Titre: Title:	Adaptation de domaine non supervisée pour la ré-identification de personnes dans des vidéos
Auteur: Author:	Yacine Khraimeche
Date:	2020
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Khraimeche, Y. (2020). Adaptation de domaine non supervisée pour la ré-identification de personnes dans des vidéos [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5324/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/5324/
Directeurs de recherche: Advisors:	Guillaume-Alexandre Bilodeau
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Adaptation de domaine non supervisée pour la ré-identification de personnes dans des vidéos

YACINE KHRAIMECHE
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Juillet 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Adaptation de domaine non supervisée pour la ré-identification de personnes dans des vidéos

présenté par **Yacine KHRAIMECHE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Michel DAGENAIS, Ph. D., président
Guillaume-Alexandre BILODEAU, Ph. D., membre et directeur de recherche
Benjamin DE LEENER, Ph. D., membre

DÉDICACE

À ma famille, mes amis,
et mes professeurs....

REMERCIEMENTS

Le travail présenté dans ce mémoire représente une étape enrichissante de ma vie, tant sur le plan académique que personnel. Il n'aurait pas été possible sans l'apport et la participation des personnes qui ont entouré ce projet.

Je tiens à remercier tout particulièrement mon directeur de recherche, le professeur Guillaume-Alexandre Bilodeau, qui m'a guidé tout au long de la réalisation de ce travail en partageant son expertise et ses intuitions toujours pertinentes.

Je remercie également toute l'équipe d'Arcturus et notamment David Steele et Harshad Mahadik pour avoir mis en place le projet et l'avoir suivi avec intérêt.

Mes plus sincères remerciement aux membres du jury, M. Michel Dagenais et M. Benjamin de Leener pour avoir pris le temps de lire et d'évaluer ce mémoire.

Enfin, un grand merci à tous mes collègues du laboratoire LITIV, tout particulièrement mes amis Jules et Medhi avec qui j'ai travaillé dans cette collaboration avec Arcturus.

RÉSUMÉ

La ré-identification de personnes (ReID) vise à faire correspondre les identités de personnes à travers un réseau de caméras de vidéosurveillance. Cette tâche de reconnaissance précise est un élément clé de plusieurs tâches d'analyse vidéo dans les réseaux de caméras connectées à grande échelle dans les villes intelligentes, telles que le suivi multi-objets multi-caméras ou la recherche de personnes. C'est une tâche difficile en raison de la grande variation intra-classe causée par plusieurs facteurs, par exemple la pose, le point de vue, l'arrière-plan, l'éclairage, les occlusions, l'étalonnage de la caméra, etc.

Malgré ces difficultés, les modèles d'apprentissage profonds ont récemment considérablement amélioré les performances des méthodes de ReID de personnes, dépassant même la précision humaine. Un réseau de ReID apprend à générer des caractéristiques discriminantes encapsulant l'identité d'une personne tout en étant indépendant des facteurs de variation intra-classe.

Cependant, cette affirmation n'est vraie que dans un cadre supervisé dans lequel nous avons accès à un ensemble d'entraînement d'images annotées du domaine. Les caractéristiques de ReID ne se généralisent pas bien. Lorsque les réseaux sont appliqués à un domaine différent de celui sur lequel ils ont été entraînés, les performances des modèles ReID s'effondrent car les vecteurs de caractéristiques en sortie ne sont pas adaptés à ce nouveau domaine. Les images du nouveau domaine peuvent être très différentes des images d'entraînement en raison de l'environnement, de l'éclairage, du point de vue, des paramètres et du modèle de caméra différents, etc. Toutes ces variations viennent s'ajouter aux variations intra-classe, et le réseau de ReID échoue car il n'est pas entraîné pour y faire face.

Malheureusement, pour les applications pratiques avec de grands réseaux de caméras, il est impossible de collecter et d'étiqueter une grande quantité de données pour chaque nouvelle configuration d'utilisation ou de caméra. Le contexte non supervisé, dans lequel les images annotées ne sont pas disponibles, est un contexte plus réaliste pour l'utilisation pratique de ReID. C'est pourquoi l'adaptation de domaine en ReID de personnes, c'est-à-dire l'optimisation pour un domaine cible non annoté d'un réseau de ReID entraîné sur un domaine source annoté, est la principale piste de recherche à explorer avant d'envisager une utilisation pratique à grande échelle de la ReID. Certaines méthodes d'adaptation de domaine récentes s'attaquent au problème de la ReID de personnes non supervisée. Cependant, il y a encore de la place pour de l'amélioration car la performance dans le cadre non supervisé est encore bien en deçà de celle supervisée.

Pour aborder le cadre non supervisé en ReID de personnes, nous introduisons un nouveau

système de réseaux de neurones qui incorpore le réseau de ReID dans un réseau adverse génératif (GAN), et qui les entraîne conjointement pour une tâche auxiliaire : la reconstruction d'images. Cela permet de désentrelacer les caractéristiques dans le domaine non annoté. Le réseau ReID apprend à extraire des caractéristiques discriminantes dans le domaine source annoté selon l'identité. Cependant, en raison de l'écart entre les domaines, ces caractéristiques ne couvrent pas tous les facteurs de variation dans le domaine cible. Nous introduisons un deuxième réseau pour apprendre les caractéristiques complémentaires du domaine. Il y a ainsi désentrelacement, car le réseau de ReID se concentre sur l'identité tandis que le deuxième réseau apprend des caractéristiques complémentaires nécessaires pour reconstruire des images. Nous opérons donc dans l'espace des caractéristiques latentes pour supprimer les caractéristiques non liées à l'identité des caractéristiques de ReID. Des travaux antérieurs ont prouvé l'efficacité du désentrelacement pour la ReID de personnes supervisée. Sans aucune annotation, notre méthode est en mesure d'imposer des contraintes efficaces pour inciter le réseau de ReID à produire des vecteurs de ReID adaptés au nouveau domaine.

Notre système de réseaux peut être utilisé avec n'importe quelle architecture pour le réseau de ReID, et ne nécessite pas de calcul supplémentaire lors de l'inférence. Nous avons mené des expériences sur les bases de données MSMT17, Market1501 et DukeMTMC. Nous dépassons la performance de l'état de l'art sur un transfert de bases de données de référence pour la ReID de personnes non supervisée.

ABSTRACT

Person re-identification (ReID) aims at matching pedestrian identities across a non-overlapping video surveillance camera network. This fine-grained recognition task is a key component for several video analysis tasks in connected large-scale camera network in smart cities, such as multi-target multi-camera tracking or person retrieval. It is a challenging task due to the large intra-class variation caused by multiple factors, for example pose, viewpoint, background, lighting, occlusions, camera calibration, etc.

Despite these difficulties, deep learning models have recently improved significantly the performance of person ReID methods, even surpassing human accuracy. A ReID network learns to generate discriminative features encapsulating the identity of a person while being independent to the intra-class variations factors.

However, this statement is only true in a supervised setting in which we have access to a training set of domain-related labelled images. ReID features do not generalize well. When the networks are applied to a domain different from the one they were trained on, the performance of ReID models collapses since the output feature vectors are not adapted to this new domain. Images from the new domain can be widely different from the training images because of different environment, lighting, viewpoint, camera settings and model, and so on. All added variations are factoring with the intra-class ones, and the ReID network fails since it is not trained to deal with these.

Unfortunately, for real-world applications with large camera networks, it is impossible to collect and label large amount of data for each new use case or camera setup. The unsupervised setting, in which labelled images are not available, is a more realistic context for practical use of ReID. That is why domain adaptive person ReID, that is to say the optimization for an unlabelled target domain of a ReID network trained on a labelled source domain, is the main research avenue to explore before considering large-scale practical use of ReID. Some recent domain adaptive methods tackle the problem of unsupervised person ReID. However, there is still room for improvement as the performance in the unsupervised setting is still far below its supervised counterpart.

To address the unsupervised setting in person ReID, we introduce a novel neural network framework that incorporates the ReID network in a Generative Adversarial Network (GAN), and that jointly trains them with an auxiliary task : image reconstruction. This allows features disentanglement in the unlabelled domain. The ReID network learns to extract discriminative features in the labelled source domain using identity annotations. However,

due to the domain gap, these features do not encompass every factor of variation in the target domain. We introduce a second network to learn the complementary features of the domain. This results in disentanglement, because the ReID network focus on identity while the second network learns additional features necessary to reconstruct images. We operate in the latent feature space to distill identity-unrelated features from the ReID features. Previous work proved the effectiveness of disentanglement for supervised person ReID. Without any labelling, our method is able to enforce efficient constraints to encourage the ReID network to produce ReID vectors adapted for the new domain.

Our framework can be used with any architecture for the ReID network, and does not require additional computing during inference. We conducted experiments on the MSMT17, Market1501 and DukeMTMC datasets. We achieve state-of-the-art performance on a reference dataset transfer benchmark for unsupervised Person ReID.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ReID	Ré-identification
GAN	Generative Adversarial Network
KL	Kullback-Leibler
mAP	Mean Average Precision
CMC	Cumulated Matching Characteristics
AdaIN	Adaptive Instance Normalization
MMD	Maximum Mean Discrepancy

CHAPITRE 1 INTRODUCTION

Depuis quelques années, les villes se dotent de systèmes intelligents visant à améliorer la qualité de vie des habitants. Ces systèmes collectent et analysent des données pour optimiser la distribution des services, faciliter la mobilité urbaine par un meilleur contrôle du trafic routier, assurer une sécurité accrue, préserver l'environnement en minimisant les pollutions de toutes natures, etc.

Les caméras de vidéosurveillance sont des éléments centraux dans le dispositif de collecte de données de ces villes intelligentes. Les caméras sont organisées en réseaux : localement à l'échelle d'un bâtiment ou d'un complexe (centre commercial, aéroport, campus universitaire...), ou plus étendus à l'échelle de la ville. Les réseaux quadrillent les villes et génèrent des énormes volumes de données sous forme de flux vidéo. Leur analyse manuelle est fastidieuse et coûteuse en ressources humaines, d'où un besoin d'automatisation des tâches d'analyse complexes de flux vidéos provenant de sources multiples.

1.1 Définitions et concepts de base

Les avancées récentes en vision par ordinateur, et en particulier l'utilisation de techniques d'apprentissage profond, répondent à ce besoin en améliorant significativement les performances en détection automatique et en analyse des vidéos. Ces algorithmes doivent résoudre des problèmes complexes : en gestion de trafic routier, on s'interrogera par exemple sur le volume du trafic routier à une intersection ; pour la sécurisation d'une station de métro, on souhaitera déclencher une alerte si une personne se trouve sur les rails ou si un colis abandonné est détecté.

Un processus classique d'analyse de vidéo par apprentissage profond se décompose en trois étapes principales : la détection, le suivi et l'analyse des trajectoires ou d'activités. La première étape de la détection des objets d'intérêts dans une vidéo consiste typiquement à tracer une boîte englobante autour de chaque objet d'intérêt dans chaque trame. Pour cela, un réseau de neurone est entraîné sur des bases de données d'images de la même nature que l'élément à détecter, par exemple des personnes ou des voitures. Le réseau appliqué à une image peut alors repérer toutes les instances d'apparition de l'élément [13]. On commence aussi à extraire les masques de segmentation pour une détection plus précise [14].

Lors de la deuxième étape, les objets détectés dans chaque trame sont mis en correspondance dans le temps pour constituer une trajectoire. L'algorithme d'association de données repose

typiquement sur deux distances : une distance spatiale entre la position de chaque boîte englobante, et une distance en terme d'apparence. Les meilleures associations sont celles qui minimisent ces distances.

Enfin, la dernière étape est l'analyse des trajectoires ou leur utilisation pour répondre au problème posé initialement.

À l'échelle d'un réseau de caméras, il faut ajouter une étape cruciale de partage des informations de chaque caméra du réseau : la ré-identification (ReID). La ReID vise à mettre en correspondance les identités des éléments d'intérêts détectés à travers un réseau de caméras dont les champs de vision ne se recoupent pas, en construisant un modèle d'apparence pour chaque identité détectée. La Figure 1.1 résume les différents modules intervenant dans la tâche de suivi multi-personnes et multi-caméra dans un réseau de caméras de vidéosurveillance.

Dans ce projet, on s'intéressera uniquement à la ré-identification de personnes dans des vidéos. On notera toutefois que les méthodes développées dans ce mémoire peuvent être appliquées sans modifications majeures à la ré-identification de véhicules, ou tout autre objet : il suffit de réaliser l'entraînement sur les bases de données adaptées. La figure 1.2 présente un exemple de ReID dans le réseau de caméras de vidéosurveillance de la base de donnée CamNet [15].

1.2 Éléments de la problématique

La ré-identification de personnes est basée sur la construction d'un modèle d'apparence robuste. Le module de ReID appliqué à chaque personne détectée génère un vecteur caractéristique de l'identité d'une personne, appelé vecteur de ReID. Le but du module de ReID est de générer des vecteurs de ReID discriminants pour l'identité, c'est à dire qui minimisent la distance entre les vecteurs de ReID de deux détections correspondant à la même personne, tout en maximisant la distance entre deux vecteurs provenant de personnes différentes.

Les vecteurs de ReID doivent être indépendants de nombreux facteurs de variation de l'apparence des personnes détectées, résumés dans la figure 1.3. On observe tout d'abord une grande diversité de poses, qui dépend du positionnement de la caméra et de la position du corps de la personne filmée. De plus, le contexte de la prise de vue introduit de nouveaux facteurs de variation, comme l'arrière-plan changeant selon l'emplacement de la caméra et l'éclairage qui varie selon le moment de la journée. En particulier, la mise en correspondance de détections réalisées en intérieur avec des prises de vue extérieures est délicate puisque le contexte est énormément modifié. À ces difficultés s'ajoutent des différences directement liées aux caméras, comme le modèle ou la calibration de chaque caméra qui peuvent varier au sein d'un réseau. Enfin, la qualité des détections est un facteur supplémentaire compliquant la

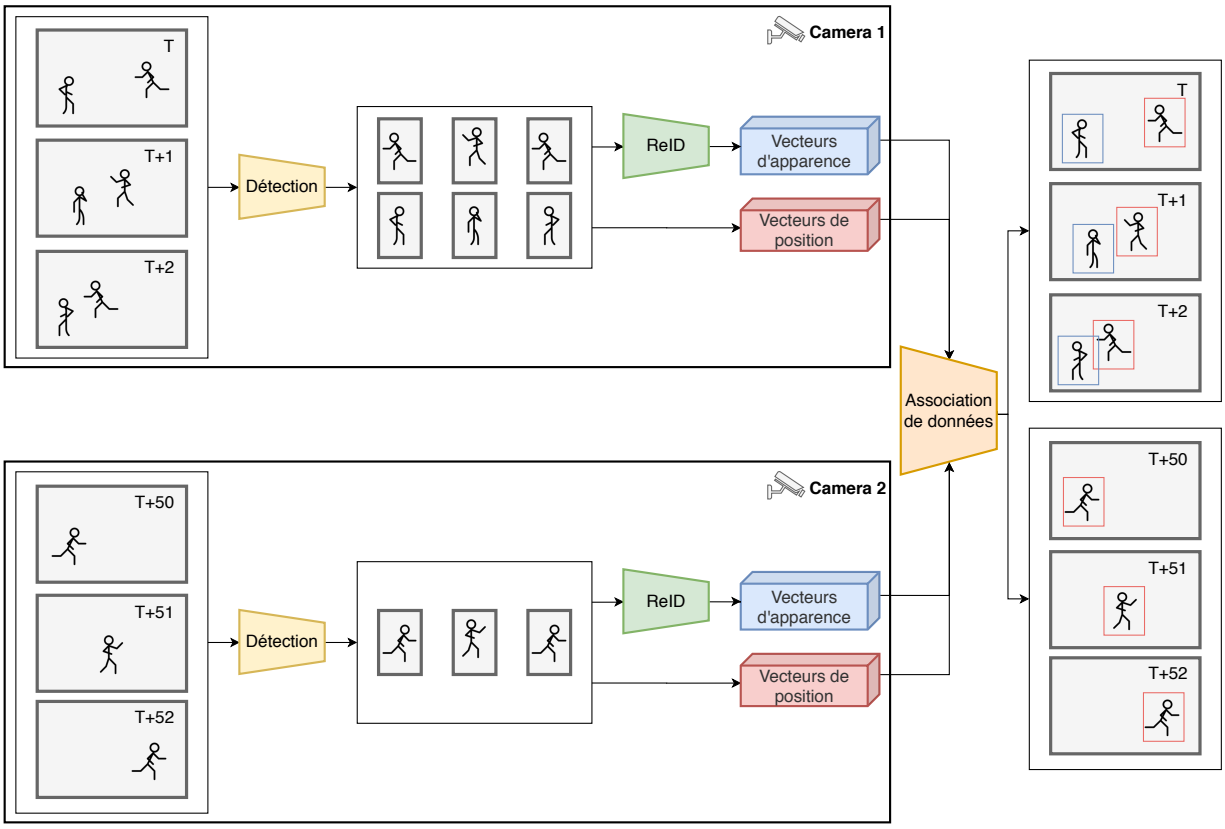

Figure 1.1 Architecture globale d'un système de suivi multi-personnes et multi-caméra dans un réseau de caméras de vidéosurveillance. Les modules de détection et de ReID sont intégrés dans les caméras, tandis que l'association de données est centralisée. Dans l'exemple, une personne apparaissant dans la vue de la première caméra est ré-identifiée dans la vue de la deuxième caméra (boîte englobante rouge).

tâche de ReID, puisqu'elles peuvent être incorrectes ou présenter des occlusions.

Malgré ces difficultés, les méthodes actuelles sont performantes dans un contexte supervisé, dans lequel on dispose d'un ensemble d'entraînement où des images du domaine sont annotées par identité. Le réseau apprend alors à correctement distinguer les identités pour un réseau de caméras donné, dans les mêmes conditions que lors de l'entraînement. Les performances des méthodes récentes dépassent même en précision celles de l'humain sur ces bases de données [16].

Cependant, si on applique le même réseau de ReID à une autre base de donnée, la performance s'effondre puisque les vecteurs de ReID générés en sortie ne sont pas adaptés au nouveau domaine. Les images proposées lors de l'entraînement peuvent en effet différer en terme

Figure 1.2 Exemple de suivi de personnes dans le réseau de caméras de la base de donnée CamNet. La personne avec un gilet rouge est ré-identifiée dans les vues de chaque caméra

d'environnement et d'éclairage, ne pas contenir les mêmes points de vue ou le même contenu, être capturées par un modèle de caméra différent, etc. Autant de difficultés s'ajoutant aux facteurs de variations énoncés précédemment et pour lesquels le réseau de ReID n'est pas entraîné. La Figure 1.4 présente une comparaison des images de plusieurs bases de données de ReID.

Or en pratique il est impossible de collecter et d'annoter des données pour chaque nouveau contexte d'utilisation de la ReID. Le contexte non supervisé, dans lequel on ne dispose pas d'annotations d'identité pour les images du domaine, est donc le contexte réaliste conforme à la mise en oeuvre pratique de la ReID. L'adaptation de domaine en ReID, qui consiste à optimiser pour le domaine cible non supervisé un réseau entraîné sur un domaine source annoté, constitue donc le défi principal à surmonter pour envisager le déploiement pratique de ces techniques.

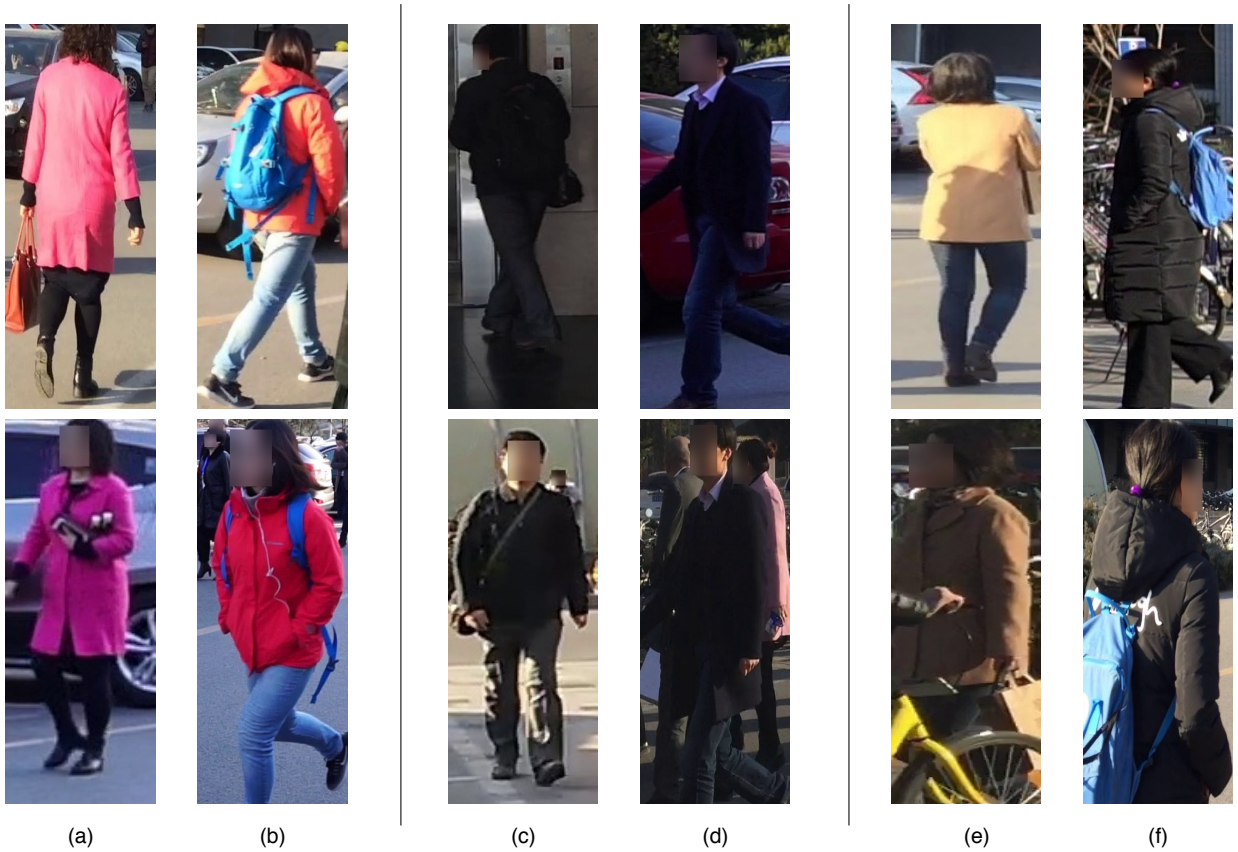

Figure 1.3 Difficultés posées par les facteurs de variations d'apparence des paires de personnes détectées dans la base de données MSMT17 [6]. (a)-(b) : variation de pose. L'apparence change suivant le point de vue et la position du corps de la personne. (c)-(d) : variation d'environnement. On observe un changement d'éclairage entre intérieur et extérieur, et des variations dans l'arrière-plan. (e)-(f) : difficultés liées à la détection. On constate la présence d'occlusions et de boîtes englobantes mal positionnées.

1.3 Objectifs de recherche

L'objectif de ce projet est de développer un algorithme efficace d'adaptation de domaine pour la ré-identification de personnes afin de répondre au problème de perte de performance en milieu non supervisé. La méthode proposée répondra aux objectifs spécifiques suivants :

- La méthode développée ne doit pas nécessiter d'annotations du domaine cible qui serait coûteuse en ressource humaine et impossible à réaliser à grande échelle.
- Le processus d'adaptation de domaine ne doit pas sacrifier la rapidité d'exécution à la précision en ajoutant des calculs supplémentaires lors de l'inférence dans le domaine cible.

Figure 1.4 Exemples de paires d'images issues de plusieurs bases de données de ReID : colonne a) Viper [7], colonne b) PRID [8], colonne c) CUHK03 [9], colonne d) Market1501 [10], colonne e) DukeMTMC [11, 12] et colonne f) MSMT17 [6]. Les images diffèrent par la résolution, la gamme de couleurs, l'éclairage, etc.

— Enfin, la méthode devra être validée sur plusieurs bases de données, et améliorer significativement et de manière consistante la performance initiale du réseau de ReID.

1.4 Contributions

Nous proposons un nouveau système de réseaux de neurones pour l'adaptation de domaine non supervisée en ReID. Nous intégrons le réseau de ReID dans un système de réseaux antagonistes génératifs, et l'entraînons dans le domaine cible à une tâche auxiliaire : la génération d'images. Sans utiliser d'annotations dans ce nouveau domaine, nous parvenons à mettre en place des contraintes efficaces pour inciter le réseau de ReID à générer des vecteurs de ReID adaptés au nouveau domaine. Nous implémentons pour cela des fonction de pertes

visant à désentrelacer les caractéristiques extraites des images relatives à l'identité de la personne des autres caractéristiques de l'image. Le modèle proposé est généralisable à tous types d'architectures pour le réseau de neurones de ReID.

À notre connaissance, il s'agit de la première tentative d'utilisation de réseaux antagonistes génératifs de désentrelacement pour l'adaptation de domaine en ReID. Des travaux antérieurs se sont intéressés au désentrelacement et ont montré l'efficacité de cette méthode dans un contexte supervisé. Nous introduisons également une méthode pour extraire des paires d'images de même identité dans un contexte non supervisé. Ces paires sont utilisées pour régulariser le processus de désentrelacement.

1.5 Plan du mémoire

Après ce chapitre introductif, le chapitre 2 établit une revue critique des concepts et travaux existants dans la littérature, dans les domaines pertinents. Nous aborderons le sujet de la ReID en milieu supervisé et non supervisé, avant d'explorer les méthodes d'adaptation de domaine, les réseaux antagonistes génératifs et le concept de désentrelacement.

Le chapitre 3 explique la méthode proposée. Nous y détaillerons chaque composante du modèle mis en place, les fonctions de pertes employées et le processus d'entraînement.

Le chapitre 4 présente les expériences et résultats validant la solution proposée. Nous discuterons ces résultats en les mettant en perspective par rapport à l'état de l'art.

Le chapitre 5 conclura sur les apports du projet et les perspectives d'améliorations.

CHAPITRE 2 REVUE DE LITTÉRATURE

Nous passons en revue dans ce chapitre la littérature relative à notre projet. La première partie sera consacrée à la mise en place du cadre d'étude. Ensuite, nous aborderons la ReID de personnes dans un cadre supervisé. Enfin, nous présenterons des méthodes de ReID dans un contexte non supervisé.

2.1 Cadre d'étude de la ReID

La ReID de personnes est un sujet de recherche très actif. Nous définissons dans cette partie les deux cadres d'études distincts principalement explorés : les cadres supervisés et non supervisés.

2.1.1 Étapes de la ReID

Un système général de ReID procède en cinq étapes, résumées à la Figure 2.1 :

1. **Collecte des données** à partir des caméras de vidéosurveillance, sous forme de flux vidéo provenant de chacune des caméras du réseau. Le type de caméras utilisées (domaine visible, infrarouge, thermique...) ainsi que leurs emplacements (intérieur/extérieur, point de vue...) sont déterminants quant à la difficulté de la tâche de ReID.
2. **Génération des boîtes englobantes** autour de chaque personne de chaque image du flux vidéo. Cette tâche étant trop laborieuse pour être effectuée manuellement, elle est généralement réalisée automatiquement par un réseau de neurones, par exemple RetinaNet [17]. Obtenir des boîtes englobantes précises est primordial, mais en pratique des impératifs relatifs au temps d'exécution peuvent obliger à faire une concession sur la qualité des détections en utilisant des méthodes plus rapides mais moins précises comme YOLO [13].
3. **Annotation des identités** des boîtes englobantes. Les variations importantes entre domaines obligent la plupart du temps à annoter des identités pour construire un ensemble d'entraînement spécifique à chaque nouveau contexte d'utilisation de la ReID. Cette tâche est effectuée manuellement, et constitue une limite pour la réalisation pratique de la ReID.
4. **Entraînement du modèle** sur les images annotées. L'utilisation des images annotées pour concevoir une représentation discriminante grâce à une métrique d'apprentissage

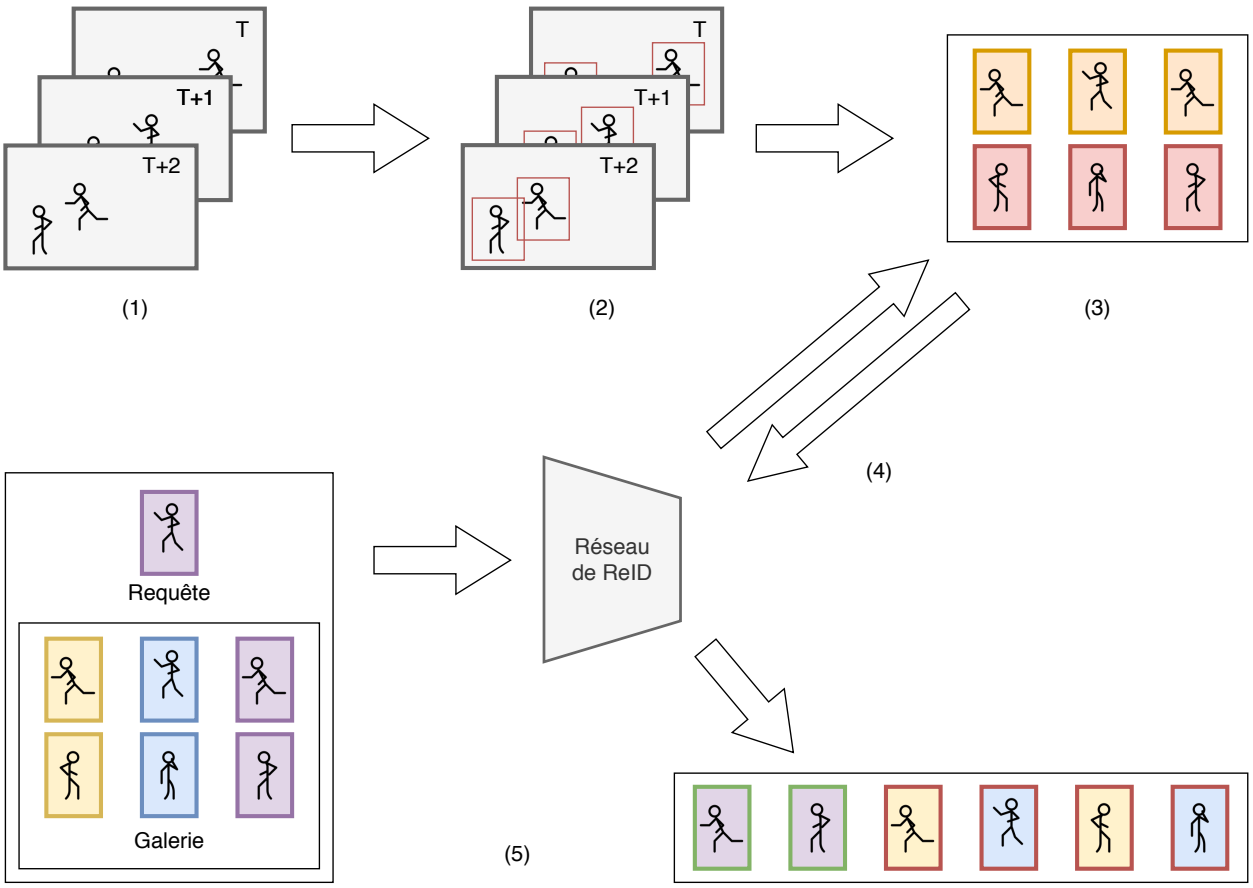

Figure 2.1 Les cinq étapes du processus général de ReID : (1) Collecte des données, (2) Génération des boîtes englobantes, (3) Annotation des identités, (4) Entraînement du modèle, (5) Évaluation du modèle.

adaptée est le sujet central de la recherche en ReID.

5. **Évaluation du modèle** par un exercice de recherche de personnes dans une galerie. Pour chaque image requête, on établit une liste triée de résultats selon la distance dans l'espace de caractéristiques de ReID. Le ré-agencement des résultats pour optimiser la performance fait également l'objet de travaux [18].

2.1.2 Définition des cadres supervisés et non supervisés

D'après [19], on définit le paradigme classique de la ReID supervisée en émettant une hypothèse sur chacune de ces cinq étapes :

1. On se limite à des réseaux de caméras fonctionnant dans le spectre visible, par opposition à des systèmes hétérogènes, par exemple une combinaison de flux vidéos

infrarouges et du spectre visible.

2. La ReID est réalisée sur des boîtes englobantes générées au préalable, et non pas sur le flux vidéo brut. On écarte ici les travaux réalisant conjointement la détection et l'extraction de caractéristiques.
3. La ReID s'effectue dans un cadre supervisé, en présence de suffisamment de boîtes englobantes annotées pour construire l'ensemble d'entraînement spécifique au domaine. Les méthodes non supervisées seront passées en revue dans la deuxième partie de ce chapitre.
4. L'ensemble d'entraînement est correctement annoté, sans erreur d'identité. La conception de modèles robustes au bruit fait également l'objet de recherches.
5. L'évaluation s'effectue par une recherche dans une galerie contenant la même identité que la requête. On classe donc simplement les résultats selon la distance de ReID pour répondre au problème. La question de la vérification d'identité se pose également et conduit à concevoir des méthodes spécifiques pour établir si oui ou non deux images contiennent la même identité, par exemple en définissant un seuil.

Les hypothèses 3 et 4 du cadre supervisé sont les plus contraignantes car elles supposent l'annotation manuelle des boîtes englobantes. Le cadre non supervisé ne présuppose pas la mise à disposition de ces annotations, mais conservera tout de même les hypothèses 1, 2 et 5.

2.2 Ré-identification de personnes supervisée

Trois axes d'amélioration se détachent dans le cadre supervisé : la construction d'une représentation discriminante dans l'espace des caractéristiques, l'optimisation de la métrique d'apprentissage, et le ré-agencement des résultats. On se limitera dans cette étude aux méthodes d'apprentissage profond, dont les performances récentes dépassent largement les résultats plus anciens obtenus par des méthodes traditionnelles résumées en [20].

2.2.1 Représentation des caractéristiques

Dans le cadre fermé, on dispose d'une base de données d'images du réseau de caméra contenant chacune une personne détectée, et annotées par identité. L'objectif du réseau de ReID est, étant donné une image d'une personne en entrée, de produire un vecteur de caractéristiques discriminant pour l'identité. Ainsi, deux images de la même personne conduiront à des vecteurs proches dans l'espace latent de caractéristiques, selon une distance adaptée (le plus souvent la distance euclidienne). Plusieurs approches sont développées pour améliorer

le vecteur de caractéristique. On peut distinguer : les approches globales qui extraient un vecteur unique pour l'ensemble de l'image, les approches locales qui agrègent des vecteurs locaux ciblant des parties de l'image, l'utilisation de données auxiliaires, et la conception d'architectures de réseaux de neurones spécialisés.

Caractéristiques globales

L'approche globale consiste à extraire un vecteur de caractéristiques unique pour toute l'image. Un réseau de neurones convolutif entraîné à la classification d'images sur une base de données d'images de type ImageNet [21] produit déjà ce vecteur de caractéristique en sortie de sa dernière couche convolutive, celle avant l'application de la couche entièrement connectée de prédiction de classe (Figure 2.2). Le réseau est ensuite entraîné sur une base de données de ReID, contentant des images de personnes annotées par identité, pour affiner la représentation latente pour la tâche de ReID.

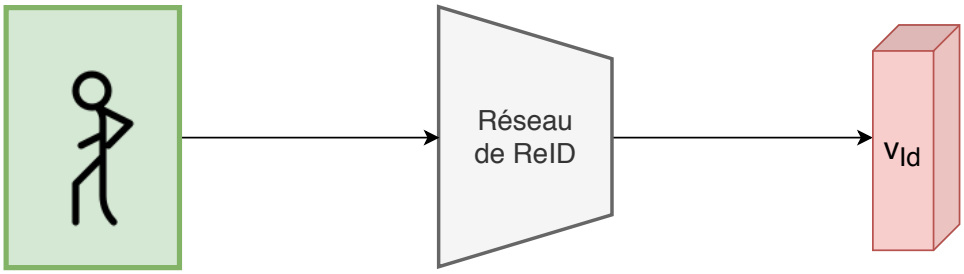

Figure 2.2 Approche globale de représentation des caractéristiques

D'une manière générale, les architectures les plus performantes pour la tâche plus générale de classification d'images sont également adaptées pour la ReID. Parmi les réseaux obtenant les meilleurs performances en se basant uniquement sur l'extraction globale de caractéristiques, on peut citer l'architecture ResNet [22] souvent utilisée comme architecture de base en ReID.

On peut ajouter des mécanismes supplémentaires pour tirer parti des spécificités de la tâche de ReID afin de produire une meilleure représentation. En effet, tous les pixels de l'images ne contribuent pas au même degré à l'identité visuelle de la personne : l'arrière-plan devrait par exemple être ignoré. Des mécanismes d'attention parviennent ainsi à focaliser le processus d'apprentissage sur l'avant-plan de l'image contenant la personne, qui contribuera davantage au vecteur de ReID produit. Le réseau HA-CNN [23] basé sur l'architecture ResNet combine deux mécanismes d'attention : une attention spatiale douce, par laquelle le réseau apprend à attribuer un score d'importance à chaque pixel, et une attention dure régionale où seules les parties discriminatives de l'image sont sélectionnées, pour corriger des détections imprécises

et recentrer les boîtes englobantes sur la personne.

Caractéristiques locales

Les approches locales produisent plusieurs vecteurs de caractéristiques locaux correspondant à différentes parties de l'image. Le vecteur de ReID final est obtenu en agrégeant les vecteurs locaux.

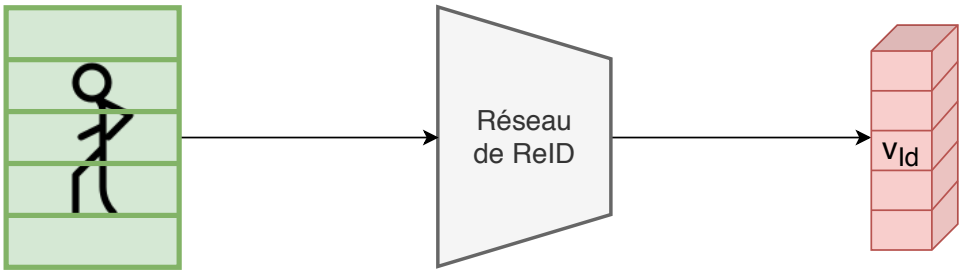

Figure 2.3 Approche locale de représentation des caractéristiques

Jusqu'à présent, les méthodes présentées ne sont pas spécifiques à la ReID de personnes : elles pourraient s'appliquer directement à la détection de véhicules par exemple, en utilisant la base de données d'apprentissage adaptée. Or, les images de personnes présentent des invariances liées à la structure du corps humain dont des modèles tirent parti pour affiner la représentation latente des personnes. Ces méthodes proposent d'agréger des caractéristiques locales extraites à partir des zones de l'image correspondant à différentes parties du corps humain (Figure 2.3).

Parmi ces méthodes, on peut distinguer deux catégories : celles utilisant explicitement la pose, comme le réseau Spindle Net [24] qui extrait sept boîtes englobantes correspondant à différentes parties du corps (tête, haut du corps, bas du corps, deux bras et deux jambes), et celles qui découpent l'image en des zones prédéfinies, par exemple en des zones horizontales pour le réseau MGN [4]. Les caractéristiques locales sont ensuite extraites de chaque partie, et agrégées pour former le vecteur global de caractéristiques.

Les modèles basés sur la pose utilisent une segmentation plus précise, mais leur performance dépend de la qualité de l'extraction de pose. De plus, l'étape supplémentaire de détection de pose réalisée par un autre réseau de neurones peut ralentir l'inférence. En comparaison, les méthodes utilisant des bandes horizontales ne rajoutent pas de calculs supplémentaires, mais sont sensibles à la qualité de la boîte englobante et aux occlusions.

Caractéristiques auxiliaires

Certains modèles tirent parti de données additionnelles pour améliorer la performance des réseaux de ReID (Figure 2.4). Il peut s'agir de données sémantiques, comme par exemple la détection des vêtements portés, etc.

Le modèle st-ReID [25] propose de compléter l'information visuelle par des données temporelles. La distribution des temps de trajet entre chaque paire de caméra du réseau est constituée à partir des images de l'ensemble d'entraînement, ce qui permet d'attribuer un score temporel à une paire d'image lors de l'évaluation. La distance entre deux images est calculée à partir du score temporel et de la distance visuelle entre les vecteurs de ReID.

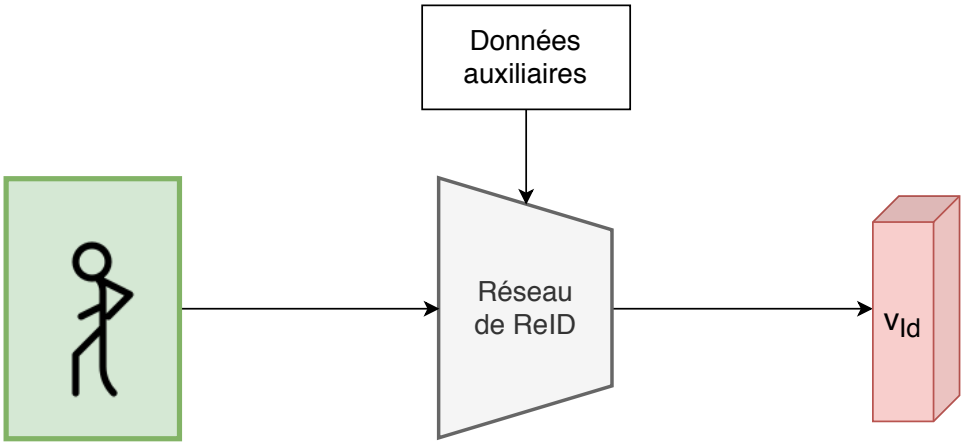

Figure 2.4 Approche de représentation des caractéristiques à partir de données auxiliaires

Des techniques d'augmentation de données sont également appliquées en ReID, notamment en utilisant des réseaux adverses génératifs (GAN) [26]. Ainsi, le modèle génératif DG-Net [2] multiplie la taille de la base de données d'entraînement en générant des images mélangeant la structure et l'apparence des images originales.

Conception d'architecture

La plupart des modèles de ReID utilisent des architectures développées pour la tâche de classification d'images. Des travaux explorent la conception d'architectures de réseaux de neurones spécifiquement pour la tâche de ReID. Parmi ces modèles, on peut citer le réseau OSNet [3], qui propose d'extraire de caractéristiques en parallèle sur plusieurs échelles. Cela permet de représenter à la fois les caractéristiques générales de la personne, et les détails plus fins.

En pratique, des blocs résiduels parallèles réalisent l'extraction, et les caractéristiques pro-

duites sont agrégées dynamiquement. Ce processus est réalisé au sein des blocs OS illustrés à la Figure 2.5 : quatre canaux extraient l'information à plusieurs échelles en appliquant une série de convolutions séparables en profondeur [27]. Le module d'agrégation prend en entrée les vecteurs générés par chaque canal et produit un vecteur de poids qui accorde plus ou moins d'importance à chaque échelle d'extraction. Cette structure résulte en un réseau léger mais particulièrement efficace, sans nécessiter de mécanismes complexes d'attention ou l'utilisation de données additionnelles.

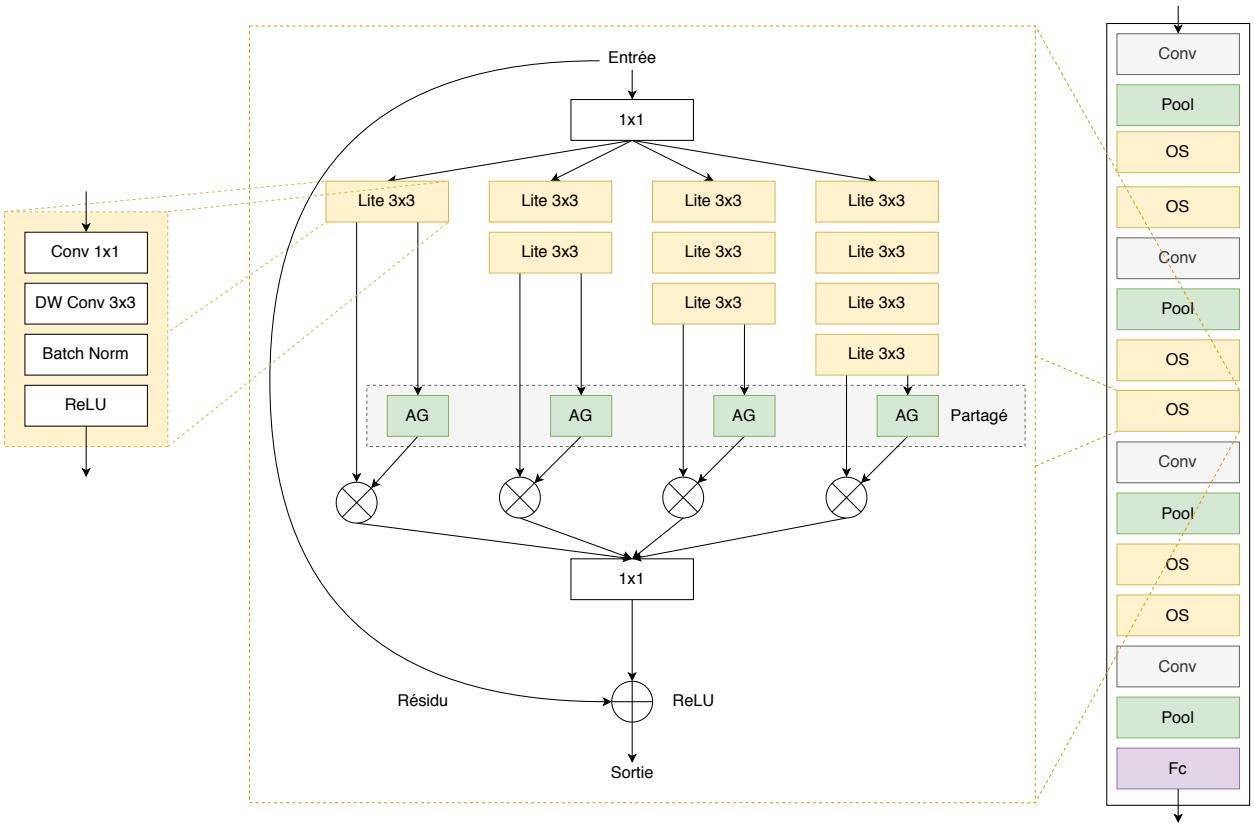

Figure 2.5 Le bloc OS et l'architecture OSNet. Le réseau complet comporte une succession de blocs OS et de blocs convolutifs classiques. Au sein du bloc OS, un canal est constitué de plusieurs couches Lite 3x3 : le nombre de couches détermine l'échelle du canal. Chaque couche Lite 3x3 contient une convolution séparable en profondeur. Les résultats des quatre canaux sont ensuite agrégés par le module AG.

2.2.2 Métrique d'apprentissage

Deux métriques d'apprentissage sont principalement employées en ReID : la fonction de coût de classification et la fonction de coût triplet. Plus récemment, une troisième métrique a été

introduite : le désentrelacement des caractéristiques par une tâche auxiliaire.

Fonction de coût de classification

On peut envisager le problème de la ReID de personnes comme une tâche de classification, où chaque identité correspond à une classe [28]. On ajoute alors une couche entièrement connectée de classification au réseau de ReID lors de l'entraînement, qui associe à chaque vecteur de caractéristiques la probabilité d'appartenir à chaque classe de l'ensemble d'entraînement (Figure 2.6). Le réseau apprend alors à générer des vecteurs v_{id} qui permettent de prédire correctement les identités.

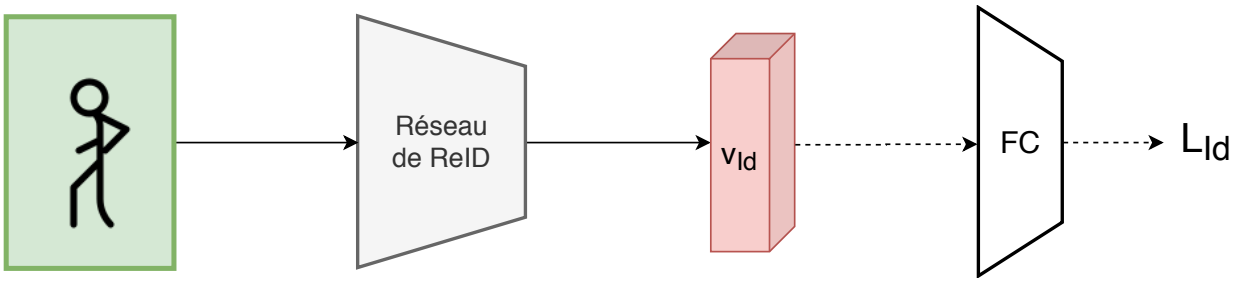

Figure 2.6 Fonction de coût de classification. Une couche entièrement connectée (FC) est ajoutée lors de l'entraînement pour prédire les classes.

Pour un batch de taille n, où les images x_i contiennent les identités y_i, on définit la fonction de coût de classification par entropie croisée $\mathcal{L}_{Id}$ par :

$$\mathcal{L}_{Id} = -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|x_i)) \tag{2.1}$$

Dans cette équation, $(p(y_i|x_i)$ est la probabilité prédite par le réseau que l'image x_i contienne l'identité y_i. Cette probabilité est calculée par la couche de classification, directement à partir du vecteur de caractéristiques. Lors de l'entraînement, la fonction de coût $\mathcal{L}_{Id}$ est minimisée ce qui revient à maximiser la probabilité correcte prédite : lorsque les probabilités sont proches de 1, $\mathcal{L}_{Id}$ se rapproche de 0. Lors de l'évaluation, la couche de classification est retirée pour garder le vecteur de caractéristiques en sortie du réseau de ReID.

Fonction de coût triplet

On peut également aborder la reID comme un problème de classement selon la distance de ReID : la distance entre deux images de la même personne dans l'espace latent des caractéris-

tiques d'identité doit être minimisée, tandis que la distance entre deux images de personnes différentes doit rester grande (Figure 2.7).

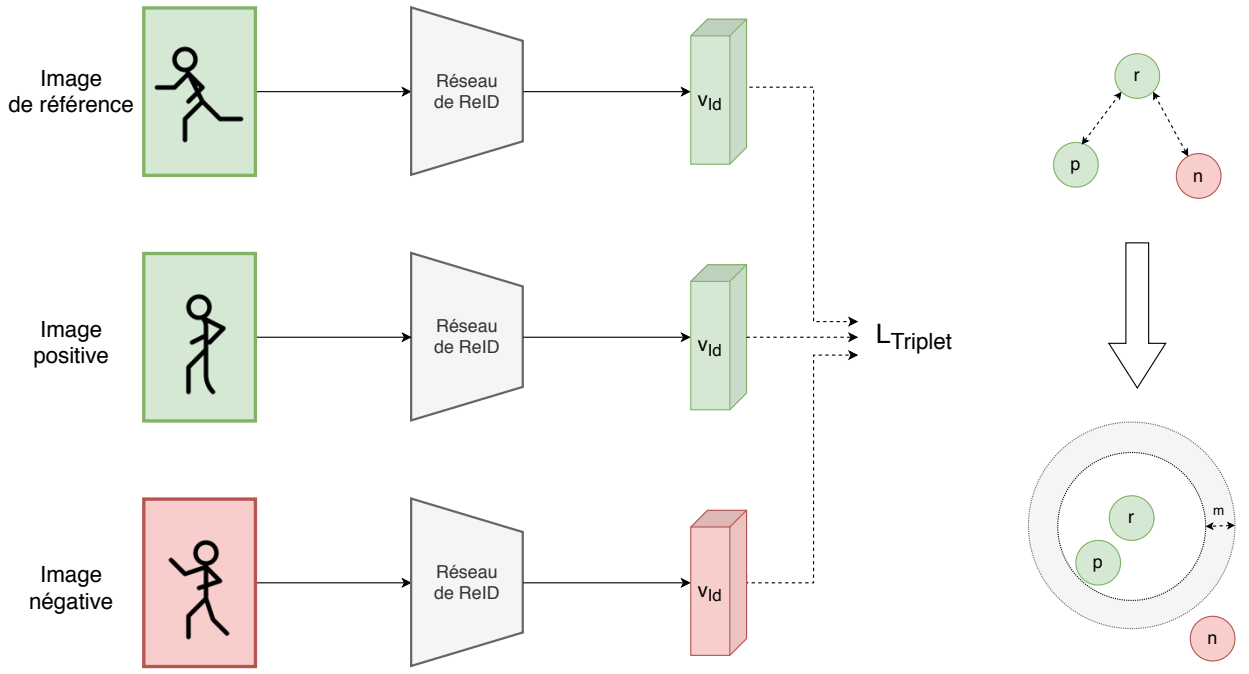

Figure 2.7 Fonction de coût triplets. La distance entre l'image de référence r et l'image positive p contenant la même identité que r est minimisée, tandis que la distance entre r et l'image négative n est maximisée.

Pour implémenter ce critère, un triplet d'images est constitué à partir d'une image de référence r, une image positive p contenant la même identité que r, et une image négative n contenant une identité différente. On fixe une marge m et on définit la fonction de coût triplet $\mathcal{L}_{Tri}$ par :

$$\mathcal{L}_{Tri} = \max(d(r,p) - d(r,n) + m, 0) \tag{2.2}$$

où d désigne la distance dans l'espace latent des vecteurs de caractéristiques d'identité, par exemple la distance euclidienne. Si la distance entre les images de la paire positive est inférieure à la distance entre la paire négative, avec un écart supérieur à la marge m, la fonction de coût triplet $\mathcal{L}_{Tri}$ sera nulle. Dans le cas contraire, la fonction de coût sera positive. Minimiser la fonction de coût triplet $\mathcal{L}_{Tri}$ lors de l'apprentissage revient donc à minimiser la distance entre les images de même identité, tout en assurant que la distance entre deux images d'identité différente reste supérieure à la marge fixée. Dans le cas optimal, toutes les distances entre les paires positives sont inférieures aux distances entre paires négatives, ce

qui est bien l'objectif recherché pour la ReID.

Le choix des images positives et négatives lors de la construction du triplet est très important. La plupart des triplets sont très rapidement correctement classés lors de l'entraînement, il faut donc extraire des triplets difficiles pour optimiser le processus. Ces triplets difficiles vérifient : $d(r,p) > d(r,n)$. Pour ces triplets, la fonction de coût triplet $\mathcal{L}_{Tri}$ ne sera pas nulle ce qui permet de poursuivre l'apprentissage. En revanche, choisir pour chaque image le triplet le plus difficile peut conduire à accorder trop d'importance à quelques images particulières de la base de données, les plus difficiles à identifier, et donc à diminuer la performance sur les images normales constituant la plus grande partie de la base de données.

Le compromis souvent adopté est alors d'extraire des triplets modérément difficiles, qui seront informatifs lors de l'entraînement sans pour autant détourner l'apprentissage vers des cas particuliers. En pratique, un algorithme en ligne de recherche du triplet le plus difficile au sein de chaque batch d'entraînement est adopté [29]. Un batch est constitué en sélectionnant P classes, et K images par classes. Les PK triplets extraits sont des triplets modérément difficiles : ce sont les plus difficiles dans un sous-ensemble de la base de données.

Désentrelacement par une tâche auxiliaire

Le désentrelacement est un moyen d'expliquer les caractéristiques latentes produites par les réseaux de neurones génératifs, en séparant les facteurs de variation indépendants. Comme montré à la Figure 2.8, deux réseaux de neurones encodent des informations indépendantes et complémentaires dans des vecteurs de caractéristiques, qu'un réseau de neurones génératif utilise pour produire des images (tâche auxiliaire). Des fonctions de coût adaptées sur les images produites imposent le désentrelacement selon les facteurs de variations voulus.

Cette méthode a fait l'objet de plusieurs applications récentes en ReID supervisée. Le réseau de ReID encode l'information d'identité dans le vecteur de ReID, tandis qu'un second réseau encode l'information qui n'est pas reliée à l'identité. Ainsi, le réseau FD-GAN [30] essaie de produire un vecteur de ReID complètement indépendant de la pose, qui est le principal facteur de variation à considérer en ReID. Pour cela, le réseau génère des images à partir des informations d'identité encodées par le réseau de ReID, et d'une pose cible encodée par un deuxième réseau. Des fonctions de coûts sur l'image produite réalisent le désentrelacement de la pose et de l'identité.

Le réseau DG-Net [2] propose quant à lui d'apprendre conjointement à générer des images où la structure et l'apparence sont désentrelacées, en utilisant des blocs Adaptive Instance Normalization [31] utilisés dans les méthodes de transfert de style. La séparation de l'infor-

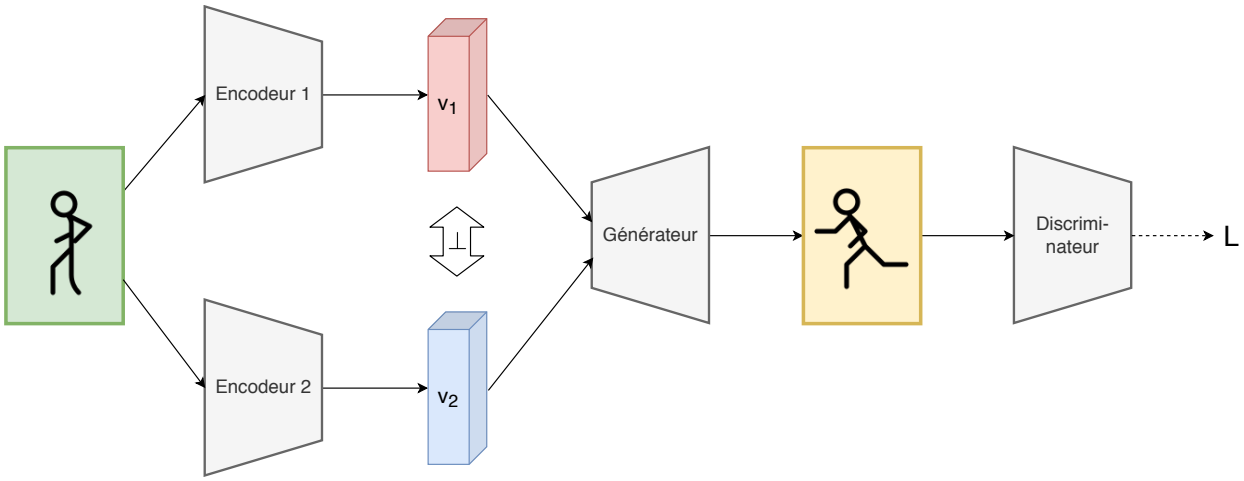

Figure 2.8 Système de réseaux de neurones de désentrelacement. L'information de l'image est encodée dans les deux vecteurs v_1 et v_2 par les deux encodeurs. Le générateur produit une image à partir des vecteurs, sur laquelle le discriminateur est appliqué. Des fonctions de coûts L contraignent le processus de désentrelacement pour faire en sorte que les deux vecteurs contiennent des informations indépendantes et complémentaires, ce qui est ici symbolisé par $\perp$.

mation est donc analogue à la distinction entre style et contenu effectuées dans ces travaux. Le vecteur de structure porte l'information de pose et d'arrière-plan, mais aussi certaines informations reliées à l'identité comme les cheveux, la carrure, le genre, etc. Le vecteur d'apparence contient les informations de couleur et de texture des vêtements. L'identité n'est donc pas totalement désentrelacées de tous les autres facteurs de variation.

En revanche, le modèle ISGAN [5] prend en considération tous les facteurs possibles de variation en désentrelaçant les caractéristiques reliées à l'identité des caractéristiques indépendantes de l'identité. Cette méthode nécessite seulement des paires d'image de même identité. En mélangeant les caractéristiques extraites par deux encodeurs, les caractéristiques communes liées à l'identité sont isolées de toutes les autres caractéristiques.

2.2.3 Ré-agencement

En ReID, la performance est mesurée lors de la tâche de recherche dans une galerie d'images, à partir d'une image requête. Le modèle calcule la distance dans l'espace des caractéristiques entre l'image requête et les images de la galerie, et renvoie une liste triée de résultats. Le ré-agencement [18] consiste à utiliser les distances entre les images de la galerie pour modifier la liste retournée, comme présenté à la Figure 2.9.

Pour chaque image de la galerie, on établit la liste des plus proches images pour la distance de ReID. Si l'image requête se retrouve dans cette nouvelle liste, il y a plus de chances que l'image soit un résultat positif contenant la même personnes. Une métrique combinant la distance de ReID et la distance réciproque est établie, ce qui permet d'améliorer la performance de la tâche de recherche.

Cependant, cette méthode n'a aucune incidence sur la qualité de la représentation latente, elle permet simplement d'augmenter la performance rapportée lors de l'évaluation.

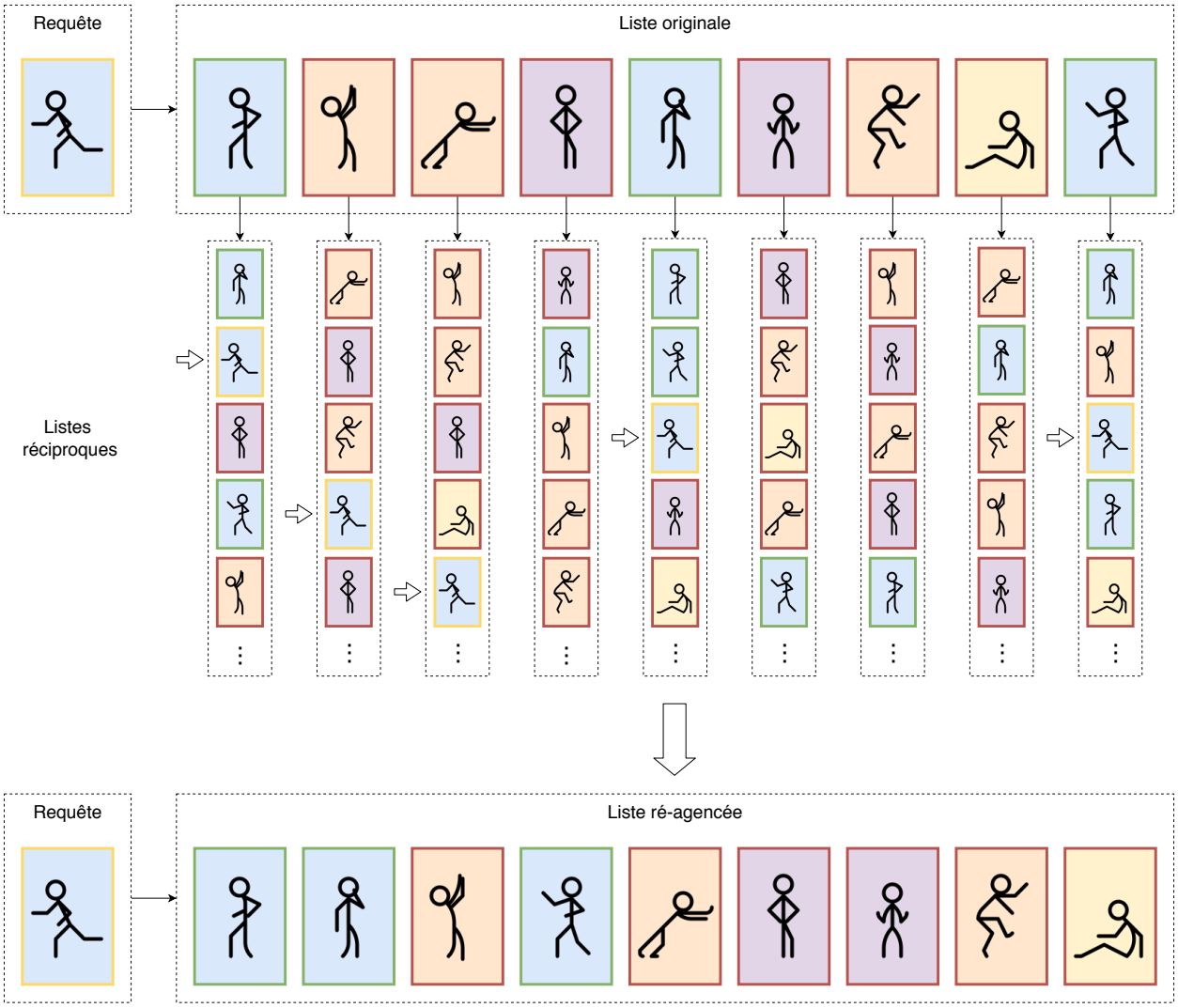

Figure 2.9 Ré-agencement. La liste de résultats originale est établie à partir des vecteurs de ReID. Chaque image de la liste est à son tour considérée comme requête pour construire les listes réciproques. Une nouvelle métrique combinant la distance originale et les distances réciproques permet d'établir la liste ré-agencée, en exploitant les similarités entre les images de la galerie

2.3 Ré-identification de personnes non supervisée

Les méthodes présentées jusqu'à présent supposent l'existence d'une base de données d'images annotées sur laquelle l'entraînement est réalisé. Dans cette partie, nous montrerons que la généralisation de ces modèles à d'autres domaines est la principale limitation pour la mise en oeuvre de la ReID à grande échelle. Pour résoudre ce problème, nous introduirons alors des méthodes de ReID dans le contexte non supervisé.

2.3.1 Limitation des modèles supervisés

L'application de l'apprentissage profond à la tâche de ReID a permis d'améliorer considérablement les performances de la ReID de personnes. Le Tableau 2.2 rassemble les résultats de quelques méthodes présentées précédemment, dans le contexte supervisé. L'évaluation est réalisée par la recherche de l'identité d'une image requête dans une galerie. On mesure la précision au rang 1, c'est à dire le pourcentage de résultats positifs pour la première image de la liste de résultats, ainsi que la Mean Average Precision (mAP) calculée à partir de la courbe de précision-rappel.

Tableau 2.1 Résultats des méthodes de ReID supervisées sur plusieurs bases de données de ReID. On reporte la précision au rang 1 (R1) et la Mean Average Precision (mAP) obtenues par les auteurs [1–5]. Certaines méthodes n'ont pas été testées sur toutes les bases de données par leurs auteurs.

Method	Backbone	Market1501		CUHK03		DukeMTMC		MSMT17	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
HA-CNN [1]	Inception	91.2	75.7	41.7	38.6	80.5	63.8	-	-
DG-Net [2]	ResNet	94.8	86.0	-	-	86.6	74.8	77.2	52.3
OSNet [3]	OSNet	94.8	86.7	72.3	67.8	88.7	76.6	79.1	55.1
MGN [4]	ResNet	95.7	86.9	66.8	66.0	88.7	78.4	-	-
ISGAN [5]	ResNet	95.2	87.1	72.3	68.8	90.0	79.5	-	-

Les performances mesurées sont très bonnes dans le contexte supervisé, dépassant 90% pour la précision au rang 1 sur certaines bases de données. Néanmoins, ces résultats sont obtenus sur l'ensemble de test des bases de données où les réseaux ont été entraînés. En pratique, on dispose cependant rarement de bases de données d'images annotées du domaine d'utilisation des réseaux. En effet, annoter les images est un processus manuel laborieux, qu'il est inenvisageable d'effectuer pour chaque nouveau contexte de déploiement d'un réseau de caméras.

Comparons donc à présent les performances de quelques méthodes dans un cadre non su-

pervisé, dans lequel on ne dispose pas d'annotations d'identité pour des images du domaine. Ce cadre expérimental correspond à une utilisation pratique de la ReID. L'approche la plus simple consiste à appliquer directement un réseau de neurones, entraîné au préalable sur une base de données d'images annotées, sur les images du domaine cible (une autre base de données). Comme le montre le Tableau 2.2, la performance est alors énormément dégradée sur toutes les bases de données. La baisse de performance observée peut être attribuée à l'écart existant entre les domaines d'images, selon le contexte d'utilisation de la ReID. Il est aussi difficile de prédire la performance quand on passe d'un domaine à l'autre : l'apparence visuelle de l'image n'est pas toujours une indication fiable pour estimer la performance des réseaux de neurones dans un nouveau domaine. Les modèles de ReID ne se généralisent donc pas bien à d'autres domaines que celui sur lequel ils ont été entraînés, ce qui motive le développement de modèles de ReID performants dans un contexte non supervisé.

Tableau 2.2 Résultats des méthodes de ReID entraînés sur une base de données de ReID annotée et appliqués directement à une autre base de données. La flèche indique le sens du transfert. On reporte la précision au rang 1 (R1) et la Mean Average Precision (mAP) obtenues par les auteurs [2, 3].

Method	Backbone	DukeMTMC ↓ Market1501		Market1501 ↓ DukeMTMC		MSMT17 ↓ Market1501		MSMT17 ↓ DukeMTMC	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
DG-Net [2]	ResNet	56.1	26.8	42.6	24.2	61.8	33.6	61.9	40.7
OSNet [3]	OSNet	52.2	24.0	44.7	25.9	66.6	37.5	66.0	45.3

2.3.2 Adaptation de domaine non supervisée en ReID

Le but de l'adaptation de domaine non supervisée est d'améliorer la performance d'un modèle entraîné sur une base de données source annotée appliqué à une base de donnée non annotée cible.

Alignement des distributions

D'une manière générale, l'adaptation de domaine est réalisée en alignant la distribution des données des deux domaines, par exemple en minimisant la Maximum Mean Discrepancy (MMD) [32] entre le domaine source et le domaine cible. Cependant, en ReID les classes des domaines sources et cibles sont complètement distinctes puisqu'elles correspondent à des identités différentes de personnes, et on ne retrouve pas les mêmes personnes dans les deux

domaines. Il faut donc développer des méthodes d'adaptation de domaine spécifiques à la ReID.

Le réseau ECN [33] propose de tirer parti des propriétés d'invariance des bases de données de ReID : le réseau est entraîné conjointement sur les deux domaines. Sur le domaine source, une fonction de coût de classification est appliquée. Sur le domaine cible, un module fait respecter trois propriétés d'invariance à travers des fonctions de coût : l'invariance individuelle par laquelle chaque vecteur de ReID doit être suffisamment différent des autres, l'invariance par caméra qui assure que le vecteur de ReID est indépendant du style de la caméra, et l'invariance de voisinage qui fait en sorte que des images similaires produisent des vecteurs de ReID proches.

Au lieu d'adapter l'image entière, des parties de l'image peuvent également être comparés pour contourner le problème de divergence entre les classes. En effet, l'apparence de patchs d'images varie moins entre les domaines que celle d'images entières. Le réseau PAUL [34] apprend à extraire des patchs discriminants pour l'identité et à classifier l'identité à partir de ces patchs.

Une base de données auxiliaire peut également être exploitée pour effectuer un apprentissage multi-annotations [1]. Les images non annotées sont comparées à un ensemble d'images de référence de la base auxiliaire pour créer un vecteur de probabilité tenant lieu d'annotation et servant à l'apprentissage.

Transfert de style

Les réseaux adverses génératifs [26] sont utilisés en ReID non supervisée, avec comme principal objectif le transfert de style entre domaines : l'objectif est de préserver les identités tout en transférant la résolution, les conditions d'éclairage, les arrière-plans, etc.

Le réseau génératif CycleGAN [35] est ainsi appliqué pour réduire l'écart entre les domaines source et cible [6], en appliquant le style du domaine cible sur les images de la base de données source lors de l'entraînement.

Dans le modèle IPGAN [36], le réseau de ReID est entraîné sur les images générées dans le style de chaque caméra du domaine cible, obtenues en utilisant le réseau StarGAN [37] pour réaliser le transfert d'image à image. On considère ainsi que chaque caméra est associée à un domaine qui lui est propre, ce qui permet un transfert plus précis pour chaque vue.

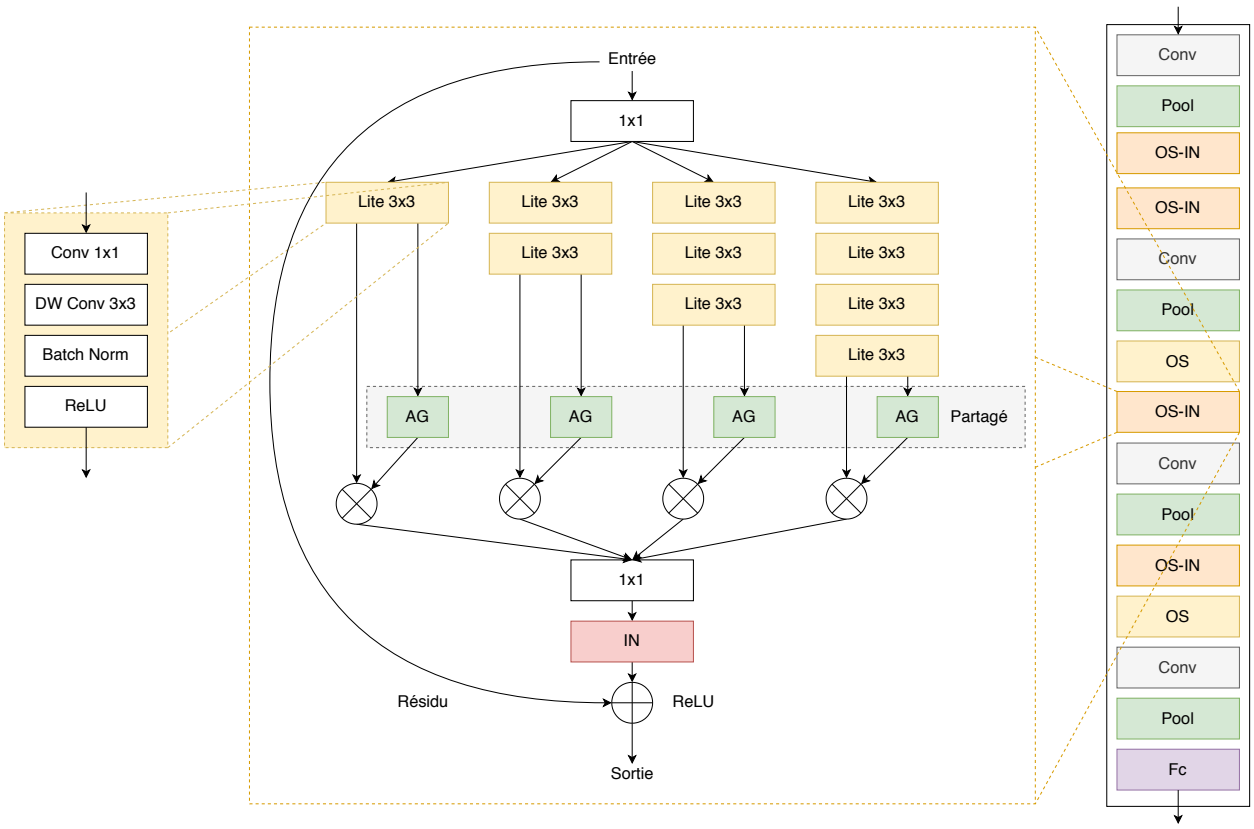

Figure 2.10 Le bloc OS-IN et l'architecture OSNet-AIN. Pour former le bloc OS-IN, une couche de normalisation d'instance IN est incorporée dans le bloc OS, avant l'ajout du résidu. Le réseau intègre des blocs OS et des blocs OS-IN modifiés.

Conception d'architecture

Une autre approche de ce problème consiste à concevoir spécifiquement un réseau de ReID généralisable à d'autres domaines. Le réseau OSNet-AIN [38] propose ainsi de modifier le réseau OSNet [3] pour améliorer la performance lors de l'application à d'autres bases de données.

Pour cela, des couches de normalisation d'instance [39] sont incorporées dans certains blocs OS pour former le bloc OS-IN illustré à la Figure 2.10. Ces couches normalisent chaque instance en standardisant la moyenne et l'écart-type, contrairement à la normalisation par batch qui opère au niveau du mini-batch. Cela a pour effet de réduire l'influence des conditions d'éclairage, de l'environnement et des caractéristiques de la caméra sur le vecteur de ReID produit.

L'architecture finale OSNet-AIN est obtenue par une recherche dans l'espace des architectures

de réseaux, pour trouver la meilleure combinaison possible des blocs OS et OS-IN. Sans adaptation supplémentaire, l'architecture modifiée OSNet-AIN obtient de bons résultats sur les domaines non annotés.

Désentrelacement

Si plusieurs méthodes récentes utilisent avec succès le désentrelacement pour la ReID de personnes supervisée, son application pour la ReID de personnes non supervisée est encore peu explorée.

Pourtant, des travaux récents ont montré que le désentrelacement nécessite peu d'annotations : la modèle de désentrelacement robuste de pose et d'apparence proposée en [40] requiert ainsi seulement des paires d'images de même identité. Cette méthode introduit un réseau de classification dont le but est d'estimer la quantité minimale de régularisation nécessaire sur l'information mutuelle partagée par les vecteurs de caractéristiques pour effectuer le désentrelacement : trop de contraintes diminue l'information portée par les vecteurs, alors que trop peu de contrainte ne permet pas de séparer l'information. Les vecteurs obtenus ainsi sont à la fois indépendants et informatifs.

Dans un contexte non supervisé, on peut citer le modèle PDA-Net [41], qui applique un désentrelacement guidé par la pose. Le réseau apprend à extraire des caractéristiques de ReID indépendantes du domaine, en générant des images inter-domaines qui mêlent la pose et le contenu de deux images de deux domaines différents.

La méthode que nous proposons réalise également le désentrelacement des caractéristiques latentes dans le contexte non supervisé, mais ne nécessite pas d'extraction de pose.

Synthèse

Ainsi, plusieurs approches ont été envisagées pour réaliser l'adaptation de domaine en ReID. Les techniques d'alignement des distributions reposent sur une construction précise de caractéristiques discriminantes sur le domaine cible à partir d'hypothèses sur la structure et les invariances du domaine. Elles sont donc efficaces pour les domaines cibles sur lesquelles elles ont été conçues, mais leur performance risque d'être moins bonne sur d'autres domaines.

Les méthodes de transfert de style et de désentrelacement sont plus générales puisqu'elles ne reposent pas sur une connaissance a priori du domaine cible. Le transfert de style opère au niveau des images, alors que le désentrelacement agit directement dans l'espace latent sur les vecteurs de caractéristiques. Comme démontré dans le contexte supervisé, cela permet aux méthodes de désentrelacement d'orienter efficacement l'apprentissage pour obtenir des

vecteurs de ReID performants.

Enfin, les améliorations apportées dans l'architecture OSNet-AIN pour l'extraction de caractéristiques plus universelles sont complémentaires : elles peuvent par exemple être combinées avec les techniques de transfert de style ou de désentrelacement.

C'est pourquoi nous proposons une méthode d'adaptation de domaine basée sur le désentrelacement des caractéristiques latentes, et qui utilise l'architecture OSNet-AIN pour le réseau de ReID.

CHAPITRE 3 MÉTHODOLOGIE

3.1 Aperçu du système de réseaux de neurones

La méthode d'adaptation de domaine proposée est basée sur le désentrelacement des vecteurs de ReID dans le domaine cible. Tout comme [5], nous proposons de désentrelacer les caractéristiques liées à l'identité de tous les autres facteurs de variation des caractéristiques latentes qui ne sont pas liées à l'identité, que l'on appellera caractéristiques de contenu.

Dans le cadre de cette nouvelle tâche, le réseau de ReID est incorporé dans un système de réseaux antagonistes génératifs [26]. En apprenant à générer des images réalistes dans le domaine cible qui préservent l'identité et le contenu par cette tâche auxiliaire, le réseau de ReID se spécialise pour l'extraction de caractéristiques liées à l'identité et adaptées à ce domaine.

Cette tâche annexe à la classification ne nécessite pas l'annotation des identités de l'ensemble des données cible. Elle utilise des paires d'image du domaine cible présentant avec une haute probabilité la même identité, que l'on extrait en utilisant le réseau de ReID pré-entraîné sur la base de donnée source. Nous proposons une méthode pour filtrer les mauvaises paires pour créer une base de donnée contenant peu de bruit, dans le domaine cible.

Le générateur prend en entrée des vecteurs de caractéristiques extraits d'une paire d'image. Il apprends à générer des images réalistes préservant à la fois l'identité et le contenu. Des contraintes portant aussi bien sur l'espace latent de caractéristiques que sur les images générées dans le domaine cible induisent l'adaptation de domaine des caractéristiques d'identité utilisées dans le processus génératif.

La méthode de désentrelacement fait intervenir un module de génération d'images (G), un module discriminateur (D) et deux modules d'encodage d'information (E_{id} et E_C), comme le montre la Figure 3.1.

3.2 Encodeur d'identité

Notre système de réseau réalise le désentrelacement des caractéristiques d'identité des autres caractéristiques qui ne sont pas liées à l'identité. Pour ce faire, il inclut deux encodeurs qui génèrent les deux espaces latents désentrelacés de caractéristiques. Le réseau de ReID E_{Id} est utilisé comme un encodeur de caractéristiques relatives à l'identité des personnes, qui extrait le vecteur de caractéristiques d'identité v_{Id}. E_{Id} n'encode que partiellement l'image puisque

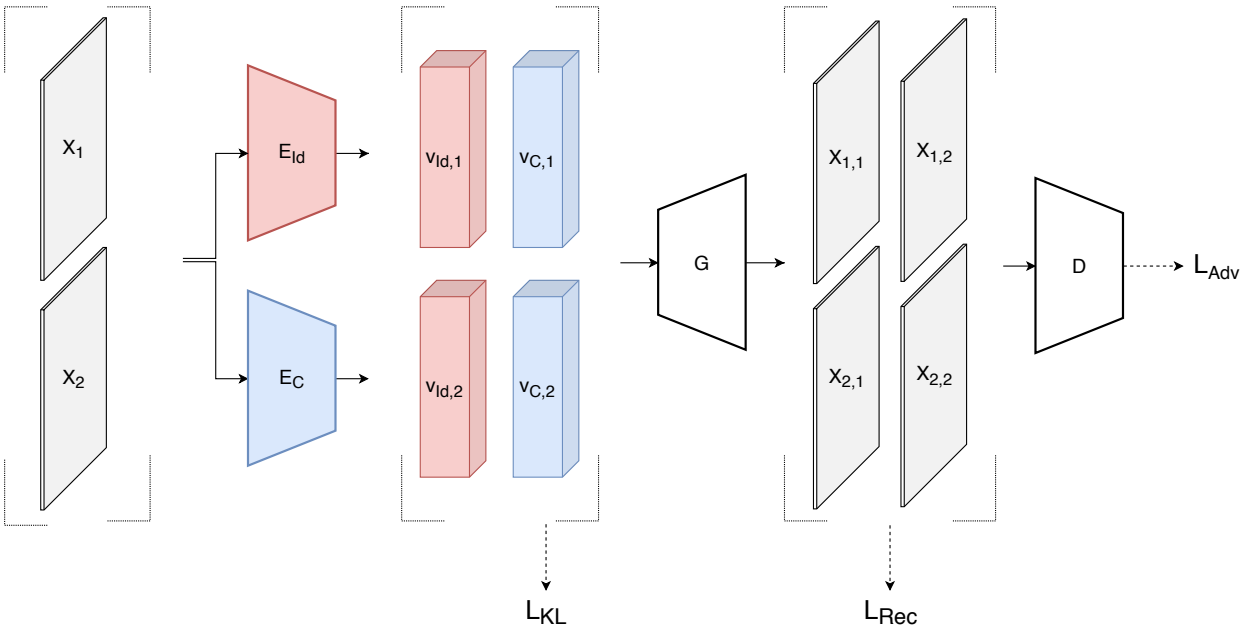

Figure 3.1 Aperçu du système de réseaux UD-GAN (Unsupervised Disentanglement GAN). L'encodeur d'identité E_{Id} et l'encodeur de contenu E_C extraient respectivement les vecteurs d'identité v_{Id} et les vecteurs de contenu v_C des paires d'images du domaine cible X_1 et X_2, présentant la même identité. Le générateur G produit quatre images en échangeant les caractéristiques d'identité et de contenu dans les images générées. Le discriminateur D fait la distinction entre les images générées et les véritables images du domaine cible.

les caractéristiques indépendantes de l'identité ne se retrouvent pas dans v_{Id}.

N'importe quelle architecture de réseau de ReID peut être intégrée dans notre système de réseaux, puisque le processus de désentrelacement est réalisé par des modules additionnels. Ainsi, notre approche est complémentaire des méthodes visant à l'optimisation des architectures des réseaux de ReID pour la construction d'une représentation discriminante dans l'espace des caractéristiques, introduites à la sous-section 2.2.1.

E_{Id} est entraîné pour la ReID de personnes sur le domaine source annoté, par une fonction de coût de classification qui présente de meilleures performances que la fonction de coût triplet dans les travaux récents [28]. Sur cette base de données, la ReID de personnes supervisée est une tâche de classification avec de multiples classes : chaque identité correspond à une classe. Pour cette tâche, nous ajoutons une couche entièrement connectée de classification au réseau de ReID (voir Figure 3.2) et nous l'entraînons pour associer les vecteurs de ReID aux classes adéquates en utilisant une perte par entropie croisée softmax.

Les annotations sont lissées [42] pour prévenir le sur-apprentissage. Pour cela, on encode

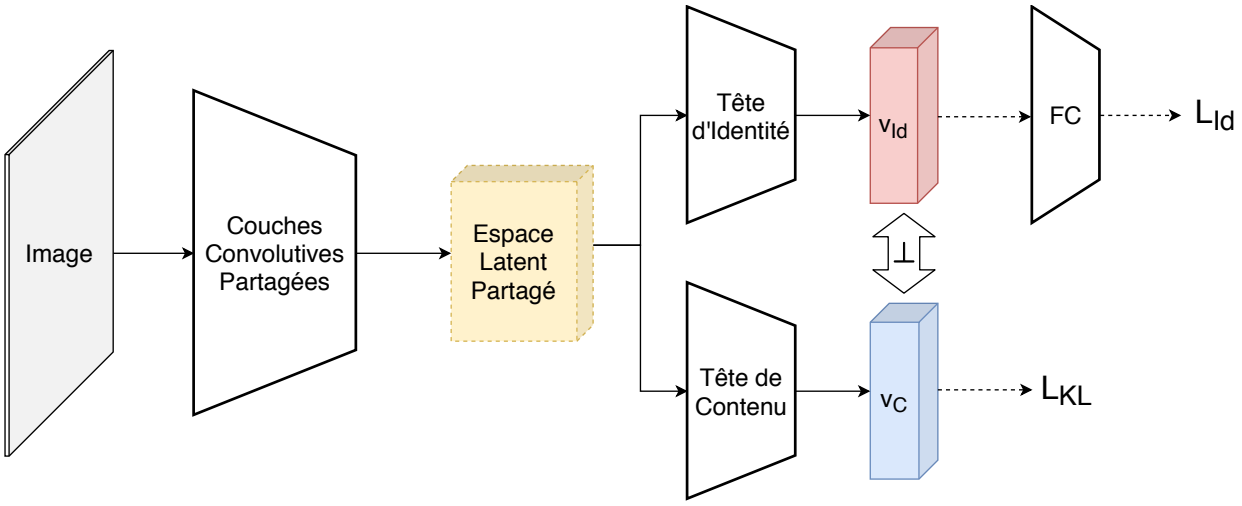

Figure 3.2 Architecture des encodeurs d'identité et de contenu. Le réseau de ReID est composé des couches partagées et de la tête d'identité, auxquelles on ajoute une couche entièrement connectée de classification pour calculer $\mathcal{L}_{Id}$. Les têtes d'identité et de contenu sont appliqués à la carte de caractéristiques commune pour produire les vecteurs v_C et v_{Id} dans les deux espaces latents désentrelacés. Ces deux vecteurs sont encouragés à porter de l'information indépendante et complémentaire, symbolisé par $\perp$ sur la figure.

l'annotation d'identité dans le vecteur y_{hot} par l'encodage one-hot : le vecteur d'annotation contient seulement des 0, et un 1 à l'indice correspondant à l'identité annotée. Le vecteur d'annotations y lissé est alors défini par :

$$y = (1 - \alpha) * y_{hot} + \frac{\alpha}{n}, \tag{3.1}$$

où n est le nombre total de classes et α une petite constante. Ainsi, la probabilité associée à la vraie identité est légèrement diminuée, et le résidu est reporté sur toutes les autres classes possibles.

Nous définissons alors la fonction de coût d'identité $\mathcal{L}_{Id}$ sur le domaine source :

$$\mathcal{L}_{Id} = -\sum_{k=1}^{n} y_k \log(\frac{\exp(x_k)}{\sum_{j=1}^{n} \exp(x_j)}), \tag{3.2}$$

où x est le vecteur de taille n contenant les prédictions pour chaque classe. Le coût est minimum si le vecteur prédit est proche du vecteur d'annotations y : si pour l'indice i où y_i est proche de 1, x_i l'est aussi et les autres valeurs de x sont petites, alors on obtient pour $\mathcal{L}_{Id}$ une valeur proche de $\log(1) = 0$.

3.3 Encodeur de contenu

On introduit un second encodeur pour prendre en charge les caractéristiques qui ne sont pas liées à l'identité, l'encodeur de contenu E_C. Comme proposé dans [5], le réseau E_C est conçu comme une deuxième tête ajoutée au réseau de ReID (Figure 3.2). Les deux réseaux partagent leurs premières couches de convolutions. Le produit de ces couches est une carte de caractéristiques commune aux deux encodeurs, qui contient toute l'information extraite de l'image. Les dernières couches de E_{Id} sont dupliquées et entraînées pour l'extraction de caractéristiques indépendantes de l'identité.

Le vecteur d'informations de contenu v_C produit par l'encodeur de contenu E_C vient compléter les informations de v_{Id}, ce qui permet de représenter de la manière la plus exhaustive possible l'image originale dans l'espace latent, à partir des caractéristiques extraites par les deux encodeurs. On cherche ainsi à définir un deuxième espace latent de caractéristiques, orthogonal à celui des caractéristiques relatives à l'identité.

La conception des deux encodeurs sous forme de deux têtes rajoutées à un socle commun permet de mieux maîtriser l'apprentissage : les deux têtes disposent en entrée des mêmes informations dans l'espace latent intermédiaire partagé. L'information est alors séparée par les deux têtes lors de l'apprentissage, dans les vecteurs v_C et v_{Id}.

Contrairement à E_{Id}, E_C est entraîné directement dans le domaine cible en minimisant la quantité d'information partagée par les vecteurs v_C et v_{Id}. Sans contrainte additionnelle sur l'encodeur de contenu E_C, on observe que le générateur a tendance à uniquement tirer parti des informations du vecteur de contenu pour générer des images. En effet, contrairement à E_{Id} dont l'entraînement est directement contraint par la fonction de coût d'identité $\mathcal{L}_{Id}$, l'encodeur de contenu est libre d'extraire n'importe quelle information des images pour produire v_C. Pour limiter la quantité d'information que peut contenir v_C, nous introduisons la fonction de coût divergence Kullback-Leibler (KL) sur l'encodeur E_C qui contraint la distribution des vecteurs de contenu v_C à rester proche d'une distribution normale :

$$\mathcal{L}_{KL} = D_{KL}(v_C \parallel \mathcal{N}(0,1)), \tag{3.3}$$

où la divergence KL est définie pour deux distributions p et q par :

$$D_{KL}(p \parallel q) = -\int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx. \tag{3.4}$$

Comme montré en [43], l'application de la fonction de coût divergence KL diminue la quantité

d'information reliée à l'identité contenue dans le vecteur v_C. En effet, le champ des distributions et donc l'information portée par le vecteur v_C est limitée. Or, le générateur a besoin des informations sur le contenu de l'image pour satisfaire les exigences de reconstruction. Ces informations ne sont pas portées par le vecteur d'identité v_{Id}, contraint par la fonction de coût d'identité $\mathcal{L}_{Id}$. Le vecteur v_C est donc encouragé à prendre en charge ces informations, au détriment des informations redondantes auxquelles le générateur a déjà accès. Cela incite donc les deux encodeurs à produire des informations complémentaires, réalisant ainsi le processus de désentrelacement.

Nous implémentons $\mathcal{L}_{KL}$ en utilisant l'astuce de reparamétrisation [44, 45] : minimiser la fonction de coût divergence KL revient à minimiser la fonction de coût :

$$\mathcal{L}_{KL} = \frac{1}{2}\sum_{i=1}^{N}(\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2) - 1), \tag{3.5}$$

où N est la longueur de v_C et μ et σ sont la moyenne et l'écart-type de v_C. Tout comme [44], v_C est calculé par : $v_C = \mu + v * \sigma$, où v est un vecteur aléatoire tiré selon la distribution normale : $v \sim \mathcal{N}(0, 1)$

3.4 Générateur et discriminateur

À partir d'une paire d'images du domaine cible X_1 et X_2 présentant la même identité (avec une haute probabilité puisque les annotations du domaine cible ne sont pas disponibles), les deux encodeurs extraient les vecteurs de ReID $v_{Id,1}$ et $v_{Id,2}$ ainsi que les vecteurs de contenu $v_{C,1}$ et $v_{C,2}$.

Le générateur G produit quatre images à partir de la paire d'images en échangeant les caractéristiques d'identité et de contenu dans les images générées (Figure 3.1). Par exemple, $X_{1,2}$ contiendra la même identité que X_1, tout en présentant le même contenu que X_2.

Le générateur G est mis en compétition dans un schéma adverse avec le discriminateur D, qui assure le réalisme des images générées dans le nouveau domaine. Le discriminateur est conçu selon le modèle PatchGAN [46] pour faire la distinction entre les véritables images et les images générées. Cette méthode opère sur des parties de l'image en appliquant une convolution, et effectue la moyenne des résultats locaux pour produire le score global correspondant à la probabilité que l'image soit réelle. La fonction de coût adverse $\mathcal{L}_{Adv}$ résultante affecte les deux encodeurs, le discriminateur et le générateur :

$$\mathcal{L}_{Adv} = \sum_{i \in \{1,2\}} \log D(X_i) + \sum_{i,j \in \{1,2\}} \log(1 - D(G(v_{Id,i}, v_{C,j})). \tag{3.6}$$

L'extraction des vecteurs de caractéristiques et la génération d'images sont également contraintes par la fonction de coût de reconstruction $\mathcal{L}_{Rec}$ portant sur les deux encodeurs et le générateur :

$$\mathcal{L}_{Rec} = \sum_{i,j \in \{1,2\}} \|X_{i,j} - X_i\|_1. \tag{3.7}$$

Lorsque $i = j$, l'image générée $X_{i,i}$ et l'image originale X_i devraient être identiques : nous voulons nous assurer que l'information extraite par les deux encodeurs est exhaustive. De la même manière, dans le cas où $i \neq j$, l'image reconstruite $X_{i,j}$ devrait alors être similaire à l'image X_j d'où $v_{C,j}$ a été extrait, puisque le réseau de ReID devrait encoder la même information d'identité pour X_i et X_j.

3.5 Fonctions de coût

Nous avons défini deux objectifs d'entraînement conjoints, un par domaine. Sur le domaine source annoté, le réseau de ReID est entraîné à extraire des caractéristiques discriminantes pour l'identité, par le biais de la fonction de coût $\mathcal{L}_{Id}$.

Sur le domaine cible non annoté, la fonction de coût $\mathcal{L}_{Id}$ ne peut pas être utilisée puisqu'elle nécessite toutes les annotations d'identité sur l'ensemble d'entraînement : les paires extraites ne suffisent pas à classifier chaque image selon l'identité. Le système de réseau génératif est entraîné à extraire des caractéristiques désentrelacées et à générer des images, en optimisant la fonction de coût totale sur le domaine cible définie par :

$$\mathcal{L}_{Cible} = \lambda_{Rec}\mathcal{L}_{Rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{Adv}\mathcal{L}_{Adv}, \tag{3.8}$$

où les poids λ_{Rec}, λ_{KL} et λ_{Adv} sont des hyperparamètres de notre méthode, qui établissent un équilibre entre les différents objectifs d'apprentissage. Par exemple, le poids λ_{KL} doit être bien choisi pour régulariser la quantité d'information portée par le vecteur de contenu v_C. Si λ_{KL} est trop petit, des caractéristiques liées à l'identité se retrouveront dans v_C, alors que s'il est trop grand, v_C sera trop limité et le vecteur d'identité v_{Id} portera également des caractéristiques de contenu.

En définissant $\mathcal{L}_{Cible}$, notre objectif est d'optimiser les vecteurs d'identité dans le domaine cible : E_{Id} est modifié pour produire des meilleures images dans le domaine cible, ce qui a

pour effet d'améliorer le vecteur généré car les caractéristiques de contenu auront bien été séparées de celles d'identité.

3.6 Extraction de paires non supervisée

Le processus génératif nécessite des paires d'image contenant la même identité dans le domaine cible. Puisque notre méthode est non supervisée, nous n'utilisons pas d'annotations d'identité. C'est pourquoi nous appliquons le réseau de ReID, pré-entraîné au préalable sur le domaine source, sur des images du domaine cible. Dans nos expérimentations, le domaine cible correspond à l'ensemble d'entraînement de la base de donnée cible, sans tenir compte des annotations. Nous sélectionnons les meilleures associations selon la distance de ReID pour chaque image pour générer des paires ayant une grande probabilité de correspondre à la même identité.

Afin de réduire le bruit, nous introduisons un critère pour filtrer les mauvaises paires potentielles inspiré des techniques de ré-agencement [18]. Nous utilisons chaque image de l'ensemble d'entraînement de la base de données cible comme image requête. Comme illustré à la Figure 3.3, nous établissons pour chaque image requête, la liste des images candidates de la base de données les plus semblables, triée selon la distance de ReID. Pour une image requête, si sa meilleure image candidate n'inclut pas l'image requête dans une de ses cinq premières images candidates, nous écartons cette paire des paires d'entraînement. En d'autres mots, les deux images de la paire doivent se sélectionner mutuellement selon la distance de ReID.

Pour rassembler plus de données, les images requête qui ne trouvent pas de paires parmi les cinq premiers candidats sont inclues dans des paires contentant deux fois la même image, reproduisant ainsi une situation où les deux images de la paire seraient très similaires. Cette étape de post-traitement conduit à une base de donnée de paires présentant moins d'erreurs, où on conduira l'entraînement de désentrelacement dans le domaine cible.

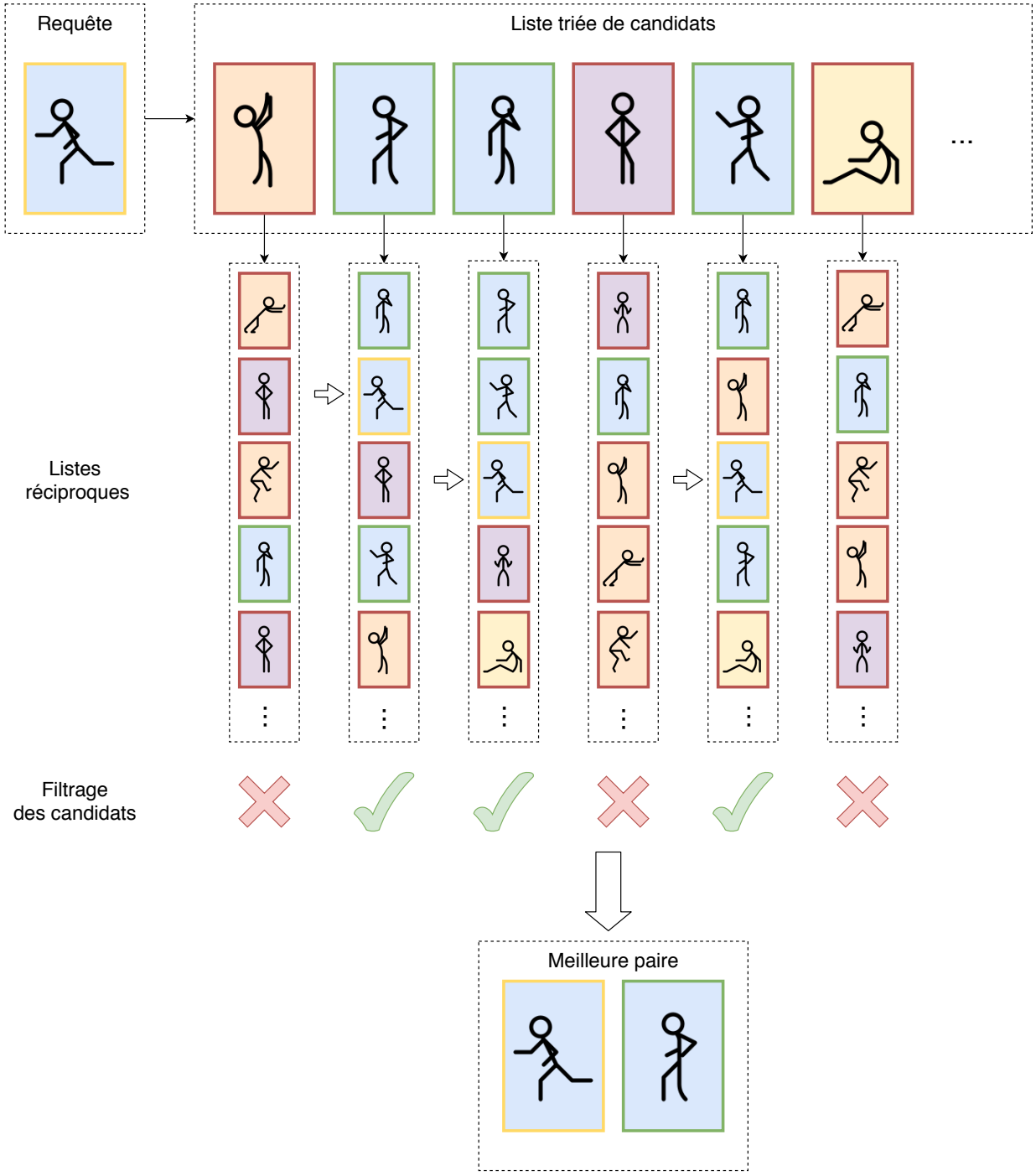

Figure 3.3 Extraction de paires d'images contenant avec une grande probabilité la même identité, dans un contexte non supervisé. La liste de candidats est établie selon la distance de ReID, en appliquant le réseau de ReID pré-entraîné sur la base de données source. Nous écartons les images qui ne comportent pas l'image requête parmi les premières images de sa liste réciproque. La meilleure paire sélectionnée est celle qui minimise la distance de ReID, après filtrage.

3.7 Entraînement

L'entraînement des réseaux de neurones s'effectue en trois étapes.

3.7.1 Étape 1 : Pré-entraînement du réseau de ReID de référence

Le réseau de ReID E_{Id} est entraîné dans un premier temps en utilisant la fonction de coût $\mathcal{L}_{Id}$ sur le domaine source annoté. Il apprend à extraire des caractéristiques discriminantes tout en étant insensible aux éléments de l'image indépendants de l'identité comme l'arrière-plan, la pose ou l'éclairage. Nous avons adopté l'architecture OSNet-AIN [38], qui est la variante de l'architecture OSNet [3] optimisée pour l'adaptation de domaine. Ce choix nous assure que la performance du réseau de ReID sera satisfaisante sur le domaine cible dès la fin de l'étape de pré-entraînement sur le domaine source, sans avoir été exposé au nouveau domaine.

3.7.2 Étape 2 : Pré-entraînement des modules génératifs

Nous appliquons sur le domaine cible le réseau de ReID pré-entraîné sur le domaine source, pour y extraire les paires d'images. Nous fixons l'encodeur d'identité E_{Id}, et entraînons uniquement l'encodeur de contenu E_C, le générateur G et le discriminateur D sur ces nouvelles paires d'images, par le biais de la fonction de coût $\mathcal{L}_{Cible}$.

3.7.3 Étape 3 : Entraînement généralisé

Finalement, tous les modules sont entraînés conjointement sur les deux domaines, en utilisant $\mathcal{L}_{Id}$ sur le domaine source et $\mathcal{L}_{Cible}$ sur le domaine cible. Nous alternons les batchs provenant de chacun des deux domaines. Ici, le domaine source impose une régularisation pour assurer que le réseau de ReID E_{Id} ne perde pas ses propriétés discriminantes en étant optimisé pour la génération d'images dans le domaine cible.

CHAPITRE 4 RÉSULTATS

4.1 Implémentation

Pour implémenter notre méthode, nous avons sélectionné une architecture adaptée pour chaque réseau de neurones, et nous fixons les paramètres de l'entraînement ainsi que les hyperparamètres du modèle.

4.1.1 Architecture

Les architectures choisies pour les différents modules de notre système de réseaux de neurones incorporent les composants les plus performants des travaux récents en vision par ordinateur.

Encodeurs d'identité et de contenu

Les encodeurs d'identité et de contenu adoptent l'architecture OSNet-AIN [38], et comportent notamment les blocs OS et OS-IN présentés précédemment (Figure 2.5 et Figure 2.10). Comme montré à la Figure 4.1, l'architecture est modifiée en dupliquant les derniers blocs convolutifs ainsi que la couche finale de pooling pour créer deux têtes distinctes correspondant à chaque encodeur. Les premières couches sont partagées par les deux encodeurs.

Nous complétons la tête d'identité par une couche entièrement connectée, pour réaliser l'entraînement sur le domaine source avec la fonction de coût entropie croisée. Nous ajoutons également deux couches entièrement connectées à la tête de contenu, pour calculer les coefficients de l'astuce de reparamétrisation intervenant dans le calcul de la divergence KL [44, 45].

Les paramètres des couches sont détaillés dans le Tableau 4.1.

Générateur

Le générateur est construit selon le modèle de DCGAN [47]. Tel qu'illustré à la Figure 4.2, il est composé de six blocs de déconvolution. Chaque bloc contient une couche convolutive transposée [47] pour augmenter la résolution spatiale des informations, ainsi que les couches suivantes : normalisation de batch [48], leaky ReLU [49] et dropout [50].

Les auteurs de DCGAN ont montré que ce schéma stabilise l'entraînement du réseau génératif, et le rend facilement explicable : les variations de l'espace latent sont répercutées de manière structurée dans l'image générée.

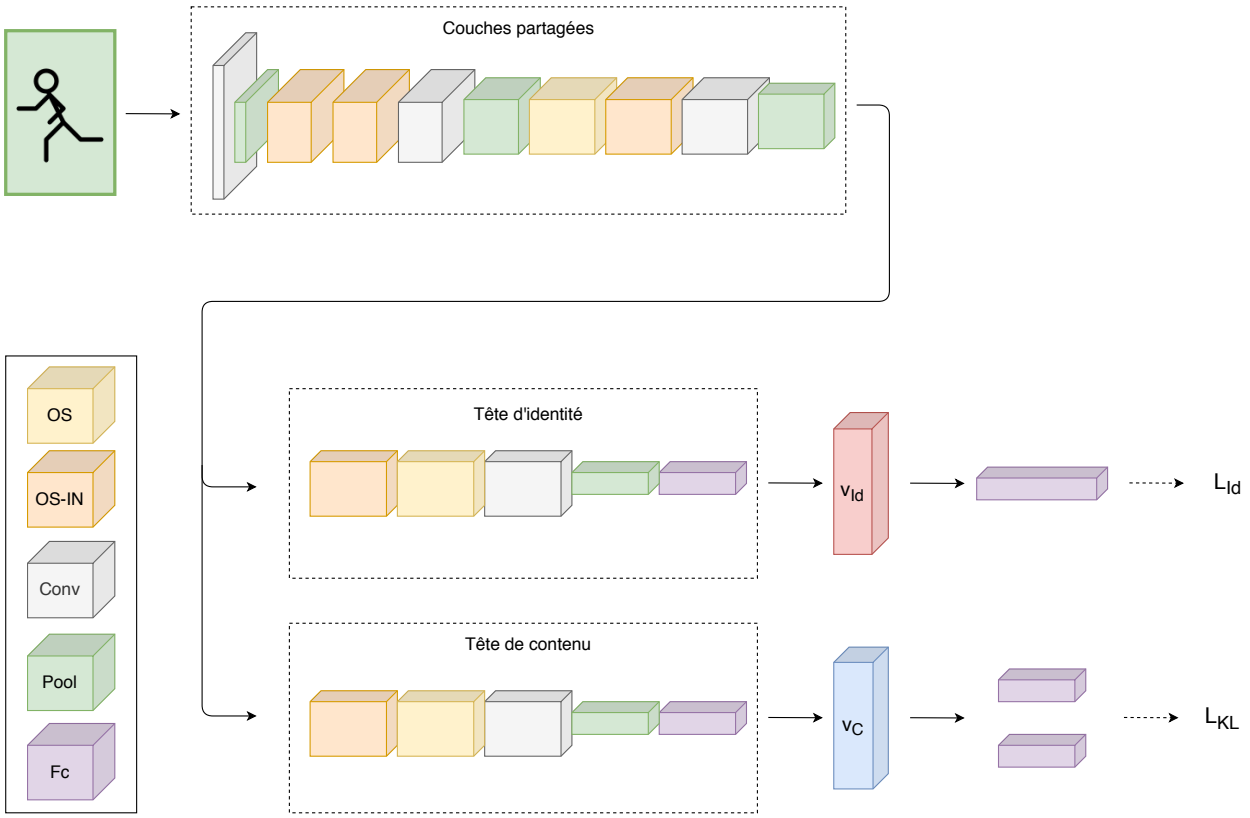

Figure 4.1 Architecture des encodeurs d'identité et de contenu, basée sur l'architecture OSNet-AIN. Les premières couches partagées par les deux encodeurs sont appliquées à l'image. Les deux têtes encodent les vecteurs d'identité v_{Id} et de contenu v_C. Afin de calculer les fonctions de coûts $\mathcal{L}_{Id}$ et $\mathcal{L}_C$ intervenant dans l'apprentissage, des couches entièrement connectées (FC) sont appliquées sur les deux vecteurs.

L'architecture du générateur est détaillée dans le Tableau 4.2.

Tableau 4.1 Architecture détaillée des encodeurs de contenu et d'identité, basée sur l'architecture OSNet-AIN.

Module	Bloc	Couche	Taille en sortie
Couches partagées	Conv-1	7x7 Convolution	[128,64,64]
		3x3 Max Pooling	[64,32,64]
	Conv-2	OS+IN Block	[64,32,256]
		OS+IN Block	[64,32,256]
	Transition	1x1 Convolution	[64,32,256]
		2x2 Average Pooling	[32,16,256]
	Conv-3	OS Block	[32,16,384]
		OS+IN Block	[32,16,384]
	Transition	1x1 Convolution	[32,16,384]
		2x2 Average Pooling	[16,8,384]
Tête dupliquée	Conv-4	OS+IN Block	[16,8,512]
		OS Block	[16,8,512]
	Conv-5	1x1 Convolution	[16,8,512]
	Gap	Global Average Pooling	[1,1,512]
	Fc	Fully Connected	[1,1,512]
Couche de classification	Fc_{Id}	Fully Connected	[1,1,$n_{classes}$]
Couches de reparamétrisation	Fc_{μ}	Fully Connected	[1,1,512]
	Fc_{σ}	Fully Connected	[1,1,512]
Nombre de paramètres		2,2M	

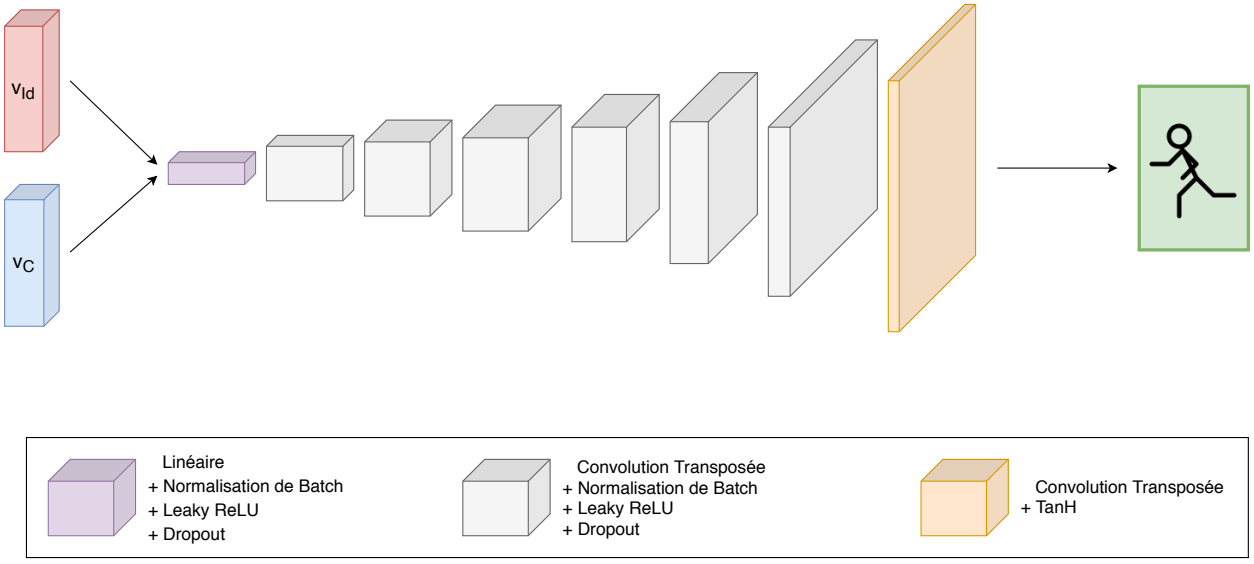

Figure 4.2 Architecture du générateur. Le réseau prend en entrée les vecteurs de caractéristique de contenu et d'identité, et génère l'image en appliquant des blocs de déconvolution (en gris).

Tableau 4.2 Architecture détaillée du générateur.

Bloc	Couche	Taille en sortie	Nombre de paramètres
Linear	Linear Batch Normalization Leaky RelU + Dropout	[512,1,1]	524,800 1024 0
Conv-1	6x2 Transposed Convolution Batch Normalization Leaky RelU + Dropout	[512, 6, 2]	3,145,728 1024 0
Conv-2	4x4 Transposed Convolution Batch Normalization Leaky RelU + Dropout	[512, 12, 4]	4,194,304 1024 0
Conv-3	4x4 Transposed Convolution Batch Normalization Leaky RelU + Dropout	[512, 24, 8]	4,194,304 1024 0
Conv-4	4x4 Transposed Convolution Batch Normalization Leaky RelU + Dropout	[256, 48, 16]	2,097,152 1024 0
Conv-5	4x4 Transposed Convolution Batch Normalization Leaky RelU + Dropout	[128, 96, 32]	524,288 1024 0
Conv-6	4x4 Transposed Convolution Batch Normalization Leaky RelU + Dropout	[64, 192, 64]	131,072 1024 0
Conv-7	4x4 Transposed Convolution Tanh	[3, 384, 128]	3,072 0
Total			14,8M

Discriminateur

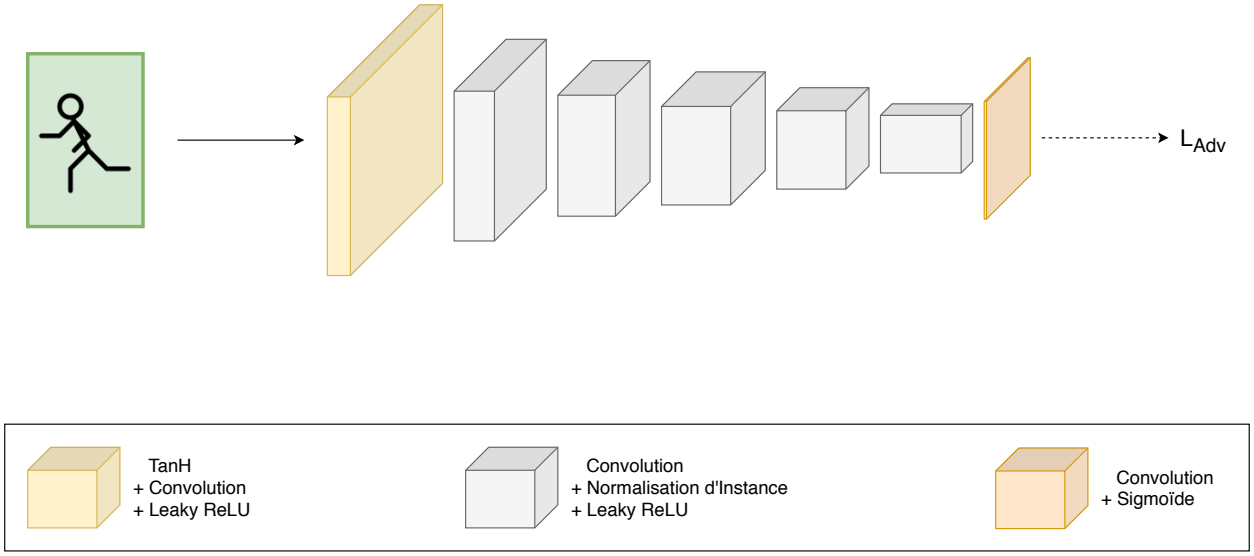

Figure 4.3 Architecture du discriminateur. La fonction de coût adverse $\mathcal{L}_{Adv}$ est calculée à partir de l'image en entrée. Les blocs convolutifs intègrent notamment des couches de normalisation d'instance.

Le discriminateur est basé sur le réseau PatchGAN [46]. Comme le montre la Figure 4.3, il consiste en une série de sept blocs composés d'une couche convolutive suivie par : normalisation d'instance [39] et leaky ReLU [49].

L'architecture est détaillée dans le Tableau 4.3.

Tableau 4.3 Architecture détaillée du discriminateur.

Bloc	Couche	Taille en sortie	Nombre de paramètres
Conv-1	Tanh 4x4 Convolution Leaky ReLU	[32, 192, 64]	0 1,568 0
Conv-2	4x4 Convolution Instance Normalization Leaky ReLU	[64, 96, 32]	32,832 0 0
Conv-3	4x4 Convolution Instance Normalization Leaky ReLU	[128, 48, 16]	131,200 0 0
Conv-4	4x4 Convolution Instance Normalization Leaky ReLU	[256, 24, 8]	524,544 0 0
Conv-5	4x4 Convolution Instance Normalization Leaky ReLU	[256, 12, 4]	1,048,832 0 0
Conv-6	4x4 Convolution Instance Normalization Leaky ReLU	[256, 11, 3]	1,048,832 0 0
Conv-7	4x4 Convolution Sigmoid	[1, 10, 2]	4,097 0
Total			2,8M

4.1.2 Entraînement

Les images sont redimensionnées à une hauteur de 384 pixels, et une largeur de 128 pixels. La taille des vecteurs de contenu et d'identité est fixée à 512, ce qui correspond à la taille de vecteur utilisée par le réseau OSNet [3].

Sur le modèle de l'entraînement du réseau OSNet-AIN [38], l'encodeur d'identité est pré-entraîné sur ImageNet [21]. Lors de l'étape 1, l'encodeur d'identité est entraîné sur la base de donnée source durant 100 epochs, sur des batchs de 32 images. Nous utilisons l'optimiseur AMSGrad [51] avec un taux d'apprentissage initial fixé à 0.00015, qui est ensuite réduit par la méthode de recuit de cosinus [52]. Lors des 20 premiers epochs, nous entraînons seulement la dernière couche de classification ajoutée au réseau pré-entraîné [53].

Lors de l'étape 2, nous entraînons le discriminateur avec une descente de gradient stochastique, où le moment est fixé à 0.9. Le générateur et l'encodeur de contenu utilisent l'optimiseur Adam [54] ($\beta_1 = 0.9, \beta_1 = 0.999$). Les modules sont entraînés pendant 200 epochs avec une taille de batch de 16 images (8 paires). Nous fixons le taux d'apprentissage à $2e - 4$.

Pour l'étape 3, nous réduisons le taux d'apprentissage à $2e-5$ et entraînons le système de réseaux au complet pendant 400 epochs, en alternant les deux domaines.

4.1.3 Hyperparamètres

Nous fixons les poids $\lambda_{KL} = 1e-4$, $\lambda_{Rec} = 10$ et $\lambda_{Adv} = 1$ dans $\mathcal{L}_{Cible}$. Notamment, les valeurs de λ_{KL} et λ_{Rec} réalisent un compromis entre effectuer une meilleure reconstruction en autorisant plus d'information dans le vecteur de contenu, et trop se fier au vecteur de contenu, ce qui aurait pour effet de retirer de l'information d'identité du vecteur d'identité. Nous avons mené une recherche par quadrillage pour la valeur de λ_{KL} pour trouver le meilleur compromis, avec $\lambda_{KL} \in \{1e-3, 1e-4\}$.

Nous avons implémenté notre modèle sur Pytorch (v1.3.1), et l'avons entraîné sur un GPU NVIDIA RTX2080. L'entraînement complet dure environ 5 jours.

4.2 Bases de données

Les bases de données spécifiques à la ré-identification sont constituées d'images de personnes extraites de vidéos capturées par des réseaux de caméra de vidéosurveillance. Les personnes sont extraites dans les vidéos par un réseau de détection de personnes, puis annotées manuellement selon l'identité de la personne détectée. Les bases de données plus anciennes Viper [7], PRID [8] et CUHK03 [9] contiennent un petit nombre d'identités réparties sur quelques caméras. Les bases de données plus récentes Market1501 [10], DukeMTMC [11,12] et MSMT17 [6] proposent un plus grand nombre d'images pour s'adapter aux méthodes récentes d'apprentissage automatique qui exploitent un grand nombre de données. Elles proposent également une difficulté accrue en variant les scènes, l'éclairage, les occlusions, les points de vues et les changements de poses. La base de donnée MSMT17 récemment publiée contient ainsi des caméras intérieures et extérieures et des prises de vues réalisées à différents moments de la journée pour varier l'éclairage.

Nous nous intéressons dans ce travail uniquement aux trois bases de données récentes Market1501, DukeMTMC et MSMT17, qui présentent un défi plus important et sont plus représentatives des cas d'usages réels. La Figure 4.4 présente des exemples d'images contenues dans ces bases de données. Nous évaluerons les transferts entre ces trois bases de données, en les considérant alternativement comme bases de données source et cible.

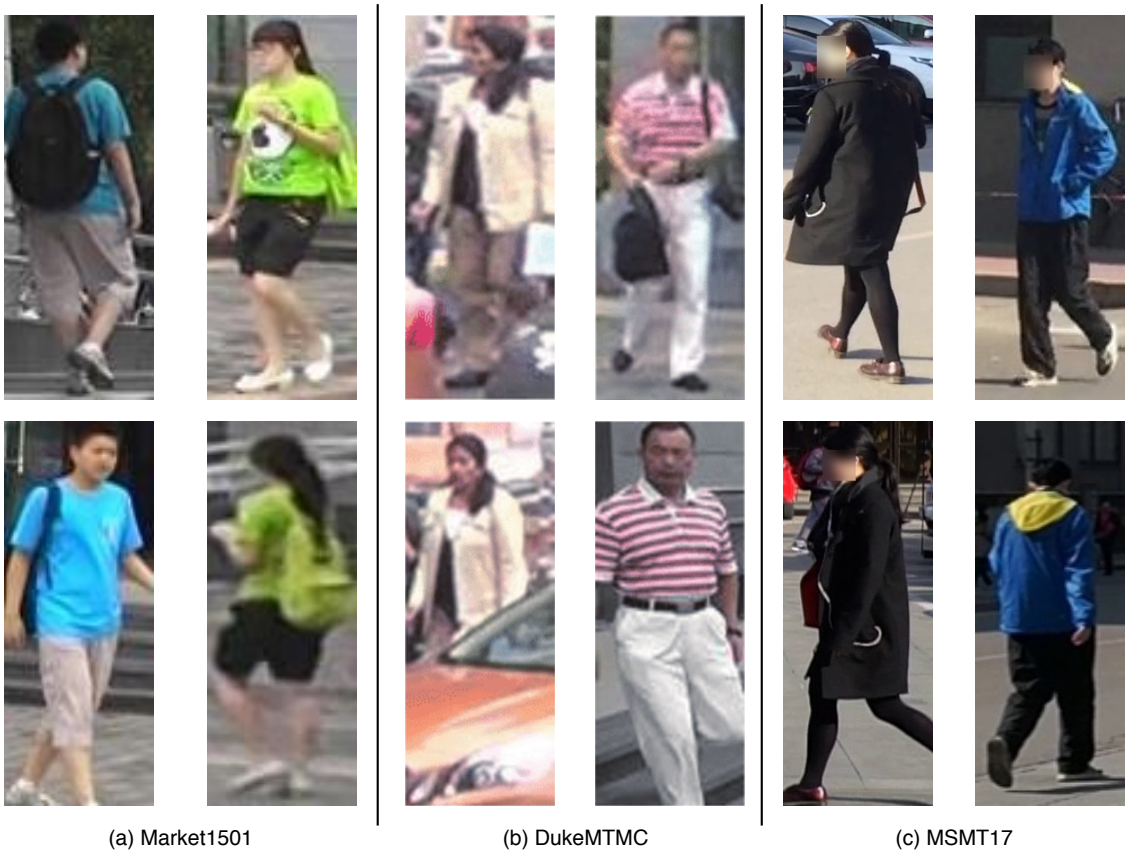

Figure 4.4 Exemples d'images des bases de données utilisées : Market1501, DukeMTMC et MSMT17.

4.3 Évaluation

Les images des bases de données sont divisées entre ensemble d'entraînement et ensemble de test, avec des identités distinctes. Les images de l'ensemble de test de la base de donnée cible ne sont pas utilisées pour l'entraînement.

Lorsque la base MSMT17 est considérée en tant que base de données source, toutes les images de la base de données sont rassemblées pour constituer un grand ensemble d'entraînement, selon le protocole expérimental usuel [1, 34, 38]. En revanche, seules les images de l'ensemble d'entraînement des bases de données Market1501 et DukeMTMC sont utilisées lorsque ces bases constituent les bases de données sources, là encore selon le protocole expérimental établi.

Les images de l'ensemble de test sont elles-même partagées en deux groupes : requête et galerie. L'ensemble de requête contient une image pour chaque caméra et chaque identité

présentes dans l'ensemble de test. Le Tableau 4.4 résume les caractéristiques détaillées des trois bases de données utilisées.

Tableau 4.4 Caractéristiques des bases de données Market1501, DukeMTMC and MSMT17. (E : Entraînement, R : Requête, G : Galerie), Cam : nombre de caméras

Base	Cam.	Images (E-R-G)	Identités (E-R-G)
Market1501 [10]	6	32668 (12936-3368-19732)	1501 (751-750-750)
DukeMTMC [11, 12]	8	36411 (16522-2228-17661)	1812 (702-702-1110)
MSMT17 [6]	15	126441 (30248-11659-82161)	4101 (1041-3060-3060)

La performance du réseau est évaluée par deux métriques : Cumulated Matching Characteristics (CMC) et Mean Average Precision (mAP). Le réseau de ReID extrait un vecteur de ReID pour chaque image de l'ensemble de test. On calcule ensuite la distance entre chaque image de l'ensemble de requête et les images de la galerie, en excluant les images provenant de la même caméra. La courbe CMC représente pour un rang k donné la probabilité de retrouver l'identité de la requête parmi les k premières images de la galerie triées par distance. On relève typiquement en ReID les valeurs de précision aux rangs $k = 1$, 5 et 10.

La courbe CMC est une mesure classique de précision lorsqu'on dispose d'une liste triée de résultats, avec un résultat positif et des résultats négatifs. Cependant, les valeurs de la courbe CMC sont biaisées lorsque plusieurs résultats positifs sont admissibles, ce qui est le cas en ReID puisque chaque identité de l'ensemble de requête apparaît à plusieurs reprises dans la galerie. Dans cette situation, la courbe CMC renseigne seulement sur la présence d'une valeur positive parmi les premiers résultats. Pour une évaluation plus précise, il faut aussi prendre en compte le nombre de résultats positifs et leur position dans la liste. C'est pourquoi une deuxième métrique, mAP, est introduite avec la base de donnée Market1501 [10]. Cette métrique est définie comme la moyenne sur toute les images requêtes de l'aire sous la courbe de précision-rappel de la liste triée de résultats correspondant à la requête.

On appelle performance de référence les valeurs de CMC et mAP à l'issue de l'entraînement préalable sur l'ensemble de données source. Cette performance dépend uniquement de l'architecture du réseau de ReID ainsi que des paramètres de l'entraînement. Pour évaluer le processus d'adaptation à l'ensemble de données cibles, on mesure notamment la différence entre la performance de référence et la performance finale à l'issue du processus d'adaptation.

4.4 Résultats

Nous évaluons notre méthode pour les transferts entre les bases de données MSMT17, Market1501 et DukeMTMC. Plus spécifiquement, nous réalisons les transferts depuis la

base MSMT17 : MSMT17 → Market1501 et MSMT17 → DukeMTMC, ainsi que les transferts DukeMTMC → Market1501 et Market1501 → DukeMTMC. Les transferts vers la base MSMT17 ne sont en général pas étudiés dans la littérature, car MSMT17 est la base de données la plus grande et qui présente le plus de diversité de contenu : elle est donc idéale pour servir de base de données source, mais les performances d'un transfert vers MSMT17 ne seraient pas bonnes puisque le réseau de ReID ne serait alors pas exposé à la même diversité de contenu lors de l'entraînement préalable sur la base de données source.

4.4.1 Comparaison avec l'état de l'art

Nous comparons notre méthode avec plusieurs approches de l'état de l'art pour l'adaptation de domaine en ReID de personnes : MAR [1], PAUL [34] et OSNet-AIN [38]. Nous avons conduit nos tests dans les mêmes conditions expérimentales, et présentons les résultats dans le Tableau 4.5.

Tableau 4.5 Résultats de l'adaptation de domaine de la base de données MSMT17 vers les bases de données Market1501 et DukeMTMC. Les meilleurs résultats sont indiqués en gras, les deuxièmes meilleurs sont soulignés.

Méthode	Publication	MSMT17 → Market1501				MSMT17 → DukeMTMC			
		R1	R5	R10	mAP	R1	R5	R10	mAP
MAR [1]	CVPR19	67.7	81.9	-	40.0	67.1	79.8	-	48.0
PAUL [34]	CVPR19	68.5	82.4	87.4	40.1	**72.0**	<u>82.7</u>	<u>86.0</u>	**53.2**
OSNet-AIN [38]	-	<u>70.1</u>	<u>84.1</u>	<u>88.6</u>	<u>43.3</u>	<u>71.1</u>	**83.3**	**86.4**	<u>52.7</u>
UD-GAN (Ours)	-	**73.6**	**86.4**	**90.7**	**47.2**	62.7	75.6	80.3	43.6

Tableau 4.6 Résultats de l'adaptation de domaine entre les bases de données Market1501 et DukeMTMC. Les meilleurs résultats sont indiqués en gras, les deuxièmes meilleurs sont soulignés.

Méthode	Publication	DukeMTMC → Market1501				Market1501 → DukeMTMC			
		R1	R5	R10	mAP	R1	R5	R10	mAP
IPGAN [36]	-	57.2	76.0	82.7	28.0	47.0	63.0	68.1	27.0
ECN [33]	CVPR19	<u>75.1</u>	**87.6**	**91.6**	<u>43.0</u>	**63.3**	<u>75.8</u>	<u>80.4</u>	<u>40.4</u>
PDA-Net [41]	ICCV19	**75.2**	<u>86.3</u>	<u>90.2</u>	**47.6**	<u>63.2</u>	**77.0**	**82.5**	**45.1**
OSNet-AIN [38]	-	61.0	77.0	82.5	30.6	52.4	66.1	71.2	30.5
UD-GAN (Ours)	-	61.6	77.2	83.0	31.4	49.7	64.7	70.5	30.2

Notre méthode surpasse de manière significative les méthodes existantes pour le transfert de la base de données MSMT17 à la base de données Market1501 : nous relevons une amélioration en valeur absolue de 3.5% sur la précision au rang 1, et de 3.9% en mAP par rapport à

la performance de référence d'OSNet-AIN sur la base de données Market1501. Pour ce transfert, le processus de désentrelacement des caractéristiques d'identité sur la base de données cible Market1501, à partir des paires d'images extraites, s'est prouvé efficace pour optimiser les performances du réseau de ReID OSNet-AIN. Pour ce transfert, nous dépassons les performances de l'état de l'art.

En revanche, notre méthode ne permet pas d'améliorer les performances du réseau de ReID de référence lors du transfert de la base de données MSMT17 à la base de données DukeMTMC. Au contraire, les performances décroissent en valeur absolue de 8.4% sur la précision au rang 1, et de 9.1% en mAP par rapport à la performance de référence d'OSNet-AIN sur la base de données DukeMTMC.

On retrouve le même constat pour les transferts entre les bases de données DukeMTMC et Market1501. Le processus d'adaptation de domaines parvient bien à améliorer les performances du réseau de référence lors du transfert de la base de données DukeMTMC à la base de données Market1501, puisqu'on observe une augmentation en valeur absolue de 0.6% sur la précision au rang 1, et de 0.8% en mAP. Mais le transfert inverse, de la base de données Market1501 à la base de données DukeMTMC, échoue : on relève une diminution de 2.7% sur la précision au rang 1, et de 0.3% en mAP.

Notre méthode est donc efficace pour l'adaptation de domaine vers la base de données cible Market1501, mais échoue vers la base de données cible DukeMTMC. On retrouve ce résultat quelle que soit la base de données source. Pour expliquer cette différence, comparons les images générées lors des transferts MSMT17 $\rightarrow$ Market1501 et MSMT17 $\rightarrow$ DukeMTMC (Figure 4.5 et Figure 4.6). On remarque que les images générées pour la base de données Market1501 contiennent plus de détails relatifs aux personnes, alors que celles de DukeMTMC sont plus floues et moins texturées. Il semble donc que notre méthode ne parvient pas aussi bien à réaliser le désentrelacement des caractéristiques d'identité sur les images de la base de données DukeMTMC. L'optimisation des paramètres d'apprentissage pour cette base, et notamment les contributions des différentes composantes de la fonction de coût $\mathcal{L}_{Cible}$, pourrait peut-être permettre d'améliorer le processus de désentrelacement, et donc la performance de la tâche de ReID.

Nous relevons également de meilleures performances pour les transferts depuis la base de données MSMT17 par rapport aux autres bases de données sources, à la fois pour la performance en valeur absolue et en ce qui concerne l'amélioration par rapport à la base de données source. Notre méthode délègue l'apprentissage de caractéristiques discriminantes à la base de données source, et son efficacité dépend donc fortement de la taille de cette base et de la diversité des images qu'elle contient. La base de données MSMT17 étant plus grande et plus

riche, il est normal d'observer les meilleurs résultats pour les transferts réalisés à partir de cette base de données.

La performance des transferts entre les bases de données DukeMTMC et Market1501 est inférieure aux méthodes ECN [33] et PDA-Net [41]. Ces méthodes sont basées plus fortement sur les spécificités de la base de données cible pour l'extraction de caractéristiques discriminantes, en exploitant des propriétés d'invariance ou encore par l'extraction de pose. Cela explique leurs bons résultats même lorsque la base de donnée source est plus petite. Cependant, les performances de ces méthodes sur ces transferts sont comparables à nos résultats pour les transferts depuis la base de données MSMT17. Les méthodes ECN [33] et PDA-Net [41] n'ont pas été testés pour les transferts depuis la base de données MSMT17.

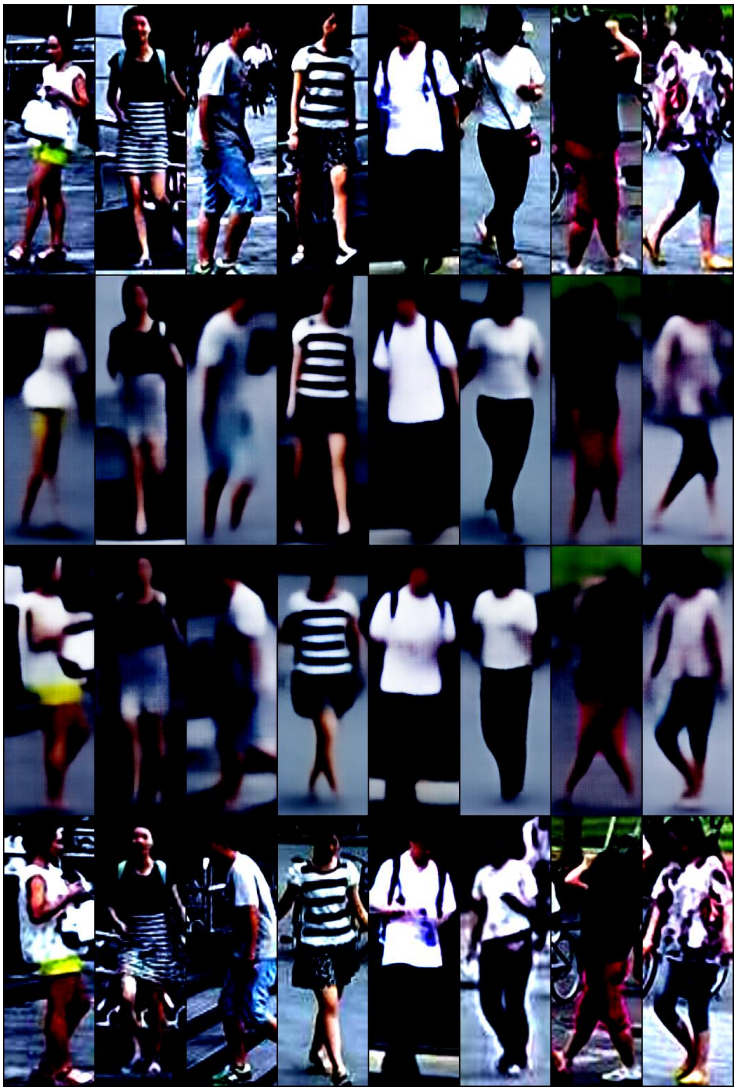

Figure 4.5 Exemple d'images générées sur la base de données Market1501. Première et dernière ligne : paires d'images originales de la base de données. Deuxième ligne : images reconstruites en utilisant des caractéristiques d'identité et de contenu extraites de la même image dans la première ligne. Troisième ligne : images générées en échangeant l'identité de la première ligne et le contenu de la dernière ligne.

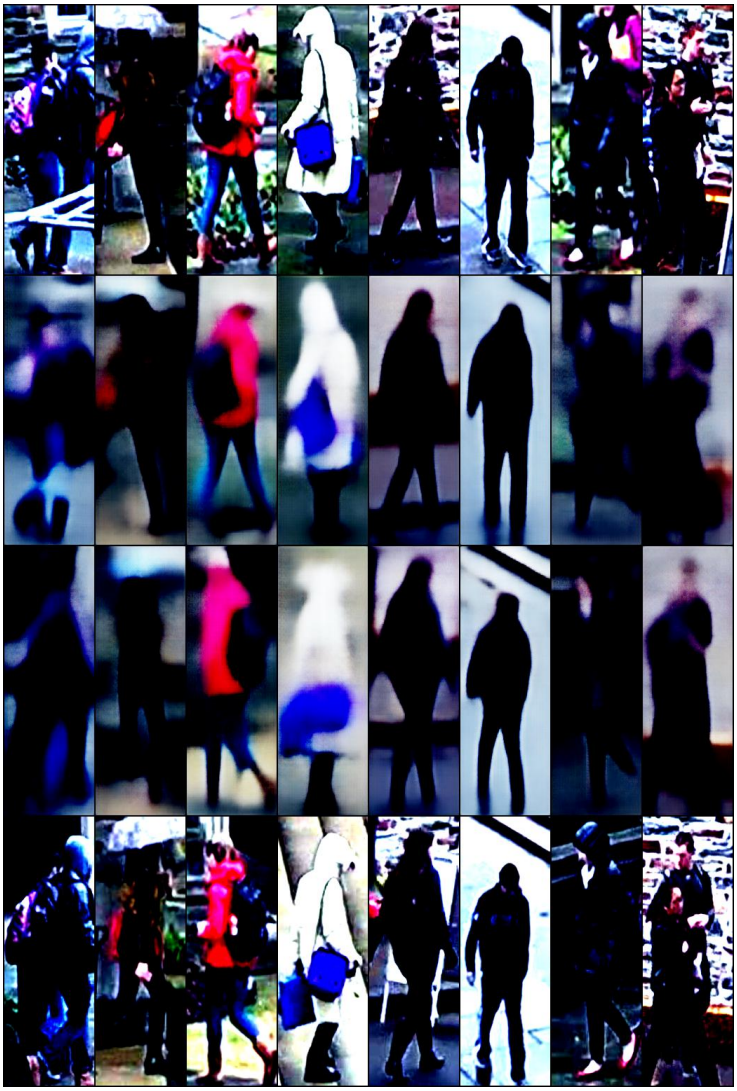

Figure 4.6 Exemple d'images générées sur la base de données DukeMTMC. Première et dernière ligne : paires d'images originales de la base de données. Deuxième ligne : images reconstruites en utilisant des caractéristiques d'identité et de contenu extraites de la même image dans la première ligne. Troisième ligne : images générées en échangeant l'identité de la première ligne et le contenu de la dernière ligne.

4.4.2 Sur-apprentissage en l'absence d'adaptation de domaine

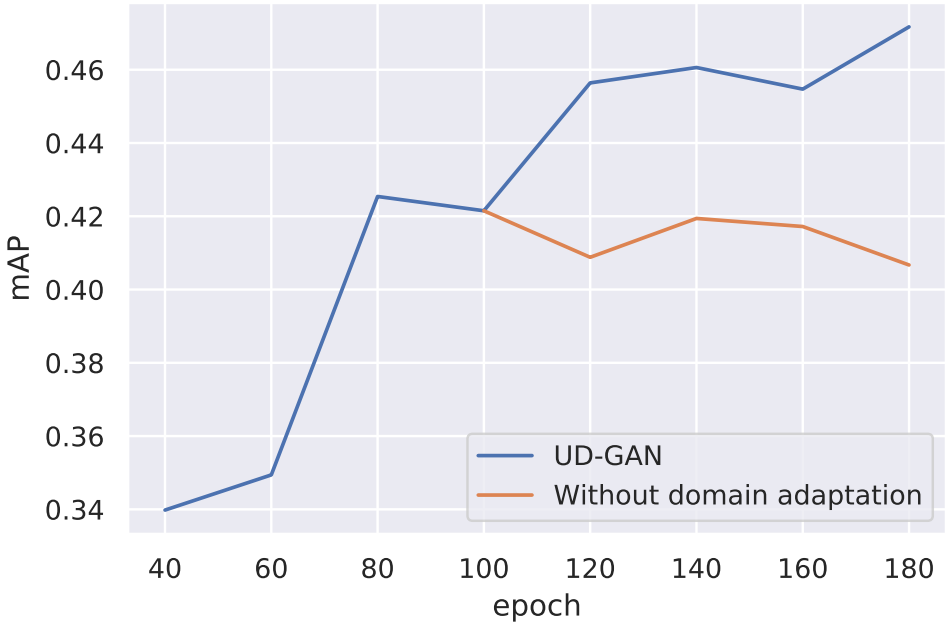

Figure 4.7 Mean Average Precision (mAP) sur la base de données cible Market1501 en fonction du nombre d'epochs d'entraînement sur la base de données source MSMT17, pour deux protocoles expérimentaux : 1) en utilisant notre méthode et 2) sans adaptation de domaine.

Nous comparons à la Figure 4.7 le transfert de la base de données MSMT17 à Market1501 sans le processus d'adaptation de domaine, et en utilisant notre méthode. Nous représentons la performance sur la base de données cible Market1501 en fonction du nombre d'epochs d'entraînement sur la base de données source MSMT17. Les epochs sont comptés sur la base de données MSMT17 : les 100 premiers epochs correspondent à l'étape 1, pendant laquelle seul le réseau de ReID est entraîné. Les epochs suivants correspondent à l'étape 3, dans laquelle les batchs sont tirés alternativement des bases de données source et cible. Nous comparons avec la performance du transfert de domaine direct lorsque la même quantité de données de la base de données source est utilisée pendant l'entraînement.

Si l'entraînement sur la base de données source MSMT17 continue sans adaptation de domaine, la performance sur la base de données cible Market1501 atteint rapidement un plateau, et diminue même en même temps que le réseau de ReID est optimisé pour la base de données source. Nous en concluons que le réseau de ReID peut perdre en capacité à discriminer en sur-apprenant la distribution de la base de données source. Notre méthode permet la poursuite de l'entraînement sans sur-apprentissage en utilisant l'information supplémentaire

provenant de la base de données cible. Ce faisant, le réseau de ReID apprend à extraire des caractéristiques pertinentes dans la base de données cible.

4.4.3 Extraction de paires non supervisée

Nous introduisons une méthode pour l'extraction non supervisée de paires d'images de même identité dans l'ensemble d'entraînement du domaine cible au début de l'étape 2. En considérant simplement la meilleure correspondance sans traitement supplémentaire, le réseau ReID pré-entraîné sur la base de donnée MSMT17 atteint déjà une précision de 89.1% au rang 1 sur l'ensemble d'entraînement non annoté de Market1501. Notez cependant que nous autorisons ici les correspondances d'images capturées par la même caméra que l'image de requête, contrairement au protocole d'évaluation où des correspondances doivent être trouvées dans les images capturées par les autres caméras du réseau. Cela explique pourquoi la précision est beaucoup plus élevée que lors de l'évaluation sur l'ensemble de test. Ce bon résultat valide le choix d'OSNet-AIN comme base de référence solide pour l'adaptation de domaine en ReID de personnes.

Après avoir filtré les paires qui pourraient être erronées selon les critères décrits à la section 3.6, nous obtenons une précision de 96.0% tout en conservant 74.2% des paires (9602 paires sur les 12936 paires possibles que nous essayons de construire avec notre approche d'extraction de pires). Ce résultat valide le bon fonctionnement du processus de filtrage, qui permet bien de garder la plupart des paires potentielles tout en diminuant fortement le nombre de paires erronées. On retrouve des performances similaires pour la précision de l'extraction de paires lors des autres transferts effectués.

CHAPITRE 5 CONCLUSION

5.1 Synthèse des travaux

Nous avons introduit un nouveau modèle génératif de désentrelacement non supervisé (UD-GAN), pour le transfert de domaine en ré-identification de personnes. Notre système de réseaux de neurones peut être utilisé avec n'importe quelle architecture de réseau de ReID, et peut être appliqué à n'importe quel domaine cible non annoté, sans modification. Nous introduisons une méthode efficace d'extraction de paires d'images contenant avec une grande probabilité la même identité, dans un contexte non supervisé, et une nouvelle méthode d'adaptation de domaine basée sur le désentrelacement des caractéristiques latentes de ReID.

Nous entraînons conjointement le réseau de ReID sur le domaine source annoté pour l'extraction de caractéristiques d'identité discriminantes, tout en conduisant le désentrelacement des caractéristiques dans le domaine cible non annoté pour adapter les caractéristiques de ReID au nouveau domaine. Nous proposons également une méthode pour extraire de manière fiable des paires d'images de la même identité dans le domaine cible, que nous utilisons en entrée de notre système de réseaux génératif de désentrelacement.

Nous avons mené des expériences sur plusieurs cadres de référence pour la ReID de personne non supervisée. Nous dépassons la performance de l'état de l'art pour la précision en mAP et au Rang 1 pour le transfert de MSMT17 à Market1501. Le processus de désentrelacement des caractéristiques d'identité sur la base de données cible, à partir des paires d'images extraites, permet d'optimiser les performances du réseau de ReID.

5.2 Limitations de la solution proposée

Cependant, notre méthode ne permet pas toujours d'améliorer les performances du réseau de ReID de référence. Malgré nos efforts, les transferts vers la base de données DukeMTMC se sont révélés infructueux pour le moment. Les très bonnes performances obtenues pour les transferts vers la base de données Market1501 laissent cependant à penser que l'amélioration des résultats serait possible pour la base de données DukeMTMC, par exemple en affinant les hyperparamètres de notre méthode.

5.3 Améliorations futures

À l'avenir, nous prévoyons de tester notre modèle avec d'autres architectures de réseau de ReID. Le modèle récent OSNet-AIN que nous utilisons est très performant pour la ReID non supervisée. Il serait intéressant d'utiliser notre méthode avec des architectures moins efficaces mais plus répandues comme ResNet.

RÉFÉRENCES

[1] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong et J.-H. Lai, "Unsupervised Person Re-Identification by Soft Multilabel Learning," dans *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA : IEEE, juin 2019, p. 2143–2152.

[2] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang et J. Kautz, "Joint Discriminative and Generative Learning for Person Re-Identification," dans *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA : IEEE, juin 2019, p. 2133–2142.

[3] K. Zhou, Y. Yang, A. Cavallaro et T. Xiang, "Omni-Scale Feature Learning for Person Re-Identification," dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South) : IEEE, oct. 2019, p. 3701–3711.

[4] G. Wang, Y. Yuan, X. Chen, J. Li et X. Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-Identification," *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*, p. 274–282, 2018.

[5] C. Eom et B. Ham, "Learning Disentangled Representation for Robust Person Re-identification," dans *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox et R. Garnett, édit. Curran Associates, Inc., 2019, p. 5298–5309.

[6] L. Wei, S. Zhang, W. Gao et Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-identification," dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA : IEEE, juin 2018, p. 79–88.

[7] D. Gray et H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," dans *Computer Vision – ECCV 2008*, ser. Lecture Notes in Computer Science, D. Forsyth, P. Torr et A. Zisserman, édit. Berlin, Heidelberg : Springer, 2008, p. 262–275.

[8] M. Hirzer, C. Beleznai, P. M. Roth et H. Bischof, "Person Re-identification by Descriptive and Discriminative Classification," dans *Image Analysis*, ser. Lecture Notes in Computer Science, A. Heyden et F. Kahl, édit. Berlin, Heidelberg : Springer, 2011, p. 91–102.

[9] W. Li, R. Zhao, T. Xiao et X. Wang, "DeepReID : Deep Filter Pairing Neural Network for Person Re-identification," dans *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA : IEEE, juin 2014, p. 152–159.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang et Q. Tian, "Scalable Person Re-identification : A Benchmark," dans *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile : IEEE, déc. 2015, p. 1116–1124.

[11] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara et C. Tomasi, "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking," *arXiv :1609.01775 [cs]*, sept. 2016.

[12] Z. Zheng, L. Zheng et Y. Yang, "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro," dans *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice : IEEE, oct. 2017, p. 3774–3782.

[13] J. Redmon et A. Farhadi, "YOLOv3 : An Incremental Improvement," *arXiv :1804.02767 [cs]*, avr. 2018.

[14] K. He, G. Gkioxari, P. Dollár et R. Girshick, "Mask R-CNN," *arXiv :1703.06870 [cs]*, janv. 2018.

[15] S. Zhang, E. Staudt, T. Faltemier et A. K. Roy-Chowdhury, "A Camera Network Tracking (CamNeT) Dataset and Performance Baseline," dans *2015 IEEE Winter Conference on Applications of Computer Vision*. Waikoloa, HI, USA : IEEE, janv. 2015, p. 365–372.

[16] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang et J. Sun, "AlignedReID : Surpassing Human-Level Performance in Person Re-Identification," *arXiv :1711.08184 [cs]*, janv. 2018.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He et P. Dollár, "Focal Loss for Dense Object Detection," *arXiv :1708.02002 [cs]*, févr. 2018.

[18] Z. Zhong, L. Zheng, D. Cao et S. Li, "Re-ranking Person Re-identification with k-Reciprocal Encoding," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI : IEEE, juill. 2017, p. 3652–3661.

[19] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao et S. C. H. Hoi, "Deep Learning for Person Re-identification : A Survey and Outlook," *arXiv :2001.04193 [cs]*, janv. 2020.

[20] L. Zheng, Y. Yang et A. G. Hauptmann, "Person Re-identification : Past, Present and Future," *arXiv :1610.02984 [cs]*, oct. 2016.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et Li Fei-Fei, "ImageNet : A large-scale hierarchical image database," dans *2009 IEEE Conference on Computer Vision and Pattern Recognition*, juin 2009, p. 248–255.

[22] K. He, X. Zhang, S. Ren et J. Sun, "Deep Residual Learning for Image Recognition," dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA : IEEE, juin 2016, p. 770–778.

[23] W. Li, X. Zhu et S. Gong, "Harmonious Attention Network for Person Re-identification," dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA : IEEE, juin 2018, p. 2285–2294.

[24] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang et X. Tang, "Spindle Net : Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI : IEEE, juill. 2017, p. 907–915.

[25] G. Wang, J.-H. Lai, P. Huang et X. Xie, "Spatial-Temporal Person Re-identification," *AAAI*, 2019.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville et Y. Bengio, "Generative Adversarial Nets," dans *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence et K. Q. Weinberger, édit. Curran Associates, Inc., 2014, p. 2672–2680.

[27] F. Chollet, "Xception : Deep Learning with Depthwise Separable Convolutions," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI : IEEE, juill. 2017, p. 1800–1807.

[28] Y. Zhai, X. Guo, Y. Lu et H. Li, "In Defense of the Classification Loss for Person Re-Identification," *arXiv :1809.05864 [cs]*, nov. 2018.

[29] A. Hermans, L. Beyer et B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv :1703.07737 [cs]*, nov. 2017.

[30] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang et H. Li, "FD-GAN : Pose-guided Feature Distilling GAN for Robust Person Re-identification," dans *NeurIPS*, 2018.

[31] X. Huang et S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," dans *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice : IEEE, oct. 2017, p. 1510–1519.

[32] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf et A. J. Smola, "A Kernel Method for the Two-Sample-Problem," dans *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt et T. Hoffman, édit. MIT Press, 2007, p. 513–520.

[33] Z. Zhong, L. Zheng, Z. Luo, S. Li et Y. Yang, "Invariance Matters : Exemplar Memory for Domain Adaptive Person Re-Identification," dans *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA : IEEE, juin 2019, p. 598–607.

[34] Q. Yang, H.-X. Yu, A. Wu et W.-S. Zheng, "Patch-Based Discriminative Feature Learning for Unsupervised Person Re-Identification," dans *2019 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA : IEEE, juin 2019, p. 3628–3637.

[35] J.-Y. Zhu, T. Park, P. Isola et A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," dans *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice : IEEE, oct. 2017, p. 2242–2251.

[36] J. Liu, W. Li, H. Pei, Y. Wang, F. Qu, Y. Qu et Y. Chen, "Identity Preserving Generative Adversarial Network for Cross-Domain Person Re-Identification," *IEEE Access*, vol. 7, p. 114 021–114 032, 2019.

[37] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim et J. Choo, "StarGAN : Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, juin 2018, p. 8789–8797.

[38] K. Zhou, Y. Yang, A. Cavallaro et T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *arXiv :1910.06827 [cs]*, oct. 2019.

[39] D. Ulyanov, A. Vedaldi et V. Lempitsky, "Instance Normalization : The Missing Ingredient for Fast Stylization," *arXiv :1607.08022 [cs]*, nov. 2017.

[40] P. Esser, J. Haux et B. Ommer, "Unsupervised Robust Disentangling of Latent Characteristics for Image Synthesis," dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South) : IEEE, oct. 2019, p. 2699–2709.

[41] Y.-J. Li, C.-S. Lin, Y.-B. Lin et Y.-C. F. Wang, "Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation," dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South) : IEEE, oct. 2019, p. 7918–7928.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens et Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA : IEEE, juin 2016, p. 2818–2826.

[43] J. Bao, D. Chen, F. Wen, H. Li et G. Hua, "Towards Open-Set Identity Preserving Face Synthesis," dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA : IEEE, juin 2018, p. 6713–6722.

[44] D. P. Kingma et M. Welling, "Auto-Encoding Variational Bayes," *arXiv :1312.6114 [cs, stat]*, mai 2014.

[45] B. Lu, J.-C. Chen et R. Chellappa, "Unsupervised Domain-Specific Deblurring via Disentangled Representations," dans *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA : IEEE, juin 2019, p. 10 217–10 226.

[46] P. Isola, J.-Y. Zhu, T. Zhou et A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI : IEEE, juill. 2017, p. 5967–5976.

[47] A. Radford, L. Metz et S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv :1511.06434 [cs]*, janv. 2016.

[48] S. Ioffe et C. Szegedy, “Batch normalization : Accelerating deep network training by reducing internal covariate shift,” dans *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML’15. Lille, France : JMLR.org, juill. 2015, p. 448–456.

[49] A. L. Maas, A. Y. Hannun et A. Y. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models,” *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, vol. 30, p. 3, 2013.

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever et R. Salakhutdinov, “Dropout : A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, n°. 56, p. 1929–1958, 2014.

[51] S. J. Reddi, S. Kale et S. Kumar, “On the Convergence of Adam and Beyond,” *ICLR*, 2018.

[52] I. Loshchilov et F. Hutter, “SGDR : Stochastic Gradient Descent with Warm Restarts,” dans *ICLR*, 2017.

[53] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian et T. Xiang, “Deep Transfer Learning for Person Re-Identification,” dans *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, sept. 2018, p. 1–5.

[54] D. P. Kingma et J. Ba, “Adam : A Method for Stochastic Optimization,” *ICLR*, 2014.