Domain Gap and Privacy in Person Re-Identification

Hamza Rami

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAT046

Thèse de doctorat

Domain Gap and Privacy in Person Re-Identification

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 25 novembre 2024, par

HAMZA RAMI

Composition du Jury :

Florence d'Alché-Buc Professeure, Télécom Paris	Présidente/Examinatrice
Jocelyn Chanussot Directeur de recherche, Centre Inria de l'Université Grenoble Alpes	Rapporteur
Michel Crucianu Professeur, Conservatoire National des Arts et Métiers	Rapporteur
Alice Caplier Professeure des universités, Grenoble INP - Phelma	Examinatrice
Marco Cagnazzo Professeur, Télécom Paris / Université de Padova	Directeur de thèse
Jhony H. Giraldo Maître de conférences, Télécom Paris	Invité
Nicolas Winckler Ingénieur de recherche, Eviden (ex Atos)	Invité

Remerciements

Ces quelques lignes ne suffiront pas à exprimer toute ma gratitude envers celles et ceux qui m'ont accompagné tout au long de cette aventure doctorale, mais elles en sont le reflet sincère.

Tout d'abord, un immense merci à mon directeur de thèse, Marco Cagnazzo, pour ses conseils avisés, son soutien constant et sa confiance tout au long de ces années.

Je souhaite également exprimer toute ma reconnaissance à mes encadrants, qui ont joué un rôle essentiel dans cette thèse. Du côté de Télécom Paris, un grand merci à Stéphane Lathuilière et Jhony Giraldo pour leur disponibilité, leurs précieux conseils et les échanges enrichissants qui ont nourri ma réflexion. Leur expertise et leur accompagnement ont été d'une aide précieuse à chaque étape de ce travail. Un immense merci également à Nicolas Winckler et Matthieu Ospici d'Atos Eviden pour leur encadrement bienveillant et leur accueil chaleureux à Grenoble. Leur soutien, tant sur le plan scientifique que logistique, m'a permis de mener cette recherche dans des conditions optimales, notamment grâce aux ressources de calcul mises à disposition. Travailler à vos côtés a été un privilège, et je vous en suis profondément reconnaissant.

Je tiens également à remercier les membres du jury, Florence d'Alché-Buc, Jocelyn Chanussot, Michel Crucianu et Alice Caplier, pour le temps consacré à l'évaluation de mon travail et pour leurs retours constructifs, qui ont permis d'enrichir cette recherche.

À mes collègues du laboratoire LTCI, qui sont devenus bien plus que de simples collègues, merci pour ces années passées ensemble, les discussions passionnées, les pauses café salvatrices et ces moments de complicité qui ont rendu ces années de thèse inoubliables et d'autant plus agréables.

Enfin, je ne saurais oublier ma famille, qui a toujours été là, dans les moments de réussite comme dans ceux de doute. Merci pour votre patience, votre soutien infaillible et surtout votre amour inconditionnel.

À tous ceux qui, de près ou de loin, ont contribué à cette belle aventure, un immense merci du fond du cœur.

Résumé en français

La ré-identification de personnes (Re-ID) est une tâche bien établie dans les systèmes de surveillance modernes qui vise à reconnaître des individus à travers des images capturées par différentes caméras de surveillance sans champs de vision se chevauchant. Cette capacité ne se limite pas à une simple avancée technique, elle joue un rôle crucial dans l'amélioration de la sécurité publique, en facilitant par exemple le suivi d'individus dans les gares, les aéroports ou les espaces urbains denses. Elle est également utilisée dans des secteurs variés, notamment pour l'analyse comportementale en commerce, la gestion des flux piétons en smart cities, ou encore le monitoring de patients en milieu hospitalier.

Cependant, le déploiement de modèles de Re-ID est généralement limité par l'écart de domaine. Il s'agit de la disparité dans la distribution entre l'ensemble des données d'entraînement (domaine source) et les données réelles rencontrées lors du déploiement (domaine cible). En effet, les performances des modèles de Re-ID tendent à dégrader lorsqu'ils sont appliqués à un environnement différent de celui dans lequel ils ont été entraînés, en raison des variations de conditions d'éclairage, d'angles de vue et de qualité des images.

Nous nous intéressons aux méthodes d'Adaptation de Domaine Non Supervisée (UDA) qui ont émergé comme des outils puissants pour combler cet écart de domaine. En exploitant les connaissances acquises à partir du domaine source labelisé, les méthodes UDA permettent aux modèles de s'adapter à de nouveaux environnements sans nécessiter de données labelisée dans le domaine cible.

De plus, le déploiement de systèmes de Re-ID est de plus en plus soumis à des réglementations strictes sur les données, telles que le Règlement Général sur la Protection des Données (RGPD) et le AI Act. Ces nouvelles réglementations ajoutent des contraintes critiques sur le stockage et le transfert de données afin de protéger la vie privée des individus. Par conséquent, les méthodes traditionnelles d'UDA pour la Re-ID qui reposent sur le transfert et le stockage de grands volumes de données de surveillance pour l'entraînement et l'adaptation des modèles font face à des défis légaux et éthiques.

Pour adhérer aux contraintes de confidentialité, nous présentons deux nouveaux settings: Online UDA (OUDA-Rid) et Distributed UDA (DUDA-Rid). OUDA-Rid se focalise sur l'adaptation des modèles de Re-ID quand les données sont continuellement transmises par des caméras de surveillance sans accès direct aux données stockées. Ce setting est crucial pour les scénarios où les contraintes de confidentialité empêchent la rétention des données. DUDA-Rid étend le concept d'adaptation de domaine à un setting distribué dans lequel le processus d'adaptation est décentralisé à travers plusieurs caméras de surveillance, abordant ainsi les défis des restrictions de transfert de données. Pour surmonter les défis imposés par les settings susmentionnés, nous proposons dans cette thèse deux méthodes : Source-Guided Similarity Preservation (S2P) et Fed-Protoid.

S2P est conçu pour relever les défis de l'UDA et de l'oubli catastrophique, en se concen-

trant sur la contrainte de confidentialité liée au stockage des données. Au cours du processus d'apprentissage continu, S2P préserve les similarités de caractéristiques essentielles en sélectionnant soigneusement un ensemble de support provenant du domaine source qui maximise la similarité avec les données cibles. Cette approche permet une adaptation continue au domaine cible tout en respectant les réglementations liées au stockage des données.

Fed-Protoid est une méthode conçue pour faire face aux restrictions de transfert de données, en particulier l'interdiction de transférer des images de surveillance en dehors des caméras. En employant une approche d'apprentissage fédéré, Fed-Protoid permet une adaptation de domaine non supervisée distribuée à travers plusieurs appareils/caméras.

Ensemble, ces méthodes présentent une solution intégrale pour le déploiement de systèmes de Re-ID de personnes qui se conforment aux lois et réglementations récentes sur la protection des données. En comblant l'écart de domaine sous les conditions strictes de contraintes de stockage et de transfert de données, S2P et Fed-Protoid ouvrent la voie à la prochaine génération de Re-ID de personnes préservant la vie privée.

Nous validons l'efficacité des frameworks proposés, S2P et Fed-Protoid, à travers divers scénarios, y compris des tâches d'adaptation de domaine réel à réel et synthétique à réel. L'évaluation est réalisée sur des ensembles de données de référence en Re-ID, tels que Market-1501, MSMT17, CUHK03 et RandPerson, et couvre différents contextes représentatifs des défis rencontrés dans des déploiements réels.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Video Surveillance Systems

1.1.1 Context

For many centuries, surveillance has been an essential component of organized civilizations. It refers to any set of devices or mechanisms that can recognize, monitor, and track the movements of one or more people. Traditionally, the primary uses of surveillance have been in the military and for security, when watchmen and guards carried out the manual procedure.

In the mid-20th century, the world witnessed big changes that steadily transformed the surveillance landscapes with the development of the first closed-circuit television (CCTV) systems. At first, the CCTV systems were mainly used to monitor rocket launches but quickly found their way into public safety and surveillance. In the latter part of the 20th century, the digital revolution spurs the rapid evolution of surveillance systems. Moving from analog to digital systems enhanced the capabilities of data storage, retrieval, and processing. Moreover, the appearance of Internet and Wireless technologies further expanded the scope of surveillance systems, making them capable of monitoring and managing data remotely and in real-time.

Recently, the integration of Artificial Intelligence (AI) and machine learning into surveillance systems has been a pivotal breakthrough, enabling new functionalities like facial recognition, gait analysis, and person re-identification. Thanks to these advancements, the ability of surveillance systems to identify and track individuals has significantly improved in diverse settings, going from controlled environments like airports and office buildings to crowded open public spaces.

1.1.2 Integration to Various Sectors

Video surveillance systems have been widely integrated into various sectors, demonstrating their adaptability and necessity in today's modern society. For example, transportation and traffic management heavily rely on surveillance to monitor traffic, manage congestion, and ensure safety in public transport. In the retail and commerce sectors, these systems are becoming essential for security, loss prevention, and even for analyzing consumer behavior to enhance store layouts and improve the shopping experience. The concept of "smart cities" also relies on surveillance systems for urban planning, contributing to safer, more effective, and sustainable urban environments. The integration of surveillance systems is not limited to these sectors, since they can also be integrated in educational institutions, industrial environments, and healthcare facilities.

1.1.3 Surveillance and Privacy in the Era of AI: Balancing Innovation with Ethical Constraints

Through the years, surveillance systems have evolved, mirroring society's changing needs and advancing capabilities. These systems have grown from their simplistic analog form to become complex digital networks that are further enhanced by AI. Modern surveillance systems, equipped with digital cameras, provide not only visual capabilities but also possess an analytical component to interpret the observed imagery and footage. This led to the development of technologies like facial recognition and person re-identification (Re-ID), which have transformed surveillance from manual review of recorded videos after incidents to active monitoring. Now, surveillance systems are capable of analyzing videos in real time without human intervention. In particular, Person Re-ID stands out as an excellent example of innovation since it enables systems to recognize and track people across various camera views. It also makes operations more effective by making the monitoring process simpler, cutting down on the need for people to watch the security camera footage all the time.

Person Re-ID refers to the task of recognizing a person of interest across different scenes or camera feeds, even when their appearance may change due to changes in viewpoint, lighting, or pose. This task is crucial for effective surveillance since it enables continuous tracking of persons of interest across extensive public or semi-public areas. However, during the deployment of Person Re-ID systems, they encounter a crucial challenge which is *Domain Shift* [1], also referred to as Domain Gap. Domain shift is an omnipresent problem across numerous computer vision tasks, but it becomes particularly crucial and evident in the deployment of Person Re-ID systems in real-world scenarios. It mainly refers to the discrepancy between the data on which a model is trained (source domain) and the data it encounters in real-world applications (target domain). This gap results in a decrease in performance since the model's learned representations may not well generalize to new and unseen environments. Moreover, addressing the domain shift comes with another issue related to the collection of labeled data from the target domain. Firstly, acquiring such labeled data is too costly and time-consuming. Secondly, and perhaps more critically, the process of collecting and annotating data in target domains can raise serious privacy concerns, since tagging captured images of individuals in public areas with unique identifiers to train Person Re-ID models creates and

stores potentially private information about individuals' locations and activities.

In many countries, the growth of surveillance systems has urgently raised ethical questions, particularly about privacy. The state of surveillance today illustrates a complex balancing act: safeguarding the interests of individual rights while meeting the larger community's need for safety. This conflict is best exemplified by the Person Re-ID task, which raises important concerns regarding individual privacy and consent while also providing substantial benefits for collective security and efficiency. It is also important to understand the more general risk of such technologies when exploited by authoritarian and totalitarian regimes [2, 3]. These forms of governance might use surveillance systems as a tool to further widespread their centralized control and limit political freedoms. Such misuse underlines the importance of implementing and deploying surveillance systems that are aligned with ethical guidelines to prevent the weakening of democratic principles and protect personal liberties.

This thesis focuses on the understanding of privacy regulations and their impact on the Re-ID models, specifically addressing the following question: How can Re-ID models be adapted to comply with the constraints imposed by these regulations? The goal is to develop novel solutions and frameworks that ensure a balance between technological advancement and the critical need for privacy protections. In the following section, we will explore in more detail the specifics of the Person Re-ID task and highlight the main challenges that this thesis aims to address.

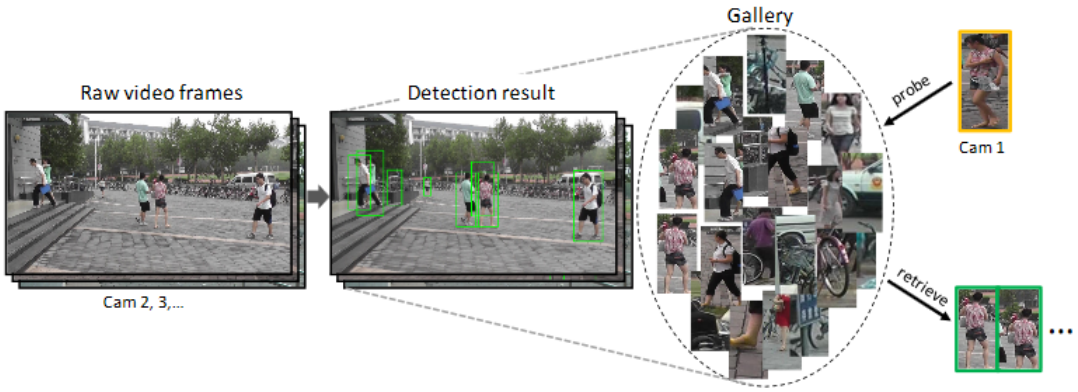
1.2 Person Re-Identification

1.2.1 Definition and Overview

The primary goal of Person Re-ID is to identify a person of interest across a network of cameras in diverse environments. More precisely, it tells whether a person, as a query, has appeared in another place, captured by a different camera, or even the same camera but at a different time instant. Technically speaking, a Person Re-ID system can be broken down into three steps (Fig. 1.1):

1. **Person Detection**: This task involves identifying and locating individuals within video frames.
2. **Person Tracking**: After individuals are detected, the person tracking task keeps monitoring their movements throughout the different frames in the video. It is akin to following a person as they move through various scenes in the video.
3. **Person Retrieval**: The last step is person retrieval, which is often considered the main and primary focus in many Person Re-ID studies. This involves recognizing and matching an individual across different cameras or periods, guaranteeing that the person of interest remains the same despite changes in viewpoint, lighting, or even clothing. In this thesis, if not specified, Person Re-ID will always refer to the person retrieval task.

Historically, the Person Re-ID was not formally recognized as a separate field and was



(a) Person Detection (b) Person Tracking (c) Person Retrieval

Figure 1.1: General Pipeline of Person Re-ID. Figure inspired from [4]

mainly integrated with multi-camera tracking, which fuses appearance models with geometry calibration across disjoint cameras [5]. The term "person re-identification" made his first appearance in 2005, when researchers started focusing on re-identifying individuals as they re-entered the camera's field of view, using dynamic Bayesian network to correlate labels and features like color and spatial-temporal cues [6]. The emergence of Convolutional Neural Networks (CNN) [7] has greatly impacted the computer vision community. The great success in tasks like image classification [8] has affected almost all the research fields including Person Re-ID. This success has spread to Person Re-ID in 2014 where [9] employs for the first time a more general way that can learn a similarity metric directly from image pixels using Siamese Networks. These first works have laid a robust foundation that has led to big improvements in the field, inspiring other researchers to come up with new ideas and better understand how to adapt neural network architectures for the task of Person Re-ID.

1.2.2 The role of Person Re-Identification in modern Surveillance

Modern intelligent video surveillance has been gaining a lot of attention in recent years. Along with detection and tracking, Person Re-ID has emerged as an essential component for having complete surveillance systems. The capability of Person Re-ID in recognizing individuals across different camera views is not only critical for tracking individuals in public spaces but also for identifying persons of interest across disjoint cameras in surveillance systems without the need for human supervision.

Over the years, the domain of Person Re-ID has witnessed considerable advancements. Lately, the emergence of end-to-end deep learning approaches and large pre-trained models has marked a remarkable leap forward, boosting the performance of Person Re-ID while addressing its main challenges such as domain shift and privacy constraints alignment. These developments have highlighted the importance of Person Re-ID, establishing it as an indispensable component and crucial element within modern surveillance systems.

The importance of Person Re-ID can be also reflected in its diverse applications across various sectors, such as:

- **Public Safety**: Person Re-ID can assist law enforcement agencies in establishing public safety. In fact, integrating Person Re-ID technology with surveillance cameras can help authorities track persons of interest and identify potential suspects. Hence, participating in monitoring public spaces more efficiently.
- **Retail Analytics**: Person Re-ID can enhance customers' experience by providing insights into preferences, frequented sections, and purchasing trends, aiding in product placement, store layout, and targeted marketing strategies.
- **Smart Cities**: Person Re-ID is a pillar in the design and the development of smart cities since it supports their main initiatives ranging from managing pedestrian flows and public safety to crowd control and handling large public gatherings.
- **Healthcare**: In healthcare facilities, the Person Re-ID can also be applied to monitor patients and manage staff workflows, contributing to overall safety and efficiency.
- **Social Robotics**: Person Re-ID plays a crucial role in the field of social robotics since it enhances a robot's ability to interact differently with each individual. In fact, by identifying and distinguishing between different individuals, social robots can tailor their interactions according to the preferences, behaviors, and needs of each person. Person Re-ID also enables robots to remember past interactions with individuals, allowing them to build on previous experiences to enhance future interactions.

1.2.3 Challenges in Person Re-Identification

Despite significant progress in recent years, Person Re-ID, along with all computer vision tasks, remains inherently complex and faces distinct challenges.

The early research mainly focused on addressing the fundamental challenges related to Person Re-ID such as variations in lighting, pose, camera views, and occlusion. For example, two disjoint cameras, within the same surveillance system, one that is directly facing the sunlight and the other one that is placed in a shaded zone where sunlight does not reach will capture the same person differently, which leads to an increase of the intra-class variation.

With the integration of end-to-end deep learning architectures, the Person Re-ID algorithms are now capable of producing robust features to those variations. Now, the focus of the community has shifted towards more complex challenges related to the deployment of Person Re-ID systems in real-world applications and can be categorized, but not limited to, domain shift and the increasing concerns regarding the standards of ethics and privacy. In what follows, we will discuss these identified challenges associated with deploying Person re-ID systems.

Domain Shift In computer vision, a model is generally trained and evaluated on images drawn from the same domain (source domain), meaning that the training set and the evaluation set are drawn from the same distribution. However, when evaluating on another domain (target domain) with images drawn from a slightly different distribution, the performance of the model usually drops drastically. Commonly referred to as domain shift, domain gap or domain drift, this disparity between source and target domains occurs very often in practice for all computer vision tasks such as semantic segmentation [10], image classification [11,

12, 13], and particularly for Person Re-ID task [14, 15].

The problem of domain shift, which arises in the case where deploying a Person Re-ID model on a target dataset, results in a significant drop in performance [1]. Typically, a Person Re-ID model trained on a set of images collected in Paris, might not produce satisfactory results in New Delhi, because the distribution of the target domain does not necessarily align with the source domain distribution. As an example, Fig. 1.2 illustrates the domain shift between two commonly used datasets in Person Re-ID (CUHK03 and PRID). It is apparent that each dataset has its distinct visual style, categorizing them into separate domains. These differences lead to a notable decrease in the performance of Person Re-ID models [14]. Such discrepancies pose a significant hurdle in practice since manually annotating data for every new domain is impractical due to time, resource, and privacy constraints.

An effective approach to deal with domain shift under the unavailability of the target domain labels is known as Unsupervised Domain Adaptation (UDA) [16]. Essentially, UDA aims to train and adapt a model on a source domain while conserving good accuracy on the target domain all without requiring any labels from the target domain. In this thesis, we focus on different UDA techniques to overcome these challenges.

Figure 1.2: Illustration of the domain shift between two Person Re-ID datasets (CUHK03 and PRID). We also show images from two different cameras within the PRID dataset. Source: [17]

Ethics and Privacy The great advances in AI technology are nowadays changing the way we work, make decisions, and interact with the world around us. At this stage of development, AI is now being integrated into various sectors, ranging from healthcare and finance to personalized advertising and surveillance. However, this rapid development and integration are raising serious ethical and privacy concerns that necessitate careful consideration. When we talk about the ethics of AI, we are systematically thinking about making sure these technologies are fair, responsible, and transparent. Moreover, privacy challenges are paramount, as AI's capability to process, analyze, and make decisions based on vast amounts of personal data poses risks to individual privacy and data protection rights.

The European Union's General Data Protection Regulation (GDPR) and the recently proposed AI Act together form a comprehensive text law to ensure the ethical use of AI, with a strong focus on data privacy and protection. In fact, the GDPR, effective since May 25, 2018, provides a foundational approach to data privacy, emphasizing the fundamental principles that are lawfulness, fairness, transparency, and accountability when processing personal data. As for the AI Act, it seeks to regulate high-risk systems by imposing requirements for adherence to ethical standards. In general, it ensures that AI systems are designed and operated in a manner that protects individuals' rights by integrating GDPR principles into the operation and functionality of the AI systems.

In this thesis, we give high importance to ethics and privacy since naturally Person Re-ID involves images of people that are categorized as sensitive data. On one hand, under the GDPR: person images are considered a special category of personal data, requiring higher protection due to their sensitivity (Chap. 2 Art. 9). On the other hand, the AI Act focuses more on regularizing the use of biometric identification systems. By considering them as a form of "high-risk" AI systems, the AI Act prohibits the use of identification systems in publicly accessible spaces for the purpose of law enforcement unless certain exceptions apply (AI Act 5.2.2).

Most of the Person Re-ID systems do not comply with the laws stated in both the GDPR and the AI Act. The two main constraints that should be addressed and respected in the Person Re-ID are related to data storage and data transfer:

- **Data Storage**: GDPR, states, through its principles on "data minimization" (Chap.2 Art. 5.c) and "storage limitation" (Chap. 2 Art. 5.e), that personal data must be kept no longer than is necessary for the purposes for which data are collected and processed. It introduces strict requirements for data handling, including the need for secure storage and the limitation of data retention periods. The AI Act further complements these principles by emphasizing the ethical use of AI technologies, ensuring they comply with data protection laws and respect individuals' privacy.
- **Data Transfer**: GDPR ensures that the transfer of personal data should only occur under conditions that fully respect the protection of the data subjects' rights. In Chap. 5 Art. 44-49, the GDPR imposes strict requirements on data transfer that include transfers based on adequacy decisions, appropriate safeguards like Binding Corporate rules (BCRs) or Standard Contractual Clauses (SSCs), or specific derogations for particular situations. The AI Act's focus on transparency and accountability in the use of AI systems necessitates clear documentation and justification of data transfers, especially when deploying AI models that have been trained on datasets collected from diverse jurisdictions.

1.3 Research Focus and Contributions

Aside from earlier challenges in Person Re-ID, we are more particularly interested in addressing two mentioned challenges: domain shift and privacy. More precisely, this thesis concentrates on bridging the domain shift and enhancing privacy in Person Re-ID applications. The main motivation of our research is to develop solutions that will not only enable efficient performance across varied domains but also respect the ethical principles and the privacy of the individuals' data being processed.

Most of the papers addressing the problem of domain shift in Person Re-ID adopt strategies developed within the Unsupervised Domain Adaptation (UDA) field. However, these approaches often do not comply with the latest GDPR laws and AI Act requirements. In this thesis, we argue that there is a need for novel and appropriate design of UDA approaches to align with existing regulations. The work conducted in this thesis has led to three papers being accepted at international conferences and has also contributed to the filling of two patents. Our main contributions can be listed as follows:

1. **Adapting and Benchmarking UDA methods under data storage constraints**: we introduce and explore the Online Unsupervised Domain Adaptation (OUDA) setting for Person Re-ID. This novel setting aims at simultaneously addressing online adaptation and privacy protection. We extend three existing offline UDA frameworks to fit the OUDA setting. Conducting evaluations across four different datasets yields insightful findings on the performance of classical UDA approaches and their limitations.
    - Hamza Rami, Matthieu Ospici, and Stéphane Lathuilière. "*Online Unsupervised Domain Adaptation for Person Re-identification.*". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022.
2. **Improving adaptation under data storage constraints**: we introduce the Source Guided Similarity Preservation (S2P) framework, designed to address the challenges of catastrophic forgetting and domain shift in the OUDA setting. The flexibility of S2P allows it to seamlessly incorporate almost any existing UDA method to adhere to the privacy requirements. We test S2P on both real-to-real and synthetic-to-real OUDA tasks using four different datasets. S2P consistently outperforms prior state-of-the-art UDA methods, showing that it is possible to achieve significant performance while respecting the data privacy related to data storage.
    - Hamza Rami, Jhony H. Giraldo, Nicolas Winckler and Stéphane Lathuilière. "*Source-Guided Similarity Preservation for Online Person Re-Identification.*". In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.
    - European Patent No. 23305260.4: "*Method, device, and computer program for adapting an ANN model for person re-identification on a target domain.*"
3. **Adaptation without data transfer**: we introduce and explore the Distributed Unsupervised Domain Adaptation (DUDA-Rid) setting that does not allow any transfer of images from the cameras, making a first in the field. Moreover, we propose the FedProtoid algorithm that employs prototypes within a federated learning framework to simultaneously tackle distributed learning and domain shift. Through rigorous testing on both real-to-real and synthetic-to-real scenarios across different datasets, Fed-

Protoid demonstrates superior performance over existing state-of-the-art methods. Additionally, we introduce an enhanced version, Fed-Protoid++, that integrates Vision Transformers (ViTs) and leverages self-supervised pre-training techniques, resulting in additional performance improvements in DUDA-Rid.

- Hamza Rami, Jhony H. Giraldo, Nicolas Winckler and Stéphane Lathuilière. "*Privacy-Preserving Adaptive Re-Identification without Image Transfer*.", In Proceedings of the European Conference on Computer Vision (ECCV), 2024.
- European Patent No. 24305312.1: "*Method for performing privacy-preserving federated learning in the framework of re-identification.*"

Chapter 2

Literature Review

As previously outlined, this thesis proposes two innovative approaches leveraging Continual Learning and Federated Learning to address the domain gap challenge in deploying Person Re-ID systems while adhering to privacy constraints. We will start the literature review on the Person Re-ID topic by discussing the difference between early traditional feature extractors pivotal in Person Re-ID and the latest Deep Learning approaches that become crucial for overcoming the core obstacles inherent in Person Re-ID. Following this we present the datasets and evaluation metrics. We then transition to the discussion on Unsupervised Domain Adaptation (UDA) to highlight the key methods and contributions in this area for general computer vision applications and specifically for the Re-ID task. Finally, we introduce Continual Learning and Federated Learning as innovative and promising methodologies. These approaches are particularly pertinent since we integrate them with Person Re-ID to forge systems that are not only effective but also prioritize privacy preservation.

2.1 Person Re-Identification

Historically, the Person Re-ID research began with its association with multi-camera tracking (Fig. 2.1). Initially, the Person Re-ID was not formally recognized as a separate field and was mainly integrated with multi-camera tracking, which fuses appearance models with geometry calibration across disjoint cameras [5]. The term "person re-identification" made his first appearance in 2005, when researchers started focusing on re-identifying individuals as they re-entered the camera's field of view, using dynamic Bayesian network to model the conditional probability of the labels given features like color and spatial-temporal cues [6]. One year later in 2006, Gheissari *et al.* [18] marked a significant shift in the field by establishing person Re-ID as an independent task, separated from multi-camera tracking. Their work mainly focused on the visual cues of persons, using color and edge histograms for visual matching. The emergence of Convolutional Neural Networks (CNNs) [7] has made a big impact on the computer vision community. The great success in tasks like image classification [8] has affected almost all the research fields including Person Re-ID. This success has spread to Person Re-ID in 2014 where Yi *et al.* [9] employs for the first time a more general

way that can learn a similarity metric directly from image pixels using Siamese Networks. The same year, Li *et al*. [19] proposed a filter pairing neural network that jointly handles misalignment, photometric and geometric transforms, occlusions, and background clutter in Person Re-ID. Moreover, they introduced the largest benchmark Re-ID dataset (CUHK03) at that time to enable training end-to-end CNNs. These first works have laid a robust foundation that has led to big improvements in the field, inspiring other researchers to come up with new ideas and better understand how to adapt neural network architectures for the task of Person Re-ID.

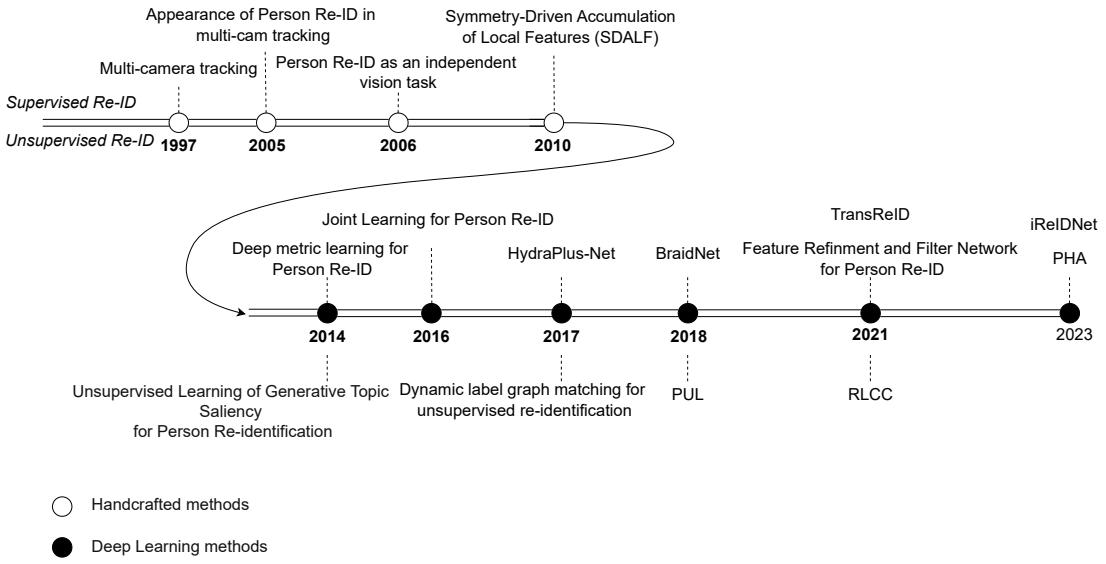


Figure 2.1: Milestones in the person re-ID history.

2.1.1 Handcrafted Feature Extractors

Before the advent of deep learning, handcrafted feature extractors were the predominant methods in computer vision tasks, including Person Re-ID. In this context, the person's appearance was mainly described by three principal visual characteristics: color, texture, and shape which serve as the foundation components for identifying individuals.

In 2006, Gheissari *et al*. [18] laid the groundwork by proposing a method that generates invariant signatures of persons by combining normalized color and salient edge histograms. Building on top of this, Gray *et al*. [20] proposed a novel approach to learn a set of viewpoint invariant features where each feature consists of a feature channel, a region, and a histogram bin, further enhancing the model's ability to recognize individuals from various viewpoints. As the research progressed, in 2010 Farenza *et al*. [21] presented an appearance-based method for Person Re-ID that consists of the extraction of features that model three complementary aspects of the human appearance: the overall chromatic content, the spatial arrangement of colors, and the presence of recurrent local motifs with high entropy. Following this, Bak *et al*. [22] developed a methodology that focuses on learning a model that

selects the most descriptive features from color, intensity, gradients, and filter responses for a specific class of objects, to optimize feature selection within a covariance metric space guided by an entropy-driven criterion. Continuing the trend of innovation, in 2014 Das *et al*. [23] improved consistency between camera pairs with appearance signatures constructed using HSV color histogram on horizontal sub-regions specifically with the torso and leg areas. One year later, Liao *et al*. [24] introduced an effective feature representation called Local Maximal Occurrence (LOMO). The LOMO feature extractor analyzes the horizontal occurrence of local features and maximizes the occurrence to make a stable representation against viewpoint changes.

The decline in the popularity of handcrafted feature extractors in favor of Deep Neural Networks (DNNs) can be attributed to several key disadvantages. First, the engineering process of handcrafted features needs considerable domain-specific knowledge which makes them more challenging to develop. Second, handcrafted feature extractors are not well-designed for large-scale datasets that exhibit a wide range of variability.

2.1.2 Deep Person Re-Identification

Recently, DNNs have emerged as the primary method for feature extraction in the Person Re-ID task, marking a significant departure from traditional handcrafted feature extractors. Unlike their handcrafted counterparts, DNNs have demonstrated remarkable performance on large-scale datasets, showcasing their robustness and versatility. Many attempts were made to improve the adaptability of DNNs in providing solutions tailored to the specific demands of person ReID, further solidifying their position as the preferred choice in the field. Similar to handcrafted feature extractors, deep learning architectures aim at learning discriminative and robust spatial feature representations to describe human appearance.

Following the publication of the CUHK-03 dataset, Wang *et al*. [25] proposed a joint Single-Image Representation (SIR) and Cross-Image Representation (CIR) learning framework based on CNN, where the SIR and CIR feature representations are jointly optimized to achieve better cross-camera person matching performance. Qian *et al*. [26] proposed MuDeep which is a novel multi-scale deep learning model for Re-ID based on Siamese network. MuDeep can learn features at different scales and evaluate their importance for cross-camera matching. However, as the number of scales increases, the model needs to learn a large number of parameters and thus has high computational burden. For this reason, Wang *et al*. [27] designed Deep Anytime Re-ID model that combines effective feature embeddings built on the four blocks of ResNet50 [28], hence resulting in the first Re-ID algorithm applicable in the presence of resource constraints. In this context, the authors of OS-Net [29] also designed a lightweight Re-ID specific network inspired by MobileNet [30].

To further enhance the ability of the feature representations to distinguish different person identities, Wang *et al*. [31] designed a deep model called BraidNet that has a specially designed cascaded WConv structure that learns to extract the comparison features of two images, which are robust to misalignments and color differences across cameras. Hou *et al*. [32] proposed a novel structure called Interaction and Aggregation (IA) to enhance the feature representation capability of CNNs. Firstly, the Spatial IA (SIA) module models the

inter-dependencies between spatial features and then aggregates the correlated features corresponding to the same body parts. Secondly, the Channel IA (CIA) module selectively aggregates channel features to enhance the feature representation, especially for small-scale visual cues.

Another research trend focused on discriminative local feature representation by partitioning a human image into multiple cells. Motivated by the idea to alleviate the problems of occlusion, boundary detection errors, view, and pose variations, local feature learning-based methods aim at extracting discriminative local features of individuals. To this end, researchers explored attention as a powerful module to find areas that have a greater impact on the feature map and to focus the model on local parts of the body appearances to correct misalignment and eliminate background perturbance. Liu *et al*. [33] proposed a new attention-based DNN, named as HydraPlus-Net (HPnet), that multi-directionally feeds the multi-level attention maps to different feature layers. In this context, Li *et al*. [34] showed the advantages of jointly learning attention selection and feature representation using a novel Harmonious Attention CNN. To further obtain better local fine-grained features of a person, Ning *et al*. [35] proposed a feature selection network that combines global and local fine-grained features to realize Person Re-ID. Recently, Xu *et al*. [36] proposed a novel Re-ID network named iReIDNet which can effectively extract local and global multi-granular feature representations by a well-designed spatial feature transform and coordinate attention mechanism together with improved global pooling.

Local feature learning can provide comprehensive information about specific pedestrian regions, but pose and occlusion fluctuations may compromise the accuracy of local features. To improve the final feature representation, several researchers frequently combine fine-grained local features with coarse-grained global features. Wang *et al*. [37] designed a multi-granularity feature learning strategy combining global and local feature representations. In order to reduce the negative effects of inaccurate bounding boxes on pedestrian matching, Zheng *et al*. [38] introduced a pyramid model that transitions from coarse-grained to fine-grained, incorporating both local and global pedestrian information along with progressive cues ranging from coarse to more detailed features.

However, these strategies often increase the learning difficulty and are not efficient or robust to real-world scenarios. In this context, He *et al*. [39] proposed for the first time a transformer-based Person Re-ID framework named TransReID. In addition to the architecture of the transformer, they designed a novel module called Jigsaw Patch Module (JPM) that rearranges the patch embeddings via shift and patch shuffle operations which generates robust features with improved discrimination ability. To further enhance the effectiveness of TransReID, Zhang *et al*. [40] proposed a Patch-wise High-frequency Augmentation (PHA) method that splits patches with high-frequency components by the Discrete Haar Wavelet Transform, then empowers the ViT to take the split patches as auxiliary input. Zhang *et al*. [41] also proposed a novel end-to-end framework that combines global and local feature representations and captures the body structural information by modeling the spatial relation between patches using graph neural networks (GNN).

Although supervised Person Re-ID approaches perform well, the high labeling costs prevent them from scaling to huge unlabeled datasets and new domains. Consequently, because of its capacity to resolve the scalability problem in person Re-ID, unsupervised person Re-ID

has gained more and more attention. In the following section, we will explore recent developments in the field of unsupervised Person Re-ID, where the reliance on labels is limited.

2.1.3 Unsupervised Person Re-Identification

The approaches to unsupervised Person Re-ID fall into two main categories: **fully unsupervised learning** (USL) and **unsupervised domain adaptation** (UDA) based methods. In this section, our focus will be on USL methods, while UDA approaches will be covered in subsequent sections.

Unsupervised Re-ID has been investigated around the same time as supervised Re-ID. Similarly, before the deep learning era USL Re-ID focused on how to construct robust feature representation manually. Traditional methods have mainly focused on feature engineering, which designs appropriate handcrafted features using prior expert knowledge. Farenzena *et al*. [42] presented an appearance-based method that consists of the extraction of features that model three complementary aspects of the human appearance: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy. Motivated by the idea that human eyes can recognize person identities based on some small salient regions, Zhao *et al*. [43] proposed a novel perspective for Person Re-ID based on unsupervised salience learning.

With the emergence of deep learning architectures, several challenges were being addressed. Regarding the lack of ground-truth identity labels, pseudo-label estimation was proposed. Early works proposed the use of graph models to represent samples and perform dynamic graph matching for cross-camera labeling [44]. An alternative method that has attracted increased attention for generating pseudo-labels is clustering-based. This method utilizes clustering algorithms, such as K-means and DBSCAN [45], to progressively group training samples into clusters. The cluster IDs are then used as pseudo-labels to train Re-ID models. For instance, Fan *et al*. [46] proposed PUL, a progressive process that iterates between clustering and fine-tuning.

However, a common problem in clustering-based pseudo-label estimation is the presence of noisy labels that harm the guidance of the model training. To refine the pseudo-labels, NRMT [47] employs a dual-network approach during training to select samples and conduct collaborative clustering with those chosen samples. This method is akin to the ACT [48] method, which also utilizes two networks to differentiate between pure and diverse samples. Noisy labels can also result from inaccuracies in estimating the number of identities when using the K-means clustering algorithm. To this end, Yutian *et al*. [49] used a hierarchical bottom-up clustering which can visit all samples and determine the similarity of samples. Despite its effectiveness, hierarchical clustering struggles to distinguish between hard samples—visually similar individuals with different identities—often merging these hard samples into the same cluster. To reduce the impact of hard samples, Zeng *et al*. [50] proposed HCT, a hierarchical clustering-guided Re-ID that utilizes PK sampling in each iteration. This involves randomly selecting K samples from P identities for training. Zhang *et al*. [51] made the first attempt to leverage the spirit of temporal ensembling to tackle the problem of pseudo-label noise. Cho *et al*. [52] proposed a Part-based pseudo-label Refinement (PPLR)

framework that exploits both the global and local context of images to alleviate the label noise. Recently, in the continuation of transformer-based Re-ID [39], Luo *et al*. [53] first trained the TransReID model in a self-supervised way on the large-scale dataset LUPerson [54] then adapted C-Contrast [55] for unsupervised fine-tuning.

Person Re-ID can also be viewed as a cross-camera retrieval task aimed at learning a model to discriminate images of individuals from different camera views. In line with this, another family of methods called Camera-Aware Feature Learning has been established, which achieves comparable results to Clustering-based methods [56]. Cross-domain Mixup [57] was applied by conducting interpolation on the data manifold, which is similar to GAN-based image style transfer. Meta-learning was introduced for the task of Re-ID in [58], where the authors proposed camera-aware meta-learning (MetaCam) aiming to learn camera invariant representations by simulating the cross-camera Re-ID process during training. Finally, the Side Information Embeddings module (SIE) was proposed for TransReID [39] that plugs in learnable embeddings to mitigate feature bias toward camera variations. Despite the effectiveness of Camera-Aware approaches, this thesis places greater emphasis on Pseudo-Labeling methods, as they have consistently demonstrated superior performance in benchmarks, achieving state-of-the-art results across multiple datasets and settings. Moreover, Pseudo-Labeling methods are highly scalable as they do not require explicit annotations from every camera perspective, which is often impractical in large-scale surveillance systems.

2.1.4 Loss functions

Before the deep learning era, metric learning, which involves methods to learn distances between pairs of images within the feature space, was the subject of many years of research. Researchers focused on techniques such as the learning of the Mahalanobis distance function [59], or the projection matrix [60]. With the advent of deep learning, the design of loss functions has become central to metric learning in guiding feature representation learning. In the context of the Person Re-ID task, three main loss functions and their modifications have been studied: Identity Loss, Triplet Loss, and Contrastive loss.

Identity loss. The training process of the Person Re-ID task can be treated as a multi-class classification problem [61]. Similar to image classification, the ID-discriminative embedding network treats each individual as a separate class and uses the ID as a classification label. Hence, the ID loss can be expressed as a cross-entropy loss as follows:

$$\mathcal{L}_{id} = -\sum_{a=1}^{K} q_a(\boldsymbol{x}) \log p(y_a|\boldsymbol{x}) \tag{2.1}$$

Where K is the number of identities, $q_a(\boldsymbol{x}) = 1$ if the label of the sample image $\boldsymbol{x}$ is a, otherwise $q_a(\boldsymbol{x}) = 0$. $p(y_a|\boldsymbol{x})$ is the probability that the picture $\boldsymbol{x}$ is predicted as ID y_a using the softmax activation function. Several studies suggested alterations to the ID loss by exploring other softmax variations such as Deep Cosine Metric [62]. Although effective, relying solely on ID loss is not enough to learn a model with sufficient generalization ability. Therefore, ID loss often requires a combination with other losses to regularize the model.

Triplet loss. Triplet loss is commonly used for Person Re-ID. Motivated by the idea to ensure that an image $\boldsymbol{x}^a$ (anchor) of a specific person is closer to all other images $\boldsymbol{x}_i^p$ (positive) of the same person than it is to any images $\boldsymbol{x}_i^n$ (negative) of any other person, the Triplet Loss [63] can be formalized as follows:

$$\mathcal{L}_{trip} = \sum_i \left[||\mathbf{f}(\boldsymbol{x}_i^a) - \mathbf{f}(\boldsymbol{x}_i^p)||_2^2 - ||\mathbf{f}(\boldsymbol{x}_i^a) - \mathbf{f}(\boldsymbol{x}_i^n)||_2^2 + \alpha \right] \tag{2.2}$$

Where α is a non-negative margin enforced between positive and negative pairs. A major drawback of this formulation is that the number of triplets grows cubically as the dataset expands. To this end, Hermans *et al*. [64] proposed batch-hard triplet loss which consists of first forming batches by randomly sampling P classes (identities), and randomly sampling K images of each person, resulting in a batch of PK images. Then, for each sample a in the batch, the hardest positive and the hardest negative samples are selected from the batch to compute the following loss:

$$\mathcal{L}_{BH-trip} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[\max_p ||\mathbf{f}(\boldsymbol{x}_i^a) - \mathbf{f}(\boldsymbol{x}_i^p)||_2^2 - \min_n ||\mathbf{f}(\boldsymbol{x}_i^a) - \mathbf{f}(\boldsymbol{x}_i^n)||_2^2 + \alpha \right] \tag{2.3}$$

Combining both Triplet loss and ID loss has proven to be effective in several Re-ID methods [65, 66, 67]. Moreover, there exist other variations of the Triplet loss, including soft triplet loss for knowledge distillation [68].

Contrastive loss. Contrastive loss aims at pulling together images from the same identity and pushing away the images from different identities in the feature space. It was traditionally designed for Siamese Network-based Person Re-ID [31, 69]. Recently, Ge *et al*. [70] proposed SpCL which is a contrastive-based learning approach that integrates a memory bank to store centroids $\mathbf{c}_k$ of the identity clusters. Given $\mathbf{v}$ a feature representation of a given image, the unified contrastive loss can be expressed as:

$$\mathcal{L}_{\text{contrastive}}(\mathbf{v}) = -\log \left(\frac{\exp(\langle \mathbf{v}, \mathbf{z}^+ \rangle / \tau)}{\sum_k \exp(\langle \mathbf{v}, \mathbf{c}_k \rangle / \tau)} \right) \tag{2.4}$$

where $\mathbf{z}^+$ indicates the positive class prototype corresponding to $\mathbf{v}$ and τ is the temperature parameter.

2.1.5 Datasets and Evaluation Metrics

In the literature, we have witnessed multiple Person Re-ID datasets that can be categorized as follows:

Single-Shot datasets. Single-shot datasets provide just one query image and one true match image in the gallery. These datasets were essential for training traditional models that relied on hand-crafted feature extractors. Among the most widely used single-shot datasets are GRID [71], ViPeR [72], and CUHK01 [73], containing thousands of images.

Historically, person images were usually identified by manually annotating the images with hand-drawn bounding boxes. This emphasizes the attention to detail needed in the early stages of Person Re-ID.

Multi-Shot datasets. The shift towards the deep learning era has impacted the Person Re-ID landscape, particularly with the introduction and use of Siamese networks [9]. These networks, which often input a pair of images including an original and its augmented counterpart, are adept at evaluating the similarity between two images of the same identity. Early multi-shot datasets like CAVIAR [74] and PRID [75] were further expanded upon by subsequent datasets such as CUHK03 [19], Market1501 [76], DukeMTMC [77], and MSMT17 [78]. These later datasets introduced a larger number of identities and images, which enriches the diversity and complexity of data available for the Person Re-ID task. As the person detection field advanced, automatic detection and tracking algorithms became integral to the process. Deformable Parts Model (DPM) [79], Faster RCNN [80], and Yolo-v5 [81] have been used to extract person bounding boxes more effectively. These advancements have not only accelerated the data preparation process but also improved the accuracy and reliability of Person Re-ID systems.

Large-Scale datasets: Unlabeled and Synthetic. Because of the privacy concerns associated with collecting real images of individuals, synthetic person datasets have been proposed. These datasets such as PersonX [82] and RandPerson [83] mitigate privacy concerns while also offering valuable resources for Re-ID studies as alternatives to real datasets. To further enhance the generalization ability of Re-ID models, large-scale unlabeled datasets collected from the internet like LUPerson [54] have been proposed. They are specifically designed to replace the pre-trained models on ImageNet [84], enabling self-supervised learning that improves the performance of the Re-ID models.

Table 2.1 presents a summary of the characteristics of the various aforementioned datasets for the Re-ID task. It illustrates the shift from single-shot to multi-shot datasets, characterized by an increase in the number of images and cameras, as well as the emergence of large-scale datasets that can be either real and unlabeled, collected from the internet [54], or synthetic [82, 83].

Evaluation metrics. The commonly used evaluation metrics for Person Re-ID are the Cumulative Matching Characteristics (CMC) and the mean Average Precision (mAP). In the case of a single-shot setting, the CMC top-k accuracy can be formulated as follows:

$$\text{Acc}_k = \begin{cases} 1 & \text{if top-}k \text{ ranked gallery samples contain the query identity} \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

The final CMC score is then computed by averaging the top-k accuracies over all the queries. For multi-shot datasets, where each identity has multiple images, we randomly sample one instance for each gallery identity and compute a CMC curve like in a single-shot setting. The random sampling is repeated for N times and the expected CMC score is reported.

Table 2.1: Person Re-Identification Datasets

Dataset	Release	# Ids	# Cams	# Imgs	Bounding Box Generation	Crop Size	Multi-shot
VIPeR [72]	2007	632	2	1264	Hand	128X48	
GRID [71]	2009	1025	8	1275	Hand	Vary	
CAVIAR4ReID [74]	2011	72	2	1220	Hand	Vary	✓
3DPeS [85]	2011	192	8	1011	Hand	Vary	✓
PRID [75]	2011	934	2	24541	Hand	128X64	✓
CUHK01 [73]	2012	971	2	3884	Hand	160X60	
CUHK02 [86]	2013	1816	2	7264	Hand	160X60	✓
CUHK03 [19]	2014	1467	2	13164	Hand/DPM	Vary	✓
Market1501 [76]	2015	1501	6	32217	Hand/DPM	128X64	✓
DukeMTMC-reID [77]	2017	1812	8	36441	Hand	Vary	✓
MSMT17 [78]	2018	4101	15	126441	Faster RCNN	Vary	✓
PersonX [82]	2019	1266	6	273456	Synthetic	Vary	✓
RandPerson [83]	2020	8000	19	1.8M	Synthetic	Vary	✓
LUPerson [54]	2021	200000	-	4M	YOLO-v5	Vary	✓

The mAP is also used to evaluate the overall performance of Re-ID models. For each query, we first calculate the area under the Precision-Recall curve, also known as the average precision (AP). In the context of Re-ID, the AP can be formulated as follows:

$$\mathrm{AP} = \frac{1}{N} \sum_{k} \mathrm{P}_k \mathrm{rel}_k \tag{2.6}$$

where N is the number of ground-truth positives, P_k refers to the precision at rank k, which can be also computed using the ratio between the number of correct matches and k. And finally, the rel_k is an indicator function that is equal to 1 if we get a correct match at rank k, and 0 otherwise. The mean value of APs of all queries is then calculated, which results in the final mAP of the Re-ID model.

2.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) is one of the special settings of transfer learning, which aims to leverage knowledge from an abundant labeled source domain to learn effective predictors for the target domain with limited or no labels. This process necessitates addressing the challenge of domain shift, ensuring that the knowledge transferred is relevant and effective despite the differences between the source and target domains. Before citing the different UDA methods, the early works of Ben-David *et al*. [87] mainly focused on the generalization bound for the problem of domain adaptation. It has been shown that the target domain error can be minimized by bounding the source domain error and the discrepancy between them. To this end, UDA methods not only optimize the model with the source domain but also ensure that the discrepancy between source and target is minimized.

The first category of UDA techniques is the discrepancy-based methods. They aim to decrease the discrepancy between the two domains and align both data distributions. This is done by usually adding different distance loss functions in the activation layers of networks.

The Maximum Mean Discrepancy (MMD) is one of the most popular distances in minimizing the discrepancy between two distributions. It measures the squared distance between the embeddings of the two distributions in a reproducing kernel Hilbert space. Based on this, Tzeng *et al*. [88] proposed a Deep Domain Confusion (DDC) model that is trained with a loss that combines both the cross-entropy and the MMD loss. The same authors extended the DDC model by introducing soft label distribution matching loss [89]. Correlation Alignment (CORAL) loss [90] has been introduced as an alternative to MMD loss. CORAL aims to align second-order statistics (co-variances) between the cross-domain distributions. Other losses like Jensen-Shannon Divergence (JSD) [91, 92] and Wasserstein Distance [93, 94] were also deployed to decrease the discrepancy between source and target domains.

With the advent of Generative Adversarial Nets (GANs) [95], adversarial learning models have been found to be effective in identifying invariant representations in domain adaptation. DANN [96] is one of the first adversarial methods for adversarial-based UDA. It consists of integrating a gradient reversal layer to enhance the discrimination of source and target domains. Similarly, ADDA [97] used an inverted label GAN loss to split the source and target domains. Combining MMD with adversarial learning also showed promising results in adaptation. For instance the Joint Adaptation Network (JAN) [98] combined MMD with adversarial learning to align the joint distribution of multiple domain-specific layers across domains. To further improve the results, several approaches incorporate image-level adaptation to preserve image consistency throughout training, aiding in feature alignment. Progressive domain adaptation, as described in [99], integrates feature alignment with image-level adaptation. This method initially employs a model to transform images from the source to an intermediate domain through image translation. The transformed images retain their original labels from the source domain and serve as simulated training images for the target domain. Subsequently, alignment is performed between the intermediate and target domains. Additionally, Zhang *et al*. [100] introduced a technique that adjusts the weights of target samples that may potentially mislead the domain discriminator.

As an alternative to adversarial and discrepancy-based methods, pseudo-labeling techniques have been proposed. These techniques leverage unlabeled data from the target domain as a training mechanism to facilitate domain adaptation [101]. This category of methods, originally designed for semi-supervised learning, follows a two-step process: (1) Generate pseudo-labels in the target domain based on the model's confidence scores and clustering, and (2) fine-tune the model using the generated pseudo-labels with target domain data. Generally, the source model is treated as the initial pseudo labeler to generate the pseudo-labels. Saito *et al*. [102] proposed a novel asymmetric tri-training method to generate pseudo-labels. Two networks are used to assign pseudo-labels to unlabeled samples and the remaining network is trained by the pseudo-labeled target samples. To include the semantic information in the images, Xie *et al*. [103] proposed a moving semantic transfer network (MSTN) to achieve semantic matching and domain adversary losses to obtain pseudo-labels. Rather than exclusively relying on the predicted class probability of the generated pseudo-labels, Zhang *et al*. [100] proposed a sample re-weighting strategy, such that a selected sample is assigned a higher weight when it is not close to the source samples, and vice versa. Recently, Litrico *et al*. [104] proposed a re-weighting of the classification loss based on the reliability of the pseudo-labels that is measured by estimating their uncertainty, which brings robustness against the noise that inevitably affects the pseudo-labels.

Pseudo-labeling techniques have also been developed for Person Re-ID as a technique to facilitate domain adaptation. Similar to broader practices in this field, these methods typically generate pseudo-labels by clustering all cross-camera samples based on visual similarity, and then a model is fine-tuned as a classification task. Wu *et al*. [105] proposed a Clustering and Dynamic Sampling (CDS) method that iteratively clusters the target samples into several centers and dynamically selects informative ones from each center to fine-tune the source domain model. Since clustering may lead to data imbalance in clusters, Ding *et al*. [106] proposed to use cluster validity as the guidance and derive a dispersion-based criterion that promotes compact and well-separated clusters. AD-Cluster [107] is an augmented discriminative clustering method that trains a generative model to generate images of different identities in different camera styles and improve the discriminative capability of the Re-ID model with the augmented clusters. Yang Fu *et al*. proposed a Self-Similarity Grouping (SSG) [67] approach that assigns different pseudo-labels to both global and local features. To mitigate the effects of noisy hard pseudo-labels, Mutual-Mean Teaching (MMT) [68] (Fig. 2.2), introduced by Yaxiao *et al*., employs a teacher-student framework, involving two networks. These networks are trained jointly, using hard pseudo-labels generated by both networks and soft pseudo-labels generated by their mean networks, to refine the pseudo-labels in the target domain. SpCL [70] is another method based on contrastive learning which employs a hybrid memory that stores and continually updates the centroids. However, the clustering result suffers from the data distribution discrepancy. To address this issue, Zhang *et al*. [108] proposed a heterogeneous graph to promote the domain adaptation and the quality of pseudo-labels simultaneously. Meanwhile, Zheng *et al*. [109] introduced a Group-aware Label Transfer (GLT) algorithm which enables the online interaction and mutual promotion of pseudo-label prediction and representation learning.

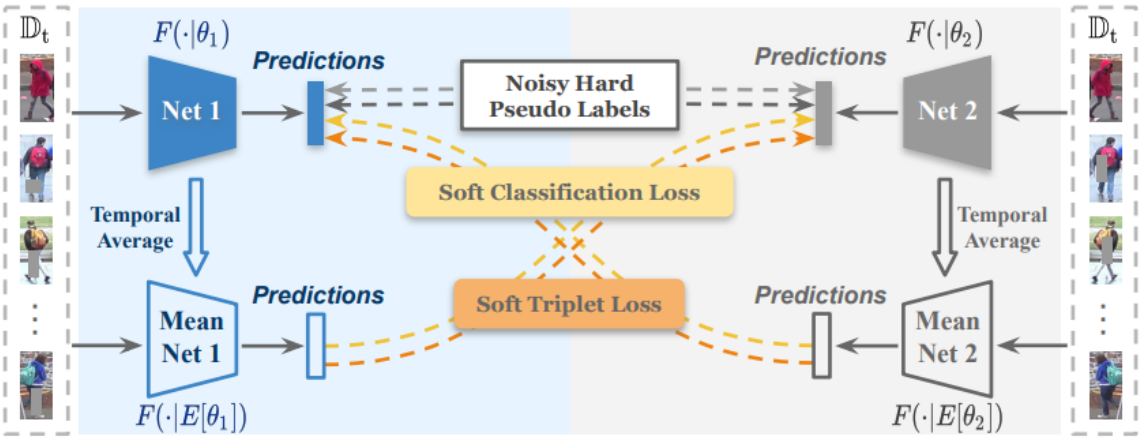


Figure 2.2: Illustration of the MMT framework. Source: [68].

An alternative approach [110, 111, 112] to the methods previously discussed attempts to narrow the gap between the source and target domains through the use of intermediate domains. Traditional techniques [110, 111] embed the source and target domains into a Grassmann manifold and learn a specific geodesic path between the two domains. In deep learning methods [113, 114], they either use GANs to generate a domain flow by reconstructing input images on pixel level [113] or learn better domain-invariant features by bridging the learning of the generator and the discriminator [114]. The same approach was explored

for Person Re-ID. Dai *et al.* [115] proposed an Intermediate Domain Module (IDM) that can learn intermediate domains' representations on the fly by mixing the source and target domains' hidden representations. In this thesis, we will conduct a thorough examination of these UDA techniques that are based on the pseudo-labeling approach. We will detail the main challenges these methods face in the context of Person Re-ID and provide insights and improvements to ensure compliance with the imposed privacy regulations.

2.3 Continual Learning

Continual Learning (CL) emerges as a foundational strategy to facilitate lifelong capabilities in the context of developing robust and adaptive AI systems. It consists of training a model to gradually acquire, retain, and refine knowledge over time across a variety of tasks and/or domains. A major challenge, known as *catastrophic forgetting* [116], emerges when training continuously a model on new data distributions, generally leading to a decline in performance on previously learned tasks. This problem is symbolic of the fundamental trade-off between plasticity and stability, where plasticity refers to the ability of the model to adapt to new tasks by modifying its parameters, and stability refers to the capacity of the model to retain previously acquired knowledge [117, 118]. In the field of CL, scenarios are generally classified according to two factors: the division of incremental batches and the availability of task identities. These factors specify how the data is presented to the model. For instance, in the *Task-Incremental Learning* (TIL) the tasks have disjoint data label spaces, and the model can access the task identities to switch between different specialized modes or parameters [119, 120]. *Domain-Incremental Learning* (DIL) is another scenario where data share the same label space but come from different input distributions [119]. Finally, *Class-Incremental Learning* (CIL) is another scenario where the data are incrementally provided in batches, with each batch introducing new classes. In CIL, no task identities are available, requiring the model to continuously adapt to recognize new classes alongside old ones [121].

CL methods can broadly be categorized into three categories. The first category is the *regularization-based* methods, where explicit regularization terms are added to balance the old and new tasks. Early works focused on selectively regularizing the changes in network parameters to minimize forgetting. For instance, EWC [116] introduced a quadratic penalty in the loss function, which penalizes parameter variations based on their importance to previous tasks, where the importance is estimated using the Fisher Information Matrix (FIM) [122]. Rather than computing offline the FIM, SI [123] proposed computing the per-synapse consolidation strength in an online fashion and over the entire learning trajectory in parameter space. Inspired by neuroplasticity, Aljundi *et al.*[124] introduced MAS which is a novel approach to CL that computes the importance of the parameters in an unsupervised and online manner. The importance of each parameter is computed based on how sensitive the predicted output function is to a change in this parameter. finally, Chaudhry *et al.* [125] presented an improved EWC++ that combines the regularization terms of both EWC [116] and SI [123] to integrate their advantages. Within the same direction of research, recent works focused on the network itself. Instead of consolidating parameters, NPC [126] estimates the importance of each neuron and reduces its learning rate accordingly. AGS-CL [127] freezes the parameters connecting the important neurons, which is equivalent to a hard version of

weight regularization. Moreover, the AGS-CL suggests re-initializing the weights associated with unimportant nodes after learning each task to prevent the negative transfer. Regularization can also target the intermediate or final output of the prediction function. This is typically done by employing knowledge distillation [128] and a teacher-student framework where the teacher is the previously learned model and the student is the currently trained model. LwF [129], a pioneer work in this direction, proposed to learn new training samples while using their predictions from the output head of old tasks to compute the distillation loss. LwM [130] presented Attention Distillation Loss that preserves the acquired knowledge without storing any data, by penalizing the changes in classifiers' attention maps. As we go through this thesis, we will discuss in more detail the teacher-student framework and how it can be adapted to representation learning in the context of Person Re-ID.

The second category of CL methods is the *replay-based* methods, which typically store a few old training samples within a small memory bank. The key challenges of this category of methods are how to *construct* and how to *exploit* the memory bank. For the construction, early works adopt fixed principles for sample selection such as reservoir sampling [131], ring buffer [132] that ensures an equal number of old samples per class, and mean-of-feature [133] which selects an equal number of old samples that are closest to the feature mean of each class. To further improve storage efficiency, GMED [134] is a proposed framework for editing stored examples in continuous input space via gradient updates, to create more challenging samples for replay. As for the exploitation, replay-based methods require an adequate use of the memory bank to recover past knowledge. Yiduo *et al*. [135] proposed a novel approach based on mutual information maximization. Sun *et al*. [136] studied the differential influence of training examples using a novel MetaSP algorithm. In the same context, Zhicheng *et al*. [137] managed to identify a new class of second-order influences that gradually amplify incidental bias in the replay buffer and compromise the selection process. As discussed earlier, these methods necessitate retaining examples from previous tasks, a practice that may be limited in many scenarios, like Person Re-ID, where privacy concerns arise. In response, generative replay has been proposed as an alternative to storing samples. However, it requires training an additional generative model to replay generated data, which can sometimes demand significant computational resources [138, 139, 140].

The methods previously discussed utilize a common set of parameters across tasks, often leading to interference between tasks. In contrast, developing task-specific parameters can directly address this issue, prompting researchers to concentrate on a different group of methods known as *architecture-based* methods [141]. Based on whether the model parameters expand with the number of tasks, architecture-based methods can be categorized into two types: fixed capacity or capacity-increasing. The first sub-category of fixed capacity frameworks usually selects for each task a sub-network from the CL model to achieve knowledge transfer [142, 143, 144]. Capacity-increasing frameworks prevent forgetting old tasks and adapt to new ones by introducing new task-specific parameters for each additional task while freezing parameters related to old tasks [145, 146, 147].

In this thesis, we will focus on the first category of CL methods that are *regularization-based*. This choice is driven by the privacy concerns related to the task of Re-ID. As discussed in the previous chapter, recent privacy regulations prohibit the storage of previously captured images that arrive as a one-pass data stream, making *replay-based* methods unsuit-

able. Additionally, our specific scenario is online incremental learning, where the task of Re-ID in the CL process remains consistent. This setting limits the applicability of *architecture-based* methods, which often require distinct parameters for different tasks.

2.4 Federated Learning

Federated Learning (FL) is a machine learning paradigm proposed to meet the increasing demand for collaborative learning across various decentralized devices or data holders while preserving privacy [148]. Fundamentally, FL preserves by design data confidentiality and privacy by allowing multiple clients (or edge devices)-each with their own local data- to participate in the development of a global model without having to exchange data directly. This is achieved by training locally in a distributed manner a model with each client's local data and then sending the updated versions to a central server. The central server then aggregates these updates to improve the global model, using an aggregation rule, which is sent back to all the clients for further training (Fig. 2.3). Let's consider N models $\{M_i\}_{i=1}^{N}$ that collaborate jointly on separated datasets $\{\mathcal{D}_i\}_{i=1}^{N}$ to optimize the learning of a centralized server model M_{server}. A federated learning strategy is effective if the predictive accuracy A_{server} of M_{server} respects the following conditions:

$$\begin{aligned} &|A_{server} - A_{center}| < \epsilon \\ &A_{server} > A_i, i = 1, ..., N \end{aligned} \tag{2.7}$$

Where A_{center} is the predictive accuracy of a model trained on the union of all the datasets, A_i is the predictive accuracy of the model M_i if it was trained separately only on D_i, and ϵ a small non-negative constant. We say that the algorithm for FL has ϵ accuracy loss [149].

FL has been applied to various computer vision tasks like image segmentation [150], classification [151], and person Re-ID [152]. Despite its growing adoption, optimizing federated learning frameworks encounters three primary challenges: 1) Non-IID data: the data distributions vary across different clients. 2) Unbalanced data: there is inconsistency in the volume of data available with each client. 3) Limited communication: practical applications favor methods that require fewer communication rounds and can operate under low bandwidth, owing to network instability and security considerations.

FederatedAveraging (FedAvg), first introduced by McMahan [148], employs a strategy where local models trained with local data are averaged on the server, and this averaged model is then redistributed to the clients. The protocol stipulates that each client model starts from the same random initialization. During each communication round, the server disseminates the aggregated model to the clients. The weighting of client models in FedAvg is based on the quantity of data each client has, although this does not necessarily reflect the variations in domain distribution across the data. However, this approach can neglect data-poor clients, since clients with larger amounts of data have a greater influence on the quality of the global model, which results in a suboptimal solution. To address this issue, researchers have proposed alternatives to adjust the weights of the clients. For instance, FedDisco, introduced by [153], proposed using the difference between local and global parameter distributions as a complementary metric for aggregation weights. Similarly, the authors of another work

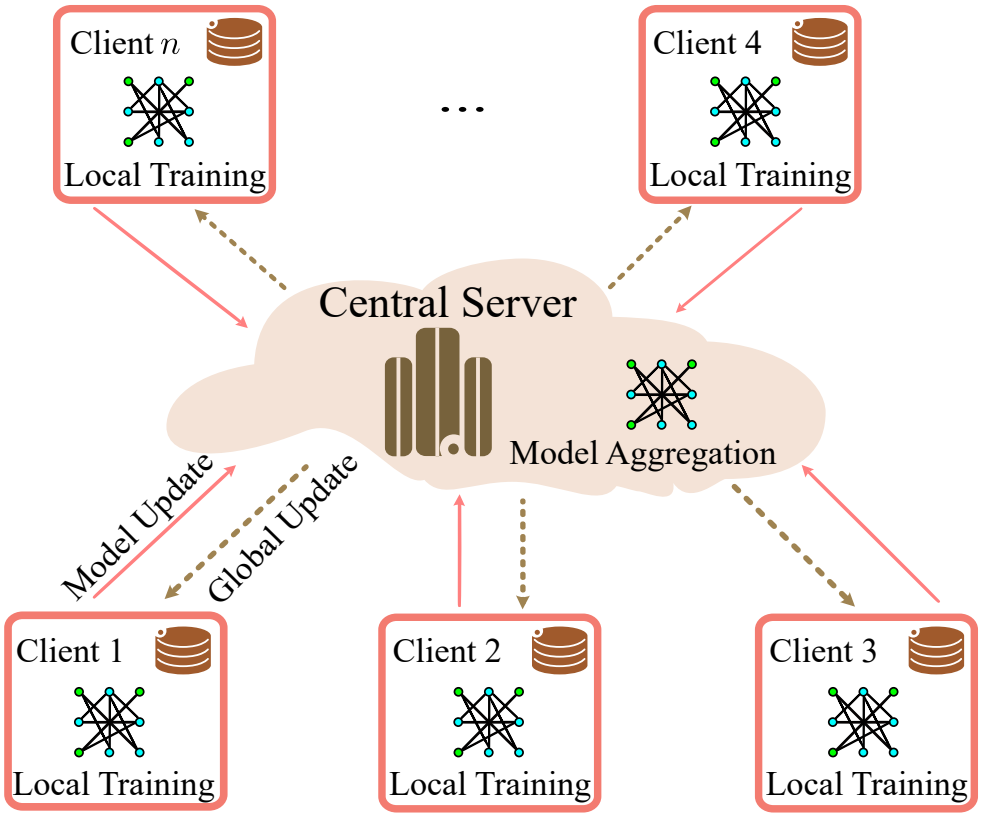


Figure 2.3: The basic framework of FL.

[154] suggested adaptively assigning different weights to clients based on their contribution in each round, which can be measured by contrasting the local gradient vector with the global gradient vector.

Building on the concepts of FedAvg, FedProx [155] introduces a proximal term to the local objective functions to better handle data heterogeneity and enhance model stability. This proximal term helps maintain local updates closer to the global model, improving consistency and reducing the impact of statistical heterogeneity. This approach allows for variable local workloads, addressing both statistical and systems heterogeneity:

$$\min_{w} F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 \tag{2.8}$$

where $\mathbf{w}^t$ represents the server model weights after round t.

Further expanding the landscape of federated learning solutions, other frameworks such as FedAsync [156] implement asynchronous model updates using a weighted average approach. Similarly, FedShare [157] introduces a shared public dataset among clients to mitigate weight divergence issues. Additionally, FedMeta [158] integrates meta-learning techniques into the federated learning framework, offering a novel approach to model training across decentralized environments. Each of these developments represents a step toward addressing the unique challenges posed by the federated learning paradigm, specifically regarding data heterogeneity, communication constraints, and model synchronization.

FL has recently been explored in the context of Person Re-ID, where privacy concerns

are critical due to the personal nature of the images involved. This makes the application of federated learning not just useful but essential. However, it raises an interesting question: How effective is it to directly apply existing federated learning approaches to the specific challenges of Person Re-ID? Research so far has tackled this by either tweaking federated learning optimization techniques to better suit Person Re-ID models or by adapting cutting-edge Person Re-ID methods to fit into the federated learning framework. The initial research in the field of Supervised Person Re-ID within a federated learning framework was introduced in *FedReID* [159]. This paper proposed a Federated Partial Averaging (FedPav) optimization technique, applying FedAvg solely to the feature extractor (backbone) while allowing the classifiers to be trained independently on each client's dataset. Observing that local models often outperformed the aggregated model, the authors enhanced their approach by incorporating knowledge distillation. This method treats all client models as teachers and the central server model as a student, aiming to minimize the $L2$ norm between the clients' average outputs and the server's output. A subsequent study, *FedUnReID* [160], adapted the well-known unsupervised baseline for Person Re-ID, Bottom-Up Clustering (BUC) [49], to the federated setting. This work included the development of a Controller to manage personalized epochs and a Profiler to aid in personalized clustering at the edge devices, while also introducing a personalized update method to better tailor the aggregated models for each client. Building further on these concepts, the study titled Federated Unsupervised Person Re-identification via Camera-aware Clustering (FedUCA [161]) adapted another unsupervised Person Re-ID method, CAP [162], to federated learning. According to the authors, federated learning significantly enhances Person Re-ID performance across various datasets, particularly in smaller-scale datasets. This indicates that the datasets from federated clients can boost the performance of purely unsupervised methods, thanks to the diverse data contributions. In chapter 5, we will explore how can a source domain further contribute to overall performance and how the knowledge from the source can be leveraged across all clients in a federated setting.

Chapter 3

Online Unsupervised Domain Adaptation for Person Re-identification

3.1 Introduction

In this chapter, we present our first contribution to the field of Person Re-ID. We had previously identified data storage as one of the critical constraints that impact the deployment of Re-ID models in real-world scenarios, due to the sensitive nature of personal images related to privacy concerns. In the subsequent sections, we will focus on data storage as the primary limitation of Re-ID models. We will thoroughly explore the main constraints related to data storage and develop a novel setting that simulates the real-world scenarios where these constraints are encountered.

In Chapter 2, we also detail the state-of-the-art methods in Person Re-ID and highlight UDA techniques as promising approaches to address the distribution shifts between training sets and images captured in test environments. Building on this insight, we propose to investigate the performance of UDA methods when adapted to environments with restricted data access. This benchmark aims to understand their effectiveness and adaptability under such constraints.

Despite their relative efficiency, UDA methods are all based on the assumption that we have access to a large set of samples from the target domain environment during the training to perform adaptation in an *offline* fashion. In this chapter, we argue that this assumption is violated in many real-world scenarios. First, when deploying a Person Re-ID system, we gather images as long as they are recorded in the form of a stream that continuously provides images from different cameras/places. The *offline* process implies that the Re-ID system requires a possibly long data collection phase before deployment. Second, since the Re-ID task evolves person identities, the system is confronted with confidentiality purposes in many countries, forcing the technology provider of such models to discard previously seen images. Therefore, we argue that to match practical scenarios, an unsupervised domain adaptation method for Person Re-ID should respect two main constraints: 1) *Online adaptation:* the target domain data is not accessible all at once, but in a stream fashion where only small

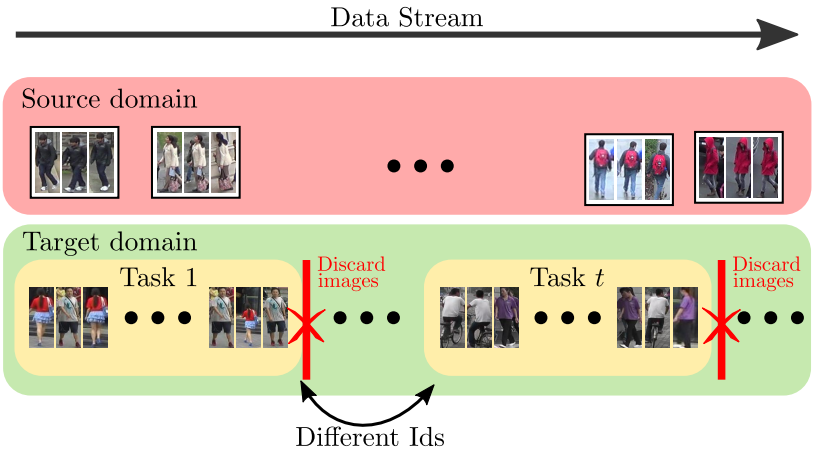


Figure 3.1: Illustration of the proposed OUDA for Re-ID setting: in Online Unsupervised Domain Adaptation for Person Re-ID, the annotated source dataset is available at any time while the target dataset is divided into multiple tasks. In between each task, the data from the previous task is discarded.

batches of images are available at a given instant of time and 2) *Privacy protection*: Images captured by the different cameras can be used to update the Re-ID model and stored for only a limited period of time. To this end, in this chapter, we propose and study a practical scenario for Unsupervised Person Re-ID which is the Online Unsupervised Domain Adaptation setting for Person Re-ID (OUDA-Rid). Fig. 3.1 gives an illustration of the proposed online setting, where we assume that the model has full access to the well-annotated source data set, however, unlike all the previous methods, the target domain dataset is fed to the Re-ID model in an online fashion. Practically, the target domain will be divided into several unlabeled subsets of images, where each subset will be viewed by the Re-ID model only once, hence respecting the two constraints: *online adaptation* and *privacy protection*. Regarding evaluation, we consider an independent and fixed target dataset with identities that do not overlap with any of the training tasks.

The contributions of this chapter can be summarized as twofold:

- We propose a new challenging yet practical scenario, the Online Unsupervised Domain Adaptation (OUDA-Rid) setting for Person Re-ID to respect two main constraints: Online Adaptation and Privacy Protection of identities.
- We adapt and evaluate three existing frameworks for *offline* UDA to the proposed OUDA-Rid setting: the Strong Baseline [66], MMT [68] and SpCL [70]. These methods are evaluated in four different adaptation settings based on three public and widely-used datasets: Market 1501 [76], DukeMTMC-reID [77] and MSMT17 [78]. Our results provide some interesting experimental conclusions regarding the performance and limitations of existing approaches. We hope that this work will stimulate the community to address domain adaptive Re-ID in the OUDA-Rid setting.

3.2 Related Work

This section expands on the previous related work discussed in chapter 2 from two perspectives. First, we give more details about UDA-based Re-ID methods with a particular focus on Domain translation and Pseudo-labeling based methods. Secondly, we review the state-of-the-art methods that have recently been proposed for addressing the lifelong learning setting for Person Re-ID, while also identifying their major limitations.

Unsupervised domain adaptation (UDA) for person Re-identification has been recently gaining a lot of attention for its practical applications. UDA methods can transfer learned knowledge from an annotated source domain to an unlabeled target domain, thus reducing the cost and discarding the need to have a well-annotated data set. Most of the existing methods and approaches in this area can be divided into two main categories: domain translation-based and pseudo-label-based methods.

Domain translation-based methods employ style transfer methods to modify the source images to obtain images with the content of the source domain but the appearance of the target. In this way, they obtain images similar to the target with the corresponding label annotations from the source images. These generated images are then used to refine the network parameters. Recent works in this category investigate the integration of generative models [163] [164] [165], as an example, we have [78] which is based on CycleGAN [166] to bridge the domain gap by transferring persons from the source domain to the target. [167] also generates images while preserving the self-similarity of the images before and after the translation and the domain-dissimilarity of the translated source images to the target images. Finally, we can cite the work of Zhong *et al.* [168] where the proposed framework learns camera-invariant features while enforcing domain connectedness, where two images, one from the source domain and the other one from the target domain, are fed to the network as a negative pair of images.

Pseudo-labeling methods, also called clustering-based methods, employ an iterative process alternating between clustering and finetuning [169, 170, 49, 171, 172]. In its simplest implementation, [66], the cluster indexes obtained in the clustering stage are used as labels to fine-tune the Re-ID network. Despite its simplicity, this simple approach obtains satisfactory results but suffers from limitations that have been addressed in recent works. For instance, Yang Fu *et al.* proposed a Self-Similarity Grouping (SSG) [67] approach that assigns different pseudo-labels to both global and local features. To mitigate the effects of noisy hard pseudo-labels, Mutual-Mean Teaching (MMT) [68], proposed by Yaxiao *et al.*, adopts a teacher-student framework with two networks that are trained jointly using hard pseudo-labels generated by the two networks and soft pseudo labels generated by their Mean Networks, to conduct pseudo-label refinement in the target domain. Moreover, we can mention the work done by Ger *et al.* referred to as SpCL [70] that, unlike previous methods, takes advantage of both labeled source domain images' centroids and un-clustered target instances, stored in a hybrid memory, in addition to the target domain clusters. The memory gives more supervision to the feature extractor during training while minimizing the unified contrastive loss over the three kinds of information available in the hybrid memory. Importantly, pseudo-label-based methods achieve better results than translation approaches and maintain up until now the state-of-the-art performances on almost all public datasets [68, 70].

In addition, these approaches avoid the computation overhead of the transfer-based approach that requires image generation. Consequently, our experimental benchmark will focus only on pseudo-labeling approaches.

Even though all the aforementioned methods have shown promising results and great capability to adapt to a new target domain data set, their training process always assumes that they can have access to the entire target domain, which is difficult to hold in a real-world application as previously discussed in Sec. 3.1.

Lifelong Learning for person Re-Identification. Lifelong Learning, also called Continual Learning or Incremental Learning[173, 174, 175], is a field that aims at mitigating the catastrophic forgetting problem, which means that the model tends to forget previous knowledge acquired during previous tasks when learning new ones. Recently, many approaches have been developed to solve this problem for common vision tasks such as object detection [176], segmentation [177], or even image generation [178]. We can categorize existing methods into three main categories. First, teacher-student frameworks [129, 179], use a teacher module to remind the student network about the knowledge acquired in the past. The second category of methods relies on the regularization of the parameters update when new tasks arrive [180]. Finally, the third category is replay methods that consist of using stored images or an image generation model to feed old-task images along with the current task images into the learning network [181].

Recently, only a few works have tackled the problem of lifelong learning in the case of Person Re-ID. [182] propose an Adaptive Knowledge Accumulation (AKA) framework, however, the training process is fully supervised and treats only the domain-incremental scenario. Zhipeng Huang *et al.* [183] address a scenario similar to ours except that storing images from the previous task is permitted. In this thesis, we consider that in real-world applications, person images might be subject to confidentiality purposes, and therefore storing images from previous tasks is not permitted.

3.3 Online Setting for UDA for Person Re-ID (OUDA-Rid)

3.3.1 Problem Definition

In this section, we describe the proposed online unsupervised domain adaptation setting for person re-identification (OUDA-Rid). We consider that we have access to an annotated source domain data set $D_S = \{(\boldsymbol{x}_i^S, \boldsymbol{y}_i^S)|_{i=1}^{Ns}\}$, where $\boldsymbol{x}_i^S$ and $\boldsymbol{y}_i^S$ denote the i^{th} training sample and its associated person identity label. We consider that we also have access to a target domain data set D_T where ground truth identity labels are not available. However, differently from the standard UDA setting, we consider that the target data set is accessible as an online stream of data. More precisely, we adopt the batch-based relaxation [184] of the online learning scenario. The model will have access to the target domain D_T as a stream of T independent batches $T_t, t \in 1..T$. In analogy with the Continual Learning (CL) setting and to avoid confusion with the *mini-batch* used in Stochastic Gradient Descent (SGD), each target batch will be called a task. Each task T_t is composed of N_t images $\{\boldsymbol{x}_i^t, i = 1..N_t\}$

that depict an unknown number of identities. We assume that there is no identity overlap between tasks even if our approach does not strictly require it. This assumption corresponds to the practical scenario where data are collected over several hours or days. Even if the same person can appear again at different times, most detections will correspond to different identities.

Importantly, we consider that at the end of the task, the images of the task T_t cannot be used for the next tasks. This corresponds to a practical scenario where sensitive data can only be stored for a short period of privacy concerns (e.g. camera images from a public area). In addition to the source domain that is accessible at any time, only the parameters of the networks can be kept in memory between two tasks. Finally, the goal is to deploy the trained model on an unknown target dataset that follows the same distribution as the training target tasks but does not share identities with the training tasks.

In this work, we adapt three frameworks for UDA to our OUDA-Rid setting. As detailed in Sec. 3.2, the UDA methods based on pseudo-labeling dominate most Person Re-ID benchmarks. Therefore, we focus our work on this paradigm. First, we employ a *Strong baseline* that is a very simple, yet effective, baseline. Then, we consider MMT [68] and SpCL [70], which are two methods that achieve state-of-the-art performance on publicly available datasets. Apart from their performance, what motivates the choice of these two frameworks is that, on the one hand, MMT has attracted a lot of attention lately and it is now considered a reference baseline for the task of UDA for person re-identification. On the other hand, SpCL is included in our benchmark since it illustrates the potential advantage of employing a memory to combine source and target data. Once adapted, the three frameworks will be evaluated and tested under four different configurations to: 1) decide which of the three frameworks is most suited to the OUDA-Rid problem 2) measure the drop in performance due to the online constraint 3) study the sensitivity of each model to its hyper-parameters.

3.3.2 Strong Baseline

The *Strong Baseline* [66] is a simple pseudo labeling pipeline. A feature extractor network F (backbone in Fig. 3.2) is pre-trained on the source labeled domain data set. After pre-training, the model is then fine-tuned on the target unlabeled data set. The fine-tuning on target consists of an iterative process where two major steps are alternated until convergence:

1. F is used to extract image features for every target domain image. Then, a standard clustering algorithm (DBSCAN [45] in our experiments) is applied to the extracted target domain features to obtain K clusters. In our case, K is automatically returned by the DBSCAN algorithm. In this way, we assign a cluster label to every image.
2. F is then fine-tuned on the target samples using their cluster labels as pseudo-labels. More precisely, a target domain classifier C with K classes is added to classify the images' features along with their assigned pseudo labels. The network is then trained via the minimization of a combination of an identity loss $\mathcal{L}_{id}^{T}(\theta)$ and a triplet loss $\mathcal{L}_{tri}^{T}(\theta)$. Assuming a sample $\boldsymbol{x}_i$ with pseudo-label $\boldsymbol{y}_i$, the identity loss is given by:

$$\mathcal{L}_{id} = \mathcal{L}_{ce}(C(F(\boldsymbol{x}_i)), \boldsymbol{y}_i), \tag{3.1}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss. Assuming the hardest positive and hardest negative features in the current mini-batch for the sample $\boldsymbol{x}_i$, denoted f_i^+ and f_i^- respectively, the triplet can be written:

$$\mathcal{L}_{tri}^T(\theta) = \max[0, ||F(\boldsymbol{x}_i) - f_i^+)|| + m - ||F(\boldsymbol{x}_i) - f_i^-)||] \tag{3.2}$$

where $||.||$ denotes the $\mathcal{L}^2$-norm and $m = 0.5$ denotes the triplet distance margin.

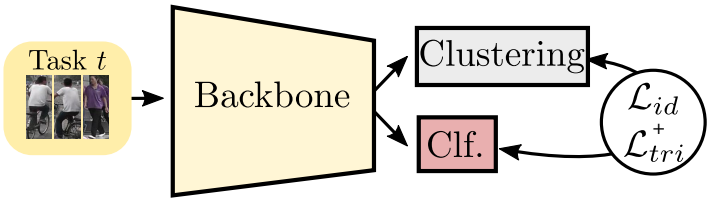


Figure 3.2: Scheme of *Strong Baseline*: training iterate between clustering and finetunning. The network is trained using a combination of cross-entropy and triplet losses.

Adaptation to OUDA-Rid.

In our setting, this baseline approach is applied to each task. Instead of using the whole target dataset in the clustering step, we use only the data of the current task. The two steps are applied iteratively for several epochs.

3.3.3 MMT

MMT is a recent framework proposed by [68], that integrates the teacher-student framework with two networks that train jointly. The main motivation is to design a framework that uses both hard and soft pseudo labels to learn better features. As shown in Fig. 3.3, MMT extends the *Strong Baseline* in several ways. First, MMT employs two networks F_1 and F_2 instead of a single feature extractor F. To enforce that the networks help each other, the classifier C_1 for the feature extractor F_1 is trained to predict the clustering labels obtained from F_2 and vice-versa. Second, mean teacher networks M_1 and M_2 are introduced. These networks are obtained by estimating the running average on the network parameters of F_1 and F_2. These networks predict more stable pseudo labels since they combine the knowledge of the networks at previous training iterations. In addition to the identity and triplet losses introduced in the *Strong baseline*, the two networks F_1 and F_2 are also optimized with respect to a soft classification loss and a soft triplet loss. Those losses are calculated for each network over the predictions of the other mean network. The losses between F_1 and M_2 are:

$$\mathcal{L}_{sid} = -M_2(\boldsymbol{x}_i).\log C_1(F_1(\boldsymbol{x}_i)) \tag{3.3}$$

$$\mathcal{L}_{stri} = -\mathcal{L}_{bce}(\tau_1^F(\boldsymbol{x}_i), \tau_2^M(\boldsymbol{x}_i)), \tag{3.4}$$

where $\mathcal{L}_{bce}$ denotes the binary cross entropy and τ_1^F and τ_2^M are given by:

$$\tau_1^F(\boldsymbol{x}_i) = \frac{e^{||F_1(\boldsymbol{x}_i) - F_1(\boldsymbol{x}_i^-)||}}{e^{||F_1(\boldsymbol{x}_i) - F_1(\boldsymbol{x}_i^+)||} + e^{||F_1(\boldsymbol{x}_i) - F_1(\boldsymbol{x}_i^-)||}} \tag{3.5}$$

$$\tau_2^M(\boldsymbol{x}_i) = \frac{e^{||M_2(\boldsymbol{x}_i) - M_2(\boldsymbol{x}_i^-)||}}{e^{||M_2(\boldsymbol{x}_i) - M_2(\boldsymbol{x}_i^+)||} + e^{||M_2(\boldsymbol{x}_i) - M_2(\boldsymbol{x}_i^-)||}} \tag{3.6}$$

Note that to encourage the two networks to learn different image representations, different random data transformation policies are used for each network pair (F_1, M_1) and (F_2, M_2).

Adaptation to OUDA-Rid.

We adapt MMT to the OUDA-Rid setting in the following way: at the end of each task, the parameters of the four networks are kept and reused for the next task.

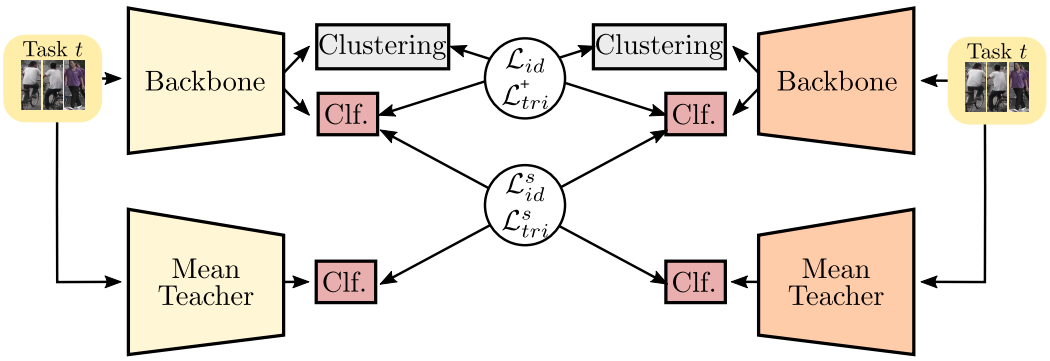


Figure 3.3: Scheme of *MMT*: two networks are trained thanks to two other momentum encoder networks. The two networks are trained using a combination of cross-entropy and triplet losses.

3.3.4 SpCL

Finally, we consider the SpCL method proposed in [70]. This framework (Fig. 3.4) employs a hybrid memory that stores and continually updates three types of feature vectors: the class centroids for every class of the source domain, cluster centroids for every cluster from the target domain, and the image feature corresponding to the target-domain samples that are not assigned to any cluster and that are considered outliers. This memory provides supervision to the feature extractor via a contrastive loss over the three types of features in the memory.

Adaptation to OUDA-Rid.

We consider two adaptations of the SpCL framework. In the first version, referred to *SpCL-SF*, we adopt a source-free SF strategy [185, 186] where we do not use any of the source data when adapting to the target domain. This version is introduced because it allows for a fair comparison with the MMT and the *Strong Baseline* that use the source dataset only for pre-training. In our second version (simply referred to as *SpCL*), we use the source dataset during the whole adaptation process in addition to the target data of the current task. In both cases, the memory is emptied, and clustering is performed at the beginning of each task.

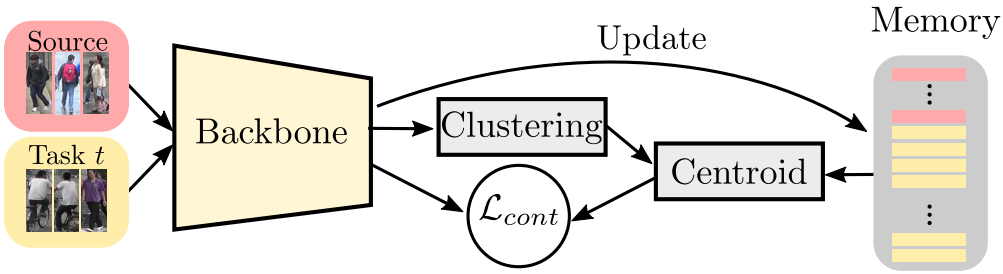


Figure 3.4: Scheme of *SpCL*: a feature memory is used to perform contrastive learning.

3.4 Experiments

3.4.1 Datasets

We evaluate the different frameworks on three widely-used real-world person benchmark datasets in the Domain Adaptive Person Re-ID:

- Market 1501 [76]: is a large-scale public dataset that contains 1501 identities that are captured by six different cameras. The total number of images is 32,668 for which 12,936 images of 751 identities are used for training and 19,732 images corresponding to the remaining 750 identities are used as a test set. We follow the official testing protocol stating that 3,368 query images should be tested and matched to 19,732 gallery images.
- DukeMTMC-reID [77]: The Duke Multi-Tracking Multi-Camera Re-Identification consists of images extracted from videos captured by 8 different cameras. It contains 16,522 training images corresponding to 702 identities, 2,228 query images of another 702 identities along 17,661 gallery images for testing.
- MSMT17 [78]: The third benchmark is the most challenging dataset since it has a greater diversity in terms of people's appearances, viewpoints, and scales. It consists of multiple hours of videos captured by 15 different cameras. This dataset is a large-scale dataset consisting of 32,621 images of 1,042 identities as a training set, and 11,659 query images along with 82,161 gallery images corresponding to 3,060 identities as a test set.

3.4.2 Evaluation Protocol

To evaluate the performance of the different frameworks on our proposed setting, we consider four source-target configurations: Duke→Market, Market→Duke, Market→MSMT17, and Duke→MSMT17. These configurations are widely used in the literature and illustrate domain shifts of diverse difficulty. For each configuration, we randomly and uniformly split the training identities into 5 subsets corresponding to 5 tasks. For the evaluation metrics, we adopt the metrics commonly used in Re-ID [68, 70]: Mean Average Precision (mAP) and Rank-1 [187] accuracies. These metrics are computed on the entire test set of the target domain after each task during the online adaptation process. The proposed testing protocol is chosen to have a global overview of the model's adaptation capability to the domain shift

between source and target, and also to see which framework is the most suited for online adaptation.

In our preliminary experiments, we observed that the number of epochs per task is a key hyperparameter. Even with a separate validation set, this hyper-parameter could not be chosen by mAP maximization since it requires identity labels and it would break the unsupervised adaptation assumption. On the contrary, using an inappropriate hyperparameter value would jeopardize the validity of the conclusions of our experiments. Therefore, we use the following procedure: we run the *strong baseline* with four different numbers of epochs ranging from 10 to 40. Then, we observed that training for 20 epochs per task leads to the best performance. Therefore, we use 20 epochs per task for all the methods. Note that we report an ablation study in Sec.3.4.6 to measure the sensitivity to this hyper-parameter and we validate that this choice remains satisfactory for the other methods.

3.4.3 Additional Baselines

To better assess the performance of the evaluated approaches, we consider two additional baselines that are not trained following our OUDA-Rid setting. First, we report the performance of the model pre-trained on the source and directly evaluated on the target. This baseline is common to all the frameworks since all the methods use the same pre-trained model and it is referred to as *Direct inference*. The second baseline is specific to each framework. It corresponds to the original method trained in the standard UDA setting and is referred to as *Offline*. It can be interpreted as an upper bound for the online methods.

3.4.4 Implementation details

We follow the common practices in the UDA Person Re-ID field by adopting ResNet50 [28] pre-trained on ImageNet [84] as a backbone. For clustering, we use DBSCAN [45] which is frequently used in the pseudo-label-based methods as it requires no prior on the number of clusters but only the maximum distance between two samples to consider one in the neighborhood of the other. We employ the maximum distance hyper-parameter set in the original papers of MMT and SpCL. Adam [188] optimizer is adopted with an initialized learning rate equal to $3.5 * 10^{-4}$ and a weight decay of 0.0005 [68, 70]. Finally, all the images were resized to 256 x 128 before being fed into the backbone (backbones for MMT), and the batch size was set to 64 corresponding to 16 different identities with 4 images per ID.

3.4.5 Results

We report in Figs. 3.5, 3.6, 3.7 and 3.8 the results of the *Strong Baseline* [66], MMT [68] and SpCL [70] on respectively four OUDA-Rid configurations: Duke→Market, Market→Duke, Market→MSMT17 and Duke→MSMT17. For every configuration, we report the final performances of each framework at the end of the adaptation process and plot the evolution of the test performance while the model is adapting to the target domain. Each experiment was

repeated 3 times with different batch sampling initializations (*i.e.*seeds). The colored area corresponds to the variance of the performance on the test set at the end of each task, where the points correspond to the mean performances of the different initializations.

First of all, in the four configurations, the results show that the pre-trained ResNet50 on the source domain gives poor performances when directly deploying it into the target domain without any finetuning (*Direct inference*) compared to when it is fine-tuned on the target domain, either in the *Offline* or *Online* setting. This big gap in terms of performance illustrates the problem of domain shift.

Then, when it comes to the 5-tasks Online setting, the conclusions differ between methods and datasets. In the case of the Duke→Market configuration (Fig. 3.5), we observe that MMT (orange line) performs best among the online methods and reaches 63.7% of mAP. This result is very satisfactory since MMT bridges most of the gap between *Direct inference* and *Offline*. The *strong baseline* obtains lower performance since it plateaus after the second task. However, it surprisingly outperforms the two SpCL variants. Indeed, the performance has not improved significantly after completing the first task. We even observe a small drop when processing the second task for the SpCL variant that uses the source domain images. We also notice that the difference between the two variants of SpCL is minor illustrating that with a straightforward adaptation of the SpCL method, initially proposed for offline UDA, SpCL does not benefit much from the availability of the source data. On the right-hand side of Fig. 3.5, we can see that the *Strong Baseline* shows the highest sensibility to random seeds. Moreover, MMT keeps reaching the best performance independently of the random seed.

	Offline		**Direct inference**		**Online**	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseline	75.6	90.9	29.6	62.4	49.4	77.1
MMT	80.9	92.9	29.6	62.4	63.7	87.5
SpCL-SF	76.7	90.3	29.6	62.4	42.9	70.2
SpCL	78.2	90.5	29.6	62.4	47.9	72.9

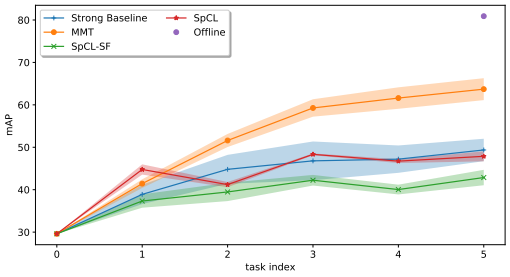


Figure 3.5: Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA-Rid Duke ->Market configuration. We report mAP and Rank-1 accuracy for each method.

Regarding the Market→Duke configuration (see Fig. 3.6), MMT is again the best performing method even though its gap with respect to the best offline method (purple dot) is larger. This behavior change can be explained by the highest difficulty of this setting as illustrated by the lower score obtained by the offline methods (*e.g.*70.4% of the map in Market →Duke vs 80.9% of the map in Duke →Market). In this more difficult configuration, the *strong baseline* does not perform well since it achieves the worst performance among all the evaluated methods. The behavior of SpCL is very instructive. At the beginning of training (until the second task), the source-free model performs better but shows degraded performance later in training. This behavior can be explained by a probable divergence of the model that forgets its initial source model and overfits the target task. On the contrary, the

SpCL variant that uses the source data needs more time to handle the domain shift but keeps slowly increasing. Concerning the variance of the performance, we can see that the four frameworks are sensitive to their random seed, especially at the beginning of the adaptation process. However, this variance decreases after a few tasks, showing that the training becomes more stable (after two tasks for most methods) except for the *Strong baseline*, where the variance of the performances becomes even higher on late tasks.

	Offline		**Direct inference**		**Online**	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseling	60.4	75.9	28.2	50.1	26.8	59.3
MMT	67.7	80.3	28.2	50.1	51.7	72.3
SpCL-SF	68.8	82.9	28.2	50.1	38.7	61.8
SpCL	70.4	83.8	28.2	50.1	42.7	66.0

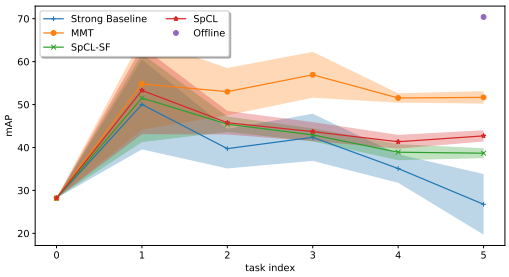


Figure 3.6: Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA-Rid Market ->Duke configuration. We report mAP and Rank-1 accuracy for each method.

	Offline		**Direct inference**		**Online**	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseling	9.7	25.8	8.9	28.9	6.1	18.0
MMT	22.9	49.2	8.9	28.9	15.1	31.5
SpCL-SF	26.3	53.4	8.9	28.9	13.1	36.5
SpCL	26.8	53.7	8.9	28.9	14.7	36.7

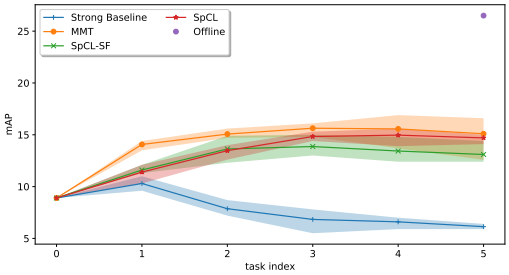


Figure 3.7: Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA-Rid Market ->MSMT configuration. We report mAP and Rank-1 accuracy for each method.

In the Market→MSMT configuration (Fig. 3.7), conclusions drastically change since SpCL has almost the same results as MMT, hence, breaking the big gap between the two methods in performance we observed in previous configurations. This change can be explained by the large training target dataset MSMT. Therefore, every target task contains more images and more identities. This difference is beneficial to both SpCL variants that perform similarly. Regarding, MMT, we see that the performance starts degrading from task 3. Again, it can be explained by the fact that in the case of a large target dataset, MMT can forget the knowledge from the source domain that is not further used during adaptation. Interestingly, the best performance of MMT (end of task 3) is higher than the best performance of SpCL. It illustrates the importance of handling the divergence problem and designing efficient consolidation mechanisms. Finally, we observe that the strong baseline worsens the performance compared to the initial pre-trained model. Regarding the variance of the performances, we can see that MMT, SpCL, and SpCL-SF finally get more or less similar results at the end of the adaptation process. Finally, in the Duke→MSMT configuration (Fig. 3.8), the conclusions remain similar to the previous setting. Nevertheless, we can mention that

	Offline		Direct inference		Online	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseline	10.9	28.6	11.1	35.2	7.2	19.9
MMT	23.3	50.1	11.1	35.2	17.0	35.0
SpCL-SF	26.3	52.6	11.1	35.2	17.1	43.1
SpCL	26.5	53.1	11.1	35.2	17.8	40.8

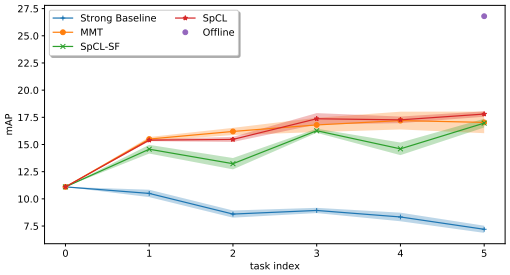


Figure 3.8: Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA-Rid Duke ->MSMT configuration. We report mAP and Rank-1 accuracy for each method.

SpCL outperforms MMT in this specific configuration, and observe higher instability on the SpCL-SF method that oscillates in the last tasks.

3.4.6 Analyses

Model sensitivity: number of training iterations.

In this section, we study the effect of the number of epochs on the performance of the four frameworks (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the following configuration: 5-task OUDA-Rid Duke→Market. In Fig. 3.9 we report the performance on the target test set (mAP) of the three frameworks while varying the number of training epochs between 0 and 40 epochs per task. Note that zero epoch corresponds to the *Direct inference* performance of the pre-trained model without any training on the target domain. It can be observed that with 20 epochs the *strong baseline* achieves the best performance on the test set. When we increase the number of the training epochs, we see a decrease in the performance on the test set of the three frameworks probably illustrating overfitting issues in the current training task. SpCL, thanks to its memory-based system, that provides supervision from the labeled source domain images to the Re-ID model, needs fewer training epochs per task to converge, compared to the *strong baseline* and MMT. We see that the four aforementioned frameworks are sensitive to the number of training epochs to some extent. These experiments illustrate the difficulty of the OUDA-Rid setting where only a few samples are available in each task and where overfitting can appear rapidly.

Impact of the number of tasks. We also conducted further experiments to show the effects of varying the number of tasks on the adaptation performances. In Fig. 3.10 we report the final performance (mAP) on the target test set of the four methods when considering 1, 3, 5, 8, and 10 tasks. Naturally, when augmenting the number of tasks during the adaptation process, the number of images per task decreases. This affects the fine-tuning of the model, where we can see in Fig. 3.10 that for all the considered frameworks, the performance drops when considering more challenging online settings by adding more tasks.

We also performed experiments with a number of tasks larger than 10 (typically 15 or 20), however, training did not succeed due to the sampling limitation. More precisely, when the number of tasks increases the number of samples becomes too small to be handled by

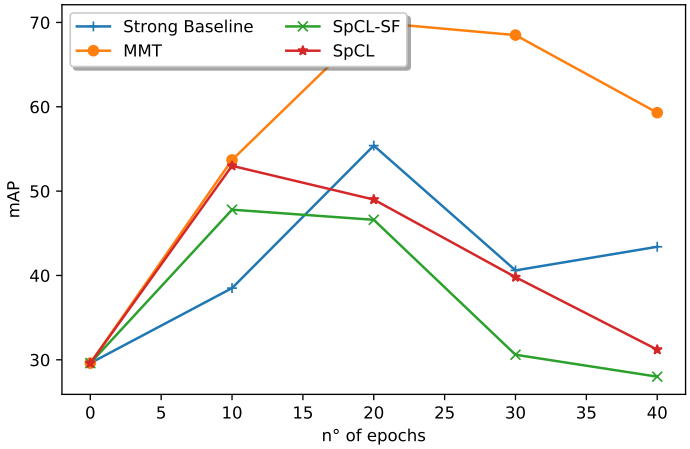


Figure 3.9: Effect of the number of training epochs per task on the Re-ID performance. At zero, we reported the results from the *direct inference* model.

DBSCAN. In such challenging configurations, only a few clusters are considered, where only a few images per cluster are sampled, hence the sampling of the 16 identities with 4 images per id, which is necessary for the optimization of the triplet loss, becomes impossible. This clustering issue shows the limitation of UDA methods to address our OUDA-Rid setting and demonstrates the need for new methods tailored for OUDA-Rid.

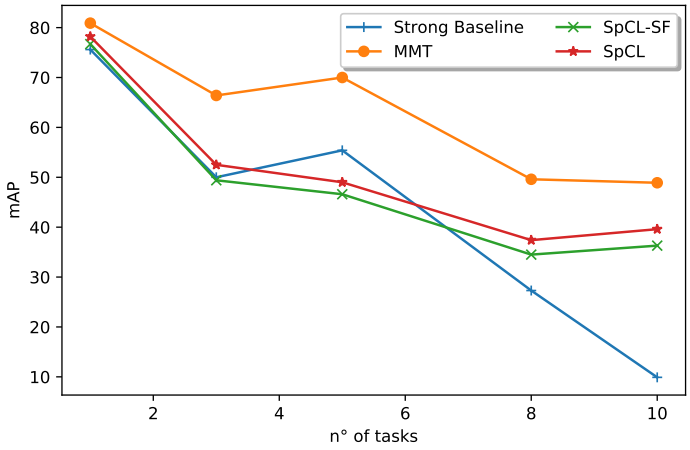


Figure 3.10: Effect of the number of tasks on the performances of the four frameworks at the end of the adaptation process. We varied the number of epochs from 1 to 10. Note that 1 epoch corresponds to the *offline* setting.

3.5 Conclusions

In this chapter, we introduced the Online Domain Adaptation Re-ID problem and presented an empirical benchmark where we adapt and evaluate three state-of-the-art methods previously introduced for the *Offline* UDA setting. Our experiments show that existing methods can achieve satisfactory results in simple online adaptation scenarios but fail to reach the performance achieved in the *Offline* setting. We also show that the best-performing methods depend on the setting. Finally, our experiments highlight the forgetting problem when the source model is not used during adaptation. These conclusions pave the way toward novel approaches for online domain adaptive Re-ID with the aim to inspire further research in this setting that matches real-world constraints and better protects privacy. Building on these findings, the next chapter will introduce a novel framework designed to effectively adapt UDA methods and maintain their performance, despite the challenges imposed by the OUDA-Rid setting.

Chapter 4

Source-Guided Similarity Preservation

4.1 Introduction

Since deploying algorithms that conform with policies of data privacy protection has become a legal obligation in a growing number of countries, the Online Unsupervised Domain Adaptation for person Re-Identification (*OUDA-Rid*) setting was introduced in the previous chapter (Chap 3) to address the limitations of traditional UDA techniques.

To recall, UDA methods combine a well-annotated dataset (*source domain*) and an unlabeled dataset corresponding to the *target domain*, aiming to train a model that can perform well in the new environment. Despite progress in recent years [68, 70], UDA for person Re-ID suffers from three main issues that prevent its practical use. First, when collecting the target data required to adapt the model, images are generally gathered as a stream that continually sends photos from various cameras/locations. Consequently, collecting a large target dataset may take time and delay deployment. In addition, in UDA, the model is frozen after deployment and does not benefit from the new data, which are continuously captured. Finally, numerous countries have adopted privacy regulations that forbid technology providers to store images of individuals. Thus, collecting a large target dataset is not possible.

In the OUDA-Rid framework, we operate under the assumption that we have access to annotated source data as well as unlabeled target data. However, in contrast to traditional UDA settings, the target dataset is treated as an online stream of data, aligning with the constraint that camera-captured images cannot be stored. In addition to complying with privacy-protection regulations, this setting also enables the person Re-ID model to be continuously updated as new target data becomes available, thereby improving its adaptability to changes in the target domain. Following this, we propose in this chapter an innovative approach to adapt UDA methods to the setting of OUDA-Rid, thereby adhering to privacy regulations related to the storage and retention of data.

In the previous chapter (Chap 3), we demonstrated that when existing UDA methods are adapted to the OUDA-Rid setting, there is a significant drop in performance compared to their deployment in the traditional offline setting. This drop can be explained by the two

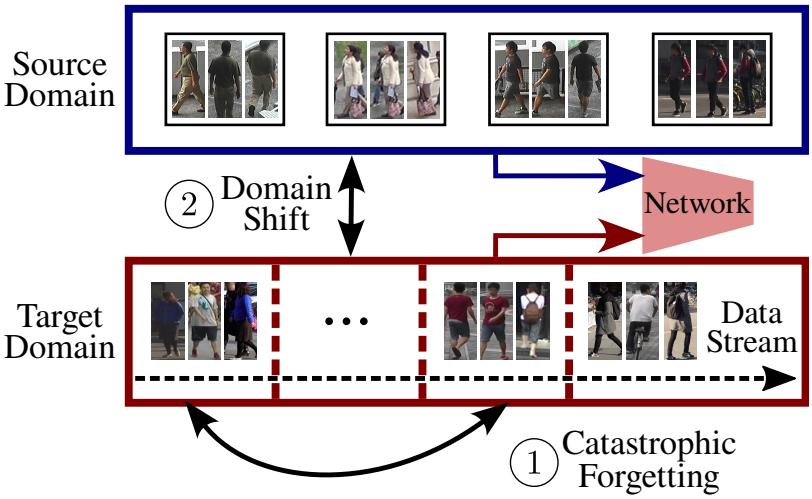


Figure 4.1: In OUDA for person Re-ID, the images of the target domain are available as a stream of data, and past images cannot be stored. Two main challenges should be addressed: 1) catastrophic forgetting and 2) domain shift.

main difficulties of OUDA-Rid illustrated in Fig. 4.1: catastrophic forgetting and domain shift. Catastrophic forgetting appears when the model only observes a few target identities, and consequently, the model tends to forget previously learned identities. Domain shift is a change in the data distribution between the source and target domains. Addressing the domain shift is especially challenging in the online setting since, at every training step, we observe only a small and possibly biased subset of the target domain.

In this chapter, we consider that these two difficulties must be addressed jointly since mitigating catastrophic forgetting can lead to target representations that better capture the full target distribution, and consequently facilitate source-target distribution alignment. We introduce a unified Source-guided Similarity Preservation (*S2P*) framework for OUDA-Rid that addresses these two challenges jointly. We take inspiration from *replay-based* strategies [189, 190] to introduce a Knowledge Distillation (*KD*) mechanism. By transferring the knowledge acquired with a teacher model to a student model, the KD [191] method enables the learning of more robust and generalizable features. However, unlike existing replay-based approaches, we do not store any target image to conform to the *privacy protection* requirement. To this end, we extract a support set composed of source images that are similar to the previously seen images of the target. This support is thus used to regularize the learning process and alleviate catastrophic forgetting. Our framework combines both explicit source-target distribution alignment and pseudo-labeling to address domain shift. S2P can easily integrate almost any existing UDA approaches [68, 70] and readily outperforms all state-of-the-art methods for OUDA-Rid in several challenging conditions in real-to-real and synthetic-to-real tasks. The main contributions of this chapter can be summarized as follows:

- We introduce a novel S2P algorithm that uses source-guided similarity preservation to jointly alleviate the *catastrophic forgetting* and *domain shift* while respecting the *privacy protection* requirements.
- S2P can easily incorporate almost any existing UDA approach. In particular, we present the integration of the *MMT* [68], *SpCL* [70] and *IDM* [115] methods into our framework,

which achieve remarkable results in the UDA setting.
- We perform extensive experiments[1] in real-to-real and synthetic-to-real OUDA tasks with four datasets. S2P readily improves previous state-of-the-art UDA methods for OUDA-Rid. A set of ablation studies validates each component of our algorithm.

4.2 Related Work

We extend the previous related work sections by including the recently published method IDM [115] as a pseudo-labeling-based approach for UDA. Additionally, we clarify how our support set selection method differentiates from previous approaches in the literature.

UDA for person Re-ID. Existing methods can be divided into *domain translation-based* [163, 164, 165] and *pseudo-labeling* [169, 170, 49, 171, 172]. Pseudo-labeling methods employ an iterative process alternating between clustering and fine-tuning [169, 170, 49, 171, 172]. In addition to the previously mentioned methods, such as the *strong baseline* [66], *MMT* [68], and *SpCL* [70], a recent work has introduced the use of an Intermediate Domain Module (IDM) [115] as means to bridge the gap between source and target domains. We adopt the pseudo-labeling framework as it has outperformed previous techniques in almost all datasets [68, 70] and avoids the computational overhead of transfer-based methods. Our S2P overall framework can incorporate existing pseudo-labeling methods toward better performance in the OUDA-Rid setting.

Lifelong learning for Re-ID. While previous methods in continual learning for person Re-ID, such as [182, 183], have adopted a less restrictive setting that allows keeping images from previous tasks, we follow the more challenging and privacy-preserving OUDA-Rid setting presented in Chapter 3. To address domain shift and catastrophic forgetting in OUDA-Rid, we introduce two key technical contributions: a source-guided knowledge distillation strategy and an explicit domain alignment. Gong *et al.* [192] introduced a technique based on landmarks that is similar to our support set selection. However, these landmarks were proposed to solve the domain gap in the context of UDA with classical machine learning techniques, while we have to also consider the catastrophic forgetting problem in OUDA-Rid using end-to-end deep learning models.

4.3 Source-Guided Similarity Preservation

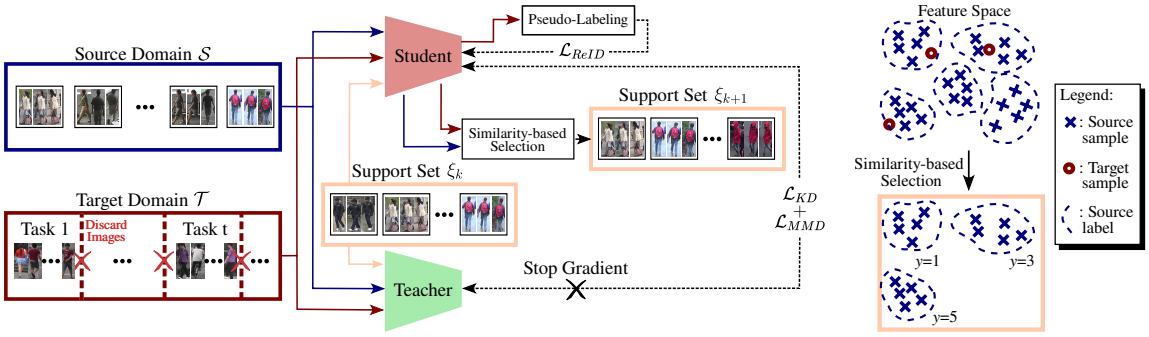
OUDA-Rid problem definition. In OUDA-Rid, we assume having access to a well-annotated source domain dataset $\mathcal{S} = \{(\boldsymbol{x}_i^S, \boldsymbol{y}_i^S)\}_{i=1}^{Ns}$, and an unlabeled target domain dataset $\mathcal{T} = \{\boldsymbol{x}_i^t\}_{i=1}^{Nt}$. Here, both domain images are not necessarily drawn from the same distribution. We consider that we have access to the target domain dataset in the form of an ongoing stream of data. Similarly to the previous chapter, we consider that we observe a sequence of N_T tasks $\{\mathcal{T}^1 \cup \mathcal{T}^2 \cup ... \cup \mathcal{T}^{N_T}\}$. Each task $\mathcal{T}^k, 1 \leq k \leq N_T$ is a set of images captured by several cameras and depicting an unknown number of identities. To align with practical

[1] Code available: https://github.com/ramiMMhamza/S2P

scenarios, we consider that each identity can be observed by different cameras simultaneously. However, it is unlikely for an identity to appear at widely separated time intervals (*e.g.* different days). Therefore we can assume that identities do not overlap across different tasks, although this assumption is not strictly required in our approach.

In the rest of this section, we present our S2P framework to alleviate the two major challenges of the OUDA setting: catastrophic forgetting and domain shift. First, our framework integrates a teacher model that distills previously acquired knowledge. The KD strategy of S2P is based on feature space similarity preservation and only requires images from the source domain, hence respecting the *privacy protection* norms. Second, we minimize the discrepancy between the source domain and the target domain to reduce the domain shift and further enhance the stability of the S2P.

4.3.1 Overview of the Approach



(a) Overall Pipeline of S2P.

(b) Construction of the support set via similarly-based selection.

Figure 4.2: The pipeline of S2P. a) S2P incorporates knowledge distillation $\mathcal{L}_{KD}$, discrepancy $\mathcal{L}_{MMD}$ loss functions, and a teacher model to mitigate the catastrophic forgetting and domain-shift problems. b) Our algorithm employs a similarity-based selection to construct the support set ξ_k from the source domain that maximizes the similarity with the target images.

Fig. 4.2 shows the pipeline of our S2P framework. In every task of the OUDA-Rid problem, the target labels are not available and we assume that the identities are different even if our S2P does not strictly require this assumption. Furthermore, we construct a *support set* that plays the role of a memory bank for *source-guided knowledge distillation*. We could keep a few samples from previous tasks if there were no privacy constraints. However, in S2P the support set only includes images from the source domain. We choose those images based on their similarities to previously seen images, ensuring a good approximation of the previously learned feature spaces during continual learning.

In this chapter, we follow an overall training scheme that was adopted by multiple UDA methods for Re-ID [68, 70]. More concretely, we use a student model that consists of a

feature extractor $\mathcal{F}(\cdot)$. First, the student model is pre-trained on source data $\mathcal{S}$, and then fine-tuned on the unlabeled target data $\mathcal{T}$ with three different loss functions:

- $\mathcal{L}_{KD}$: the knowledge distillation loss in the feature similarity space is proposed to preserve the previously acquired knowledge. To this end, a *similarity-based selection* strategy is applied to the source domain to construct the support set, and a teacher model $\bar{\mathcal{F}}(\cdot)$ is added to the main pipeline (Sec. 4.3.2).
- $\mathcal{L}_{MMD}$: the Maximum Mean Discrepancy (*MMD*) loss is minimized to reduce explicitly the domain shift. In other words, we want to construct a feature space that is domain invariant and can regroup features from both the source and the target domains (Sec. 4.3.3).
- $\mathcal{L}_{ReID}$: this loss corresponds to the loss of the UDA method that is integrated into our framework. This loss is jointly minimized on the source domain $\mathcal{S}$ and the target domain images $\mathcal{T}$ together with their pseudo-labels. The pseudo labels are estimated by a clustering algorithm assigning each image to a cluster label (Sec. 4.3.4).

4.3.2 Source-Guided Knowledge Distillation

When learning a new task $\mathcal{T}^k$, the model must be updated to better discriminate the appearance of the new individuals. However, the model should also preserve the knowledge acquired on previous tasks $\mathcal{T}^i \; \forall \; 1 \leq i \leq k-1$. Therefore, we employ a teacher model that progressively distills the knowledge to the student model. Distillation is performed in the feature space over a set of source-based support images. Since target images cannot be stored, we propose to use images from the source domain as the support set. More precisely, we select images that are similar to the images from the target domain seen in previous tasks. This solution encourages the student model to project the images into a common feature space, resulting in more discriminant and task-invariant representations.

Support set collection. Fig. 4.2 (b) depicts the construction of the support set in S2P. We construct the support set based on the cosine similarity in the feature space between the current target images and the source domain images. For each image $\boldsymbol{x}^t$ in the target task $\mathcal{T}^k$, we identify the image $\xi_x(\boldsymbol{x}^t)$ and its corresponding identity label $\xi_y(\boldsymbol{x}^t)$ from the source domain that maximizes the cosine similarity in the feature space:

$$\big(\xi_x(\boldsymbol{x}^t), \xi_y(\boldsymbol{x}^t)\big) = \underset{(\boldsymbol{x}^s, \boldsymbol{y}) \in \mathcal{S}}{\operatorname{argmax}} \frac{\mathcal{F}(\boldsymbol{x}^s) \cdot \mathcal{F}(\boldsymbol{x}^t)}{\|\mathcal{F}(\boldsymbol{x}^s)\| \|\mathcal{F}(\boldsymbol{x}^t)\|}. \tag{4.1}$$

Then, we add to the support set all the images from the source that correspond to the selected identity $\xi_y(\boldsymbol{x}_t)$:

$$\xi_k = \bigcup_{\boldsymbol{x}^t \in \mathcal{T}^k} \{(\boldsymbol{x}^s, \boldsymbol{y}) \in \mathcal{S}, \boldsymbol{y} = \xi_y(\boldsymbol{x}^t)\}. \tag{4.2}$$

While learning a new task $\mathcal{T}^{k+1}$, ξ_k is used as a memory that best approximates previously seen images.

Teacher-student framework. As a teacher, we need a model that has accumulated knowledge from previous tasks and can effectively guide the student's learning on a new task. We use the Exponential Moving Average (*EMA*) parameters update [193, 194] of the current

model. At every iteration i, the parameters $\bar{\boldsymbol{\theta}}_i$ of the teacher model are given by:

$$\bar{\boldsymbol{\theta}}_i = \alpha \bar{\boldsymbol{\theta}}_{i-1} + (1-\alpha)\boldsymbol{\theta}, \tag{4.3}$$

where $\boldsymbol{\theta}$ denotes the current parameters of the student model and $\alpha \in [0, 1)$ is the weighting factor. At the first iteration of our framework, $\bar{\boldsymbol{\theta}}_0$ is initialized using a model pre-trained on the source dataset. Once the adaptation process is performed on a specific task, only the teacher is used for inference on the test set.

KD loss. Knowledge distillation commonly uses softened softmax labels from the teacher in training the student network [191, 129]. However, we argue that this formulation is not suitable for Re-ID. In classification problems, the absolute position of the samples in the feature space must be preserved to remain compatible with the learned classifiers. On the contrary, in Re-ID, we are interested only in preserving the relative distance between samples. Therefore, we employ a distillation loss that acts on similarity matrices to offer the model more freedom to adjust the position of the features in the learned space.

Assuming an input tensor $\mathbf{X}$ corresponding to a mini-batch of n images from the support set $\{\boldsymbol{x}_i\}_{i=1}^n$, we use the student network $\mathcal{F}$ to compute the feature representations $\mathbf{F} = \mathcal{F}(\mathbf{X}) \in \mathbb{R}^{n \times c}$, where c is the dimension of the feature space. Similarly, we compute the features with the teacher network $\bar{\mathbf{F}} = \bar{\mathcal{F}}(\mathbf{X}) \in \mathbb{R}^{n \times c}$. Then, we calculate the similarity matrices $\mathbf{S} \in \mathbb{R}^{n \times n}$ and $\bar{\mathbf{S}} \in \mathbb{R}^{n \times n}$ containing the pairwise scalar product between the current features of all images in the current batch of the support set:

$$\mathbf{S} = \mathbf{F}\mathbf{F}^\mathsf{T}, \text{ and } \bar{\mathbf{S}} = \bar{\mathbf{F}}\bar{\mathbf{F}}^\mathsf{T}. \tag{4.4}$$

Moreover, we minimize the Frobenius norm $\|\cdot\|_F$ between the similarity matrices of the teacher and the student. The source-guided knowledge distillation loss can thus be formulated as follows:

$$\mathcal{L}_{KD}(\bar{\mathbf{S}}, \mathbf{S}) = \left\| \frac{\bar{\mathbf{S}}}{\|\bar{\mathbf{S}}\|} - \frac{\mathbf{S}}{\|\mathbf{S}\|} \right\|_F^2. \tag{4.5}$$

4.3.3 Source-Target Distribution Alignment

To achieve successful knowledge distillation over the support set, it is crucial to ensure that the selected images from the source domain are visually similar to the previously seen target images. To this end, we introduce an additional training loss that explicitly aligns the source and the target feature distribution. We use the Maximum Mean Discrepancy (MMD) loss [195] to reduce the domain shift by minimizing the discrepancy between the source and target domains. Formally, given an input batch of images $\{\boldsymbol{x}_i^s\}_{i=1}^n, \{\boldsymbol{x}_j^t\}_{j=1}^n$ coming from both $\mathcal{S}$ and $\mathcal{T}^k$, we compute the feature representations from both the teacher and the student models: $\bar{\mathbf{B}} = (\bar{\mathbf{b}}_i)_{i=1}^n, \mathbf{B} = (\mathbf{b}_j)_{j=1}^n \in \mathbb{R}^{n \times c}$, where:

$$\bar{\mathbf{b}}_i = \bar{\mathcal{F}}(\boldsymbol{x}_i^s), \text{and } \mathbf{b}_j = \mathcal{F}(\boldsymbol{x}_j^t). \tag{4.6}$$

As shown in [195], assuming a positive semi-definite kernel K, the MMD loss can be empirically estimated as follows:

$$\mathcal{L}_{MMD}(\bar{\mathbf{B}}, \mathbf{B}) = \frac{1}{n^2} \sum_{i,j=1}^{n} [K(\bar{\mathbf{b}}_i, \bar{\mathbf{b}}_j) + K(\mathbf{b}_i, \mathbf{b}_j) - 2\mathrm{K}(\bar{\mathbf{b}}_i, \mathbf{b}_j)]. \tag{4.7}$$

We follow the common practice and employ the Gaussian kernel [196] with bandwidth parameter σ:

$$K(\bar{\mathbf{b}}_i, \mathbf{b}_j) = \exp\left(-\frac{\left\|\bar{\mathbf{b}}_i - \mathbf{b}_j\right\|^2}{2\sigma^2}\right), \tag{4.8}$$

where we set the bandwidth σ to the estimated variance of each minibatch as in [196].

4.3.4 Incorporating Pseudo-Labeling into S2P.

We now detail how we integrate three state-of-the-art pseudo-labeling-based frameworks into S2P: MMT [68], SpCL [70] and IDM [115].

MMT employs two networks $\mathcal{F}_1$ and $\mathcal{F}_2$ instead of a single feature extractor $\mathcal{F}$ as discussed above. The classifier C_1 for the feature extractor $\mathcal{F}_1$ is trained to predict the clustering labels obtained from $\mathcal{F}_2$ and vice-versa. Mean teacher networks $\bar{\mathcal{F}}_1$ and $\bar{\mathcal{F}}_2$ are introduced. In addition to the cross-entropy loss $\mathcal{L}_{ce}$, and the triplet loss $\mathcal{L}_{tri}$ introduced in the *strong baseline* [66], the two networks $\mathcal{F}_1$ and $\mathcal{F}_2$ are also optimized using a soft classification loss $\mathcal{L}_{sce}$ and a soft triplet loss $\mathcal{L}_{stri}$ with their mean networks [129]. Finally, $\mathcal{L}_{ReID}$ is a weighted sum of the four aforementioned losses. To integrate MMT into S2P, the two similarity matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are estimated using respectively $\mathcal{F}_1$ and $\mathcal{F}_2$ as student networks from a support set mini-batch. Similarly, two teacher similarity matrices $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are estimated from the two mean teachers. The total knowledge-distillation loss is defined as the sum of $\mathcal{L}_{KD}(\bar{\mathbf{S}}_1, \mathbf{S}_1)$ and $\mathcal{L}_{KD}(\bar{\mathbf{S}}_2, \mathbf{S}_2)$. In the same way, $\mathcal{L}_{MMD}$ is jointly optimized on the source and the target domains in the feature spaces of both student-teacher couples $(\mathcal{F}_1, \bar{\mathcal{F}}_1)$ and $(\mathcal{F}_2, \bar{\mathcal{F}}_2)$.

SpCL adopts a contrastive training scheme in the feature space over a hybrid memory that is continually updated by the estimated pseudo-labels. The hybrid memory stores three types of feature representations: 1) the centroids for every class of the source domain, 2) the centroids for every cluster from the target domain, and 3) the feature representations of the outliers. Finally, $\mathcal{L}_{ReID}$ is a contrastive loss that jointly distinguishes classes, clusters, and unclustered instances in the feature space of the hybrid memory. For more details, the readers are referred to [70]. The integration of SpCL into our S2P is straightforward. We first add the teacher model, which is the EMA of the fine-tuned model. Then, for each new task, the support set is constructed to add $\mathcal{L}_{KD}$ and $\mathcal{L}_{MMD}$ to the S2P pipeline.

IDM is based on a module designed to generate intermediate domain representations by mixing the hidden representations of the source and target domains. Network training is regularized with additional losses, which promote diversity among the domain variables and ensure that the intermediate domain lies between the source and target domains. To integrate

IDM into our S2P framework, we first add a teacher model which is obtained through EMA over the model's weights, including the IDM module. Then, during the optimization, we sum the two losses of S2P, $\mathcal{L}_{KD}$ and $\mathcal{L}_{MMD}$, to the IDM losses.

4.4 Experiments and Results

This section introduces the datasets used in the current work, the evaluation protocol, the implementation details, as well as the results and discussions of S2P. We compare our algorithm against four state-of-the-art approaches for UDA for person Re-ID: the *strong baseline* [66], MMT [68], SpCL [70], and IDM [115]. Finally, we perform a set of ablation studies to analyze each component of S2P, including the construction of the support set, the choice of the teacher, and the loss functions. In particular, we compare our KD loss $\mathcal{L}_{KD}$ with alternatives [197, 198] previously introduced in the literature for similar tasks.

Method	MS → M		MS → C		M → MS		RP → M	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseline [66]	$51.4_{\pm 1.8}$	$72.3_{\pm 0.5}$	$5.3_{\pm 1.2}$	$4.3_{\pm 1.9}$	$6.1_{\pm 0.1}$	$18.1_{\pm 0.3}$	$43.1_{\pm 1.3}$	$67.6_{\pm 1.6}$
MMT [68]	$65.8_{\pm 0.1}$	$83.7_{\pm 0.1}$	$32.2_{\pm 1.6}$	$32.2_{\pm 2.4}$	$15.1_{\pm 1.9}$	$36.9_{\pm 0.1}$	$58.7_{\pm 0.7}$	$77.5_{\pm 0.1}$
SpCL [70]	$53.5_{\pm 0.4}$	$76.0_{\pm 0.3}$	$15.6_{\pm 3.1}$	$15.7_{\pm 1.7}$	$14.7_{\pm 0.2}$	$36.7_{\pm 2.3}$	$50.5_{\pm 2.8}$	$72.1_{\pm 3.5}$
IDM [115]	$57.5_{\pm 0.2}$	$78.6_{\pm 0.2}$	$8.3_{\pm 0.2}$	$7_{\pm 0.3}$	$7.9_{\pm 0.5}$	$21.5_{\pm 0.1}$	$60.8_{\pm 0.2}$	$80.4_{\pm 0.1}$
S2P-MMT (ours)	$\underline{70}_{\pm 0.4}$	$\underline{87.1}_{\pm 0.4}$	$\mathbf{40.4}_{\pm 0.8}$	$\mathbf{42.4}_{\pm 0.9}$	$\underline{19.5}_{\pm 0.1}$	$\underline{43.3}_{\pm 0.7}$	$\underline{61.4}_{\pm 0.1}$	$\underline{81}_{\pm 0.2}$
S2P-SpCL (ours)	$69.1_{\pm 0.1}$	$87.1_{\pm 0.1}$	$\underline{34.3}_{\pm 0.3}$	$\underline{35.1}_{\pm 0.5}$	$\mathbf{20.2}_{\pm 0.1}$	$\mathbf{46.1}_{\pm 0.2}$	$59_{\pm 0.1}$	$80.5_{\pm 0.2}$
S2P-IDM (ours)	$\mathbf{71.3}_{\pm 0.1}$	$\mathbf{88.0}_{\pm 0.1}$	$17.5_{\pm 0.5}$	$16.6_{\pm 0.5}$	$14.2_{\pm 0.3}$	$33.9_{\pm 0.2}$	$\mathbf{70.2}_{\pm 0.2}$	$\mathbf{86.1}_{\pm 0.4}$

Table 4.1: Performance of S2P and four state-of-the-art methods in the last task in three real-to-real and one synthetic-to-real OUDA-Rid tasks. The best and second-best methods on each dataset are highlighted in **bold** and underlined, respectively.

Datasets. We evaluate S2P on four widely used person Re-ID datasets in domain adaptation:

- *CUHK03* (C) [19] comprises $14,097$ photos of $1,467$ individual identities captured by six cameras, with each identity recorded by two cameras. The dataset includes manually annotated and automatically generated bounding boxes. We utilize manually annotated bounding boxes for training and testing. The dataset provides a random train/test split in which 100 identities are selected for testing and the rest for training.
- *RandPerson* (RP) [83] is a synthetic dataset containing $8,000$ identities and $1,801,816$ images. We use a subset of $132,145$ images from the original $8,000$ identities.
- *Market 1501* (M) [76] and *MSMT17* (MS) [78] that were both introduced and used in the previous chapter.

Evaluation protocol. We follow the experimental protocol introduced in Chapter 3. We evaluate the performance of all methods using the standard training/testing splits proposed by the original authors for *Market 1501* and *MSMT17*. In *CUHK03*, we use a more challenging testing protocol proposed in [199], which consists of splitting the dataset into 767 and 700 identities for training and testing, respectively. RP is always used as a source dataset in this chapter.

We evaluate S2P for OUDA-Rid in several real-to-real and synthetic-to-real configurations: MS→M, MS→C, M→MS, and RP→M. These configurations are widely used in the literature [68, 70, 83] and illustrate domain shifts of diverse difficulties. For each configuration, we randomly and uniformly split the training identities into five subsets, corresponding to five tasks for OUDA-Rid, each having a distinct set of identities. We also perform additional experiments where we increase the number of tasks in the target domain, which are detailed in the supplementary material A.

We adopt the commonly used metrics for evaluation in Re-ID [68, 70]: mean Average Precision (mAP) and CMC Rank-1 [76] accuracies. These metrics are computed on the entire test set of the target domain after each task during the online adaptation process. We report the average mAP and Rank-1 over three repetitions with different seeds.

Implementation details. We follow the common practices in the UDA person Re-ID field by adopting ResNet50 [28] pre-trained on ImageNet [84] as a backbone. We employ the features computed after the global average pooling layer. We use DBSCAN for clustering, which is commonly employed in pseudo-labeling methods because it requires no prior assumption on the number of clusters. For each new task, Adam [188] optimizer is adopted with an initial learning rate (LR) equal to $3.5e-4$, a linear LR scheduler, and weight decay of $5e-4$ [68, 70]. Same as the previous chapter (Chap 3), the number of epochs per task is set to 20. For the EMA, we follow [68] and set α to 0.999 to update the teacher model parameters. Finally, all the images are resized to 256×128 before being fed into the backbone (or backbones for MMT), and the batch size was set to 64 corresponding to 16 different identities with 4 images per ID.

4.4.1 Quantitative Results

Comparison with the state of the art. Table 4.1 reports the mAP accuracy and CMC Rank-1 score obtained at the end of training with all methods in three *real-to-real* configurations: MS→M, MS→C, M→MS, and one *synthetic-to-real* configuration RP→M. The reported metrics are computed at the end of the adaptation process in each case. The low scores of the *strong baseline* are due to the presence of the domain shift, which cannot be appropriately addressed with this method. The state-of-the-art UDA methods MMT, SpCL and IDM struggle when deployed in the OUDA-Rid setting. The drop in performances of MMT, SpCL, and IDM is partially explained by the presence of catastrophic forgetting. Furthermore, MMT outperforms SpCL and IDM in almost all configurations, showing that their student-teacher framework is well suited to OUDA-Rid.

Table 4.1 shows that S2P-MMT, S2P-SpCL, or S2P-IDM outperforms all previous state-of-the-art UDA methods in OUDA-Rid over all configurations. For example, S2P improves the mAP of SpCL from 15.6 to 34.3 and from 14.7 to 20.2 in MS→C and M→MS, respectively.

As for IDM, our S2P significantly improves its performances, from 8.3 to 17.5 and from 7.9 to 14.2, in the same configurations: MS→C and M→MS. Finally, for MMT, S2P improves the mAP, from 32.2 to 40.4 and from 15.1 to 19.5, in MS→C and M→MS, respectively. The gain for SpCL and IDM is greater than for MMT because MMT already integrates

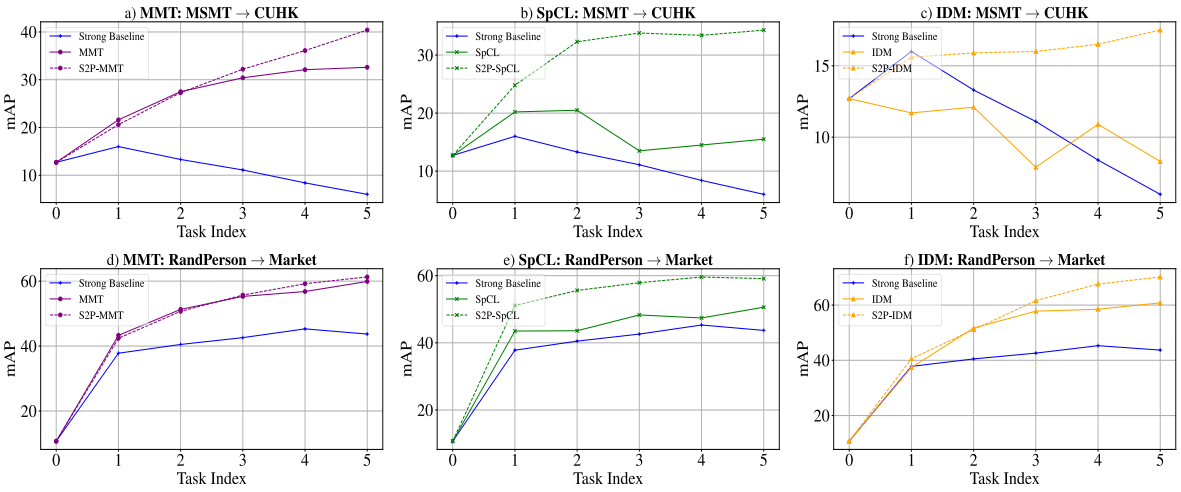


Figure 4.3: Comparison of S2P with four state-of-the-art methods in terms of mAP vs. task index in two different OUDA-Rid tasks, MSMT→CUHK and RandPerson→Market.

a teacher in its knowledge distillation loss function (soft cross entropy and soft triplet loss), whereas SpCL and IDM are only optimized using hard pseudo labels without any refinement.

Similarly, we can see that in the *synthetic-to-real* scenario RP→M, S2P noticeably improves the performance of the three state-of-the-art methods. S2P improves: from 58.7 to 61.4, from 50.5 to 59, and from 60.8 to 70.2 the performances of MMT, SpCL, and IDM, respectively. These results demonstrate that S2P can be successfully deployed in OUDA-Rid applications where we cannot have access to a real and well-annotated dataset for the source domain[2].

Continual behavior. To explore the analysis of the continual behavior of the different methods, we compare in Fig. 4.3 the mAP at the end of each task before and after incorporating the three state-of-the-art methods MMT, SpCL, and IDM into our S2P framework. For this analysis, we choose two different configurations: MS→C (Fig. 4.3-a, -b, and -c) and RP→M (Fig. 4.3-d, -e and -f). In general, the low performances of the direct inference (*i.e.* the mAP at task 0) and the *strong baseline* show that the chosen configurations are of varying degrees of difficulty.

Fig. 4.3 also shows the effect of catastrophic forgetting as a drop in performance in new tasks in several situations. For example, the *strong baseline* presents degradation of performance in both configurations in new tasks. Similarly, SpCL and IDM both lose accuracy when confronted with new incoming data due to catastrophic forgetting and domain shifts in the later tasks. For MS→C configuration: in b) the mAP of SpCL goes from 20.5 in the second task to 13.5 in the third task, while in c) the performance of IDM drops from 10.9 to 8.3 in the last task. Finally, for MMT we can notice in a) that the performance reaches an undesirable plateau after the third task in the same configuration. This shows that the knowledge acquired during the first stages of OUDA-Rid is lost during the adaptation process. Furthermore, the fluctuations of the mAP of SpCL and IDM in b), c), e), and f) in Fig. 4.3 illustrate the inability of the models to maintain a general structure of the feature space

[2]Additional experiments in different configurations can be found in the supplementary material A.

that captures the whole target domain distribution.

On the contrary, S2P-MMT, S2P-SpCL and S2P-IDM show a steady improvement in performance on the two configurations. Specifically, all three methods achieve better performance when learning later tasks when incorporated into our S2P framework and deliver consistent results across the different configurations.

Moreover, it is clear from the learning curves across all the different tasks that S2P successfully adapts UDA methods to the continual setting OUDA-Rid, resulting in a superior learning process evolution and a solid accumulation of prior knowledge.

4.4.2 Ablation Studies

We perform three ablation studies about: 1) the loss functions, 2) the knowledge distillation design, and 3) the choice of the teacher model. We run those experiments with S2P-SpCL as the pseudo-labeling method in OUDA-Rid configurations, namely, MS→C and RP→M.

The impact of the two main losses of S2P. The two main loss functions (KD and MMD) of S2P were introduced in Sec. 4.3.2 and 4.3.3. In this ablation, we study the influence of different configurations of the losses $\mathcal{L}_{MMD}$ and $\mathcal{L}_{KD}$ in the performance of S2P as shown in Table 4.2. The performance of the baseline significantly improves in almost all the configurations by only integrating either the $\mathcal{L}_{MMD}$ or $\mathcal{L}_{KD}$. For example, the configuration MS→ M shows a gain in performance. The mAP goes from 53.5 to 62.4 with $\mathcal{L}_{MMD}$ and from 53.5 to 65.1 with $\mathcal{L}_{KD}$ for S2P-SpCL. Furthermore, combining both losses leads to an additional overall improvement in performance in all cases.

		S2P-SpCL								S2P-MMT							
$\mathcal{L}_{MMD}$	$\mathcal{L}_{KD}$	MS → M		MS → C		M → MS		RP → M		MS → M		MS → C		M → MS		RP → M	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
✗	✗	53.5	76.0	15.6	15.7	14.7	36.7	50.5	72.1	65.8	83.7	32.2	32.2	15.1	36.9	58.7	77.5
✓	✗	62.4	82.9	24.1	23.6	15.2	38.5	55.4	77.5	62.6	81.4	27.4	26.4	15.3	37	60.8	80.2
✗	✓	65.1	85.1	28.2	26.7	16	40	55.5	78.9	67	85.5	35.2	35.1	17.8	41.1	60.4	80.1
✓	✓	**69.1**	**87.1**	**34.3**	**35.1**	**20.2**	**46.1**	**59**	**80.5**	**70**	**87.1**	**40.4**	**42.4**	**19.5**	**43.3**	**61.4**	**81**

Table 4.2: Ablation study on the effectiveness of the $\mathcal{L}_{MMD}$ and $\mathcal{L}_{KD}$ loss functions using S2P-SpCL and S2P-MMT.

Knowledge Distillation Design. We examine our knowledge distillation mechanism focusing on two key factors: the loss function and the selection of the support set.

Regarding the support set construction, our similarity-based selection relies on a cosine similarity function ξ given in Eq. (4.2). We explore two different approaches to compute the support set as shown in Table 4.3. The first strategy employs all the images of the source domain $\mathcal{S}$ to construct the support set. The second (Rank-1 NN) selects only the most similar image from the source domain to each previously seen image, without considering its identity's other images. The similarity-based selection strategy ξ shows the best results in almost all cases as shown in Table 4.3. Furthermore, we compare our $\mathcal{L}_{KD}$ with two different losses that are widely used in the literature: $\mathcal{L}_{SP}$ [197] which uses pairwise activation similarities to supervise the training of the student model, and $\mathcal{L}_{AT}$ [198] where only the activations

are used to compute a mean squared error between the student and the teacher models. The results of Table 4.3 allow us to draw the conclusion that our knowledge distillation design better suits the setting of OUDA-Rid and outperforms both the other knowledge distillation losses and support set selection strategies.

Dist. Loss	Support Set	MS → C		RP → M	
		mAP	Rank-1	mAP	Rank-1
$\mathcal{L}_{KD}$	Source Domain $\mathcal{S}$	29.3	28.1	56.3	78.3
$\mathcal{L}_{KD}$	Rank-1 NN	29.8	29.6	56.4	78.9
$\mathcal{L}_{KD}$	Similarity-based ξ	**34.3**	**35.1**	**59**	**80.5**
$\mathcal{L}_{SP}$ [197]	Similarity-based ξ	26.5	25	55.4	78.8
$\mathcal{L}_{AT}$ [198]	Similarity-based ξ	26.4	25.6	55.5	78.8

Table 4.3: Ablation study on the design of our knowledge distillation mechanism using S2P-SpCL. We assess the impact of two key factors: the loss function and the selection function of the support set.

To qualitatively illustrate the construction of our support set, in Fig. 4.4, we show some random samples of the support set for MS→C and RP→M, where $\boldsymbol{x}^t$ is the image in the target domain and $\xi_x(\boldsymbol{x}^t)$ is its most similar image in the source domain.

The choice of the teacher. As described in Sec. 4.3.2 for S2P, knowledge distillation is performed with a teacher network obtained via EMA updates. In this ablation study, we investigate alternative solutions for the choice of the teacher model as shown in Table 4.4. We analyze three teacher models: 1) at the start of each task t, the teacher is frozen and initialized by the weights of the fine-tuned model on the previous task $\mathcal{F}_{t-1}$; 2) the teacher is an EMA of the student model, being updated only at the end of the previously seen tasks $\bar{\mathcal{F}}_{t-1}$; and 3) the mean teacher $\bar{\mathcal{F}}$ obtained via EMA after each iteration (*i.e.*, one mini-batch pass) as in Sec 4.3.2. The results in Table 4.4 suggest that the choice of the teacher model is highly critical to alleviating the problem of catastrophic forgetting and that the proposed solution outperforms other alternatives.

Figure 4.4: The support set construction based on the similarities between the source domain MS (RP respectively) and the target domain C (M respectively).

Teacher Model	MS → C		RP → M	
	mAP	Rank-1	mAP	Rank-1
Task-specific $\mathcal{F}_{t-1}$	14.3	14.9	28.7	57
EMA of task-specific $\bar{\mathcal{F}}_{t-1}$	14.8	15.1	28.3	55.7
EMA of the student $\bar{\mathcal{F}}$	**34.3**	**35.1**	**59**	**80.5**

Table 4.4: Ablation study on the choice of the teacher model for Knowledge Distillation using S2P-SpCL.

4.5 Conclusions

In this chapter, we introduced a new Source-guided Similarity Preservation (S2P) algorithm for the problem of Online Unsupervised Domain Adaptation for person Re-identification (OUDA-Rid). S2P jointly addresses catastrophic forgetting and domain shift with a knowledge distillation mechanism that respects data privacy regulations. This mechanism is based on a support set composed of source images similar to previously seen identities in the target dataset. We also introduced an explicit source-target distribution alignment and a pseudo-labeling strategy to alleviate the domain shift. We performed extensive experiments where S2P straightforwardly incorporates existing state-of-the-art UDA methods and consistently outperformed them by significant margins. This chapter concludes our discussion and contributions in integrating data storage regulation constraints into the deployment of Re-ID models. In the following chapter, we will shift our focus to another form of privacy regulations concerning data transfer, as outlined in Sec. 1.2.3.

Chapter 5

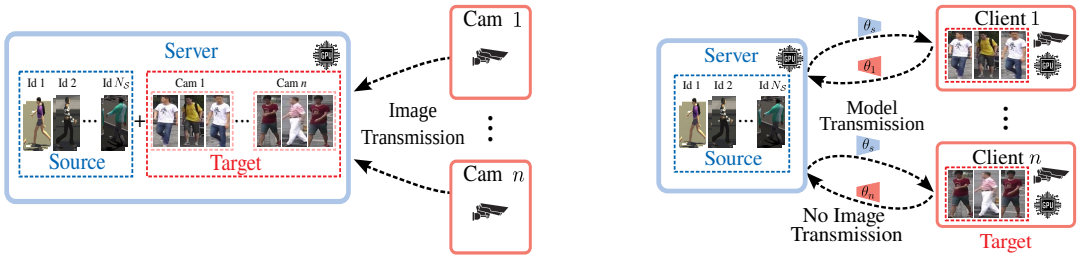
Privacy-Preserving Adaptive Re-Identification With no Image Transfer

5.1 Introduction

In the previous chapters (Chaps 3,4), we introduced the OUDA-Rid as a novel setting that adequately simulates real-world Re-ID scenario by incorporating constraints related to data storage. Alongside this, another crucial aspect involves dealing with another form of privacy regulations related to data transfer. Despite remarkable advancements in recent years [68, 70], applying UDA to person Re-ID (*UDA-Rid*) encounters privacy concerns due to the need to collect, store and transfer images of individuals in public areas. Rigorous privacy regulations in many countries restrict technology providers from retaining images of people. For example, within the European Union, the General Data Protection Regulation (GDPR) obligates technology providers to adhere to the principles of "*Data Minimization*" [200] and "*Purpose Limitation*" [201], requiring that personal data be processed only when it is necessary for a designated purpose. These general principles prompt the following question: *What minimal data usage is truly "necessary" for Re-ID systems?*

An initial response to this question can be derived from the findings of the previous chapter (Chap 4), where we presented the S2P framework designed to eliminate the necessity for storing images when deploying Re-ID models. We have shown that S2P effectively adapts UDA methods to align with privacy regulations, thereby clarifying GDPR's practical implications. However, these methods typically require transferring all captured images to a central server, which also poses privacy challenges [202]. Our work explores an alternative perspective on the question of minimal data usage: *Is transferring images outside the cameras truly "necessary" for Re-ID?* Our goal is to demonstrate that adaptation can be performed exclusively within edge devices, ensuring no image data is transmitted beyond its capture point as illustrated in Fig. 5.1b. This paradigm provides a privacy-compliant solution while leveraging the benefits of advanced Re-ID models.

To avoid the need for transmitting images, we approach this privacy-preserving Distributed UDA for person Re-ID (DUDA-Rid) task as a federated learning problem which in-



(a) Traditional UDA for Person Re-ID (UDA-Rid) (b) Privacy-preserving Distributed UDA for person Re-ID (DUDA-Rid)

Figure 5.1: In traditional Unsupervised Domain Adaptation (UDA) as depicted in Fig. (a), images are transmitted to a centralized server, which combines the unlabeled target images with the annotated source samples to train a model. In contrast, Distributed UDA for person re-identification (DUDA-Rid) shown in Fig. (b) keeps target images exclusively on edge devices. The learning process is divided between the server and cameras, the latter being equipped with local computational resources (). Only model parameters are exchanged between the clients and the server.

herently entails two interconnected challenges: (i) training the model in a distributed setup, and (ii) addressing the domain gap between the source and target datasets. Therefore, the key challenge behind the proposed setting is to simultaneously tackle the domain gap while working within a federated learning framework.

To jointly address the privacy and domain shift challenges in DUDA-Rid, we introduce a novel Federated Prototype-based learning for person Re-ID (*Fed-Protoid*) algorithm that enables domain adaptation without transmitting any image over the camera network. Fed-Protoid integrates a pseudo-labeling framework within the federated learning setup, and we propose a distributed version of the Maximum Mean Discrepancy (*MMD*) technique to enhance alignment between the source and target domains. Usually, MMD is calculated in a reproducing kernel Hilbert space using the kernel trick, which involves comparing source and target samples. Instead, we compute source prototypes and only share these prototypes with clients to adhere to privacy constraints. This approach for domain adaptation achieves high adaptation capabilities while keeping communication requirements to a minimum. Fed-Protoid readily outperforms all evaluated methods for DUDA-Rid in various challenging conditions in real-to-real and synthetic-to-real tasks. Furthermore, we show that using self-supervised pre-training [54] coupled with a Vision Transformer (*ViT*) significantly enhances performance across most scenarios for DUDA-Rid. We refer to this architecture as Fed-Protoid++.

The main contributions of this chapter can be summarized as follows:

- To our knowledge, we are the first to introduce and address the DUDA-Rid problem.
- We introduce a novel Fed-Protoid algorithm that uses prototypes to jointly address distributed learning and domain shift in DUDA-Rid. To this end, we propose a distributed version of the MMD loss to solve the domain gap in the federated setting.
- We perform extensive experiments in real-to-real and synthetic-to-real tasks with four datasets for DUDA-Rid. Fed-Protoid outperforms previous state-of-the-art person Re-

ID methods in federated learning under similar conditions. A set of ablation studies validates each component of our algorithm.
- We further propose a Fed-Protoid++, which uses ViT and recent self-supervised pre-training techniques to achieve additional gains[1].

5.2 Related Work

This section extends the previously discussed related work in three ways. first, we introduce domain-invariant feature learning based methods as another form of UDA techniques, and we discuss their limitations within our specific DUDA-Rid setting. Second, we recall briefly the previously adapted methods from the field of federated learning to the Re-ID task, showing their limitations. Finally, we present a literature review on Prototypical learning and detail how our approach compares with these methods.

Domain adaptation for person Re-ID. The current methods for domain adaptation can be broadly classified into three categories. The first is the *domain translation-based* methods [163, 164, 165], which use style transfer techniques such as CycleGAN [166] to modify the source domain to match the appearance of the target set. Recent studies in this category have focused on enhancing the translation process via self-similarity preservation [167] or camera-specific translation [168]. These types of methods are not well-suited for the DUDA-Rid problem since current federated learning methods with generative models are limited to toy datasets such as MNIST or CIFAR-10 [203, 204].

The second category is based on *domain-invariant* feature learning. Shan *et al.* [205] proposed a framework for Re-ID by minimizing the distribution variation of the source's and target's mid-level features based on the MMD loss. Huang *et al.* [206] designed a novel domain adaptive module to separate the feature map, while Liu *et al.* [207] introduced a coupling optimization method for domain adaptive person Re-ID. Despite their effectiveness, these methods assume unrestricted access to the target domain on the server, relying on continuous image transmission and storage between cameras and the central server, an assumption that conflicts with privacy constraints in real-world applications.

The third category is the *pseudo-labeling* methods that utilize an iterative process alternating between clustering and fine-tuning [169, 170, 49, 171, 172], as described in the previous chapters. We opt for the pseudo-labeling framework as it outperforms previous techniques on most datasets and since it is compatible with our DUDA-Rid setting. Nevertheless, naively using a pseudo-labeling framework like MMT [68] in the federated scenario incurs high communication costs. Therefore, we design our approach to reduce communication requirements between the clients and the server. Furthermore, our pseudo-labeling approach is enhanced with an explicit feature alignment mechanism based on MMD minimization.

Federated learning for person Re-ID. Federated Learning (FL) [148] aims at learning separately from multiple models trained on edges local data. FL restricts the sharing of data

[1]Code available: https://github.com/ramiMMhamza/Fed-Protoid

between clients and the server, as well as between clients to protect data privacy. FL has been applied to various computer vision tasks like image segmentation [150], classification [151], and person Re-ID [152]. Federated Averaging (*FedAvg*) [148] was first proposed by McMahan *et al.* based on averaging local models trained with local data and redistributing the averaged server model to the edges. Since FedAvg requires all the models in the edges to be identical to the server model, Federated Partial Averaging (*FedPav*) [152] was proposed to leverage only the common part of the clients' models (for example the backbones). In this work, we adapt the FedPav to include also the weights of the model being trained on the labeled source domain.

FL has also been investigated in the task of person Re-ID. FedReID [159] was first proposed to solve the task of supervised person Re-ID, which incorporates the FedPav optimization technique. A second work that also tackles the problem of FL in person Re-ID is FedUnReID [160], where the authors proposed an adaptation of the well-known unsupervised baseline for person Re-ID BUC [49]. In this spirit, FedUCA [161] was recently introduced to address the challenge of FL for person Re-ID. The authors draw inspiration from CAP [162], adopting both inter- and intra-camera losses to update a memory bank for each client. These methods focus on the setting of federated by dataset. This setting represents client-edge architecture, where clients are defined as the edge servers. Each edge server collects and processes images from a network of multiple cameras. In contrast, our work focuses on a more restricted federated setting which does not allow the transmission of images between the cameras and any edge server. Finally, adapting FedUCA to our context is impractical. This is because, in our setting, each client possesses images from a single camera device, rendering the optimization of the inter-camera loss unfeasible.

Prototypical learning. The concept of prototypes in modern machine learning was first introduced in the field of few-shot learning to learn a metric space where classification can be performed by computing distances to prototype representations of each class [208]. Following this spirit, prototypical networks were applied to various computer vision tasks, such as semantic segmentation [209, 210] and continual learning [133, 211]. Prototypical learning has also made its way into federated learning, initially applied to diverse domains unrelated to person Re-ID. For instance, Federated Prototype learning (*FedProto*) [212] strives to align features globally using prototypes. Classifier Calibration with Virtual Representations (*CCVR*) [213] generates virtual features by leveraging an approximated Gaussian mixture model. More recently, Federated Prototypes Learning (*FPL*) [214] incorporates cluster prototypes and unbiased prototypes to mitigate the domain gap between the data in the server and clients. Notably, these previous methods are tailored for scenarios where prior information about the number of classes is available, such as in MNIST and CIFAR-10 datasets. Fed-Protoid is the first attempt to leverage prototypes in DUDA-Rid, which brings new challenges due to the unsupervised nature of the problem.

5.3 Federated Prototype-based Re-ID

Problem definition. The objective of this chapter is to train a model $F_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ to identify individuals in a collection of n cameras deployed in a target environment. To

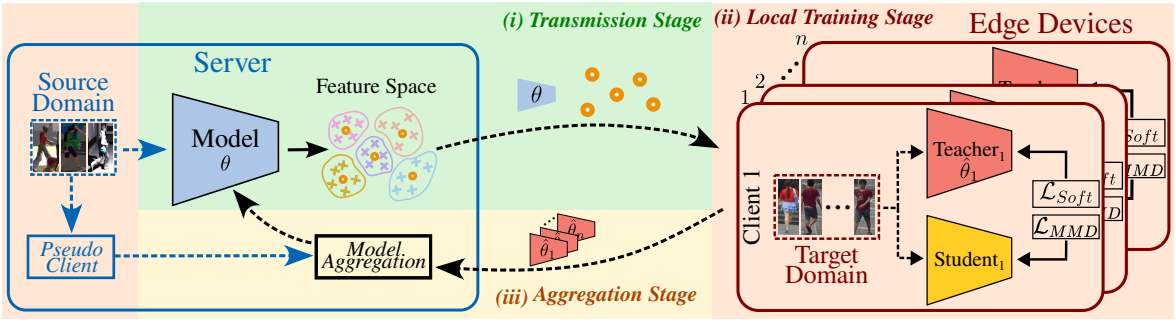


Figure 5.2: The pipeline of Fed-Protoid. Our algorithm aggregates n edge-client models and one pseudo-client model in the server. Therefore, prototypes are computed with the aggregated model from the feature space of the source domain. The prototypes and aggregated model are then distributed to all edge devices for local unsupervised training and adaptation. This local training on each client involves cross-entropy, triplet, and Maximum Mean Discrepancy (MMD) loss functions.

this end, we have at our disposal n unlabeled datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ associated to each camera-client. Each dataset is composed of N_i training samples (images): $\mathcal{D}_i = \{\boldsymbol{x}_j^{(i)}\}_{j=1}^{N_i}$. Each target dataset $\mathcal{D}_i$ is confined to its respective edge camera device and cannot be transmitted, with each camera functioning as a client that interacts solely with a centralized server. We also have an annotated source dataset $\mathcal{S} = \{(\boldsymbol{x}_j^{\mathcal{S}}, \boldsymbol{y}_j^{\mathcal{S}})\}_{j=1}^{N_s}$ available on the server, where N_s represents the number of instances in the source dataset. The main challenge in this DUDA-Rid setting is to align the distributions of the different clients with the source domain in a distributed and privacy-preserving manner, *i.e.*, without sharing images at any point.

In classical UDA-Rid joint learning, the training objective commonly involves two main loss terms: the source domain loss $\mathcal{L}_s$, and the target domain loss $\mathcal{L}_t$. In non-distributed UDA-Rid, learning is commonly performed via the minimization of a linear combination of both source and target domain datasets as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{S}}\mathcal{L}_s(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\mathcal{L}_t(\boldsymbol{\theta}, \boldsymbol{x}), \tag{5.1}$$

where $\mathcal{D} = \bigcup_{i=1}^{n} \mathcal{D}_i$. Typically, this loss is minimized using stochastic gradient descent. However, in our DUDA-Rid setting, the gradient of this total loss cannot be estimated without important communication costs. This is because the source term can be accessible only on the server via the source model, which we designate as the *pseudo-client*. Meanwhile, each device i is limited to compute only its local target loss term: $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_i}\mathcal{L}_t(\boldsymbol{\theta}, \boldsymbol{x})$. In the following, we outline our training strategy to minimize the total loss $\mathcal{L}$ in a distributed manner. Additionally, we describe the specifications of each loss term to facilitate efficient communication and robust learning.

5.3.1 Overview of Fed-Protoid

Fig. 5.2 shows the pipeline of Fed-Protoid for the DUDA-Rid setting. Our algorithm aggregates n client models along with the pseudo-client in a distributed setting. It adheres to

standard practices in FL and functions in rounds. Each round is composed of three stages: (i) **transmission stage**: the aggregated model is distributed to every client and pseudo-client; (ii) **local training stage**: each client, as well as the pseudo-client, adapts their local model; (iii) **aggregation stage**: the local models are transmitted back to the server for aggregation.

At the beginning of each new round, the *transmission stage* also includes the transfer of source prototypes that are later used for source-target alignment. The aggregated model $F_{\boldsymbol{\theta}}$ computes the features of the source samples and the prototypes of each individual as the centroid of its feature representations. The prototypes of $\mathcal{S}$ are then transmitted along with the aggregated model $F_{\boldsymbol{\theta}}$ to all clients. Note that we assume the server utilizes either synthetic data or real data gathered in compliance with relevant legislation. Consequently, the transmission of source prototypes does not breach the privacy-preserving constraints.

In the *local training stage*, we use a teacher-student architecture to adapt $\boldsymbol{\theta}$ to the unlabeled target dataset $\mathcal{D}_i$ on each device i, and to the labeled source dataset $\mathcal{S}$. The server updates the pseudo-client via supervised training, while the local adaptation on each client involves cross-entropy, triplet, and MMD loss functions. Considering the use of the cross-entropy loss and the variation of the number of identities for each client, we add to each local device i a personalized classifier head $\mathcal{C}_i$. This classifier is designed to match the number of classes to the respective number of identities in each client, including the pseudo-client.

Finally, in the *aggregation stage*, the server gathers and aggregates the n client models $\{F_{\hat{\boldsymbol{\theta}}_1}, F_{\hat{\boldsymbol{\theta}}_2}, \ldots, F_{\hat{\boldsymbol{\theta}}_n}\}$ obtained in the *local training stage* and the model $F_{\hat{\boldsymbol{\theta}}_s}$ trained on the source dataset using a weighted average sum as follows:

$$\boldsymbol{\theta} = \alpha\hat{\boldsymbol{\theta}}_s + (1-\alpha)\sum_{i=1}^{n} w_i\hat{\boldsymbol{\theta}}_i, \tag{5.2}$$

where $\{\hat{\boldsymbol{\theta}}_s, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_n\}$ are the parameters of the client models after adaptation, α is the weight contribution of the pseudo-client model $\hat{\boldsymbol{\theta}}_s$, and w_i is the weight assigned to the ith client model given by $w_i = \frac{N_i}{\sum_{i=1}^{n} N_i}$.

5.3.2 Teacher-student architecture

All clients, encompassing the pseudo-client, employ the same teacher-student architecture. This framework is chosen for its effectiveness in enabling self-training techniques, which have been shown to yield optimal performance in UDA-Rid scenarios. While self-training is not required in the source domain due to its labeled nature, the use of the teacher-student framework favors similar training dynamics across both clients and the pseudo-client, facilitating more efficient model aggregation.

For simplicity, we assume here that we are in the ith client. Firstly, we initialize at each round the parameters of the teacher model $\bar{\boldsymbol{\theta}}_i$ and student model $\boldsymbol{\theta}_i$ with the parameters of the aggregated model $\boldsymbol{\theta}$. During adaptation, the student model is updated through the minimization of the target loss function $\mathcal{L}_t(\cdot)$ which are later detailed in Sections 5.3.3 and 5.3.4. After back-propagation through the student, we use the Exponential Moving Average (*EMA*) parameters update [193, 194] to compute the teacher model. At every iteration t, the

parameters $\bar{\boldsymbol{\theta}}_i^{(t+1)}$ of the teacher model are given by:

$$\bar{\boldsymbol{\theta}}_i^{(t+1)} = \tau \bar{\boldsymbol{\theta}}_i^{(t)} + (1-\tau)\boldsymbol{\theta}_i, \tag{5.3}$$

where $\tau \in [0, 1)$ is a weighting factor. The model $\hat{\boldsymbol{\theta}}_i$, which is sent back to the server for model aggregation, is assigned to the final teacher model $\bar{\boldsymbol{\theta}}_i^{(t)}$.

5.3.3 Prototype estimation and server training

Source prototypes. In the *transmission stage*, the server sends prototypes to all the target clients. These prototypes are defined as the mean feature representation for each identity from the source domain. Formally, the prototype $\boldsymbol{p}_k$ of the kth identity is given by:

$$\boldsymbol{p}_k = \frac{1}{|\mathcal{S}_k|} \sum_{l \in \mathcal{S}_k} F_{\boldsymbol{\theta}} \left(\boldsymbol{x}_l^{\mathcal{S}} \right) \ \forall\, 1 \leq k \leq K, \tag{5.4}$$

where K is the number of identities in $\mathcal{S}$, $\mathcal{S}_k \subset \mathcal{S}$ is the set of images of the kth identity, and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset \ \forall i \neq j$. With enough diverse identities and images per identities from the source domain, the set of all source prototypes can serve as an approximation of the source domain distribution which can be transmitted with little cost. Subsequently, we use them to align the source and target distributions in the edge devices in the *local training stage*.

Pseudo-client loss. The source domain is treated as a pseudo-client in the *local-training stage*. Since the pseudo-client has access to the source domain dataset with labeled samples $\mathcal{S} = \{(\boldsymbol{x}_j^{\mathcal{S}}, \boldsymbol{y}_j^{\mathcal{S}})\}_{j=1}^{N_s}$, we can compute a supervised source loss $\mathcal{L}_s$ for the jth sample as:

$$\mathcal{L}_s(\boldsymbol{x}_j^{\mathcal{S}}, \boldsymbol{y}_j^{\mathcal{S}}) = \mathcal{L}_{CEs} + \mathcal{L}_{Tris}, \tag{5.5}$$

with

$$\begin{aligned}
\mathcal{L}_{CEs} =& \beta_1 \mathcal{L}_{CE} \left(\mathcal{C}_s \circ F_{\boldsymbol{\theta}_s}(\boldsymbol{x}_j^{\mathcal{S}}), \boldsymbol{y}_j^{\mathcal{S}} \right) + \beta_2 \mathcal{L}_{CE} \left(\mathcal{C}_s \circ F_{\boldsymbol{\theta}_s}(\boldsymbol{x}_j^{\mathcal{S}}), \bar{\mathcal{C}}_s \circ F_{\bar{\boldsymbol{\theta}}_s}(\boldsymbol{x}_j^{\mathcal{S}}) \right), \\
\mathcal{L}_{Tris} =& \gamma_1 \mathcal{L}_{Tri} \left(F_{\boldsymbol{\theta}_s}(\boldsymbol{x}_j^{\mathcal{S}}), \boldsymbol{y}_j^{\mathcal{S}} \right) + \gamma_2 \mathcal{L}_{Tri} \left(F_{\boldsymbol{\theta}_s}(\boldsymbol{x}_j^{\mathcal{S}}), F_{\bar{\boldsymbol{\theta}}_s}(\boldsymbol{x}_j^{\mathcal{S}}) \right),
\end{aligned}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\bar{\mathcal{C}}$ is the teacher classifier head, $\mathcal{L}_{Tri}$ is the triplet loss, $\beta_1 + \beta_2 = 1$, and $\gamma_1 + \gamma_2 = 1$.

5.3.4 Local training on edge devices

We now detail the *local training stage* for the clients. A key difficulty of the target domain training is the estimation of the number of identities from an unlabeled set of images $\mathcal{D}_i = \{\boldsymbol{x}_j^{(i)}\}_{j=1}^{N_i}$. To this end, we apply the pseudo-labeling technique [169, 170, 49] consisting of an iterative process between clustering with the DBSCAN [45] method and fine-tuning. After this pseudo-labeling process we get an augmented dataset $\tilde{\mathcal{D}}_i = \{\boldsymbol{x}_j^{(i)}, \tilde{\boldsymbol{y}}_j^{(i)}\}_{j=1}^{N_i}$, where $\tilde{\boldsymbol{y}}_j^{(i)}$ is the pseudo-label associated to the jth sample.

Target client loss. In the edge devices, the teacher model generates soft labels that guide the student model to be less confident about the hard pseudo-labels [66]. This results in a refinement of the wrong predictions of the student model. Specifically, for a given target client dataset $\mathcal{D}_i$, the local loss function $\mathcal{L}_i$ in a mini-batch is given by:

$$\mathcal{L}_i = \frac{1}{m} \sum_{j \in \mathcal{D}_{i,m}} \mathcal{L}_p(\boldsymbol{x}_j^{(i)}) + \lambda \mathcal{L}_{MMD}\left(\mathcal{D}_{i,m}, \mathcal{P}_m\right), \tag{5.6}$$

where $\mathcal{D}_{i,m} \subseteq \mathcal{D}_i$ is the set of images in the mini-batch with $|\mathcal{D}_{i,m}| = m$, $\mathcal{L}_p(\boldsymbol{x}_j^{(i)})$ is a pseudo-label loss for the jth sample, λ is a weighting factor, and $\mathcal{L}_{MMD}\left(\mathcal{D}_{i,m}, \mathcal{P}_m\right)$ is the MMD loss between $\mathcal{D}_{i,m}$ and a subset of the prototypes $\mathcal{P}_m \subseteq \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_K\}$ with $|\mathcal{P}_m| = m$. The pseudo-label loss $\mathcal{L}_p(\boldsymbol{x}_j^{(i)})$ is the same as in (5.5), but since the true labels are not available in the clients, we use the pseudo-labels $\tilde{\boldsymbol{y}}_j^{(i)}$ instead. The local loss is used to update the student parameters $\boldsymbol{\theta}_i$.

Personalized pseudo-epoch. A significant challenge in federated learning scenarios is determining the optimal number of training epochs for each client. This decision is crucial to achieve the best balance between learning efficiency and transmission overhead. In our task, this problem is also crucial to prevent over-fitting in clients with only a few identities or images. To ensure equal usage of all identities within a client during a federated training round, we introduce the *Personalized Pseudo-Epoch* (*PPE*).

For a specific client i, let K_i represent the count of identities in $\mathcal{D}_i$ as identified by the DBSCAN algorithm. In every iteration, mini-batches are constructed by randomly selecting I identities. From each chosen identity, B images are sampled, as in previous works [68, 70]. Consequently, we define the number of iterations required for one PPE as $P_i = \frac{K_i}{I}$. By doing so, we ensure that, during a federated training round, each identity is presented an equal number of times, irrespective of the varying number of identities present in each client's dataset.

5.4 Experiments and Results

In this section, we detail our experimental setup, covering datasets, implementation details, and evaluation metrics. Subsequently, we compare Fed-Protoid against two categories of approaches: (i) FL + UDA, wherein we adapt the UDA methods MMT [68] and SpCL [70] to DUDA-Rid, and (ii) federated learning approaches for person Re-ID, namely FedReID [159] and FedUnReID [160]. Finally, we conduct a series of ablation studies to (i) validate the teacher-student architecture and aggregation choice, (ii) confirm the suitability of the MMD loss, and (iii) demonstrate the efficacy of the transformer-based architecture coupled with self-supervised pre-training (Fed-Protoid++).

5.4.1 Experimental setup

Datasets. We evaluate our method in real-to-real and synthetic-to-real scenarios. For the source domain, we use two datasets: *MSMT* (MS) [78] and *RandPerson* (RP) [83]. For the target domain, we use the *Market* (M) dataset [76], previously introduced, alongside the new protocol *CUHK03-np* (C) [215] which consists of splitting the CUHK03 dataset into 767 identities for training and 700 identities for testing. In testing, each query identity is selected by both cameras to ensure the evaluation of the cross-camera Re-ID.

Evaluation protocol. In the DUDA-Rid setting, we assume the cameras are equipped with embedded devices that can train the teacher-student models of the clients. To mimic this scenario, we split *Market* into six clients and *CUHK03-np* into two clients, where each client contains images from a single camera viewpoint. We adopt the commonly used metrics for evaluation in person Re-ID [68, 70]: mean Average Precision (mAP) and CMC Rank-1 [76] accuracies. During each round of the federated learning, each client performs a number of PPEs. Therefore, we compute the metrics on a separate test set related to the target domain using the aggregated model from the server. We report for each method the highest average mAP and Rank-1, with the number of rounds required to reach these top scores.

Implementation details. For a fair comparison with the state-of-the-art methods, we follow the common practices in the UDA for person Re-ID field by adopting ResNet-50 [28] pre-trained on ImageNet [84] as a backbone. We train every method for 800 rounds of federated learning. Except FedUnReID, where we follow its implementation details and set the training number of rounds to 200. We stop the training process upon observing any signs of divergence, specifically when there is a considerable decline in the test mAP over the training rounds. We present a sensibility analysis of the hyper-parameters of Fed-Protoid in the supplementary material B.

To stress the practicality of the adopted setting, we also consider a variant of Fed-Protoid called Fed-Protoid++, where we employ a stronger backbone architecture and leverage as initialization a model pre-trained on a large-scale Re-ID dataset. Concerning the architecture, we transition from the traditional ResNet-50 to a ViT [39] backbone. We complement the backbone improvement with the adoption of self-supervised pre-trained models on the large-scale unlabeled dataset LUPerson [53].

5.4.2 Comparison with the state-of-the-art

Since Fed-Protoid is the first method that addresses DUDA-Rid, we adapt various methods initially designed for alternative settings to facilitate a comparison with Fed-Protoid.

Competitive methods. We assess the performance of Fed-Protoid against two federated frameworks for person Re-ID: FedReID [159] and FedUnReID [160]. On one hand, FedReID is a Fully Supervised (*FS*) method that incorporates dynamic weight adjustment, knowledge distillation, and FedPav as its aggregation rule. The original study of FedReID also explores our dataset partition in the edge devices, *i.e.*, each client contains images from a single camera. However, a key distinction is that FedReID requires the target dataset $\mathcal{D}_i$ to be labeled, while we do not have such a constraint. On the other hand, FedUnReID

Method	Type	MS → M			MS → C			RP → M			RP → C		
		mAP	Rank-1	#R	mAP	Rank-1	#R	mAP	Rank-1	#R	mAP	Rank-1	#R
FedReID [159]	FS	38.9	61.9	800	11.6	11.7	750	38.9	61.9	800	11.6	11.7	750
FedReID+$\mathcal{S}$	FS	39.5	63.8	790	12.0	12.3	800	40.0	64.4	800	11.4	11.6	780
FedUnReID [160]	PU	19.5	43.6	190	6.8	7.0	170	19.5	43.6	190	6.8	7.0	170
FedUnReID+$\mathcal{S}$	PU	31.0	61.7	170	10.5	11.1	170	31.5	31.8	170	10.6	11.6	160
FedAvg+SpCL [70]	UDA	39.1	67.3	8	19.7	18.9	1	36.1	62.9	9	21.2	21.6	3
FedPav+MMT* [68]	UDA	45.8	73.6	70	22.4	21.9	9	30.2	58.9	9	19.0	19.7	9
Fed-Protoid (ours)	UDA	51.0	76.8	288	23.8	23.1	22	39.2	66.4	22	25.1	24.7	253
Fed-Protoid++ (ours)	UDA	**61.7**	**82.6**	170	**43.8**	**42.4**	24	**45.2**	**71.8**	186	**25.7**	**24.9**	212

Table 5.1: Comparison of mAP, Rank-1 accuracy, and number of rounds (#R) for four adaptation configurations. The different methods range from Fully Supervised (FS), Purely Unsupervised (PU) to Unsupervised Domain Adaptation (UDA). *The communication cost for a single round in MMT is four times greater than that in the other ResNet-based models.

is a framework that adapts the Purely Unsupervised (*PU*) baseline Bottum-Up-Clustering (*BUC*) [49] for Federated person Re-ID. We also compare Fed-Protoid with FedReID+$\mathcal{S}$ and FedUnReID+$\mathcal{S}$. These variants are improved versions of the original frameworks where we initialize the models with supervised source pre-training, offering a fairer comparison with Fed-Protoid that leverages the source domain's knowledge.

Since our DUDA-Rid setting combines both UDA and FL, we extend our comparison to include Fed-Protoid against UDA methods for person Re-ID. For the UDA methods, we adapt the state-of-the-art pseudo-labeling approaches SpCL and MMT to suit the DUDA-Rid setting. In this process, during each federated learning round, we send copies of these UDA frameworks to all the edge clients for local training. Additionally, for a fair comparison with Fed-Protoid, we train in the server the pseudo-source client on the labeled source domain $\mathcal{S}$. The aggregation rule for these adapted UDA methods, denoted FedAvg+SpCL and FedPav+MMT is consistent with Eq. (5.2). The main objective of this comparison is to evaluate the effectiveness of traditional UDA methods when confronted to privacy constraints, where the target domain is distributed over multiple edge devices (cameras).

Quantitative results and discussions. Table 5.1 reports the best mAP accuracy and CMC Rank-1 score alongside the number of rounds (#R) required to achieve these top scores. We include two real-to-real configurations MS $\rightarrow$ M, MS $\rightarrow$ C, and two syntetic-to-real RP $\rightarrow$ M, RP $\rightarrow$ C.

Fed-Protoid demonstrates good results against the supervised and unsupervised federated learning methods for person Re-ID, FedReID and FedUnReID as shown in Table 5.1. For example, Fed-Protoid obtains 23.8 of mAP in MS $\rightarrow$ C, outperforming FedReID with 11.6 mAP and FedUnReID with 6.8 mAP. More interestingly, Fed-Protoid reaches this performance after only 22 rounds, whereas FedReID and FedUnReID require 750 and 170 rounds, respectively. Fed-Protoid also reaches superior performance in the RP $\rightarrow$ C configuration with 25.1 mAP compared to the other federated learning methods. We also evaluate the improved versions FedReID+$\mathcal{S}$ and FedUnReID+$\mathcal{S}$ where both models start with a pre-training on the source domain. Even though starting from the source pre-trained models improves slightly the original models' performances, they are below the performances obtained by Fed-Protoid in almost all configurations.

Fed-Protoid also improves significantly the performance of the adapted UDA baselines SpCL and MMT as shown in Table 5.1. In MS $\rightarrow$ M, SpCL and MMT achieve a mAP of 39.1 and 45.8, respectively, while Fed-Protoid achieves a mAP accuracy of 51. This observation can also be generalized to the synthetic-to-real configurations like RP $\rightarrow$ M, where Fed-Protoid reaches a performance of 39.2 mAP, while SpCL and MMT achieve 36.1 and 30.2 mAP, respectively. Even though Fed-Protoid requires more communication rounds to reach its optimal performance compared to MMT, it is important to notice that Fed-Protoid transmits approximately only a quarter of the data weights per round. This is because the MMT architecture sends four backbones to the server, whereas Fed-Protoid needs to share only one (the teacher model) and the transmission cost of the prototypes is almost negligible compared to the weights of the models. Overall, Fed-Protoid is more effective in the DUDA-Rid scenario than the adapted UDA baselines SpCL and MMT as shown in Table 5.1.

Fed-Protoid++ Recent work [53] has shown the suitability and effectiveness of self-supervised pre-training methods for transformer-based methods [39] in person Re-ID, yielding substantial enhancements across a variety of Re-ID benchmarks. In the context of our DUDA-Rid setting, the performance of Fed-Protoid++ is consistent with the aforementioned findings as shown in Table 5.1. Particularly, transitioning from the ResNet-50 to a ViT backbone pre-trained in a self-supervised way leads to remarkable performance enhancements in all the configurations. For instance, we observe an increase in the mAP from 51 to 61.7 in MS $\rightarrow$ M. Similarly, we have an improvement from 39.2 to 45.2 in RP $\rightarrow$ M in the mAP, showing Fed-Protoid++ enhanced effectiveness. The improvement in the performance of using transformer-based models in person Re-ID comes from three main reasons [39]: (i) the multi-head self-attention effectively captures long-range dependencies and drives the model to focus on diverse human-body parts, (ii) transformer-based models have the ability to extract fine-grained features which is essential in person Re-ID, and (iii) the rich variety and volume of the LUPerson dataset provide the model with the capability of extracting more robust features that are generalizable across small downstream datasets. We perform an ablation study in Section 5.4.3 to empirically validate these points.

Training dynamics In Fig. 5.3, we illustrate the progression of the mAP of the different methods in the MS $\rightarrow$ M configuration. Notably, there is a difference in the evolution of the mAP between the methods designed for the FL FedReID+$\mathcal{S}$ and FedUnReID+$\mathcal{S}$, and the UDA-based methods FedAvg+SpCL and FedPav+MMT. Specifically, while FedReID+$\mathcal{S}$ and FedUnReID+$\mathcal{S}$ exhibit a consistent improvement during the training, this trend is not mirrored in the performance of FedAvg+SpCL and FedPav+MMT. In fact, both UDA-based methods tend to converge rapidly at the early stage of FL training. This is because initially, the local models are relatively close to the source model, allowing for easier leveraging of the source domain knowledge in the first rounds of FL. However, as training progresses, the local models start diverging from the source model, leading to a decrease in performance. Conversely, our methods demonstrate a stable progression, effectively managing to mitigate domain shift during training. This highlights the effectiveness of our approach in maintaining consistent performance in the DUDA-Rid setting.

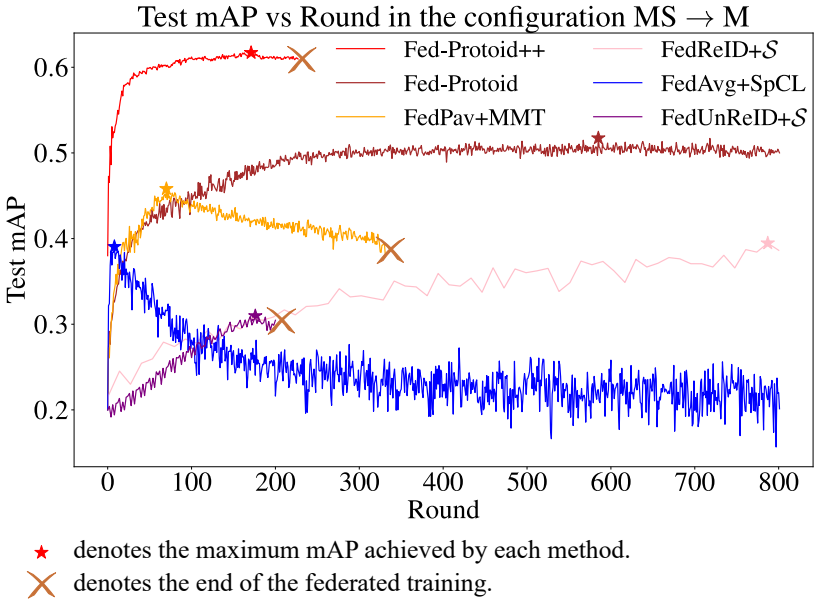


Figure 5.3: Test mAP vs Round of the different methods in the real-to-real configuration MS → M.

5.4.3 Ablation studies

This section presents ablation study results to assess the contribution of Fed-Protoid's main components.

On the effectiveness of the teacher-student framework. The integration of the teacher-student architecture gives multiple possibilities to the design of Fed-Protoid. Table 5.2 shows the results of the ablation study where (i) we do not have the teacher-student framework in the pseudo-client, (ii) we have the teacher-student framework in the pseudo-client and we transmit the students for aggregation, and (iii) we have the teacher-student framework in the pseudo-client and we transmit the teachers for aggregation. Table 5.2 shows a considerable drop in performance when the teacher model is omitted from the pseudo-client in both MS → M and MS → C. Specifically, the mAP decreases from 51 to 37.4 in MS → M, and from 23.8 to 22.5 in MS → C, underscoring the crucial role of the teacher model in the pseudo-client. These findings align with our claim in Section 5.3.2 regarding the use of the teacher-student architecture in the pseudo-client to keep similar training dynamics with the other clients. The teacher-student architecture gives another alternative of transmitting the student instead of the teacher models for aggregation. Table 5.2 illustrates that this alternative yields reasonable performance. However, using students for aggregation falls marginally short of the performance achieved by aggregating the teacher models.

On the effectiveness of the MMD loss. The MMD loss, serving as a measure of domain discrepancy, offers a variety of options for the reproducing kernel Hilbert space where we minimize the distance between source prototypes and target feature representations. Table 5.3 shows a comparison between different kernel functions including linear, order 2, and Gaussian kernels. The linear kernel minimizes the mean average of prototypes and target features distributions, while the order 2 kernel minimizes the mean average and the standard

Teach.-Stud. on $\mathcal{S}$	Transmission	MS $\rightarrow$ M	MS $\rightarrow$ C
✗	Teacher	37.4	22.5
✓	Student	<u>49.3</u>	**23.8**
✓	Teacher	**51.0**	**23.8**

Table 5.2: Ablation study of the teacher-student framework: comparing teacher vs. student model aggregation from edge devices.

MMD	Kernel	MS $\rightarrow$ M	MS $\rightarrow$ C
✗	–	42.6	22.4
✓	Linear	38.1	22.1
✓	Order 2	27.8	13.1
✓	Gaussian	**51.0**	**23.8**

Table 5.3: Impact of the kernel function choice on the effectiveness of the Maximum Mean Discrepancy (MMD) loss.

deviation of these distributions. Table 5.3 suggests that the linear and order 2 kernels are not effective in the DUDA-Rid setting. This can be attributed to potentially biased estimations of the true mean (linear) and variance (2nd order) within relatively small and diverse batches of images. Furthermore, using the MMD loss with a Gaussian kernel achieves superior performance in all cases, including when MMD loss is not used at all. We further evaluate the MMD's effectiveness by examining its performance with limited prototypes and comparing the proposed distributed MMD with the original MMD in the supplementary material B, demonstrating its robustness against device storage and communication limitations and proving it to be effective and suitable for our setting.

On the effectiveness of the backbones and pre-training datasets. The final ablation study focuses on the impact of the different modifications done to design Fed-Protoid++. Table 5.4 shows the ablation study for different backbones, pre-training strategies, and warm-up. For the backbones, we have the option to use either the classical ResNet-50 or ViT Small (S). For the pre-training dataset and strategy, we can either use fully supervised on ImageNet or self-supervised in LUPerson. The warm-up consists of adding an additional supervised pre-training on the source domain $\mathcal{S}$. We observe in Table 5.4 that adopting the ViT backbone combined with an appropriate pre-training dataset significantly enhances the performance. Fed-Protoid corresponds to a ResNet-50 backbone pre-trained on ImageNet, while Fed-Protoid++ corresponds to a ViT model pre-trained on the large-scale LUPerson dataset in a self-supervised way.

A key finding in Table 5.4 is that using ViT (S) instead of ResNet-50 with the same pre-training strategy consistently results in performance improvements. For instance, when comparing ResNet-50 and ViT (S) pre-trained on ImageNet, the mAP slightly improves from 51 to 52.4 in MS $\rightarrow$ M, and from 23.8 to 27.5 in MS $\rightarrow$ C. This suggests that the ViT-based

Backbone	Pre-tr.	Warm-up	MS → M	MS → C
ResNet-50	ImageNet	✗	41.5	23.7
ResNet-50	ImageNet	✓	**51.0**	**23.8**
ResNet-50	LUPerson	✗	44.0	13.6
ResNet-50	LUPerson	✓	46.0	16.0
ViT (S)	ImageNet	✓	52.4	27.5
ViT (S)	LUPerson	✗	59.7	23.9
ViT (S)	LUPerson	✓	**61.7**	**43.8**

Table 5.4: Impact of the backbone architecture and pre-training datasets on the performance of Fed-Protoid.

backbone learns more robust features in the target domain. Additionally, ViT (S) captures more generalizable features when pre-trained in a self-supervised way thanks to the large and diverse set of unlabeled images in LUPerson. As a final remark, the warm-up generally enhances the performances across all the scenarios and configurations.

5.5 Conclusion

In this chapter, we presented a novel approach for the task of UDA for person Re-ID that addresses both problems of domain shift and privacy preservation. To comply with privacy standards, our method Fed-Protoid learns a person Re-ID model across multiple edge devices without transmitting target images from the cameras where they were captured. By integrating a teacher-student architecture and a source-client model, trained in the server side on labeled source domain, and introducing a distributed version of the Maximum Mean Discrepancy (MMD) loss, Fed-Protoid ensures effective domain adaptation with the target clients while keeping communication requirements minimal. Our experiments show the superiority of Fed-Protoid compared to existing methods under various challenging configurations, including real-to-real and synthetic-to-real DUDA-Rid tasks.

Chapter 6

Conclusion and Future Work

Throughout this thesis, our focus was mainly on tackling the major challenges of Person Re-ID, with a specific emphasis on the domain gap and privacy concerns that arise in modern surveillance systems. The primary contribution of this research work has been the development of robust methods that adapt Re-ID models to changing environments while adhering to strict data privacy regulations, such as RGPD and the AI Act. In what follows, we will summarize our main contributions and their limitations before discussing future perspectives of this thesis.

6.1 Summary and Discussion

In Chapter 1, we introduce the context of video surveillance systems with a highlight on the transformative role of AI and machine learning in improving the functionality and integration of these systems across different sectors. We also describe the task of Person Re-ID and its crucial role in modern surveillance. A historical overview shows how Person Re-ID has changed over time, starting with its original integration with multi-camera tracking to its adaptation and integration with deep learning techniques. Additionally, we discuss the significant challenges that Person Re-ID faces in the new era of AI, particularly focusing on domain gap and privacy constraints, which are pivotal in shaping the development and deployment of Re-ID systems in real-world scenarios.

Chapter 2 consists of a literature review that extensively covers the evolution and the current state of research in the Person Re-ID field. It highlights also the significant advancements in related fields to this thesis which are Unsupervised Domain Adaptation (UDA), Continual Learning, and Federated Learning. One of the most significant advances in the field of Person Re-ID is the transition from handcrafted feature extractors to sophisticated neural network architectures that improve the robustness and discrimination capabilities of Re-ID systems. The literature review also explores the UDA techniques, which have been proven effective in adapting Person Re-ID models to new unlabeled environments, minimizing the need for extensive labeling. Other than pseudo-labeling methods, that rely solely on the iterative process of clustering and fine-tuning, in this thesis, we have also developed

statistical techniques based on the Maximum Mean Discrepancy (MMD) to further reduce explicitly the domain gap under privacy constraints. Finally, we detail the recent advances in Continual Learning and Federated Learning, since both fields present innovative solutions that combine well with UDA techniques to respect the privacy of data. On one side, Continual Learning is explored as a promising approach to deal with catastrophic forgetting in the OUDA setting (Chap 3) where images are collected as a stream of data and where no images are allowed to be stored for an unlimited amount of time. On the other side, Federated Learning is evaluated for its relevance in privacy-preserving collaborative learning, where data remains decentralized. In the following chapters, we attempt to integrate these areas to propose a unified approach that leverages the adaptive ability of UDA methods and privacy-preserving characteristics of Continual Learning and Federated Learning to address the challenges related to both the domain gap and the privacy constraints in deploying effective Person Re-ID systems.

Chapter 3 introduces our first contribution to the field of Person Re-ID. We propose a new setting called the OUDA-Rid, which involves adapting a Re-ID model trained on a labeled source dataset to an unlabeled target dataset collected in a sequential and online fashion. We outline the limitations of existing UDA methods that typically assume having access to a large set of target domain data for offline training. This is an assumption that is often violated since it relies on storing potentially large amounts of person images, whereas Re-ID systems are confronted with confidentiality purposes forcing them to discard previously collected and seen images. Furthermore, we present adaptations of different UDA frameworks to the OUDA setting covering different techniques: the strong baseline [66], teacher-student frameworks [68], and contrastive-based UDA methods [70]. These methods are evaluated in an experimental setup that simulates real-world conditions where data is provided in batches without identity overlap between them. The results show that when adapting UDA methods to the OUDA-Rid setting, the performance results in a significant drop compared to offline training, which underscores the need to design new frameworks that can effectively handle the stream of data and continuously adapt to new environments while preserving the previously acquired knowledge.

The next contribution, which is detailed in Chapter 4, explores the challenges of the OUDA-Rid setting which are domain gap and catastrophic forgetting. To this end, we propose the S2P framework which incorporates Knowledge Distillation to address the dual challenges present in the OUDA-Rid setting. S2P employs a teacher-student model and utilizes a support set that is constructed by images derived from the source domain that are similar to the target domain to preserve the previously acquired knowledge. Furthermore, it minimizes the MMD loss in the feature space between the source and target domains to facilitate the continual adaptation to the new target domain data streams. Our extensive experiments demonstrate that S2P adapts effectively existing UDA methods to the relatively complex yet practical setting of OUDA-Rid. This adaptation results in a significant performance boost compared to the results obtained by straightforwardly applying these UDA methods to the OUDA-Rid setting. Moreover, we believe that the privacy-preserving aspect of the S2P framework could serve as a model for developing similar frameworks in other fields where data is confronted with privacy concerns.

Chapter 5 explores an orthogonal yet complementary direction to the previous work while

maintaining adherence to privacy regulations. This chapter shifts our focus toward different constraints associated this time with data transfer. We challenge the need for transferring captured images to a central server for training and adapting Re-ID models, as this approach raises significant privacy concerns and increases the risks of data breaches. We started by introducing the DUDA-Rid setting which is a task that is based on performing adaptation directly within the edge devices, ensuring no image is transmitted beyond its capture point. Next, we propose a novel Federated Prototype-based learning method, named Fed-Protoid, that leverages federated learning to train the Re-ID model in a distributed manner while tackling the domain shift problem. The proposed Fed-Protoid framework integrates a pseudo-labeling framework with a distributed MMD technique for aligning the source and target domains without transmitting any image data. To adhere to the primary constraint of data transfer, Fed-Protoid computes and shares only source prototypes with target clients, thereby achieving high adaptation capabilities while ensuring data safety and minimizing communication requirements. Additionally, the source domain typically consists of an academic or synthetic publicly available dataset that is collected in a way that does not compromise privacy.

This thesis tackles the critical challenges related to privacy with the deployment of Re-ID systems. It proposes novel solutions and improvements to existing Re-ID systems to comply with privacy regulations while maintaining robust performance. In a world increasingly focused on ethical principles, particularly regarding data privacy, the findings of this thesis are timely.

Consequently, this thesis has great potential to attract industry players looking to commercialize advanced Re-ID technologies and government structures interested in deploying these technologies for public security purposes. This interest could open possible collaborative efforts that would strengthen the interactions between industrial and governmental sectors, enhance compliance with privacy requirements, and facilitate the real-world testing and refinement of Re-ID technologies. On one side, industry entities, that are continually in search of cutting-edge technologies, can build upon our findings to further develop, scale, and finally commercialize the innovations presented in this thesis. On the other side, government bodies may initiate pilot projects, grants, or partnerships that would allow for field testing and eventual deployment. This cooperative effort is essential for ensuring the development and deployment of the next generation of Re-ID systems, ensuring they are not only effective but also ethically and socially responsible.

6.2 Future Directions

Unified Framework for Federated Continual Learning: A natural direction for future work is to combine the two settings, OUDA-Rid and DUDA-Rid, to create a unified framework for Federated Continual Learning. In this context, recent work from Shenaj *et al.* [216] proposes an Asynchronous Federated Continual Learning, where the continual learning of multiple tasks occurs at each client with different orderings. This approach, which has yet to be applied to the Re-ID task, would address both data storage and data transfer regulations simultaneously, offering an environment to develop even more privacy-preserving

solutions for Re-ID. However, when considering a complete solution, a significant challenge may arise. It concerns mainly the deployment of the models on edge devices with limited storage capacity to host the Re-ID models. Techniques for efficient learning like pruning [217] and quantization [218] have shown remarkable effectiveness in decreasing the number of parameters in deep learning models, which in turn could reduce the size of the Re-ID models deployed on these devices.

Advancements in Generative AI: The recent rise of diffusion models in Generative AI highlights an exciting opportunity to generate large-scale high-quality images conditioned on text or other modalities [219, 220]. In chapter 5, our experiments indicate decent adaptation performance when the source domain is fully synthetic, which supports the potential of this approach. Moreover, combining those datasets with appropriate unsupervised and self-supervised pre-training techniques [54, 53, 221] could lead to the development of more robust and effective pre-trained models, thereby improving the initialization and the generalization of Re-ID models. This approach can be a promising alternative to pre-trained models on ImageNet [84], which contains images manually assigned one-hot labels from a pre-defined set, thereby completely ignoring the rich semantic content beyond these categories.

Integration of Foundation Models: Foundation models (*e.g.* DALL-E [222], GPT-4, and Llama-3 [223]) are models trained on extensive datasets using generally self-supervision at scale. These models are versatile and capable of being adapted to a wide range of downstream tasks. Incorporating Large Language Models (LLMs) and Vision Language Models (VLMs) into Re-ID systems will offer a new avenue for enhancing model training with the two modalities, images and text. The simplistic idea is to fine-tune the visual model of VLMs, like CLIP [224] or Llava [225] on Re-ID datasets, which already obtained competitive performances in various Re-ID tasks [226]. To go beyond this basic idea, exploring techniques for efficient learning such as text prompt learning [227] and multi-modal prompt learning [228] could be investigated to minimize further the computation cost of deploying Re-ID models on edge devices.

6.3 Limitations

A significant limitation of the research presented in this thesis is the possible violation of privacy when the Re-ID model is being tested. This problem is outside the main scope of this thesis, yet it is still significant to talk about it and its potential fixes and solutions.

Technically speaking, sending the person of interest's anchor image to every camera in a distributed network of cameras (DUDA-Rid) could be one way to protect privacy when testing the Re-ID system. Next, every camera ranks all of its previously stored gallery photos based on a similarity metric. Finally, only the closest matches from each camera are sent back to the server for approval, in order to protect the privacy of those who are not involved.

Furthermore, to use Re-ID systems, governments, and organizations must balance safety

needs against ethical concerns like equality, transparency, and access (Section 1.1.3). Without clear communication and strict management of data access and usage, there will always be a significant chance of misuse and privacy violations. Using specialized government teams to establish stringent access controls that guarantee only authorized personnel can access raw images or model updates, as well as implementing anomaly detection systems to identify unusual spikes in data transmission, are possible solutions to detect and prevent privacy violations during Re-ID inference.

Appendix A

Source-Guided Similarity Preservation for Online Person Re-Identification Supplementary Materials

This supplementary material contains additional results, analysis, and details about the S2P framework. The following items are included:

- We provide more details about the implementation of MMT, SpCL and IDM in the OUDA setting (Sec. A.1).
- We present the results of two dataset configurations that are not discussed in Chapter 4, namely, Market→CUHK and RandPerson→CUHK (Sec. A.2).
- We conduct additional ablation studies. First, we show the impact of increasing the number of tasks in the OUDA setting on the performance of S2P. Second, we validate the choice of hyperparameters and the model used for inference. And finally, we compare the performance of S2P and other UDA methods in the source domain while adapting to the target domain (Sec. A.3). Note that, considering the consistent performances of MMT and SpCL across datasets (refer to Tab. 4.1), we conduct our additional ablation studies with those two methods to derive conclusions that more likely hold true across multiple use cases.

A.1 Additional Implementation Details

In this section, we provide additional details about the implementation of the different frameworks. In Tab. 4.1, we compare the performance of S2P with three state-of-the-art UDA methods, namely MMT, SpCL and IDM. We follow [229] to implement and adapt those methods to the OUDA setting. In [229], the authors run experiments of the *strong baseline* in the OUDA setting while varying the number of epochs. They observed that the best performance is achieved using 20 epochs. We keep the same hyper-parameters for MMT, S2P-MMT, SpCL, S2P-SpCL, IDM and S2P-IDM in Tab. 4.1.

In contrast to [229, 68], we do not consider the DukeMTMC-ReID dataset, because it has been retracted by its original authors in response to a report[1] that shows that the Duke dataset was used by a few companies for some research projects that violate Human Rights.

A.2 Additional Comparison with the State-of-the-art

In this section, we provide results of complementary configurations that are not reported in Chapter 4: RP → C, and M→C. In Tab. A.1, we report the mAP and the Rank-1 of the *strong baseline*, SpCL, MMT, and our framework S2P-MMT and S2P-SpCL. We can see that S2P has the best performances on both configurations. For example, the mAP of SpCL goes from 13.2 to 27.8, and from 12.4 to 31.2, in both configurations, respectively, when integrated into our S2P framework.

Method	**RP → C**		**M → C**	
	mAP	**Rank-1**	**mAP**	**Rank-1**
Strong Baseline [66]	$2.5_{\pm 0.1}$	$1.6_{\pm 0.1}$	$6.9_{\pm 1.4}$	$6.1_{\pm 1.2}$
MMT [68]	$21.0_{\pm 0.4}$	$21.5_{\pm 0.4}$	$\underline{32.9}_{\pm 0.3}$	$\underline{32.9}_{\pm 0.4}$
SpCL [70]	$13.2_{\pm 2.2}$	$12.3_{\pm 2.4}$	$12.4_{\pm 0.5}$	$11.9_{\pm 1.1}$
S2P-MMT (ours)	$\underline{23.8}_{\pm 1.9}$	$\underline{23.8}_{\pm 2.1}$	$\mathbf{34.8}_{\pm 2.1}$	$\mathbf{35.9}_{\pm 0.3}$
S2P-SpCL (ours)	$\mathbf{27.8}_{\pm 0.3}$	$\mathbf{28.1}_{\pm 0.4}$	$31.2_{\pm 0.2}$	$31.5_{\pm 0.1}$

Table A.1: Performance of S2P and three state-of-the-art methods in two additional OUDA-Rid tasks.

A.3 Additional Ablation Studies

Increasing the number of tasks. We conduct an additional experiment using MSMT as the target dataset to evaluate the impact of increasing the number of tasks. Since MSMT is a much larger dataset than Market and CUHK, increasing the number of tasks to ten results in smaller data partitions, but the task partitions remain comparable to those of Market or CUHK in a five-task OUDA setting in terms of the number of images. This experiment extends the results of Chapter 4 by augmenting the number of tasks while keeping the same number of images per task. S2P outperforms the UDA state-of-the-art in terms of mAP for MMT and SpCL as shown in Fig. A.1. These results demonstrate the effectiveness of S2P even when additional tasks are introduced, highlighting the efficacy of our technique in increasingly complex scenarios.

Weights of the losses. Here, we validate the weights of the different losses in S2P. When implementing $\mathcal{L}_{ReID}$, we employ the weighting parameters provided in the respective origi-

[1] https://exposing.ai/duke_mtmc

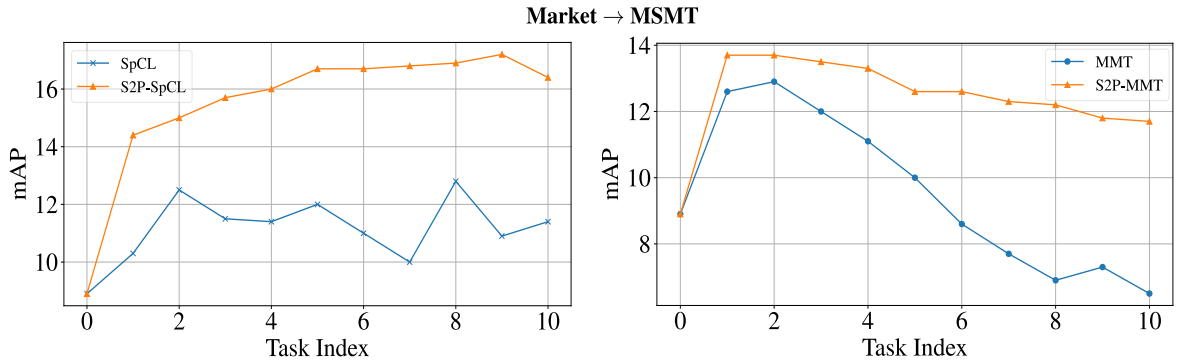


Figure A.1: Comparison of S2P with other state-of-the-art methods in terms of mAP vs. task index in a 10-tasks OUDA Market→MSMT configuration

nal papers [70, 68, 115] in S2P-SpCL, S2P-MMT and S2P-IDM. For the sake of simplicity, the weight of $\mathcal{L}_{ReID}$ is then set to 1 in all our experiments regardless of the chosen pseudo-labeling method. For $\mathcal{L}_{KD}$ and $\mathcal{L}_{MMD}$, we show in Tab. A.2 a comparison in mAP of S2P-SpCL when varying the corresponding weights λ_{KD} and λ_{MMD}. We observe that the best performances are obtained with $\lambda_{KD} = 1$ and $\lambda_{MMD} = 0.03$. We also notice that S2P is not very sensitive to the weights of the losses and remains, in all cases, above the performance of the original SpCL (15.6 in Tab. 4.1).

λ_{MMD} \ λ_{KD}	0.1	1	10
0.003	31.7	33.3	31
0.03	31.1	**34.3**	31.4
0.3	28.4	32.2	29.7

Table A.2: Ablation study on the weights of the two main losses of S2P λ_{KD} and λ_{MMD}. The table shows the mAP of S2P-SpCL in the MSMT→CUHK configuration. The best performing configuration is shown in **bold**.

Performance of the teacher. In what follows, we justify the choice of the teacher model for inference. Tab. A.3 shows the performance of S2P-MMT and S2P-SpCL using the student and teacher networks in inference. In the OUDA setting, we show that the teacher model in S2P guides the training of the student model. Furthermore, the teacher in turn benefits from the accuracy of the student models by leveraging the previously acquired knowledge, hence giving more accurate predictions. Tab. A.3 shows that in all the aforementioned configurations, we get better results when deploying the teacher model for inference, rather than the student model.

Performance on the source domain. In Fig. A.2, we show the evolution of the mAP of both MMT and S2P-MMT on the source domain when considering two different configurations: a) Market→MSMT and b) Market→CUHK. We observe that the performance on the

Method	MS → M		MS → C		M → MS		M → C	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
S2P-MMT (student)	63.1	82	28.3	28.5	14.1	35.2	26.2	26.3
S2P-MMT (teacher)	**70**	**87.1**	**40.4**	**42.4**	**19.5**	**43.3**	**34.8**	**35.9**
S2P-SpCL (student)	61.9	81.9	30.7	31.9	17.5	41.5	21.4	21.6
S2P-SpCL (teacher)	**69.1**	**87.1**	**34.3**	**35.1**	**20.2**	**46.1**	**31.2**	**31.5**

Table A.3: Ablation study on the choice of the inference model in the S2P framework. We compare the performance of S2P in the last task in four real-to-real OUDA-Rid tasks when using the student and the teacher models at inference time. The best performing method on each dataset is shown in **bold**.

source domain is improved during the first task of OUDA. After the first task, the performance of MMT on the source domain drops in both configurations, showing that the model focuses more on capturing the distribution of the target domain, hence overfitting the upcoming tasks and forgetting the previously acquired knowledge on the source domain. On the contrary, the performance of S2P-MMT on the source domain remains relatively high and stable after the first task, showing that our S2P framework effectively maintains a common feature space for the source and target domains.

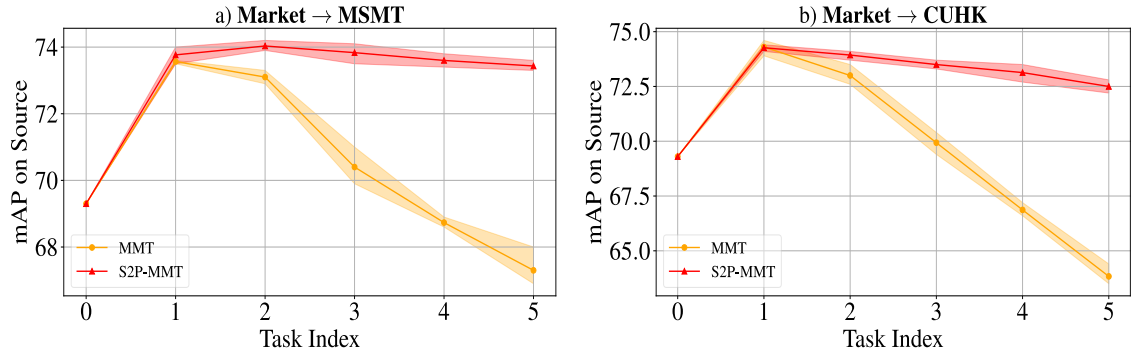


Figure A.2: Performance of MMT and S2P-MMT on the source domain in two OUDA tasks: a) Market→MSMT and b) Market→CUHK.

Appendix B

Privacy-Preserving Adaptive Re-Identification without Image Transfer Supplementary Materials

In this supplementary material, we provide results and analysis of additional experiments and present additional details about the Fed-Protoid.

- We provide the pseudo-code of Fed-Protoid to give more details about its algorithmic structure.
- We present a set of experiments focused on the source prototypes, initially showing the significance of utilizing the global model for their computation instead of the pseudo client. Subsequently, we highlight the impact of reducing the number of source prototypes in the performance of Fed-Protoid.
- We include a detailed analysis regarding the sensibility of the hyper-parameters of Fed-Protoid.
- We compare the distributed MMD with the original MMD and some Domain Generalization (DG) Re-ID methods.

B.1 Fed-Protoid: Algorithm

For completeness, we detail the Fed-Protoid algorithm as follows:

Algorithm 1 Fed-Protoid algorithm

Input: n unlabeled datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$, an annotated dataset $\mathcal{S}$, and the source pre-trained weights $\boldsymbol{\theta}_s$
Initialize model $F_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}_s$
for each training round **do**
    Transmission Stage:
    Transmit $F_{\boldsymbol{\theta}}$ and source prototypes to all clients
    Local Training Stage:
    Update pseudo-client model $F_{\hat{\boldsymbol{\theta}}_s}$ using $\mathcal{S}$ and $\mathcal{L}_s$
    for each client i **do**
        Update client model $F_{\hat{\boldsymbol{\theta}}_i}$ using local dataset $\mathcal{D}_i$ and $\mathcal{L}_i$
    end for
    Aggregation Stage:
    Aggregate models using equation $\boldsymbol{\theta} = \alpha\hat{\boldsymbol{\theta}}_s + (1-\alpha)\sum_{i=1}^{n} w_i\hat{\boldsymbol{\theta}}_i$
end for
Output: Trained federated model $F_{\boldsymbol{\theta}}$

B.2 Additional experiments on the source prototypes: computation and communication

B.2.1 Impact of Global model in source prototype computation

In this section, we present experimental results for both Fed-Protoid and Fed-Protoid++, showing the advantages of utilizing the global model for computing source prototypes. The results shown in Tab. B.1 indicate that across the two configurations MS $\rightarrow$ M and MS $\rightarrow$ C, we obtain superior performance when the prototypes are derived from the global model instead of the pseudo-client model. The effectiveness of using the global model in prototype computation can be attributed to its ability to bridge the gap between the source and target domain distributions. Essentially, the prototypes generated by the global model represent a median distribution that lies between those of the source and target domains. This intermediary positioning facilitates more efficient optimization of the Maximum Mean Discrepancy (MMD). Instead of directly aligning the target domain with the source domain, the global model provides features that are equidistant to both domains. Consequently, this approach converges all distributions towards a central, unified distribution, rather than skewing them towards the source domain distribution alone.

B.2.2 The impact of the number of source prototypes

The following experiments aim to assess the effectiveness of Fed-Protoid in scenarios with stronger memory and communication limitations. Such situations occur when a large source domain with many identities is deployed in the server, resulting in an increased number

Method	Source Prototypes computed with	MS → M	MS → C
Fed-Protoid	Pseudo-client	43.1	23.6
	Global model	**51.0**	**23.8**
Fed-Protoid++	Pseudo-client	60.5	43.7
	Global model	**61.7**	**43.8**

Table B.1: Ablation study of the choice of the model that computes the source prototypes.

of prototypes, thus requiring higher communication bandwidth. For instance, the synthetic RP dataset contains $8,000$ identities which is 8 times the number of identities in the MS dataset. To address this challenging scenario, we propose investigating whether a simple uniform sub-sampling of prototypes reduces the transmission cost without impacting the ReID performance. We evaluate Fed-Protoid with a reduced number of source prototypes in both configurations: RP→M and RP→C. Fig. B.1 shows that Fed-Protoid remains effective despite a significant decrease in the number of source prototypes for the MMD optimization on edge devices. Fed-Protoid shows a stable performance when the number of source prototypes varies between 2% and 100% in RP→M and between 20% and 10% in RP→C configuration. Therefore, we argue that Fed-Protoid is robust against device storage and communication limitations.

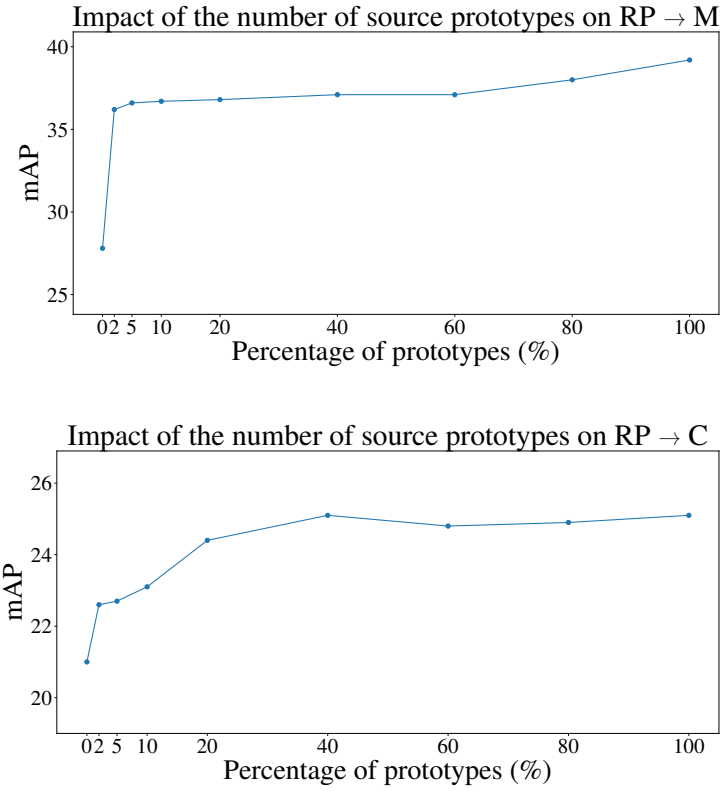


Figure B.1: The impact of the number of source prototypes in the Fed-Protoid performance in two configurations: RP→M and RP→C

B.3 Variability of Fed-Protoid and hyper-parameters

B.3.1 Variability of the performance of Fed-Protoid across different initialization

We conduct multiple experiments of our Fed-Protoid and Fed-Protoid++ with three different seeds in all the configurations presented in Tab. 5.1. We report in Tab. B.2 the mean of the mAP and Rank-1 across those runs. Alongside these metrics, we also report the standard deviation to illustrate the variability in the results. We can state that we have consistency and minimal variance across all the different configurations, demonstrating the robustness and reliability of our method under different initializations.

Configuration	Fed-Protoid		Fed-Protoid++	
	mAP	Rank-1	mAP	Rank-1
MS $\rightarrow$ M	$51.0_{\pm 0.3}$	$76.8_{\pm 0.2}$	$61.7_{\pm 0.2}$	$82.6_{\pm 0.1}$
MS $\rightarrow$ C	$23.8_{\pm 0.2}$	$23.1_{\pm 0.1}$	$43.8_{\pm 0.2}$	$42.4_{\pm 0.4}$
RP $\rightarrow$ M	$39.2_{\pm 0.1}$	$66.4_{\pm 0.0}$	$45.2_{\pm 0.1}$	$71.8_{\pm 0.1}$
RP $\rightarrow$ C	$25.1_{\pm 0.4}$	$24.7_{\pm 0.3}$	$25.7_{\pm 0.2}$	$24.9_{\pm 0.1}$

Table B.2: Standard deviation of both Fed-Protoid and Fed-Protoid++ with varying seeds.

B.3.2 Hyper-parameters ablation study

Fig. B.2 illustrates the results of the sensitivity analysis conducted on the hyper-parameters of Fed-Protoid. In this analysis, all hyper-parameters except the specific one under investigation are maintained at their default values. We observe that changing the hyper-parameters β_1 and γ_1 results in a slight impact on the accuracy with only minimal variances, showing that our method is stable and robust. As for λ, which is the hyper-parameter controlling the importance of the MMD loss in the final objective, we determined that a value of 0.1 yields optimal results and have therefore set it to this fixed value for subsequent experiments.

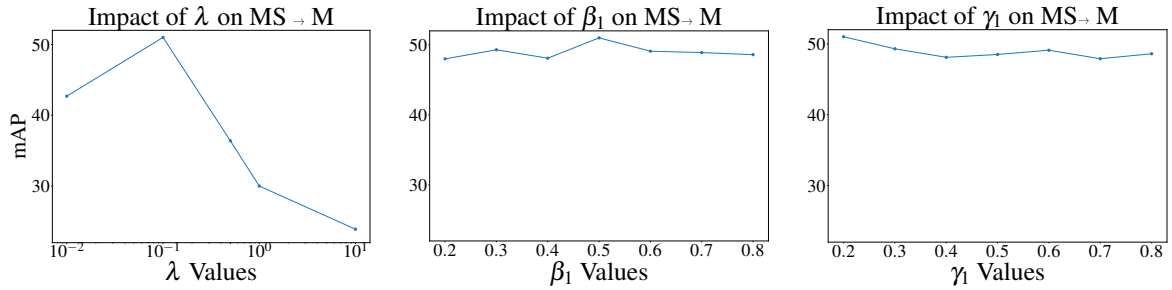


Figure B.2: Ablation study on the sensibility of the different hyper-parameters of Fed-Protoid.

B.4 Distributed MMD vs. Original MMD

Tab. B.3 compares the distributed MMD with the original MMD. For the original MMD, we use the same DUDA-Rid setting, but the MMD loss is computed over the entire target dataset. The results show that the distributed MMD outperforms the original MMD in both configurations. Note that the original MMD violates DUDA-Rid privacy constraints, making it unsuitable for our problem.

Table B.3: Comparison between original and distributed MMD.

Method	**MS → M**		**RP → M**	
	mAP	**Rank-1**	**mAP**	**Rank-1**
Fed-Protoid + orig. MMD	47.1	75.3	30.2	59.2
Fed-Protoid + dist. MMD (ours)	**51.0**	**76.8**	**39.2**	**66.4**

B.5 Comparison with DG and additional experiments.

Tab. B.4 includes additional results from two SOTA methods in DG Re-ID: TransMatcher [230] and PAT [231]. We compare these methods with Fed-Protoid (ViT) presented in Tab. 5.4. In all configurations, Fed-Protoid (ViT) outperforms the other methods. These results are further improved using Fed-Protoid++, which incorporates the LUP large-scale dataset during pre-training instead of being initialized by ImageNet.

Table B.4: Comparison between Fed-Protoid and DG methods.

Method	**Type**	**MS → M**		**MS → C**	
		mAP	**Rank-1**	**mAP**	**Rank-1**
TransMatcher [230]	DG	52.0	80.1	22.5	23.7
PAT [231]	DG	47.3	72.2	25.1	24.2
Fed-Protoid (ViT) (ours)	UDA	52.4	80.6	27.5	26.6
Fed-Protoid++ (ours)	UDA	**61.7**	**82.6**	**43.8**	**42.4**

References

[1] Yawei Luo et al. “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation”. In: *CVPR*. 2019.

[2] Robert Morgus and Justin Sherman. “How U.S. Surveillance Technology is Propping Up Authoritarian Regimes”. In: *The Washington Post*. 2019.

[3] Lowy Institute. *Digital Authoritarianism, China and COVID*. 2023. URL: https://www.lowyinstitute.org/.

[4] Nazia Perwaiz, Muhammad Fraz, and Muhammad Shahzad. “Person Re-Identification Using Hybrid Representation Reinforced by Metric Learning”. In: vol. PP. 2018.

[5] Timothy Huang and Stuart Russell. “Object identification in a bayesian context”. In: *IJCAI*. 1997.

[6] Wojciech Zajdel, Zoran Zivkovic, and Ben JA Krose. “Keeping track of humans: Have I seen this person before?” In: *ICRA*. 2005.

[7] Y. LeCun et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation*. 1989.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *NIPS*. 2012.

[9] Dong Yi et al. “Deep metric learning for person re-identification”. In: *ICPR*. 2014.

[10] Yi-Hsuan Tsai et al. “Domain Adaptation for Structured Output via Discriminative Patch Representations”. In: *ICCV*. 2019.

[11] Xingchao Peng et al. “Moment Matching for Multi-Source Domain Adaptation”. In: *ICCV*. 2019.

[12] Baochen Sun, Jiashi Feng, and Kate Saenko. “Return of frustratingly easy domain adaptation”. In: *AAAI*. 2016.

[13] Guoliang Kang et al. “Contrastive Adaptation Network for Unsupervised Domain Adaptation”. In: *CVPR*. 2019.

[14] Hehe Fan et al. “Unsupervised Person Re-Identification: Clustering and Fine-Tuning”. In: *ACM (TOMM)*. 2018.

[15] Fengxiang Yang et al. “Asymmetric Co-Teaching for Unsupervised Cross Domain Person Re-Identification”. In: *AAAI*. 2020.

[16] Yaroslav Ganin and Victor Lempitsky. “Unsupervised domain adaptation by back-propagation”. In: *ICML*. 2015.

[17] Longhui Wei et al. “Person transfer gan to bridge domain gap for person re-identification”. In: *CVPR*. 2018.

[18] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. “Person reidentification using spatiotemporal appearance”. In: *CVPR*. 2006.

[19] Wei Li et al. “Deepreid: Deep filter pairing neural network for person re-identification”. In: *CVPR*. 2014.

[20] Douglas Gray and Hai Tao. “Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features”. In: *ECCV*. 2008.

[21] Michela Farenzena et al. “Person re-identification by symmetry-driven accumulation of local features”. In: *CVPR*. 2010.

[22] Sławomir Bąk et al. “Learning to Match Appearances by Correlations in a Covariance Metric Space”. In: *ECCV*. 2012.

[23] Abir Das, Anirban Chakraborty, and Amit K. Roy-Chowdhury. “Consistent Re-identification in a Camera Network”. In: *ECCV*. 2014.

[24] Shengcai Liao et al. “Person re-identification by Local Maximal Occurrence representation and metric learning”. In: *CVPR*. 2015.

[25] Faqiang Wang et al. “Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification”. In: *CVPR*. 2016.

[26] Xuelin Qian et al. “Multi-scale deep learning architectures for person re-identification”. In: *CVPR*. 2017.

[27] Yan Wang et al. “Resource aware person re-identification across multiple resolutions”. In: *CVPR*. 2018.

[28] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CVPR*. 2016.

[29] Kaiyang Zhou et al. “Omni-scale feature learning for person re-identification”. In: *ICCV*. 2019.

[30] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861*. 2017.

[31] Yicheng Wang et al. “Person Re-identification with Cascaded Pairwise Convolutions”. In: *CVPR*. 2018.

[32] Ruibing Hou et al. “Interaction-and-aggregation network for person re-identification”. In: *CVPR*. 2019.

[33] Xihui Liu et al. “HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis”. In: *ICCV*. 2017.

[34] Wei Li, Xiatian Zhu, and Shaogang Gong. “Harmonious attention network for person re-identification”. In: *CVPR*. 2018.

[35] Xin Ning et al. “Feature Refinement and Filter Network for Person Re-Identification”. In: *IEEE Trans. Circuits Syst. Video Technol.* 2021.

[36] Ruyu Xu et al. “Person re-identification based on improved attention mechanism and global pooling method”. In: *J. Vis. Commun. Image Represent.* 2023.

[37] Guanshuo Wang et al. “Learning discriminative features with multiple granularities for person re-identification”. In: *ACM*. 2018.

[38] Feng Zheng et al. “Pyramidal person re-identification via multi-loss dynamic training”. In: *CVPR*. 2019.

[39] Shuting He et al. “TransReID: Transformer-Based Object Re-Identification”. In: *ICCV*. 2021.

[40] Guiwei Zhang et al. “PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification”. In: *CVPR*. 2023.

[41] Ji Zhang et al. “Learning global and local features using graph neural networks for person re-identification”. In: *Signal Processing: Image Communication*. 2022.

[42] M. Farenzena et al. “Person re-identification by symmetry-driven accumulation of local features”. In: *CVPR*. 2010.

[43] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. “Unsupervised Salience Learning for Person Re-identification”. In: *CVPR*. 2013.

[44] Mang Ye et al. “Dynamic label graph matching for unsupervised video re-identification”. In: *ICCV*. 2017.

[45] Martin Ester et al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *ACM (KDD)*. 1996.

[46] Hehe Fan et al. “Unsupervised person re-identification: Clustering and fine-tuning”. In: *ACM (TOMM)*. 2018.

[47] Fang Zhao et al. “Unsupervised Domain Adaptation with Noise Resistible Mutual-Training for Person Re-identification”. In: *ECCV*. 2020.

[48] Fengxiang Yang et al. “Asymmetric co-teaching for unsupervised cross-domain person re-identification”. In: *AAAI*. 2020.

[49] Yutian Lin et al. “A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification”. In: *AAAI*. 2019.

[50] Kaiwei Zeng et al. “Hierarchical clustering with hard-batch triplet loss for person re-identification”. In: *CVPR*. 2020.

[51] Xiao Zhang et al. “Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification”. In: *CVPR*. 2021.

[52] Yoonki Cho et al. “Part-based pseudo label refinement for unsupervised person re-identification”. In: *CVPR*. 2022.

[53] Hao Luo et al. “Self-Supervised Pre-Training for Transformer-Based Person Re-Identification”. In: *arXiv preprint arXiv:2111.12084*. 2021.

[54] Dengpan Fu et al. “Unsupervised Pre-training for Person Re-identification”. In: *CVPR*. 2021.

[55] Zuozhuo Dai et al. “Cluster contrast for unsupervised person re-identification”. In: *ACCV*. 2022.

[56] Wangmeng Xiang et al. “Second-Order Camera-Aware Color Transformation for Cross-Domain Person Re-identification”. In: *ACCV*. 2021.

[57] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. “Generalizing Person Re-Identification by Camera-Aware Invariance Learning and Cross-Domain Mixup”. In: *ECCV*. 2020.

[58] Fengxiang Yang et al. “Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification”. In: *CVPR*. 2021.

[59] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. “Person re-identification by probabilistic relative distance comparison”. In: *CVPR*. 2011.

[60] Shengcai Liao and Stan Z. Li. “Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification”. In: *ICCV*. 2015.

[61] Liang Zheng, Yi Yang, and Alexander G Hauptmann. “Person re-identification: Past, present and future”. In: *arXiv preprint arXiv:1610.02984*. 2016.

[62] Nicolai Wojke and Alex Bewley. “Deep cosine metric learning for person re-identification”. In: *WACV*. 2018.

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *CVPR*. 2015.

[64] Alexander Hermans, Lucas Beyer, and Bastian Leibe. “In defense of the triplet loss for person re-identification”. In: *arXiv preprint arXiv:1703.07737*. 2017.

[65] Hao Luo et al. “Bag of tricks and a strong baseline for deep person re-identification”. In: *CVPR Workshops*. 2019.

[66] Hehe Fan et al. “Unsupervised person re-identification: Clustering and fine-tuning”. In: *ACM TOMM*. 2018.

[67] Yang Fu et al. “Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification”. In: *ICCV*. 2019.

[68] Yixiao Ge, Dapeng Chen, and Hongsheng Li. “Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification”. In: *ICLR*. 2020.

[69] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. "Gated siamese convolutional neural network architecture for human re-identification". In: *ECCV*. 2016.

[70] Yixiao Ge et al. "Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID". In: *NIPS*. 2020.

[71] Chen Change Loy, Tao Xiang, and Shaogang Gong. "Multi-camera activity correlation analysis". In: *CVPR*. 2009.

[72] Douglas Gray and Hai Tao. *Viewpoint Invariant Pedestrian Recognition (VIPeR) Dataset v1.0*. 2014.

[73] Wei Li, Rui Zhao, and Xiaogang Wang. "Human Reidentification with Transferred Metric Learning". In: *ACCV*. 2012.

[74] Chen Change Loy, Tao Xiang, and Shaogang Gong. "Custom pictorial structures for re-identification". In: *BMVC*. 2011.

[75] Martin Hirzer et al. "Person Re-identification by Descriptive and Discriminative Classification". In: *SCIA*. 2011.

[76] Liang Zheng et al. "Scalable Person Re-identification: A Benchmark". In: *(ICCV)*. 2015.

[77] Ergys Ristani et al. "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking". In: *ECCV*. 2016.

[78] Longhui Wei et al. "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification". In: *CVPR*. 2018.

[79] Pedro F. Felzenszwalb et al. "Object Detection with Discriminatively Trained Part-Based Models". In: *IEEE TPAMI*. 2010.

[80] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *NIPS*. 2015.

[81] Ultralytics. *YOLOv5: A state-of-the-art real-time object detection system*. `https://docs.ultralytics.com`. 2021.

[82] Xiaoxiao Sun and Liang Zheng. "Dissecting person re-identification from the viewpoint of viewpoint". In: *CVPR*. 2019.

[83] Yanan Wang, Shengcai Liao, and Ling Shao. "Surpassing real-world source training data: Random 3D characters for generalizable person re-identification". In: *ACM MM*. 2020.

[84] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *CVPR*. 2009.

[85] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. "3DPes: 3D People Dataset for Surveillance and Forensics". In: *ACM Workshops*. 2011.

[86] Wei Li and Xiaogang Wang. "Locally Aligned Feature Transforms across Views". In: *CVPR*. 2013.

[87] Shai Ben-David et al. “A theory of learning from different domains”. In: *Machine Learning*. 2010.

[88] Eric Tzeng et al. “Deep domain confusion: Maximizing for domain invariance”. In: *arXiv preprint arXiv:1412.3474*. 2014.

[89] Eric Tzeng et al. “Simultaneous deep transfer across domains and tasks”. In: *ICCV*. 2015.

[90] Baochen Sun and Kate Saenko. “Deep coral: Correlation alignment for deep domain adaptation”. In: *ECCV Workshops*. 2016.

[91] B. Fuglede and F. Topsoe. “Jensen-Shannon divergence and Hilbert space embedding”. In: *ISIT*. 2004.

[92] Junguang Jiang et al. “Resource Efficient Domain Adaptation”. In: ACM, 2020.

[93] S. S. Vallender. “Calculation of the Wasserstein Distance Between Probability Distributions on the Line”. In: *Theory of Probability & Its Applications*. 1974.

[94] Chen-Yu Lee et al. “Sliced wasserstein discrepancy for unsupervised domain adaptation”. In: *CVPR*. 2019.

[95] Ian Goodfellow et al. “Generative adversarial nets”. In: *NIPS*. 2014.

[96] Yaroslav Ganin et al. “Domain-adversarial training of neural networks”. In: *JMLR*. 2016.

[97] Eric Tzeng et al. “Adversarial discriminative domain adaptation”. In: *CVPR*. 2017.

[98] Mingsheng Long et al. “Deep transfer learning with joint adaptation networks”. In: *ICML*. 2017.

[99] Han-Kai Hsu et al. “Progressive domain adaptation for object detection”. In: *WACV*. 2020.

[100] Weichen Zhang et al. “Collaborative and Adversarial Network for Unsupervised Domain Adaptation”. In: *CVPR*. 2018.

[101] Yang Zou et al. “Confidence regularized self-training”. In: *ICCV*. 2019.

[102] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. “Asymmetric tri-training for unsupervised domain adaptation”. In: *ICML*. 2017.

[103] Shaoan Xie et al. “Learning Semantic Representations for Unsupervised Domain Adaptation”. In: *ICML*. 2018.

[104] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. “Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation”. In: *CVPR*. 2023.

[105] Jinlin Wu et al. “Clustering and Dynamic Sampling Based Unsupervised Domain Adaptation for Person Re-Identification”. In: *ICME*. 2019.

[106] Guodong Ding et al. “Towards better validity: Dispersion based clustering for unsupervised person re-identification”. In: *arXiv preprint arXiv:1906.01308*. 2019.

[107] Yunpeng Zhai et al. “Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification”. In: *CVPR*. 2020.

[108] Minying Zhang et al. “Unsupervised Domain Adaptation for Person Re-identification via Heterogeneous Graph Alignment”. In: *AAAI*. 2021.

[109] Kecheng Zheng et al. “Group-aware label transfer for domain adaptive person re-identification”. In: *CVPR*. 2021.

[110] Gong Boqing et al. “Geodesic flow kernel for unsupervised domain adaptation”. In: *CVPR*. 2012.

[111] Gopalan Raghuraman, Li Ruonan, and Chellappa Rama. “Unsupervised adaptation across domain shifts by generating intermediate data representations”. In: *IEEE TPAMI*. 2013.

[112] Zhen Cui et al. “Flowing on riemannian manifold: Domain adaptation by shifting covariance”. In: *IEEE Trans. Cybern.* 2014.

[113] Gong Rui et al. “Dlow: Domain flow for adaptation and generalization.” In: *CVPR*. 2019.

[114] Shuhao Cui et al. “Gradually vanishing bridge for adversarial domain adaptation”. In: *CVPR*. 2020.

[115] Yongxing Dai et al. “IDM: An Intermediate Domain Module for Domain Adaptive Person Re-ID”. In: *ICCV*. 2021.

[116] James Kirkpatrick et al. “Overcoming catastrophic forgetting in neural networks”. In: *PNAS*. 2017.

[117] Guoliang Lin, Hanlu Chu, and Hanjiang Lai. “Towards better plasticity-stability trade-off in incremental learning: A simple linear connector”. In: *CVPR*. 2022.

[118] Sanghwan Kim et al. “Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning”. In: *CVPR*. 2023.

[119] Aaron Van Den Oord, Oriol Vinyals, et al. “Neural discrete representation learning”. In: *NIPS*. 2017.

[120] Yen-Chang Hsu et al. “Re-evaluating continual learning scenarios: A categorization and case for strong baselines”. In: *arXiv preprint arXiv:1810.12488*. 2018.

[121] Da-Wei Zhou et al. “Deep class-incremental learning: A survey”. In: *arXiv preprint arXiv:2302.03648*. 2023.

[122] Hippolyt Ritter, Aleksandar Botev, and David Barber. “Online structured laplace approximations for overcoming catastrophic forgetting”. In: *NIPS*. 2018.

[123] Friedemann Zenke, Ben Poole, and Surya Ganguli. “Continual learning through synaptic intelligence”. In: *ICML*. 2017.

[124] Rahaf Aljundi et al. “Memory aware synapses: Learning what (not) to forget”. In: *ECCV*. 2018.

[125] Arslan Chaudhry et al. “Riemannian walk for incremental learning: Understanding forgetting and intransigence”. In: *ECCV*. 2018.

[126] Inyoung Paik et al. “Overcoming catastrophic forgetting by neuron-level plasticity control”. In: *AAAI*. 2020.

[127] Sangwon Jung et al. “Continual learning with node-importance based adaptive group sparse regularization”. In: *NIPS*. 2020.

[128] Jianping Gou et al. “Knowledge distillation: A survey”. In: *IJCV*. 2021.

[129] Zhizhong Li and Derek Hoiem. “Learning without Forgetting”. In: *ECCV*. 2016.

[130] Prithviraj Dhar et al. “Learning without memorizing”. In: *CVPR*. 2019.

[131] Arslan Chaudhry et al. “On tiny episodic memories in continual learning”. In: *arXiv preprint arXiv:1902.10486*. 2019.

[132] David Lopez-Paz and Marc’Aurelio Ranzato. “Gradient episodic memory for continual learning”. In: *NIPS*. 2017.

[133] Sylvestre-Alvise Rebuffi et al. “icarl: Incremental classifier and representation learning”. In: *CVPR*. 2017.

[134] Xisen Jin et al. “Gradient-based editing of memory examples for online task-free continual learning”. In: *arXiv preprint arXiv:2006.15294*. 2020.

[135] Yiduo Guo, Bing Liu, and Dongyan Zhao. “Online Continual Learning through Mutual Information Maximization”. In: *ICML*. 2022.

[136] Qing Sun et al. “Exploring example influence in continual learning”. In: *NIPS*. 2022.

[137] Zhicheng Sun, Yadong Mu, and Gang Hua. “Regularizing Second-Order Influences for Continual Learning”. In: *CVPR*. 2023.

[138] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. “Learning to imagine: Diversify memory for incremental learning using unlabeled data”. In: *CVPR*. 2022.

[139] Shengbang Tong et al. “Incremental learning of structured memory via closed-loop transcription”. In: *arXiv preprint arXiv:2202.05411*. 2022.

[140] Sergi Masip et al. “Continual Learning of Diffusion Models with Generative Distillation”. In: *arXiv preprint arXiv:2311.14028*. 2023.

[141] Johannes Von Oswald et al. “Continual learning with hypernetworks”. In: *arXiv preprint arXiv:1906.00695*. 2019.

[142] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. “Piggyback: Adapting a single network to multiple tasks by learning to mask weights”. In: *ECCV*. 2018.

[143] Arun Mallya and Svetlana Lazebnik. “Packnet: Adding multiple tasks to a single network by iterative pruning”. In: *CVPR*. 2018.

[144] Ching-Yi Hung et al. “Compacting, picking and growing for unforgetting continual learning”. In: *NIPS*. 2019.

[145] Xilai Li et al. “Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting”. In: *ICML*. 2019.

[146] Kai Zhu et al. “Self-sustaining representation expansion for non-exemplar class-incremental learning”. In: *CVPR*. 2022.

[147] Ju Xu et al. “Adaptive Progressive Continual Learning”. In: *IEEE TPAMI*. 2021.

[148] Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *AISTATS*. 2017.

[149] Jiehan Zhou et al. “A survey on federated learning and its applications for accelerating industrial internet of things”. In: *arXiv preprint arXiv:2104.10501*. 2021.

[150] Quande Liu et al. “Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space”. In: *CVPR*. 2021.

[151] Tian Li et al. “Federated optimization in heterogeneous networks”. In: *MLSys*. 2020.

[152] Weiming Zhuang et al. “Performance optimization of federated person re-identification via benchmark analysis”. In: *ACM MM*. 2020.

[153] Rui Ye et al. “Feddisco: Federated learning with discrepancy-aware collaboration”. In: *ICMLg*. 2023.

[154] Zuoqi Tang et al. “Optimizing Federated Learning on Non-IID Data Using Local Shapley Value”. In: *Artificial Intelligence*. 2021.

[155] Tian Li et al. “Federated optimization in heterogeneous networks”. In: *Proc. Mach. Learn. Res.* 2020.

[156] Cong Xie, Sanmi Koyejo, and Indranil Gupta. “Asynchronous Federated Optimization”. In: *arXiv e-prints*. 2019.

[157] Yue Zhao et al. “Federated Learning with Non-IID Data”. In: *CoRR*. 2018.

[158] Fei Chen et al. “Federated Meta-Learning with Fast Convergence and Efficient Communication”. In: *arXiv: Learning*. 2018.

[159] Weiming Zhuang et al. “Performance optimization of federated person re-identification via benchmark analysis”. In: *ACM MM*. 2020.

[160] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. “Joint optimization in edge-cloud continuum for federated unsupervised person re-identification”. In: *ACM MM*. 2021.

[161] Jiabei Liu et al. “Optimizing Federated Unsupervised Person Re-identification via Camera-aware Clustering”. In: *MMSP*. 2022.

[162] Menglin Wang et al. “Camera-aware proxies for unsupervised person re-identification”. In: *AAAI*. 2021.

[163] Jiawei Liu et al. “Adaptive Transfer Network for Cross-Domain Person Re-Identification”. In: *CVPR*. 2019.

[164] Yixiao Ge et al. “Structured Domain Adaptation with Online Relation Regularization for Unsupervised Person Re-ID”. In: *arXiv preprint arXiv:2003.06650*. 2020.

[165] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. “Instance-Guided Context Rendering for Cross-Domain Person Re-Identification”. In: *ICCV*. 2019.

[166] Jun-Yan Zhu et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: *ICCV*. 2017.

[167] Weijian Deng et al. “Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification”. In: *CVPR*. 2018.

[168] Zhun Zhong et al. “Generalizing A Person Retrieval Model Hetero- and Homogeneously”. In: *ECCV*. 2018.

[169] Liangchen Song et al. “Unsupervised Domain Adaptive Re-Identification: Theory and Practice”. In: *Pattern Recognition*. 2020.

[170] Mang Ye et al. “Dynamic Graph Co-Matching for Unsupervised Video-Based Person Re-Identification”. In: *IEEE TIP*. 2019.

[171] Hao Feng et al. “Complementary pseudo labels for unsupervised domain adaptation on person re-identification”. In: *IEEE TIP*. 2021.

[172] Guillaume Delorme et al. “CANU-ReID: a conditional adversarial network for unsupervised person re-identification”. In: *ICPR*. 2021.

[173] Prakhar Kaushik et al. “Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping”. In: *arXiv preprint arXiv:2102.11343*. 2021.

[174] Steven C. Y. Hung et al. “Compacting, Picking and Growing for Unforgetting Continual Learning”. In: *NIPS*. 2019.

[175] Cheng-Hao Tu, Cheng-En Wu, and Chu-Song Chen. “Extending Conditional Convolution Structures For Enhancing Multitasking Continual Learning”. In: *APSIPA ASC*. 2020.

[176] Wang Zhou et al. “Lifelong Object Detection”. In: *arXiv preprint arXiv:2009.01129*. 2020.

[177] Fabio Cermelli et al. “Incremental Learning in Semantic Segmentation from Image Labels”. In: *arXiv preprint arXiv:2112.01882*. 2021.

[178] Mengyao Zhai et al. “Lifelong GAN: Continual Learning for Conditional Image Generation”. In: *ICCV*. 2019.

[179] Fei Ye and Adrian G. Bors. “Lifelong Teacher-Student Network Learning”. In: *IEEE TPAMI*. 2021.

[180] Jeongtae Lee et al. “Lifelong Learning with Dynamically Expandable Networks”. In: *arXiv preprint arXiv:1708.01547*. 2017.

[181] Chenshen Wu et al. “Memory Replay GANs: learning to generate images from new categories without forgetting”. In: *NIPS*. 2018.

[182] Nan Pu et al. “Lifelong Person Re-Identification via Adaptive Knowledge Accumulation”. In: *CVPR*. 2021.

[183] Zhipeng Huang et al. “Lifelong Unsupervised Domain Adaptive Person Re-identification with Coordinated Anti-forgetting and Adaptation”. In: *arXiv preprint arXiv:2112.06632*. 2021.

[184] Enrico Fini et al. “Online continual learning under extreme memory constraints”. In: *ECCV*. 2020.

[185] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. “Universal source-free domain adaptation”. In: *CVPR*. 2020.

[186] Cristiano Saltori et al. “Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection”. In: *IEEE 3DV*. 2020.

[187] Liang Zheng et al. “Scalable Person Re-identification: A Benchmark”. In: *ICCV*. 2015.

[188] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *ICLR*. 2015.

[189] Andreea Bobu et al. “Adapting to Continuously Shifting Domains”. In: *ICLR*. 2018.

[190] Hao Chen, Benoit Lagadec, and Francois Bremond. “Unsupervised Lifelong Person Re-identification via Contrastive Rehearsal”. In: *arXiv preprint arXiv:2203.06468*. 2022.

[191] Geoffrey Hinton, Jeff Dean, and Oriol Vinyals. “Distilling the knowledge in a neural network”. In: *NIPS Workshops*. 2014.

[192] Boqing Gong, Kristen Grauman, and Fei Sha. “Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation”. In: *ICML*. 2013.

[193] Antti Tarvainen and Harri Valpola. “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”. In: *NIPS*. 2017.

[194] Samuli Laine and Timo Aila. “Temporal Ensembling for Semi-Supervised Learning”. In: *ICLR*. 2017.

[195] Arthur Gretton et al. “A kernel method for the two-sample-problem”. In: *NIPS*. 2006.

[196] Yujia Li, Kevin Swersky, and Rich Zemel. “Generative moment matching networks”. In: *ICML*. 2015.

[197] Frederick Tung and Greg Mori. “Similarity-preserving knowledge distillation”. In: *ICCV*. 2019.

[198] Sergey Zagoruyko and Nikos Komodakis. “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer”. In: *ICLR*. 2016.

[199] Zhun Zhong et al. “Re-ranking person re-identification with k-reciprocal encoding”. In: *CVPR*. 2017.

[200] European Parliament and Council of the European Union. *General Data Protection Regulation (GDPR)*. Chapter 2, inproceedings 5.c. 2016.

[201] European Parliament and Council of the European Union. *General Data Protection Regulation (GDPR)*. Chapter 2, inproceedings 5.b. 2016.

[202] European Parliament and Council of the European Union. *General Data Protection Regulation (GDPR)*. Chapter 5, inproceedings 44-49. 2016.

[203] Mohammad Rasouli, Tao Sun, and Ram Rajagopal. “Fedgan: Federated generative adversarial networks for distributed data”. In: *arXiv preprint arXiv:2006.07228*. 2020.

[204] Pranvera Kortoçi et al. “Federated Split GANs”. In: *ACM MobiComWorkshops*. 2022.

[205] Shan Lin et al. “Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification”. In: *BMVC*. 2018.

[206] Yangru Huang et al. “Domain adaptive attention model for unsupervised cross-domain person re-identification”. In: *arXiv preprint arXiv:1905.10529*. 2019.

[207] Xiaobin Liu and Shiliang Zhang. “Domain adaptive person re-identification via coupling optimization”. In: *ACM MM*. 2020.

[208] Jake Snell, Kevin Swersky, and Richard Zemel. “Prototypical networks for few-shot learning”. In: *NIPS*. 2017.

[209] Nanqing Dong and Eric P Xing. “Few-shot semantic segmentation with prototype learning.” In: *BMVC*. 2018.

[210] Khoi Nguyen and Sinisa Todorovic. “Feature weighting and boosting for few-shot segmentation”. In: *ICCV*. 2019.

[211] Xu Han et al. “Continual relation learning via episodic memory activation and reconsolidation”. In: *ACL*. 2020.

[212] Yue Tan et al. “FedProto: Federated prototype learning across heterogeneous clients”. In: *AAAI*. 2022.

[213] Mi Luo et al. “No fear of heterogeneity: Classifier calibration for federated learning with non-iid data”. In: *NIPS*. 2021.

[214] Wenke Huang et al. “Rethinking federated learning with domain shift: A prototype view”. In: *CVPR*. 2023.

[215] Zhun Zhong et al. “Re-ranking Person Re-identification with k-reciprocal Encoding”. In: *CVPR*. 2017.

[216] Donald Shenaj et al. “Asynchronous federated continual learning”. In: *CVPR*. 2023.

[217] Zhu Liao et al. “NEPENTHE: Entropy-Based Pruning as a Neural Network Depth’s Reducer”. In: *arXiv preprint arXiv:2404.16890*. 2024.

[218] Tailin Liang et al. “Pruning and quantization for deep neural network acceleration: A survey”. In: *Neurocomputing*. 2021.

[219] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *NIPS*. 2020.

[220] Jonathan Ho and Tim Salimans. “Classifier-free diffusion guidance”. In: *arXiv preprint arXiv:2207.12598*. 2022.

[221] Weihua Chen et al. “Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks”. In: *CVPR*. 2023.

[222] Aditya Ramesh et al. “Zero-shot text-to-image generation”. In: *ICML*. 2021.

[223] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971*. 2023.

[224] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *ICML*. 2021.

[225] Haotian Liu et al. “Visual instruction tuning”. In: *NIPS*. 2024.

[226] Siyuan Li, Li Sun, and Qingli Li. “CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels”. In: *AAAI*. 2023.

[227] Kaiyang Zhou et al. “Learning to prompt for vision-language models”. In: *IJCV*. 2022.

[228] Muhammad Uzair Khattak et al. “Maple: Multi-modal prompt learning”. In: *CVPR*. 2023.

[229] Hamza Rami, Matthieu Ospici, and Stéphane Lathuilière. “Online Unsupervised Domain Adaptation for Person Re-identification”. In: *CVPR Workshops*. 2022.

[230] Shengcai Liao and Ling Shao. “TransMatcher: Deep image matching through transformers for generalizable person re-identification”. In: *NeurIPS* (2021).

[231] Hao Ni et al. “Part-aware transformer for generalizable person re-identification”. In: *ICCV*. 2023.

ECOLE DOCTORALE

Titre : Ecart de domaine et confidentialité pour la réidentification de personnes.

Mots clés : Ré-identification de personnes, Adaptation de Domaine Non Supervisé, Apprentissage Continu, Apprentissage Fédéré.

Résumé : La ré-identification de personnes (Re-ID) vise à identifier des individus à travers des caméras de surveillance non superposées. Malgré leur potentiel de sécurité, les modèles de Re-ID restent limités par l'écart de domaine, c'est-à-dire une divergence entre les données d'entraînement (domaine source) et de déploiement (domaine cible). L'adaptation de domaine non supervisée (UDA) permet d'atténuer ce problème sans nécessiter de labels dans le domaine cible.
Cependant, les réglementations sur la confidentialité, comme le RGPD et l'AI Act, imposent des restrictions strictes sur le stockage et le transfert des données, rendant les approches UDA classiques, qui reposent sur la centralisation des données, inapplicables.
Pour répondre à ces contraintes, nous introduisons l'UDA continue (OUDA-Rid), qui adapte les modèles sur un flux continu de données sans stockage, et l'UDA distribuée (DUDA-Rid), qui décentralise l'adaptation sur plusieurs caméras pour éviter le transfert de données. Nous proposons Source-Guided Similarity Preservation (S2P) et Fed-Protoid. S2P atténue l'oubli catastrophique dans l'OUDA-Rid en préservant les similarités essentielles entre domaines source et cible, assurant ainsi une adaptation continue conforme à la confidentialité. Fed-Protoid utilise l'apprentissage fédéré pour répondre aux restrictions de transfert dans le DUDA-Rid, permettant une adaptation distribuée sans partage d'images sensibles.
Nos frameworks offrent une solution de Re-ID respectueuse de la vie privée tout en réduisant l'écart de domaine. Nous les validons sur plusieurs scénarios, incluant l'adaptation réel à réel et synthétique à réel, avec des jeux de données tels que Market-1501, MSMT17, CUHK03 et RandPerson. Les résultats montrent que S2P et Fed-Protoid assurent des performances robustes dans des conditions réelles.

Title : Domain Gap and Privacy in Person Re-Identification.

Keywords : Person Re-ID ; UDA ; Continual Learning ; Federated Learning.

Abstract : Person Re-Identification (Re-ID) aims to recognize individuals across non-overlapping surveillance cameras. Despite its potential for security, Re-ID models suffer from the domain gap—the discrepancy between training (source domain) and real-world deployment (target domain). Unsupervised Domain Adaptation (UDA) mitigates this issue, enabling adaptation without labeled target data.
However, privacy regulations like GDPR and the AI Act impose strict limits on data storage and transfer, making traditional UDA methods, reliant on centralized data, legally and ethically problematic.
To address this, we introduce Online UDA (OUDA-Rid), which adapts models from continuous data streams without storing past data, and Distributed UDA (DUDA-Rid), which decentralizes adaptation across multiple cameras to prevent data transfer. We propose Source-Guided Similarity Preservation (S2P) and Fed-Protoid to meet these constraints. S2P mitigates catastrophic forgetting in OUDA-Rid by preserving critical feature similarities from the source domain, ensuring privacy-preserving continual adaptation. Fed-Protoid leverages federated learning to address data transfer restrictions in DUDA-Rid, allowing distributed adaptation without sharing sensitive images.
Our frameworks offer a privacy-preserving solution for Person Re-ID while bridging the domain gap. We validate them across multiple scenarios, including real-to-real and synthetic-to-real adaptation, on datasets such as Market-1501, MSMT17, CUHK03, and RandPerson. The results confirm that S2P and Fed-Protoid achieve strong performance under real-world constraints.

Institut Polytechnique de Paris
91120 Palaiseau, France