cependant, si l'un d'eux porte un sac à dos rouge et l'autre non, cette différence peut être discriminante. Exploiter ces informations nécessite des modèles capables de raisonner à la fois au niveau visuel et sémantique.

Dans les approches classiques, la supervision attributaire est souvent réduite à une classification multi-label indépendante, ignorant les co-occurrences ou les relations hiérarchiques entre attributs [14, 17]. De même, la re-identification repose typiquement sur des pertes contrastives (triplet loss, circle loss, etc.) qui maximisent la distance entre descripteurs, mais sans alignement explicite avec des propriétés sémantiques observables [18]. Cette séparation méthodologique limite la capacité à construire des représentations globales robustes.

Un autre défi central réside dans la **généralisation** des modèles. Dans des contextes industriels, les modèles doivent être déployés dans des environnements variés, avec peu ou pas de données annotées. Cette variabilité de domaine est une source majeure de dégradation des performances ("domain gap") [19]. Les approches récentes cherchent donc

à concevoir des modèles plus robustes au transfert de domaine, souvent en intégrant des signaux externes comme le langage [11, 20].

En résumé, l'analyse de personnes dans les systèmes de vidéoprotection combine plusieurs difficultés :

— annotations bruitées et classes déséquilibrées en PAR ;
— forte variabilité intra-classe et ambigüité inter-classe en ReID ;
— séparation artificielle entre supervision sémantique et supervision identitaire ;
— difficulté à généraliser sans adaptation sur de nouveaux domaines.

Ces limites justifient l'exploration de paradigmes d'apprentissage conjoints, exploitant simultanément des signaux visuels et sémantiques, comme ceux offerts par les modèles vision-langage.

1.3 Limites des approches classiques

Les approches traditionnelles pour la reconnaissance d'attributs (PAR) et la re-identification de personnes (ReID) reposent essentiellement sur des architectures convolutionnelles entraînées de manière supervisée. Ces stratégies ont certes permis des avancées notables en vision par ordinateur, mais elles présentent des limitations structurelles qui restreignent leur efficacité dans des environnements complexes comme ceux rencontrés en vidéoprotection.

Dans les systèmes de PAR, la stratégie classique consiste à associer à chaque attribut un classifieur binaire indépendant, généralement construit sur les dernières couches d'un CNN tel que ResNet ou MobileNet. Cette décomposition ignore les relations d'interdépendance ou de co-occurrence fréquentes entre attributs (ex. "porte une jupe" est souvent associé à "femme") et n'exploite pas la structure sémantique des propriétés visuelles [6, 7]. De plus, la supervision attributaire se heurte à des étiquettes parfois bruitées, ambiguës ou subjectives, ce qui impacte la qualité de l'apprentissage.

Côté ReID, les modèles classiques apprennent une fonction d'embedding à partir d'images entières ou découpées en parties, en utilisant des pertes contrastives telles que la *Triplet Loss*, la *Softmax loss* ou des variantes comme la *Circle Loss* [21, 22]. Si ces approches permettent de projeter les images d'une même identité dans un voisinage latent, elles restent fortement dépendantes de la diversité des poses et de la qualité des échantillons d'entraînement. L'apprentissage repose le plus souvent sur une seule modalité (visuelle) sans recours à des descripteurs textuels ou contextuels.

Par ailleurs, les tentatives de combiner PAR et ReID dans un cadre multi-tâches se heurtent à plusieurs obstacles :

— Les objectifs sont parfois conflictuels : la classification des attributs favorise des représentations sémantiquement structurées, tandis que la ReID nécessite une discrimination maximale entre identités visuellement proches.

— Les pertes associées à chaque tâche ne sont pas triviales à pondérer et leur co-optimisation peut conduire à des compromis sous-optimaux [10].
— Les représentations apprises restent ancrées dans un espace visuel peu interprétable, ce qui limite leur transférabilité à d'autres domaines ou tâches.

Enfin, un problème central réside dans le manque d'alignement explicite entre les attributs et l'identité. Deux individus peuvent partager un grand nombre d'attributs sans être la même personne. À l'inverse, une même personne peut apparaître sous des tenues très différentes à travers les caméras. Les modèles actuels ne disposent pas de mécanismes explicites pour intégrer cette ambivalence sémantique dans l'apprentissage de leurs représentations.

Dans un contexte de surveillance à large échelle, où les identités sont souvent inconnues et les annotations rares ou bruitées, ces limitations deviennent critiques. Il est donc nécessaire d'explorer des approches capables d'unifier la supervision attributaire et identitaire, dans un cadre plus souple et sémantiquement structuré.

C'est précisément ce que permettent les modèles vision-langage récents comme CLIP [11], qui projettent simultanément des images et des textes dans un même espace latent. En exploitant ces capacités de projection croisée, il devient possible de redéfinir les relations entre images, attributs et identités, et d'intégrer plus finement la structure sémantique du problème dans l'apprentissage.

1.4 Hypothèse de travail et intuition scientifique

Dans un système de vidéoprotection opérationnel, il est essentiel de pouvoir relier efficacement des personnes à partir d'indices visuels hétérogènes, qu'ils soient exprimés sous forme d'images, de requêtes textuelles ("homme en veste rouge avec un sac à dos noir") ou d'attributs annotés. Cela suppose de dépasser l'approche strictement discriminative centrée sur les identités visuelles, pour intégrer une composante sémantique plus explicite dans les représentations apprises.

L'hypothèse centrale que nous posons dans ce travail est que les attributs visuels d'une personne — typiquement utilisés pour la classification multi-label — peuvent être mobilisés comme **descripteurs latents intermédiaires** servant à guider l'apprentissage des représentations identitaires. Ces attributs textuels, en tant que vecteurs sémantiques, sont capables de structurer l'espace d'embedding de manière plus robuste, en y injectant une géométrie sémantique complémentaire à celle des descripteurs purement visuels.

Ce positionnement se situe à la croisée de deux tendances récentes :

— D'une part, les travaux en *attribute-based re-identification* montrent que les propriétés sémantiques (couleurs, vêtements, accessoires) peuvent être exploitées pour construire des descripteurs plus stables et explicables [7, 10, 19].
— D'autre part, les modèles vision-langage comme CLIP [11] permettent d'aligner

efficacement des images et des descriptions textuelles dans un espace partagé. Ce paradigme rend possible l'apprentissage à partir de paires image–attributs, sans nécessiter une supervision pixel par pixel ou strictement catégorielle.

Nous faisons donc l'hypothèse qu'un modèle entraîné simultanément :

1. à rapprocher les images de leurs attributs textuels (**apprentissage sémantique**),
2. à discriminer des identités à travers des images positives et négatives (**apprentissage métrique**),

sera plus robuste et généralisera mieux dans des contextes non supervisés ou faiblement annotés. Ce cadre permet de construire des représentations jointes où les attributs structurent l'espace des identités, et où les identités servent de régularisation pour éviter les confusions sémantiques.

Afin de rendre cette intuition exploitable en pratique, nous proposons une architecture fondée sur CLIP, étendue par :

— une tête de projection adaptative pour la prédiction attributaire supervisée (via Focal Cross-Entropy),
— un module de génération de **centroïdes attributaires**, calculés à partir des embeddings textuels de chaque image,
— une fonction de **Triplet Loss entre centroïdes et images** positives/négatives, favorisant un ancrage sémantique dans l'espace latent.

Ce schéma vise à combiner la précision locale des descripteurs visuels avec la structure globale imposée par les attributs. Il s'inscrit dans la continuité de travaux comme Cerberus [23], MP-ReID [24] et AGPR [25], tout en apportant une simplification computationnelle : **un seul pipeline d'entraînement**, dans lequel les supervisions attributaire et identitaire sont co-optimisées au sein d'un même modèle.

1.5 Objectifs de la thèse

L'objectif global de cette thèse est de proposer une approche unifiée pour la reconnaissance d'attributs et la re-identification de personnes, intégrée dans un cadre vision-langage fondé sur le modèle CLIP. Ce travail s'inscrit dans le contexte industriel de la vidéoprotection, où la capacité à extraire automatiquement des informations sémantiques et à suivre des individus à travers plusieurs caméras constitue un enjeu majeur de sécurité et de performance opérationnelle.

Plus précisément, nous visons à concevoir, implémenter et évaluer une architecture d'apprentissage conjointe permettant :

1. de projeter des images de personnes et leurs attributs textuels dans un espace sémantique commun,

2. de structurer cet espace à l'aide de centroïdes attributaires définis comme moyenne des représentations textuelles associées à chaque image ancre,
3. d'optimiser simultanément l'alignement image–texte (via Focal Cross-Entropy) et la discrimination inter-identité (via Triplet Loss),
4. de garantir une compatibilité métrique des embeddings, facilitant leur utilisation dans des tâches de recherche ou d'appariement inter-caméras,
5. et de valider la pertinence du modèle sur des jeux de données multi-attributs et multi-identités, en tenant compte de la scalabilité et des contraintes de généralisation.

Ce positionnement se distingue des approches existantes par sa volonté de réunir, au sein d'un même modèle, la supervision sémantique (description en langage naturel) et la supervision identitaire (person ID), tout en exploitant les capacités de généralisation et de transfert du modèle CLIP. Contrairement aux méthodes à pipeline séquentiel — où les attributs sont prédit d'abord puis exploités pour guider la ReID —, nous proposons un entraînement unique, où chaque lot (batch) combine des images ancres, positives, négatives, et leurs attributs textuels, en un apprentissage commun.

Ce cadre ouvre plusieurs perspectives :

— une meilleure robustesse aux variations d'apparence (vue, pose, éclairage),
— une meilleure interprétabilité des résultats grâce à la structure sémantique de l'espace latent,
— une réduction de la dépendance aux annotations identitaires exhaustives,
— et un potentiel de déploiement plus direct dans des environnements complexes ou non annotés.

En résumé, cette thèse ambitionne de poser les bases d'un modèle multimodal sémantique pour l'analyse de personnes, exploitant à la fois les attributs comme descriptions textuelles et les identifiants comme supervision métrique, dans une optique de réutilisabilité, robustesse et explicabilité.

1.6 Plan du manuscrit

Ce manuscrit est structuré en plusieurs chapitres, organisés de manière progressive, du cadre théorique jusqu'aux contributions expérimentales proposées.

— **Chapitre 2** — État de l'art des approches d'apprentissage profond en vision par ordinateur : rappels sur l'apprentissage supervisé, non supervisé, présentation des architectures convolutionnelles (CNN), des mécanismes d'attention (transformers) et des modèles vision–langage récents (dont CLIP).
— **Chapitre 3** — Reconnaissance d'attributs de personnes (PAR) : jeux de données, méthodes classiques basées sur les CNN, approches récentes exploitant la

co-occurrence d'attributs, l'attention et les modèles vision–langage.

— **Chapitre 4** — Apprentissage conjoint guidé par les attributs pour la ré-identification : combinaison Focal Loss / Triplet Loss dans un espace latent aligné image–texte, avec l'introduction de centroïdes attributaires comme ancrages sémantiques pour la ReID.

— Enfin, le **Chapitre 5** conclut la thèse en récapitulant les contributions principales, en identifiant les limites actuelles, et en ouvrant des perspectives de recherche pour les systèmes de vidéo-analyse multimodaux et explicables.

Publications associées à cette thèse

Les travaux présentés dans ce manuscrit ont donné lieu aux publications suivantes :

— **Publications acceptées** :

M. A. Sedeh, A. Benbihi, R. Martin, M. Clausel, C. Pradalier,
AttriVision : Advancing Generalization in Pedestrian Attribute Recognition using CLIP,
the Winter Conference on Applications of Computer Vision (WACV) [26].

Luis F. Batista, S. Khazem, M. A. Sedeh, S. Hutchinson, C. Pradalier
PoTATO : A Dataset for Analyzing Polarimetric Traces of Afloat Trash Objects,
European Conference on Computer Vision (ECCV), Milan, 2024 [27].

— **Article soumis** :

M. A. Sedeh, A. Benbihi, R. Martin, M. Clausel, C. Pradalier,
Multimodal Identity Learning via Attribute-Guided CLIP and Triplet-Centroid Loss for Person Re-Identification,
The Conference on Robots and Vision (CRV 2026).

2

Fondamentaux de l'apprentissage pour la vision par ordinateur

Ce chapitre présente les fondements scientifiques et techniques sur lesquels repose notre approche. Il introduit d'abord les bases de l'apprentissage automatique appliqué à la vision par ordinateur, puis détaille les travaux existants sur la reconnaissance d'attributs de personnes, la ré-identification (ReID), et l'utilisation conjointe d'approches multimodales via des modèles vision-langage (VLMs). Enfin, une synthèse critique positionne notre proposition dans cet écosystème.

2.1 Fondamentaux de l'apprentissage pour la vision par ordinateur

2.1.1 Apprentissage supervisé

L'apprentissage supervisé constitue l'un des paradigmes les plus fondamentaux de l'intelligence artificielle moderne. Il repose sur le principe selon lequel un système peut apprendre à prédire une sortie désirée à partir d'entrées observables, en exploitant un ensemble de données annotées. Dans le cadre de la vision par ordinateur, cela se traduit par des modèles capables de reconnaître des objets, de catégoriser des scènes, de détecter des visages ou encore d'inférer des attributs visuels à partir d'images.

Formellement, le but de l'apprentissage supervisé est de construire une fonction de prédiction $f : \mathcal{X} \to \mathcal{Y}$ à partir d'un ensemble fini de données d'entraînement $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, où $x_i \in \mathcal{X}$ représente une observation (par exemple une image) et $y_i \in \mathcal{Y}$ la sortie attendue (ex. : une classe ou un vecteur d'attributs). L'hypothèse sous-jacente est qu'il existe une distribution conjointe inconnue $P(X, Y)$, et que les exemples fournis sont des échantillons indépendants issus de cette distribution.

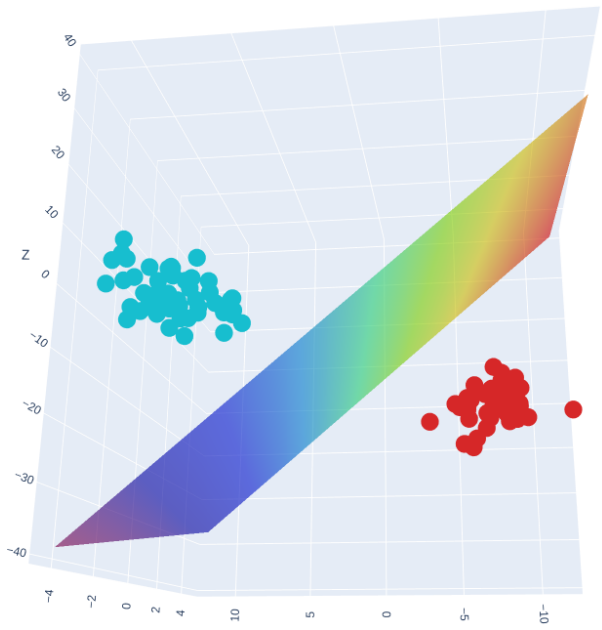


FIGURE 2.1 – Illustration d'une séparation linéaire entre deux classes de données dans un espace tridimensionnel. Chaque couleur représente une classe différente, et le plan coloré représente l'hyperplan de décision appris par un classifieur.

L'objectif de l'algorithme d'apprentissage est de minimiser l'erreur de prédiction sur des données encore non vues, c'est-à-dire de généraliser au-delà du jeu d'entraînement. Ce principe est illustré en figure 2.1. Pour cela, une fonction de coût $\mathcal{L}(f(x), y)$ est définie pour évaluer la performance du modèle sur une paire donnée. En pratique, on cherche à minimiser le risque empirique :

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i) \tag{2.1}$$

L'optimisation de cette quantité repose sur des algorithmes de descente de gradient (ou de ses variantes), ajustant les paramètres internes du modèle afin de réduire progressivement la perte mesurée.

Espace des données d'entrée Dans les tâches de vision, les entrées x_i sont généralement des images représentées sous forme de tenseurs de dimensions $(H \times W \times C)$, où H est la hauteur, W la largeur, et C le nombre de canaux (généralement 3 pour les images RGB). Ces tenseurs sont ensuite vectorisés ou traités dans des espaces de grande dimension $\mathbb{R}^d$ où $d = H \cdot W \cdot C$, ou bien encodés dans des représentations intermédiaires via des réseaux profonds.

Ainsi, une base d'entraînement typique est constituée d'un ensemble de points $\{x_i\}_{i=1}^N$ dans $\mathbb{R}^d$, chacun associé à un label y_i. Le défi principal est que d est souvent de l'ordre de plusieurs milliers voire millions (ex. : pour des images $224 \times 224 \times 3$, on a $d = 150528$), ce qui rend l'apprentissage sujet à la malédiction de la dimension. Cette complexité

justifie le recours à des méthodes paramétriques puissantes comme les réseaux de neurones convolutifs, capables d'extraire des représentations compactes et discriminantes.

Exemple : classification d'images dans $\mathbb{R}^d$ Un classifieur f entraîné sur un ensemble d'images peut être vu comme une fonction qui partitionne l'espace $\mathbb{R}^d$ en régions décisionnelles associées à différentes classes (illustré dans figure 2.2). La frontière entre ces régions dépend des représentations internes construites par le modèle (features) et des critères de décision appliqués en sortie.

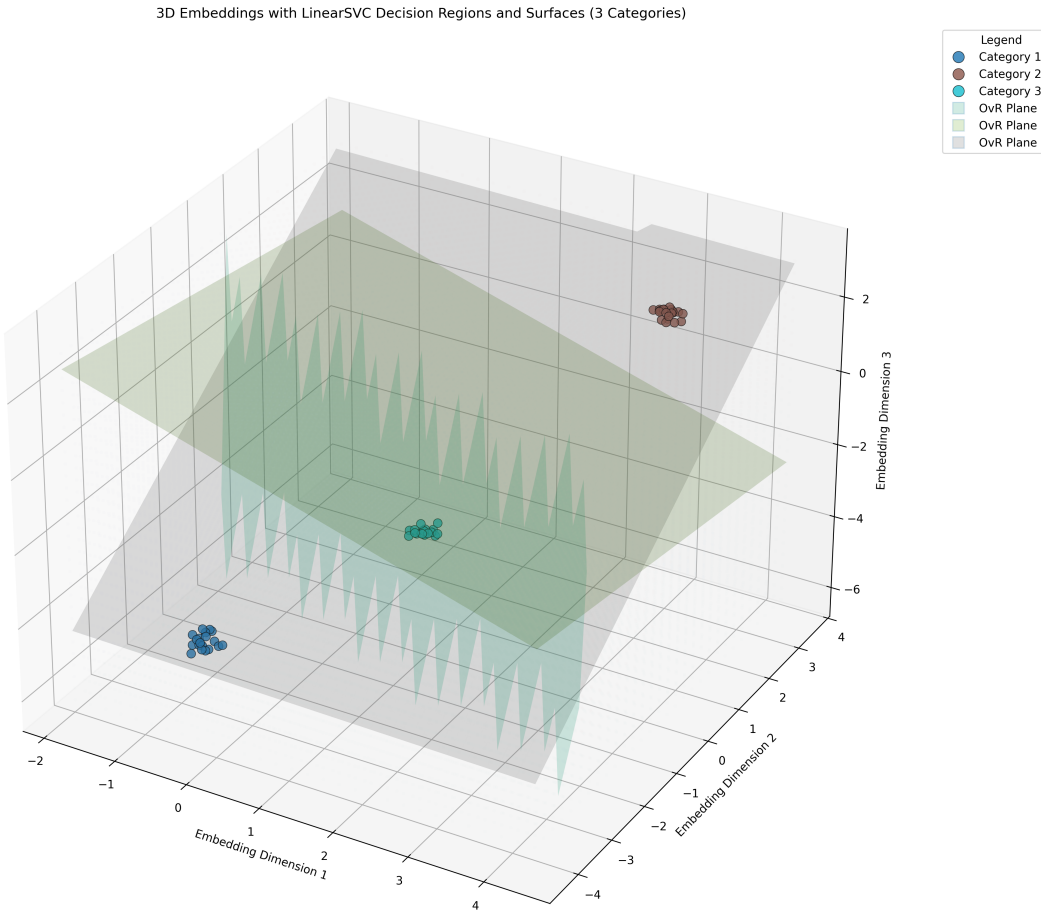


FIGURE 2.2 – Visualisation des embeddings dans un espace 3D pour trois catégories distinctes. Chaque point représente une instance projetée dans l'espace latent, colorée selon sa classe. Les plans colorés représentent les hyperplans de séparation calculés par un classifieur linéaire (LinearSVC) en mode un-versus-le-reste (OvR), illustrant les régions de décision apprises. Cette figure met en évidence la séparabilité linéaire des classes dans l'espace projeté.

2.1.2 Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne dispose d'aucune annotation explicite pour guider l'apprentissage. L'algorithme explore les structures, régularités ou regroupements naturels présents dans les données en entrée. Il s'agit donc d'un processus de découverte de représentations ou de motifs latents à partir de l'espace des observations seules, sans connaissance préalable des étiquettes de classe.

Dans le contexte de la vision par ordinateur, ce type d'apprentissage est particulièrement utile dans des cas où l'annotation manuelle est coûteuse ou impossible (ex. : segmentation non supervisée, détection d'anomalies, préentraînement de réseaux [28]).

Objectif mathématique Soit un ensemble de données $\mathcal{X} = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^d$ sans labels associés. L'objectif est de découvrir une structure $Z = \{z_1, z_2, \ldots, z_N\}$, avec $z_i \in \mathcal{Z}$, révélant des propriétés sous-jacentes : classes latentes, facteurs explicatifs, organisation topologique ou distributionnelle.

Clustering : partitionnement des données L'une des approches classiques de l'apprentissage non supervisé est le *clustering*, qui consiste à regrouper les exemples similaires dans des classes latentes illustrée dans 2.3. La méthode k-means est une des plus connues. Elle vise à minimiser l'inertie intra-cluster :

$$\mathcal{L}_{\text{k-means}} = \sum_{i=1}^{N} \|x_i - \mu_{z_i}\|^2 \tag{2.2}$$

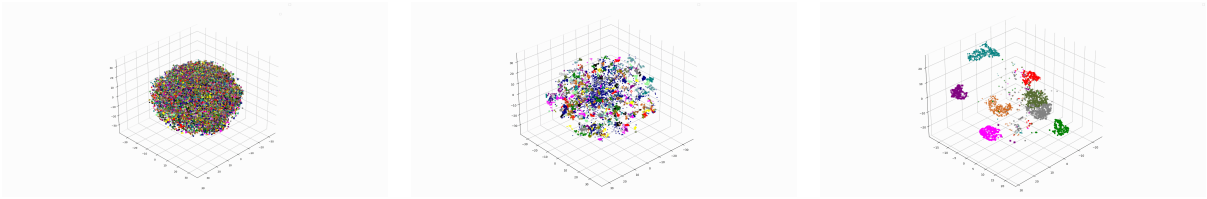
où μ_{z_i} désigne le centroïde du cluster auquel appartient x_i. Des variantes comme DBSCAN ou les algorithmes hiérarchiques permettent de découvrir des structures plus complexes [29].

Réduction de dimension Les données visuelles sont souvent de très haute dimension. Des techniques comme l'analyse en composantes principales (PCA) [30], t-SNE [31] ou UMAP [32] permettent de projeter les données dans un espace de plus faible dimension tout en conservant leur structure géométrique ou topologique :

- PCA : préserve la variance globale. - t-SNE : conserve les proximités locales. - UMAP : équilibre global/local avec meilleure scalabilité.

Ces méthodes facilitent la visualisation et la compréhension des distributions d'images non annotées. Un exemple de la PCA est illustrée dans 2.4

Auto-encodeurs et représentation latente Les **auto-encodeurs** sont des réseaux de neurones conçus pour apprendre à *compresser* des données d'entrée en une représenta-



(a) Début de l'entraînement (b) Milieu de l'entraînement (c) Fin de l'entraînement

FIGURE 2.3 – Évolution de la distribution des embeddings au cours de l'entraînement. On observe une structuration progressive de l'espace latent avec un regroupement par classe de plus en plus marqué.

tion latente de faible dimension, puis à les *reconstruire*. On peut les comparer à un artiste qui réalise un croquis simplifié (l'encodeur), et un autre qui tente de recréer le portrait original à partir de ce croquis (le décodeur). L'objectif est de minimiser la différence entre l'image d'origine et sa reconstruction.

Un auto-encodeur est composé de deux modules principaux :

— **Encodeur** ($f_\theta : \mathbb{R}^d \to \mathbb{R}^k$) : transforme une donnée d'entrée $\mathbf{x} \in \mathbb{R}^d$ en une représentation latente $\mathbf{z} \in \mathbb{R}^k$, avec $k \ll d$.

— **Décodeur** ($g_\phi : \mathbb{R}^k \to \mathbb{R}^d$) : reconstruit une approximation $\hat{\mathbf{x}}$ de la donnée d'origine à partir de la représentation latente $\mathbf{z}$.

L'apprentissage consiste à minimiser une **perte de reconstruction**, mesurant l'écart entre l'entrée et la sortie. La fonction de perte couramment utilisée est l'erreur quadratique moyenne (MSE) comme illustrée dans 2.5 :

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \|x_i - g_\phi(f_\theta(x_i))\|^2 \tag{2.3}$$

En contraignant le modèle à encoder l'essentiel de l'information dans un espace latent compressé, les auto-encodeurs apprennent des **représentations compactes et informatives**. Celles-ci peuvent être exploitées pour des tâches variées comme la réduction de dimension, la détection d'anomalies ou la génération de nouvelles données [33, 34].

2.1.3 Apprentissage auto-supervisé

L'apprentissage auto-supervisé constitue une évolution récente des approches non supervisées en vision par ordinateur. Il repose sur l'idée de générer automatiquement les labels à partir des données elles-mêmes, sans intervention humaine. L'entraînement des modèles se fait alors sur des tâches dites *prétextes*, conçues pour forcer le réseau à apprendre des représentations utiles et transférables. Parmi les tâches les plus courantes, on trouve la prédiction de la rotation appliquée à une image, comme dans RotNet [35], ou la reconstitution de parties manquantes dans une image, une stratégie d'inpainting proposée par [36]. D'autres méthodes, telles que SimCLR [37] et MoCo [38], visent à

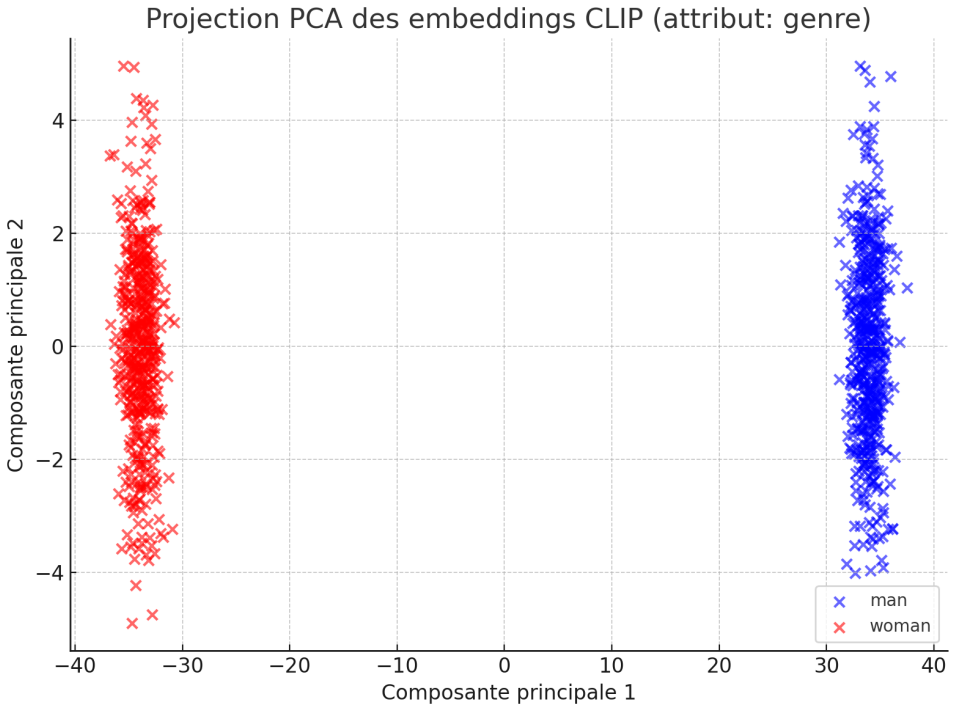


FIGURE 2.4 – Projection PCA des embeddings visuels extraits par CLIP pour 1000 images. Chaque point représente une image, colorée selon l'attribut "genre" (*man* en bleu, *woman* en rouge).

maximiser la similarité entre différentes augmentations d'une même image dans un espace latent. Enfin, des approches plus récentes comme BYOL [39] et DINO [40] reposent sur la prédiction de vues latentes, sans recours explicite à un objectif de contraste négatif.

Ces méthodes ont démontré leur efficacité en produisant des représentations visuelles de grande qualité, souvent comparables, voire supérieures, à celles obtenues par des modèles supervisés classiques sur plusieurs benchmarks standards en vision par ordinateur, tels qu'ImageNet ou COCO [37, 40].

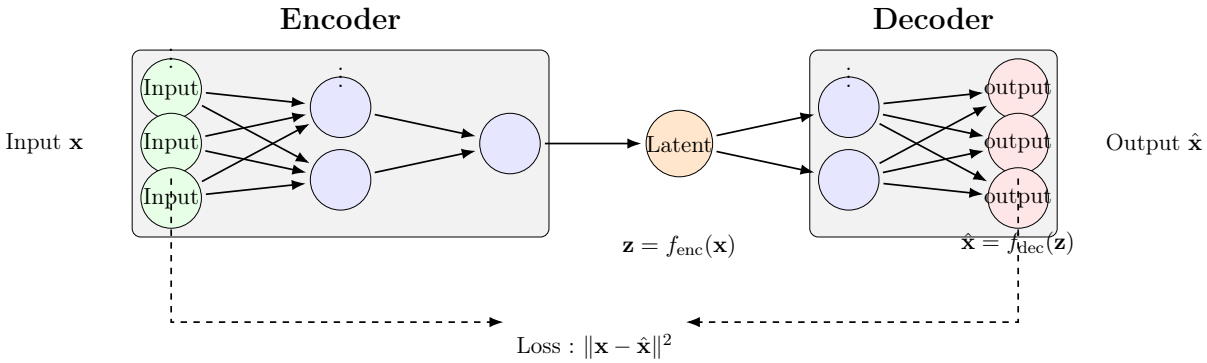


FIGURE 2.5 – Architecture d'un auto-encodeur feed-forward. L'encodeur projette l'entrée $\mathbf{x}$ dans un espace latent $\mathbf{z}$, puis le décodeur reconstruit $\hat{\mathbf{x}}$.

Résumé L'apprentissage non supervisé est indispensable lorsque l'annotation est indisponible ou pour préentraîner des modèles génériques. Il fournit des représentations utiles en aval, pour des tâches supervisées (classification, détection), ou non supervisées (retrieval, génération, compression).

2.1.4 Synthèse des paradigmes d'apprentissage

Les paradigmes d'apprentissage abordés – supervisé et non supervisé – fournissent les bases théoriques et méthodologiques sur lesquelles reposent la majorité des avancées en vision par ordinateur.

L'apprentissage **supervisé** reste aujourd'hui la méthode la plus utilisée pour les tâches de classification, détection, ou segmentation, en raison de sa précision et de son efficacité lorsque des annotations sont disponibles en grande quantité. Il s'appuie sur la minimisation d'une fonction de perte entre prédictions et vérités terrain, et tire sa puissance des réseaux neuronaux profonds tels que les CNN ou les transformers.

L'apprentissage **non supervisé**, quant à lui, permet d'exploiter les vastes volumes de données non étiquetées. Il est devenu incontournable pour le préentraînement de modèles génériques et l'extraction de représentations pertinentes. Les méthodes auto-supervisées les plus récentes (SimCLR, BYOL, DINO) ont montré des performances compétitives par rapport à leurs homologues supervisés sur plusieurs benchmarks standards [37, 39, 40].

Tableau comparatif des paradigmes :

Critère	**Supervisé**	**Non supervisé**
Données nécessaires	Données + labels	Données seules
Objectif	Prédire des labels explicites	Découvrir des structures latentes ou des représentations
Exemples en vision	Classification, segmentation	Clustering, représentation, pré-entraînement
Modèles emblématiques	CNN, ViT [41], ResNet [42]	Autoencodeurs, SimCLR [37], DINO [40]
Limites	Dépendance aux labels	Ambiguïté sémantique, évaluation difficile

Vers des approches hybrides. Dans la pratique, les frontières entre ces paradigmes tendent à s'estomper. Des modèles modernes combinent supervision explicite et auto-supervision (ex. : vision-language models). La compréhension fine de ces paradigmes reste donc essentielle pour explorer efficacement les méthodes actuelles en reconnaissance d'attributs ou re-identification.

2.2 Architectures fondamentales

2.2.1 Réseaux de neurones convolutionnels (CNN)

L'émergence des réseaux de neurones convolutionnels (*Convolutional Neural Networks*, CNN) a transformé la vision par ordinateur moderne. Leur architecture tire parti des propriétés spatiales des données visuelles, notamment la localité des motifs, l'invariance aux translations, et la redondance des structures visuelles à travers l'image. Ces caractéristiques en font une architecture idéale pour traiter des images, où l'on cherche à extraire des motifs hiérarchiques — des bords simples jusqu'à des concepts sémantiques complexes.

Motivations fondamentales : invariance locale et partage de poids L'idée principale d'un CNN repose sur deux propriétés clés :

- **Invariance locale** : les motifs visuels (ex. bords, textures) apparaissent localement dans une image et doivent être reconnus indépendamment de leur position exacte.
- **Partage des poids** : au lieu d'apprendre un filtre distinct pour chaque position dans l'image (comme dans un multilayer perceptron), un CNN utilise le même noyau de convolution à travers tout le champ visuel, ce qui réduit drastiquement le nombre de paramètres et améliore la généralisation.

Fonctions d'activation La fonction d'activation introduit une non-linéarité après la convolution. L'activation la plus couramment utilisée est la **ReLU** (Rectified Linear Unit) définie par :

$$\mathrm{ReLU}(x) = \max(0, x) \tag{2.4}$$

Elle permet de supprimer les activations négatives, accélère l'apprentissage et évite le phénomène de saturation observé avec des fonctions comme la sigmoïde.

Couches de pooling (sous-échantillonnage spatial) Les couches de *pooling* jouent un rôle essentiel dans les architectures convolutionnelles en permettant de réduire progressivement la taille spatiale des cartes de caractéristiques tout en conservant l'information discriminante la plus saillante. Elles opèrent localement, à l'échelle de petites régions, en résumant l'activité des neurones dans chaque voisinage. La forme la plus courante est le *max pooling*, qui sélectionne la valeur maximale à l'intérieur d'une fenêtre glissante, favorisant ainsi les activations les plus fortes et les motifs les plus dominants. Une alternative souvent utilisée est le *average pooling*, qui calcule la moyenne des activations dans la région considérée, produisant une représentation plus lissée de la réponse locale.

En réduisant le nombre de paramètres et la complexité computationnelle, ces opérations contribuent également à limiter le risque de sur-apprentissage. De plus, elles introduisent une forme d'invariance aux petites translations, ce qui améliore la robustesse du modèle face à des variations locales de positionnement dans les images.

Batch Normalization La **Batch Normalization** standardise les activations d'une couche en cours d'entraînement. Pour chaque mini-lot, on centre et on réduit les activations :

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{2.5}$$

où μ_B et σ_B^2 désignent respectivement la moyenne et la variance des activations calculées sur le mini-lot B, définies par :

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2, \tag{2.6}$$

et où $\epsilon > 0$ est un terme de régularisation introduit pour garantir la stabilité numérique de l'opération.

puis on applique une affine $y_i = \gamma\hat{x}_i + \beta$ où γ, β sont apprises. Cela accélère l'apprentissage, stabilise les gradients, et permet d'utiliser des taux d'apprentissage plus élevés [43].

Rétro-propagation et taille des tenseurs Chaque couche modifie la taille des données selon les paramètres de filtre de taille k, de stride s, et de padding p. La taille de sortie (H_{out}, W_{out}) d'une couche de convolution est donnée par :

$$H_{out} = \left\lfloor \frac{H - k + 2p}{s} \right\rfloor + 1 \quad ; \quad W_{out} = \left\lfloor \frac{W - k + 2p}{s} \right\rfloor + 1 \tag{2.7}$$

Pour la rétro-propagation (backpropagation), il est crucial de comprendre comment les gradients sont calculés et propagés à travers ces couches. Le processus inverse du passage avant (forward pass) permet de déterminer l'impact de chaque paramètre sur l'erreur finale.

Dans une couche de convolution, la rétro-propagation implique le calcul du gradient par rapport à l'entrée et aux poids du filtre. Ceci nécessite des opérations de convolution transposée (aussi appelée déconvolution ou convolution fractionnée) pour reconstruire une taille de tenseur plus grande. La taille de sortie de la rétro-propagation pour le gradient par rapport à l'entrée est souvent l'inverse de la formule du passage avant.

Pour les opérations de pooling (max pooling, average pooling), la rétro-propagation distribue le gradient de manière différente. Pour le max pooling, le gradient est envoyé uniquement à la position du maximum de la région de pooling. Pour l'average pooling, le gradient est distribué uniformément sur toutes les positions de la région.

La compréhension de ces transformations de taille est fondamentale pour l'implémentation correcte des réseaux de neurones profonds, en particulier lors de la construction d'architectures complexes telles que les encodeur-décodeurs où les données sont successivement réduites puis agrandies.

Principe de la convolution discrète Chaque couche d'un réseau de neurones convolutionnel applique une opération locale entre les entrées (images ou cartes de caractéristiques) et un ensemble de filtres appris. Illustré dans 2.6, en pratique, l'opération implémentée correspond à une *corrélation croisée* discrète. Dans le cas simplifié d'une image d'entrée mono-canal $X \in \mathbb{R}^{H \times W}$, d'un noyau $K \in \mathbb{R}^{k \times k}$, sans padding et avec un pas unité, la valeur de sortie en une position spatiale (i, j) est donnée par :

$$Y(i,j) = \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} K(u,v)\, X(i+u, j+v), \tag{2.8}$$

où $0 \leq i \leq H-k$ et $0 \leq j \leq W-k$. La carte de sortie $Y \in \mathbb{R}^{(H-k+1)\times(W-k+1)}$ est appelée *carte de caractéristiques*.

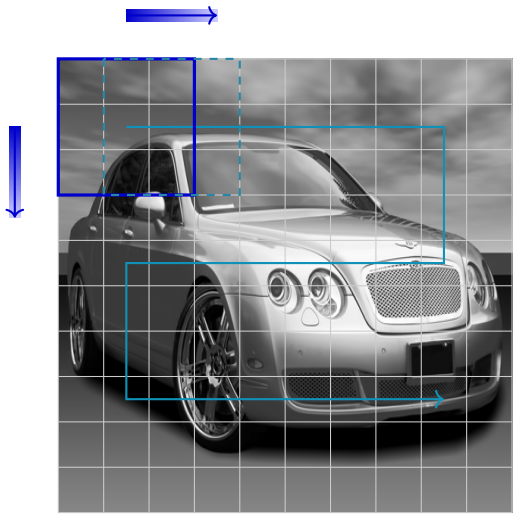


FIGURE 2.6 – Illustration du mécanisme de convolution 2D sur une image issue du jeu de données [1] : un filtre 3×3 (encadré bleu foncé) est appliqué à partir du coin supérieur gauche de l'image, puis se déplace de manière convolutive.

Architecture d'un CNN typique Un CNN est composé de plusieurs blocs empilés :

- **Convolution 2D** : extraction de motifs locaux.
- **Fonction d'activation** (ReLU) : non-linéarité pour capturer des motifs complexes.
- **Pooling (max ou average)** : sous-échantillonnage pour réduire la dimensionnalité et obtenir de l'invariance aux translations.

— **Couches entièrement connectées** : en fin de réseau, elles permettent une classification globale.

Ces blocs sont répétés en profondeur, permettant une hiérarchie de représentations allant des contours vers les objets. Comme illustré à la figure 2.7, l'architecture CNN considérée enchaîne trois blocs *Conv + ReLU + Pool* qui extraient des caractéristiques de plus en plus abstraites tout en réduisant la résolution spatiale (de $H \times W \times C$ à $\frac{H}{8} \times \frac{W}{8} \times C_3$). Le tenseur obtenu est ensuite aplati (*Flatten*) puis projeté par deux couches entièrement connectées avant la sortie *Softmax* qui fournit la distribution sur les classes. Les annotations sous chaque bloc précisent les dimensions intermédiaires, et la flèche courbée matérialise la continuité du flux entre la dernière couche convolutives (ligne supérieure) et la phase de classification (ligne inférieure).

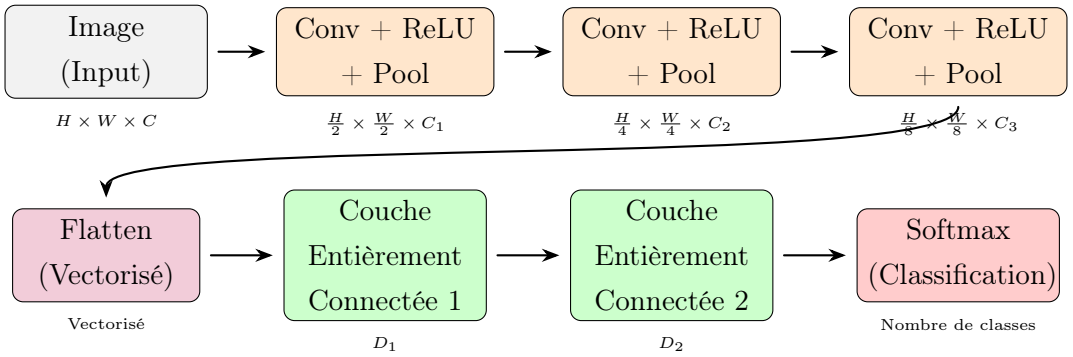


FIGURE 2.7 – CNN inspiré de VGG sur deux lignes : convolution/pooling (ligne 1), puis Flatten, couches entièrement connectées et Softmax (ligne 2). Les sous-textes indiquent les dimensions ou tailles de sorties.

Réseaux CNN profonds : ResNet, VGG, EfficientNet Au fil des années, de nombreuses architectures de réseaux de neurones convolutionnels (CNN) ont été proposées, chacune apportant des innovations majeures ayant marqué l'évolution de la vision par ordinateur. L'une des premières architectures notables est LeNet-5, introduite en 1998 par LeCun et al. [44]. Ce réseau, conçu pour la reconnaissance de chiffres manuscrits sur la base MNIST, a jeté les bases de l'apprentissage supervisé par rétropropagation dans des réseaux convolutifs multicouches.

La véritable explosion du deep learning en vision est survenue en 2012 avec l'introduction d'AlexNet [45], qui a largement surpassé les méthodes traditionnelles lors du défi ImageNet. Ce réseau profond de huit couches, combinant convolutions, ReLU et dropout, a démontré que les CNN pouvaient tirer parti des accélérations GPU pour traiter de vastes ensembles de données visuelles complexes.

Deux ans plus tard, VGGNet [46] a proposé une architecture plus profonde et régulière, constituée de blocs de convolutions 3×3 empilés. Malgré un nombre de paramètres élevé, VGG s'est distinguée par sa simplicité conceptuelle et sa capacité à capturer des

représentations hiérarchiques fines, devenant une référence dans les tâches de transfert d'apprentissage.

En 2015, ResNet [47] a introduit un changement de paradigme majeur avec les connexions résiduelles. Ces connexions permettent à l'information de circuler plus facilement à travers le réseau en contournant certaines couches, ce qui facilite l'optimisation de très grands réseaux (plus de 100 couches) sans dégradation des performances.

Plus récemment, EfficientNet [48] a proposé une approche systématique de mise à l'échelle des architectures CNN. En équilibrant la profondeur, la largeur et la résolution d'entrée à l'aide d'un unique coefficient d'échelle, ce modèle atteint un compromis remarquable entre précision et efficacité computationnelle, devenant un choix privilégié pour de nombreuses applications sur des dispositifs à ressources limitées.

Ainsi, chaque génération d'architectures CNN a contribué à repousser les limites du traitement visuel automatique, en améliorant à la fois la profondeur d'analyse, la stabilité d'apprentissage et la rentabilité des calculs.

Avantages et limites Les réseaux de neurones convolutionnels présentent plusieurs avantages fondamentaux qui expliquent leur succès en vision par ordinateur. Leur capacité à extraire automatiquement des invariants visuels, tels que les motifs locaux et les contours, en fait des outils particulièrement efficaces pour le traitement d'images naturelles. Grâce au partage des poids à travers les différentes régions de l'image, les CNN offrent également une excellente efficacité computationnelle, ce qui permet d'entraîner des modèles profonds avec un nombre relativement restreint de paramètres par rapport aux réseaux entièrement connectés. De plus, leur architecture hiérarchique, structurée autour de couches convolutives, de regroupements spatiaux (pooling) et de non-linéarités, est naturellement bien adaptée à la structure des données visuelles, où la localisation et l'agencement spatial jouent un rôle crucial.

Cependant, malgré ces qualités, les CNN classiques présentent plusieurs limitations importantes. Ils peinent notamment à modéliser efficacement les dépendances spatiales de longue portée entre différentes régions d'une image, ce qui les rend moins performants lorsqu'il s'agit de capturer des relations complexes à l'échelle de la scène. Par ailleurs, leur capacité à intégrer explicitement la sémantique des relations objets-contexte ou objets-objets reste limitée, car l'information y est souvent traitée localement sans mécanisme de raisonnement global. Enfin, les architectures CNN traditionnelles sont étroitement liées au domaine visuel, et leur extension à d'autres modalités comme le langage ou l'audio nécessite des adaptations structurelles importantes. Ces limitations ont contribué à l'émergence de modèles plus flexibles et multi-modaux, capables de traiter conjointement des informations issues de sources hétérogènes et de capturer des relations sémantiques riches au-delà du simple traitement d'image.

Évolutions des CNN modernes Les architectures convolutionnelles traditionnelles ont connu de nombreuses évolutions structurelles au fil des années afin d'améliorer leur expressivité, leur stabilité d'apprentissage et leur efficacité computationnelle. Parmi les contributions majeures figure l'introduction des connexions résiduelles avec les réseaux ResNet [47]. Celles-ci reposent sur l'ajout de chemins d'identité permettant de propager directement l'information et les gradients entre des couches non adjacentes, via une opération d'addition résiduelle.

Ce mécanisme facilite l'optimisation des réseaux profonds en incitant les couches empilées à apprendre des résidus par rapport à une fonction identité, plutôt qu'une transformation complète. Il permet ainsi de contourner le phénomène de dégradation des performances observé lorsque la profondeur du réseau augmente, et rend possible l'entraînement stable de modèles comportant plusieurs centaines, voire milliers, de couches.

D'autres travaux se sont concentrés sur la normalisation des activations intermédiaires, en particulier lorsque la taille des mini-lots devient trop faible pour que la Batch Normalization reste efficace. La normalisation de groupe [49], qui divise les canaux en groupes indépendants avant de normaliser chacun d'eux séparément, s'est imposée comme une alternative robuste, notamment dans des contextes de faible taille de batch ou de calcul distribué.

Enfin, des mécanismes d'attention légère ont été intégrés dans les architectures convolutionnelles pour renforcer leur capacité de sélection des caractéristiques pertinentes. Les modules Squeeze-and-Excitation (SE) [50], par exemple, permettent de recalibrer dynamiquement les canaux de caractéristiques en fonction de leur importance contextuelle, introduisant ainsi une attention adaptative sur les filtres à travers les dimensions spatiale et de canaux.

2.2.2 Introduction aux transformeurs

Les transformeurs (*Transformers*) ont été introduits initialement pour le traitement automatique du langage naturel (NLP) par Vaswani et al. dans le travail fondateur "Attention is All You Need" [51]. Ils ont profondément modifié le paradigme des architectures séquentielles traditionnelles, telles que les réseaux récurrents (RNN) ou LSTM, en se basant exclusivement sur un mécanisme d'attention auto-adaptative (*self-attention*), sans utiliser ni convolutions, ni récursivité.

L'architecture standard d'un transformeur repose sur une série de blocs empilés, comprenant un mécanisme de self-attention multi-tête, des couches de normalisation avec connexions résiduelles, et un réseau *feedforward* entièrement connecté. L'entraînement massivement parallèle rend le modèle hautement scalable d'un point de vue computationnel.

Le principe de l'attention consiste à pondérer dynamiquement l'importance de chaque

élément d'une séquence d'entrée relativement aux autres via des projections en requêtes (Q), clés (K) et valeurs (V) à partir d'une même entrée $X \in \mathbb{R}^{n \times d}$ (où n est la longueur de séquence et d la dimension des représentations) — voir la figure 2.9. La variante *multi-head*, illustrée dans la figure 2.8, parallélise plusieurs têtes apprenant des dépendances complémentaires à différentes échelles.

Dans un bloc transformeur complet, l'attention multi-tête est suivie d'un réseau *feed-forward* ; des normalisations et résidus facilitent l'optimisation de modèles profonds. La formalisation détaillée du mécanisme d'attention (ainsi que les équations associées) est présentée en section 2.2.3. La modularité, le parallélisme natif et la capacité à modéliser des dépendances longues ont rapidement conduit à l'adoption des transformeurs au-delà du NLP, notamment en vision par ordinateur.

2.2.3 Principe du self-attention

Le mécanisme de *self-attention* est la brique fondamentale des réseaux transformeurs. Il permet à chaque élément d'une séquence d'interagir avec les autres en apprenant à pondérer leur contribution selon leur pertinence contextuelle. Contrairement aux convolutions locales, l'attention auto-adaptative traite de manière globale les relations entre toutes les positions de la séquence en une seule opération parallèle.

Soit une entrée $X \in \mathbb{R}^{n \times d}$, représentant une séquence de n vecteurs de dimension d (par exemple, des tokens ou des patchs d'image). Trois matrices de projection apprises sont utilisées pour transformer chaque vecteur x_i en trois représentations distinctes :

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

où $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ sont des matrices de poids apprises. Q (requêtes), K (clés), et V (valeurs) sont toutes de dimension $n \times d_k$. Le paramètre d_k désigne la dimension de l'espace latent dans lequel sont projetées les requêtes (Q), les clés (K) et les valeurs (V).

Le score d'attention entre un élément i et un élément j est calculé comme le produit scalaire entre leurs représentations projetées :

$$s_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}}$$

où la normalisation par $\sqrt{d_k}$ stabilise la variance des gradients. Ces scores sont ensuite convertis en poids d'attention via une fonction softmax :

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{j'=1}^{n} \exp(s_{ij'})}$$

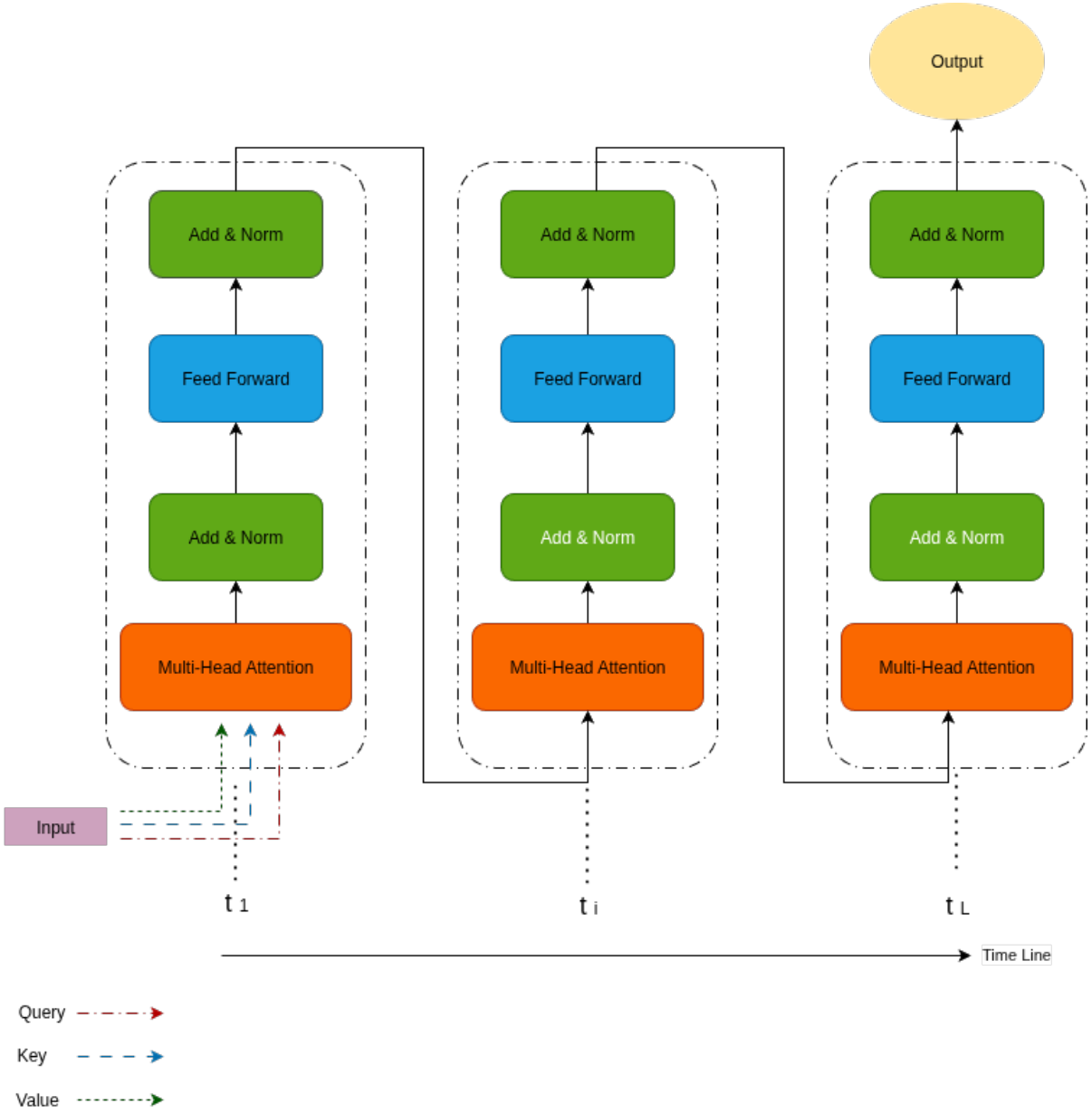


FIGURE 2.8 – Architecture d'un transformeur déplié temporellement. Chaque bloc contient une attention multi-tête suivie d'un réseau feed-forward, avec normalisation et connexions résiduelles.

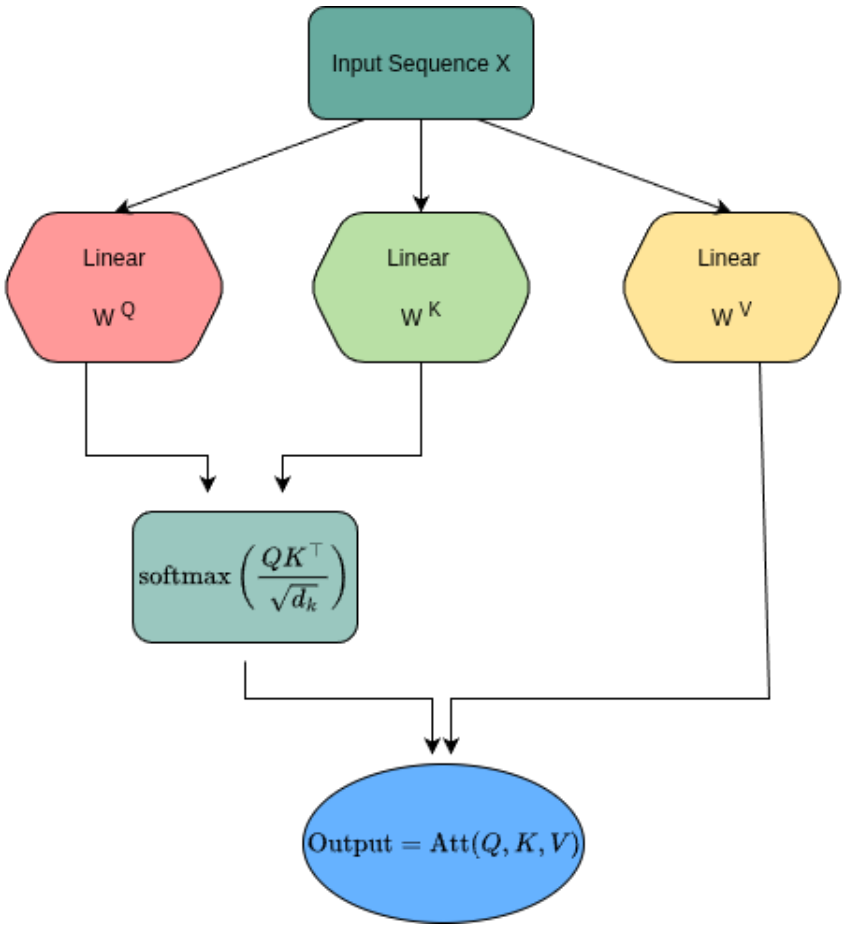


FIGURE 2.9 – Schéma simplifié de l'attention multi-tête : l'entrée se divise en Q, K et V, qui sont combinés par la *Multi-Head Attention* pour produire une sortie contextualisée.

Chaque vecteur de sortie est obtenu comme une combinaison pondérée des valeurs v_j :

$$z_i = \sum_{j=1}^{n} \alpha_{ij} v_j$$

Au niveau matriciel, le mécanisme complet s'écrit :

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

Ce mécanisme est ensuite étendu à l'*attention multi-tête*, où plusieurs têtes indépendantes apprennent des sous-espaces différents en parallèle. Chaque tête h utilise des matrices spécifiques W_h^Q, W_h^K, W_h^V, et les sorties sont concaténées et projetées dans l'espace d'origine :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \quad \text{où} \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

avec $W^O \in \mathbb{R}^{hd_k \times d}$ une matrice de sortie. Cette redondance permet au modèle de capturer simultanément plusieurs types de relations sémantiques, comme les connexions syntaxiques locales et les dépendances globales.

Dans une architecture de transformeur complète, chaque couche est composée d'un sous-bloc d'attention multi-tête suivi d'un bloc feedforward positionnel. Ces composants sont insérés dans un schéma d'apprentissage profond avec résidus et normalisation. Concrètement, chaque bloc suit la structure :

$$\begin{aligned} \tilde{X} &= \text{LayerNorm}(X) \\ Y &= X + \text{MultiHead}(\tilde{X}) \\ Z &= Y + \text{FFN}(\text{LayerNorm}(Y)) \end{aligned}$$

où FFN est un réseau de neurones à deux couches avec activation non linéaire, généralement de type GELU ou ReLU. Les connexions résiduelles améliorent la propagation du gradient, et les couches de normalisation stabilisent les distributions internes au cours de l'apprentissage.

L'ensemble de ce mécanisme permet aux transformeurs de dépasser les approches traditionnelles en capturant des relations complexes, longues et non-locales dans les données, ce qui explique leur succès aussi bien en traitement du langage que dans la vision par ordinateur, notamment avec les Vision Transformers.

2.2.4 Vision Transformer (ViT)

L'introduction du *Vision Transformer* (ViT) par Dosovitskiy et al. [52] a marqué une rupture dans le traitement des images en vision par ordinateur. Cette architecture transpose le paradigme des transformeurs, initialement conçus pour les séquences textuelles, à des données visuelles. Le principal défi était de définir une représentation séquentielle à partir d'images bidimensionnelles. ViT propose pour cela un découpage des images en petits blocs appelés *patchs*, chaque patch étant traité comme un token d'une séquence, analogue aux mots dans le traitement du langage.

Formellement, une image d'entrée $x \in \mathbb{R}^{H \times W \times C}$ est divisée en une grille régulière de N patchs non chevauchants de taille $P \times P$. Chaque patch est aplati et projeté dans un espace vectoriel de dimension fixe D via une couche linéaire :

$$z_0^i = x_{\text{patch}}^i W^E + E^i \quad \text{pour} \quad i = 1, \ldots, N$$

où $W^E \in \mathbb{R}^{(P^2C) \times D}$ est une matrice de projection linéaire, et E^i est un embedding positionnel ajouté pour injecter une notion d'ordre, absente des transformeurs par construction.

En plus des N patchs, ViT introduit un token spécial $[CLS]$, initialisé aléatoirement, qui joue le rôle d'agrégateur global de la représentation de l'image. Ce token est concaténé à la séquence des embeddings des patchs, portant ainsi la longueur totale de la séquence à $N + 1$:

$$Z_0 = [z_0^{[CLS]}; z_0^1; \ldots; z_0^N] + E_{\text{pos}}$$

où $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ contient les embeddings positionnels apprenants.

La séquence Z_0 est ensuite passée à travers une série de L blocs transformeurs empilés. Chaque bloc contient une couche de multi-head self-attention, suivie d'un réseau feedforward positionnel, encadré par des normalisations et des connexions résiduelles, selon la structure standard introduite précédemment. La sortie finale du token $[CLS]$ est extraite comme représentation globale de l'image pour la classification :

$$y = \text{MLP}(z_L^{[CLS]})$$

où MLP désigne une tête de classification composée de couches linéaires.

Le principal avantage de ViT est sa capacité à modéliser globalement les interactions entre toutes les régions de l'image dès les premières couches. Contrairement aux CNNs, où la croissance du champ réceptif est progressive, les transformeurs accèdent directement à une vue globale de l'image. Cette propriété leur confère une meilleure capacité à raisonner sur des relations sémantiques longues distance, comme les symétries ou les relations entre objets spatialement éloignés.

Cependant, ViT présente aussi des limitations importantes. Le coût computationnel

du self-attention est quadratique en la taille de la séquence, ce qui pose problème pour les images haute résolution. De plus, l'absence d'inducteurs de biais locaux (comme le partage de poids et la translation invariance des convolutions) rend les modèles ViT très dépendants de l'échelle des données d'entraînement. Dans l'article original [52], les meilleures performances ne sont atteintes que lorsque le modèle est préentraîné sur le très grand corpus JFT-300M.

Depuis, de nombreuses variantes ont été proposées pour atténuer ces limitations : patchs hiérarchiques [53], attention locale restreinte [54], hybridation avec des couches convolutives [55], ou encore compression adaptative des tokens [56].

Malgré ces défis, ViT constitue désormais une alternative crédible aux CNNs dans de nombreuses tâches, et sa structure modulaire l'a rendu particulièrement adapté aux modèles vision-langage comme CLIP.

2.3 Modèles vision-langage

2.3.1 CLIP et dérivés

Les modèles vision-langage visent à apprendre une représentation commune entre des images et des descriptions textuelles associées. CLIP (*Contrastive Language–Image Pretraining*), proposé par OpenAI [57], constitue une avancée majeure dans ce domaine. Ce modèle a pour objectif d'aligner les représentations visuelles et textuelles dans un espace latent partagé, facilitant ainsi les tâches de correspondance image-texte, telles que la recherche par requête naturelle, la reconnaissance d'attributs, ou encore la classification en zero-shot.

L'architecture de CLIP repose sur deux encodeurs indépendants : un encodeur visuel, souvent basé sur un Vision Transformer (ViT), et un encodeur textuel, basé sur un transformeur de type GPT. Ces deux branches prennent respectivement en entrée une image x_I et une phrase x_T décrivant cette image (par exemple, *"une personne avec un sac à dos noir"*). Les sorties de ces deux encodeurs sont projetées dans un espace de dimension commune d :

$$f_I = \phi_I(x_I) \in \mathbb{R}^d, \quad f_T = \phi_T(x_T) \in \mathbb{R}^d$$

ù ϕ_I et ϕ_T sont les encodeurs visuel et textuel, respectivement. Les vecteurs sont ensuite normalisés par leur norme ℓ_2, ce qui permet de mesurer la similarité à l'aide du produit scalaire ou de la distance cosinus :

$$s(x_I, x_T) = \cos(f_I, f_T) = \frac{f_I \cdot f_T}{\|f_I\|\|f_T\|}$$

L'apprentissage repose sur une perte contrastive appelée InfoNCE, qui pousse les ap-

pariements corrects (image, texte) à avoir une similarité plus élevée que les appariements incorrects. Pour un batch de taille N, la perte totale s'écrit :

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ij}/\tau)} - \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ji}/\tau)} \right]$$

où $s_{ij} = \cos(f_I^i, f_T^j)$ est la similarité entre l'image i et le texte j, et τ est une température apprenable. Cette formulation symétrique force à la fois les images à se rapprocher de leur texte correspondant, et les textes de leur image associée.

L'un des atouts majeurs de CLIP réside dans son pré-entraînement sur 400 millions de paires image-texte extraites du web. Cette diversité lui confère une capacité de généralisation impressionnante en zero-shot. Contrairement aux modèles traditionnels entraînés sur des classes fermées (comme ImageNet), CLIP peut être directement utilisé pour des tâches nouvelles en générant les descriptions textuelles associées aux classes cibles.

Dans le contexte de la reconnaissance d'attributs de personnes (PAR), cette capacité à comparer une image à une description comme *"personne portant un pantalon rouge"* permet une supervision textuelle fine. Des travaux comme AttriCLIP [26] ou Cerberus [23] exploitent ce principe pour projeter chaque attribut dans l'espace sémantique de CLIP, et former des prédictions par similarité.

Dans les tâches de re-identification (ReID), l'idée d'associer des identités visuelles à des représentations textuelles sémantiques gagne également en popularité. Des frameworks comme LVLM-ReID [58] utilisent la puissance de CLIP pour apprendre des représentations discriminantes supervisées par des descriptions attributaires ou contextuelles. D'autres approches, comme CLIP-FGDI [59], introduisent des stratégies de génération de prompts pertinents afin de renforcer l'alignement sémantique.

L'intérêt de CLIP réside également dans son architecture symétrique, qui facilite les adaptations multi-tâches. Il est ainsi possible d'exploiter le même réseau de base pour des tâches de classification, de détection d'attributs et de mise en correspondance identitaire, au moyen de projections additionnelles ou de têtes spécialisées. Par ailleurs, son apprentissage contrastif se prête naturellement à des stratégies d'affinage supervisé, telles que l'utilisation conjointe de la *Focal Loss* ou de la *Triplet Loss*, approche que nous avons explorée dans notre propre travail.

Malgré ses avantages, CLIP présente certaines limitations. Son efficacité dépend fortement de la qualité des prompts textuels utilisés, ce qui justifie l'émergence de méthodes de génération adaptative de textes. De plus, les biais présents dans les données web peuvent être transférés au modèle, affectant son comportement en conditions critiques. Enfin, son coût mémoire reste supérieur à celui des CNNs classiques, en particulier lors du traitement de grands lots d'images ou de longues séquences textuelles.

2.3.2 Développements récents et adaptations à la PAR et à la ReID

Depuis sa publication, CLIP a suscité de nombreuses extensions visant à l'adapter à des contextes plus spécialisés, notamment en reconnaissance d'attributs de personnes (PAR) et en re-identification (ReID). Ces deux tâches partagent un besoin fondamental de représentation robuste, à grain fin et interprétable — un domaine dans lequel l'alignement image/texte de CLIP s'avère particulièrement pertinent.

Dans le cadre de la reconnaissance d'attributs, plusieurs approches ont proposé d'utiliser les attributs comme des requêtes textuelles à comparer aux images via les embeddings CLIP. Par exemple, AttriCLIP [26] reformule la prédiction d'attributs comme une tâche de matching texte-image, où chaque attribut est exprimé sous forme d'un prompt textuel, tel que *"une personne portant un pantalon gris"*. Cette formulation permet de contourner les limitations des classifieurs binaires traditionnels, notamment en gérant plus naturellement les co-occurrences sémantiques et les incertitudes d'annotation.

Dans le même esprit, Cerberus [23] introduit la notion de *semantic ID* en combinant les labels d'attributs associés à une personne. Un vecteur textuel est généré à partir de cette combinaison, servant à guider l'apprentissage de la représentation globale. Le modèle intègre ainsi à la fois la tâche de PAR, la recherche attributaire (APS) et la ReID, tout en utilisant une perte de guidage sémantique entre les embeddings visuels et textuels.

Concernant la re-identification, plusieurs travaux ont tiré parti de la capacité des modèles vision-langage à représenter des identités à travers des descriptions riches. LVLM-ReID [58] propose une approche dans laquelle un grand modèle vision-langage (Large Vision-Language Model, LVLM) génère un *semantic token* spécifique à chaque individu. Ce vecteur sémantique est ensuite utilisé comme ancre dans un module d'interaction guidée, permettant d'affiner les représentations identitaires.

De même, CLIP-FGDI [59] introduit une stratégie à trois phases pour affiner les représentations textuelles : génération automatique de descriptions, alignement inter-domaines, puis raffinement des projections visuelles. L'objectif est de produire des descripteurs robustes aux variations de domaine (par exemple, entre différentes caméras), tout en préservant une granularité suffisante pour la discrimination inter-individuelle.

Pour la ReID en contexte difficile, comme la variation d'apparence (changement de vêtements, angles extrêmes, etc.), plusieurs travaux utilisent des prompts dynamiques ou des centres sémantiques comme supervision. Par exemple, VLPSR [60] montre comment exploiter la robustesse de CLIP en zero-shot pour des scénarios où peu d'identités sont annotées. Le modèle utilise une adaptation des embeddings textuels pour relier les attributs connus à des instances visuelles inconnues.

Ces développements démontrent que l'usage de CLIP et ses variantes ne se limite pas à la recherche sémantique ou au zero-shot, mais peut être exploité pour des ap-

prentissages supervisés complexes à fort signal sémantique. En combinant des modules textuels génératifs avec des contraintes discriminatives (ex. Triplet Loss, Focal Loss), les modèles vision-langage s'intègrent désormais dans des pipelines hybrides de classification, recherche et identification.

Toutefois, certains défis subsistent. Le contrôle des prompts reste en grande partie manuel, ce qui peut affecter la robustesse en production. L'injection explicite d'attributs discriminants sous forme de supervision textuelle soulève également des questions sur le déséquilibre de leur occurrence. Enfin, le coût mémoire et la latence induits par les encodeurs lourds freinent encore leur déploiement temps réel, particulièrement dans des systèmes embarqués.

2.3.3 Synthèse critique : CNNs, ViT et modèles vision-langage

L'évolution des architectures de vision par ordinateur peut être vue comme une progression vers des représentations de plus en plus sémantiques, globales et adaptatives. Les réseaux de neurones convolutionnels (CNNs) ont longtemps constitué l'épine dorsale des systèmes de reconnaissance visuelle. Leur efficacité repose sur l'exploitation de l'invariance locale et du partage de poids, ce qui les rend particulièrement performants pour détecter des motifs spatiaux hiérarchiques. Néanmoins, leur capacité à modéliser des relations longues distances reste limitée sans ajout explicite de mécanismes non locaux.

L'introduction des Vision Transformers (ViTs) a permis de surmonter certaines de ces limitations en remplaçant les convolutions par des opérations d'attention auto-adaptative. Les ViTs capturent globalement les relations entre régions dès les premières couches, ce qui améliore leur capacité à raisonner de manière contextuelle. Toutefois, cette puissance d'expression se paie en données et en ressources computationnelles : sans pré-entraînement massif ou augmentation de données sophistiquée, les ViTs ont tendance à surajuster ou à ne pas converger.

Les modèles vision-langage tels que CLIP marquent une nouvelle étape dans cette évolution. En apprenant un espace latent partagé entre images et textes, ils facilitent l'alignement sémantique supervisé par des descriptions naturelles. Cette capacité d'alignement multimodal ouvre de nouvelles perspectives pour les tâches à supervision faible ou ambigüe, comme la reconnaissance d'attributs ou la re-identification sous contraintes sémantiques.

En résumé, si les CNNs offrent des modèles compacts et efficaces pour des tâches classiques, les ViTs et CLIP introduisent une capacité sémantique qui les rend mieux adaptés aux tâches à granularité fine et sémantique élevée. Le tableau 2.1 résume les caractéristiques comparatives majeures de ces architectures :

Critère	CNN	ViT	CLIP
Modèle inductif	Fort (locale)	Faible	Faible
Réceptif global dès les premières couches	Non	Oui	Oui
Capacité d'alignement sémantique	Non	Non	Oui
Besoin en données	Modéré	Élevé	Très élevé (pré-entraînement)
Adaptabilité zero-shot	Non	Rare	Oui
Interprétabilité (attributs)	Faible	Moyenne	Élevée
Utilisation pour PAR	Fréquente	En hausse	En hausse
Utilisation pour ReID	Standard	En développement	Récente mais prometteuse

TABLE 2.1 – Comparaison synthétique des architectures CNN, ViT et CLIP.

Cette synthèse met en lumière la tendance actuelle de la recherche : combiner les biais inductifs (comme dans les CNNs) avec des mécanismes globaux et sémantiques (comme dans ViT et CLIP), ouvrant la voie à des architectures hybrides pour les tâches complexes telles que la reconnaissance d'attributs ou la re-identification multimodale.

2.4 Optimisation des modèles profonds

2.4.1 Fonctions de perte en classification

Dans les tâches d'apprentissage supervisé en classification, la fonction de perte joue un rôle central : elle guide l'optimisation du réseau neuronal en quantifiant l'écart entre les prédictions et les étiquettes réelles. Le choix de cette fonction dépend du type de tâche (mono-label ou multi-label), de la distribution des classes, et parfois de la nature des représentations apprises (apprentissage métrique).

Entropie croisée pour la classification à classe unique La fonction de perte la plus courante dans les tâches de classification à classe unique est l'**entropie croisée** (*Cross-Entropy*, CE). Elle est dérivée de la théorie de l'information et mesure la dissimilarité entre la distribution prédite par le modèle $p_\theta(x)$ (issue d'une softmax) et la distribution réelle y. C désigne le nombre total de classes considérées dans la tâche de classification :

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(p_\theta(x)_i) \tag{2.9}$$

Cette formulation est utilisée dans de nombreux contextes : classification d'images (ImageNet [61]), détection d'objets (YOLO [62]), ou encore reconnaissance de scènes.

Elle a l'avantage d'être convexe dans le cas linéaire et de produire des gradients bien conditionnés dans le cas des réseaux neuronaux profonds.

Binary Cross-Entropy (BCE) pour la classification multi-label Dans un cadre multi-label, chaque échantillon peut posséder plusieurs classes simultanément. Les sorties du réseau sont indépendantes et activées par une fonction sigmoïde $\sigma(z_i)$ sur chaque neurone de sortie. La fonction de perte s'écrit comme une somme d'entropies croisées binaires :

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{C} [y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))] \tag{2.10}$$

où C est le nombre de classes et $y_i \in \{0, 1\}$. Elle est largement utilisée en reconnaissance d'attributs de personne (PAR) [63, 64], détection multi-étiquette, et détection d'objets avec plusieurs catégories.

Focal Loss : pour lutter contre le déséquilibre des classes Les jeux de données réels présentent souvent un déséquilibre fort entre classes (ex. : attributs vestimentaires rares, classes minoritaires en détection). Dans ce contexte, la **Focal Loss**, proposée dans [65], introduit un terme modulateur qui diminue l'impact des exemples faciles sur l'apprentissage :

$$\mathcal{L}_{FL} = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{2.11}$$

où :

— p_t est la probabilité prédite correcte ($p_t = p$ si $y = 1$, $1 - p$ sinon),

— γ est le paramètre de focalisation ($\gamma > 0$),

— α est un facteur de pondération facultatif pour gérer le déséquilibre inter-classe.

Lorsque p_t est élevé (exemple bien classé), le facteur $(1-p_t)^{\gamma}$ devient petit, réduisant la perte associée à cet exemple. À l'inverse, les exemples mal classés (petit p_t) sont davantage mis en avant. L'effet du changement de γ est illustré dansla figure 2.10.

Cette fonction a été utilisée dans des contextes variés, allant de la détection d'objets déséquilibrée (RetinaNet [65]) à la reconnaissance d'attributs déséquilibrés [66].

Triplet Loss : apprentissage métrique Lorsque l'objectif est d'apprendre un espace de représentation dans lequel les instances similaires sont proches et les dissimilaires éloignées, comme en re-identification ou en *recherche par similarité* (*retrieval*), on utilise des pertes dites **contrastives**. La plus répandue est la **Triplet Loss** [67].

On considère un triplet (a, p, n) composé d'une image ancre, d'une image positive (même classe) et d'une image négative (classe différente). La perte cherche à imposer que l'image ancre soit plus proche du positif que du négatif d'une certaine marge m :

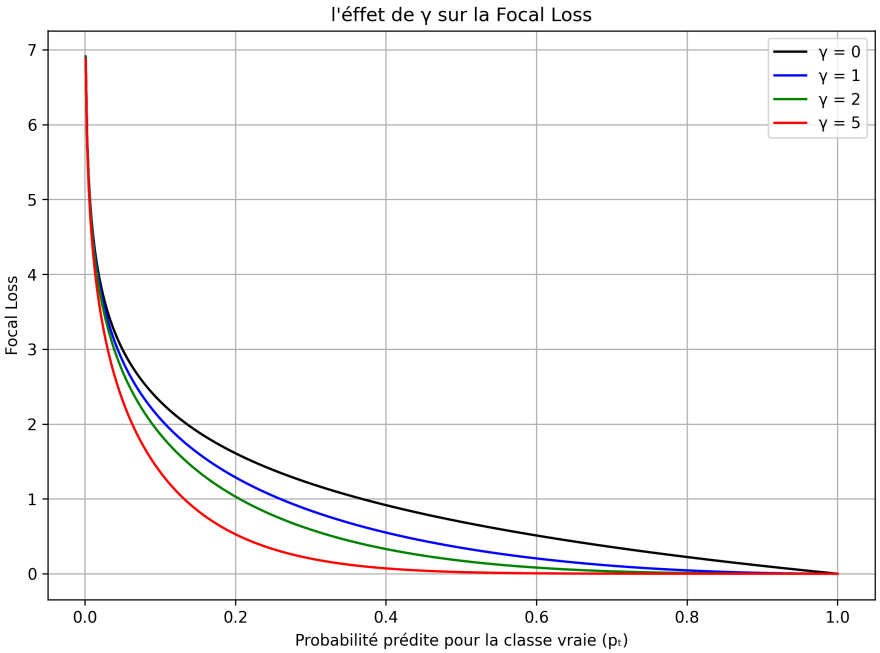


FIGURE 2.10 – Effet du paramètre de focalisation γ sur la Focal Loss. Lorsque $\gamma = 0$, la Focal Loss correspond à l'entropie croisée classique. Des valeurs plus élevées de γ diminuent l'impact des exemples faciles (avec une probabilité prédite élevée pour la classe correcte) et concentrent l'apprentissage sur les exemples difficiles.

$$\mathcal{L}_{Triplet} = \max\left(\|f(a) - f(p)\|_2^2 - \|f(a) - f(n)\|_2^2 + m, 0\right) \tag{2.12}$$

où $f(x)$ est l'encodage de l'image par le réseau.

Cette formulation est utilisée dans FaceNet [67], DeepSort [68], et de nombreux modèles de Person ReID.

Triplet Loss à distance cosinus Une variante utilise la **similarité cosinus** plutôt que la distance euclidienne. Elle est particulièrement adaptée lorsque les embeddings sont L_2-normalisés (comme dans CLIP [57]) :

$$\mathcal{L}_{Triplet}^{\cos} = \max\left(\cos(f(a), f(n)) - \cos(f(a), f(p)) + m, 0\right) \tag{2.13}$$

L'utilisation d'une distance angulaire améliore la stabilité de l'apprentissage dans les modèles multi-modaux où les distributions sont hétérogènes (images et texte).

Synthèse Le choix de la fonction de perte est déterminant pour la performance du modèle. Dans le cas des architectures jointes (PAR + ReID), les travaux récents combinent souvent des pertes de classification (comme la Focal Loss) avec des pertes métriques (Triplet ou ArcFace loss) [18, 69] afin d'optimiser simultanément la discrimination des attributs et la structure de l'espace latent.

2.4.2 Stratégies d'optimisation des réseaux profonds

L'optimisation des réseaux neuronaux profonds constitue une composante essentielle de l'apprentissage supervisé. L'objectif est de minimiser la fonction de perte définie précédemment en mettant à jour les poids du modèle de manière itérative. L'efficacité et la stabilité de l'apprentissage dépendent fortement du choix de l'algorithme d'optimisation et de l'ajustement des hyperparamètres.

Descente de gradient et rétropropagation La méthode de base utilisée pour optimiser les réseaux profonds est la **descente de gradient stochastique** (SGD) combinée à l'algorithme de **rétropropagation** des gradients. Soit une fonction de perte $\mathcal{L}(\theta)$ où θ désigne les paramètres du réseau. À chaque itération, les paramètres sont mis à jour selon :

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta) \tag{2.14}$$

où η est le **taux d'apprentissage** (*learning rate*) et $\nabla_\theta \mathcal{L}$ est le gradient de la perte par rapport aux paramètres.

La rétropropagation (backpropagation) consiste à appliquer la règle de dérivation en chaîne pour calculer efficacement les gradients en partant de la couche de sortie et en remontant couche par couche jusqu'aux entrées [70].

Stochastic Gradient Descent (SGD) La version standard de la descente de gradient calcule le gradient sur l'ensemble des données. En pratique, pour des raisons de performance, on utilise une version *stochastique*, où le gradient est estimé sur un **mini-lot** (mini-batch) de données :

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathcal{B}}(\theta) \tag{2.15}$$

avec $\mathcal{B}$ un mini-batch aléatoire. L'approximation stochastique du gradient introduit un bruit contrôlé dans le processus d'optimisation. Ce bruit agit comme une forme de régularisation implicite, empêchant le modèle de converger prématurément vers des minima trop étroits fortement dépendants des données d'entraînement. En favorisant l'exploration de régions de l'espace des paramètres associées à des minima plus plats, la descente de gradient stochastique tend ainsi à améliorer la capacité de généralisation du modèle.

Ajout du Momentum L'optimisation peut être accélérée en accumulant les gradients précédents via le **momentum** :

$$v_t = \beta v_{t-1} + \nabla_\theta \mathcal{L}(\theta) \quad ; \quad \theta \leftarrow \theta - \eta v_t \tag{2.16}$$

Cela permet de lisser les mises à jour et de sortir plus facilement des minima locaux ou des plateaux. Ce mécanisme est à l'origine des optimisations comme Nesterov Momentum [71].

Adam et AdamW : méthodes adaptatives **Adam** (Adaptive Moment Estimation) [72] est aujourd'hui l'optimiseur le plus répandu. Il adapte dynamiquement le taux d'apprentissage pour chaque paramètre :

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \quad ; \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \\ \theta_t &= \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{aligned} \tag{2.17}$$

où g_t est le gradient, β_1 et β_2 sont des coefficients de moyenne mobile, $\beta_i \in [0, 1]$ et ϵ est une constante de stabilisation.

Cependant, Adam standard introduit un problème d'interaction entre la normalisation adaptative et la régularisation de type L2. Pour corriger cela, **AdamW** [73] propose de découpler le poids de décroissance (weight decay) du calcul du gradient :

$$\theta \leftarrow \theta - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta \right) \tag{2.18}$$

où λ est le coefficient de décroissance.

Hyperparamètres critiques La qualité de l'apprentissage dépend fortement de plusieurs hyperparamètres clés :

- **Taux d'apprentissage** η : trop faible, l'apprentissage stagne ; trop élevé, il devient instable.
- **Taille du mini-batch** : affecte la stabilité des gradients. Les tailles entre 16 et 128 sont courantes. Les très gros batches nécessitent une adaptation du taux d'apprentissage [74].
- **Décroissance de poids (Weight Decay)** : régularisation implicite pour éviter le sur-apprentissage.
- **Planificateurs de taux (schedulers)** : la réduction progressive du taux d'apprentissage (ex. *cosine annealing*, *plateau*, *warmup*) permet d'améliorer la convergence [75].

Visualisation de la trajectoire d'optimisation On peut illustrer la différence entre les méthodes en traçant les trajectoires qu'elles empruntent sur une surface de perte en 3D comme illustré dans la figure 2.11.

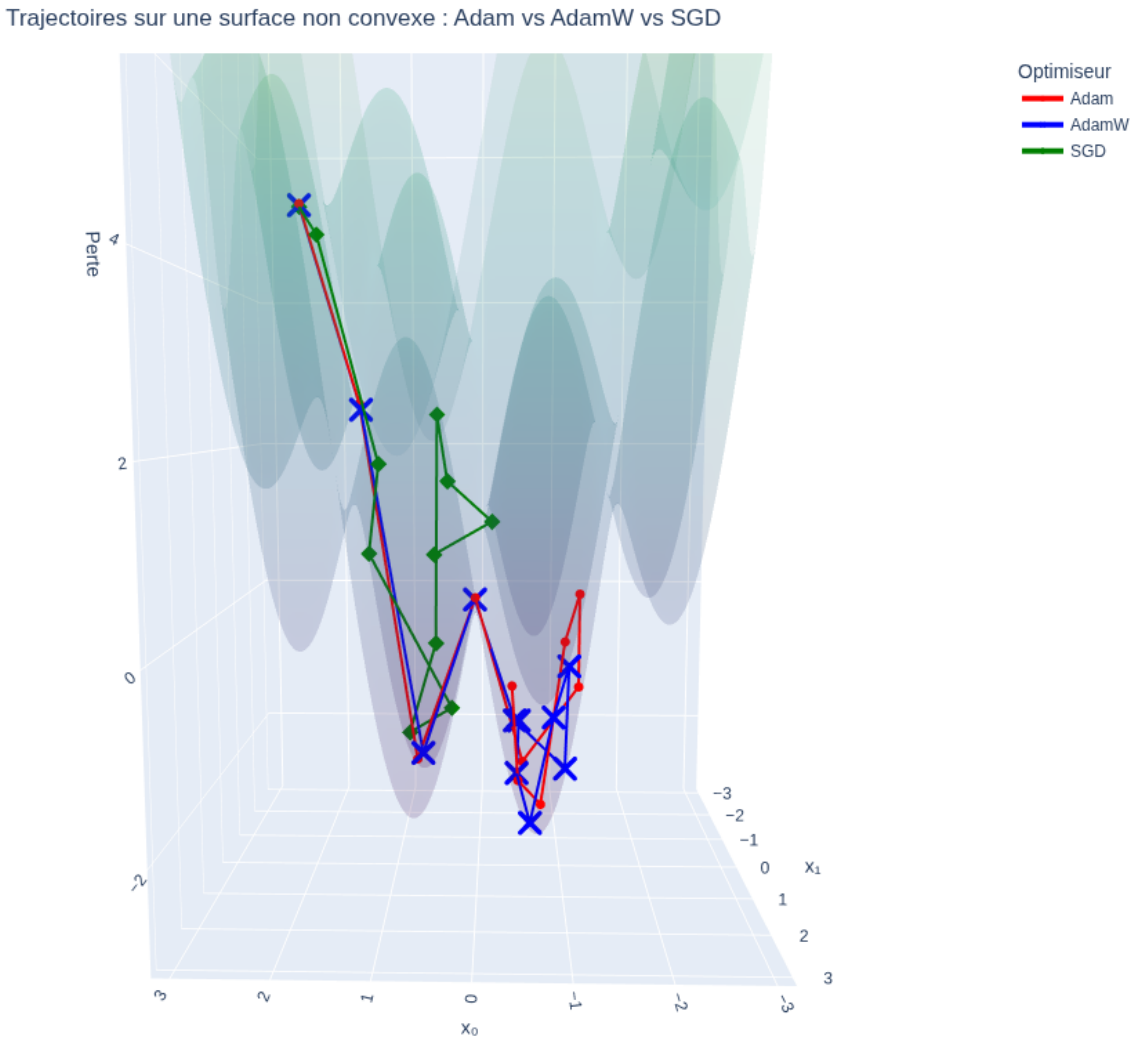


FIGURE 2.11 – Visualisation des trajectoires d'optimisation des algorithmes *SGD*, *Adam* et *AdamW* sur une surface de perte non convexe. Ce graphique illustre les différences de comportement entre ces méthodes : SGD progresse avec des oscillations liées à l'absence d'adaptation du taux d'apprentissage, tandis qu'Adam et AdamW suivent des trajectoires plus lisses grâce à leur mécanisme de moyenne mobile. AdamW, en particulier, dissocie le terme de régularisation du gradient, ce qui lui confère une meilleure stabilité.

SGD converge plus lentement, mais explore mieux la surface ; Adam converge rapidement mais peut rester bloqué dans des zones sous-optimales en absence de régularisation adéquate.

Synthèse Le choix de l'algorithme d'optimisation dépend du type de tâche, de la taille des données, et des ressources disponibles. Dans les modèles de grande taille (ex. CLIP, transformers), AdamW avec warmup et scheduler cosinus est la norme [52, 57]. Une attention particulière doit être portée aux mécanismes de régularisation implicite induits par l'optimiseur.

2.4.3 Améliorations pratiques et stabilité d'entraînement

L'entraînement stable et efficace d'un réseau de neurones profonds repose non seulement sur le choix des architectures et des fonctions de perte, mais également sur un ensemble de techniques pratiques visant à améliorer la convergence, éviter les instabilités numériques et optimiser l'utilisation des ressources matérielles.

Learning Rate Scheduling Le taux d'apprentissage (*learning rate*) est l'un des hyperparamètres les plus sensibles de l'entraînement. Une valeur mal ajustée peut conduire soit à une stagnation de l'apprentissage, soit à des divergences.

Pour cette raison, il est courant d'adopter une **politique d'évolution dynamique** du taux d'apprentissage à travers un *scheduler*. Les stratégies les plus répandues sont :

- **Cosine Annealing** : le taux d'apprentissage suit une courbe cosinusoïdale décroissante, sans plateaux intermédiaires, favorisant une descente progressive vers un minimum stable [75].
- **Reduce on Plateau** : la valeur est réduite automatiquement (souvent d'un facteur 10) lorsque la performance (perte ou métrique de validation) stagne pendant un certain nombre d'époques.
- **Warmup** : avant d'appliquer la décroissance, un préchauffage initial (quelques centaines ou milliers d'itérations) permet d'amorcer l'entraînement à petit pas [74].

Ces mécanismes améliorent la stabilité initiale et la convergence terminale, en particulier dans des contextes d'apprentissage joint (multi-tâches) ou multi-modal.

Gradient Clipping Dans certaines situations (notamment avec des architectures profondes ou des pertes complexes), les gradients peuvent devenir instables ou exploser. Le **clipping des gradients** consiste à imposer une borne supérieure sur la norme des gradients :

$$g \leftarrow \frac{\tau}{\|g\|} g \quad \text{si } \|g\| > \tau \tag{2.19}$$

où g désigne le vecteur de gradients et τ le seuil choisi (typiquement entre 0.5 et 5). Cette technique est largement utilisée dans les architectures RNN [76] mais aussi en vision dans les entraînements à forte instabilité (transformers, apprentissage contrastif, etc.).

Early Stopping L'**early stopping** est une stratégie simple mais efficace pour éviter le sur-apprentissage. Il consiste à surveiller la performance sur un ensemble de validation, et à interrompre l'entraînement si aucune amélioration significative n'est observée après un nombre prédéfini d'époques (*patience*).

Cette méthode est utilisée en combinaison avec un mécanisme de sauvegarde des meilleurs poids du modèle pour garantir la conservation du point optimal (checkpointing). Elle permet également de réduire les coûts computationnels.

Gestion mémoire et précision mixte (FP16) L'entraînement en **précision mixte** repose sur l'utilisation conjointe de calculs en simple précision (FP32) et demi-précision (FP16). Cette approche permet :

— d'**accélérer** les calculs sur les GPU compatibles (via Tensor Cores),
— de **réduire la consommation mémoire** (jusqu'à 50%),
— de **permettre des batchs plus grands**.

La figure 2.12 illustre ces phénomènes. Les bibliothèques comme `Apex` (NVIDIA) ou `torch.cuda.amp` (PyTorch 1.6+) facilitent l'implémentation de cette technique en garantissant une stabilité numérique, notamment via un mécanisme de **gradients dynamiques mis à l'échelle** (*dynamic loss scaling*) [2].

Résumé L'utilisation de ces méthodes ne modifie pas l'architecture du modèle, mais a un impact direct sur la vitesse de convergence, la qualité de l'optimisation et la robustesse de l'entraînement. Elles sont aujourd'hui incontournables dans les pipelines modernes, y compris dans les tâches de vision complexes comme la Re-Identification ou l'analyse multi-attributs.

2.4.4 Surapprentissage et régularisation

Le surapprentissage (*overfitting*) est l'un des problèmes majeurs rencontrés lors de l'entraînement de modèles d'apprentissage automatique, en particulier dans les réseaux profonds dotés d'un très grand nombre de paramètres.

Définition du surapprentissage On parle de surapprentissage lorsque le modèle apprend à mémoriser les données d'entraînement, y compris leurs bruit et anomalies, au lieu d'extraire des régularités généralisables. Cela se traduit par une très faible perte sur

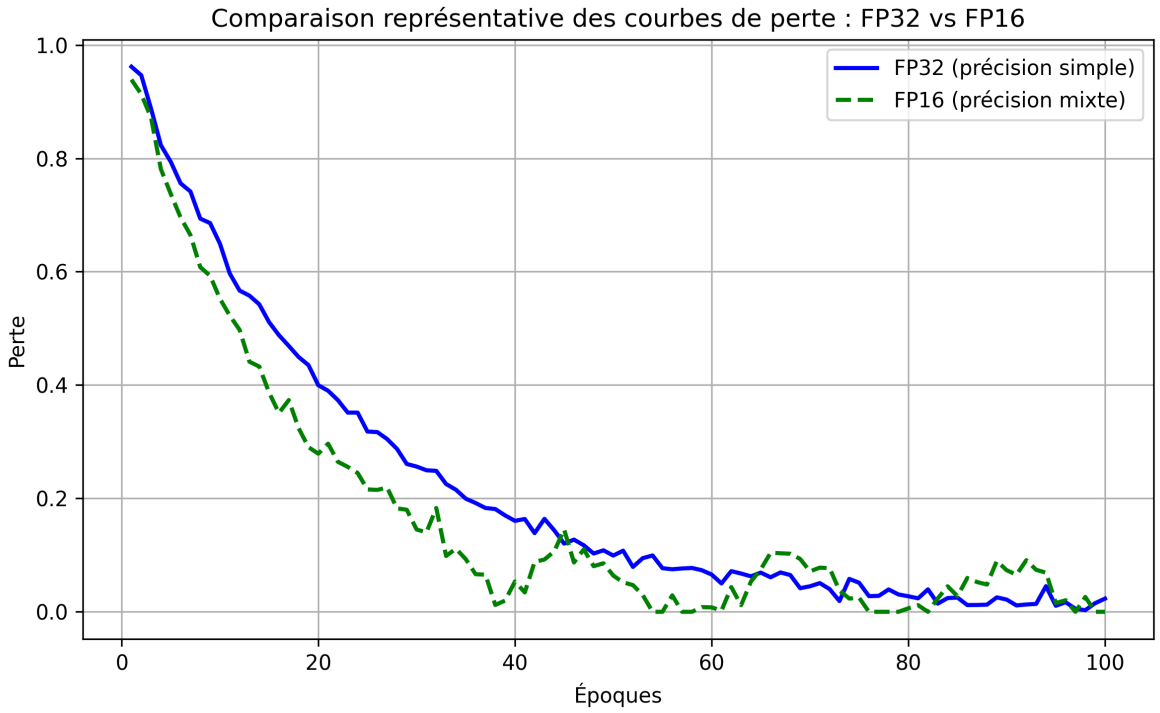


FIGURE 2.12 – Comparaison simulée des courbes de perte pour l'entraînement en précision **FP32** et **FP16**. La précision mixte (FP16) permet une convergence plus rapide dans les premières étapes, mais introduit une légère instabilité et un plateau de performance plus élevé, reflet des limitations numériques potentielles (par exemple : perte de précision dans le calcul de gradients de faible amplitude, entraînant un apprentissage moins efficace sur certaines couches du réseau). Ce phénomène est notamment documenté dans les travaux de [2].

l'ensemble d'entraînement, mais une performance dégradée sur les données de validation ou de test.

Formellement, si l'on note L_{train} la perte sur l'ensemble d'entraînement et L_{val} celle sur validation, on observe un surapprentissage lorsque :

$$L_{train} \downarrow \quad \text{et} \quad L_{val} \uparrow$$

Ce phénomène illustré dans la figure 2.13, est d'autant plus probable lorsque :
— le modèle est trop complexe (surparamétrisé),
— les données d'entraînement sont peu nombreuses ou peu variées,
— l'entraînement se prolonge trop longtemps (nombre d'époques excessif).

Régularisation explicite : Dropout Introduit par Srivastava et al. [77], le **Dropout** consiste à désactiver aléatoirement des neurones du réseau pendant l'entraînement avec une probabilité p, empêchant ainsi la co-adaptation entre neurones.

Soit h une activation intermédiaire, on applique :

$$\tilde{h}_i = h_i \cdot z_i \quad \text{avec} \quad z_i \sim \text{Bernoulli}(1-p)$$

À l'inférence, le réseau est restauré avec tous ses neurones, mais les activations sont multipliées par $(1-p)$ pour compenser.

Dropout est efficace dans les couches entièrement connectées et parfois en convolution (SpatialDropout), bien qu'il soit souvent redondant avec la Batch Normalization.

Régularisation implicite : Data Augmentation La **data augmentation** consiste à générer dynamiquement de nouvelles images d'entraînement en appliquant des transformations aléatoires (géométriques, photométriques, etc.) aux échantillons originaux.

Exemples de transformations usuelles :
— *Flips horizontaux*, *rotations*, *recadrages aléatoires*,
— *Jittering* de couleur, de luminosité, de contraste,
— *Erasing aléatoire*, *MixUp* [78], *CutMix* [79].

Ces méthodes améliorent la généralisation en exposant le réseau à une distribution plus large et en l'empêchant de mémoriser des artefacts spécifiques.

Normalisation interne : Batch Normalization La **Batch Normalization** (BN), introduite par Ioffe et Szegedy [43], stabilise l'entraînement en normalisant les activations d'une couche pour chaque mini-batch :

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mu_B^{(k)}}{\sqrt{\sigma_B^{(k)2} + \epsilon}}, \quad y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)}$$

où $\mu_B^{(k)}$ et $\sigma_B^{(k)}$ sont la moyenne et l'écart-type du batch pour la k-ième activation, et $\gamma^{(k)}, \beta^{(k)}$ sont des paramètres appris.

BN accélère la convergence et joue un rôle régularisant, réduisant la sensibilité aux initialisations et au choix du learning rate.

2.4.5 Métriques d'évaluation

Pour évaluer la capacité de généralisation du modèle, plusieurs métriques sont utilisées selon la tâche :

— **Accuracy** : proportion de prédictions correctes parmi l'ensemble des exemples. C'est la métrique la plus simple et la plus intuitive, mais elle peut être trompeuse lorsque les classes sont très déséquilibrées. Pour une classification multi-classes, elle est définie par :

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}$$

où N est le nombre total d'échantillons, y_i le label réel, et $\hat{y}_i$ la prédiction du modèle. Cette métrique mesure donc la proportion d'exemples correctement classés.

— **Top-k Accuracy** : extension de l'accuracy classique utilisée dans les tâches avec un grand nombre de classes (par ex. ImageNet). Elle vérifie si la bonne classe figure parmi les k prédictions les plus probables du modèle :

$$\text{Top-k Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y_i \in \text{Top-}k(\hat{y}_i)\}$$

- **Top-1 Accuracy** : équivalente à l'accuracy standard. - **Top-5 Accuracy** : une prédiction est correcte si la bonne classe est dans les cinq premières.

— **AUC (Area Under the Curve)** : utilisée principalement en classification binaire et multi-label, en particulier lorsque les classes sont déséquilibrées. Elle mesure la surface sous la courbe ROC (*Receiver Operating Characteristic*), qui trace le *taux de vrais positifs* (TPR) en fonction du *taux de faux positifs* (FPR) [80].

— Un AUC de 0,5 correspond à une performance aléatoire (comme tirer à pile ou face).

— Un AUC de 1,0 indique une séparation parfaite entre classes positives et négatives.

— **Précision, Rappel et F1-score** : métriques particulièrement utiles lorsque les données sont fortement déséquilibrées.

— **Précision** : proportion de prédictions positives correctes parmi toutes les prédictions positives faites par le modèle.

$$\text{Prcision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

— **Rappel** : proportion des vrais positifs correctement détectés parmi toutes les instances positives réelles.

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

où TP, FP, TN et FN correspondent respectivement aux vrais positifs, faux positifs, vrais négatifs et faux négatifs.

— **F1-score** : harmonise précision et rappel via leur moyenne harmonique, offrant une mesure unique équilibrée.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Rappel}}{\text{Precision} + \text{Rappel}}$$

Ces métriques permettent de mieux évaluer un modèle lorsque certaines classes sont rares et que l'accuracy seule serait insuffisante.

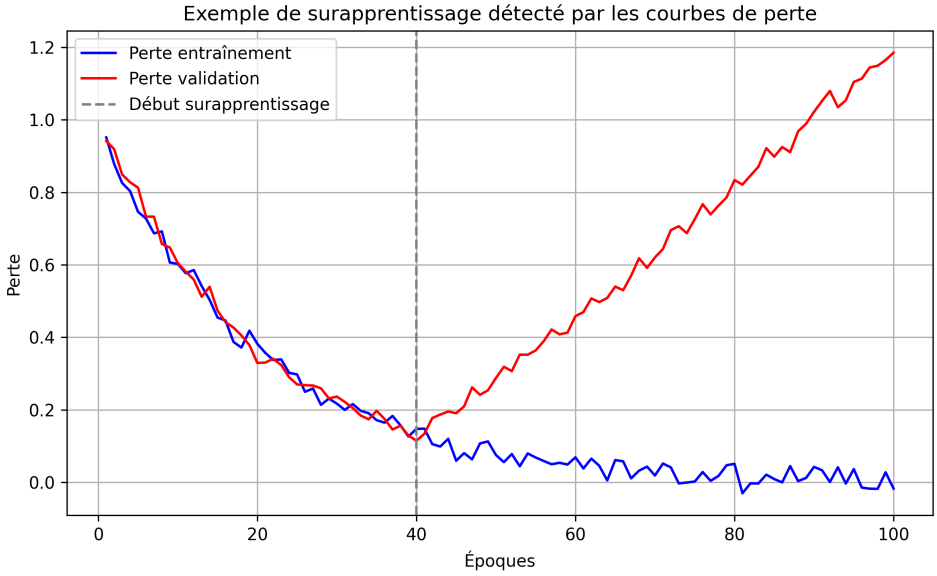


FIGURE 2.13 – Exemple de surapprentissage : la perte d'entraînement continue de diminuer tandis que la perte de validation augmente après un certain nombre d'époques, signalant une perte de généralisation.

Conclusion En somme, l'optimisation de l'entraînement des réseaux de neurones profonds est une tâche multidimensionnelle qui s'étend bien au-delà de la simple définition de l'architecture. La mise en œuvre de techniques d'amélioration pratique telles que l'adaptation dynamique du taux d'apprentissage (Learning Rate Scheduling), le l'écrêtage des gradients (Gradient Clipping), l'arrêt précoce (Early Stopping) et l'utilisation de la précision mixte (FP16) est devenue indispensable. Ces méthodes ne se contentent pas

d'accélérer la convergence ; elles stabilisent le processus d'entraînement, optimisent l'utilisation des ressources et permettent l'exploration d'architectures toujours plus profondes et complexes.

Parallèlement, la lutte contre le surapprentissage est cruciale pour garantir la capacité de généralisation des modèles. Des stratégies de régularisation explicites, comme le Dropout, ainsi que des techniques implicites, telles que la Data Augmentation et la Normalisation par Batch (Batch Normalization), s'avèrent fondamentales. Elles permettent au modèle de développer une compréhension robuste des régularités sous-jacentes des données, plutôt que de mémoriser les spécificités de l'ensemble d'entraînement.

Enfin, l'évaluation rigoureuse de la performance du modèle s'appuie sur un ensemble de métriques adaptées à la nature de la tâche. Que ce soit l'Accuracy, le Top-k accuracy, l'AUC, ou le F1-score et ses composantes (précision et rappel), le choix pertinent de ces indicateurs est essentiel pour refléter fidèlement la capacité du modèle à généraliser sur des données non vues.

L'intégration synergique de ces "astuces" d'entraînement, des techniques de régularisation et d'une évaluation méticuleuse constituent le fondement d'une démarche solide pour concevoir, entraîner et déployer des modèles d'apprentissage automatique fiables et performants dans des applications concrètes.

3

AttriVision : La reconnaissance d'attributs avec CLIP

3.1 Introduction

La reconnaissance d'attributs de personnes (PAR) est un domaine de la vision par ordinateur qui consiste à identifier et catégoriser des attributs visuels tels que le genre, l'âge apparent, le type et la couleur des vêtements, ainsi que la présence d'accessoires, à partir d'images ou de vidéos.

La PAR est utilisée dans des contextes de surveillance [81], d'infrastructures de ville intelligente, de systèmes de sécurité et de commerce en ligne [82], où elle améliore la prise de décision et l'efficacité opérationnelle.

Une exigence fondamentale pour déployer la PAR dans des scénarios réels est la nécessité d'une forte capacité de généralisation, en raison de la diversité des environnements, des capteurs d'image, des conditions d'acquisition, de la densité de personnes et des distributions d'attributs souvent déséquilibrées.

Les méthodes récentes s'appuient sur l'apprentissage profond, notamment sur les réseaux de neurones convolutionnels (CNN) [63, 83, 84, 85], et entraînent des modèles PAR à extraire des caractéristiques hiérarchiques qui capturent à la fois des détails bas niveau — contours et textures — et des informations génériques de haut niveau, telles que les formes et les motifs.

D'autres avancées vers la généralisation de la PAR passent par la constitution de jeux de données exhaustifs, à l'image du *Unified Pedestrian Attribute Recognition* (UPAR) [86], qui agrège des corpus tels que PA100K [87], PETA [88], RAPv2 [89] et Market1501 [90]. Le jeu de données UPAR uniformise quarante attributs binaires répartis en douze catégories afin d'améliorer la robustesse et la capacité de généralisation des modèles PAR, en atténuant les biais propres à chaque base et en augmentant le volume d'exemples d'entraînement. Par ailleurs, des techniques innovantes, comme la décomposition tenso-

rielle [91], sont utilisées pour optimiser les CNN avec une complexité de calcul réduite sans dégrader significativement la précision. Ces progrès facilitent le déploiement de modèles PAR en temps réel sur des dispositifs à ressources limitées, élargissant ainsi l'applicabilité et l'impact de la technologie dans divers domaines [92, 93]. Cependant, la généralisation des modèles PAR reste un verrou majeur [86, 92] : les méthodes actuelles, bien qu'efficaces sur leur jeu d'apprentissage, peinent à se transférer de façon fiable à des données inédites [92, 94].

Comme l'ont montré les modèles fondationnels récents [57, 95], l'amélioration de la généralisation en PAR repose sur deux leviers principaux : (1) faire évoluer l'architecture du réseau ou accroître son échelle [84, 96, 97], *e.g.,* pour modéliser plus finement les relations spatiales entre attributs [98, 99] ; (2) améliorer les données d'entraînement [86], *e.g.,* en augmentant leur volume et leur diversité [10, 87, 88, 90].

Toutefois, collecter de nouvelles données n'est pas aussi trivial pour la PAR que pour d'autres tâches d'apprentissage automatique telles que la segmentation sémantique ou la synthèse vocale [86, 92]. La raison majeure tient à la grande variabilité des scénarios de déploiement, qui impose une curation minutieuse des échantillons afin de couvrir la distribution réelle des images et des étiquettes.

Parmi les facteurs de variabilité figurent l'environnement, les conditions d'éclairage, l'angle de vue et la résolution, qui diffèrent fortement d'un domaine à l'autre et influent sur les performances du modèle. Par exemple, un modèle entraîné sur des images haute résolution sous éclairage homogène se dégrade face à des images basse résolution ou à des éclairages changeants, typiques des situations réelles [92]. S'ajoutent enfin les différences culturelles et régionales, qui modifient la distribution d'attributs tels que les styles vestimentaires ou les accessoires.

Un autre défi majeur réside dans la collecte des données : la plupart des informations pertinentes pour la PAR proviennent de flux de vidéosurveillance captés dans des espaces publics, ce qui soulève d'importantes questions de confidentialité lorsque ces données sont exploitées à des fins de recherche ou commerciales, comme ce fut le cas pour plusieurs ensembles antérieurs [100, 101]. Dans ce contexte, l'option (1), consistant à améliorer l'architecture du réseau pour maximiser l'utilité des données existantes, apparaît donc plus réaliste que l'acquisition de nouveaux corpus.

Nous proposons ainsi de tirer parti des propriétés de généralisation déjà présentes dans le modèle fondationnel image–texte CLIP [57] afin de produire des représentations génériques d'images et d'attributs, adaptées à la généralisation de la PAR. Le schéma général de notre approche est illustré dans la figure 3.1.

CLIP [57] apprend des plongements communs pour les images et les textes dans un même espace latent ; il est entraîné sur des couples image–texte de sorte que leurs représentations soient proches lorsqu'ils sont sémantiquement corrélés. Grâce à une base d'apprentissage massive, ces embeddings se révèlent extrêmement robustes et se généra-

lisent à une large variété de distributions, ce qui rend CLIP particulièrement adapté au transfert *zero-shot* vers de nombreuses tâches aval.

Nous exploitons ces propriétés de généralisation, ainsi que la richesse multimodale des représentations CLIP, pour pallier la faible capacité de transfert observée en PAR. La classification multi-étiquettes image–attributs est reformulée comme un problème de similarité : chaque libellé binaire est converti en une phrase descriptive, par exemple « a photo of a person with sunglasses », pour laquelle CLIP [57] produit un vecteur. Ce vecteur attribut est alors comparé au plongement de l'image, et l'optimisation apprend à rapprocher les deux représentations lorsqu'il existe effectivement l'attribut dans l'image.

Afin de tenir compte du fort déséquilibre de classes propre aux jeux de données PAR et d'accroître la robustesse ainsi que la précision du modèle, nous concevons une fonction de perte *Focal Cross-Entropy* (FCE) spécifiquement adaptée à CLIP ; inspirée de la focal loss pour la détection d'objets [65], elle accentue le poids des exemples difficiles.

Les contributions de ce travail se résument comme suit :

— **Définition de la FCE pour CLIP.** Nous introduisons la perte Focal Cross-Entropy, qui augmente le poids accordé aux exemples les plus difficiles à classer, atténue le déséquilibre de classes et améliore à la fois la précision et la robustesse du modèle.

— **Cadre simple et polyvalent pour la PAR.** Notre approche exploite les capacités du modèle CLIP [57] pour la reconnaissance d'attributs piétons et démontre la flexibilité des encodeurs image et texte de CLIP. Elle assure de hautes performances en PAR et se généralise efficacement à d'autres tâches d'attributs, telles que la reconnaissance d'attributs de véhicules, offrant ainsi une solution robuste et largement applicable.

— **Performances de pointe.** Nous montrons que notre méthode, baptisée *AttriVision*, atteint l'état de l'art sur plusieurs jeux de données de référence pour la PAR et présente une remarquable capacité de généralisation, attestée par ses résultats supérieurs sur le benchmark UPAR.

3.2 Travaux connexes

La PAR est généralement formulée comme un problème de classification multi-étiquettes : un modèle reçoit une image en entrée et produit plusieurs libellés en sortie. Les travaux récents abordent cette tâche en entraînant des réseaux de neurones de manière supervisée [86, 92].

Les *modèles globaux* extraient de l'image une représentation unique, ensuite transmise au classifieur [63, 83], et proposent diverses stratégies de *pooling* pour injecter dans

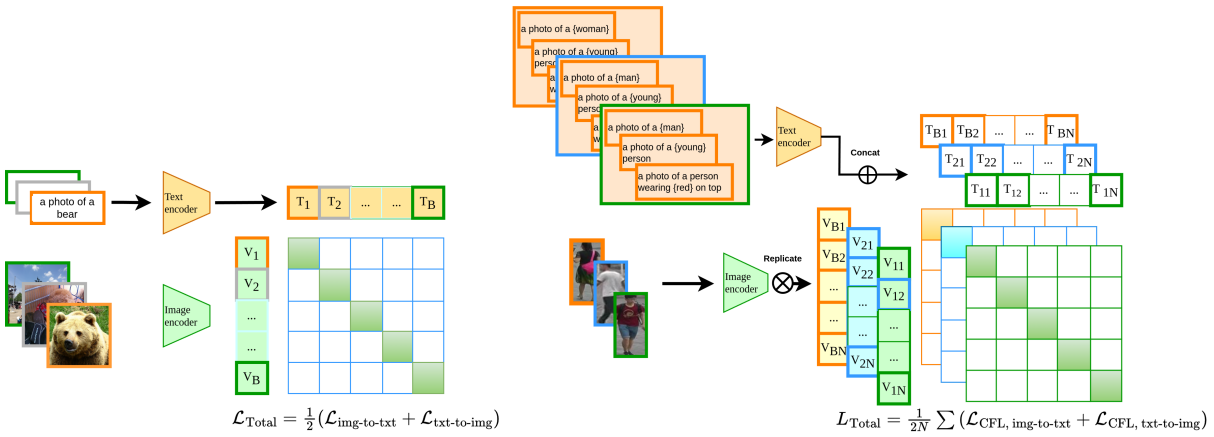


FIGURE 3.1 – Comparaison entre notre approche et CLIP : pour une image donnée d'un lot, nous extrayons les plongements visuels de l'image ainsi que ceux des N attributs associés, puis nous faisons correspondre chaque paire *(images, attribute)*.

ce vecteur global des informations locales plus fines [84, 85]. Les régions d'intérêt locales peuvent être définies à l'aide de poses corporelles prêtes à l'emploi [102, 103], de la distribution spatiale de caractéristiques locales [97, 103], ou, plus récemment, grâce à des mécanismes d'attention [104].

À partir d'une image représentant une personne et de ses attributs associés, nous entraînons CLIP de manière à rapprocher, dans un espace latent commun, la représentation de l'image et celles des attributs correspondants. L'apprentissage est supervisé à partir de paires image–texte (image, attribut).

Dans l'esprit de [57], nous recourons à des invites (*prompts*) en langue naturelle plutôt qu'à des étiquettes brutes : chaque label binaire est converti en une phrase compatible avec CLIP. Pour l'attribut « headwear », la présence est formulée par *"a photo of a person wearing a hat"* et l'absence par *"a photo of a bareheaded person"*.

Nous mettons en place une mise en correspondance systématique permettant d'exprimer, en langue naturelle, la présence comme l'absence de chaque attribut, sans énumérer toutes les combinaisons possibles ; les formulations sont générées dynamiquement pour chaque image.

3.2.1 Perte *Focal Cross-Entropy* (FCE)

La perte FCE combine une similarité cosinus, mise à l'échelle par une température, avec une pondération de type *focal* afin d'ajuster la contribution des paires selon leur difficulté. L'objectif est d'aligner les plongements image–texte pour les paires positives tout en introduisant, via la normalisation, des paires négatives issues du même ensemble de candidats. Nous présentons ci-dessous la formulation utilisée.

Soient f et g les encodeurs d'images et de textes de CLIP. Pour un lot de M images $\{x_i\}_{i=1}^{M}$ et un ensemble de K candidats textuels $\{t_j\}_{j=1}^{K}$ (obtenus à partir des attributs

sous forme d'invites en langue naturelle), on définit les plongements normalisés

$$\hat{\imath}_i \coloneqq \frac{f(x_i)}{\|f(x_i)\|}, \qquad \hat{s}_j \coloneqq \frac{g(t_j)}{\|g(t_j)\|}.$$

Les *logits* sont alors :

$$Z_{ij} \coloneqq \frac{\hat{\imath}_i^\top \hat{s}_j}{\tau},$$

où $\tau > 0$ est un paramètre de température (apprenable). Les probabilités associées aux candidats texte, pour une image donnée, résultent d'une normalisation *softmax* sur les K textes :

$$p_{i,j} \coloneqq \frac{\exp(Z_{ij})}{\sum_{\ell=1}^{K} \exp(Z_{i\ell})}.$$

On note $\mathcal{P}_i \subset \{1, \dots, K\}$ l'ensemble des indices de textes positifs (attributs vrais) pour x_i. La composante contrastive (image→texte) s'écrit :

$$\mathcal{L}_i \coloneqq -\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log p_{i,j}, \qquad \mathcal{L} \coloneqq \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_i.$$

La version FCE introduit une pondération de type *focal* sur les probabilités :

$$\mathcal{L}_i^{\mathrm{FCE}} \coloneqq -\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} (1 - p_{i,j})^{\gamma} \, \log p_{i,j}, \qquad \gamma \geq 0,$$

et la perte totale est $\mathcal{L}^{\mathrm{FCE}} \coloneqq \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} \mathcal{L}_i^{\mathrm{FCE}}$. Dans cette écriture, la similarité cosinus correspond au produit scalaire des plongements normalisés, la température contrôle l'échelle des logits, et la pondération $(1 - p_{i,j})^{\gamma}$ accentue la contribution des paires peu probables du point de vue du modèle. Une extension symétrique (texte→image) peut être ajoutée en définissant un *softmax* par colonne et en moyennant les deux termes.

CLIP [57] est optimisé avec l'objectif contrastif symétrique suivant, qui favorise des scores de similarité plus élevés pour les paires positives (image–texte correspondantes) que pour les paires négatives :

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{2N} \left(\sum_{i=1}^{N} \log \frac{\exp(\mathbf{L}_{ii})}{\sum_{j=1}^{N} \exp(\mathbf{L}_{ij})} + \sum_{i=1}^{N} \log \frac{\exp(\mathbf{L}_{ii})}{\sum_{j=1}^{N} \exp(\mathbf{L}_{ji})} \right), \tag{3.1}$$

où $\mathbf{L}_{ij}$ sont des logits construits à partir de la similarité cosinus entre les plongements normalisés $f(x_i)$ et $g(t_j)$, mise à l'échelle par une température $\tau > 0$, par exemple

$$\mathbf{L}_{ij} \coloneqq \frac{f(x_i)^\top g(t_j)}{\tau \, \|f(x_i)\| \, \|g(t_j)\|}.$$

En posant

$$p_{i,j} \coloneqq \frac{\exp(\mathbf{L}_{ij})}{\sum_{k=1}^{N} \exp(\mathbf{L}_{ik})} \quad \text{et} \quad q_{i,j} \coloneqq \frac{\exp(\mathbf{L}_{ij})}{\sum_{k=1}^{N} \exp(\mathbf{L}_{kj})},$$

on a $\mathcal{L}_{\text{CE}} = -\frac{1}{2N} \sum_{i=1}^{N} \big(\log p_{i,i} + \log q_{i,i}\big)$.

3.3 Implémentation

Nous utilisons le jeu de données du *UPAR Challenge 2024* [86] pour évaluer la généralisation des modèles de reconnaissance d'attributs piétons. Sauf mention contraire, nos expériences suivent le protocole d'évaluation officiel du challenge et les métriques associées.

L'atout majeur d'UPAR réside dans son orientation vers l'adaptation de domaine et la généralisation *cross-dataset*. En proposant des images issues de multiples sources et présentant une large variété de distributions d'attributs et de conditions, le jeu met les modèles au défi de conserver un haut niveau de performances quels que soient les domaines. Cet aspect est crucial pour les applications réelles, où les modèles doivent se montrer fiables au-delà des conditions spécifiques rencontrées lors de l'entraînement. UPAR constitue donc un banc d'évaluation rigoureux et complet pour mesurer l'efficacité des systèmes de reconnaissance d'attributs piétons.

3.3.1 Détails d'implémentation

Association des labels numériques à des phrases descriptives. Au cours de l'entraînement, chaque image d'entrée est associée à plusieurs libellés textuels correspondant aux attributs de vérité terrain. Ces attributs, initialement encodés sous forme de vecteurs binaires, sont d'abord convertis en étiquettes lisibles par l'humain, puis reformulés en phrases descriptives grammaticalement correctes. Concrètement, chaque label PAR, encodé numériquement, est traduit en une phrase qui explicite l'attribut représenté. Par exemple, un label numérique relatif au *genre* de valeur 1 devient « a photo of a man », tandis qu'une valeur 2 est convertie en « a photo of a woman ». Cette association est appliquée systématiquement à l'ensemble des attributs, couvrant un large spectre de catégories : couleur des vêtements, accessoires (lunettes, chapeaux, etc.) ou caractéristiques physiques (âge, longueur des cheveux, *etc.*).

3.3.2 Garantir la cohérence des données entre jeux de données

Normalisation sémantique et textuelle. Soit $\mathcal{D}$ l'ensemble des jeux de données PAR considérés. Pour chaque $d \in \mathcal{D}$, on note $\mathcal{A}_d$ l'inventaire d'attributs et $\mathcal{L}_d$ l'espace d'étiquettes (encodages binaires, multi-classes ou hiérarchiques). Nous définissons une

ontologie commune Ω (attributs canoniques) et une *fonction d'alignement*

$$\rho_d : \mathcal{A}_d \longrightarrow \Omega,$$

qui met en correspondance chaque attribut spécifique à d avec son concept canonique. La textualisation est assurée par un générateur $\tau : \Omega \to \Sigma^*$ fondé sur un ensemble de gabarits contrôlés Π (lexique et morphosyntaxe normalisés). Ainsi, pour une image x et un attribut $a \in \mathcal{A}_d$ positif pour x, la phrase associée est $t = \tau\big(\rho_d(a)\big)$. Ce dispositif unifié (i) rend comparables les descriptions issues de jeux hétérogènes et (ii) réduit la variance linguistique en imposant un lexique et des patrons de surface constants.

Appariement univoque dans les lots d'entraînement. Lors de l'apprentissage contrastif, un lot est constitué d'un ensemble de paires

$$\mathcal{B} = \{(x_i, t_i)\}_{i=1}^{N} \quad \text{avec} \quad t_i = \tau\big(\rho_{d(i)}(a_i)\big),$$

où, pour chaque image x_i, une unique étiquette positive a_i est échantillonnée (les autres positives éventuelles sont reportées à des lots ultérieurs). Nous imposons une contrainte d'injectivité *dans le lot* :

$$i \neq j \;\Rightarrow\; x_i \neq x_j \text{ et } t_i \neq t_j,$$

de sorte que chaque ancre visuelle possède un unique texte positif, et réciproquement. Cette contrainte évite les collisions de positifs au sein du dénominateur de la perte de type InfoNCE (la perte contrastive) et prévient les ambiguïtés dues à la co-présence d'attributs identiques. La perte contrastive s'écrit alors, pour des encodeurs f (image) et g (texte) et une température γ,

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big(\langle f(x_i),\, g(t_i)\rangle/\gamma\big)}{\sum_{j=1}^{N} \exp\big(\langle f(x_i),\, g(t_j)\rangle/\gamma\big)},$$

où, par construction, le seul positif pour x_i dans le lot est t_i. Ce cadre préserve l'intégrité de l'objectif contrastif tout en permettant, au fil des lots, de couvrir l'ensemble des attributs positifs d'une même image.

3.3.3 Augmentation des données et variabilité des phrases

Pour renforcer la robustesse du modèle, nous appliquons des techniques d'augmentation aux images *et* aux phrases descriptives. Côté images, nous recourons aux augmentations classiques telles que la rotation et le retournement horizontal. Côté texte, nous introduisons de la variabilité en sélectionnant aléatoirement un sous-ensemble d'attributs pour décrire chaque image, de sorte que le réseau soit exposé à un éventail diversifié de

descriptions linguistiques durant l'entraînement. Les caractéristiques et la diversité des jeux de données utilisés dans ce cadre sont résumées dans le Tableau 3.1, mettant en évidence les différences en termes de nombre d'attributs, de groupes et de distributions d'entraînement et de test.

Statistiques	**PETA**	**Market-1501**	**PA100K**	**UPAR2024**
Attributs	65	27	26	40
Groupes	17	10	15	12
Images d'entraînement	15067	12936	80000	97669
Personnes test (catégories)	1552	484	849	351
Catégories inédites (test)	1242	315	315	151
Images de test	3933	16483	10000	28095

TABLE 3.1 – Statistiques des jeux de données PETA, Market-1501, PA100K et UPAR2024.

3.3.4 Configuration du modèle

Nous adoptons la version *vanilla* de CLIP fondée sur ViT-B/32, en conservant son pipeline de pré-traitement d'origine : les images sont redimensionnées à 224×224 px, recadrées au centre (*center-crop*) puis normalisées à l'aide des statistiques d'ImageNet. Les tailles des jetons produits par les encodeurs image et texte restent identiques à celles du CLIP initial.

— **Volet image.** L'entrée est découpée en patchs de 32×32 px (configuration ViT-B/32) ; chaque patch est ensuite projeté dans un espace vectoriel de 512 dimensions.

— **Volet texte.** La phrase est tokenisée en une séquence pouvant atteindre 77 jetons, chacun étant à son tour plongé dans un vecteur de 512 dimensions.

3.3.5 Processus d'évaluation

Pour mesurer les performances du modèle, nous calculons la *Mean Accuracy* (mA) et le score *F1*. Comme dans les approches traditionnelles fondées sur des labels numériques ou catégoriels, nous considérons les phrases descriptives générées par le modèle et les phrases de vérité terrain comme deux entités comparables. La mA évalue la proportion moyenne de prédictions correctes par attribut, tandis que le F1 combine précision et rappel afin de fournir une mesure équilibrée de la qualité des prédictions textuelles par rapport aux références.

3.3.6 Résultats

La Table 3.2 présente les performances d'*AttriVision* sur plusieurs jeux de données, et met en évidence une amélioration nette par rapport aux méthodes existantes en PAR. Sur PA100K, *AttriVision* devance largement les autres modèles : +7,2 points de mA et +6,3 points de F1 par rapport à MSCC, et +2,6 mA ainsi que +2,1 F1 face à CLEAR. Les jeux PETA et RAPv2 confirment cette tendance, l'approche atteignant de nouveaux records avec des marges respectives de +2,5 et +2,6 points de mA sur le précédent meilleur modèle, CLEAR. Enfin, sur UPAR, si le gain en mA reste notable, l'amélioration notable de 1,9 point en F1 démontre une meilleure capacité de généralisation que CLEAR.

Méthodes	Backbone	PA100K		PETA		RAPv2		Market1501		UPAR2024	
		mA	F1	mA	F1	mA	F1	mA	F1	mA	F1
MSCC [85]	ResNet50	82.1	86.8	80.8	87.4	80.2	83.0	79.1	83.0	84.1	85.7
VFA [105]	ResNet50	81.3	87.0	86.5	87.3	–	–	–	–	–	–
JLAC [106]	ResNet50	82.3	87.6	87.0	87.5	–	–	–	–	–	–
DAFL [107]	ResNet50	83.5	87.1	89.6‡	88.0	79.5	82.2	–	–	–	–
Strong Baseline	ResNet50	84.0	86.3	81.6	88.1	77.4	78.5	76.5	83.6	82.3	86.4
UPAR2023 [92]	ResNet50	83.5	88.1	89.6‡	88.0	79.5	82.2	–	–	–	–
UPAR2023 [92]	ConvNeXt-B	84.8	90.2	88.4	89.9‡	79.9	81.0	81.5	87.6	–	–
FRDL + GOAT [108]	ConvNeXt-B	88.6‡	89.0	89.4	88.0	–	–	–	–	–	–
CLEAR [109]	Cross-Transformer	87.2	91.0‡	88.2	89.8	81.6‡	87.4‡	83.0‡	87.9‡	85.9‡	90.0‡
AttriVision (nous)	ViT-B/32	**89.8**†	**93.1**†	**90.7**†	**91.5**†	**84.2**†	**89.1**†	**83.8**†	**88.8**†	**89.4**†	**91.9**†

TABLE 3.2 – Comparaison des méthodes PAR sur plusieurs jeux de données. Les meilleures valeurs sont en **gras**† et les secondes meilleures sont marquées d'un ‡.

Le Tableau **??** met en évidence la difficulté intrinsèque de la tâche : les modèles concurrents n'affichent que des gains marginaux entre les tout meilleurs résultats. Par exemple, les méthodes *fanttec* et *hdcolab* présentent des performances équilibrées avec des scores moyens de 72.2 et 71.8, respectivement, soulignant le défi que représente l'optimisation simultanée de la précision et du rappel dans ce domaine. À l'inverse, notre modèle **AttriVision** opère un saut significatif, atteignant une mA de 78.9 et un score F1 de 81.0, soit une moyenne de 79.9. Cet écart net vis-à-vis des autres approches démontre l'efficacité de notre méthode et indique que notre méthode est à même de relever les défis inhérents au concours UPAR 2024 [86], en conservant un rappel et une précision élevés malgré le déséquilibre des données. Cette performance reflète enfin la capacité supérieure du modèle à se généraliser sur des ensembles complexes, surpassant l'ensemble des méthodes concurrentes par une marge significative.

La tâche de *person retrieval* fondée sur les attributs, proposée dans le challenge UPAR 2024, se révèle particulièrement complexe : il s'agit de faire correspondre des individus sur la base d'attributs détaillés, dans des environnements très variés et non contraints. Cette difficulté se manifeste par les faibles scores de Rank-1 (R-1) et de précision moyenne (mA) obtenus par les meilleurs modèles : les trois premières solutions affichent des R-1 compris entre 16.1 et 20.7, et des mA allant de 6.7 à 7.4, comme l'illustre le Tableau 3.3. Même des méthodes plus élaborées, telles qu'ASMR et SAL, peinent à dépasser ces limites :

ASMR atteint un R-1 de 16.8 pour une mA de 5.3, tandis que SAL se situe encore plus bas avec un R-1 de 13.4 et une mA de 3.8.

Malgré la complexité de la tâche, notre modèle AttriVision surpasse nettement toutes les autres approches, obtenant un R-1 de 30.72 et une mA de 21.8. Cet écart important vis-à-vis des méthodes concurrentes — notamment CLEAR (R-1 = 29.1, mA = 14.8) et la ligne de base UPAR (R-1 = 26.2, mA = 13.4) — démontre la supériorité d'AttriVision pour la recherche de personnes par attributs. La capacité du modèle à augmenter simultanément les scores de R-1 et de mA témoigne de son aptitude accrue à se généraliser et à retrouver avec précision les individus dans cette tâche hautement exigeante.

Recherche de personnes par attributs		
Méthode	**R-1**	**mA**
#1 solution	20.7	7.4
#2 solution	16.1	6.7
#3 solution	16.1	6.8
ASMR [110]	16.8	5.3
SAL [111]	13.4	3.8
UPAR2024 baseline [86]	26.2	13.4
CLEAR [109]	29.1	14.8
AttriVision	**30.72**	**21.8**

TABLE 3.3 – Comparaison de notre approche avec d'autres solutions pour la tâche de *person retrieval* du défi UPAR 2024.

3.3.7 Indépendance vis-à-vis du domaine

Les expériences précédentes ont démontré la capacité d'AttriVision à extraire des représentations discriminantes pour la reconnaissance d'attributs piétons. Afin de vérifier si ces représentations conservent leur pertinence au-delà du domaine humain, nous évaluons ici la robustesse et la transférabilité du modèle sur un corpus appartenant à un contexte visuel radicalement différent.

L'évaluation de l'indépendance vis-à-vis du domaine a pour objectif de mesurer la stabilité des représentations apprises lorsque le modèle est exposé à des distributions d'images distinctes de celles rencontrées durant l'apprentissage. Le jeu de données *Stanford Cars* [112], composé d'images de véhicules annotées par type et par marque, constitue à ce titre un banc d'essai particulièrement approprié. Contrairement aux scènes piétonnes, ce corpus présente des structures rigides, des textures homogènes et une sémantique entièrement non anthropocentrée. L'application d'AttriVision sur cet ensemble permet ainsi d'examiner la portée de ses représentations multimodales lorsqu'elles sont confrontées à des régularités visuelles et conceptuelles étrangères à son domaine d'entraînement initial.

Cette expérimentation vise à apprécier la capacité du modèle à préserver un pouvoir discriminant élevé sans adaptation préalable, en distinguant des catégories d'objets

proches sur le plan visuel mais distinctes sur le plan sémantique. Les performances obtenues sur *Stanford Cars* offrent une mesure empirique du degré d'*indépendance au domaine*, c'est-à-dire de la faculté du modèle à produire des représentations invariantes aux variations de contexte et transférables entre domaines visuels hétérogènes.

Méthode	**Backbone**	**Précision (%)**
VehiClassNet-ViT [113]	CNN-SPP	93.0
CSQA-Net [114]	ResNet-50	**95.6**†
SwinTrans [115]	Swin-B	94.2
DCL [116]	R50-ViT-B	94.8
TransFG [117]	ViT-B/16	94.9
ViT-Net [41]	ViT-B/16	95.0
MpT-Trans [118]	ViT-B/16	95.3‡
ACC-ViT [119]	ViT-B/16	94.9
AttriVision	ViT-B/32	<u>95.0</u>

TABLE 3.4 – Comparaison des performances de reconnaissance de type et de marque de véhicule sur le jeu *Stanford Cars* [112]. Les meilleures valeurs sont en **gras**† et les secondes meilleures sont marquées d'un ‡.

Les résultats du Tableau 3.4 indiquent que notre méthode atteint une précision de 95 %, comparable à celle des approches les plus performantes de la littérature. Cette performance, obtenue sans modification spécifique, atteste la robustesse des représentations apprises et confirme leur transférabilité à des tâches de reconnaissance fine-grainée situées en dehors du domaine sémantique initial.

3.4 Étude d'ablation

3.4.1 Analyse

Les résultats obtenus avec la version *off-the-shelf* de CLIP sur le jeu de validation UPAR montrent que le modèle ne parvient pas à capturer efficacement les attributs fins propres aux personnes. Son apprentissage généraliste, fondé sur un large corpus image–texte, ne permet pas une spécialisation suffisante pour des catégories visuelles détaillées comme la couleur ou la longueur des vêtements, les accessoires ou les caractéristiques capillaires. Un ajustement du modèle est donc nécessaire afin d'améliorer la sélectivité et la robustesse de ses représentations visuelles.

Le Tableau 3.5 met en évidence une progression nette sur l'ensemble des catégories du jeu UPAR, avec une précision moyenne de **78,9 %**, contre **32,2 %** pour la version non adaptée de CLIP.

Cette amélioration confirme la capacité du modèle ajusté à mieux isoler les indices visuels discriminants, tout en maintenant une cohérence sémantique globale entre les différentes classes d'attributs.

3.4.2 Représentation efficace des attributs

Formuler une étiquette sous la forme « a person $\{a_1, a_2, a_3, \ldots, a_n\}$ », c'est-à-dire concaténer plusieurs attributs a_i dans une seule phrase, s'avère sous-optimal pour plusieurs raisons. Premièrement, la distribution déséquilibrée des attributs fait que certains a_i sont peu représentés tandis que d'autres dominent, entraînant un apprentissage biaisé lorsque ces informations sont fusionnées dans une même étiquette. Deuxièmement, dans l'espace latent de CLIP, apparier une image à une phrase longue et complexe (« a person $\{a_1, a_2, a_3, \ldots, a_n\}$ ») génère une représentation unique et clairsemée qui affaiblit la capacité de généralisation du modèle, comme l'illustre le Tableau 3.5.

À l'inverse, utiliser plusieurs étiquettes du type « a photo of a a_i » — une par attribut individuel a_i — exploite la force première de CLIP : l'alignement d'images avec des descriptions textuelles courtes et bien définies. Chaque attribut est ainsi appris indépendamment, évitant d'encombrer l'espace latent avec des phrases trop spécifiques et uniques.

Attribut	**Précision AttriVision (%)**	**Précision OTS (%)**
Âge	87.2	9.16
Genre	83.5	57.09
Cheveux	79.8	14.83
Longueur haut du corps	81.3	44.73
Couleur haut du corps	76.5	36.25
Longueur bas du corps	80.4	62.65
Couleur bas du corps	75.2	21.68
Type de bas	79.9	48.84
Chapeaux	82.6	2.80
Sacs	74.7	30.57
Lunettes	70.5	25.54
Précision moyenne	**78.9**	**32.19**

TABLE 3.5 – Précisions de notre méthode comparée à la version *off-the-shelf* de CLIP (OTS) pour chaque attribut du jeu de données UPAR.

3.5 Conclusion

En résumé, *AttriVision* établit un nouveau standard en reconnaissance d'attributs grâce à la combinaison robuste des embeddings multimodaux pré-entraînés de CLIP et d'une perte *Focal Cross-Entropy* (FCE) spécialement conçue pour remédier simultanément au déséquilibre de classes et à la variabilité de domaine. Outre ses excellentes performances en PAR, notre méthode se distingue par ses capacités remarquables en recherche d'images fondée sur les attributs, tirant parti d'un alignement image–texte de très grande finesse. Ses résultats supérieurs sur des jeux de données et des tâches variés témoignent de sa grande généralisabilité, faisant d'*AttriVision* une solution pertinente tant pour la

reconnaissance de personnes que pour des applications plus larges de recherche d'images dans des environnements complexes et réels.

Cependant, la reconnaissance d'attributs n'exploite qu'une partie du potentiel discriminant des représentations visuelles, sans relier directement ces attributs à l'identité des individus. Afin d'étendre ce cadre et de renforcer la cohérence sémantique entre apparence et identité, le chapitre suivant introduit une approche unifiée de **reconnaissance conjointe des attributs et de ré-identification de personnes (PAR–ReID)**. Cette nouvelle formulation vise à apprendre un espace latent partagé où les attributs décrivent non seulement les caractéristiques visibles, mais participent également à la différenciation des identités, ouvrant ainsi la voie à une compréhension plus complète et contextualisée des représentations humaines.

4

Re-Identification conjointe des personnes et de leurs attributs

4.1 Introduction

La re-identification de personnes (ReID) constitue une tâche centrale en vision par ordinateur, notamment dans les contextes de vidéosurveillance, de sécurité urbaine et de gestion intelligente des flux piétons. Elle consiste à retrouver un même individu observé par différentes caméras ou à des instants distincts, malgré les variations d'apparence induites par les changements d'éclairage, d'angle de vue, d'occlusion ou d'habillement. La robustesse d'un système de ReID repose sur sa capacité à apprendre des descripteurs invariants, capables de préserver l'identité d'un individu au-delà de ces transformations visuelles.

En parallèle, la PAR offre une description sémantique complémentaire de l'apparence humaine, en identifiant des propriétés discriminantes telles que le type ou la couleur des vêtements, la présence d'accessoires, le genre ou d'autres caractéristiques physiques. Comme présenté dans le chapitre 3, le modèle **AttriVision** exploite la dimension sémantique au moyen d'un apprentissage multimodal image–texte basé sur CLIP, ce qui permet d'améliorer la reconnaissance d'attributs dans des conditions variées, y compris en présence de données issues de domaines différents de celui utilisé pour l'entraînement (scénarios *out-of-distribution*).

Cependant, la reconnaissance d'attributs ne capture qu'une partie des informations discriminantes nécessaires à l'identification d'une personne. En effet, deux individus peuvent partager un ensemble d'attributs similaires tout en demeurant distincts sur le plan identitaire. Réciproquement, la re-identification seule ignore souvent la structure sémantique des attributs qui, pourtant, constituent des descripteurs interprétables et transférables.

Ces deux tâches sont donc intimement liées : l'une encode des indices sémantiques, l'autre encode des invariants d'identité. Les traiter conjointement permet de renforcer à

la fois la robustesse de la ReID et la cohérence sémantique de la PAR.

Dans cette optique, nous proposons une méthode d'**apprentissage conjoint des attributs et de l'identité**, qui prolonge les principes établis dans le chapitre précédant. Notre approche repose sur le modèle vision–langage CLIP, mais introduit une tête intermédiaire destinée au calcul de centroïdes de caractéristiques pour chaque identité. Cette représentation structure l'espace latent de manière à rapprocher les instances partageant une même identité tout en maintenant un alignement sémantique avec les attributs textuels. Le résultat est un espace unifié où les entités visuelles et linguistiques cohabitent, garantissant une compatibilité entre la représentation d'un individu et la description de ses attributs.

Sur le plan de l'optimisation, le modèle combine la Focal Cross-Entropy, introduite dans la secton 3.2.1 du chapitre précédant, pour gérer les déséquilibres inter-attributs, avec la *Triplet Loss* décrite dans 2.4.1, utilisée ici pour contraindre la cohérence métrique entre identités. Cette combinaison permet de conserver les avantages du cadre AttriVision tout en ajoutant une supervision supplémentaire spécifique à la ReID, aboutissant à des embeddings à la fois discriminants et sémantiquement interprétables.

Ce chapitre présente en détail cette extension conjointe, en décrivant successivement les fondements théoriques de l'apprentissage partagé, l'architecture proposée et les résultats expérimentaux obtenus sur plusieurs jeux de données publics, dont *PETA* [12], *UPAR* [15] et *RAPv2* [13]. L'analyse met en évidence la capacité de cette approche à maintenir des performances élevées tout en favorisant une compréhension unifiée des attributs et des identités. Enfin, la discussion conclut sur les apports de cette intégration multimodale et les perspectives ouvertes vers une représentation plus éthique, complète et cohérente de l'identité humaine.

4.2 État de l'art

4.2.1 Introduction générale à la tâche de Re-Identification

La re-identification de personnes, ou Person Re-Identification (ReID), est une tâche de vision par ordinateur qui consiste à associer des images d'un même individu prises à différents moments ou depuis des caméras distinctes. Dans les systèmes de surveillance multi-caméras, cette capacité permet notamment le suivi inter-caméras, la recherche de personnes et l'analyse de flux de population. Contrairement à la reconnaissance faciale, fondée sur des traits biométriques, la ReID s'appuie sur des données souvent capturées dans des environnements non contrôlés, où le visage peut être partiellement ou totalement occulté.

La difficulté principale de la ReID réside dans la variabilité des apparences visuelles pour une même identité. Plusieurs facteurs compliquent la tâche : (i) les variations d'éclai-

rage et de conditions météorologiques entre les caméras, (ii) les différences d'angle de vue et de posture, (iii) les occlusions partielles causées par des objets, d'autres personnes ou des éléments de l'environnement, et (iv) la qualité hétérogène des dispositifs de capture. Par ailleurs, la similarité visuelle entre individus, liée par exemple à des morphologies proches ou à des habits similaires, exige du système une capacité de discrimination précise.

La ReID peut être formulée comme un problème de recherche dans une base d'images : à partir d'une image requête (query), il s'agit de retrouver, dans une galerie (gallery), les images représentant la même personne. Cette tâche peut être abordée selon deux scénarios : (i) fermé (closed-set), où toutes les identités sont présentes lors de l'entraînement, ou (ii) ouvert (open-set), où le modèle doit généraliser à des identités non vues. Les protocoles d'évaluation courants reposent notamment sur la Mean Average Precision (mAP) et les courbes de taux de reconnaissance aux premiers rangs (CMC@1, CMC@5, CMC@10). La courbe CMC (*Cumulative Matching Characteristic*) représente, pour chaque rang k, le taux de requêtes pour lesquelles au moins une image correspondant à la bonne identité est retrouvée parmi les k premières positions de la liste classée.

Historiquement, la ReID s'est d'abord appuyée sur des méthodes exploitant des caractéristiques manuelles, telles que les histogrammes de couleur ou les descripteurs texturaux (LBP, HOG) [120, 121]. Cependant, ces approches ont montré des limites en termes de robustesse aux conditions réelles. Avec l'avènement de l'apprentissage profond, les réseaux convolutionnels (CNN) ont permis des avancées majeures, notamment via des approches supervisées par classification d'identités [122, 123] ou via apprentissage métrique [124]. Plus récemment, les modèles basés sur l'attention et les Transformers [125] ont montré leur capacité à modéliser des dépendances globales dans l'image, surpassant souvent les performances des CNN classiques.

Malgré les avancées récentes, certaines limites demeurent. Les modèles de ReID présentent une sensibilité importante au changement de domaine : un modèle entraîné sur un jeu de données donné peut perdre en performance lorsqu'il est appliqué à un autre environnement de capture. Par ailleurs, les attributs sémantiques tels que le genre, le style vestimentaire ou la présence d'accessoires restent peu exploités dans les approches classiques, alors qu'ils peuvent fournir des indices complémentaires utiles à la différenciation entre identités. Dans cette perspective, certaines méthodes récentes intègrent des descripteurs d'attributs [10, 126], ou s'appuient sur des modèles vision-langage préentraînés comme CLIP [11] ou d'autres variantes plus récentes [58].

La ReID peut être abordée selon deux modalités : image-par-image (image-based) ou à partir de séquences vidéo (video-based). Dans ce dernier cas, les vidéos permettent de combiner des informations spatiales et temporelles en agrégeant les caractéristiques extraites sur plusieurs images. Cette approche permet de modéliser des dynamiques plus complexes, au prix d'un coût computationnel plus élevé [127, 128]. Ce chapitre se concentre

sur la ReID à partir d'images individuelles, tout en tenant compte du développement des méthodes vidéo, en particulier dans les contextes de surveillance à grande échelle.

Cette section se poursuit par une analyse critique des méthodes classiques de ReID, mettant en évidence leurs fondements, leurs performances, mais aussi leurs limitations en contexte réel.

4.2.2 Méthodes traditionnelles à base de descripteurs manuels

Avant l'émergence des approches par apprentissage profond, la re-identification de personnes reposait sur des descripteurs visuels extraits manuellement. Ces descripteurs visaient à représenter l'apparence des individus en s'appuyant sur des caractéristiques simples et interprétables telles que la couleur, la texture ou la forme. Bien qu'ils aient constitué une première étape dans le développement de la ReID, ces dispositifs se sont révélés sensibles aux variations des conditions de capture et limités en termes de généralisation.

Histogrammes de couleurs Les histogrammes de couleur figurent parmi les premières méthodes utilisées pour modéliser l'apparence des personnes. Ils décrivent la distribution des couleurs dans une image ou une région d'image, généralement dans des espaces tels que RGB, HSV ou YCbCr. Ce type de descripteur permet de représenter la dominante chromatique des vêtements. Gray et Tao [120] ont montré que l'utilisation de l'espace HSV pouvait offrir une certaine robustesse aux variations d'intensité lumineuse. Toutefois, ces descripteurs présentent des limites en cas de changements d'éclairage entre caméras ou lorsque plusieurs individus portent des tenues visuellement proches.

Descripteurs de texture Pour pallier les limites des histogrammes de couleur, des descripteurs texturaux tels que les Local Binary Patterns (LBP) et les Histograms of Oriented Gradients (HOG) ont été introduits. Les LBP encodent les motifs locaux en comparant chaque pixel à ses voisins, ce qui permet de générer une signature relativement stable face aux variations d'éclairage local [129]. Les HOG, proposés par Dalal et Triggs [121], décrivent la distribution des orientations de gradient dans l'image et sont conçus pour capturer la forme globale des objets. Ces descripteurs permettent de détecter certains motifs visuels (plis, coutures, textures), mais restent limités lorsqu'il s'agit de différencier des individus ayant une apparence similaire, en particulier dans des environnements denses.

Descripteurs fréquentiels et filtrage multi-orientation Les filtres de Gabor, inspirés de la réponse du cortex visuel humain, permettent d'analyser des régions d'image selon différentes fréquences spatiales et orientations. Petkov et Westenberg [130] ont montré que

ces filtres sont utiles pour extraire des motifs répétitifs dans les textures vestimentaires. Lorsqu'ils sont appliqués sur une pyramide multi-échelle, ils offrent une certaine robustesse aux variations de taille et d'orientation. Cependant, leur redondance élevée et leur coût computationnel freinent leur usage dans des systèmes en temps réel.

Combinaisons et encodages hybrides Certaines approches ont combiné des descripteurs de couleur et de texture afin d'enrichir la représentation visuelle. Par exemple, ColorLBP concatène des histogrammes de couleur avec des motifs LBP extraits de différentes parties du corps. Ces représentations peuvent ensuite être normalisées, agrégées par moyenne, ou encodées à l'aide de méthodes comme Bag-of-Words (BoW) ou Fisher Vectors [131]. De telles techniques ont été utilisées dans les premières évaluations sur des jeux de données comme VIPeR ou i-LIDS. Leur performance reste toutefois limitée face aux méthodes récentes, notamment sur des benchmarks de plus grande échelle ou plus variés.

Limitations générales Les méthodes basées sur des descripteurs manuels présentent plusieurs limites importantes. Elles dépendent fortement de l'alignement précis des silhouettes, et une erreur de détection peut altérer significativement la représentation obtenue. Leur capacité à capturer des informations sémantiques de haut niveau est très limitée. Par ailleurs, ces approches ne permettent pas d'apprendre une représentation spécifique à la tâche de re-identification, ce qui réduit leur capacité à généraliser à des contextes variés.

L'émergence des bases de données de grande taille et des capacitées de calcul accélerées par GPU ont progressivement conduit à l'abandon des descripteurs manuels au profit de méthodes d'apprentissage profond, que nous détaillons dans la section suivante.

Apprentissage supervisé par CNN

Avec la disponibilité croissante de données annotées à grande échelle et les progrès matériels liés au calcul parallèle sur GPU, l'apprentissage profond a profondément modifié les approches en re-identification. Les réseaux de neurones convolutionnels (CNN) se sont imposés comme une méthode de référence, en apprenant automatiquement des représentations discriminantes à partir d'images brutes. Contrairement aux descripteurs manuels, les CNN peuvent extraire des informations hiérarchiques pertinentes pour la tâche, allant des motifs bas niveau aux configurations globales du corps. Des architectures comme ResNet-50 [47], combinées à des têtes de classification ou à des modules métriques, ont montré une nette amélioration des performances sur des benchmarks standard tels que Market-1501 [90] et DukeMTMC-ReID [100].

Classification d'identités avec supervision totale La manière la plus directe d'utiliser un CNN pour la re-identification consiste à formuler la tâche comme une classification supervisée, où chaque identité est associée à une classe. Le réseau est entraîné à prédire l'étiquette correcte à l'aide d'une couche softmax et d'une fonction de perte de type cross-entropy. Cette approche, introduite dans le modèle IDE (ID-discriminative Embedding) [122], est généralement mise en œuvre avec des architectures de base telles que ResNet-50 [47] ou Inception-v3 [132]. Les représentations extraites avant la couche de classification peuvent ensuite être utilisées pour la comparaison entre individus.

Une fois l'apprentissage terminé, la couche softmax est retirée, et les activations issues du dernier bloc convolutif ou de la couche de pooling global sont utilisées comme vecteurs d'embedding. Ces vecteurs servent ensuite à comparer les images entre elles, à l'aide d'une distance cosinus ou euclidienne, pour réaliser la re-identification sur les ensembles de test.

Part-based CNN et modélisation locale Pour intégrer des informations locales plus détaillées, certaines architectures CNN intègrent un partitionnement spatial explicite des cartes de caractéristiques. Le modèle PCB (Part-based Convolutional Baseline) [123] divise la carte de caractéristiques en bandes horizontales de taille fixe, chacune étant traitée indépendamment avant agrégation. Cette structure permet de maintenir une forme de correspondance entre les différentes régions du corps, telles que la tête, le torse ou les jambes. D'autres modèles, comme MGN (Multiple Granularity Network) [133], combinent plusieurs niveaux de granularité (globale, locale, multi-part) en parallèle pour produire des représentations plus diversifiées. Ces architectures ont permis d'améliorer les performances sur des jeux de données courants tels que Market-1501 ou DukeMTMC-reID.

Augmentations et stratégies régularisatrices Les CNN utilisés pour la ReID tirent parti d'opérations d'augmentation de données visant à améliorer leur capacité de généralisation. Une technique fréquemment utilisée est le Random Erasing [134], qui consiste à masquer aléatoirement une région rectangulaire de l'image pendant l'entraînement. Cette opération simule des occlusions partielles et limite la dépendance du réseau à des régions spécifiques de l'image. D'autres transformations classiques, comme le retournement horizontal, les ajustements de luminosité ou l'application de la normalisation par batch (BatchNorm), sont également courantes. Des techniques plus avancées, telles que le mixup, l'augmentation par GAN, ou encore le dropout spatial, sont parfois employées pour accroître la diversité des configurations rencontrées en phase d'apprentissage.

Limites des approches CNN supervisées Bien qu'efficaces, les approches CNN supervisées présentent plusieurs limites. Leur formulation en classification impose un cadre fermé, dans lequel les identités vues à l'entraînement ne correspondent pas à celles de test. Le réseau apprend ainsi des représentations génériques sans supervision explicite sur

les distances inter-identités. Par ailleurs, ces méthodes requièrent de grandes quantités de données annotées, ce qui en limite l'utilisation dans des contextes où l'annotation est coûteuse ou difficile à obtenir. Enfin, les architectures CNN traditionnelles se concentrent sur des motifs locaux et peinent à capturer des relations globales ou des dépendances contextuelles, comme les associations entre accessoires ou les éléments de l'arrière-plan. Ces limitations ont conduit au développement de modèles intégrant des mécanismes d'attention, qui seront abordés dans les sections suivantes.

4.2.3 Apprentissage métrique et pertes contrastives

Au-delà des approches supervisées classiques qui traitent la tâche de Re-Identification comme un problème de classification fermée, les méthodes fondées sur l'**apprentissage métrique** visent à apprendre une fonction d'encodage projetant les images de personnes dans un espace latent où les distances reflètent la similarité d'identité. Cette stratégie est particulièrement adaptée aux scénarios de REID ouverts, où les identités rencontrées lors du test diffèrent de celles vues à l'entraînement.

Dans le cadre de notre approche, nous exploitons la **Triplet Loss** [124], présentée en détail dans la section 2.4.1, comme composante principale de la supervision métrique. Cette fonction de perte favorise la proximité des représentations appartenant à une même identité tout en maximisant la distance avec celles d'identités distinctes, structurant ainsi l'espace des embeddings selon la cohérence identitaire. Nous combinons cette contrainte avec la *Focal Cross-Entropy*, déjà introduite dans la section 3.2.1, afin de renforcer la stabilité de l'apprentissage dans les contextes à distribution déséquilibrée d'attributs.

Cette combinaison assure un équilibre entre discrimination inter-identités et spécialisation intra-catégorie, tout en permettant à l'espace latent d'intégrer simultanément des signaux sémantiques (issus des attributs) et discriminants (liés aux identités).

L'efficacité de la Triplet Loss dépend fortement de la stratégie de *sélection des exemples* (*mining*) des exemples. En particulier, le *hard negative mining* consiste à choisir les triplets les plus difficiles (négatif proche ou positif éloigné), ce qui accélère la convergence mais peut provoquer une instabilité de l'apprentissage si mal contrôlé.

Apprentissage métrique avec Contrastive Loss Une autre fonction de perte couramment utilisée en apprentissage métrique est la **Contrastive Loss** [135], définie sur des paires d'images. Elle encourage les paires positives (même identité) à se rapprocher dans l'espace d'embedding, et pousse les paires négatives (identités différentes) à s'éloigner au-delà d'un certain seuil. La formulation est la suivante :

$$\mathcal{L}_{\text{contrastive}} = y \cdot d^2 + (1 - y) \cdot \max(0, m - d)^2$$

où $y = 1$ pour les paires positives et $y = 0$ pour les paires négatives, d est la distance

entre les embeddings, et m une marge. Cette perte est plus simple à mettre en œuvre que la Triplet Loss, mais elle est souvent considérée comme moins expressive car elle ne prend en compte que deux images à la fois.

Pertes unifiées : Circle Loss Des variantes plus récentes ont été proposées pour combiner les avantages des pertes contrastive et triplet dans un cadre unifié. La **Circle Loss** [22] modélise simultanément les relations positives et négatives à l'aide d'un mécanisme de pondération circulaire basé sur la similarité entre les vecteurs :

$$\mathcal{L}_{\text{circle}} = \log \left[1 + \sum_{(i,j)^+} \exp(-\gamma \cdot \alpha_{ij} \cdot (s_{ij} - \Delta_p)) + \sum_{(i,j)^-} \exp(\gamma \cdot \alpha_{ij} \cdot (s_{ij} - \Delta_n)) \right]$$

où s_{ij} désigne la similarité normalisée entre les représentations des échantillons i et j. Les ensembles $(i,j)^+$ et $(i,j)^-$ correspondent respectivement aux paires positives (même identité) et négatives (identités différentes). Le coefficient α_{ij} est un poids adaptatif dépendant de la difficulté de la paire considérée, permettant d'accorder une importance accrue aux paires les plus informatives. Le paramètre $\gamma > 0$ contrôle l'échelle de la fonction de perte et influence la vitesse de convergence, tandis que Δ_p et Δ_n définissent respectivement les marges cible pour les paires positives et négatives.

Techniques de mining avancées Plusieurs stratégies de sélection des triplets ou des paires ont été proposées pour améliorer l'efficacité de l'apprentissage métrique. Le *batch-hard mining* sélectionne, dans chaque mini-lot, les exemples les plus difficiles pour chaque ancre. Le *semi-hard mining*, utilisé par FaceNet [67], choisit les négatifs qui sont plus proches que le positif, mais encore correctement classés. D'autres approches comme le *distance-weighted sampling* [136] pondèrent les échantillons négatifs en fonction de leur distance à l'ancre pour équilibrer l'influence des exemples faciles et difficiles.

Des pertes alternatives comme la Lifted Structured Loss [137] ou la N-pair Loss [138] exploitent plus efficacement les mini-lots en considérant simultanément plusieurs paires utiles au lieu de se limiter à des triplets indépendants.

Limites et perspectives L'apprentissage métrique permet de produire des représentations discriminantes sans dépendre explicitement du nombre de classes. Toutefois, plusieurs limites subsistent dans sa mise en œuvre. La qualité de l'apprentissage dépend fortement de la stratégie de sélection des exemples (mining), ce qui peut affecter la stabilité et la convergence du modèle. Par ailleurs, en présence d'attributs fortement déséquilibrés ou absents dans les données d'entraînement, les distances apprises peuvent manquer de robustesse.

Des travaux récents ont proposé des améliorations structurelles pour dépasser ces contraintes. Le modèle introduit dans [128] propose un backbone optimisé pour les séquences vidéo, réduisant la dépendance à la densité des frames. La méthode *Disentangled ReID* [139] vise quant à elle à séparer explicitement les composantes liées à l'identité de celles associées au contexte (comme la pose ou l'arrière-plan), rendant l'espace d'embedding plus interprétable et potentiellement plus robuste.

Conclusion Les fonctions de perte métrique ont joué un rôle central dans l'évolution des systèmes de ReID, en permettant un apprentissage fondé sur la similarité plutôt que sur une classification stricte. Elles offrent ainsi un cadre plus flexible, particulièrement adapté aux scénarios ouverts où les identités diffèrent entre l'entraînement et le test. Toutefois, ces approches reposent souvent sur une supervision fortement structurée, comme la constitution de triplets ou de paires équilibrées, et restent limitées dans leur capacité à exploiter des informations sémantiques explicites, telles que les attributs ou les descriptions textuelles. Dans cette perspective, les modèles vision-langage et l'apprentissage conjoint des identités et des attributs constituent des pistes prometteuses, que nous explorerons dans les sections suivantes.

4.2.4 Attention et Transformers

L'introduction des mécanismes d'attention dans les modèles de Re-Identification a marqué une étape importante pour dépasser les limitations structurelles des réseaux convolutionnels. Contrairement aux CNN, dont le champ de réception est local et dépend du noyau de convolution, l'attention permet de modéliser des dépendances globales au sein de l'image. Ce mécanisme renforce les régions visuellement discriminantes tout en atténuant les éléments contextuels non pertinents. Cette capacité est particulièrement utile en ReID, où les indices de différenciation sont souvent localisés (par exemple, un sac, des chaussures, ou des motifs vestimentaires spécifiques). Pour un rappel des mécanismes d'attention et de leur application aux ViT, voir chapitre 2, sections 2.2.2 et 2.2.3.

Modules d'attention dans les CNN Des modèles hybrides ont introduit des blocs d'attention dans des architectures convolutionnelles classiques pour améliorer la représentation visuelle. Le *Residual Attention Network* [42] intègre des modules d'attention spatiale et de canal qui permettent de pondérer dynamiquement les activations selon leur importance contextuelle. Le modèle *ABDNet* (Attention-Based Dual-branch Network) [140] combine, quant à lui, une attention spatiale globale avec un alignement local dans deux branches parallèles. Ces approches renforcent la qualité des embeddings en incitant le réseau à se concentrer sur les régions pertinentes pour la distinction des identités.

Transformers pour la ReID L'introduction des Transformers en vision par ordinateur, en particulier avec le *Vision Transformer* (ViT) [141], a conduit à leur intégration progressive dans les systèmes de Re-Identification. Pour une présentation détaillée de l'architecture ViT et de ses principes de projection par patchs, on renverra au chapitre 2, section 2.2.4. He et al. ont proposé *TransReID* [125], une architecture spécifiquement adaptée aux images de personnes. Ce modèle incorpore des mécanismes de positionnement ainsi qu'un renforcement explicite de la structure spatiale. TransReID repose sur un backbone ViT enrichi par des embeddings de parties corporelles alignées, et introduit un *camera-aware position embedding* permettant de mieux modéliser les variations de distribution induites par le changement de caméra, phénomène courant dans les systèmes multi-vues.

Capacités et ouverture vers le vision-langage Les Transformers offrent la capacité de modéliser des dépendances à longue portée, ce qui les rend sensibles à des motifs globaux tels que la cohérence vestimentaire ou l'alignement des poses corporelles. Cette propriété permet de dépasser les limites des approches purement locales. Par ailleurs, la structure flexible des Transformers à l'entrée les rend particulièrement adaptés à la fusion multi-modale, notamment texte-image. Cette compatibilité naturelle avec les représentations linguistiques ouvre la voie à des approches de ReID fondées sur des modèles vision-langage.

Part Attention Transformer Le modèle PAT (Part Attention Transformer) [142] propose une attention localisée sur les différentes parties anatomiques du corps humain. En combinant une segmentation explicite avec un mécanisme d'attention hiérarchique, PAT améliore la localisation et la díscrimination des attributs locaux tout en capturant les relations globales.

Vision multi-caméra et attention contextuelle Certains travaux récents cherchent à modéliser explicitement les biais introduits par les dispositifs de capture hétérogènes. La méthode présentée dans [143] propose un mécanisme d'attention conditionné par les caractéristiques propres à chaque caméra, telles que la résolution, l'angle de vue ou la dominante colorimétrique. Cette attention contextuelle permet d'adapter dynamiquement les représentations extraites en fonction du domaine de capture, améliorant ainsi la capacité de généralisation des modèles de ReID dans des environnements multi-caméras.

Avantages et limitations Les modèles fondés sur les Transformers surpassent généralement les architectures convolutionnelles sur les grands jeux de données de ReID, en particulier en termes de précision moyenne (mAP) et de taux de reconnaissance en top-1. Leur capacité à modéliser globalement la structure corporelle et à capturer les corréla-

tions visuelles réduit le besoin de modules explicites de type part-aware. Cependant, ces modèles présentent également certaines limitations : ils nécessitent des ressources computationnelles élevées (notamment en mémoire GPU), et peuvent se montrer sensibles au bruit et au surapprentissage dans le cas de bases de données de petite taille ou faiblement annotées.

Conclusion Les mécanismes d'attention, et en particulier les Transformers, représentent une avancée significative pour la Re-Identification. Ils permettent de dépasser les limites des architectures convolutionnelles classiques, en intégrant des relations globales et contextuelles essentielles pour la distinction fine entre identités. Par ailleurs, leur structure modulaire et flexible en entrée constitue une base pertinente pour le développement d'approches multi-modales. Cette propriété ouvre la voie à l'intégration de signaux visuels et linguistiques, comme le proposent les récents travaux en vision-langage, et sert de fondement à l'approche conjointe explorée dans ce manuscrit.

Intégration des attributs comme supervision auxiliaire

Les attributs visuels humains — tels que le genre, la couleur des vêtements ou la présence d'accessoires comme des sacs ou des chapeaux — représentent des informations sémantiques intermédiaires pouvant enrichir la représentation visuelle d'un individu dans le cadre de la re-identification. Contrairement aux identifiants d'identité, qui sont spécifiques au jeu de données et arbitraires, les attributs sont partageables entre contextes, interprétables, et généralement plus robustes à la variabilité des domaines. Leur intégration dans le processus d'apprentissage permet d'orienter l'encodage des caractéristiques vers des dimensions discriminantes et d'améliorer la généralisation du modèle dans des contextes non vus.

Approches multi-tâches et guidage par attributs Une première manière d'exploiter les attributs dans la ReID consiste à introduire une supervision multi-tâches. Lin et al. [10] proposent de combiner deux objectifs : une perte de classification d'identité (cross-entropy) et une perte d'attributs formulée comme une classification multi-label (binary cross-entropy). Ce cadre permet au backbone convolutionnel d'apprendre des caractéristiques partagées tout en étant contraint par des signaux sémantiques supplémentaires. Cette régularisation améliore la convergence et favorise une meilleure généralisation aux données non vues.

Tay et al. [126] introduisent un encodeur d'attributs explicite, couplé à une branche dédiée à l'identité. Les deux représentations sont fusionnées dans un espace latent commun. Ce type d'architecture favorise l'alignement entre images partageant les mêmes attributs, ce qui renforce la cohérence intra-classe, même en présence de variations visuelles importantes.

Fusion attribut-identité La fusion des attributs avec les représentations d'identité peut se faire à plusieurs niveaux :

- **Fusion précoce (early fusion)** : les descripteurs attributaires sont concaténés directement aux embeddings visuels.
- **Fusion tardive (late fusion)** : les scores d'attributs servent à pondérer les similarités ou à réordonner les galeries.
- **Fusion intermédiaire (cross-modal fusion)** : les attributs textuels ou vectoriels servent à conditionner des modules d'attention, comme dans les modèles vision-langage.

Des travaux récents, tels que FusionSegReID [144], ont montré qu'une combinaison d'informations visuelles et sémantiques, associée à l'utilisation de masques de segmentation, permet d'améliorer significativement la qualité des embeddings. Ce couplage rend les représentations plus robustes face aux variations de contexte, notamment en réduisant l'influence des éléments de fond.

Avantages et défis L'intégration des attributs en ReID présente plusieurs bénéfices. Elle renforce la robustesse du modèle aux changements d'apparence, en s'appuyant sur des caractéristiques stables telles que le genre, la morphologie ou certains accessoires (sacs, lunettes), qui persistent à travers les séquences. Elle apporte également une meilleure interprétabilité, en permettant une lecture sémantique des représentations et des décisions du système. Enfin, en contraignant le réseau à apprendre des concepts généraux et partageables, elle limite les risques de sur-apprentissage liés à une supervision strictement identitaire.

Ces avantages s'accompagnent néanmoins de défis, notamment liés à la disponibilité et à la qualité des annotations d'attributs, souvent coûteuses à acquérir. De plus, l'intégration multi-source (identité, attributs, contexte) requiert des stratégies d'alignement ou de fusion soigneusement conçues pour éviter les conflits de supervision.

De plus, ces approches présentent certaines limitations. La qualité des annotations attributaires a un impact direct sur les performances, or de nombreux jeux de données disponibles comportent des étiquettes partielles, bruitées ou ambiguës. Par ailleurs, la représentation des attributs sous forme binaire ou discrète ne permet pas de capturer l'ensemble des nuances visuelles nécessaires à une distinction fine entre identités visuellement proches.

Vers une fusion attribut-identité profonde Les approches récentes basées sur des modèles vision-langage, tels que CLIP ou des Large Vision-Language Models (LVLM) [58], permettent de dépasser certaines limitations des méthodes précédentes. En formulant les attributs comme des requêtes textuelles en langage naturel, ces modèles établissent un alignement sémantique entre image et description attributaire, sans dépendre d'annota-

tions explicites ou binaires pour chaque attribut. Cette formulation rend possible une supervision faible, plus souple et mieux adaptée aux contextes variés.

Cette capacité d'alignement croisé constitue un élément central de notre approche. Nous proposons un apprentissage conjoint des représentations d'attributs et d'identités à l'aide d'un modèle vision-langage, entraîné à partir de pertes combinées — une Focal Cross-Entropy pour la robustesse aux déséquilibres, et une Triplet Loss pour la discrimination métrique. Cette optimisation est renforcée par l'utilisation de centroïdes d'identité en mémoire, servant d'ancrage explicite pour structurer l'espace d'embedding.

Travaux récents et paradigmes émergents

Les travaux récents dans le domaine de la re-identification de personnes ont enrichi les paradigmes classiques en introduisant des modèles multi-modaux, des architectures adaptées à des contextes extrêmes (par exemple les environnements sportifs ou les variations importantes d'apparence), ainsi que des méthodes exploitant le langage naturel ou l'attention contextuelle. Cette section propose un aperçu structuré de contributions notables publiées entre 2024 et 2025.

FusionSegReID : Segmentation et récupération multi-modale Dans [144], les auteurs présentent FusionSegReID, un modèle combinant des données visuelles et textuelles pour la tâche de re-identification. Le système intègre également un module de segmentation afin de générer des masques précis de la silhouette, supprimant ainsi les zones de fond potentiellement nuisibles. Cette segmentation ciblée permet d'améliorer la robustesse aux occlusions et à la variabilité du contexte visuel. Le modèle repose sur un double encodeur image/texte entraîné avec une fonction de perte contrastive, illustrant les avantages de la fusion vision–langage dans un cadre supervisé.

Skeleton-based ReID pour le changement de tenue Le travail présenté dans [145] propose un paradigme inédit en ReID en exploitant uniquement la dynamique squelettique pour reconnaître un individu. Cette approche est motivée par les limitations des modèles basés sur l'apparence dans des scénarios de changement vestimentaire. Les auteurs montrent que la biomécanique de la marche encode suffisamment d'information pour généraliser à travers différentes tenues. Ce type de représentation est par ailleurs peu sensible aux conditions d'éclairage.

ReID-SAM : Application à des scénarios sportifs extrêmes Le rapport technique associé au challenge SkiTB 2025 présente ReID-SAM, un modèle spécifiquement conçu pour la re-identification de skieurs à partir de vidéos capturées en conditions sportives [146]. Le modèle combine des indices vestimentaires avec des informations posturales

dynamiques pour produire une représentation robuste, adaptée à des contextes caractérisés par des vitesses élevées, des angles de vue inhabituels et des obstructions fréquentes (neige, chevauchements entre athlètes). Ce travail illustre la capacité d'adaptation des principes de la ReID à des domaines d'application éloignés des contextes traditionnels de surveillance urbaine.

Effet des biais caméras sur la ReID Dans [143], les auteurs explorent systématiquement l'impact des caractéristiques propres aux caméras (résolution, couleurs, point de vue, calibration) sur la performance des modèles. Ils montrent que les biais de domaine induits par le matériel de capture peuvent dégrader fortement la robustesse inter-caméras. Ils introduisent des modules d'adaptation spécifiques et des embeddings caméra-aware pour corriger ces déviations.

Backbones efficaces pour ReID vidéo Le travail présenté dans [128] examine l'utilisation des Transformers pour la re-identification vidéo, et remet en question la nécessité de traiter l'ensemble des patchs extraits par les modèles de type ViT. Les auteurs observent qu'une portion significative des patchs visuels contribue peu à la discrimination d'identité, et proposent une stratégie de sélection dynamique visant à ne conserver que les régions informatives. Cette approche permet de réduire la charge computationnelle sans perte notable de performance, ouvrant la voie à des architectures plus légères, adaptées à des systèmes de ReID temps réel.

Mix-Modality ReID : galerie hétérogène visible/infrarouge L'article [147] introduit un protocole de re-identification multi-modal, dans lequel la galerie contient simultanément des images capturées en lumière visible et en imagerie infrarouge. Ce cadre expérimental vise à refléter des conditions d'usage plus réalistes, notamment dans les environnements de surveillance nocturne ou faiblement éclairés, où les capteurs multispectraux sont couramment utilisés. Les auteurs proposent un modèle d'apprentissage d'un espace latent partagé, dans lequel les représentations d'identité sont alignées quelles que soient les modalités d'entrée. Cette stratégie permet d'assurer une correspondance inter-modale tout en maintenant la séparation entre individus.

Vision-Language Models pour la ReID L'étude de [58] examine l'intégration de grands modèles vision–langage (Large Vision-Language Models, LVLMs) dans le cadre de la re-identification de personnes. Ces modèles, pré-entraînés sur de vastes corpus de paires image–texte, permettent de modéliser des relations sémantiques complexes entre modalités. L'alignement multi-modal est utilisé pour enrichir les représentations visuelles par l'intermédiaire d'un encodage textuel des attributs humains. Cette stratégie permet d'améliorer les performances de ReID sans recourir à une supervision explicite sur les at-

tributs, tout en renforçant la généralisation à des contextes variés. *En contraste avec [58], notre contribution n'extrait pas de "tokens" sémantiques via un LVLM ; elle adapte CLIP à un* apprentissage conjoint en un seul stage *(FCE pour les attributs + Triplet pour l'identité) et* structure l'espace *par des* centroïdes textuels *utilisés comme ancres métriques, afin d'optimiser* simultanément *PAR et ReID dans un même espace vision–langage.*

Représentations désentrelacées pour la ReID longue durée L'approche proposée dans [139] vise à séparer explicitement, dans l'espace latent, les composantes liées à l'identité (telles que la morphologie ou la silhouette générale) de celles liées au contexte d'acquisition (comme la pose, l'arrière-plan ou les conditions d'éclairage). Ce désentrelacement permet de renforcer la stabilité des représentations visuelles face aux variations contextuelles, un enjeu central pour la re-identification longue durée dans des séquences vidéo étendues ou à travers des sessions capturées à plusieurs jours d'intervalle.

Synthèse Les travaux récents mettent en évidence une évolution vers des modèles plus flexibles et mieux adaptés aux contraintes opérationnelles des systèmes de re-identification. Cette dynamique se traduit par l'intégration de modalités hétérogènes (visible/infrarouge), la prise en compte de la variabilité vestimentaire, l'optimisation de l'inférence pour des applications temps réel, ainsi que l'exploitation croissante des signaux sémantiques issus du langage. L'ensemble de ces avancées souligne la pertinence d'approches multi-modales et contextuelles, et justifie le choix méthodologique adopté dans ce manuscrit : une modélisation conjointe de l'identité et des attributs dans un cadre unifié fondé sur les modèles vision–langage.

4.2.5 Synthèse critique et limites des approches actuelles

À l'issue de cette revue des approches classiques et contemporaines en *Person Re-Identification*, plusieurs constats peuvent être formulés. Ces éléments mettent en lumière les tendances structurantes du domaine, tout en révélant un certain nombre de limites persistantes sur le plan méthodologique et applicatif. Cette analyse critique sert de point d'appui pour motiver l'approche conjointe proposée dans ce manuscrit, qui vise à surmonter certaines de ces limitations par une modélisation unifiée des identités et des attributs dans un cadre multi-modal.

Évolution des paradigmes L'évolution des approches en re-identification de personnes s'articule autour de trois grandes transitions méthodologiques. La première correspond au passage des descripteurs manuels—tels que les histogrammes de couleur, les motifs texturaux ou les filtres fréquentiels—à l'apprentissage supervisé par réseaux de neurones, qui permet d'extraire automatiquement des représentations plus robustes et

discriminantes. La deuxième transition a vu l'émergence de modèles fondés sur l'apprentissage métrique, tels que les réseaux optimisés par Triplet Loss ou Contrastive Loss, qui remplacent la classification stricte par un apprentissage basé sur la similarité inter-images. Enfin, une troisième évolution majeure réside dans l'intégration de mécanismes d'attention, avec l'essor des Transformers, et dans l'ouverture vers des paradigmes multimodaux, combinant vision et langage. Ces avancées successives ont contribué à améliorer la robustesse, la transférabilité inter-domaine et la capacité à modéliser des scènes complexes.

Forces actuelles Les modèles récents de re-identification atteignent aujourd'hui des performances élevées sur les benchmarks de référence (Market-1501, DukeMTMC-reID, MSMT17). Ces résultats reposent sur plusieurs avancées clés :

- **Encodage global et local** : des architectures comme PCB [123] ou MGN [133] exploitent simultanément des descripteurs globaux et régionaux pour mieux capturer la structure corporelle ;
- **Mécanismes d'attention et Transformers** : des modèles comme TransReID [125] ou PAT [148] améliorent la sélection des zones discriminantes et modélisent les dépendances globales dans l'image ;
- **Apprentissage métrique supervisé** : l'usage de pertes de type triplet ou contrastive permet d'optimiser directement les distances entre identités dans l'espace d'embedding ;
- **Stratégies d'augmentation de données** : des techniques comme le Random Erasing, la génération par GAN, ou le sampling basé sur la difficulté enrichissent la diversité des données vues à l'entraînement.

Limites persistantes Malgré les progrès récents, plusieurs limitations structurelles subsistent dans les approches actuelles de la re-identification. La première concerne la fermeture du paradigme d'entraînement : la plupart des modèles sont développés dans un cadre où les identités d'entraînement sont figées, ce qui limite leur capacité à généraliser à des individus non vus, notamment dans des contextes ouverts ou changeants.

Un autre point concerne la sous-exploitation des attributs sémantiques. Bien que des caractéristiques comme le genre, la couleur des vêtements ou la présence d'accessoires soient informatives, elles sont encore rarement intégrées de manière structurée dans les représentations apprises. Lorsqu'elles sont prises en compte, c'est souvent de manière auxiliaire, sans que leur potentiel discriminant ne soit pleinement exploité.

Par ailleurs, les architectures traditionnelles reposent quasi exclusivement sur le canal visuel, sans tirer parti d'autres modalités potentiellement disponibles, telles que le langage ou le comportement. Cette absence de multimodalité native freine l'adaptation des modèles à des scénarios complexes, comme la surveillance multi-capteurs ou les environ-

nements bruyants.

Enfin, les approches récentes restent sensibles à certaines variations : le changement de tenue, les poses inhabituelles ou les angles de vue extrêmes induisent encore des dégradations notables de performance. De plus, les modèles basés sur des Transformers, bien que performants, restent très coûteux en ressources, ce qui en limite l'utilisation dans des environnements embarqués ou à contraintes temps réel.

Au regard de ces limites, il apparaît nécessaire d'adopter une approche de ré-identification de personnes, à la fois multitâche, multimodale et guidée par les attributs.

Justification de notre approche Face aux limites identifiées dans les travaux existants, il devient pertinent d'envisager une approche de ReID plus flexible, reposant sur des principes multi-tâches, multi-modaux et guidés par les attributs. Notre proposition s'inscrit dans cette direction. Elle s'appuie sur un modèle pré-entraîné de type CLIP afin d'établir un alignement sémantique entre les images et les attributs formulés en langage naturel, ce qui permet une meilleure correspondance entre la perception visuelle et les descriptions humaines. L'apprentissage repose sur une combinaison de deux objectifs complémentaires : une perte Focal Cross-Entropy pour la classification des attributs, et une Triplet Loss pour structurer l'espace métrique des identités. Afin de renforcer la discrimination inter-identités, une tête de mémoire intermédiaire introduisant des centroïdes par classe est intégrée à l'architecture. Enfin, notre cadre d'évaluation est conçu pour rendre compte à la fois de la précision des prédictions attributaires et de la performance en ReID, assurant ainsi une validation complète et cohérente du modèle proposé.

Représentations sémantiques guidées par les attributs Cerberus [23] concatène des vecteurs d'attributs en *identifiants sémantiques* utilisés comme classes cibles et régularisés par une *semantic guidance loss*. Ce dispositif améliore PAR, APS et ReID. À l'inverse, nous n'érigeons pas les attributs en classes : des centroïdes d'identité guident l'espace image–texte de CLIP, évitant l'explosion combinatoire (détails en §4.2.6).

Apprentissage par prompts multiples pour les attributs MP-ReID [24] génère des *prompts* textuels riches pour superviser l'alignement image/texte. Nous retenons l'idée de signaux textuels expressifs, mais préférons des gabarits normalisés (ontologie + patrons) et un guidage par centroïdes dans un schéma *one-stage* (voir §4.2.6).

Variabilité interne des attributs et ReID contextuel Dans [25], les auteurs introduisent le cadre **AGPR (Attribute-Guided Pedestrian Retrieval)** qui aborde spécifiquement la variabilité interne des attributs au sein d'un même individu — une problématique rarement considérée dans les approches traditionnelles. Contrairement aux hypothèses classiques de stabilité attributaire, AGPR modélise les transitions possibles

(par exemple, avec ou sans sac selon les scènes) et ajuste dynamiquement les distances attributaires en fonction du niveau de confiance estimé sur chaque attribut. Ainsi, les dimensions attributaires de l'espace de similarité sont pondérées de manière adaptative, ce qui permet de réduire les pénalités inutiles liées à des variations normales de contexte.

Synthèse des tendances récentes Les approches basées sur les attributs ont évolué du simple signal auxiliaire vers des composants centraux du pipeline de ReID. Trois tendances majeures se dégagent. La première repose sur l'usage d'identifiants sémantiques pour structurer l'espace de représentation, comme le propose Cerberus [23], où des vecteurs d'attributs servent directement de classes cibles. La deuxième tendance met à profit l'alignement vision–langage via des prompts textuels générés automatiquement, à l'image de MP-ReID [24], qui relie descriptions attributaires et embeddings visuels. Enfin, une troisième direction importante vise à modéliser la variabilité intra-identité des attributs, notamment dans des contextes dynamiques ou ambigus, comme l'illustre le cadre AGPR [25].

Ces lignes de travaux convergent vers une ReID sémantiquement enrichie et plus robuste en contexte ouvert, ce qui motive notre choix d'un apprentissage conjoint attributs–identités dans un espace vision–langage structuré par centroïdes.

4.2.6 Méthodes conjointes d'attributs et de Re-Identification

À la suite des travaux exposés dans le chapitre précédent sur la reconnaissance d'attributs de personnes , l'intégration conjointe de cette tâche avec celle de la ReID apparaît comme une évolution naturelle et prometteuse. Ce paradigme vise à tirer parti de la complémentarité entre les deux objectifs : les attributs offrent une supervision sémantique riche et généralisable, tandis que la ReID exige des représentations discriminantes robustes. Plutôt que de reléguer les attributs au rang de simples signaux auxiliaires, les approches récentes tendent à les incorporer au cœur même des architectures d'apprentissage, de manière synergique, afin d'améliorer à la fois la précision des identifications et la robustesse aux variations d'apparence ou de contexte. Plusieurs architectures ont été proposées pour implémenter ce paradigme conjoint. Nous présentons ici quatre contributions représentatives illustrant différentes stratégies de fusion entre identité et attributs.

Modèle multi-tâche APR (2017) Le modèle pionnier APR (*Attribute-Person Recognition*) [149] introduit un apprentissage conjoint de la re-identification et des attributs. L'architecture repose sur un backbone CNN partagé, suivi de deux branches distinctes : une pour la classification d'identité (avec une couche softmax), l'autre pour la prédiction multi-label d'attributs (via activation sigmoïde). Une pondération adaptative est appliquée pour moduler l'impact des attributs selon leur redondance ou complémentarité. Ce

cadre multi-tâche permet de régulariser l'apprentissage des représentations visuelles, en améliorant à la fois la stabilité de l'optimisation et la séparation inter-identités.

Cerberus : Identifiants sémantiques Le modèle **Cerberus** [23] introduit la notion d'*identifiants sémantiques*, construits par concaténation des vecteurs binaires d'attributs associés à chaque individu. Ces identifiants servent de classes cibles dans une tâche de classification, permettant de structurer l'espace d'apprentissage à la fois localement (via les attributs partiels) et globalement (par leur combinaison cohérente). Une *semantic guidance loss* régularise le modèle en rapprochant les représentations de personnes partageant des attributs similaires. Cerberus réalise simultanément les tâches de ReID, PAR et APS, et démontre une nette supériorité sur les benchmarks PETA, RAP et Market-1501-Attribute. En contraste, notre méthode n'encode pas les attributs comme classes (*semantic IDs*) : elle s'appuie sur des centroïdes guidant l'espace image–texte de CLIP et optimise conjointement PAR et ReID sans explosion combinatoire du nombre de classes.

MP-ReID : Prompts multiples auto-générés La méthode **MP-ReID** (*Multi-Prompts ReID*) [24] propose de tirer parti des capacités des modèles de langage pour transformer automatiquement les attributs en descriptions textuelles riches. À partir des annotations attributaires disponibles, le système génère plusieurs prompts du type *"a young man wearing a black hoodie and white sneakers"*, qui sont utilisés comme supervision dans un espace vision–langage commun. L'apprentissage repose sur une perte contrastive inter-modale, favorisant l'alignement entre les représentations visuelles et textuelles. Ce cadre permet une supervision plus flexible et expressive, rendant possible un apprentissage par requête sémantique, sans nécessité de supervision stricte sur l'identité. Cette ligne de travaux nous inspire l'usage de signaux textuels riches ; toutefois, plutôt qu'une ingénierie de *prompts* multiples, nous adoptons une textualisation parcimonieuse et reproductible (gabarits + ontologie) couplée à un guidage par centroïdes, afin d'optimiser simultanément PAR et ReID en un seul stage.

AGPR : Gestion de la variabilité intra-attribut Le cadre **AGPR** (*Attribute-Guided Pedestrian Retrieval*) introduit dans [25] s'attaque à une problématique rarement explorée : la variabilité intra-individuelle des attributs. Dans des scénarios réalistes, un même piéton peut apparaître avec ou sans sac, avec ou sans lunettes, selon le contexte. AGPR propose un mécanisme attentionnel qui évalue dynamiquement la stabilité de chaque attribut et ajuste l'importance accordée à chacun lors du calcul de similarité. L'espace d'embedding est ainsi réorienté vers des proximités sémantiques plutôt que vers des concordances strictes. L'architecture repose sur un module Transformer dédié (ATPR), capable de fusionner efficacement les représentations visuelles et les requêtes attributaires.

Synthèse Les approches conjointes d'apprentissage attribut–identité présentent plusieurs bénéfices notables. En premier lieu, elles permettent d'enrichir la représentation visuelle en la structurant autour de dimensions sémantiques interprétables, ce qui renforce la cohérence interne des embeddings. Elles offrent également une meilleure robustesse face aux variations d'apparence ou de conditions environnementales, souvent problématiques pour les méthodes classiques. Par ailleurs, en intégrant des attributs génériques et partageables, ces modèles deviennent plus efficaces dans des scénarios faiblement supervisés, voire en zéro-shot, notamment dans les cas de recherche guidée par attributs (APS). Enfin, la formulation attributaire ouvre naturellement la voie à des systèmes interactifs, capables d'exploiter des requêtes formulées en langage naturel pour effectuer des recherches plus flexibles et personnalisées.

Limites des méthodes conjointes attributs–ReID Malgré leurs résultats prometteurs, les approches conjointes souffrent encore de limitations significatives. Tout d'abord, la qualité des annotations attributaires reste un facteur critique : dans de nombreux ensembles de données, ces annotations sont souvent partielles, bruitées, voire incohérentes, ce qui fragilise la supervision. Par ailleurs, certains attributs sont sujets à de fortes variations intra-personnelles — un individu peut apparaître avec ou sans sac, lunettes ou chapeau selon les contextes — rendant difficile l'apprentissage de représentations stables sans mécanisme de pondération adaptative. À cela s'ajoute le coût computationnel des architectures récentes, notamment celles reposant sur des stratégies multi-prompts ou des modules multi-têtes, qui alourdissent l'entraînement et ralentissent l'inférence. Enfin, la généralisation inter-domaine reste un défi majeur : les modèles entraînés sur une distribution d'attributs spécifique tendent à perdre en robustesse lorsqu'ils sont confrontés à des environnements réels hétérogènes ou à des jeux de données aux styles visuels variés. Ces constats motivent la nécessité de concevoir une architecture capable d'exploiter les attributs de manière plus flexible, robuste au bruit et sémantiquement informée. C'est précisément dans cette optique que notre modèle s'inscrit, en combinant une fusion vision–langage guidée par prompts, une supervision hybride intégrant à la fois la classification d'attributs et l'apprentissage métrique, ainsi qu'un mécanisme de centroïdes attributaires stabilisant les représentations identitaires.

4.2.7 Intégration de modèles vision–langage dans la ReID

Les modèles vision–langage, tels que CLIP, introduisent une nouvelle manière de lier des représentations visuelles à des descriptions textuelles. En REID, cette capacité d'alignement ouvre des possibilités pour intégrer une supervision sémantique différente des approches classiques fondées sur les identifiants. Ces modèles permettent notamment d'exploiter des requêtes textuelles, de mieux structurer l'espace d'embedding, et d'aborder des

scénarios faiblement annotés ou inter-domaines. Cette section présente des contributions récentes explorant leur usage explicite en ReID.

CLIP-FGDI : Représentations invariantes et prompts guidés Le modèle **CLIP-FGDI** [59] explore l'usage de CLIP dans un cadre ReID. Les auteurs s'appuient sur la capacité des représentations de CLIP à encoder des informations sémantiques transférables. Leur approche repose sur trois composantes principales : un affinage supervisé des prompts textuels, l'apprentissage d'un espace commun incluant des prompts *invariants au domaine* et *représentatifs du domaine*, et une légère adaptation de l'encodeur visuel. Cette stratégie vise à améliorer la robustesse des représentations face aux variations de domaine, sans nécessiter d'adaptation spécifique sur les données cibles.

LVLM-ReID : Génération sémantique et interaction guidée Le modèle **LVLM-ReID** [58] explore l'usage des grands modèles vision–langage (LVLMs) dans le cadre de la ReID. Ces modèles, capables de générer du texte à partir d'une image, sont utilisés pour produire un *token sémantique* décrivant chaque individu. Ce token est ensuite injecté dans un module d'interaction sémantique afin de renforcer l'encodage identitaire. Cette approche démontre que la génération d'attributs implicites (même sans supervision attributaire) permet d'enrichir les représentations pour la ReID. LVLM-ReID offre de bonnes performances sur Market-1501 et DukeMTMC-reID, en particulier dans les configurations à faible supervision.

Benchmark et évaluation des LVLMs en ReID Le travail *Human Re-ID Meets LVLMs* [150] réalise une étude comparative des LVLMs les plus avancés face aux modèles ReID spécialisés. Sur Market-1501, les résultats montrent que les LVLMs généraux comme GPT-4V ou BLIP-2 capturent bien les attributs visibles, mais rencontrent des difficultés sur les images de basse résolution ou à fort désalignement de pose. L'étude met en évidence l'importance d'adapter les LVLMs à la spécificité de la tâche de ReID plutôt que de les utiliser directement en zéro-shot.

VLPSR : ReID en zéro-shot avec prompts optimisés Le cadre **VLPSR** [151], bien qu'orienté objet, présente des éléments transposables à la ReID. Il repose sur l'optimisation de prompts contextuels (CoOp) pour adapter les VLMs à des tâches sans supervision directe. En alignant les images avec des descriptions textuelles via des prompts appris, le modèle améliore la performance de récupération en zéro-shot. Cette approche suggère un potentiel pour la ReID dans des environnements à supervision limitée.

Alignement modal visible/infrarouge via CLIP Le travail **Textual Modality Alignment Learning (TMAL)** [152] explore l'utilisation de CLIP pour réduire l'écart

entre images visibles et infrarouges dans un cadre multi-spectral. En apprenant des jetons textuels spécifiques à chaque modalité, TMAL construit un espace partagé dans lequel les représentations des individus restent cohérentes malgré les différences de capteur. Cette approche illustre comment les VLMs peuvent être adaptés à des configurations de surveillance hétérogènes, comme les conditions jour/nuit.

Considérations sur les biais de caméra L'étude sur les **biais de caméra** [143] met en évidence que les modèles vision–langage peuvent hériter de biais issus de leur pré-entraînement (par exemple, des distributions dominantes dans les images web). Ces biais peuvent interagir avec ceux introduits par les dispositifs de capture (résolution, angle, éclairage), influençant la qualité de l'alignement image–texte. Les auteurs proposent d'intégrer des informations dépendantes de la caméra dans le processus d'apprentissage afin de mieux contextualiser l'optimisation des représentations vision–langage.

Conclusion L'intégration des modèles vision–langage dans la ReID offre des perspectives nouvelles, notamment en facilitant l'alignement sémantique entre images et descriptions textuelles, en améliorant la généralisation inter-domaine grâce à l'utilisation de prompts adaptatifs, et en permettant des scénarios d'apprentissage faiblement supervisés. Néanmoins, ces approches requièrent une attention particulière quant à la conception des prompts, au contrôle des représentations générées, ainsi qu'à la prise en compte des biais liés aux domaines ou aux capteurs. Dans ce contexte, notre méthode propose d'adapter le cadre de CLIP à la ReID via un apprentissage supervisé hybride, combinant pertes attributaires et métriques, tout en intégrant une mémoire structurée par centroïdes dynamiques pour stabiliser les représentations d'identité.

4.2.8 Conclusion : nécessité d'un apprentissage conjoint

Limites des approches vision–langage et justification d'un cadre unifié Bien que les modèles vision–langage offrent une capacité d'alignement sémantique prometteuse pour la ReID, plusieurs limites structurelles persistent dans les approches actuelles :

- **Intégration insuffisante des attributs dans la représentation d'identité** : les attributs sont généralement utilisés comme supervision textuelle externe, sans intervention explicite dans l'optimisation de l'espace d'embedding des identités.
- **Pipeline fragmenté** : des modèles tels que CLIP-FGDI [59] ou LVLM-ReID [58] reposent sur des étapes successives (prétraitement, génération et adaptation de prompts, réentraînement partiel), ce qui complexifie l'apprentissage et introduit des dépendances inter-étapes.
- **Objectifs d'optimisation disjoints** : les fonctions de perte associées à l'alignement image–texte et à la discrimination entre identités sont souvent traitées de

manière indépendante, limitant la cohérence de l'espace latent appris.

— **Dépendance à la formulation des prompts** : la qualité des représentations issues des VLMs est sensible aux formulations textuelles choisies, ce qui peut induire des biais ou une variabilité lors de l'inférence.

Ces limites soulignent la nécessité d'un apprentissage conjoint, dans lequel les attributs et les identités sont modélisés simultanément au sein d'un espace partagé. La section suivante présente une telle approche, combinant supervision attributaire, contrainte métrique et intégration vision–langage dans un cadre unifié.

La méthode proposée repose sur un entraînement unifié en un seul stage, combinant simultanément l'optimisation de la prédiction des attributs à l'aide d'une Focal Cross-Entropy et l'apprentissage d'un espace discriminant pour l'identité à l'aide d'une Triplet Loss. Par ailleurs, l'introduction d'un module de centroïdes attributaires contribue à stabiliser les représentations visuelles, tout en assurant une supervision sémantique explicite. Cette formulation favorise une interaction cohérente entre les descripteurs visuels et sémantiques, et s'intègre naturellement dans le cadre des modèles vision–langage comme CLIP. Le schéma général de notre approche est illustré dans la figure 4.1. Empiriquement, cette formulation *one-stage* se traduit par des améliorations systématiques et robustes des métriques attributaires (mA, F1) et d'identification (Rank-1, mAP) sur plusieurs jeux publics, comme l'attestent les résultats des Tab. 3.2, **??** et 3.3.

4.3 Méthodologie proposée

Dans cette section, nous présentons le cadre méthodologique développé pour répondre aux limites identifiées dans l'état de l'art. L'approche proposée repose sur un apprentissage conjoint des identités et des attributs, structuré autour d'un modèle vision–langage supervisé par des pertes hybrides. Nous commençons par formaliser mathématiquement le problème de ReID multi-attributs.

4.3.1 Formulation du problème de ReID multi-attributs

Soit un ensemble d'images annotées $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{N}$, où :

— x_i désigne une image d'entrée,

— $y_i \in \{1, \ldots, C\}$ correspond à l'identité de la personne, avec C le nombre total d'identités distinctes,

— $a_i \in \{0, 1\}^K$ est le vecteur d'attributs sémantiques binaires associés à l'individu, avec K le nombre total d'attributs considérés.

L'objectif est d'apprendre une fonction d'encodage paramétrée f_θ qui projette chaque image dans un espace de représentation latent, telle que :

1. les représentations $f_\theta(x_i)$ soient discriminantes vis-à-vis des identités y_i, afin de permettre la correspondance fiable des individus ;
2. la géométrie de l'espace latent soit cohérente avec la structure sémantique induite par les attributs a_i, de sorte que des individus partageant des attributs similaires soient rapprochés ;
3. l'apprentissage combine de manière conjointe et coordonnée ces deux objectifs, en intégrant des signaux d'identité et d'attributs au sein d'un cadre d'optimisation unifié.

Pour garantir un apprentissage cohérent entre la structure identitaire et l'information sémantique, nous définissons une fonction de perte composée, combinant plusieurs objectifs complémentaires :

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{triplet}} + \lambda_2 \cdot \mathcal{L}_{\text{FCE}} + \lambda_3 \cdot \mathcal{L}_{\text{centroid}} \tag{4.1}$$

où :

— $\mathcal{L}_{\text{triplet}}$ est une perte triplet visant à maximiser la séparation entre les identités dans l'espace d'embedding ;
— $\mathcal{L}_{\text{FCE}}$ est une Focal Cross-Entropy appliquée à la prédiction multi-label des attributs visuels, utile en cas de déséquilibre ;
— $\mathcal{L}_{\text{centroid}}$ est un terme de régularisation qui pousse les embeddings à se rapprocher dynamiquement de centroïdes attributaires calculés en mémoire ;
— λ_1, λ_2, λ_3 sont des coefficients scalaires contrôlant l'influence relative de chaque composant. Ces coefficients sont optimisés par descente de gradient, ce qui évite un réglage manuel des pondérations et favorise un équilibre adaptatif entre les objectifs identitaires et sémantiques.

4.3.2 Encodage partagé et structuration par centroïdes attributaires

L'approche proposée repose sur un encodeur visuel unique f_θ, chargé de projeter chaque image x_i dans un espace latent partagé $\mathcal{Z} \subset \mathbb{R}^d$, selon une représentation $z = f_\theta(x_i)$. Cette représentation vise à capturer simultanément les caractéristiques discriminantes liées à l'identité de la personne et les attributs sémantiques décrivant son apparence. Contrairement aux méthodes séparant les espaces identitaires et attributaires, cette conception impose une cohabitation contrôlée de ces deux dimensions dans un vecteur latent unique.

Une première contrainte est introduite par une *Triplet Loss*, qui organise l'espace selon les relations d'identité entre les images. Les triplets (x_a, x_p, x_n) sont construits pour rapprocher les représentations d'images issues d'une même personne tout en éloignant

celles provenant d'individus différents. Le principe théorique de cette fonction, ainsi que ses variantes de *hard negative mining*, ont été détaillés dans la section 2.4.1 du chapitre 2 et la section 4.2.3 du présent chapitre.

En parallèle, une supervision attributaire est assurée par une *Focal Cross-Entropy* adaptée aux tâches multi-label déséquilibrées. Cette fonction pondère dynamiquement la contribution des classes rares et fréquentes, garantissant une meilleure robustesse aux déséquilibres inter-attributs (voir section 3.2.1 du chapitre 3).

Enfin, un troisième niveau de structuration repose sur l'introduction de **centroïdes attributaires**, dérivés des descriptions textuelles associées à chaque image. Ces centroïdes, dont le calcul détaillé est présenté en section 4.3.5, agissent comme des ancrages sémantiques stables dans l'espace latent, favorisant l'alignement entre les modalités visuelle et textuelle. Ils renforcent la cohérence des représentations au sein de l'identité tout en préservant la variabilité intra-classe.

Pour chaque image x_i, les centroïdes correspondant aux attributs sémantiques actifs exercent une force d'attraction sur la représentation extraite. Cette contrainte est formalisée par une perte définie au niveau de chaque échantillon :

$$\mathcal{L}_{\text{centroid}}^{(i)} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{a_i^{(k)} = 1\} \, \|f_\theta(x_i) - c_k\|_2^2,$$

où $a_i^{(k)} \in \{0, 1\}$ indique la présence de l'attribut k pour l'image x_i, $f_\theta(x_i)$ désigne l'embedding visuel extrait par le modèle, et $\mathcal{C} = \{c_k\}_{k=1}^{K}$ représente l'ensemble des centroïdes associés aux attributs binaires.

Les centroïdes c_k sont maintenus de manière *dynamique* au cours de l'apprentissage et mis à jour progressivement à partir des embeddings des images présentant l'attribut correspondant, par exemple via une moyenne mobile exponentielle. Cette stratégie permet d'adapter continuellement la position des centroïdes à l'évolution de l'espace de représentation induite par l'optimisation conjointe des différentes fonctions de perte.

La perte globale associée aux centroïdes est obtenue en moyennant ce terme sur l'ensemble des échantillons considérés (ou sur chaque mini-lot) :

$$\mathcal{L}_{\text{centroid}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{centroid}}^{(i)}.$$

Ce terme de régularisation encourage les embeddings à se regrouper autour de régions sémantiques cohérentes, définies par les attributs, tout en préservant la séparation inter-identités imposée par la perte triplet. Il contribue ainsi à structurer l'espace latent de manière sémantiquement informative, sans compromettre la capacité discriminante nécessaire à la ré-identification de personnes.

4.3.3 Construction des triplets guidée par les attributs

Chaque triplet est formé à partir d'une image ancre x_a, pour laquelle une image positive x_p est sélectionnée parmi les exemples appartenant à la même identité (Person IDentity - PID) :

$$\text{PID}(x_p) = \text{PID}(x_a) \quad \text{et} \quad x_p \neq x_a \tag{4.2}$$

Cette sélection permet d'introduire des variations visuelles intra-identité tout en assurant la cohérence du signal supervisé.

L'image négative x_n est choisie parmi les images associées à des identités différentes, selon la condition suivante :

$$\text{PID}(x_n) \neq \text{PID}(x_a) \tag{4.3}$$

Plutôt que d'effectuer un tirage aléatoire, la sélection est guidée par la similarité attributaire : x_n est choisie pour maximiser le recouvrement sémantique avec x_a. Soient A_a et A_i les ensembles d'attributs de x_a et x_i respectivement, alors :

$$x_n = \arg \max_{x_i \in \mathcal{D},\ \text{PID}(x_i) \neq \text{PID}(x_a)} |A_a \cap A_i| \tag{4.4}$$

Ce critère produit des négatifs présentant une apparence globale similaire à l'ancre, mais appartenant à une identité distincte. Lorsqu'aucun exemple ne satisfait un recouvrement attributaire suffisant, le seuil de similarité est progressivement abaissé afin de garantir la formation du triplet.

Cette construction permet d'introduire une variabilité contrôlée dans les exemples négatifs, en tenant compte de la proximité sémantique tout en respectant la contrainte identitaire.

Logique de construction des triplets

Le triplet (x_a, x_p, x_n) est construit de manière à encadrer l'ancre entre :

— une image positive x_p représentant une vue différente du même individu,
— une image négative x_n appartenant à une identité différente, mais présentant une forte similarité d'attributs avec l'ancre.

La formulation contraint le modèle à produire des représentations discriminatives capables de différencier des individus visuellement proches sur la base de détails subtils, au-delà de la simple correspondance globale d'attributs.

4.3.4 Principes sous-jacents

La stratégie de formation des triplets repose sur une structuration contrôlée de l'espace d'embedding, visant à équilibrer cohérence intra-classe et séparation inter-classe. Les images positives, issues d'une même identité mais présentant des variations d'apparence (vue, posture, éclairage), permettent de préserver une certaine tolérance aux changements naturels tout en renforçant la stabilité des représentations. À l'inverse, les images négatives sont sélectionnées de manière à introduire une proximité visuelle ou sémantique avec l'ancre, tout en appartenant à une identité distincte. Ce contraste contribue à contraindre efficacement l'apprentissage, en forçant le modèle à distinguer des individus visuellement proches sur la base de caractéristiques identitaires plus fines. Une telle organisation favorise une meilleure généralisation dans des contextes variés, y compris ceux où les scènes comportent des ambiguïtés visuelles importantes.

4.3.5 Génération des centroïdes attributaires

L'intégration de l'information textuelle attributaire repose sur la construction de centroïdes sémantiques à partir des descriptions textuelles associées à chaque image ancre x_a. Soit $\{t_1, t_2, ..., t_{N_a}\}$ l'ensemble des attributs textuels correspondant à x_a. Chaque attribut t_i est encodé par le modèle texte de CLIP en un vecteur $g_\psi(t_i) \in \mathbb{R}^d$, où d désigne la dimension de l'espace d'embedding textuel.

Le centroïde attributaire $c_a \in \mathbb{R}^d$ est alors défini comme la moyenne des embeddings textuels normalisés :

$$c_a = \frac{1}{N_a} \sum_{i=1}^{N_a} \frac{g_\psi(t_i)}{\|g_\psi(t_i)\|_2},$$

où $\| \cdot \|_2$ désigne la norme euclidienne. Cette normalisation préalable empêche qu'un attribut domine le centroïde en raison d'une magnitude intrinsèque élevée.

Après agrégation, le centroïde est projeté dans l'espace unitaire :

$$c_a \leftarrow \frac{c_a}{\|c_a\|_2},$$

afin d'assurer la compatibilité avec les embeddings visuels $f_\theta(x)$, également normalisés.

Le centroïde c_a représente ainsi une synthèse vectorielle des informations attributaires de l'image dans l'espace latent partagé. Cette représentation est ensuite exploitée dans le calcul des pertes d'appariement image–texte et dans la contrainte de régularisation présentée en section 4.3.2.

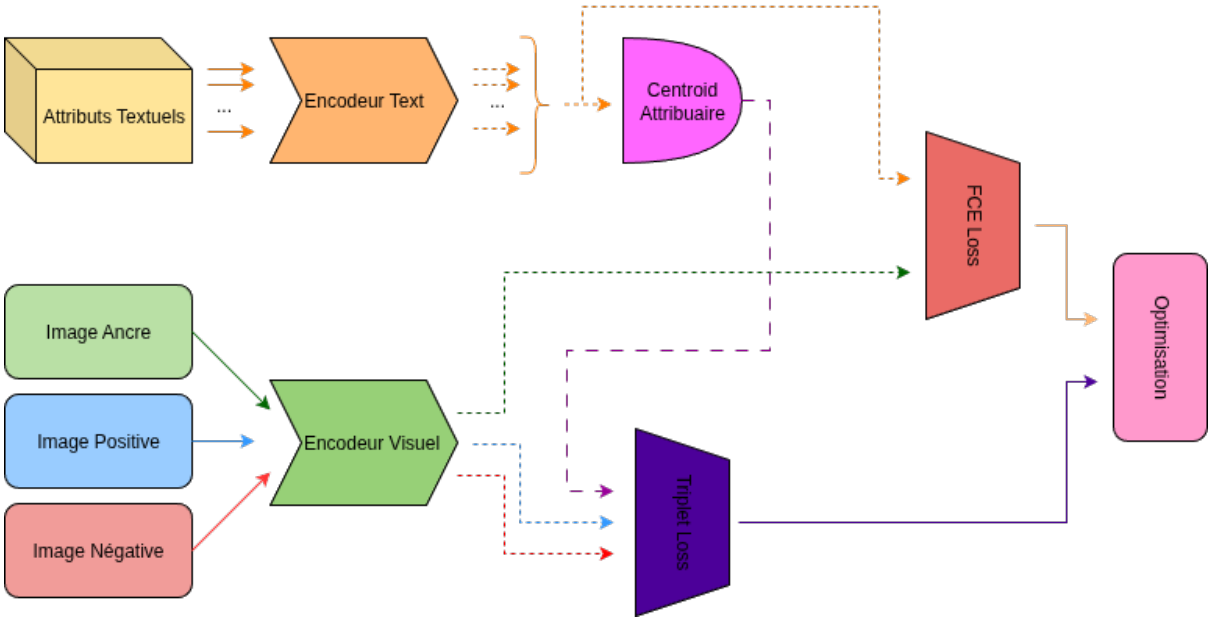


FIGURE 4.1 – Architecture CLIP adaptée pour l'apprentissage conjoint attributs–ReID avec centroïdes et double supervision.

4.3.6 Calcul de la Triplet Loss entre centroïde et images

Afin de contraindre les représentations visuelles et textuelles dans un espace commun, une Triplet Loss est appliquée entre le centroïde attributaire d'une image ancre et ses images positive et négative correspondantes. Cette stratégie vise à exploiter de manière complémentaire l'information sémantique dérivée des attributs et l'information visuelle extraite des images.

Formulation de la Triplet Loss avec centroïdes

Soient c_a le centroïde calculé à partir des attributs textuels associés à l'ancre x_a, $f(x_p)$ l'embedding de l'image positive et $f(x_n)$ l'embedding de l'image négative. La Triplet Loss utilisée est définie par :

$$\mathcal{L}_{\text{triplet}}(c_a, f(x_p), f(x_n)) = \max\left(0, d(c_a, f(x_p)) - d(c_a, f(x_n)) + m\right) \tag{4.5}$$

où $d(\cdot, \cdot)$ est une fonction de distance et m est la marge imposée entre les distances positive et négative.

Le centroïde c_a est conçu comme une représentation sémantique agrégée de l'image ancre, tandis que $f(x_p)$ et $f(x_n)$ correspondent respectivement aux embeddings visuels d'une instance de même identité et d'une instance d'identité différente.

Rôle de la marge m

La marge $m > 0$ introduit une séparation stricte entre les distances positives et négatives. Elle impose que la distance entre le centroïde c_a et l'image négative $f(x_n)$ soit supérieure à la distance entre c_a et l'image positive $f(x_p)$ d'au moins m. Formellement, l'objectif est de satisfaire :

$$d(c_a, f(x_p)) + m < d(c_a, f(x_n)) \tag{4.6}$$

La marge agit comme un paramètre de régularisation, contrôlant la sévérité de la contrainte imposée sur l'espace d'embedding. Un choix approprié de m permet d'éviter l'effondrement des représentations (collapse) tout en garantissant une capacité discriminative élevée.

Choix de la distance cosinus

La fonction de distance $d(\cdot, \cdot)$ adoptée est la distance cosinus définie par :

$$d_{\cos}(x, y) = \frac{1}{2}\left(1 - \frac{x \cdot y}{\|x\|_2 \|y\|_2}\right) \tag{4.7}$$

où $x \cdot y$ est le produit scalaire et $\|\cdot\|_2$ désigne la norme euclidienne. Le facteur $\frac{1}{2}$ permet de normaliser la distance dans l'intervalle $[0, 1]$, rendant son échelle compatible avec les autres métriques utilisées dans le modèle, telles que la distance euclidienne normalisée ou les mesures contrastives exploitées dans la Triplet Loss. Cette formulation assure également une interprétation géométrique claire, puisqu'elle correspond à $\sin^2(\frac{\theta}{2})$, où θ désigne l'angle entre les deux vecteurs.

Ce choix est motivé par plusieurs considérations :

— Les embeddings issus de CLIP sont préalablement normalisés, ce qui rend la distance cosinus particulièrement adaptée.

— La distance cosinus est insensible aux changements d'échelle dans les représentations, ce qui est important dans le cas de comparaisons entre modalité visuelle et textuelle.

— Elle favorise une comparaison sur la directionnalité des vecteurs, ce qui correspond au comportement attendu pour l'appariement sémantique.

Motivations de l'utilisation de centroïdes

L'utilisation du centroïde attributaire c_a comme ancre dans la Triplet Loss vise à améliorer la robustesse du modèle face à la variabilité visuelle au sein d'une même identité. En agrégeant plusieurs attributs textuels liés à l'image ancre, le centroïde capture une représentation sémantique stabilisée qui reste cohérente malgré les variations d'apparence (exemples : posture, éclairage, occlusion partielle).

La supervision croisée entre attributs et images impose donc une structuration de l'espace latent fondée simultanément sur la sémantique textuelle et sur la similarité visuelle, renforçant ainsi la capacité du modèle à généraliser en présence de variations intra-classe importantes.

4.3.7 Avantage d'un apprentissage en une seule étape

Contrairement aux méthodes reposant sur des pipelines en plusieurs étapes — incluant la génération de prompts, des phases d'adaptation spécifiques ou l'ajout de modules externes — la présente approche procède à une optimisation conjointe en un seul stage, selon un schéma d'apprentissage de bout en bout. L'encodeur visuel $f_\theta(x)$ est supervisé simultanément par deux objectifs : la discrimination identitaire via la perte triplet $\mathcal{L}$triplet, et la prédiction attributaire via la Focal Cross-Entropy $\mathcal{L}$FCE.

L'intégration directe de ces deux sources de supervision, sans phase intermédiaire, contribue à une formulation plus homogène du problème. Cette configuration réduit la complexité du modèle, tout en permettant un encodage partagé susceptible de s'adapter à d'autres modalités (telles que le texte, la vidéo ou l'imagerie infrarouge), sans modification substantielle du cadre d'apprentissage.

4.4 Configuration d'entraînement détaillée

Dans cette section, nous détaillons les choix d'optimisation et de configuration adoptés pour entraîner le modèle CLIP modifié dans le cadre du joint training Attributs–ReID.

4.4.1 Hyperparamètres d'entraînement

L'entraînement est effectué sur 100 époques avec une taille de batch fixée à 16 pour les données d'apprentissage. Le taux d'apprentissage initial est défini à 1×10^{-5}, sans phase de préchauffage explicite. Le scheduler utilisé est un **Cosine Annealing** classique, avec une décroissance progressive du taux d'apprentissage vers une valeur minimale de 5×10^{-6}.

Le tableau 4.1 récapitule l'ensemble des hyperparamètres principaux utilisés :

4.4.2 Optimiseur et stratégie de régularisation

L'optimiseur sélectionné est AdamW, qui introduit une correction explicite sur la décroissance du poids (*weight decay*) en dehors de la mise à jour par moment. Cette variante est particulièrement adaptée pour stabiliser l'apprentissage des modèles de type transformer, en limitant la dérive des poids dans les couches de projection.

Paramètre	Valeur
Nombre d'époques	100
Batch size	16
Learning rate initial	1×10^{-5}
Scheduler	CosineAnnealingLR (T_max = 100, $\eta_{min} = 5 \times 10^{-6}$)
Optimiseur	AdamW
Bêtas d'AdamW	(0.9, 0.98)
Poids de décroissance	1×10^{-4}
Epsilon (stabilité numérique)	1×10^{-8}
Clip Grad Norm	1.0
Precision mixte FP16	Utilisée (GradScaler)

TABLE 4.1 – Résumé des hyperparamètres utilisés pour l'entraînement.

La régularisation L2 est appliquée avec un coefficient fixé à 10^{-4}, ce qui permet de limiter les risques de sur-ajustement compte tenu de la taille modérée du dataset UPAR utilisé.

4.4.3 Gestion de la précision mixte et du GPU

Afin d'optimiser la vitesse d'apprentissage et de réduire l'empreinte mémoire, le modèle est entraîné en précision mixte FP16 à l'aide du module `torch.cuda.amp`. Le mécanisme de `GradScaler` est activé pour stabiliser les mises à jour de gradients en faible précision.

Toutes les opérations d'entraînement sont effectuées sur GPU (dispositif CUDA détecté automatiquement).

De plus, afin d'éviter les explosions de gradient, une opération de `clip_grad_norm_` est appliquée à hauteur de 1.0 sur tous les paramètres du modèle avant chaque mise à jour.

4.5 Jeux de données utilisés

L'évaluation de notre approche conjointe pour la Person Re-Identification (ReID) et la reconnaissance d'attributs (PAR) repose sur des ensembles de données publics largement utilisés dans la littérature. Nous sélectionnons trois jeux de données représentatifs : **PETA**, **UPAR** et **RAPv2**, chacun offrant des annotations d'identité et d'attributs, et permettant l'évaluation dans un cadre multi-tâches réaliste.

4.5.1 Description des jeux de données

PETA (PEdesTrian Attribute) Le dataset **PETA** [12] contient environ 19 000 images issues de 10 bases de vidéosurveillance urbaines différentes. Chaque image est annotée

avec jusqu'à 61 attributs (certains binaires et d'autres multiclasses), dont 35 sont généralement sélectionnés pour les expériences. Bien que PETA n'ait pas été conçu pour la ReID, chaque image est associée à une identité unique. Sa diversité inter-domaine, sa densité d'attributs et son déséquilibre de classes en font un excellent benchmark pour la reconnaissance multi-attributs et l'apprentissage robuste.

UPAR (Unified Pedestrian Attribute and ReID) **UPAR** [15] est un dataset de nouvelle génération conçu pour unifier les tâches de ReID et de reconnaissance d'attributs dans un cadre cohérent. Il contient plus de 130 000 images issues de plusieurs sous-datasets (Market-1501, DukeMTMC, MSMT17, PETA, RAP, etc.), avec 46 attributs harmonisés et plus de 21 000 identités uniques. UPAR est particulièrement adapté à l'apprentissage multi-tâches car il :

— couvre une grande diversité d'environnements et de caméras,
— contient des identités annotées de manière cohérente à travers les attributs,
— permet une évaluation conjointe avec des splits définis pour chaque sous-tâche.

RAPv2 (Richly Annotated Pedestrian) Le dataset **RAPv2** [13] est une extension du RAP original. Il comprend plus de 84 000 images provenant de 26 caméras intérieures simulant des conditions de surveillance réalistes. Les images sont annotées avec 72 attributs incluant des informations liées à la pose, à la vue (angle), et au contexte d'apparition (indoor/outdoor). Ce dataset est intéressant pour sa granularité annotative et sa complexité contextuelle (variations de pose, occlusions, angles extrêmes).

4.5.2 Protocole expérimental

Pour garantir la comparabilité et l'équité des évaluations, nous adoptons un protocole expérimental unifié, compatible avec les tâches de classification d'attributs et d'identification.

— **Splits** : Pour chaque dataset, nous utilisons les partitions officielles : par exemple, PETA (9 500 train / 1 900 val / 7 600 test), RAPv2 (33 268 train / 8 000 test), UPAR suit des splits personnalisés par sous-dataset, alignés avec les benchmarks d'origine.
— **ReID** : La performance est évaluée selon les métriques standard : **CMC** (Cumulative Matching Characteristic) à top-1, top-5 et top-10, et **mAP** (mean Average Precision).
— **Attributs** : Nous reportons les métriques classiques : précision moyenne (mA), précision, rappel, F1-score, sur l'ensemble des attributs binaires.
— **Équilibre inter-tâches** : Toutes les métriques sont mesurées simultanément pour analyser l'effet d'un apprentissage conjoint sur les performances de chaque tâche.

4.5.3 Justification des choix dans un cadre multi-tâches

Ces jeux de données ont été choisis pour leur complémentarité et leur richesse dans un cadre multi-tâches :

— **PETA** est un benchmark robuste pour la reconnaissance d'attributs, et permet d'évaluer la sensibilité des modèles à la rareté et au déséquilibre inter-classes.

— **RAPv2** introduit une complexité contextuelle essentielle pour tester la robustesse face à la variabilité de scène, de pose et de caméra.

— **UPAR**, en tant que base unifiée, est le plus représentatif des besoins d'un apprentissage multi-objectif moderne. Il permet d'évaluer simultanément les tâches de PAR, ReID et APS (Attribute-based Person Search).

Ces datasets permettent ainsi de valider notre proposition dans des conditions variées, avec des scénarios de transfert, de supervision faible et de diversité attributaire importante — conditions représentatives d'un déploiement réel dans des systèmes de surveillance intelligents.

4.6 Résultats expérimentaux et analyses approfondies

4.6.1 Résultats quantitatifs

Nous présentons ici les performances quantitatives du modèle proposé sur plusieurs jeux de données standards. Les résultats sont évalués en utilisant les métriques classiques de la littérature en Person Re-Identification, à savoir le mean Average Precision (mAP) et les taux de reconnaissance cumulés (CMC).

Dans un premier temps, nous comparons notre approche aux méthodes de Re-Identification pure, c'est-à-dire optimisées exclusivement pour la discrimination d'identité sans supervision attributaire explicite. Ces méthodes reposent principalement sur des architectures convolutionnelles ou attentionnelles, incluant des variantes à base de parties corporelles (PCB, Part-Aligned), de mécanismes d'attention spatiale (PAT, RGA-SC) ou encore des approches récentes de type CLIP-ReID et Cerberus. Les résultats présentés dans le Tableau 4.2 montrent que notre modèle atteint des performances compétitives, surpassant la plupart des architectures existantes sur les jeux de données Market1501 et DukeMTMC-ReID.

Afin d'évaluer plus spécifiquement la contribution de la supervision attributaire dans notre cadre d'apprentissage conjoint, nous confrontons ensuite notre méthode aux approches de Person Attribute Recognition (PAR) appliquées à la Re-Identification. Ces modèles exploitent explicitement les annotations sémantiques (vêtements, accessoires, posture, etc.) pour renforcer la représentation d'identité, à l'instar de A^3M, CA^3Net, AttKGCN ou GPS. Comme le montre le Tableau 4.3, notre modèle surpasse ces méthodes basées attributs sur l'ensemble des métriques et des jeux de données, démontrant

Méthode	Market1501 mAP	Market1501 R@1	Duke mAP	Duke R@1
PCB [123]	77.4	92.3	66.1	81.7
Part-Aligned [153]	79.6	91.2	69.3	84.4
DG-Net [154]	86.0	94.8	74.8	86.6
PAT [148]	88.3	95.4	79.9	88.9
RGA-SC [155]	88.4	96.1	77.4	81.1
CLIP-ReID [20]	89.6	95.5	82.5	90.0
Notre méthode	**88.9**	**96.5**	**81.2**	**91.8**

TABLE 4.2 – Comparaison des performances des méthodes de Re-Identification (ReID) pure sur Market1501 et DukeMTMC-ReID. Les résultats sont en %.

la pertinence du couplage attribut–identité dans un espace latent unifié.

Méthode	Market1501		DukeMTMC-ReID	
	mAP	R-1	mAP	R-1
UPAR [92]	40.6	55.2	40.6	55.2
ARN [10]	62.6	83.6	50.2	72.6
A^3M [156]	66.8	87.8	52.0	73.9
UF [157]	70.1	87.1	52.8	74.0
UCAD [158]	79.5	92.6	–	–
CA^3Net [159]	79.6	92.1	–	–
APDR [160]	80.1	93.2	56.3	75.8
AANet [161]	82.5	93.6	59.6	78.1
AttKGCN [162]	85.5	94.4	59.4	77.4
GPS [163]	87.8	95.2	78.7	88.2
Cerberus [23]	89.8	96.1	80.7	91.1
Notre méthode	**90.1**	**96.3**	**84.2**	**93.9**

TABLE 4.3 – Comparaison des performances mAP / Rank-1 pour les méthodes basées attributs sur Market1501 et DukeMTMC-reID.

Afin d'évaluer la capacité du modèle à capturer simultanément les informations attributaires et identitaires, nous comparons notre approche à plusieurs méthodes de Person Attribute Recognition (PAR), incluant APR S-CNN et AGPR, sur le jeu de données PETA. Contrairement aux approches classiques centrées sur la seule reconnaissance d'attributs, notre modèle réalise un apprentissage conjoint qui conserve une forte cohérence entre les représentations visuelles et sémantiques. Les résultats du tableau 4.4 montrent que notre approche atteint un mAP et un Rank-1 comparables à ceux d'APR, tout en restant conceptuellement plus simple et mieux généralisable à des contextes multi-domaines.

Méthode	**Référence**	Rank-1	Rank-5	Rank-10	mAP
APR	CVPR 2017	87.04	**95.10**	**96.42**	**66.89**
APR (w/o ARM)	-	85.71	94.32	96.46	66.59
APR (w/o attri)	-	81.93	92.28	94.58	58.74
Baseline-2 (Res50)	-	49.76	70.07	77.67	23.95
S-CNN	ECCV 2016	65.88	-	-	39.55
AGPR	TOMM 2022	71.3	62.0	84.6	81.0
Notre méthode		**88.02**	94.8	95.72	66.64

TABLE 4.4 – Comparaison des performances sur la tâche de PAR évaluée sur le jeu de données PETA. Les méthodes APR, S-CNN et AGPR reposent sur un apprentissage guidé par les attributs, tandis que notre approche unifiée atteint des performances comparables, tout en intégrant la supervision attributaire et identitaire dans un même espace latent.

4.6.2 Recherche d'images par image témoin

Nous avons également évalué les performances de notre modèle dans un scénario de recherche d'images par similarité visuelle, dans lequel une image témoin est utilisée comme requête pour retrouver des images correspondantes dans une galerie. Ce protocole, conceptuellement proche de la Re-Identification classique, est ici interprété comme une tâche plus générale de recherche visuelle indépendante de l'identité.

Méthode	**Market-1501**		**DukeMTMC-reID**	
	mAP	Rank-1	mAP	Rank-1
AAIPR [164]	20.7	40.3	15.7	46.6
AIHM [165]	24.3	43.3	17.4	50.5
SAL [166]	29.8	49.0	–	–
ASMR [110]	31.0	49.6	–	–
UPAR [92]	32.3	45.0	–	–
Cerberus [23]	31.7	49.3	23.0	53.5
Notre méthode	**34.4**	**53.2**	**25.2**	**56.7**

TABLE 4.5 – Comparaison des performances pour la tâche de recherche par image témoin (*image-to-image retrieval*) sur Market-1501 et DukeMTMC-reID. Les scores sont exprimés en mAP (%) et Rank-1 (%). Les valeurs absentes correspondent à des expériences non rapportées dans les publications originales.

Le Tableau 4.5 met en évidence une amélioration systématique apportée par notre approche sur les deux jeux de données. Sur *Market-1501*, le modèle atteint un mAP de 34.4 % et un Rank-1 de 53.2 %, soit un gain absolu d'environ 3 points de mAP et 4 points

de Rank-1 par rapport à la meilleure méthode concurrente, Cerberus. Sur *DukeMTMC-reID*, les écarts sont similaires, avec une amélioration de +2.2 mAP et +3.2 Rank-1.

Ces résultats montrent que les représentations apprises conservent une cohérence suffisante pour permettre la recherche d'images même lorsqu'aucune identité explicite n'est supervisée. L'organisation conjointe de l'espace latent, contrainte à la fois par les pertes métriques et attributaires, facilite la comparaison inter-images et renforce la robustesse des correspondances visuelles. Cela illustre la capacité du modèle à généraliser au-delà du cadre strict de la Re-Identification, ouvrant la voie à des scénarios de recherche multimodale où la requête peut être indifféremment textuelle ou visuelle.

4.6.3 Études d'ablation

Nous analysons ici l'impact de chacune des composantes principales du modèle proposé sur les performances globales. Le Tableau 4.6 présente les résultats obtenus lorsque certains modules sont activés ou désactivés : l'encodeur visuel basé sur CLIP, la fonction de classification Focal Cross-Entropy (FCE) pour les attributs, la perte triplet appliquée aux identités, et enfin le mécanisme d'alignement sémantique par centroïdes textuels.

Configuration	**Market1501**		**DukeMTMC-ReID**	
	mAP	R-1	mAP	R-1
CLIP seul (backbone)	48.8	50.1	37.7	46.5
+ FCE (classification attributs)	55.5	56.5	42.2	51.8
+ Triplet (supervision identité)	67.8	70.2	55.0	59.4
+ Centroïdes (alignement sémantique)	**90.1**	**96.3**	**84.2**	**93.9**

TABLE 4.6 – Étude d'ablation des composants de notre méthode sur Market1501 et DukeMTMC-ReID.

Les résultats confirment la contribution progressive de chaque composant. L'encodeur visuel CLIP, utilisé seul, fournit des représentations génériques mais insuffisamment discriminantes pour la Re-ID, en raison de son entraînement initial sur des associations image–texte générales. L'ajout de la Focal Cross-Entropy introduit une supervision explicite sur les attributs, renforçant la sélectivité des représentations visuelles. La Triplet Loss améliore ensuite la structuration métrique de l'espace latent, favorisant la séparation entre identités et la cohérence intra-classe.

L'effet du module de centroïdes est particulièrement marqué : il induit un bond de plus de 20 points de mAP et près de 25 points de Rank-1 sur Market1501. Cette amélioration s'explique par le rôle du mécanisme d'alignement sémantique, qui ancre les représentations visuelles dans des régions stables de l'espace latent associées aux descriptions textuelles. En contraignant les embeddings d'images à se regrouper autour de centroïdes textuels cohérents, le modèle réduit la variance intra-identité tout en préservant la diversité inter-

attributs. Autrement dit, les centroïdes agissent comme une *régularisation structurelle multimodale*, stabilisant la géométrie de l'espace et favorisant un apprentissage conjoint plus homogène entre identité et attributs.

Cette interaction entre supervision métrique, supervision attributaire et alignement sémantique constitue l'un des points clés du gain observé, démontrant que la fusion contrôlée des modalités visuelle et textuelle peut significativement améliorer la discriminabilité dans la Re-ID.

4.6.4 Analyse de sensibilité aux hyperparamètres

Une évaluation de la sensibilité aux hyperparamètres a également été menée. La figure 4.2 présente les effets de variations du coefficient γ de la FCE, de la marge α de la triplet loss, et du facteur de pondération λ de la FCE. Les résultats indiquent que les performances ne varient que modérément dans une plage de valeurs autour des réglages choisis. Toutefois, certains extrêmes peuvent entraîner une baisse de performance, par exemple une marge trop faible réduisant la séparation entre classes, ou un γ élevé accentuant la pondération des échantillons rares au détriment des autres.

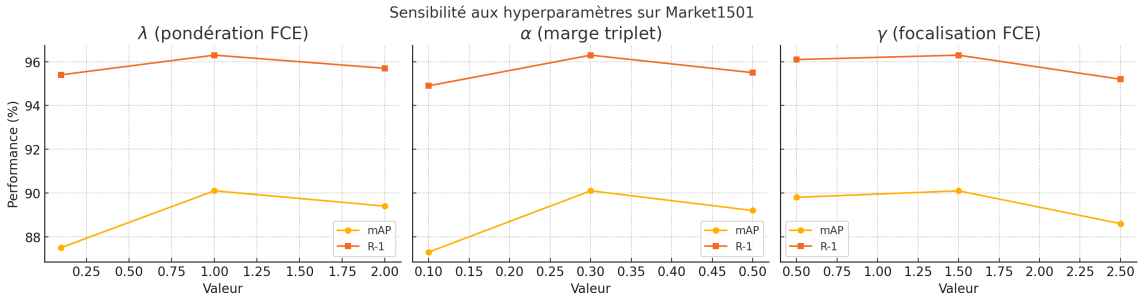


FIGURE 4.2 – Impact des hyperparamètres sur les performances (mAP et Rank-1) sur le dataset Market1501. Trois paramètres sont analysés : λ (pondération de la perte Focal Cross-Entropy), α (marge de la triplet loss), et γ (paramètre de focalisation de la FCE).

Analyse de sensibilité. La Figure 4.2 présente une étude de sensibilité aux principaux hyperparamètres influençant la performance du modèle. Les résultats montrent que la performance varie modérément en fonction de la valeur choisie pour chacun des paramètres. Pour λ, la performance atteint un maximum autour de 1.0, suggérant un équilibre entre la pondération des attributs et les autres composantes de la perte. Concernant α, une marge trop faible ou trop élevée semble nuire à la capacité discriminative de la triplet loss, la valeur de 0.3 produisant les meilleurs résultats. Enfin, le paramètre γ régule la focalisation sur les exemples difficiles dans la Focal Cross-Entropy ; une valeur intermédiaire de 1.5 apparaît comme un bon compromis. Ces résultats suggèrent une certaine robustesse du modèle autour de valeurs centrales pour chacun de ces hyperparamètres.

L'ensemble de ces résultats souligne l'utilité de la combinaison des différents signaux de supervision pour structurer l'espace latent selon des critères à la fois visuels et sémantiques.

4.6.5 Analyse qualitative

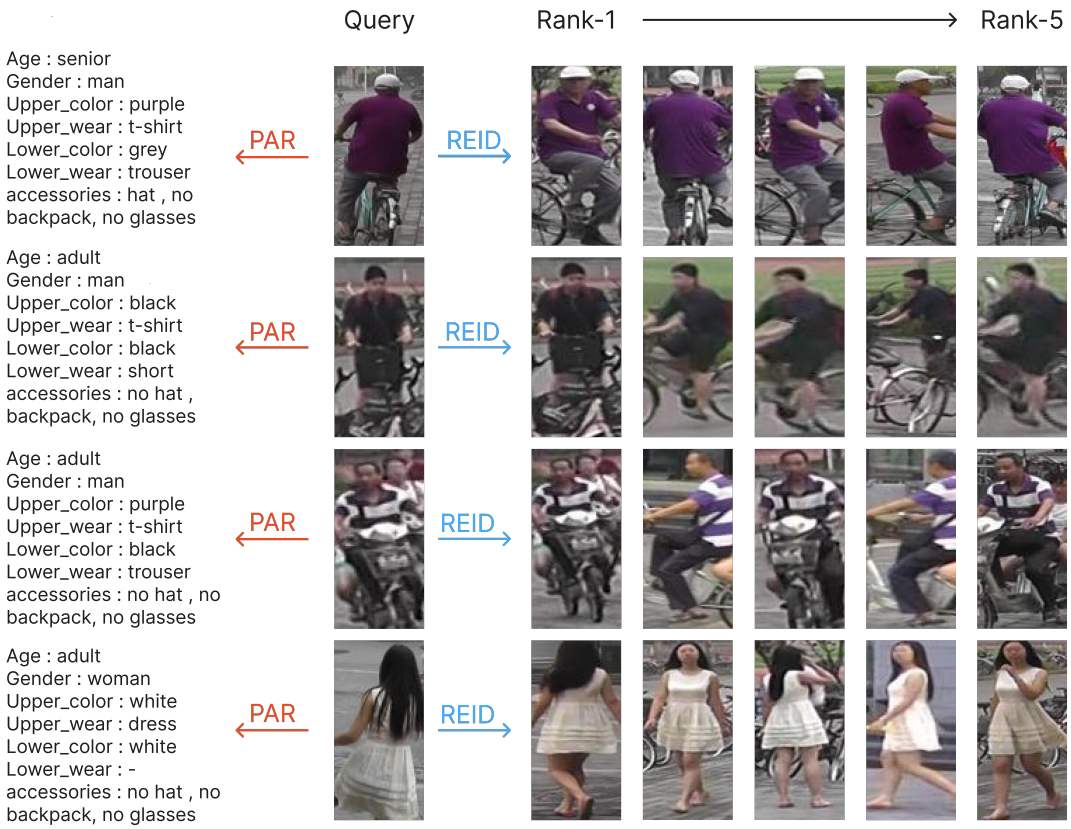


FIGURE 4.3 – Résultats qualitatifs sur le dataset Market1501. Chaque ligne montre une image requête (colonne "Query") accompagnée des cinq images les plus similaires selon deux branches du modèle : PAR et ReID.

La Figure 4.3 illustre plusieurs cas dans lesquels les représentations issues de la reconnaissance d'attributs (PAR) et celles issues de la ré-identification (ReID) permettent d'extraire les voisins les plus proches d'une image requête selon différentes modalités. On observe que la branche PAR a tendance à regrouper les individus partageant des attributs visuels similaires (âge, genre, couleurs et types de vêtements), ce qui peut renforcer la cohérence sémantique des résultats. En revanche, la branche ReID privilégie une correspondance plus fine d'apparence globale, incluant la posture, l'orientation ou des détails non codés explicitement dans les attributs. Cette complémentarité suggère que les deux types d'indices – attributaires et visuels globaux – capturent des informations distinctes mais potentiellement synergiques pour la tâche de ré-identification.

4.6.6 Robustesse hors-distribution

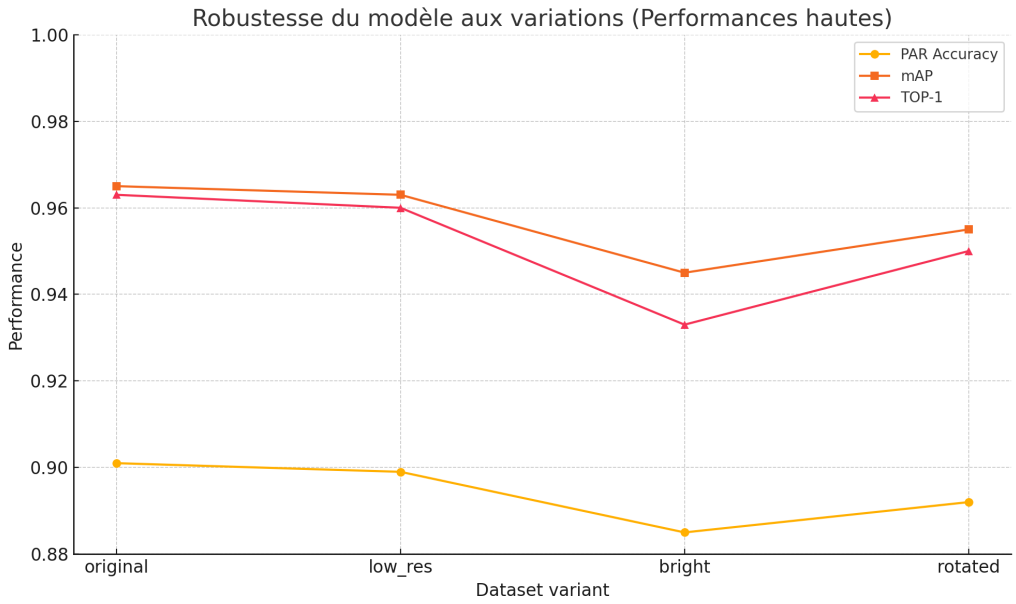


FIGURE 4.4 – Évolution des performances du modèle selon différentes variantes des données (résolution, luminosité, orientation).

Les résultats présentés en figure 4.4 montrent les variations des performances du modèle sur plusieurs versions altérées du jeu de test : réduction de la résolution, augmentation de la luminosité, et rotation des images. Trois métriques sont reportées : la précision des attributs (PAR Accuracy), le mAP pour la tâche de ré-identification, et le taux de reconnaissance en rang 1 (TOP-1). Une baisse est observée sur les données fortement éclairées, tandis que les performances restent comparables dans les autres cas.

4.7 Discussion approfondie des résultats

Les résultats précédents montrent que la combinaison d'une supervision identitaire (Triplet Loss) et attributaire (Focal Cross-Entropy) dans un espace latent partagé constitue un levier essentiel pour améliorer à la fois la robustesse et la généralisation. Contrairement aux approches séquentielles qui dissocient les espaces d'identité et d'attributs, le modèle proposé apprend une représentation unifiée dans laquelle les deux dimensions interagissent de manière explicite. Cette cohabitation des contraintes favorise une meilleure régularisation de l'espace : les attributs stabilisent les représentations intra-identité, tandis que la supervision métrique préserve la séparation inter-identités.

Le rôle des centroïdes textuels apparaît déterminant dans cet équilibre. En introduisant des points d'ancrage sémantiques calculés à partir des descriptions textuelles, ils imposent une géométrie partagée entre modalités, qui oriente la distribution des embeddings visuels vers des régions sémantiquement cohérentes. Ce mécanisme permet d'assurer

la compatibilité entre la vision et le langage sans sacrifier la discrimination visuelle : les performances obtenues sur la tâche de recherche image–à–image (Tab. 4.5) en témoignent. Les centroïdes agissent en pratique comme des *repères conceptuels* qui organisent la topologie de l'espace latent, réduisant la variance intra-classe et renforçant la consistance des voisinages locaux.

Les analyses qualitatives (Fig. 4.3) confirment cet effet structurant : les clusters issus de la branche attributaire révèlent une organisation sémantique fine (par couleur, vêtement, posture), tandis que ceux de la branche ReID demeurent centrés sur la similarité d'apparence globale. La complémentarité entre ces deux perspectives illustre la principale force de l'approche proposée : la capacité à concilier une logique de discrimination et une logique d'interprétation, dans un cadre d'apprentissage véritablement multimodal.

4.7.1 Limitations identifiées

Plusieurs points de vigilance apparaissent à l'analyse conjointe des figures et tableaux.

1. **Coûts mémoire et calcul.** L'usage d'un backbone de type ViT, combiné à un apprentissage métrique nécessitant un *mining* informatif, induit une pression mémoire notable. La gestion des centroïdes textuels ajoute un passage dans l'encodeur texte et une banque de prototypes à maintenir, ce qui complexifie l'entraînement à grande échelle.

2. **Sensibilité hyperparamétrique.** La figure de sensibilité (Fig. 4.2) met en évidence des zones de stabilité autour de $\lambda \approx 1.0$, $\alpha \approx 0.3$ et $\gamma \approx 1.5$, avec une dégradation lorsque ces valeurs s'éloignent. Cela plaide pour une calibration a priori et un suivi régulier en validation.

3. **Attributs peu fréquents et bruit.** Malgré la FCE, les attributs *ultra-rares* restent sensibles au déséquilibre et au bruit d'annotation. L'impact se manifeste par une variabilité accrue des métriques attributaires selon les classes.

4. **Robustesse photométrique.** La figure de robustesse (Fig. 4.4) indique une tenue correcte aux altérations de résolution et d'orientation, avec une fragilité plus marquée en forte surbrillance. L'alignement intermodal ne compense pas entièrement ces effets photométriques.

5. **Dépendance aux formulations et aux capteurs.** Les centroïdes dérivés du texte et les prompts peuvent capter des biais liés aux caméras ou aux domaines. En l'absence de mécanismes *camera-aware*, des décalages systématiques entre sous-domaines peuvent apparaître.

4.7.2 Perspectives méthodologiques

Plusieurs pistes apparaissent exploitables à court terme, en continuité avec les constats précédents.

1. **Centroïdes dynamiques et hiérarchiques.** Remplacer les moyennes fixes par des prototypes apprenables mis à jour par moyenne exponentielle (EMA), organisés hiérarchiquement (attributs, combinaisons d'attributs, identité). Ajouter une projection *camera-aware* pour découpler les variations induites par les capteurs tout en imposant une contrainte d'invariance inter-caméras.
2. **Distillation et efficience.** Mettre en place une distillation croisée vision–langage pour transférer à un étudiant compact des cibles souples (logits attributaires, similarités texte–image, voisinages métriques). Compléter par une sélection parcimonieuse de patchs côté ViT, une quantification post-entraînement et des adaptateurs légers afin de réduire mémoire et temps d'inférence sans modifier le protocole d'évaluation.
3. **Prompts apprenables et calibration.** Introduire des prompts paramétriques conditionnés par métadonnées (caméra, scène) et par la confiance attributaire. Calibrer les logits (température, *logit adjustment* selon la fréquence) et pondérer l'influence des attributs dans les contraintes métriques en fonction de l'incertitude estimée par le classifieur attributaire.
4. **Recettes pour attributs ultra-rares.** Combiner pondération par *effective number*, marges adaptatives dans les pertes métriques et *distance-weighted sampling* pour stabiliser le *mining* lorsque la distribution est très déséquilibrée. Documenter l'impact par attribut dans un protocole de validation dédié.
5. **Évaluation étendue.** Étendre les tests inter-domaines (visible/IR, indoor/outdoor, nuit/jour), intégrer des scénarios de changement de tenue et des protocoles *continual/open-set*, et analyser les coûts d'entraînement et d'inférence sous contraintes mémoire.

4.7.3 Synthèse

Les résultats suggèrent que la co-optimisation des objectifs attributaires et métriques, associée à un ancrage sémantique via centroïdes, contribue à structurer l'espace latent pour des usages hétérogènes (ReID, retrieval image–à–image, requêtes textuelles). Les limites observées concernent surtout l'efficience, la sensibilité aux déséquilibres extrêmes et la robustesse photométrique. Des centroïdes dynamiques adaptés aux caméras, la distillation préservant la structure des voisinages et des prompts apprenables calibrés fournissent des leviers concrets pour améliorer la tenue hors-distribution et le coût opérationnel, tout en préservant l'alignement intermodal mis en évidence par l'ablation et les analyses qualitatives.

4.8 Conclusion du chapitre

Ce chapitre a présenté un cadre d'apprentissage conjoint pour la ReID et la reconnaissance d'attributs, fondé sur un modèle vision–langage de type CLIP. La combinaison d'une Focal Cross-Entropy pour la supervision attributaire, d'une Triplet Loss pour la structuration métrique et d'un guidage sémantique par centroïdes textuels a permis d'aligner de manière cohérente les espaces visuel et textuel. Les études d'ablation (Tab. 4.6) ont montré l'apport spécifique de chaque composante et leur effet cumulatif, tandis que les résultats quantitatifs sur plusieurs jeux de données (Tab. 4.2, Tab. 4.3, Tab. 4.4) et l'analyse image–à–image (Tab. 4.5) ont indiqué que l'alignement intermodal se fait sans dégrader la discrimination visuelle. Les analyses de sensibilité et de robustesse (Fig. 4.2, Fig. 4.4) ont mis en évidence des zones de stabilité et des limites connues, notamment en forte surbrillance, qui orientent des optimisations ciblées.

Dans une perspective applicative, le cadre étudié présente un intérêt pour des systèmes de surveillance intelligents. L'unification des représentations facilite des scénarios d'usage hétérogènes : recherche guidée par attributs formulés en langage naturel, appariement image–image entre caméras, interrogation rapide de galeries larges. La présence d'attributs rend l'inspection des résultats plus interprétable, ce qui peut accélérer l'expertise humaine. Les contraintes opérationnelles demeurent néanmoins structurantes : empreinte mémoire et coût d'inférence liés aux backbones de type ViT et à la banque de centroïdes, hétérogénéité des capteurs et des conditions photométriques, déséquilibres de fréquence des attributs. Les pistes discutées dans la section précédente (distillation, prompts apprenables et calibration, gestion dynamique des centroïdes, sélection parcimonieuse de patchs) s'inscrivent dans une logique d'industrialisation, avec pour objectif de contenir les coûts tout en préservant la robustesse inter-domaines.

Plusieurs prolongements scientifiques se dessinent. Sur le plan multi-modal, l'extension de l'alignement au-delà du visible (infrarouge, profondeur, flux vidéo) et l'intégration de signaux contextuels (métadonnées caméra, temps, scène) offrent des marges d'amélioration pour la tenue hors-distribution. En apprentissage fédéré, l'adaptation des centroïdes et des prompts à des nœuds hétérogènes, couplée à des mécanismes de régularisation inter-sites, permettrait d'explorer des protocoles respectueux de la confidentialité tout en maintenant une cohérence globale de l'espace latent. En auto-supervision, le pré-entraînement par objectifs contrastifs ou masqués, suivi d'un affinement conjoint attribut–identité, pourrait réduire la dépendance aux annotations rares et améliorer la transférabilité. Ces directions, combinées à une évaluation élargie aux changements de tenue et aux protocoles *open-set*/*continual*, constituent un cadre méthodologique pour la suite, où l'équilibre entre efficacité, interprétabilité et robustesse demeure central.

5

Conclusion générale et perspectives

Récapitulatif des contributions

Ce travail s'inscrit dans l'analyse de personnes en vision par ordinateur, avec un accent sur la reconnaissance d'attributs (Person Attribute Recognition, PAR) et la ré-identification d'individus (Re-Identification, ReID) en contexte de vidéoprotection. L'objectif a été de proposer un cadre d'apprentissage unifié exploitant conjointement des signaux visuels et sémantiques au moyen d'un modèle vision–langage pré-entraîné. Le schéma général de la démarche est présenté en Fig.4.1, et les jeux utilisés sont récapitulés en *Tab.* 3.1.

Modélisation conjointe des attributs et de l'identité

Nous adaptons CLIP [57] à un entraînement supervisé *en un seul stage*, combinant une Focal Cross-Entropy pour la PAR et une Triplet Loss pour la ReID. Les embeddings image et texte sont appris dans un espace partagé de sorte que les attributs contribuent à structurer les identités, et réciproquement. Pour limiter les ambiguïtés en multi-label, le passage label $\rightarrow$ texte est normalisé (gabarits + ontologie), et les lots sont construits avec un appariement *univoque* des paires image–attribut.

Centroïdes attributaires pour la structuration métrique

Nous introduisons des *centroïdes sémantiques* calculés comme moyennes normalisées d'embeddings textuels d'attributs. Ces centroïdes servent de pivots dans l'objectif métrique et agissent comme régularisation sémantique complémentaire à la Triplet Loss, en rapprochant les vues d'un même individu et en éloignant des individus visuellement proches mais distincts. La formulation et son intégration au pipeline sont détaillées en §4.3. Leur effet est mis en évidence par l'ablation (Tab.4.6), qui montre une amélioration nette du mAP et du Rank-1 sur **Market1501** et **DukeMTMC-ReID**.

Standardisation inter-jeux et protocole d'évaluation

Une procédure de mise en correspondance label → texte (détaillée en §3.3.1) est définie pour harmoniser les attributs hétérogènes issus de PETA, PA100K, RAPv2, Market1501 et UPAR2024. Le protocole retient mA/F1 pour la PAR et Rank-1/mAP pour la ReID et la recherche par image témoin.

Implémentation expérimentale et validation

Les résultats expérimentaux, présentés dans ce texte, portent sur deux volets complémentaires : la reconnaissance d'attributs (PAR) et la ré-identification (ReID).

Sur les jeux de données attributaires, le modèle atteint des gains compris entre +0,8 et +3,5 points de mA et entre +0,9 et +2,1 points de F1 selon le protocole considéré. Concernant la recherche de personnes par description textuelle, les résultats montrent une amélioration d'environ +1,6 point en Rank-1 et +7 points en mAP par rapport aux meilleures références rapportées (Tab.3.3). Ces performances traduisent l'efficacité de l'apprentissage supervisé guidé par des représentations textuelles pour renforcer la cohérence sémantique et atténuer les déséquilibres inter-classes.

Quant à la ré-identification, deux configurations expérimentales ont été évaluées. La première compare le modèle proposé aux approches ReID purement visuelles, sans supervision attributaire (Tab.4.2), tandis que la seconde confronte notre méthode à des modèles exploitant les attributs au sein de l'apprentissage (Tab.4.3). Dans les deux cas, la combinaison de la Triplet Loss et du mécanisme de centroïdes attributaires permet d'atteindre des performances élevées, jusqu'à 96,3 % de Rank-1 et 90,1 % de mAP sur **Market1501**, et 93,9 % / 84,2 % sur **DukeMTMC-ReID**. Ces résultats montrent que l'intégration de signaux sémantiques ne dégrade pas la discrimination visuelle et qu'elle améliore la structuration métrique des identités.

Enfin, dans le cadre d'un entraînement conjoint PAR–ReID, le modèle conserve des performances stables sur les deux tâches selon les résultats rapportés dans la section.4.4. Les représentations issues de l'apprentissage unifié demeurent discriminantes pour la ReID tout en préservant la précision sur les attributs, ce qui confirme la complémentarité des deux objectifs dans la construction d'un espace d'embeddings partagé.

Extension à d'autres catégories d'objets (étude de cas : véhicules)

Le cadre proposé n'est pas limité à l'analyse de personnes et peut être transposé à d'autres catégories d'objets, dès lors qu'un ensemble d'attributs descriptifs et des gabarits textuels adaptés sont disponibles. Pour illustrer la capacité à généraliser, la branche attributaire du modèle a été appliquée à l'analyse de **véhicules** sur le jeu *Stanford Cars*, en adaptant uniquement la taxonomie des attributs (*type, marque, couleur*) et les formu-

lations textuelles correspondantes. L'expérimentation conserve la même structure d'apprentissage supervisé fondée sur la Focal Cross-Entropy et le même encodeur visuel que pour la PAR humaine, mais sans intégrer la Triplet Loss ni les centroïdes sémantiques.

Les résultats obtenus (Tab.3.4) sont comparables à ceux des méthodes dédiées à la classification d'attributs de véhicules, confirmant la transférabilité du pipeline vision–langage à d'autres domaines d'analyse visuelle. La ré-identification de véhicules n'a pas été implémentée dans ce travail, mais le cadre conceptuel reste applicable sous réserve d'une adaptation du protocole métrique et d'annotations identitaires suffisantes. Cette extension illustre la capacité du modèle à s'adapter à des ontologies différentes tout en conservant la cohérence de son espace de représentation.

Discussion critique des résultats

L'utilisation de CLIP comme base pour la PAR fournit des performances compétitives sur les attributs visuellement bien ancrés (*i.e.* « porte un sac », « cheveux longs », « jupe »), avec des gains quantifiés selon les jeux de données (cf. Tab. 3.2). Sur **PETA** et **PA100K**, les améliorations atteignent environ +2 pts de mA et +1 pt de F1, tandis que sur **RAPv2**, plus déséquilibré, les gains sont plus modestes (+0,8 pt de mA) pa rapport à l'état de l'art. L'alignement image–texte, combiné à une formulation normalisée du passage *label* $\rightarrow$ *texte* (décrite dans le Chap. 3) et à une Focal Cross-Entropy, réduit la dépendance à des classifieurs binaires spécialisés et atténue les effets du déséquilibre de classes. Les attributs rares conservent une variance plus élevée, cohérente avec la longue traîne d'occurrences, et l'estimation par seuil unique peut sous- ou sur-détecter certains attributs selon les contextes de prise de vue. La qualité d'annotation (bruit, visibilité partielle) affecte également la stabilité des mesures. Ces constats plaident pour des calibrations par attribut, un enrichissement contrôlé des gabarits textuels et une prise en compte explicite de la visibilité/occlusion dans l'évaluation. Des exemples typiques de comportements du modèle sont visualisés dans la figure de synthèse qualitative (cf. Fig. 4.3).

Discrimination identitaire et ré-identification Pour la ReID, l'introduction d'une contrainte de *centroïdes attributaires* combinée à la Triplet Loss s'accompagne d'améliorations de Rank-1 et de mAP, en particulier dans des situations de forte proximité visuelle entre individus ou d'occlusions partielles. Sur le protocole *attribute-based person retrieval* du benchmark **UPAR 2024**, les résultats montrent un gain d'environ +1,6 pt en Rank-1 et +7,0 pts de mAP par rapport aux meilleures références (cf. Tab. 3.3). Les ablations présentées dans le Chap. 4 (Tab. 4.6) mettent en évidence un effet cumulatif des composantes du modèle. Les échecs résiduels concernent surtout les variations de pose, d'éclairage ou d'uniforme. Le modèle reste sensible à la qualité et à la consistance des

annotations : des labels bruités peuvent déplacer les centroïdes et ralentir la convergence, en particulier lors du *mining* de *hard negatives*. Dans ces configurations, une normalisation stricte des embeddings, une marge modérée et un *sampling* équilibré contribuent à stabiliser l'optimisation (Chap. 4, §4.6.3).

Apprentissage conjoint PAR–ReID L'entraînement conjoint produit des embeddings plus informatifs pour les deux tâches, avec des gains mesurés sur les métriques attributaires (mA/F1) et identitaires (Rank-1/mAP) selon les jeux considérés (cf. Tab. 3.2, **??**). L'amélioration est plus marquée sur les ensembles à forte variabilité d'apparence comme **UPAR** ou **RAPv2**, où les signaux sémantiques stabilisent la représentation. Les ablations confirment qu'une pondération intermédiaire, la présence des centroïdes et une phase de *warm-up* courte limitent les interférences de gradients, au prix d'un léger surcoût mémoire. Les analyses qualitatives (Fig. 4.3) montrent que les embeddings issus du modèle conjoint présentent des voisinages visuels cohérents, où les regroupements reflètent à la fois la similarité d'apparence et la compatibilité sémantique. Enfin, les gains observés en recherche image–image (cf. Tab. 4.5) confirment que l'intégration du guidage sémantique n'altère pas la discrimination visuelle, mais améliore la structure globale de l'espace latent.

Dans l'ensemble, ces éléments confirment l'intérêt d'un espace vision–langage partagé contraint par des signaux sémantiques et métriques, tout en soulignant des points d'attention relatifs à la qualité des attributs, au réglage des pertes et à la robustesse inter-domaines. Ces observations prolongent les analyses de la section. 4 et préparent les pistes exploratoires développées dans la section suivante, notamment en matière de faible supervision, d'adaptation de domaine, d'efficience et d'intégration en contexte industriel.

Conclusion finale

Ce manuscrit a présenté un cadre multimodal unifié pour la reconnaissance d'attributs (PAR) et la ré-identification (ReID), fondé sur un espace vision–langage partagé. L'approche combine, au sein d'un entraînement en une seule étape, une supervision attributaire (Focal Cross-Entropy) et un objectif métrique (Triplet Loss), complétés par un mécanisme de *centroïdes attributaires* qui organise l'espace latent et relie des instances visuellement distinctes à une base sémantique commune. Le schéma méthodologique général est présenté en 4.1, et les jeux utilisés sont récapitulés en Tab. 3.1. Ces contributions s'inscrivent dans la continuité du cheminement amorcé au Chap. 1 et développé successivement dans les chapitres consacrés à l'état de l'art (Chap. 2), à la reconnaissance d'attributs (Chap. 3) et à la modélisation conjointe des attributs et de l'identité (Chap. 4).

Les expériences menées sur plusieurs jeux de données confirment l'intérêt du couplage PAR–ReID : les tableaux de synthèse (Tab. 3.2, **??**, 3.3) montrent des améliorations observéess sur les deux volets. Les résultats, détaillés dans le Chap. 4, indiquent que l'ajout d'un guidage sémantique ne dégrade pas la discrimination visuelle et améliore la cohérence interne de l'espace latent. Les analyses qualitatives (Fig. 4.3) confirment la complémentarité des deux branches supervisées : la branche attributaire organise les voisinages selon des régularités sémantiques, tandis que la branche métrique renforce la distinction entre identités visuellement proches.

Sur le plan pratique, l'architecture demeure compacte (un encodeur visuel unique, deux objectifs complémentaires et un mécanisme de guidage par centroïdes) et se prête à une intégration progressive en contexte industriel, notamment dans le cadre CIFRE avec Two-i. Le découpage modulaire des composants et la standardisation du passage $\textit{label} \rightarrow \textit{texte}$ facilitent la réutilisation du modèle, le suivi opérationnel et l'extension à d'autres domaines d'analyse. Les limites identifiées (disponibilité et qualité des annotations, robustesse inter-sites et explicabilité) ouvrent des perspectives ciblées, détaillées dans la section précédente : faible supervision et *zero-shot*, adaptation de domaine, efficience d'inférence et audit des décisions.

En synthèse, ce travail rapproche deux lignes de recherche longtemps traitées séparément : la description sémantique par attributs et la discrimination identitaire au sein d'un même espace vision–langage. Il propose une méthodologie unifiée, validée empiriquement sur plusieurs jeux, et ouvre une voie vers une exploitation plus générale et plus robuste des modèles multimodaux dans les tâches d'analyse de personnes et, au-delà, d'objets en environnement complexe.

Bibliographie

[1] Afshin Dehghan, Syed Zain Masood, Guang Shu, and Enrique. G. Ortiz. View independent vehicle make, model and color recognition using convolutional neural network, 2017.

[2] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training, 2018.

[3] Nikhil Kumar Singh, Manish Khare, and Harikrishna B. Jethva. A comprehensive survey on person re-identification approaches : various aspects. *Multimedia Tools Appl.*, 81(11) :15747–15791, May 2022.

[4] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection : A survey, 2019.

[5] Igor Ryabchikov, Nikolay Teslya, and Nikita Druzhinin. Integrating computer vision technologies for smart surveillance purpose. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 392–401, 2020.

[6] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, 2015.

[7] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl : Deep surroundings-person separation learning for text-based person retrieval, 2021.

[8] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification : Past, present and future, 2016.

[9] Zhangqiang Ming, Min Zhu, Xiangkun Wang, Jiamin Zhu, Junlong Cheng, Chengrui Gao, Yong Yang, and Xiaoyong Wei. Deep learning-based person re-identification methods : A survey and outlook of recent works, 2022.

[10] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95 :151–161, November 2019.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[12] Kai Han, Yunhe Wang, Han Shu, Chuanjian Liu, Chunjing Xu, and Chang Xu. Attribute aware pooling for pedestrian attribute recognition, 2019.

[13] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4) :1575–1590, 2019.

[14] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net : Attentive deep features for pedestrian analysis, 2017.

[15] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar : Unified pedestrian attribute recognition and person retrieval, 2022.

[16] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification : A survey and outlook, 2021.

[17] Shan Huang, Wenlong Hu, Bin Lu, Qiang Fan, Xinyao Xu, Xiaolei Zhou, and Hao Yan. Application of label correlation in multi-label classification : A survey. *Applied Sciences*, 14(19), 2024.

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017.

[19] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification : Theory and practice, 2018.

[20] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification, 2022.

[21] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models : Person retrieval with refined part pooling (and a strong convolutional baseline), 2018.

[22] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss : A unified perspective of pair similarity optimization, 2020.

[23] Chanho Eom, Geon Lee, Kyunghwan Cho, Hyeonseok Jung, Moonsub Jin, and Bumsub Ham. Cerberus : Attribute-based person re-identification using semantic ids, 12 2024.

[24] Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification, 2023.

[25] Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, and Liang Wang. Attribute-guided pedestrian retrieval : Bridging person re-id with internal attribute variability. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17689–17699, 2024.

[26] Mehran Adibi Sedeh, Assia Benbihi, Romain Martin, Marianne Clausel, and Cédric Pradalier. Attrivision : Advancing generalization in pedestrian attribute recognition using clip. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 354–365, 2025.

[27] Luis Felipe Wolf Batista, Salim Khazem, Mehran Adibi, Seth Hutchinson, and Cedric Pradalier. Potato : A dataset for analyzing polarimetric traces of afloat trash objects, 2024.

[28] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks : A survey, 2019.

[29] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, volume 5, pages 281–298. University of California press, 1967.

[30] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3) :37–52, 1987.

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[32] Leland McInnes, John Healy, and James Melville. Umap : Uniform manifold approximation and projection for dimension reduction, 2020.

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

[34] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[35] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders : Feature learning by inpainting, 2016.

[37] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[39] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent : A new approach to self-supervised learning, 2020.

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[41] Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. ViT-NeT : Interpretable vision transformers with neural tree decoder. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11162–11172. PMLR, 17–23 Jul 2022.

[42] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[43] Sergey Ioffe and Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift, 2015.

[44] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[48] Mingxing Tan and Quoc V. Le. Efficientnet : Rethinking model scaling for convolutional neural networks, 2020.

[49] Yuxin Wu and Kaiming He. Group normalization, 2018.

[50] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale, 2021.

[53] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer : Hierarchical vision transformer using shifted windows, 2021.

[54] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[55] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit : a vision transformer in convnet's clothing for faster inference, 2021.

[56] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner : Adaptive space-time tokenization for videos. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12786–12797. Curran Associates, Inc., 2021.

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[58] Qizao Wang, Bin Li, and Xiangyang Xue. When large vision-language models meet person re-identification, 2024.

[59] Huazhong Zhao, Lei Qi, and Xin Geng. Cilp-fgdi : Exploiting vision-language model for generalizable person re-identification, 2025.

[60] Mingzhe Hu. Vlpsr : Enhancing zero-shot object reid with vision-language model. In George Bebis, Vishal Patel, Jinwei Gu, Julian Panetta, Yotam Gingold, Kyle Johnsen, Mohammed Safayet Arefin, Soumya Dutta, and Ayan Biswas, editors, *Advances in Visual Computing*, pages 56–69, Cham, 2025. Springer Nature Switzerland.

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma,

Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

[62] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection, 2016.

[63] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115. IEEE, 2015.

[64] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *2015 International Conference on Biometrics (ICB)*, pages 535–540, 2015.

[65] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[66] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation, 2018.

[67] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet : A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE, June 2015.

[68] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.

[69] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface : Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10) :5962–5979, October 2022.

[70] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088) :533–536, 1986.

[71] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[72] Diederik P. Kingma and Jimmy Ba. Adam : A method for stochastic optimization, 2017.

[73] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[74] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd : Training imagenet in 1 hour, 2018.

[75] Ilya Loshchilov and Frank Hutter. Sgdr : Stochastic gradient descent with warm restarts, 2017.

[76] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks, 2013.

[77] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1) :1929–1958, 2014.

[78] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup : Beyond empirical risk minimization, 2018.

[79] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix : Regularization strategy to train strong classifiers with localizable features, 2019.

[80] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8) :861–874, 2006.

[81] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search : Lessons learnt from a practical surveillance system. In *Proceedings of International Conference on Multimedia Retrieval*, pages 153–160, 2014.

[82] Pigi Kouki, Ilias Fountalis, Nikolaos Vasiloglou, Nian Yan, Unaiza Ahsan, Khalifeh Al Jadda, and Huiming Qu. Product collection recommendation in online retail. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 486–490, 2019.

[83] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015.

[84] Kai Han, Yunhe Wang, Han Shu, Chuanjian Liu, Chunjing Xu, and Chang Xu. Attribute aware pooling for pedestrian attribute recognition. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2456–2462, 2019.

[85] Jiabao Zhong, Hezhe Qiao, Lin Chen, Mingsheng Shang, and Qun Liu. Improving pedestrian attribute recognition with multi-scale spatial calibration. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[86] Mickael Cormier, Andreas Specker, Julio Junior, CS Jacques, Lennart Moritz, Jürgen Metzler, Thomas B Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jürgen Beyerer. Upar challenge 2024 : Pedestrian attribute recognition and attribute-based person retrieval-dataset, design, and results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 359–367, 2024.

[87] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net : Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[88] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014.

[89] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4) :1575–1590, 2018.

[90] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification : A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[91] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11) :1949–1959, 2015.

[92] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar : Unified pedestrian attribute recognition and person retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 981–990, 2023.

[93] Ashish Jha, Dimitrii Ermilov, Konstantin Sobolev, Anh Huy Phan, Salman Ahmadi-Asl, Naveed Ahmed, Imran Junejo, Zaher AL Aghbari, Thar Baker, Ahmed Mohamed Khedr, and Andrzej Cichocki. Lightweight attribute localizing models for pedestrian attribute recognition, 2023.

[94] Esube Bekele and Wallace Lawson. The deeper, the better : Analysis of person attributes recognition, 2019.

[95] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[96] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition : A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv :2107.03576*, 2021.

[97] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019.

[98] Xinwen Fan, Yukang Zhang, Yang Lu, and Hanzi Wang. Parformer : Transformer-

based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1) :411–423, 2023.

[99] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 962–971, 2021.

[100] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[101] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m : A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[102] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *Computer Vision–ECCV 2012 : 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pages 609–623. Springer, 2012.

[103] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018.

[104] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[105] Ming Chen, Guijin Wang, Jing-Hao Xue, Zijian Ding, and Li Sun. Enhance via decoupling : Improving multi-label classifiers with variational feature augmentation. 09 2021.

[106] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 :12055–12062, 04 2020.

[107] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. Learning disentangled attribute representations for robust pedestrian attribute recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 :1069–1077, 06 2022.

[108] Yibo Zhou, Hai-Miao Hu, Yirong Xiang, Xiaokang Zhang, and Haotian Wu. Pedestrian attribute recognition as label-balanced multi-label learning, 2024.

[109] Doanh C. Bui, Thinh V. Le, Ba Hung Ngo, and Tae Jong Choi. Clear : Cross-

transformers with pre-trained language model is all you need for person attribute recognition and retrieval, 2024.

[110] Boseung Jeong, Jicheol Park, and Suha Kwak. Asmr : Learning attribute-based person search with adaptive semantic margin regularizer, 2021.

[111] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. Symbiotic adversarial learning for attribute-based person search, 2020.

[112] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5546–5555, 2015.

[113] Xinjin Li, Jinghao Chang, Tiexin Li, Wenhan Fan, Yu Ma, and Haowei Ni. A vehicle classification method based on machine learning, 07 2024.

[114] Qin Xu, Sitong Li, Jiahui Wang, Bo Jiang, and Jinhui Tang. Context-semantic quality awareness network for fine-grained visual categorization, 2024.

[115] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer : Hierarchical vision transformer using shifted windows, 2021.

[116] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification, 2022.

[117] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg : A transformer architecture for fine-grained recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 :852–860, 06 2022.

[118] Chuanming Wang, Huiyuan Fu, and Huadong Ma. Multi-part token transformer with dual contrastive learning for fine-grained image classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 7648–7656, New York, NY, USA, 2023. Association for Computing Machinery.

[119] Zi-Chao Zhang, Zhen-Duo Chen, Yongxin Wang, Xin Luo, and Xin-Shun Xu. A vision transformer for fine-grained classification by reducing noise and enhancing discriminative information. *Pattern Recognition*, 145 :109979, 2024.

[120] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 262–275, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[121] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[122] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn : A decade survey of instance retrieval, 2017.

[123] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models : Person retrieval with refined part pooling (and a strong convolutional baseline), 2018.

[124] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv :1703.07737*, 2017.

[125] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid : Transformer-based object re-identification, 2021.

[126] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet : Attribute attention network for person re-identifications. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7134–7143, 2019.

[127] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[128] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Not every patch is needed : Towards a more efficient and effective backbone for video-based person re-identification, 2025.

[129] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns : Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12) :2037–2041, 2006.

[130] Nicolai Petkov and Easwar Subramanian. Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal gabor filters with surround inhibition. *Biol. Cybern.*, 97(5–6) :423–439, December 2007.

[131] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification : A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[132] Xiaoling Xia, Cui Xu, and Bing Nan. Inception-v3 for flower classification. In *2017 2nd international conference on image, vision and computing (ICIVC)*, pages 783–787. IEEE, 2017.

[133] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, MM '18, page 274–282. ACM, October 2018.

[134] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[135] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[136] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017.

[137] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[138] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[139] Yongkang Ding, Yinghao Wu, Anqi Wang, Tiantian Gong, and Liyan Zhang. Disentangled body features for clothing change person re-identification. *Multimedia Tools and Applications*, 83(27) :69693–69714, 2024.

[140] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net : Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8351–8361, 2019.

[141] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale, 2021.

[142] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit : Towards deeper vision transformer. *arXiv preprint arXiv :2103.11886*, 2021.

[143] Myungseo Song, Jin-Woo Park, and Jong-Seok Lee. Exploring the camera bias of person re-identification, 2025.

[144] Jincheng Yan, Yun Wang, Xiaoyan Luo, and Yu-Wing Tai. Fusionsegreid : Advancing person re-identification with multimodal retrieval and precise segmentation, 2025.

[145] Asaf Joseph and Shmuel Peleg. Clothes-changing person re-identification based on skeleton dynamics, 2025.

[146] Kunjun Li, Cheng-Yen Yang, Hsiang-Wei Huang, and Jenq-Neng Hwang. Technical report for reid-sam on skitb visual tracking challenge 2025, 02 2025.

[147] Wei Liu, Xin Xu, Hua Chang, Xin Yuan, and Zheng Wang. Mix-modality person re-identification : A new and practical paradigm, 2024.

[148] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery : Occluded person re-identification with part-aware transformer, 2021.

[149] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95 :151–161, November 2019.

[150] Qizao Wang, Bin Li, and Xiangyang Xue. When large vision-language models meet person re-identification, 11 2024.

[151] Mingzhe Hu. Vlpsr : Enhancing zero-shot object reid with vision-language model. In *International Symposium on Visual Computing*, pages 56–69. Springer, 2024.

[152] Bingyu Duan, Wanqian Zhang, Dayan Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. Exploiting vision-language model for visible-infrared person re-identification via textual modality alignment. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[153] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification, 2018.

[154] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification, 2021.

[155] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification, 2020.

[156] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning, 2019.

[157] Chenxin Sun, Na Jiang, Lei Zhang, Yuehua Wang, Wei Wu, and Zhong Zhou. *Unified Framework for Joint Attribute Classification and Person Re-identification : 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I*, pages 637–647. 09 2018.

[158] Yuming Yan, Huimin Yu, Shuzhao Li, Zhaohui Lu, Jianfeng He, Haozhuo Zhang, and Runfa Wang. Weakening the influence of clothing : Universal clothing attribute disentanglement for person re-identification. pages 1498–1504, 07 2022.

[159] Jiawei Liu, Zheng-Jun Zha, Hongtao Xie, Zhiwei Xiong, and Yongdong Zhang. Ca3net : Contextual-attentional attribute-appearance network for person re-identification, 2018.

[160] Shuzhao Li, Huimin Yu, Wei Huang, and Jing Zhang. Attributes-aided part detection and refinement for person re-identification, 2019.

[161] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet : Attribute attention network for person re-identifications, 2019.

[162] Bo Jiang, Xixi Wang, and Jin Tang. Attkgcn : Attribute knowledge graph convolutional network for person re-identification, 2019.

[163] Binh X. Nguyen, Binh D. Nguyen, Tuong Do, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Graph-based person signature for person re-identifications, 2021.

[164] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification, 2018.

[165] Qi Dong, Xiatian Zhu, and Shaogang Gong. Person search by text attribute query as zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3651–3660, 2019.

[166] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. Symbiotic adversarial learning for attribute-based person search, 2020.