Universidade Estadual de Campinas
Instituto de Computação

Gabriel Capiteli Bertocco

Self-supervised learning for fully unsupervised re-identification in real-world applications

Aprendizado auto-supervisionado para re-identificação totalmente não-anotada em aplicações no mundo real

CAMPINAS
2024

Gabriel Capiteli Bertocco

Self-supervised learning for fully unsupervised re-identification in real-world applications

Aprendizado auto-supervisionado para re-identificação totalmente não-anotada em aplicações no mundo real

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Anderson Rocha
Co-supervisor/Coorientadora: Dra. Fernanda Alcântara Andaló

Este exemplar corresponde à versão final da Tese defendida por Gabriel Capiteli Bertocco e orientada pelo Prof. Dr. Anderson Rocha.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

B462s

Bertocco, Gabriel Capiteli, 1995-
Self-supervised learning for fully unsupervised re-identification in real-world applications / Gabriel Capiteli Bertocco. – Campinas, SP : [s.n.], 2024.

Orientador: Anderson de Rezende Rocha.
Coorientador: Fernanda Alcântara Andaló.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Aprendizagem não-supervisionada (Aprendizado do computador). 2. Identificação biométrica. 3. Visão por computador. 4. Aprendizado profundo. I. Rocha, Anderson de Rezende, 1980-. II. Andaló, Fernanda Alcântara, 1981--. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações Complementares

Título em outro idioma: Aprendizado auto-supervisionado para re-identificação totalmente não-anotada em aplicações no mundo real
Palavras-chave em inglês:
Unsupervised learning
Biometric identification
Computer vision
Deep learning
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Anderson de Rezende Rocha [Orientador]
Sébastien Marcel
Vitomir Štruc
Esther Luna Colombini
Patrick Flynn
Data de defesa: 18-04-2024
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-7701-7420
- Currículo Lattes do autor: https://lattes.cnpq.br/3770211980444132

Universidade Estadual de Campinas
Instituto de Computação

Gabriel Capiteli Bertocco

Self-supervised learning for fully unsupervised re-identification in real-world applications

Aprendizado auto-supervisionado para re-identificação totalmente não-anotada em aplicações no mundo real

Banca Examinadora:

- Prof. Dr. Anderson de Rezende Rocha
  IC/UNICAMP
- Profa. Dra. Esther Luna Colombini
  IC/UNICAMP
- Prof. Dr. Sébastien Marcel
  IDIAP/Switzerland
- Prof. Dr. Vitomir Štruc
  University of Ljubljana/Slovenia
- Prof. Dr. Patrick Flynn
  University of Notre Dame/USA

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 18 de abril de 2024

Agradecimentos

Começo meus agradecimento àquele que é o Alfa e o Ômega, o Começo e o Fim, ao qual todas as almas buscam e esperam: Nosso Senhor Jesus Cristo. Ele que me concedeu a vida, me forneceu inspirações, e colocou pessoas tão importantes e especiais em minha vida. Também agradeço todas as inspirações e intercessões que Sua Mãe Santíssima, Nossa Senhora, e Seu pai São José me proveram durante toda a caminhada deste doutorado e em minha vida. Também agradeço aos meus santos intecessores, em particular São Josemaria Escrivá, São Padre Pio, Santa Rita de Cássia, Santo Antônio e a todos aqueles que intercederam por mim.

Agradeço com muito amor à minha esposa, Ana Flávia, que, desde quando namorávamos, me apoiou diariamente nesta jornada com sua paciência, carinho e amor.

Agradeço aos meus pais, Murilo e Maria do Carmo, que me ensinaram desde pequeno por palavras e exemplos o valor do estudo e do trabalho duro diário, e ao meu irmão Felipe, que sempre me apoiou com palavras e ações nesta jornada.

Agradeço com carinho à minha coorientadora Dra. Fernanda Andaló, que desde de meu início na vida acadêmica, em Setembro de 2015, ainda na iniciação científica, dedicou seu tempo e sua paciência em meu aprendizado e desenvolvimento enquanto cientista e profissional.

Agradeço ao meu orientador Dr. Terrance Boult por ter me dado a chance e me aceitado como seu aluno, bem como todo o apoio pessoal que ele e sua esposa, Ginger, nos deram quando nos mudamos para os Estados Unidos. Sua ajuda, seja emprestando quase tudo que tínhamos em nossa casa, como acordando às 4:00 da manhã do dia 05 de abril de 2023 para levar minha esposa ao hospital para que ela desse a luz ao nosso primeiro filho, Bento, nosso maior tesouro, é algo que será impossível retribuir. Em nome também de minha esposa, muito obrigado por todo esse acolhimento e carinho, Ginger e Terry! (*Thank you very much, Ginger and Terry, for everything you have done for us! You are very special and we hold you guys deeply in our hearts!*). Agradeço também a minha sogra Flávia, por toda sua dedicação e ajuda nas primeiras semanas de nascimento do Bento, por ter nos amparado e ajudado minha esposa nos cuidados de nosso filho, permitindo que eu dedicasse mais tempo para o desenvolvimento desta pesquisa.

E, claro, agradeço ao meu orientador Prof. Dr. Anderson Rocha, que permitiu que eu conhecesse a Dra. Fernanda e Prof. Terrance Boult, e que abriu as portas em minha jornada profissional; me aceitou e deu a chance para um garoto de 19 anos de ser seu aluno de Iniciação Científica. Graças a ele, esse garoto se desenvolveu, apresentou sua pesquisa na Itália, Estados Unidos, Eslovênia e Emirados Árabes, e agora finaliza o doutorado. Obrigado, Prof. Anderson, pela amizade, parceria e confiança!

Agradeço aos meus amigos que sempre estiveram ao meu lado, e a todos aqueles que, de forma direta e indireta, me auxiliaram nesta jornada. Infelizmente, não consigo nominalmente agradecer a cada um nestas breves linhas.

Agradeço ao financiamento pelos processos nº 2017/12646-3, nº 2019/15825-1, nº

2022/02299-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.

Por fim, espero que esta pesquisa inspire aqueles que farão seu uso e desenvolvimento na área de tecnologia a sempre usarem-na para o bem, preservando as liberdades e privacidades individuais.

Nosso Senhor Jesus Cristo, Sua Mãe Santíssima, minha esposa Ana Flávia, meu filho Bento e aos nossos futuros filhos: essa pesquisa é por vocês!

Resumo

Um dos problemas mais complexos em Aprendizado de Máquina é lidar com dados não rotulados. A maioria dos modelos com alto desempenho depende de massiva quantidade de dados rotulados para obter os melhores resultados. No entanto, rotulação não é fácil nem confiável por ser uma tarefa altamente demorada, custosa e propensa a erros. Além disso, vieses nos dados rotulados podem ser propagados para o modelo, prejudicando seu desempenho e generalização. Assim, é primordial desenvolver métodos que possam encontrar padrões em cenários totalmente não supervisionados, permitindo uma implementação rápida e menos propensa a vieses. Esses modelos podem ser usados em diversas aplicações, como investigações forenses, biometria e compreensão de eventos. Esta pesquisa propõe algoritmos de aprendizado auto-supervisionado para lidar com dados não rotulados em cenários desafiadores. Um cenário desafiador pode conter alta disparidade intraclasse (representações da mesma classe estão distantes umas das outras no espaço vetorial) e alta similaridade interclasse (amostras de classes diferentes podem estar mais próximas umas das outras). Para instanciar esse complexo requisito com os desafios mencionados, nossa exploração se concentra em duas aplicações: ReIdentificação (ReID) Não Supervisionada de Pessoas e Objetos, devido à sua aplicabilidade em compreensão de eventos, e Atribuição de Autoria em Texto. Considerando essas aplicações, nesta tese, propomos quatro métodos que lidam com níveis variados de complexidade em cenários não supervisionados. Nossas três primeiras soluções visam a tarefa de ReID Não Supervisionado de Pessoas, onde assumimos que não temos a anotação de identidade, ou seja, não sabemos "quem" foi detectado na imagem. A primeira solução considera meta-informações, como anotação de câmera, para auxílio na resolução da tarefa. Como existem cenários onde informação de câmera não está disponível, nossa segunda solução é totalmente não supervisionada, ou seja, não requer nenhuma informação adicional. Assim, pode-se aplicá-la em outras tarefas, em diferentes modalidades, como Atribuição de Autoria em Texto em postagens em redes sociais. O terceiro método também lida com cenários de reidentificação não supervisionada, mas com conjuntos de dados em grande escala. Mostramos também que podemos estendê-la para reidentificação de objetos, como, por exemplo, veículos. A quarta solução considera o problema de reconhecimento de longo alcance por meio de treinamento supervisionado. O modelo aprende com imagens distorcidas devido à turbulência atmosférica, e alcança resultados estado-da-arte em ambas as tarefas de ReID de Pessoas e Reconhecimento Facial. As soluções propostas nesta pesquisa podem ser acopladas em *pipelines* de aplicações forenses e de biometria. Elas podem ser empregadas para compreensão de eventos, em que as autoridades visam encontrar suspeitos e investigar o comportamento das pessoas, bem como relações com objetos em uma cena. As soluções podem ser usadas para obter uma compreensão do que ocorreu e propor caminhos de investigação. Elas também podem ser empregadas em modelos de biometria baseados em IA para proteção em lugares que exigem alta segurança, como instalações governamentais, segurança de fronteiras, infraestrutura crítica e anti-terrorismo.

Abstract

One of the most complex problems in Machine Learning is dealing with unlabeled data. Most top-ranking models rely on massive labeled data to achieve state-of-art results. However, data labeling is not easy nor reliable to obtain due to the highly time-consuming, costly, and error-prone task of annotation. Moreover, bias in the labeled data might be propagated to the model, hindering its performance and generalization. It is paramount to develop methods that can mine patterns in a fully-unsupervised scenario allowing a fast and bias-alleviated deployment. These models could be used in a range of applications, such as forensic investigations, biometrics, and event understanding. This research proposes self-supervised learning algorithms to deal with unlabeled data for deployment in challenging label-absent scenarios. A challenging setup might contain high intra-class disparity (features from the same class are far away from each other in the feature space) and high inter-class similarity (samples from different classes might be closer to each other). To instantiate this complex requirement with applications that capture the mentioned challenges, our exploration focuses on two applications: Unsupervised Re-Identification (ReID) of People and Objects, due to their applicability to event understanding, and on the Text Authorship Verification task. Considering these applications, in this thesis, we propose four methods that deal with varied levels of complexity in unsupervised scenarios. Our first three solutions target the Unsupervised Person ReID task where we assume we do not have identity labeling, i.e., we do not know "who" is detected in the image. The first solution considers meta-information, such as camera labels, to effectively address the task. As there are scenarios where it is not applicable, our second solution is fully unsupervised, i.e., it does not require any side information. Because of this, it can be applied to further tasks than Person ReID in different modalities, such as Text Authorship Attribution in social media posts. The third method also deals with fully unsupervised re-identification scenarios but in large-scale datasets. We also show that this solution can be applied to object re-identification, specifically vehicles. The fourth solution changes the setup by considering supervised training, however targeting long-range recognition. It learns from images mainly distorted by atmospheric turbulence and achieves state-of-the-art results in both Person ReID and Face Recognition tasks. The proposed solutions can be implemented as part of forensic and biometrics pipelines. For instance, they can be employed for event understanding where authorities aim to find possible suspects and investigate people's behavior as well as their possible relationships with objects in a scene. They can be used to get an understanding of what happened and possible investigation insights. The solutions can be also employed in AI-powered biometrics for security-sensitive protection in places such as government facilities, border security, critical infrastructure, and counterterrorism.

Contents

Chapter 1

Introduction

Machine learning and Artificial Intelligence (AI) have been extensively applied in different contexts in society: face recognition, speech recognition, geo-localization, medical diagnostics, activity recognition, credit score analysis, synthetic realities, among others [36, 213, 11]. Most of the success relies on a careful annotation of the available data to supervise the learning process. However, the labeling task is time-consuming, expensive, and error-prone. Moreover, the annotation process is subject to the judgment of the annotator to say if two or more samples belong to the same class or not, which might introduce biases on the labels and hinder the generalization of the trained model.

In this context, increasing attention has been witnessed on Self-Supervised Learning techniques due to their capacity to mine patterns and learn features from a dataset without requiring human supervision. Recent works on image recognition, such as SimCLR [22], CPC v2 [63], MoCo [57], and Dino [14], deal with this problem in a contrastive manner by generating augmented versions of the same image by occlusion, cropping, color jittering, to name a few. Other strategies are based on feature disentangling (Barlow Twins [222]), and clustering (DeepCluster [12], SwAV [13], DinoV2 [135]). Despite the comparable performance to the supervised counterparts, most works rely on the ImageNet dataset in which the majority of the classes have significant differences in semantics (e.g., it contains classes representing airplanes, animals, sports, etc.), allowing the models to rely more on coarse details than on fine-grained details to distinguish classes. In this case, the mere use of well-known augmentation strategies (such as cropping, blurring, erasing, and flipping) provides sufficient variation for optimization. However, there are many real-case problems where models need to strongly rely on fine-grained details for learning, for instance, Person Re-Identification.

Person Re-Identification (PReID) enables a broad range of applications in Computer Vision, Forensic Science, and Biometrics, such as person tracking, crime investigation and surveillance. PReID aims to retrieve the same person seen in one camera from all the other cameras present in a camera system. The same person might be seen by many cameras located in different positions in an environment. From one camera to the other, the same person might be under different illumination (from inside of a building to outside), occlusions, points of view (in one camera his/her back is recorded while in the other is his/her front), resolutions (cameras are at different distances from people or have different pixel density), and background. Besides, people in the same environment (e.g.,

at an airport) usually dress similarly (most people with backpacks, bags, and luggage), making different classes semantically closer among themselves. For this reason, PReID usually faces a high intra-class discrepancy and high inter-class similarity, as depicted in Figure 1.1.

Figure 1.1: Illustration of the high intra-class variance and inter-class disparity usually faced in Person Re-Identification (PReID) scenarios. Each set of seven images belongs to the same identity. In the first row we see that, for the same identity, their whole-body images are in different positions, resolutions, backgrounds, and illumination conditions giving a high variation in the samples of the same class. In the second row, we see three different identities wearing similar clothes. Just minor details in the image make people distinguishable from each other, such as shoes, hair, backpacks, and, when available, their faces.

Considering a similar task, Vehicle Re-Identification (VReID), in which the "identities" are vehicles, the same challenges are faced. A vehicle can be recorded under different environmental conditions and the license plate is not always visible. Image samples for VReID are shown in Figure 1.2.

Figure 1.2: The Vehicle Re-Identification (VReID) task faces similar challenges to PReID. Images show the same vehicle in different positions, resolutions, illuminations, backgrounds, and sometimes under occlusion, which hardens the re-identification task. Image reproduced from [119].

Besides the aforementioned challenges, PReID and VReID are intrinsically open-world problems: identities on the test set (query and gallery sets) are disjointed from those on the

training set, while the training and test sets of `ImageNet` [31] and many other datasets contain the same classes. Moreover, the number of classes on PReID datasets is also as large as `ImageNet`: `Market1501` [242] has 751 identities in the training set, `DukeMTMC-ReID` [151] has 702, and `MSMT17` [194] even surpasses ImageNet with 1,041 classes in the training set. The same happens for large-scale VReID datasets. `VehicleID` [108] comprises 13,164 vehicle identities in the training set in a total of 113,346 images, and `Veri-Wild` [119] comprises 30,671 identities in the training set in a total of 277,797 images.

This clearly shows that PReID and VReID are more demanding tasks than ImageNet classification, and general state-of-the-art self-supervised learning methods are not suitable for the task. These methods result in much less accurate models in comparison to prior unsupervised learning methods tailored specifically to PReID, even with ImageNet weight initialization [230, 48]. Since on real deployment, the person and vehicle images come from heterogeneous domains (surveillance, mobiles, or media cameras) and are unlabeled (i.e., we do not know "who" or "what" is recorded by the cameras), it is paramount to develop a system that can handle data from multiple domains in a fully-unsupervised way, requiring a more robust feature learning.

The problem under investigation in this Ph.D. research holds significant societal applications and impacts. It has been faced, for example, by U.S. authorities after the Capitol Invasion in January 2021 [127]. In that case, the main challenge was to find the rioters at the event and identify people recorded by cameras inside and outside the Capitol. A similar challenge was faced by the Federal Police of Brazil to identify rioters and people that stormed and destroyed Brazil's Congress, Supreme Court, and presidential offices [27] on January 8^{th} 2023. The results presented in this thesis indicate that our solutions are promising to assist in the resolution of these real-world tasks. Another example is the 2018 fire at the National Museum in Brazil, prompting Brazilian authorities to question whether it was a criminal act or caused by negligence [159]. Our pipelines could be applied to determine suspects around the museum, moments before the fire, to propose a possible answer to that question and, even after the event, to check the dynamics of the firemen to combat the fire.

In Biometrics, we could leverage our proposed solutions to assess how people behave in public places (airports, banks, shopping centers) and identify the objects they interact with. In this case, we can aid studies on crowd behavior to avoid possible tragedies, as in 2013, when a Hindu festival in India faced a stampede leaving about 115 deaths. As reported by BBC [178], "... better crowd management could have prevented the tragedy." Also, our solutions can be deployed in sensitive-security scenarios such as energy infrastructure security, surveillance systems, or counter-terrorism [74].

This thesis introduces four Unsupervised Re-Identification (U-ReID) solutions, each evolving in performance, generalization, and the complexity of the considered scenario. In summary, our solutions comprise the techniques to find possible people, or groups of people, and objects involved in an event and to, ultimately, propose candidate suspects for further investigation [137].

In the first proposed method, presented in Chapter 2, we aim at re-identifying people in a camera system and we assume that we know from which camera each person has been recorded, i.e., we do not know who is present in a given frame, but we know the camera

label. We start only from a few (usually three) Deep Convolutional Neural Networks (DCNN) models previously trained on another Re-ID dataset to learn initial features related to the problem. We propose an Unsupervised Domain Adaptation algorithm to train these networks over the target unlabeled data leveraging a novel Cross-Camera Triplet creation strategy on training, a self-ensembling strategy, and backbone ensembling on the evaluation phase. This setup has been published [7] in the IEEE Transaction on Information Forensics and Security (T-IFS) in August 2021. This solution was also presented in the 13^{th} IEEE International Workshop on Information Forensics and Security in December 2021, and in the Workshop of Long-Range Recognition[1] during the Winter Application in Computer Vision (WACV) 2023.

The second solution, presented in Chapter 3, we approach a real deployment by disregarding the camera information and taking the same backbones but without pre-training on any task-specific dataset. We only have the person's bounding box without identity or camera annotation, and the backbones have their weights initialized over ImageNet. We propose a novel ensemble-based strategy to combine neighborhood-based distances between samples from each manifold in a single distance matrix, ensembling different knowledge from each backbone. Moreover, we also present a novel ensembled-based clustering strategy that combines clustering results for different hyper-parameter values to obtain clusters with lower false-positive rates.

This second solution does not consider any task-related meta information, being generalizable to further tasks than PReID. To analyze this generalization ability, we consider a second task in the Natural Language Processing (NLP) field. The task regards the Text Authorship Attribution (TAA) for short messages. More specifically, the goal is to group tweets (X^2 - former Twitter - short messages) of the same author in a fully unsupervised manner, considering raw texts as input. To the best of our knowledge, we are the first ones to apply the same self-supervised learning pipeline to different modalities with minor adjustments in two forensics tasks facing high intra-class disparity and inter-class similarity.

This pipeline outperforms the state-of-the-art in U-ReID and obtained promising performance on text analysis. The article proposing this pipeline has been published in IEEE T-IFS [8] and also accepted in the journal track session of the IEEE International Joint Conference on Biometrics (IJCB) 2023[3]. Besides, the solution was also presented in the InterForensics2023[4], the largest Forensics conference in Latin America.

The third proposed method is presented in Chapter 4. We keep the same constraints from the second solution, however with novel strategies to deal with large-scale scenarios and with an extension for Unsupervised Vehicle Re-Identification (U-VReID). The solution presents some improved components: a self-supervised model pre-initialization over the target data, a new sampling technique to reduce data size in each iteration, a more efficient ReRanking technique to meet the large-scale learning requirements, a new clustering hyperparameter scheduling, and a novel co-training label method. It outperforms the

[1]https://sites.google.com/kitware.com/lrr-workshop-2023/home
[2]https://twitter.com/?lang=en
[3]https://ijcb2023.ieee-biometrics.org/accepted-papers/
[4]https://interforensics.com/site/interforensics2023/apresentacao

state-of-art methods that use whole datasets to perform ReRanking and also test-selected hyperparameters, i.e., they select the best hyperparameters for each dataset based on the final query/gallery sets split. We argue this is unrealistic since the common assumption is that the data is fully unlabeled, so it would be impossible to perform grid-searching to find the optimal hyperparameters in a real-world scenario.

To summarize, our third pipeline aims to tackle large-scale learning with a local ReRanking, with less sensibility to hyperparameter choices, and a novel co-training label strategy to improve clustering performance. The current solution is under review in the top-tier IEEE Transactions on Image Processing, and it was presented in InterForensics 2023. Finally, it was employed in a research consultancy for the Federal Police of Dubai in December, 2023.

The fourth solution in this thesis, presented in Chapter 5, addresses an expanding field in biometrics: long-range recognition. It was designed to perform Face Recognition and Person Re-Identification with images under different distortion levels caused mainly by atmospheric turbulence. The Ph.D. candidate was one of the designers of the solution and performed all experiments, analyses, and conclusions regarding the method dealing with the Person Re-Identification task. It is a co-authored work developed during his internship at the University of Colorado Colorado Springs (UCCS), USA, when he was member of the Biometric Recognition and Identification at Altitude and Range (BRIAR) program[5], a United States Government-supported project devoted to counterterrorism, protection of critical infrastructure, and transportation facilities, military force protection, and border security. The solution, despite not being as unsupervised as previous solutions, is part of a larger solution that considers unsupervised techniques during evaluation to improve the whole-body person matching performance. The article describing our solution has been published in the IEEE Access journal [152], and it is also part of an end-to-end identification method that has been published in a joint paper with other BRIAR members in the IEEE IJCB 2023 [39].

1.1 Research Questions

Our ultimate research goal was to develop self-supervised learning algorithms that perform robust feature learning from data to effectively mine patterns on problems with a high intra-class dissimilarity and inter-class similarity, focusing on Re-Identification. In this research project, we aim to answer the following questions:

1. What are the constraints or meta-information that can help improving self-supervised learning on ReID tasks?

2. How to design a general self-supervised learning algorithm, considering fully unlabeled data, to deal with complex problems: high intra-class dissimilarity and inter-class similarity, and identity-disjoint train and test sets?

3. How to scale self-supervised solutions to handle thousands of data samples?

[5]https://www.iarpa.gov/research-programs/briar

4. Which strategies can potentially help to perform long-range re-identification?

1.2 Contributions of this thesis

We envision self-supervised learning algorithms that can group samples of the same class on fully unlabeled scenarios on problems where the classes potentially have high inter-class similarity and intra-class dissimilarity. Considering the Re-Identification as a case study, the contributions of this research are:

- A method that effectively considers meta-information (e.g., camera label), when available, and knowledge learned from a different source domain to adapt to an unknown (unlabeled) target domain. More specifically, the designed model has:
    1. A novel cross-camera triplet-based strategy to encourage camera-invariant feature learning.
    2. A new self-ensembling strategy that combines checkpoints generated throughout the training without human intervention nor a validation set to meet unsupervised learning requirements.
    3. A new ensembling strategy during validation that does not require complex cross-supervision or hyperparameters but increases search performance.
- A fully unsupervised solution that deals in more complex scenarios where no meta-information is available nor source domain to perform weight initialization. Under these constraints, our model employs:
    1. A novel neighborhood-based ensembling strategy that combines knowledge from different architectures during training without complex cross-supervision, co-training, weighting, hyperparameters, or human intervention.
    2. A clustering fusing strategy that combines knowledge results from different clustering runs with different parameter definitions. It avoids selecting specific hyperparameter values for each dataset, which is unrealistic in a fully unsupervised scenario and can bring human biases to the model.
    3. A final solution that can be employed in Computer Vision (Unsupervised ReID) and NLP (Text Authorship Attribution) with minor adjustments.
- A large-scale fully-unsupervised solution to learn from large-scale unlabeled data in different re-identification tasks, without human intervention for clustering hyperparameter definition or knowledge sharing among the backbones. The solution has:
    1. A new neighborhood-based sampling strategy to decrease dataset size and consequently the training time in each epoch.
    2. A novel neighborhood-based re-ranking strategy that does not rely on the entire distance matrix among the datapoints for distance refinement.

3. A new noise-aware scheduling for the clustering hyper-parameter definition that follows the dynamic of the feature space, and alleviates the impact of noisy labeling while keeping the diversity in the clusters. It also avoids dataset-specific hyperparameters, which is unrealistic in a fully unsupervised scenario.
4. A co-training method that allows knowledge sharing among the architectures without relying on complex cross-supervision, human intervention and hyper-parameter tuning.

- A method that is able to learn from distorted data, mainly caused by atmospheric turbulence, to improve the robustness of the model to distortions and its performance in long-range recognition scenarios. The model employs:

  1. A new atmospheric turbulence-based augmentation that better simulates real distortions than the well-known Gaussian blur and down-sampling processing.
  2. A distortion-adaptive training strategy where we dynamically weigh different levels of distortions along the training in an easy-to-hard manner to encourage better optimization.
  3. A novel feature magnitude-based model ensembling to effectively combine knowledge from two backbones trained with and without distorted data.

Chapter 2

Unsupervised and Self-Adaptative Techniques for Cross-Domain Person Re-Identification

The labeling of massive datasets demanded by deep learning is time-consuming and error-prone, especially when targeting forensic and biometric applications. In this context, Unsupervised Domain Adaptation (UDA) aims to adapt a model trained on a source dataset to a target domain without the need for identity information of the target samples. Most ReID methods that follow this approach are based on label proposing, in which feature vectors of target images are extracted and clustered. Upon unsupervised training, these clusters receive pseudo-labels for the adaptation to the target domain. Prior works [42, 162, 45, 233, 224, 213] apply the pseudo-labeling principle by developing different ways to propose and refine clusters on the target domain. The aim is to alleviate noisy labels, which can harm feature learning.

In a high-level view, UDA methods first pre-train a model in some labeled domain (i.e., a dataset that has identity annotation for detected people), and after that, they learn from an unknown target domain (i.e., they assume the dataset does not have the identity annotation for the detected people). To perform unsupervised learning, most methods consider three steps iteratively: feature extraction, clustering, and fine-tuning. In the feature extraction step, features for the unlabeled images are extracted. As image labels are not available, the features are clustered and pseudo-labels are assigned to them, which are employed for supervised-like learning in the fine-tuning step. Prior works propose novel strategies in one, two, or all steps, to enhance performance over the target domain. The first method proposed in this research, and presented in this chapter, follows this pipeline, with novelties in each of the three steps.

As we are dealing with data from an unknown target domain, clusters can have different degrees of reliability, i.e., contain different quantities of noisy labels. We need to select the most reliable clusters to optimize the model at each iteration of the clustering process. The trained model must also be camera-invariant to yield the same feature representation for an identity, regardless of the camera point of view. Based on these observations, we hypothesize that clusters with more cameras might be more reliable to optimize the model. Suppose that a cluster contains images of the same identity seen from two or more

cameras. In this case, the model was able to embed these images close to each other in the feature space, overcoming differences in illumination, pose, and occlusion, which are inherently present in different camera vantage points.

We argue that the greater the number of different cameras in a cluster, the more reliable this cluster is to optimize the model. Following this idea, we propose a new way to create triplets of samples in an offline manner. We select one sample as an anchor for each camera represented in a cluster and two others as positive and negative examples. As a positive example, we choose a sample from one of the other represented cameras. In contrast, the negative example is a sample from a different cluster but with the same camera as the anchor. Consequently, the greater the number of cameras in a cluster, the more diverse the triplets to train the model. With this approach, we give more importance to the more reliable clusters, regularize the model, and alleviate the dependency on hyperparameters by using a single-term and single-hyper-parameter triplet loss function. This technique brings robustness and generability to the final model, easing its adaptation to different scenarios.

Another important observation is that, at different points of the adaptation from a source to a target domain, the model holds different levels of knowledge as different portions of the target data are considered each time. Thus, we argue that the model has complementary knowledge in different iterations during training. Based on this, we propose a self-ensembling strategy to summarize the knowledge from various iterations into a unique final model.

Finally, based on recent advances in ensemble-based methods for ReID [47, 226], we propose to combine the knowledge acquired by different architectures. Unlike prior work, we avoid complex training stages by simply assembling the results from different architectures only during evaluation time.

To summarize, the contributions of this first method are:

- A new approach to creating diverse triplets based on the variety of cameras represented in a cluster. This approach helps the model to be camera-invariant and more robust in generating a person's features from different perspectives. It also allows us to leverage a single-term and single-hyper-parameter triplet loss function to be optimized.

- A novel self-ensembling fusion method, which enables the final model to summarize the complementary knowledge acquired during training. This method relies upon the knowledge held by the model at different checkpoints of the adaptation process.

- A novel ensemble technique to take advantage of the complementary knowledge from different backbones trained independently. Instead of applying the typical knowledge distilling [65] or co-teaching [54, 18] methods, which add complexity to the training process, we propose using an ensemble-based prediction.

This first solution was published [7] in the IEEE Transaction on Information Forensics and Security (IEEE T-IFS) in August 2021. This solution was also presented in the 13^{th} IEEE International Workshop on Information Forensics and Security in December

2021[1], and in the Workshop of Long-Range Recognition[2] during the Winter Conference on Applications of Computer Vision (WACV) 2023.

2.1 Related Work

Prior works address Unsupervised Domain Adaptation for Person Re-Identification. They can be roughly divided into three categories: generative, attribute alignment, and label-proposing methods.

2.1.1 Generative Methods

ReID generative methods aim to synthesize data by translating images from a source to a target domain. Once data from the source dataset is labeled, the translated images in the target context receive the same labels as the corresponding original images. The main idea is to transfer low- and mid-level characteristics from the target domain, such as background, illumination, resolution, and even clothing, to the images in the source domain. These methods create a synthetic dataset of labeled images with the same conditions as the target domain. To adapt the model, they apply supervised training. Some works in this category are SPGAN [33], PTGAN [194], AT-Net [109], CR-GAN [23], PDA-Net [100], and HHL [246]. Besides transferring the characteristics from the source to the target domain for image-level generation, DG-Net++ [258] also applies label proposing through clustering. The final loss is the aggregation of the GAN-based loss function to generate images, along with the classification loss defined for the proposed labels. By doing this, they perform the disentangling and adaptation of the features on the target domain.

CCSE [105] performs camera mining and, using a GAN-based model, generates synthetic data for an identity considering the point of view of each other camera, increasing the number of images available for training. They leverage new clustering criteria to avoid creating massive clusters comprising most of the dataset and potentially having two or more true identities assigned to the same pseudo-label. Finally, they train directly from ImageNet, without considering any specific source domain. In comparison, our solution does not require synthetic images since we explore the cross-camera information inside each cluster using only real images. This leads our method to outperform CCSE considering the same training conditions (unsupervised scenario).

2.1.2 Attribute Alignment Methods

These methods seek to align common attributes in both domains to easily transfer knowledge from source to target. Such features can be clothing items (backpacks, hats, shoes) and other soft-biometric attributes that might be common to both domains. These works align mid-level features and enable the learning of higher semantic features on the target domain. Works such as TJ-AIDL [184] consider a fixed set of attributes. However,

[1]https://wifs2021.lirmm.fr/session-spl-tifs-papers-1/
[2]https://sites.google.com/kitware.com/lrr-workshop-2023/home

source and target domains can have substantial context differences, leading to potentially different attributes. For example, the source domain could be recorded in an airport and the target domain in a shopping center. To obtain a better generalization, in [103], the authors propose the Multi-task Mid-level Feature Alignment (MMFA) technique to enable the method to learn attributes from both domains and align them for a better generalization on the target domain. Other methods, such as UCDA [148] and CASCL [197], aim to align attributes by considering images from different cameras on the target dataset.

2.1.3 Label Proposing Methods

Methods in this category predict possible labels for the unlabeled target domain by leveraging clustering methods (K-means [117], DBSCAN [41], among others). Once the target data is pseudo-labeled, the next step is to train models to learn discriminative features in the new domain. PUL [42] applies the Curriculum Learning technique to adapt a model learned on a source domain to a target domain. However, as K-means is used to cluster the features, it is not possible to account for camera variability. As K-means generates only convex clusters, it cannot find more complex cluster structures, hindering the performance. UDAP [162] and ISSDA-ReID [169] utilize DBSCAN as the clustering algorithm along with labeling refinement. SSG [45] also applies DBSCAN to cluster features of the whole, upper, and low-body parts of identities of interest. The final loss is the sum of individual triplet losses in each feature space (body part). Similar to our work, they use a source domain to pre-train the model and the target domain for adaptation. However, they do not perform cross-camera mining, cluster filtering, or ensembling. These elements of our solution allow it to outperform SSG in all adaptation scenarios.

ECN [247], ECN-GPP [248], MMCL [179], and Dual-Refinement [28] use a memory bank to store features, which is updated along the training to avoid the direct use of features generated by the model in further iterations. The authors aim to avoid propagating noisy labels to future training steps, contributing to keeping and increasing the discrimination of features during training.

PAST [233] applies HDBSCAN [10] as the clustering method, which is similar to OPTICS [1] — the algorithm of choice in our work. However, the memory complexity of OPTICS is $O(n)$, while for HDBSCAN is $O(n^2)$, making our model more memory efficient in the clustering stage.

MMT [47], MEB-Net [226], ACT [207], SSKD [111], and ABMT [18] are ensemble-based methods. They consider two or more networks and leverage mutual teaching by sharing one network's outputs with the others, making the whole system more discriminative on the target domain. However, training models in a mutual-teaching regime increase the complexity of needed memory and the general training process. Besides that, noisy labels can be propagated to other ensemble models, hindering the training process. Nonetheless, ensemble-based learning provides the best performance among state-of-the-art methods. We propose using ensembles only during inference to simultaneously eliminate the complexity added to the training, still taking advantage of knowledge complementary between the models.

Our first solution is also based on Curriculum Learning with Diversity [78], a schema

whereby the model starts learning with easier examples, i.e., samples that are correctly classified with a high score early in training. However, in a multi-class problem, one of the classes might have more examples correctly classified early on, making it easier than the other classes. Therefore, in Curriculum Learning with Diversity, the method selects the most confident samples (easier samples) from the easier classes, including some examples from the harder ones. In this way, it enables the model to learn in an easy-to-hard manner, avoiding local minima and allowing better generalization.

Even though recent work achieves competitive performances, there are some limitations that we aim to address in our work. First, generative methods bring complexity by considering GANs to translate images from one domain to the other. Second, attribute Alignment methods only tackle the alignment of low and mid-level features. Third, methods in both categories need images from source and target domains during adaptation. Finally, the last Label Proposing methods consider mutual learning or co-teaching, which brings complexity to the training stage.

Similarly, we assume to have only camera-related information, i.e., we know from which camera (viewpoint) an image was taken. In all steps, we use pseudo-identity information exclusively given by the clustering algorithm without relying on any ground-truth information. We differ from the prior art by using a new diversity learning scheme and generating triplets based on each cluster's diversity of points of view. As we train the whole model, the method also learns high-level features on the target domain. We simplify the training process by considering one backbone at a time, without mutual information exchange during adaptation. Finally, we apply model ensembling for inference after the training process.

2.2 Proposed Method

Our approach to Person ReID comprises two phases: training and inference. Figure 2.1 depicts the training process, while Table 2.1 shows the variables used in this work.

During training, we independently optimize n_b different backbones to adapt the model to the target domain. This phase is divided into five main stages that are performed iteratively: feature extraction from all data; clustering; cluster selection; cross-camera triplet creation and fine-tuning; and feature extraction from pseudo-labeled data.

After training, we perform the proposed self-ensembling phase to summarize the training parameters in a single final model based on the weighted average of model parameters from each different checkpoint. We perform this step for each backbone independently and, in the end, we have n_b self-ensembled models.

During inference, for a pair query/gallery image, we calculate the distance between them considering feature vectors extracted by each of the n_b models. Hence, for each query/gallery pair, we have n_b distances, one for each of the trained models. We then apply our last ensemble technique: the n_b distances are averaged to obtain a final distance. Finally, based on this final distance, we take the label of the closest gallery image as the query label.

Table 2.1: Variables' meaning in this work

Variable	Meaning
n_b	Number of different backbones in the Ensemble
M	Model backbone
K_1	Number of iterations of the blue flow in Figure 2.1
K_2	Number of iterations of the orange flow in Figure 2.1
c_i	i-th cluster in the feature space
n_i	Number of cameras in cluster c_i
cam_j	j-th camera in a cluster
x_i^s	i-th image in the source domain
x_i^t	i-th image in the target domain
y_i^s	Label of the i-th image in the source domain
N_s	Number of images in the source domain
N_t	Number of images in the target domain
m	Number of anchors per camera in a cluster
α	Margin parameter of the Triplet Loss
B	Batch of triplets in an iteration

2.2.1 Training Stages 1 and 2: Feature Extraction from all data and Clustering

Let $D^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ be a labeled dataset representing the source domain, formed by N_s images x_i^s and their respective identity labels y_i^s; and let $D^t = \{(x_i^t)\}_{i=1}^{N_t}$ be an unlabeled target dataset representing the target domain, formed by N_t images x_i^t. Before applying the proposed pipeline, we first train a model M in a supervised way, with source dataset D^s and its labels. After training, assuming source dataset D^s is not available anymore, we perform transfer learning, updating M to the target domain, only considering samples from unlabeled target dataset D^t.

With model M trained on D^s, we first extract all feature vectors from images in D^t and create a new set of feature vectors $\{M(x_i^t)\}_{i=1}^{N_t}$. We remove possible duplicates by checking if there is a replacement from one of them, which might be caused by duplicate images on target data. The remaining feature vectors are L2-normalized to embed them into a unit hypersphere. The normalized feature vectors are clustered using the OPTICS algorithm to obtain pseudo labels.

The OPTICS algorithm [1] leverages the principle of dense neighborhood, similarly to DBSCAN [41]. DBSCAN defines the neighborhood of a sample as being formed by its closest feature vectors, with distances lower than a predefined threshold. Clusters are created based on these neighborhoods, and samples not assigned to any cluster are considered outliers. If the threshold changes, other clusters are discovered, and current clusters can be split or combined to create new ones. In other words, if we change the threshold, other clusters might appear, creating a different label proposing for the samples. However, clusters that emerge from real labels often have different distributions and densities, indicating that a generally fixed threshold might not be sufficient to detect them. In this sense, OPTICS relaxes DBSCAN by ordering feature vectors in a manifold based on the distances between them, which allows the construction of a *reachability plot*.

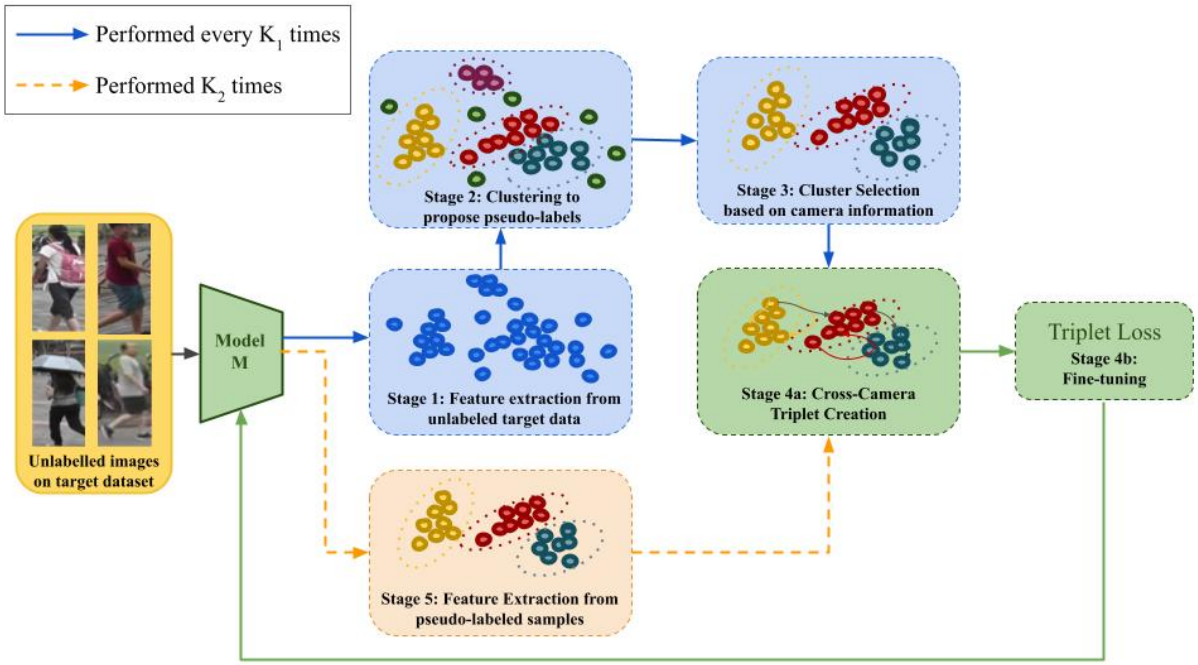


Figure 2.1: Overview of the training phase. We assume to have camera-related information, i.e., we know the camera used to acquire each image, and we do not rely on any ground-truth label information about the identities on the target domain. The pipeline has two flows: the blue flow is executed every K_1 times, and the orange flow is executed K_2 times. Both flows share steps in green. In Stage 1, we initially extract feature vectors for each training image in the target domain using model M, and cluster them using the OPTICS algorithm in Stage 2 to propose pseudo-labels. Afterward, we perform cluster selection in Stage 3, removing outliers and clusters with only one camera. Then, triplets are created based on each cluster's diversity in Stage 4a and used to train the model in Stage 4b. These steps are denoted by the blue flow in which the Clustering and Cluster Selection are performed. Instead of going back to Stage 1, the method follows the orange flow. In Stage 5, we extract feature vectors of the samples selected in Stage 3, and the process continues to Stages 4a and 4b again. The blue flow marks an iteration, while the orange flow is called an epoch. Therefore, in each iteration, we have K_2 epochs.

Probable clusters with different densities are revealed as valleys in this plot and can be detected by their steepness. With this formulation, we are more likely to propose labels closer to real label distribution on the target data.

2.2.2 Training Stage 3: Cluster Selection

After the first and second stages, feature vectors are either assigned to a cluster or considered outliers. As people can be captured by one or more cameras in a ReID system, the produced clusters are naturally formed by samples acquired by different devices. We hypothesize that clusters with samples obtained by two or more cameras are more reliable than clusters with only one camera.

If an identity is well described by model M, its feature vectors should be closer in the feature space regardless of the camera. Therefore, clusters with only one camera might be created due to bias to a particular device or viewpoint, and different identities captured by the same camera can be assigned to the same cluster. Besides, if a feature vector is predicted as an outlier by the clustering algorithm, it means that it does not have a good description of its image identity to be assigned to a cluster.

Based on these observations and for optimization purposes, we filter the feature vectors by discarding outliers and clusters with a single camera type. With camera-related information, it is possible to count the number of images from each camera in a cluster. If all samples in a cluster come from the same camera, it is removed from the feature space. By doing this, we keep in the feature space only clusters with images from at least two cameras. Figure 2.1 depicts this process, from Stage 2 to Stage 3, in which the outlier samples (green points) and clusters with only one camera (magenta points) are removed from the feature space.

The remaining clusters (the ones with two or more cameras) are considered reliable to fine-tune model M. Furthermore, different clusters have different degrees of reliability based on the number of represented cameras. Suppose images captured by several cameras form a cluster. In that case, it means model M can embed samples of the corresponding identity captured by all of these cameras in the feature space, eliminating point-of-view bias. In contrast, the fewer images from different points of view, the more complex the identity definition. In this sense, we propose a new approach of creating cross-camera triplets of samples to optimize the model by emphasizing cluster diversity and forcing samples of the same identity to be closer in the feature space regardless of their acquisition camera.

2.2.3 Training Stage 4: Cross-Camera Triplet Creation and Fine-tuning

Figure 2.2 shows the triplet creation process. A triplet is formed by an anchor, a positive, and a negative sample. During optimization, the distance from the anchor to the positive sample should be minimized, while the distance to the negative sample should be maximized. Ideally, positive and negative samples should be hard-to-classify examples for the current model M as easy examples do not bring diversity to the learning process.

We initially select, as the anchor, one random sample in cluster c captured by camera cam_j. For each camera $cam_k \neq cam_j$ in cluster c, we sort all feature vectors from camera cam_k based on their distance to the anchor. The positive sample is then selected as the median feature vector. The median is considered instead of the farthest sample (the hardest example) to avoid selecting a noisy example. We do not choose an easy example (the closest one) to avoid slowing down the model convergence or even getting stuck on a local minimum. To select the negative sample, we first sort all feature vectors from camera cam_j belonging to other clusters $\neq c$ based on their distance to the anchor. As the negative sample, we pick the closest feature vector that has not been assigned yet to a triplet. In this way, we avoid selecting the same negative sample, which brings diversity to the triplets and alleviates the harmful impact if one of the negative samples shares the anchor's real identity.

For a cluster c_i with n_i cameras, we generate a total of $n_i - 1$ triplets with the same anchor. If we select m anchors for one camera in c_i, a total of $m(n_i - 1)$ triplets are created. Considering that this process is repeated for each camera in c_i, we have a total of $n_i m(n_i - 1)$ triplets for cluster c_i. Note that the triplets are created in an offline manner. The offline creation enables us to choose triplets considering a global view of the target

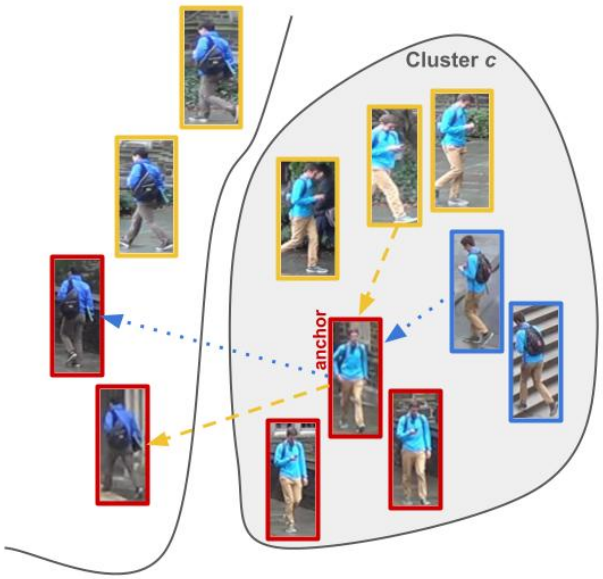


Figure 2.2: Cross-Camera Triplet Creation. For each selected cluster, we have at least two cameras. Suppose the represented cluster c has images from three cameras (represented with red, blue, and yellow contours). For each camera, we select m anchors. For each anchor, we create triplets with a positive sample from other cameras in the same cluster and a negative sample with the same camera in other clusters. For instance, for camera red, we select an anchor and we sort, based on the distance, all feature vectors from cameras yellow and blue. Then we select the median feature vector from each one (represented by the arrows coming to the anchor). To select the negative sample, we sort all feature vectors from the same camera but from a cluster $\neq c$, and we choose the closest and not previously selected sample. For the triplet with a yellow median sample as positive, we select as negative the closest sample to the red anchor from another cluster (represented by the yellow arrow leaving the anchor). For the triplet with a blue median sample as positive, we select the second closest feature vector to the red anchor from another cluster (since the first closest has already been picked). This explanation assumes $m = 1$ and is repeated for cameras yellow and blue.

data instead of creating them in a batch, which would bring a limited view of the target feature space.

The number m of anchors of a camera is the same for all clusters. Consequently, the number of triplets generated for a cluster c_i is $\mathcal{O}(n_i^2)$. The greater the diversity of cameras in a cluster, the greater its representativeness on the triplets. By emphasizing the clusters with more camera diversity during training, the model learns from easy-to-hard identities and is more robust to different viewpoints. In our experiments, we set $m = 2$ for all adaptation scenarios.

Due to this new approach of creating cross-camera triplets, we can optimize the model by using the triplet loss [156] without the need for weight decay or any other regularization term and hyper-parameters. This also suggests that cross-camera triplets help to regularize the model during training.

After creating the triplets in an offline manner, we optimize the model using the standard triplet loss function:

$$L = \frac{1}{|B|} \sum_{(x_a, x_p, x_n) \in B} \left[d(x_a, x_p) - d(x_a, x_n) + \alpha \right]_+ , \tag{2.1}$$

where B is a batch of triplets, x_a is the anchor, x_p is the positive sample and x_n is the

negative one. α is the margin that is set to 0.3 and $[.]_+$ is the $max(0, .)$ function. This is illustrated in Figure 2.1, Stage 4b.

2.2.4 Stage 5: Feature Extraction from Pseudo-Labeled Samples

This stage is part of the orange flow performed after Fine-tuning (Stage 4b). The main idea is to keep the pseudo-labeled clusters from Stage 3, recreating a new set of triplets based on the new distances between samples after the model update in Stage 4b, bringing more diversity to the training phase. To do so, we extract feature vectors only for samples of the pseudo-labeled clusters selected in Stage 3. The orange flow is performed K_2 times, and a complete cycle defines an epoch. The blue flow is performed every K_1 times and a complete cycle defines an iteration. Therefore, in each iteration, we have K_2 epochs. This concludes the training phase.

Unlike the five best state-of-the-art methods proposed in the prior art (DG-Net++, MEB-Net, Dual-Refinement, SSKD, and ABMT), our solution is trained with a single-term loss, which contains only one hyper-parameter. Even the weight decay has been removed, as the proposed method can already calibrate the gradient to avoid overfitting, as we show in Section 2.3. Moreover, prior work performs clustering on the training phase through k-reciprocal Encoding [245], which is a more robust distance metric than Euclidean distance. However, it has a higher computational footprint, as it is necessary to check the neighborhood of each sample whenever distances are calculated. For training simplicity, we opt for standard Euclidean distance to cluster the feature vectors. However, as k-reciprocal encoding gives the model higher discrimination, we adopt it during inference time. Therefore, different from previous works, we calculate k-reciprocal encoding only once during inference.

2.2.5 Self-ensembling

Our last contribution relies upon the curriculum learning theory. Different iterations of the training phase consider different amounts of reliable data from the target domain, as shown in Section 2.3. This property leads us to hypothesize that knowledge obtained at different iterations is complementary. Therefore, we propose to summarize knowledge from different moments of the optimization in a unique final model. However, as the model discrimination ability increases as more iterations are performed (the model is able to learn from more data), we propose combining the model weights of different iterations by weighting their importance with the amount of reliable data used in the corresponding iteration. We perform this weighted average of the model parameters as:

$$\theta_{final} = \frac{\sum\limits_{p \in P} p_i.\theta_i}{\sum\limits_{p \in P} p_i}, \tag{2.2}$$

where θ_i represents the model parameters after the i-th iteration and p_i is the weight assigned to θ_i. Weight p_i is obtained based on the reliability of the target domain; if more data from the target domain is considered in an iteration, it means that the model

is more confident, and then it can have more discrimination power on the target domain. Hence, p_i is equal to the percentage of reliable target data in the i-th iteration. Consequently, a model that takes more data from the target to train will have a higher weight p_i. Self-Ensembling is illustrated in Figure 2.3. **Note that we directly consider the models' learned parameters and create a new model by averaging these weights.** Therefore, we could even delete all the checkpoints after performing self-ensembling, as the combination is carried out at the parameters level.

We end up with a single model containing a combination of knowledge from different adaptation moments, which significantly boosts performance, as shown in Section 2.4.

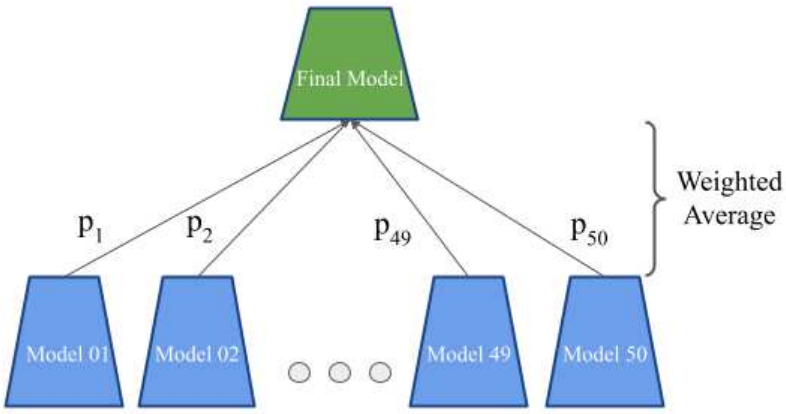


Figure 2.3: Self-Ensembling scheme after training. Different amounts of the target data (with no label information whatsoever) are used to fine-tune the model during the adaptation process. Different models created along the adaptation can be complementary. We create a new final model by weight averaging the models' parameters from different iterations. Weight p_i is based on the amount of reliable data from the target domain on the i-th iteration. We end up with a single model encoding knowledge from different moments of the adaptation.

2.2.6 Ensemble-based prediction

After training and performing the self-ensemble fusion, we have a single model adapted from the source to the target domain. However, due to the high performance of ensemble-based methods in recent ReID literature [47, 226], as a last measure, we leverage a combination of n_b different architectures to make a final prediction considering even more learned knowledge, which improves performance on the target dataset. We apply the ensemble technique only for inference, different from [47, 226] that leverage a mutual-teaching regime on training time. In turn, we avoid bringing complexity to the training but still take advantage of the complementarity from different architectures during inference.

To perform the ensemble-based prediction, we first calculate the feature distance of the query to each image on the gallery for each of the n_b final models. Let $f_k(x) = M_k(x)$ be the L2-normalized feature vector of image x obtained with model M_k and $d(f_k(q), f_k(g_i))$ be the distance between the feature vectors of the query q and of the i-th image gallery g_i extracted using the k-th model on ensemble. The final distance between query q and gallery image g_i is given by:

$$d_{final}(q, g_i) = \frac{1}{K} \sum_{k=1}^{K} d(f_k(q), f_k(g_i)), \tag{2.3}$$

where K is the number of models in the ensemble. In this way, we can incorporate knowledge from different models encoded as the distance between two feature vectors. After obtaining the distance between query q and all images in the gallery, we take the label of the closest gallery image as the query label.

We consider an equal contribution from each backbone. Without labels on the target domain, it is impossible to evaluate the impact of the individual models and give them proportional weights on the combination.

We picked three backbones that are commonly used in prior works. ResNet50 [58] and Dense- Net121 [70] have been employed in a ensemble-based strategy in [47, 226, 225], and OSNet [251] is a lightweight model commonly employed for Person Re-Identification.

2.3 Experiments and Results

This section presents the datasets we adopt in this work and compares the proposed method with the prior art with a comprehensive set of experiments considering different, and challenging, source/target domains.

2.3.1 Datasets

To validate our pipeline, we use three well-known large-scale datasets: `Market1501`, `DukeMTMC-ReID`, and `MSMT17`, which are described in Appendix B. As done in previous work in the literature, we remove from the gallery images with the same identity and camera of the query to assess the model performance in a cross-camera matching. Feature vectors are L2-normalized before calculating distances. For evaluation, we calculate the Cumulative Matching Curve (CMC), from which we report Rank-1 (R1), Rank-5 (R5), and Rank-10 (R10), and mean Average Precision (mAP).

2.3.2 Implementation details

In terms of deep-learning architectures, we adopt ResNet50 [58], OSNet [251], and Dense-Net121 [70], i.e., $n_b = 3$, all of them pre-trained on ImageNet [31]. To test them on an adaptation scenario, we choose one of the datasets as the source and another as the target domain. We train the backbone over the source domain and the adaptation pipeline over the target domain. We consider `Market1501` and `DukeMTMC-ReID` as source domains, leaving `MSMT17` only as the target dataset (the hardest one in the prior art). This way, we have four possible adaptation scenarios: `Market` $\rightarrow$ `Duke`, `Duke` $\rightarrow$ `Market`, `Market` $\rightarrow$ `MSMT17`, and `Duke` $\rightarrow$ `MSMT17`. We keep those scenarios (without `MSMT17` as a source) to have a fair comparison with state-of-the-art methods. Besides, the most challenging scenario is `MSMT17` as the target dataset: we train backbones on simpler datasets (`Market` and `Duke`) and adapt their knowledge to a harder dataset, with almost the double number of cameras and with many more identities recorded in different moments of the day and

the year. This enables us to test the generalization of our method in adaptation scenarios where the source and target domains have substantial differences in the number of identities, camera recording conditions, and environment.

We used the code available at [250] to train OSNet and at [226] to train ResNet50 and DenseNet121 over the source domains. Our source code is based on PyTorch [142] and it is freely available at `https://github.com/Gabrielcb/Unsupervised_selfAdaptative_ReID`.

After training, we remove the last classification layer from all backbones and use the last layer's output as our feature embedding. We trained our pipeline using the three backbones independently in all scenarios of adaptation. Considering the flows depicted in Figure 2.1, we perform $K_1 = 50$ cycles of the blue flow (50 iterations), and, in each one, we perform $K_2 = 5$ cycles of the orange flow (5 epochs). We consider Adam [88] as the network optimizer and set the learning rate to 0.0001 in the first 30 iterations. After the 30^{th} iteration, we divided it by ten and kept it unchanged until reaching the maximum number of iterations. As we show in our experiments, we can set the weight decay to zero since our proposed Cross-Camera Triplet Creation can regularize the model without extra hyperparameters. The triplet batch size is set to 30; batches with 30 triplets are used to update the model in each epoch. The margin in Equation 2.1 is set to 0.3, and the number of anchors is set to $m = 2$. We resize the images to $256 \times 128 \times 3$ and apply Random Flipping and Random Erasing as data augmentation strategies during training.

2.3.3 Comparison with the Prior Art

Tables 2.2 and 2.3 show results comparing the proposed method to the state of the art. The proposed method outperforms the other methods regarding mAP and Rank-1 in `Market` $\rightarrow$ `Duke` by improving those values in 1.8 and 1.7 percentage points (p.p.), respectively, and without re-ranking. In the `Duke` $\rightarrow$ `Market` scenario, we obtain a solid competitive performance by having values 0.1 p.p. lower only in Rank-1, also without re-ranking.

In turn, ABMT applies k-reciprocal encoding during training, which is more robust than Euclidean distance. However, it is more expensive to calculate as it is necessary to search for k-reciprocal neighbors of each feature vector in each iteration of the algorithm before clustering. In our case, we only apply the standard Euclidean distance during training, reducing the training time and complexity of adaptation, but still obtaining performance gains. Moreover, we have a single-term and single-hyper-parameter loss function, while ABMT depends on a loss with three terms and more hyper-parameters. They apply a teacher-student strategy to their training while we perform ensembling only for inference. Therefore, with a more direct pipeline and ensemble prediction, the proposed method has a Rank-1 only 0.1 p.p. lower in the `Duke` $\rightarrow$ `Market`, while outperforming all methods in all other adaptation scenarios.

However, to benefit from the k-reciprocal encoding, we also apply it during inference to keep a simpler training process. In this case, the proposed method outperforms the methods in the prior art regarding mAP and Rank-1 in all adaptation scenarios.

Compared to SSKD in `Duke` $\rightarrow$ `Market` scenario, we are below it by 0.3 and 0.4 p.p. in Rank-5 and Rank-10, respectively. Considering the closest actual gallery match image

to the query (R1), our ensemble retrieves more correct matches, as Table 2.2 shows, with our method outperforming SSKD by 1.2 p.p. in Rank-1 without re-ranking. Even with fewer hyper-parameters than SSKD and a more straightforward training process (no co-teaching, simpler loss function, and late ensembling), our method shows competitive results considering the training complexity trade-off.

Interestingly, the proposed method performs better under more difficult adaptation scenarios. We measure the difficulty of a scenario based on the number of different cameras it comprises. `Market`, `Duke`, and `MSMT17` have 6, 8, and 15 cameras, respectively. Hence the most challenging adaptation scenario is from `Market` to `MSMT17`. We adapt a model from a simpler scenario (6 cameras, all videos recorded in the same day period and the same season of the year) to a more complex target domain (15 cameras – 12 outdoors and 3 indoors – recorded at 3 different day periods – morning, afternoon and noon – in 4 different days – each day on a different season of the year). `Market` $\rightarrow$ `MSMT17` is the most challenging adaptation and close to real-world conditions where we might have people recorded throughout the day and in different locations (indoors and outdoors). In this case, as shown in Table 2.3, we obtained the highest performance even without re-ranking techniques. The proposed method outperforms the state of the art by 1.5 and 2.1 p.p. in mAP and Rank-1, respectively, on `Duke` $\rightarrow$ `MSMT17`, and by 2.2 and 4.2 p.p. on the most challenge scenario, `Market` $\rightarrow$ `MSMT17`.

There are several reasons why our method performs well. We explicitly designed a model to deal with the diversity of cameras and viewpoints by creating a set of triplets based on the different cameras in a cluster. We also keep a more straightforward training, with only one hyper-parameter in our loss function (triplet loss margin). Most works in the ReID literature optimize a loss function with many terms and hyperparameters. They usually consider the `Duke` $\rightarrow$ `Market` or the `Market` $\rightarrow$ `Duke` scenarios (or both of them) to perform grid-searching over hyper-parameter values. Once they find the best values, they keep them unchanged for all adaptation setups.

In ABMT [18], the authors do not provide a clear explanation on how they define the hyper-parameter values for their loss function. However, they perform an ablation study over `Duke` $\rightarrow$ `Market` and `Market` $\rightarrow$ `Duke` scenarios, so their results might be biased to those specific setups, which gives them one of the best performances. However, when they keep the same values for different and more challenging scenarios, such as `Market` $\rightarrow$ `MSMT17` or `Duke` $\rightarrow$ `MSMT17`, they obtain worse results than ours by a large margin. This shows that our method provides a better generalization capability brought by a simpler loss function and more diverse training. It prevents us from choosing specific hyper-parameter values and being biased to a specific adaptation setup. Consequently, we achieve the best performances, especially in the most challenging scenarios.

LOMO [102], BOW [242], and UMDL [145] are hand-crafted-based methods. They directly compute feature vectors over pixel values without using a neural network. UMDL also learns a shared dictionary to mine meaningful attributes from the target dataset, however, in a much simpler setup than any deep-learning method. They then calculate the distance between query and gallery images. This makes them scalable and fast deployable. However, since hand-crafted features usually do not describe high-level features from images, the methods fail when used to match the same person from different camera

Table 2.2: Results on `Market1501` to `DukeMTMC-ReID` and `DukeMTMCRe-ID` to `Market1501` adaptation scenarios. We report mAP, Rank-1, Rank-5, and Rank-10, comparing different methods. The best result is shown in **blue**, the second in **green**, and the third in **orange**. Works with (*) do not pre-train the model in any source dataset before adaptation.

		Duke → Market				Market → Duke			
Method	reference	mAP	R1	R5	R10	mAP	R1	R5	R10
LOMO [102]	CVPR'15	8.0	27.2	41.6	49.1	4.8	12.3	21.3	26.6
BOW [242]	ICCV'15	14.8	35.8	52.4	60.3	8.3	17.1	28.8	34.9
UMDL [145]	CVPR'16	12.4	34.5	52.6	59.6	7.3	18.5	31.4	37.6
PTGAN [194]	CVPR'18	-	38.6	-	66.1	-	27.4	-	50.7
PUL [42]	TOMM'18	20.5	45.5	60.7	66.7	16.4	30.0	43.4	48.5
MMFA [103]	ArXiv'18	27.4	56.7	75.0	81.8	24.7	45.3	59.8	66.3
SPGAN [33]	CVPR'18	22.8	51.5	70.1	76.8	22.3	41.1	56.6	63.0
TJ-AIDL [184]	CVPR'18	26.5	58.2	74.8	81.1	23.0	44.3	59.6	65.0
SPG+LMP [33]	CVPR'18	26.7	57.7	75.8	82.4	26.2	46.4	62.3	68.0
HHL [246]	ECCV'18	31.4	62.2	78.8	84.0	27.2	46.9	61.0	66.7
ATNet [109]	CVPR'19	25.6	55.7	73.2	79.4	24.9	45.1	59.5	64.2
CamStyle [249]	TIP'19	27.4	58.8	78.2	84.3	25.1	48.4	62.5	68.9
MAR [217]	CVPR'19	40.0	67.7	81.9	-	48.0	67.1	79.8	-
PAUL [209]	CVPR'19	40.1	68.5	82.4	87.4	53.2	72.0	82.7	86.0
ECN [247]	CVPR'19	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
ISSDA [169]	CVPR'19	63.1	81.3	92.4	95.2	54.1	72.8	82.9	85.9
PDA-Net [100]	ICCV'19	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
CR-GAN [23]	ICCV'19	54.0	77.7	89.7	92.7	48.6	68.9	80.2	84.7
PCB-PAST [233]	ICCV'19	54.6	78.4	-	-	54.3	72.4	-	-
UCDA [148]	ICCV'19	30.9	60.4	-	-	31.0	47.7	-	-
SSG [45]	ICCV'19	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
CASCL [197]	ICCV'19	35.5	65.4	80.6	86.2	37.8	59.3	73.2	77.8
SSL [106]*	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9
CCSE [105]*	TIP'20	38.0	73.7	84.0	87.9	30.6	56.1	66.7	71.5
UDAP [162]	PR'20	53.7	75.8	89.5	93.2	49.0	68.4	80.1	83.5
MMCL [179]	CVPR'20	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
ACT [207]	AAAI'20	60.6	80.5	-	-	54.5	72.4	-	-
ECN-GPP [248]	TPAMI'20	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4
HCT [223]*	CVPR'20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4
SNR [81]	CVPR'20	61.7	82.8	-	-	58.1	76.3	-	-
AD-Cluster [224]	CVPR'20	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
MMT [47]	ICLR'20	71.2	87.7	94.9	96.9	65.1	78.0	88.8	*92.5*
CycAs [193]*	ECCV'20	64.8	84.8	-	-	60.1	77.9	-	-
DG-Net++ [258]	ECCV'20	61.7	82.1	90.2	92.7	63.8	78.9	87.8	90.4
MEB-Net [226]	ECCV'20	76.0	89.9	96.0	97.5	66.1	79.6	88.3	92.2
Dual-Ref [28]	TIP'21	78.0	90.9	**96.4**	**97.7**	67.7	82.1	90.1	**92.5**
SSKD [111]	NDIC'21	**78.7**	91.7	**97.2**	**98.2**	67.2	80.2	**90.6**	**93.3**
ABMT [18]	WACV'20	**80.4**	**93.0**	-	-	**70.8**	**83.3**	-	-
Ours (w/o Re-Ranking)*	This Work	67.7	89.5	94.8	96.5	68.8	82.4	**90.6**	**92.5**
Ours (w/o Re-Ranking)	This Work	78.4	**92.9**	**96.9**	**97.8**	**72.6**	**85.0**	**92.1**	**93.9**
Ours (w/ Re-Ranking)	This Work	**88.0**	**93.8**	**96.4**	97.4	**82.7**	**87.2**	**92.5**	**93.9**

Table 2.3: Results on `Market1501` to `MSMT17` and `DukeMTMCRe-ID` to `MSMT17` adaptation scenarios. We report mAP, Rank-1, Rank-5, and Rank-10, comparing different methods. The best result is shown in **blue**, the second in **green**, and the third in **orange**. Works with (*) do not pre-train the model in any source dataset before adaptation.

		Duke → MSMT17				Market → MSMT17			
Method	reference	mAP	R1	R5	R10	mAP	R1	R5	R10
PTGAN [194]	CVPR'18	3.3	11.8	-	27.4	2.9	10.2	-	24.4
ECN [247]	CVPR'19	10.2	30.2	41.5	46.8	8.5	25.3	36.3	42.1
CCSE [105]*	TIP'20	9.9	31.4	41.4	45.7	9.9	31.4	41.4	45.7
SSG [45]	ICCV'19	13.3	32.2	-	51.2	13.2	31.6	-	49.6
ECN-GPP [248]	TPAMI'20	16.0	42.5	55.9	61.5	15.2	40.4	53.1	58.7
MMCL [179]	CVPR'20	16.2	43.6	54.3	58.9	15.1	40.8	51.8	56.7
MMT [47]	ICLR'20	23.3	50.1	63.9	69.8	22.9	49.2	63.1	68.8
CycAs [193]*	ECCV'20	26.7	50.1	-	-	26.7	50.1	-	-
DG-Net++ [258]	ECCV'20	22.1	48.8	60.9	65.9	22.1	48.4	60.9	66.1
Dual-Ref [28]	TIP'21	26.9	55.0	68.4	73.2	25.1	53.3	66.1	71.5
SSKD [111]	NDIC'21	26.0	53.8	**66.6**	**72.0**	23.8	49.6	63.1	68.8
ABMT [18]	WACV'20	**33.0**	**61.8**	-	-	27.8	55.5	-	-
SpCL [48]	NeurIPS'20	-	-	-	-	**31.0**	**58.1**	**69.6**	**74.1**
Ours (w/o Re-Ranking)	This Work	**34.5**	**63.9**	**75.3**	**79.6**	**33.2**	**62.3**	**74.1**	**78.5**
Ours (w/ Re-Ranking)	This Work	**46.6**	**69.6**	**77.1**	**80.4**	**45.2**	**68.1**	**76.0**	**79.2**

views. The substantial differences caused by changes in illumination, resolution, and pose of the identities bring a high non-linearity to the feature space that is not captured by hand-crafted-based methods. We surpass UMDL by 65.3 and 66.5 percentage points (p.p) on mAP and Rank-1 when considering `Market` → `Duke` scenario and by 66.0 and 58.4 p.p. considering `Duke` → `Market`. This shows the power of deep neural networks, which effectively describe identities in a non-overlapping camera system under different points of view.

MMFA [103] and TJ-AIDL [184] are methods based on low- and mid-level attribute alignment by leveraging deep convolutional neural networks. Since they do not encourage the networks to be robust to different points of view, their performance is lower than more recent proposed pseudo-labeling methods (PCB-PAST [233], SSG [45], UDAP [162], AD-Cluster [224], among others) and ensemble-based methods (ACT [207], MMT [47], MEB-Net [226], SSKD [111], ABMT [18]).

The same can be observed for PTGAN [194], SPGAN, and SPGAN+LMP [33], which are GAN-based methods that aim to transfer images from source to target domain, replicating the same camera conditions of the target domain in the labeled source images. However, transferring only camera-level features, such as color, contrast, and resolution, is not enough. People in the source domain might be in different poses and contexts from the ones in the target domain, and then those methods cannot fully describe images on the target domain considering these constraints. In more recent works, researchers have proposed further processing, such as pseudo-labeling (DG-Net++ [258]), pose alignment (PDA-Net [100]), and context-alignment (CR-GAN [23]). Our method can surpass all these GAN-based methods by a large margin. Compared to the most powerful of them, DG-Net++, we outperform it by 16.7 and 10.8 p.p on mAP and Rank-1 in the `Duke` →

Market scenario, and in Market → Duke by 8.8 and 6.1 p.p.

SpCL [48] is similar to ours in the sense that it increases the cluster reliability during the clustering stage as the training progresses. However, it does not apply any strategy considering diversity as we do by creating diverse triplets considering all cameras comprised in a cluster. Besides, they leverage both source and target domain images on adaptation stages and enable their model to use the source labeled identity to bring some regularization to the adaptation process. Differently, our method does not use the source domain images after fine-tuning and leverages the adaptation process relying only on target images. We outperform them by 1.2 and 4.2 p.p. in the most challenging Market → MSMT17 in mAP and Rank-1, respectively.

2.3.4 Discussion

As we aim to re-identify people in a camera system in an unsupervised way, we must be robust to hyper-parameters that require adjustments based on grid-searching using true label information, keeping the training process (and adaptation to a target domain) as simple as possible. If a pipeline is complex and too sensitive to hyper-parameters, it might be challenging to train and deploy it on a real investigation/biometric scenario, where we do not have prior knowledge about the people of interest. This complexity leads to sub-optimal performance. This has already been pointed out in [40]. The authors claim that most works rely on many hyper-parameters during the adaptation stage, which can help or hinder the performance, depending on the value assigned to them and which adaptation scenario is considered.

SSKD[111] is an ensemble-based method leveraging three deep models in a co-teaching training regime with a four-term loss function with three hyper-parameters. One of the terms of their final loss function is a multi-similarity loss [190], with three extra hyper-parameters to train the model.

MEB-Net has complex training by relying on a co-training technique with three deep neural networks in which each one learns with the others. Each of these three networks has its separate loss function with six terms, and their overall loss function is a weighted average of the individual loss functions from each model on the ensemble.

ABMT also leverages a teacher-student model where the teacher and student networks share the same architecture, increasing time and memory complexity during training. Moreover, they utilize a three-term loss function to optimize both models with three hyper-parameters controlling the contribution of each term to the final loss. They update the teacher weights based on the exponential moving average (EMA) of the student weights, in order to avoid error label amplification on training. This also adds another parameter to control the inertia in the teacher weights' EMA. The authors do not perform an ablation study regarding the hyper-parameter value variation to assess their impact on final performance.

Based on these observations, our proposed model better captures the diversity of real cases, by considering a loss function with a single term and that is less sensitive to hyper-parameters (only margin α needs to be selected). In such setups, it is difficult to select hyper-parameter values correctly, as we might not know any information about

the identities on the target domain. The self-ensembling also summarizes the whole training into a single model by using each checkpoint's confidence values over the target data, without using any hyper-parameter or human-defined value. Even adopting a more straightforward formulation, we still obtain state-of-the-art performance on the `Market` $\rightarrow$ `Duke` scenario and competitive performance on the `Duke` $\rightarrow$ `Market` scenario. Each architecture in our work is trained in parallel without any co-teaching strategy. After self-ensembling, the joint contribution from different backbones is applied only on evaluation time, avoiding label propagation of noisy examples (e.g., potential outliers) but still taking advantage of the complementarity between them.

Our assumptions are the same as recent prior art [224, 248, 105]. We assume to know from which camera an image of a person was recorded but not the identity. We rely on camera information to filter out cluster elements captured by only one camera and create the cross-camera triplets.

We also assume that at least two cameras have captured most identities and all of them have non-overlapping vantage points. All prior art holds this assumption as defined by the datasets and train/test split division.

Finally, we assume that training on a source domain related to Person Re-Identification gives the model basic knowledge to adapt to the target domain. This knowledge enables the model to propose better initial clusters on early iterations, grouping feature vectors from the same identity recorded from different cameras. The pipeline starts the adaptation with more reliable pseudo-labels in the clustering step and progressively creates more clusters representing more identities on the target domain. All works hold this assumption in Table 2.2 that do not have the (*) after their name.

Section 2.3.5 shows that our pipeline still performs well even without pre-training in a source dataset. In other words, we take the backbone trained over ImageNet and directly apply it without any previous ReID-related knowledge. Even in this setup, we can achieve competitive performance.

2.3.5 Results in the Unsupervised Scenario

This section explores the possibilities of our method when not performing any pre-training on a source domain. The method starts with backbones trained over ImageNet directly. This is a harder case as we eliminate the possibility of having prior knowledge of the person re-identification problem. This requires the backbones to adapt to the target, not relying on any identity-related annotation coming from the source domain. Table 2.2 shows the results denoted by "Ours(w/o Re-Ranking)*". In this case, we keep $\xi = 0.05$ when `Duke` is the target, as in previous results, and $\xi = 0.03$ when `Market` is the target. The value $\xi = 0.05$ was too strict, leading to clusters with images from only one camera for the Market dataset. Section 2.4 presents a deeper analysis of different choices of ξ on the clustering process.

However, when we consider `Duke` as the target domain, the model without source pre-training is the third best. We lose 3.8 and 2.6 p.p. to the equivalent pre-trained model in mAP and Rank-1, respectively, and we lose 2.0 and 0.9 p.p. compared to ABMT, outperforming all other methods. This shows that, although our model is not completely

robust to the backbone initialization, it is still capable of mining discriminative features, even without pre-training, proving comparative or better results when compared to the state of the art.

The proposed method outperforms all others in the same conditions (no pre-training, denoted with a star in Table 2.2). The difference to the best one (CycAs) is 2.9 and 4.7 p.p. on mAP and Rank-1 when Market is the target, and in 8.7 and 4.5 p.p. on mAP and Rank-1 when Duke is the target.

We conclude that the previous training on a ReID source-related dataset is important to achieve a better performance on the task. However, when no ReID source domain is available, our methods can still provide competitive results, mainly in the more challenging scenario (`Duke` as target). Inspired by these findings, in Chapters 3 and 4, we introduce methods that do not employ any ReID source dataset or camera information.

2.3.6 Qualitative Analysis

We now provide qualitative analysis by highlighting regions of the top 10 gallery images returned for a given query image. The redder the color of a region, the more important it is to the ranking. As explained in Section 2.3.1, the correct matches always come from cameras different from the query's camera. The green contour denotes a true positive, the red contour a false positive, and the blue color the query image. We present successful cases (when the first gallery image is a true positive) and failure cases (when the first gallery image is a false positive) for each camera on `Market1501` and `DukeMTMC-ReID` datasets. `MSMT17` was not considered as the dataset agreement does not allow the reproduction of the images in any format.

In Figures 2.4 and 2.6, we observe a subset of the successful cases with the activation maps for the top-10 closest gallery images to the query. We adapted the implementation from [163] to visualize the activation maps. In both scenarios, we see that our model is able to find fine-grained details on the images, enabling it to correctly match the query to the gallery images. For instance, Figures 2.4c and 2.6c depict two successful cases on `Market` $\rightarrow$ `Duke` and `Duke` $\rightarrow$ `Market` scenarios, respectively. In both cases, we see that our model finds fine-grained details on the image leading to a correct match. Figure 2.4c shows the model focusing on the red jacket, even in a different pose and under occlusion (7^{th} and 10^{th} image from left to right). Figure 2.6c shows that the model can overcome pose changes of the query on a cross-view setup. The query only shows the person's back, but the closest image is a true match showing the person from the front. The same happens on the second closest image, where the identity has its back recorded by another camera; and on the fourth and fifth closest images, only the right side is captured. The third closest image not only records a different position of the query but also has a different resolution. As another example, in Figure 2.4d, the model tends to focus on shoes and parts of the head, while in Figure 2.4a the focus is on regions depicting hair and pants. There are some hard cases such as the ones in Figures 2.4h and 2.6f, where we see they are mainly caused by similar clothes, but the method is still able to recover at least the closest gallery image (Rank-1 image). We conclude that our method is able to distinguish the semantic parts of the body and soft-biometric attributes which are vital to

Person Re-Identification. It is also important to remember that the query and the correct matches are always from different cameras, which also confirms that our model is able to overcome different camera conditions, such as identity pose changes and resolution on cross-view cameras.

Despite the state-of-the-art performance and most of the successful cases, our solution is also prone to errors and mismatching under some circumstances. Figures 2.5 and 2.7 also depict some failure cases, showing the limitations of the method. The failures are mainly related to similar clothes or soft-biometric attributes. There are scenarios where people can have similar clothes, for instance in a school where the students have the same uniform, or in a hospital where doctors and nurses usually dress similarly. However, considering that `Market1501` is composed of people in front of supermarket stores, it is marginal the chance of them being dressed the same way, however, this is the situation shown in Figure 2.7b.

Another source of errors is strong occlusion. Figure 2.5a is an example where the person has been fully occluded by a car and an umbrella. As there is no person, the method does not have any specific region to focus on and then the gallery images are almost fully activated. In Figure 2.7d, the target identity is on a motorcycle together with another person, which led them to be in the same bounding box. In this case, the method erroneously retrieves images with no identity on them (distractor images on the gallery) or images with parts of a bike.

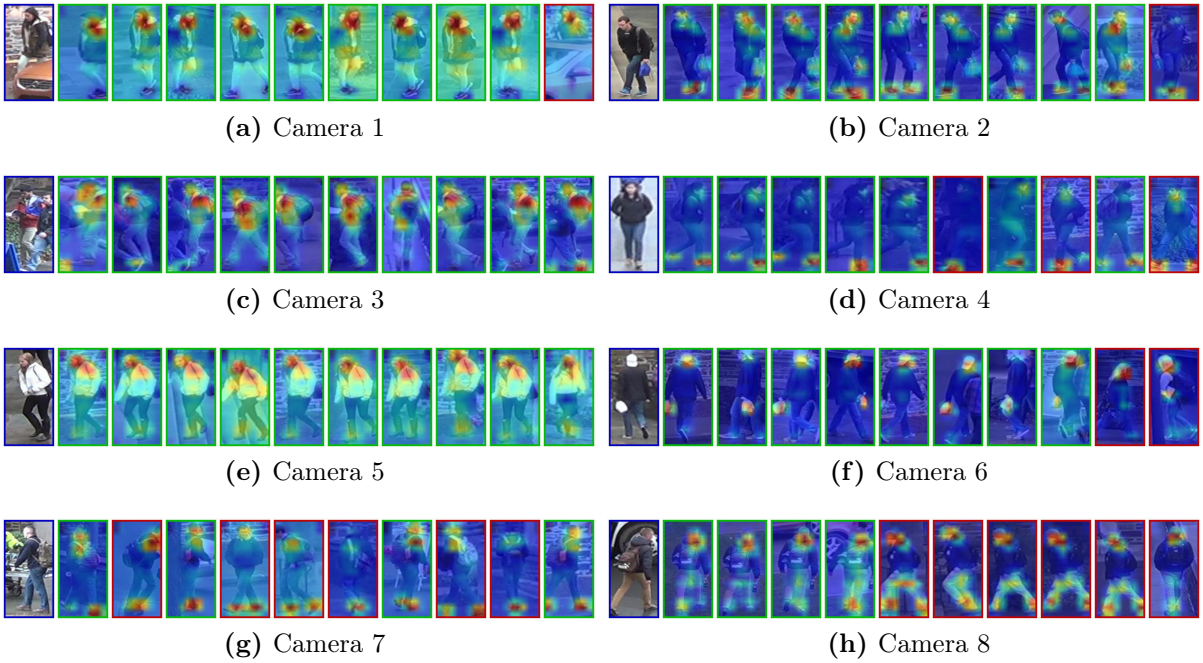
Figure 2.7a shows an interesting failure case, where the model focuses uniquely on the drawing on the person's shirt in the query image. The method returns gallery images of other identities with similar shirt drawings. Despite the failure, it is interesting to note that our method was able to focus on fine-grained details to find matches, and not activate the whole image or large parts of it.

2.4 Ablation Study

This section shows the contribution of each part of the pipeline to the final result. In each experiment, we change one of the parts and keep the others unchanged. If not explicitly mentioned, we consider ResNet50 as the backbone, OPTICS with hyper-parameter $\xi = 0.05$, and self-ensembling applied after training.

2.4.1 Impact of the Clustering Hyper-parameter

Although we have only one hyperparameter in the loss function, we still need to set hyperparameter ξ of the OPTICS clustering algorithm, which is a threshold in the range $[0, 1]$. The closer ξ is to 1, the stronger the criteria to define a cluster; that is, we might have many samples not assigned to any cluster, which leads to several detected outliers (if $\xi = 1$, all feature vectors are detected as outliers). In contrast, the closer ξ is to 0, the more relaxed the criteria, and more samples are assigned to clusters (if $\xi = 0$, all feature vectors are grouped into a single cluster). In Figure 2.8, we show the impact of the threshold ξ for the `Market` $\rightarrow$ `Duke` and `Duke` $\rightarrow$ `Market` scenarios.



(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

(e) Camera 5 **(f)** Camera 6

(g) Camera 7 **(h)** Camera 8

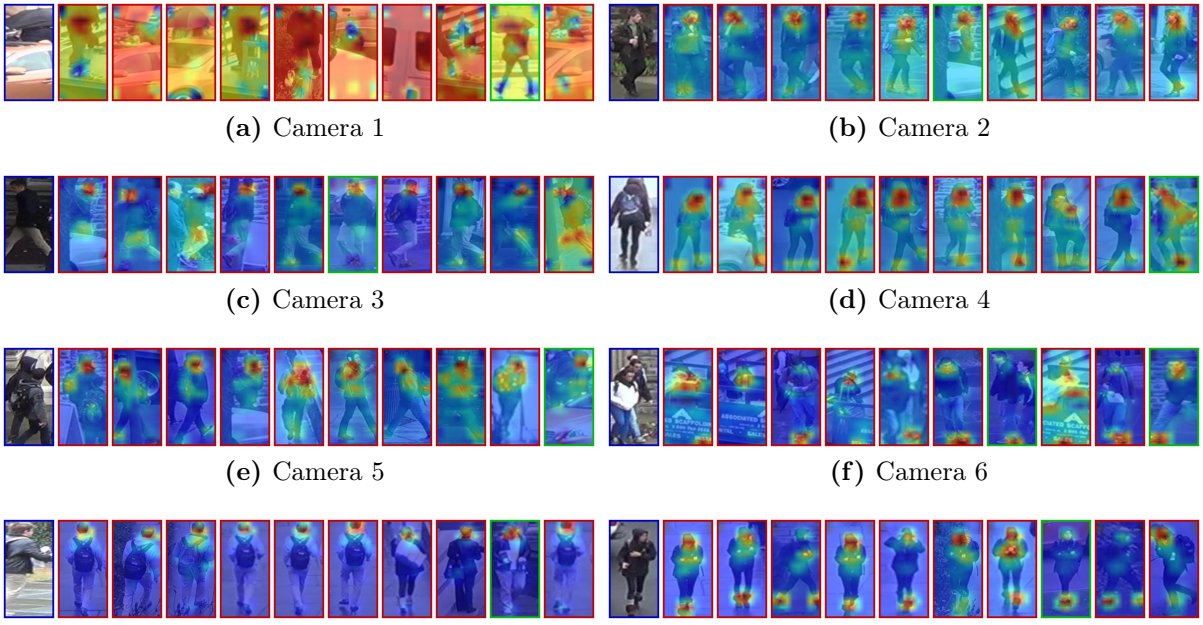
Figure 2.4: Successful cases considering one query from each camera on `Duke`. These results are obtained with the ResNet50 backbone after the `Market` → `Duke` adaptation.

The best value for ξ changes according to the adaptation scenario. This is expected when dealing with different unseen target domains. In both cases, Rank-1, Rank-5, and Rank-10 curves are more stable than the mAP curve, showing that the parameter does not impact the retrieval of true positive images. The best Rank-1 values are obtained for ξ between 0.04 and 0.08 considering both scenarios and, in the more challenging one (`Market` → `Duke`), it achieves the second-best value when $\xi = 0.05$, for both mAP and Rank-1. Although the best performance is achieved when $\xi = 0.07$ (best mAP and Rank-1), it relies on an unstable point in the setup of `Duke` → `Market`, and it is only marginally better than $\xi = 0.05$ for `Market` → `Duke`. Rank-5 and Rank-10 tend to be more stable in both cases. Thus we adopt $\xi = 0.05$ in all scenarios.

2.4.2 Impact of Curriculum Learning

In our pipeline, Stage 3 is responsible for cluster selection. After running the clustering algorithm, a feature vector can be an outlier, assigned to a cluster with only one camera, or assigned to a cluster with two or more cameras. We argue that feature space cleaning is essential for better adaptation, and that feature vectors in a cluster with at least two cameras are more reliable than ones assigned as outliers or clusters with a single camera. Then, we consider the curriculum learning principle to select the most confident samples and learn in an easy-to-hard manner. To achieve this, we remove the outliers and the clusters with only one camera. To check the impact of this removal, we performed four experiments in which we alternated between keeping the outliers and the clusters with only one camera. The results are summarized in Table 2.4.

We observe a performance gain on most metrics, especially on mAP and Rank-1, when we apply our cluster selection strategy. If we keep the outliers in the feature space



(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

(e) Camera 5 **(f)** Camera 6

(g) Camera 7 **(h)** Camera 8

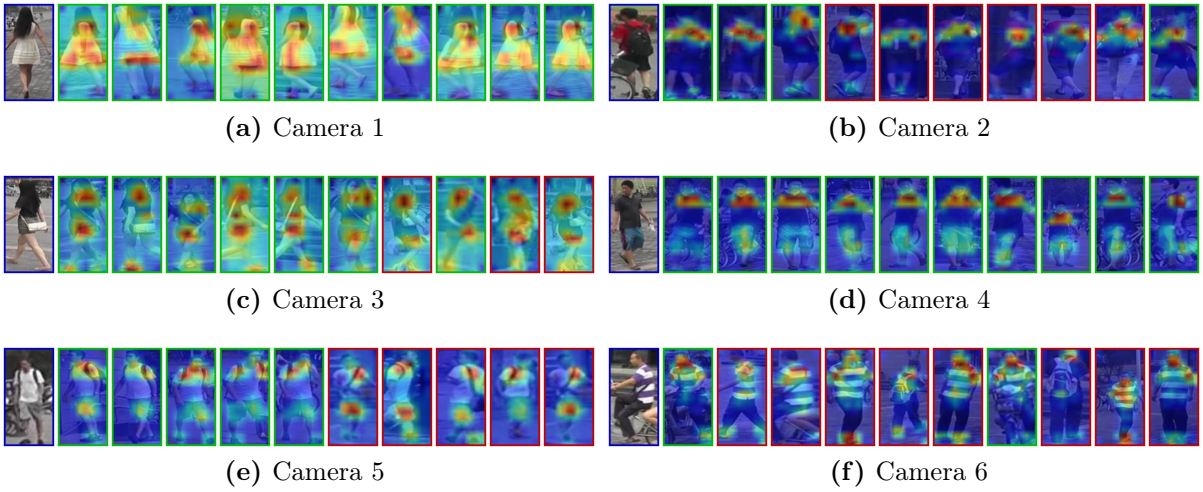
Figure 2.5: Failure cases considering one query from each camera on `Duke`. These results are obtained with the ResNet50 backbone after the `Market` → `Duke` adaptation.

Table 2.4: Impact of curriculum learning, when considering different cluster selection criteria. We tested our method with and without outliers and with and without clusters with only one camera in the feature space. All experiments consider ResNet50 as the backbone with self-ensembling applied after training.

		Duke → Market				Market → Duke			
w/o outliers	w/o cluster with one camera	mAP	R1	R5	R10	mAP	R1	R5	R10
-	-	50.9	79.2	89.5	92.8	32.7	56.7	68.5	72.9
✓	-	72.4	89.5	95.2	96.7	66.8	81.1	**90.2**	92.4
-	✓	49.1	79.8	89.5	92.6	32.7	57.2	68.4	72.3
✓	✓	**74.1**	**89.6**	**95.3**	**97.1**	**67.8**	**81.7**	90.0	**92.6**

(first and third rows in Table 2.4), we face the most significant performance drop in both adaptation scenarios. It shows the importance of removing outliers after the clustering stage; otherwise, they can be considered in the creation of triplets, increasing the number of false negatives (for instance, selecting negative samples of the same real class) and, consequently, hindering performance. We see a lower performance drop by keeping clusters with only one camera but without outliers (second row), indicating that those clusters do not hinder the performance much, but might contain noisy samples for model updating. It is more evident when we verify that the most gains were over mAP and lower gains over Rank-1 in the last row. This demonstrates that if we keep one-camera clusters, the model can still retrieve most of the gallery's correct images but with lower confidence. Hence, the cluster selection criteria effectively improves our model generalization and we apply it in all adaptation scenarios.

With this strategy, we observe that the percentage of feature vectors from the target domain kept in the feature space increases during the adaptation, as shown in Figures 2.9c and 2.10c. In fact, reliability, mAP and Rank-1 increase during training (Figures 2.9

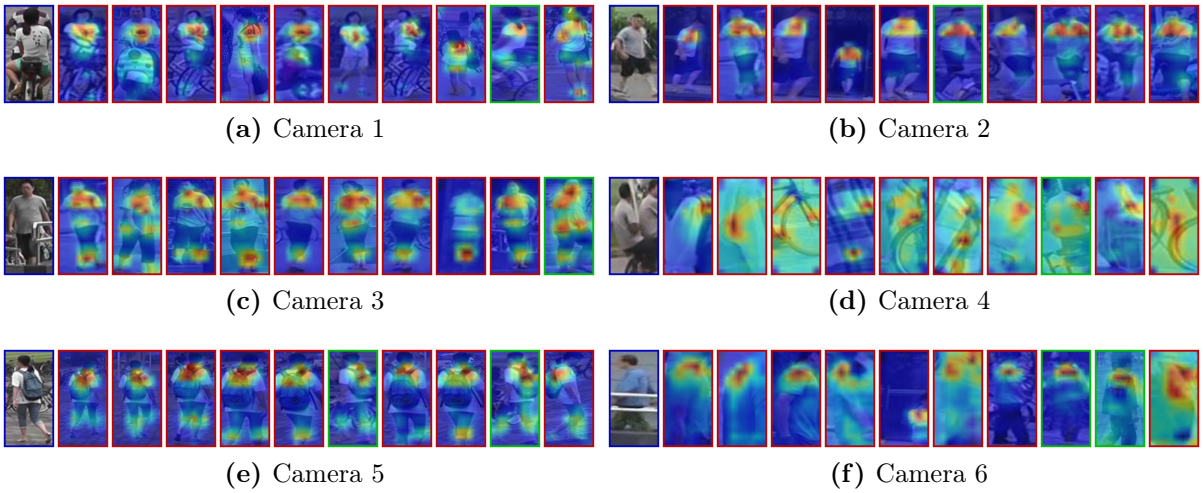


(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

(e) Camera 5 **(f)** Camera 6

Figure 2.6: Successful cases considering one query from each camera on `Market`. These results are obtained with the ResNet50 backbone after the `Duke` → `Market` adaptation.



(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

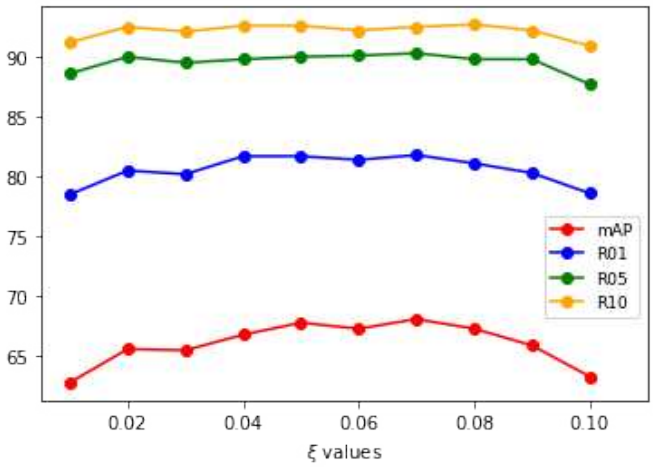
(e) Camera 5 **(f)** Camera 6

Figure 2.7: Failure cases considering one query from each camera on `Market`. These results are obtained with the ResNet50 backbone after the `Duke` → `Market` adaptation.

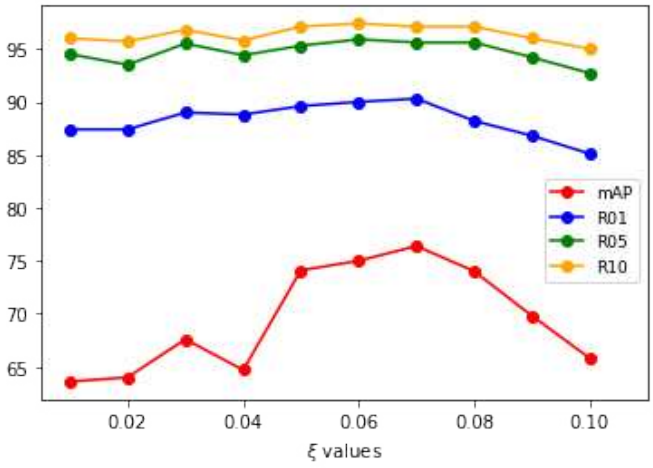
and 2.10), which means that the model becomes more robust in the target domain as more iterations are performed. This demonstrates the curriculum learning importance, where easier examples at the beginning of the training (images whose feature vectors are assigned to clusters with at least two cameras in early iterations) are used to give initial knowledge about the unseen target domain and allow the model to increase its performance gradually.

As a direct consequence, the number of clusters with only one camera removed from the feature space decreases, as shown in Figure 2.11. This means that the model learns to group cross-view images in the same cluster.

For the `Market` → `Duke` scenario, the initial percentage of removed clusters is higher than on `Duke` → `Market`. This is expected as the former is a more complex case, so initial clusters tend to have several images grouped due to the camera bias, which leads to a higher number of clusters comprising images recorded from only one camera. For the same



(a)



(b)

Figure 2.8: Impact of clustering hyper-parameter ξ. Results on (a) `Market` $\rightarrow$ `Duke`, and (b) `Duke` $\rightarrow$ `Market`.

reason, the final percentage for `Market` $\rightarrow$ `Duke` is higher than `Duke` $\rightarrow$ `Market`. In this last case, all backbones tend to stabilize between 20% and 30% of clusters removed in the last iterations. What if all identities are captured by only one camera? In this extreme case, we hypothesize that the model can still adapt to the target domain. However, the performance will be limited, as different identities could be grouped in the same cluster, increasing the false positive rate. This happens because one of our assumptions is that each identity should be captured by at least two cameras. In fact, this is inherited directly from the Person Re-Identification problem. Moreover, our method utilizes this assumption to create the triplets, enabling a better adaptation to the target domain.

2.4.3 Impact of self-ensembling

To check the contribution of our proposed self-ensembling method explained in Section 2.2.5, we take the best checkpoint of our model during adaptation in both scenarios, considering all backbones, and compare it with the self-ensembled model. Note that we

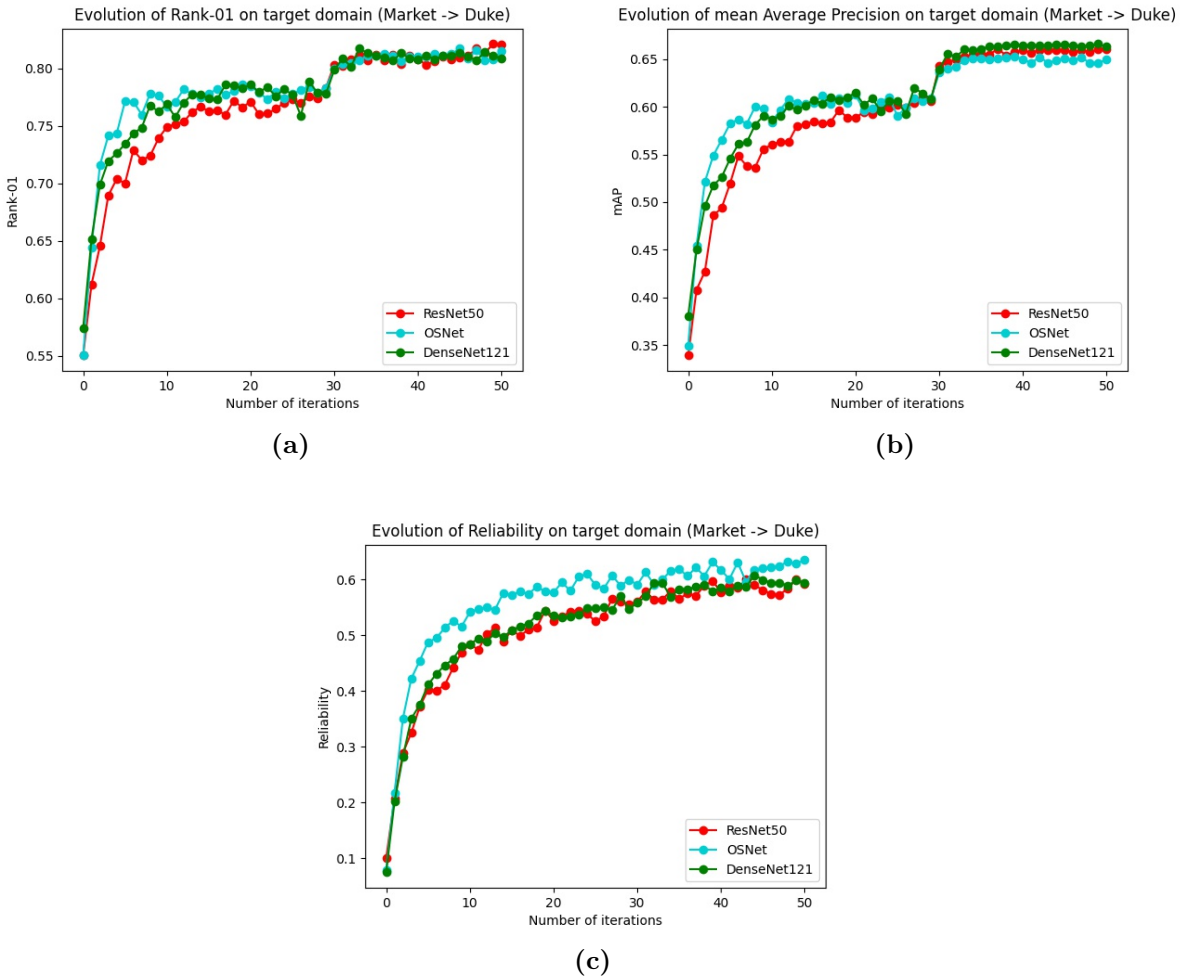


Figure 2.9: Progress on Rank-1, mean Average Precision and Reliability on target dataset, in the `Market1501` to `DukeMTMC-ReID` scenario.

select the best model only for reference. In practice, we do not know the best checkpoint during training since we do not have any identity-label information. Our goal here is merely to show that our self-ensembling method leads to a final model that outperforms any checkpoint individually. Even if we do not have any label information to choose the best one during training, the self-ensembling can summarize the whole training process in a final model, which is better than all checkpoints. Table 2.5 shows these results.

Our proposed self-ensembling method can improve discriminative power over the target domain by summarizing the whole training during adaptation. The method outperforms the best models in mAP by 2.0, 4.5, and 4.3 p.p., on `Duke` → `Market`, for ResNet50, OSNet, and DenseNet121, respectively. Similarly, for `Market` → `Duke` we achieve an improvement of 1.6, 2.2, and 3.3 p.p. in mAP for ResNet50, OSNet, and DenseNet121, respectively. We can also observe gains for all backbones in both scenarios considering Rank-1. Therefore, our proposed self-ensembling strategy increases the number of correct examples retrieved from the gallery and their confidence. It shows that different checkpoints trained with different percentages of the data from the target domain have complementary information. Besides, as the self-ensembling is performed at the parameter level, without human super-

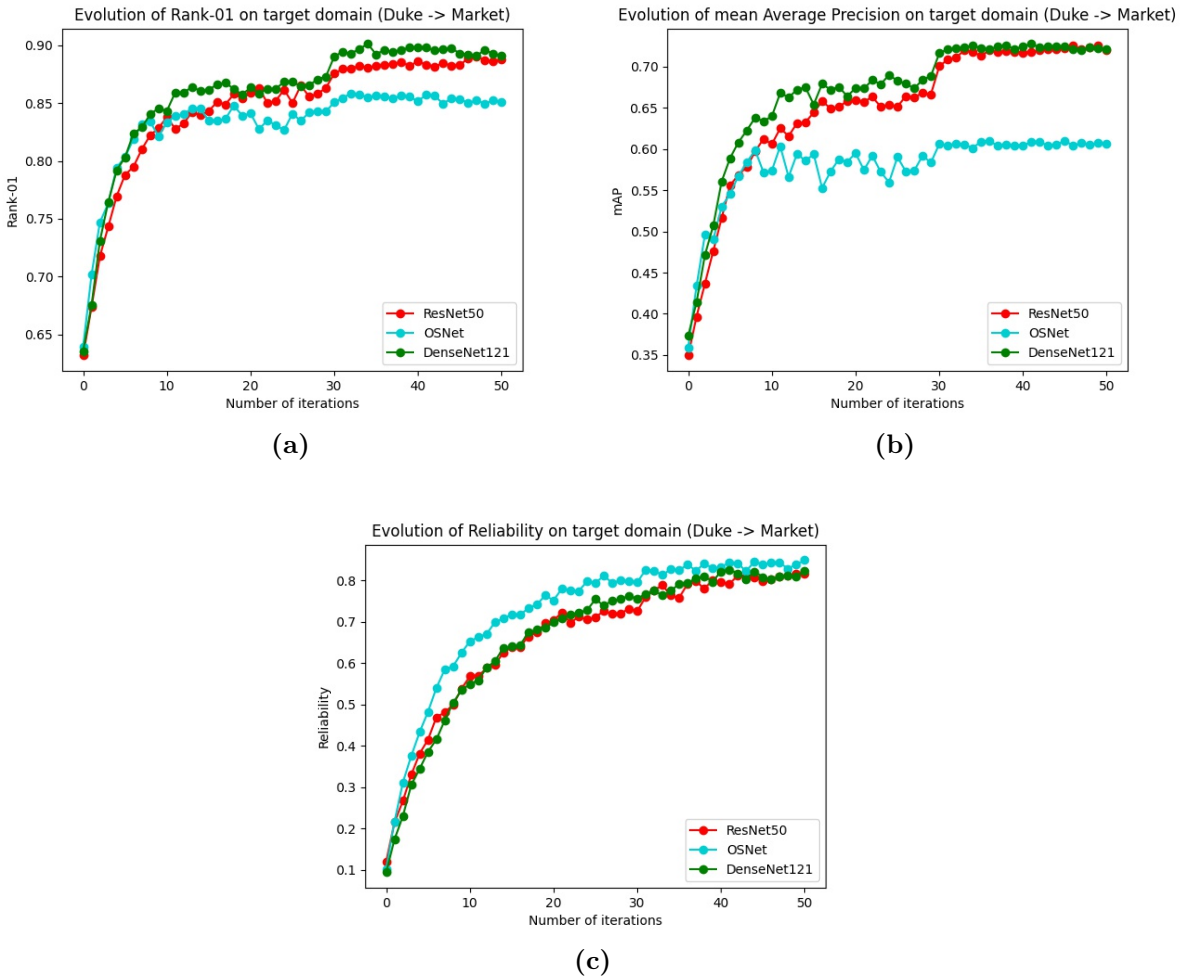


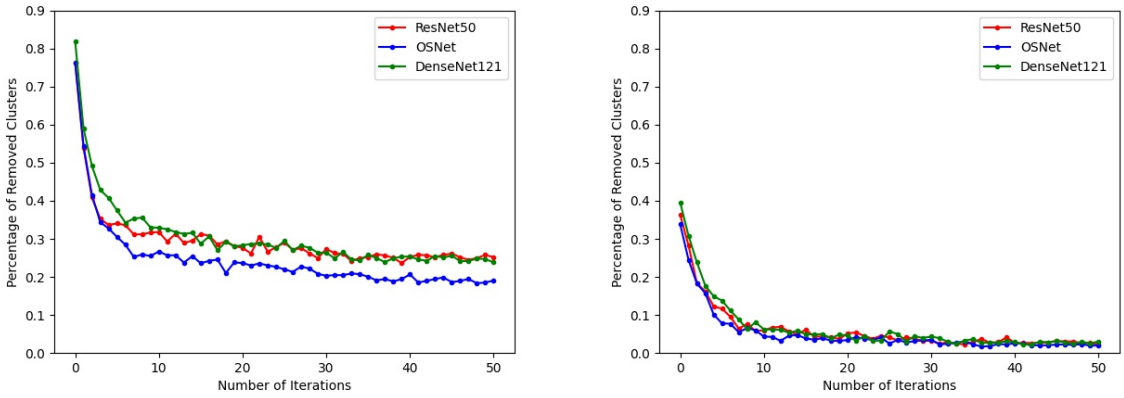
Figure 2.10: Progress on Rank-1, mean Average Precision and Reliability on target dataset on `DukeMTMC-ReID` to `Market1501` scenario.

vision and considering each checkpoint's confidence, it reduces the memory footprint by eliminating all unnecessary checkpoints and keeping only the self-ensembled final model.

2.4.4 Impact of Ensemble-based prediction

To increase discrimination ability, we combine distances computed by all considered architectures (Equation 2.3) for the final inference. Results are shown in Table 2.6.

The ensemble model outperforms the individual models by 3.3, 5.2, and 0.9 p.p. regarding Rank-1, on `Duke` $\rightarrow$ `Market`, for ResNet50, OSNet, and DenseNet, respectively. The same can be observed for `Market` $\rightarrow$ `Duke`, in which Rank-1 is improved by 3.3, 2.9, and 1.6 p.p. for ResNet50, OSNet, and DenseNet121, respectively. Results for all the other metrics also increase for both adaptation scenarios. Therefore, we can effectively combine knowledge encoded in models with different architectures. By performing it only for inference, we keep a simpler training process and still can take advantage of the ensembled knowledge from different backbones.



(a) **(b)**

Figure 2.11: Percentage of cluster removed along the training iterations on (a) `Market` → `Duke` and (b) `Duke` → `Market` scenarios considering the three backbones trained independently.

Table 2.5: Impact of self-ensembling. We consider a weighted average of the parameters of the backbone in different moments of the adaptation. "Best" refers to results obtained with the checkpoint with the highest Rank-1 during adaptation. "Fusion" is the final model created through the proposed self-ensembling method. The best results are in **blue**.

	Duke → Market				Market → Duke			
	mAP	R1	R5	R10	mAP	R1	R5	R10
ResNet (Best)	72.1	89.0	**95.5**	**97.1**	66.2	81.5	89.5	92.2
ResNet (Fusion)	**74.1**	**89.6**	95.3	**97.1**	**67.8**	**81.7**	**90.0**	**92.6**
OSNet (Best)	60.7	85.8	93.5	95.9	65.1	81.7	90.3	92.1
OSNet (Fusion)	**65.2**	**87.7**	**94.8**	**96.6**	**67.3**	**82.1**	**90.5**	**92.4**
DenseNet (Best)	72.6	90.1	95.6	97.1	66.0	81.7	90.1	92.4
DenseNet (Fusion)	**76.9**	**92.0**	**96.5**	**97.7**	**69.3**	**83.4**	**91.3**	**93.0**

2.4.5 Processing footprint

To measure the processing footprint of our pipeline (training and inference), we consider two representative adaptation scenarios: `Market` → `Duke` and `Market` → `MSMT17`. As explained, the first setup represents a mildly difficult case and the second is the most challenging one. Table 2.7 shows the time measurements.

The overall time to execute the pipeline and the whole training on `Market` → `Duke` scenario is smaller than `Market` → `MSMT17`'s, as expected, given that the latter is a more complex setup. As the number of training images is higher, the number of proposed clusters is also higher on `MSMT17`. This leads to an increase in clustering, filtering, and overall training times.

OSNet is the backbone that takes less time on both adaptation setups, because of its feature embedding size. For ResNet50 and DenseNet121, the embeddings have 2,048 dimensions while OSNet has 512. This allows a faster clustering, as Table 2.7 shows. Considering the same adaptation scenario, the clustering step is the most affected by the backbone and its respective embedding size. This is why ResNet50 and DenseNet121 present more similar training times and OSNet is the fastest one.

Table 2.6: Impact of ensemble-based prediction. Performance with and without model ensemble during inference. The best values are in **blue**.

	Duke → Market				Market → Duke			
	mAP	R1	R5	R10	mAP	R1	R5	R10
ResNet (Fusion)	74.1	89.6	95.3	97.1	67.8	81.7	90.0	92.6
OSNet (Fusion)	65.2	87.7	94.8	96.6	67.3	82.1	90.5	92.4
DenseNet (Fusion)	76.9	92.0	96.5	97.7	69.3	83.4	91.3	93.0
Ensembled model	**78.4**	**92.9**	**96.9**	**97.8**	**72.6**	**85.0**	**92.1**	**93.9**

Table 2.7: Time Evaluation. We calculate each time in HH:MM:SS for training, in MM:SS for each step, and in milliseconds (ms) for inference. On training, we analyze the time taken to cluster and filter (Stages 2 and 3), one round of fine-tuning (Finet. - Stage 4b), one epoch (time taken to perform K_2 iterations of orange flow), and the whole pipeline training. On inference (Infer.), we calculate the time to predict the identity of a query image given the gallery feature vectors.

	Market → Duke					Market → MSMT17				
	Clustering + filtering	Finet.	Epoch	Whole Training	Infer.	Clustering + Filtering	Finet.	Epoch	Whole training	Infer.
ResNet	03:55	08:55	13:34	11:31:19	5ms	16:45	09:36	28:08	23:00:55	13ms
OSNet	01:53	08:56	11:14	09:33:04	4ms	07:41	12:20	20:59	17:49:40	11ms
DenseNet	04:06	08:33	13:36	11:33:14	4ms	16:46	11:27	31:13	26:32:08	13ms
Ensemble	-	-	-	-	6ms	-	-	-	-	22ms

The inference time is calculated assuming that all gallery feature vectors have been extracted and stored. It is the average time to predict the label of one query based on the ranking of the gallery images, following the protocol presented in Section 2.3.2. The difference between both adaptation scenarios is due to the gallery size. As explained in Section 2.3.1, `MSMT17` has a gallery size more than 4× bigger than `Duke`'s.

For all experiments, we used two GTX 1080 Ti GPUs. One of them is used exclusively for clustering with an implementation based on [128], and the other for pipeline training, for each backbone.

2.5 Final Remarks

In this chapter, we presented the first solution proposed in this research. We tackle the problem of cross-domain Person Re-Identification (ReID) with non-overlapping cameras, especially targeting forensic and biometric scenarios with fast deployment requirements. We propose an Unsupervised Domain Adaptation (UDA) pipeline, with three novel techniques: (1) cross-camera triplet creation aiming at increasing diversity during training; (2) self-ensembling, to summarize complementary information acquired at different iterations during training; and (3) an ensemble-based prediction technique to take advantage of the complementary knowledge from different trained backbones.

Our cross-camera triplet creation technique increases invariance to different points of view and types of cameras in the target domain, and increases the regularization of the model, allowing the use of a single-term single-hyper-parameter triplet loss function. Moreover, we showed the importance of having this more straightforward loss function. It is less biased towards specific scenarios and helps us achieve state-of-the-art results in

the most complex adaptation setups, surpassing prior art by a large margin in most cases.

The self-ensembling technique helps us increase the final performance by aggregating information from different checkpoints throughout the training process, without human or label supervision. This is inspired by the reliability measurement, which shows that our models learn from more reliable data as more iterations are performed. Furthermore, this process is done in an easy-to-hard manner to increase model confidence gradually.

Finally, our last ensemble technique takes advantage of the complementary knowledge from different backbones, enabling us to achieve state-of-the-art results without adding complexity to the training, differently from the mutual-learning strategies used in current methods [226, 111, 18]. It is important to note that both ensembling strategies are done after training to generate a final model and a final prediction.

Because the training process is more straightforward than other state-of-the-art methods and does not need information on the target domain's identities, our work is easily extendable to other adaptation scenarios and deployed in actual investigations and other forensic and biometrics contexts, such as Vehicle Re-Identification.

A key aspect of our method also shared with other recent methods in the literature [197, 224, 248], is that it requires information about the camera used to acquire each sample. That is, in the presented solution we suppose we know, *a priori*, the device that captured each image. This information does not need to be the specific type of camera but, at least, information about different camera models. Without this information, our model could face suboptimal performance, as it would not be able to take advantage of the diversity introduced by the cross-camera triplets.

To address this drawback, we designed our second solution presented in the next chapter, which keeps the same three backbones employed in this first solution (ResNet50, OSNet, and DenseNet121) but without considering any camera or point-of-view information. In other words, our second solution relies solely on the bounding boxes without any identity, camera, or side annotation.

Regarding the clustering process, our first solution presented in this chapter requires that all selected samples are considered during this phase, which demands pairwise distance calculation between all feature vectors. Therefore, this approach may introduce higher processing times to the pipeline if large-scale datasets are employed.

Both the large-scale limitation and applicability to other Re-Identification tasks were addressed in the third solution proposed in this research (Chapter 4). It also relies solely upon person and vehicle bounding boxes without further annotation, and it is more suitable for large-scale datasets.

Chapter 3

Leveraging Ensembles and Self-Supervised Learning for Fully-Unsupervised Person Re-Identification and Text Authorship Attribution

In the previous chapter, we presented a method to tackle the Unsupervised Person Re-Identification task that requires camera annotation for each person bounding box, that is, we do not "who" is on a given bounding box but we assume we know which camera captured that image. In addition, we also employ backbones (Deep Convolutional Neural Networks) that have been trained in a supervised way in some labeled source datasets in order to obtain initial knowledge about the Person Re-Identification task before adaptation to the unlabeled target domain. Despite the best performances among similar competitors, it still faces limitations to realistic scenarios as it requires labeled source datasets for initialization and camera annotation.

In some real-world forensic and biometrics applications, the camera annotation might not be available, or we can have moving cameras such as smartphone cameras, cameras on drones, and wearable devices, which results in different capturing conditions for the same camera. In this situation, the camera annotation might be ambiguous since the same camera can have different viewpoints, resolutions, and changes in the background. While a camera identification method based on fingerprints in noise level could be employed, this approach may not fully consider environmental characteristics relevant to Person Re-Identification tasks, such as illumination, background variation, and viewpoints. Furthermore, if the source dataset employed for model initialization is too different from the target one, the model might face a cross-domain disparity, and the adaptation might be suboptimal.

In this context, we designed our second solution to tackle the Unsupervised Person Re-Identification (U-PReID) task but without considering any task-related source dataset or any camera annotation. In other words, our second solution relies on backbones initialized

with general knowledge of ImageNet and trained solely with the people bounding boxes **without any further annotation**. This solution has a broader application in real-world scenarios and can be extended to other tasks beyond U-ReID, such as Text Authorship Attribution.

Text Authorship Attribution (TAA) faces similar challenges as U-PReID. TAA aims to recover, from a gallery, texts from the same author of a query text. Multiple authors writing a social media post about the same topic (e.g., politics, sports, economics, etc.) might use similar vocabulary, which can make the texts similar, resulting in high inter-class similarity. Conversely, the same author can write about different topics, which leads to high intra-class disparity. Moreover, we address a more challenging scenario than the usual evaluation scenario for TAA, since we adopt the open-set scenario where training and test sets are disjoint in terms of identities.

In this context, we propose a novel self-supervised learning approach to handling the fully unsupervised Person ReID and Text Authorship Attribution tasks, which requires a robust distance measure and a fine-grained analysis. We start by considering a common approach: **clustering steps** to propose pseudo-labels to unlabeled samples and **optimization steps** to update the model supervised by those pseudo-labels [48, 18, 7]. However, prior methods that consider this approach often overlook two aspects: the quality of the features and the choice of hyper-parameters for the clustering algorithm. If the features are not too descriptive, samples from different classes might end up closer in the feature space, leading them to be clustered together, increasing the number of false positives and ultimately hindering model updates. Even when the features are adequate, a bad choice of hyper-parameters for the clustering process might yield suboptimal groups.

To address these problems, we take inspiration from re-ranking techniques [245] to filter out false-positive samples. The proposed method starts by calculating pairwise distances for unlabeled samples, considering features extracted by M Deep Neural Networks (we refer to them as *backbones*). Those distances are normalized by considering the mutual neighbors in each of the M feature spaces. As a second distance refinement, we average the M distances between two samples, as each backbone can provide a complementary description. As mentioned in Chapter 2, recent works consider ensemble techniques for Unsupervised Domain Adaptation [47, 226]. These studies, however, apply mutual learning by leveraging complex loss functions with one backbone supervising the other, which brings complexity to the training process. Our method, in turn, ensembles models by only taking the average distances, allowing the amalgamation of complementary information from each manifold, but with a much more straightforward setup.

The second aspect is the choice of hyper-parameters for the clustering process. We take DBSCAN as our clustering algorithm and, instead of fixing a value for the ε parameter, we scan different clustering densities — the lower the value, the denser the cluster. If a sample is identified as an outlier in any of these levels, it is marked as an outlier in the other levels. By tracking the state of each sample (inliers and outliers) through clustering runs, we can produce clusters that connect different dense regions while disregarding noisy samples.

Our method is designed for two critical applications that operate with different types of data: Person re-identification (images) and Authorship Attribution for social media

(texts). To the best of our knowledge, we are the first to apply the same self-supervised learning method to different forensics task modalities (one in Computer Vision and another in Natural Language Processing) with minor adjustments.

The main contributions of our work are:

- An effective distance averaging strategy to combine distances between feature vectors generated by independent backbones, taking advantage of complementary information. This simplifies the training process, as previously proposed complex techniques are unnecessary.
- An ensemble-based clustering strategy in which we scan a set of hyper-parameter values and combine intermediate clustering results in a unique, more robust result. By doing this, we can connect dense regions without the effect of noisy samples.
- A novel self-supervised learning formulation that can be applied to different problems such as fully-unsupervised Person ReID and Text Authorship Attribution with minor adjustments.

Similarly to the first solution presented in the previous chapter, the second proposed solution was published in the IEEE T-IFS [8]. It was also presented in the journal track session of the IEEE International Joint Conference on Biometrics (IJCB) 2023[1], and in the InterForensics 2023[2], the largest Forensics conference in Latin America.

3.1 Related Work

In this section, we describe related methods for self-supervised learning, with a more detailed exploration of person re-identification and text analysis.

3.1.1 Self-Supervised Learning

Self-supervised learning is usually done by generating two or more views of the same sample through augmentation techniques. A contrastive loss is minimized to pull together different views of the same image while pushing original images of different classes apart.

MoCo [57] generates two random augmented versions for each image in a batch. One is fed to a key encoder, and the other to a query encoder for feature extraction. The features from the key encoder are added to a dictionary that stores features from previous batches. The method minimizes a contrastive loss to pull together both augmented versions while keeping other features in the dictionary apart. SimCLR [22] also adopts two augmented versions of each image but without a dictionary of features. Instead, it minimizes a contrastive loss by considering both augmented images as a positive pair and the other images (and their augmented versions) as negative.

[1] https://ijcb2023.ieee-biometrics.org/accepted-papers/
[2] https://interforensics.com/site/interforensics2023/apresentacao

SWaV [13] applies a multi-cropping strategy by considering two standard-resolution crops and several low-resolution crops for optimization, along with a clustering assignment. Like SWaV, Dino [14] adopts different levels of cropping to optimize a teacher-student loss function. They feed the global crops to the teacher network to get a final probability distribution used to supervise the student network, which is fed with local crops. For a deeper review of self-supervised learning models, we refer the reader to this survey [83].

These methods obtain competitive performance when compared to their supervised learning equivalents. However, as they are generally tested on ImageNet, they tend to fail on problems with high intra-class dissimilarity and inter-class similarity, such as person ReID and Text Authorship Attribution of short messages.

3.1.2 Unsupervised Person Re-Identification

To tackle unsupervised person re-identification, some methods rely on pre-training a model on a source ReID dataset to acquire prior knowledge of the problem. This model is then adapted to the unlabeled target domain. A review of those works was presented in Chapter 2 in which we presented our first solution operating in the same scenario. We highlight that, in our first solution, we considered ensembles only during evaluation. We generated cross-camera triplets using camera information of samples in the generated clusters. We also proposed a self-ensembling strategy in which the training of each backbone is summarized by weight averaging the checkpoints.

Instead of considering pre-training in the ReID domain, another set of methods relies on other pieces of information, such as camera labels (as we also did for our first solution). IICS [204] leverages intra-camera training by dividing samples into sets according to their camera labels and performing clustering on each one. A backbone is trained in a multi-task manner (one task per camera), and clustering is run for the whole dataset, grouping samples of the same identity seen from different cameras. CAP [185] performs global clustering by assigning pseudo-labels for each sample of the dataset. They obtain camera proxies on each cluster for intra- and inter-camera training. ICE [17] has two versions: camera-aware and camera-agnostic. The first one considers the camera proxy features similar to CAP. The second considers only the cluster proxy, obtained by averaging features regardless of the camera label. They use a proxy-based loss along with a hard- and a soft-instance loss.

Instead of camera labels, some works rely on tracklets. CycAs [193] aims to identify the same person in frames of a video for intra-sampling. They also find the same person in other videos by considering the overlapping field-of-view between two videos for inter-sampling. With both intra- and inter-sampling, they optimize the backbone to match the same person from different points of view. UGA [197] averages the features of the same person in the same tracklet and performs cross-camera feature association creating a Cross-View-Graph (CVG) to encourage the matching of tracklets of the same person from different points of view.

Other works assume that only person bounding boxes are available. These are considered fully unsupervised. ABMT [18] relies on source pre-training, but the authors also

present results when no prior knowledge is considered. SpCL [48] proposes a self-paced strategy that introduces metrics to measure cluster reliability: cluster independence and cluster compactness. If both are higher than predefined thresholds, the cluster is kept within the feature space. They also minimize the loss function considering cluster centroids and samples stored in feature memory. RLCC [230] refines clusters by a consensus among iterations. Pseudo-labels on a certain iteration are created by considering the ones generated on a previous iteration, keeping the training stable. CACL [96] proposes a strategy to suppress the dominant colors on images, providing a more robust feature description, and a novel pseudo-label refinement method. ISE [234] synthesizes novel feature examples from real ones to refine the sample distribution, aiming to generate clusters with a higher true positive rate, as well as avoiding subdivision of the samples from the same identity in different clusters. PPLR [26] employs a part-based model that creates feature spaces from different parts of the feature map. In each one, they calculate the nearest neighbors of each sample and propose a cross-agreement metric to refine the proposed pseudo-labels.

Compared to such methods, our method also does not rely on any extra information, requiring only the bounding boxes of the people in the dataset; therefore, it best fits this last category of methods. Nonetheless, prior art often relies on clustering methods with manually chosen optimal hyper-parameter values, which might be impractical when working with unlabeled datasets. Our method differs from the rest by proposing a clustering criterion, which alleviates the burden of choosing optimal hyper-parameters.

We summarize the pros and cons of the main Unsupervised Person Re-Identification works in Table 3.1. The main advantage of our method is that it operates in a fully unsupervised scenario without relying on dataset-specific clustering hyper-parameter tuning. It leverages ensemble-based feature spaces and clustering, and it is the only one that has been designed for multiple modalities (images and text).

3.1.3 Unsupervised Text Analysis

Text Analysis is another application that can be explored with unsupervised learning methods. The Natural Language Processing (NLP) community witnessed a significant development with the introduction of models based on Attention and the Transformer architecture [177].

BERT [34] is one of the most successful models, applying an encoder-only architecture to solve many NLP tasks. The authors propose a self-supervised pre-training regime using masked language modeling and next-sequence prediction tasks, followed by a fine-tuning step using supervised data. Several works followed BERT, proposing variations using more targeted data. One example is BERTweet [131], in which the authors propose an extension to deal with tweets (short messages from Twitter).

T5 [149] takes a step forward and proposes a single architecture to solve any NLP problem that can be modeled as a text-to-text task. They apply the architecture presented by Vaswani et al. [177] with small changes in the normalization, dropout, and embedding layers.

More specifically, Text Authorship Attribution (TAA) is the task of finding the author

of a text solely by analyzing the textual information. We can reframe the task similar to ReID, in which we aim to recover, from a gallery, texts from the same author of a query text. Despite most of the research in TAA being in a closet-set scenario, in this work, we follow the more realistic and challenging scenario of open-set where the model may have never seen texts from the actual author, like in ReID. Despite current methods achieving good results for lengthy texts, authorship attribution is still challenging for short texts [175].

Nowadays, Authorship Attribution over social media data is an extremely compelling and relevant scenario involving textual information. Recently, Kirkpatrick [89] presented an interesting episode that supports this statement and the problem's relevance in this paper's context. The author shows how two teams of researchers applied AI techniques over social media messages to find the authors of QAnon messages — an anonymous creator of far-right political conspiracy theories. Despite the success of applying the Authorship Attribution techniques, this episode also showed how difficult and challenging the task is. The researchers dealt with a small set of 13 suspects, a considerable supervised dataset of 100,000 words from QAnon, and at least 12,000 words from each suspect. Furthermore, parallel investigations outside the textual universe reinforced the results (e.g., messages' timestamps from the suspects claiming they had discovered the QAnon existence).

In this work, we target the Authorship Authorship task in a fully unsupervised way, considering a dataset of tweets (short text messages from the X - former Twitter - social media platform). We consider a challenging and less explored setup, in which the authors from the training set are unlabeled and disjoint from those in the test set. As our baseline, we consider AdHominem [9], an attention- and LSTM-based model for Authorship Attribution **originally trained in a supervised manner** using social media posts. The pros and cons of AdHominem are highlighted in Table 3.1.

3.2 Proposed Method

The training pipeline is composed of seven steps: feature extraction and neighborhood-based distance computation, ensemble-based clustering, learning rate update, proxy selection, batch creation, optimization, and Mean Teacher averaging. Figure 3.1 depicts an overview of these steps.

In the first step, features are extracted for each sample, considering different backbones. We compute pairwise distances between samples based on their neighborhood for each backbone and average them across backbones to obtain a more refined and unique distance matrix. The second step is the application of our ensemble-based clustering technique to obtain pseudo-labels for the samples. We perform the learning rate calculation in the third step. In the fourth step, a proxy feature vector is selected for each cluster and, in the fifth step, sample batches are created. This information is used during the sixth step, which is the optimization of each backbone, independently. In the last step, a Mean Teacher technique is used to combine the weights of a backbone over training steps in a *momentum* model, which is used later for inference.

Table 3.1: Comparison of the prior art and our method in terms of pros and cons.

Method	Pros	Cons
SSL [106]	Does not require clustering	Requires camera labels and relies solely on the k-NN of each image for positive samples mining
CCSE [105]	Performs clustering regularization to balance the number of samples in each clustering	Requires camera labels and artificial GAN-generated images that might include biases in model learning
MPRD [76]	Does not require clustering and explores the neighborhood of image features through Graph Convolutional Networks	Requires camera labels for augmentation, and the optimization of two dependents networks (no decoupling)
DSCE-MC [208]	Leverages a symmetric and dynamic cross-entropy loss and a camera-based meta training	Requires the camera labels and nearest neighbors-based outlier reassignment which might introduce noise
JVTC [95]	Employs local and global view on loss function to regularize model learning	Requires camera labels, artificial GAN-generated images that might include biases in the learning, and frame annotation for temporal consistency calculation
JGCL [19]	Generates images of the same person in different poses, which regularizes the model training	Requires camera labels and the training of a GAN together with the main backbone, which brings complexity to the training process
IICS [204]	Employs the AIBN to regularize model learning to achieve cross-view invariant features	Requires multi-task training per camera where the complexity grows depending on the number of cameras and pseudo-identities found per camera
CAP [185]	Regularizes the model with intra- and inter-camera losses	Requires camera labels and multiple proxies per cluster to be constantly updated in an epoch
UST [7]	Leverages a cross-camera triplet creation and self-ensembling technique for checkpoints summary	Requires camera labels and an offline generation of all the triplets before training
ICE [17]	Employs a soft consistency loss to be robust to augmentation and has a camera-agnostic version	Leverages camera labels in its best version, presents a loss function with many hyper-parameters, and it shows results with a specific clustering parameter ϵ for each dataset
PPLR [26]	Leverages a part-based guided label and loss refinement and presents a camera-agnostic version	Requires camera labels in the best version, and the part-based agreement and loss function are complex
Star-Dac [154]	Provides an analysis about the time each identity is recorded by each camera and how they transit from one to another	Requires camera and timestamp annotation for each frame and a high-complex spatiotemporal-based clustering
TAUDL [97]	Does not require clustering and performs cross-tracklet association in batch level	Requires camera labels and tracklet estimation, not considering any global view of the feature space
TSSL [196]	Does not require camera labels and leverages a distribution-aware cluster distance for clustering	Assumes tracklet annotation but lacks the clustering for a global view
UTAL [98]	Does not require clustering and leverages tracklet-based soft labels for learning	Requires camera and tracklet annotations, and the training complexity grows linearly with the number of cameras
CycAs [193]	Does not require clustering and leverages a self-adaptive temperature parameter	Requires camera and timestamp annotation, and information if two or more camera fields of view overlap
UGA [197]	Does not require clustering and leverages a Cross-View Graph for inter-camera tracklet association	Requires camera and tracking annotation, and a multi-task training that grows linearly with the number of cameras
BUC [104]	Does not require camera labels and leverages a diversity regularization term on clustering	Evaluates and gets the best checkpoint in the test set, which is unrealistic
GPUFL [167]	Does not require clustering nor camera labels	Requires to keep m+1 memory banks, where m is the number of parts of the feature, and only local-neighborhood mining is performed
MV-ReID [216]	Does not require camera labels and leverages multi-patch optimization	Requires calculating the feature distance for each extracted patch, and the best checkpoint over a validation set is selected, which is unrealistic
MMCL [179]	Does not require camera labels nor clustering	Does not consider a global view of the feature space for positive and negative mining
HCT [223]	Does not require camera labels	The best checkpoint over a validation set is selected, which is unrealistic, and the hyper-parameters might change for different datasets
ABMT [18]	Exploits both global average and max pooling for feature learning	The loss function is complex, with five terms
SpCL [48]	Does not require camera labels and reliable samples are selected through independence and compactness measures	Introduces more clustering hyper-parameters which might be challenging to tune for different datasets
RLCC [230]	Does not rely on camera labels for learning and leverages pseudo-label consensus across iterations	Same as SpCL, and adds one more parameter for pseudo-label generation; however, it is sensitive to it
CACL [96]	Does not rely on camera labels and performs ensemble learning	The best checkpoint is selected, which is unrealistic, and the clustering requires different optimal hyper-parameter values for each dataset
ISE [234]	Does not rely on camera labels and uses sample extension	Leverages dataset-specific clustering hyper-parameters
AdHominem [9]	Predicts if the same author wrote two short-message posts in social media	Relies on fully **supervised** training
Ours	**Does not rely on camera labels, nor tracklets, and no hyper-parameters need to be tuned for clustering. It exploits ensembles and is the only one employed for both image and text-based tasks**	**Requires the training of more than one model, but it can be performed in parallel.**

We use the term *iteration* to refer to one complete iteration of the pipeline (blue flow in Figure 3.1), and *epoch* to refer to when the proposed clusters are used for optimization in the current iteration (green flow in Figure 3.1). We perform K_1 iterations and K_2 epochs per iteration.

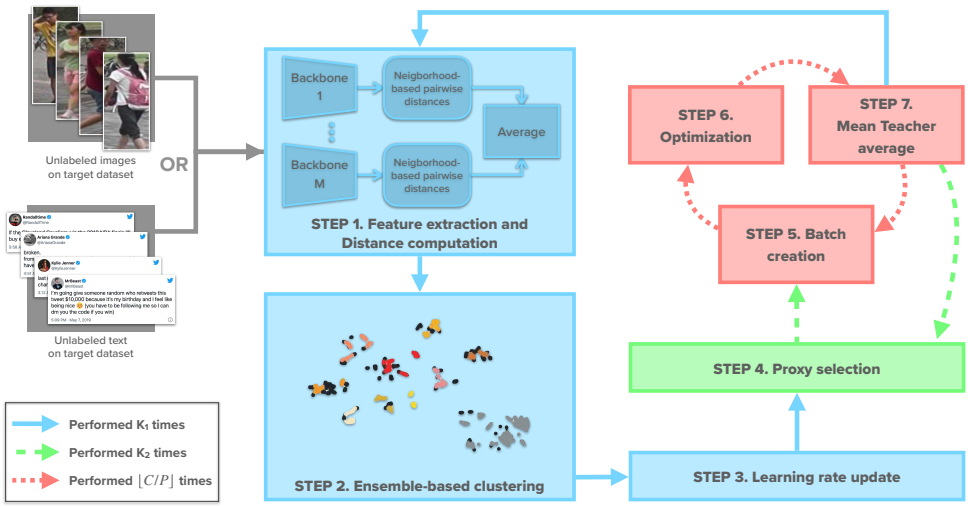


Figure 3.1: Overview of the proposed approach, comprising seven steps. Step 1: we extract feature vectors from all samples in the target dataset for each backbone, perform distance calculation followed by neighbor-based refinement, and combine the distances across backbones. Step 2: our ensemble-based clustering algorithm is performed to propose pseudo-labels and filter outliers. Step 3: we update the learning rate following a warm-up strategy [124]. Step 4: for each cluster obtained in Step 2, we randomly select a sample as a proxy. Steps 5, 6, and 7: a set of batches are created to optimize the backbones, each backbone is independently optimized and momentum models are updated based on the backbones' weights using a Mean Teacher strategy [173]. The red flow is performed $\lfloor C/P \rfloor$ times, where C is the number of clusters in the current iteration, and P is the number of clusters per batch. The cluster proxies are redefined K_2 times (green flow), after the red flow. The blue flow (entire pipeline) is performed K_1 times. Best viewed in color.

3.2.1 Step 1: Feature extraction and neighborhood-based distance computation

Consider a set $X = \{x_i\}_{i=1}^{N}$ of unlabeled data points in the target domain, consisting of N samples; and M backbones that generate feature representations for these samples.

In prior art for image representation, the output of the last global max or average pooling layer is commonly used as the final feature vector. However, global max and average pooling operations produce distinct and complementary descriptions and, when used together, they can increase the quality of the representation [18]. Following this idea, we perform both global max and average pooling after the last feature map and add the resulting vectors element-wise for the final feature vector (Figure 3.2). It is important to note that this is only done for images. For text representation, the output of the last layer is directly used as the final representation.

After extracting features for all samples, we L2-normalize them so that they are projected onto a unit hyper-sphere. Therefore, we have a set $F^m \in \mathbb{R}^{N \times d_m}$ of normalized feature vectors extracted by the m-th backbone, where d_m represents the dimension of features in this set.

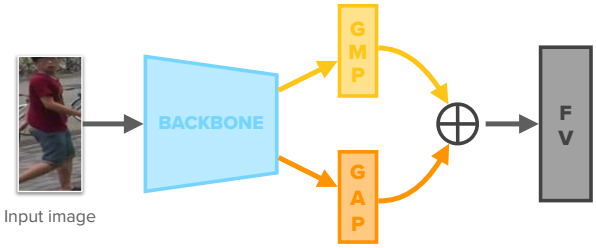


Figure 3.2: The final feature vector (FV) is obtained by extracting both global max pooling (GMP) and global average pooling (GAP) of the previous layer's output and then adding the two resulting vectors element-wise.

For each F^m, we calculate pairwise distances for all samples in the set. Inspired by re-ranking techniques [245], we refine these distances by considering the neighborhood of the samples; i.e., we normalize the distance between two samples based exclusively on the number of neighbors in common. Each distance is a number between 0 (no neighbors in common) to 1 (all neighbors in common). Consequently, we have M refined distance matrices $\{D_1, D_2, ..., D_m, ..., D_M\}$, one for each backbone. Distance matrix D_m is a general representation of the knowledge obtained by the m-th backbone over the input data, as it is calculated based on the samples' feature vectors. To explore potential complementary knowledge, we propose averaging all distance matrices:

$$\overline{\mathrm{D}} = \frac{1}{M} \sum_{m=1}^{M} D_m. \tag{3.1}$$

The final distance matrix $\overline{\mathrm{D}}$ is used as input to the proposed ensemble-based clustering technique.

3.2.2 Step 2: Ensemble-based Clustering

DBSCAN [41] clustering is the basis for our ensemble-based clustering. It relies on two hyper-parameters: $minPts$ – the minimum number of samples on a point's neighborhood to consider it as a core point – and ε – the radius of the neighborhood. Two data points p and q are considered neighbors if the distance between them is less than ε. A data point p is a core point if it has at least $minPts$ neighbors. If it has less than $minPts$ but is neighbor to a core point, then p is a border point. Otherwise, it is considered an outlier. Two points p and q are within the same cluster if a path exists $P = \{p_0, p_1, ..., p_n\}$, where $\forall_{1 \leq i \leq n-1}$ p_i is a core point, $p = p_0$ and $q = p_n$.

The performance impact of the two hyper-parameters has been studied [66, 18], and the conclusion is that DBSCAN is more sensitive to ε than to $minPts$. A wrong choice of ε can substantially hinder the performance, requiring domain knowledge to select its optimal value. A dataset with high intra-class variability might yield non-convex and sparse clusters, rendering the intra-class data points far away from each other while inter-class samples are closer. To account for this, a higher ε would be needed to group sparse samples in the same cluster. In turn, datasets with lower intra-class variability might require a lower ε, as a larger value could introduce false positives in the same cluster.

The described problem is common to several Person ReID benchmarks. `Market1501` [242]

is a dataset that comprises 751 identities recorded from six different cameras in the training set, while `MSMT17` [194] has 1,041 identities recorded from fifteen different cameras. Identities on `MSMT17` have larger intra-class variability than on `Market1501` as the number of different views of an identity is prone to be higher. Thus, different datasets require different values of ε, and this has also been pointed out in prior art. In [17], the authors use a lower value for Market1501 ($\varepsilon = 0.5$) and a larger value for MSMT17 ($\varepsilon = 0.6$) to account for dataset complexity. In [234], the authors adjust the value of ε to obtain better results. However, in a fully unsupervised scenario, it is impossible to select an optimal value for ε as there is no prior knowledge of the target data. Therefore, it is paramount to develop a clustering algorithm that does not depend on hyperparameter tuning.

We propose an ensemble-based clustering algorithm. As different values of ε yield different clusters, we run DBSCAN with different ε values and combine their results into a single final result. The proposed method effectively deals with noisy cases, allowing different but closer dense regions to be assigned to the same cluster, avoiding false positives and alleviating the burden of choosing the proper value for ε.

Considering the feature space defined by the refined distance matrix $\overline{\mathrm{D}}$ from Step 1, we perform DBSCAN with five ε values: 0.5, 0.55, 0.6, 0.65, and 0.7. As the neighborhood increases, more samples are assigned to a cluster. This does not mean that all samples are true positives and we need a way to detect false positives. If a sample has been detected as an outlier with $\varepsilon = 0.5$, it is kept as an outlier on further runs. We then can better filter out false positive samples while grouping closer dense true positive regions in the same cluster.

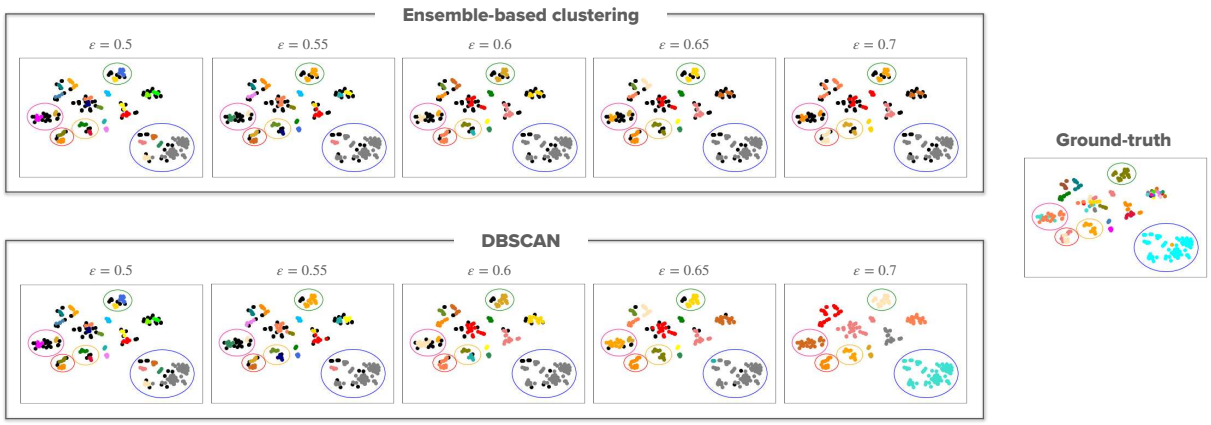


Figure 3.3: Comparison between our proposed ensemble-based clustering method and DBSCAN. The data points are identities randomly sampled from the `DukeMTMC-ReID` dataset and projected into a 2D space by T-SNE [176]. Data points of the same color represent an identity. Black points are outliers. Circles highlight regions of interest to be analyzed. In our ensemble-based clustering, the results for an ε value intrinsically combine results from all previous runs by keeping the data points' status if they were previously assigned as outliers or inliers. The last run ($\varepsilon = 0.7$) groups its results and all the previous ones, and produces more robust clusters than DBSCAN when we compare them to the final generated clusters with the ground truth. Best viewed in color.

Our ensemble-based clustering is illustrated and compared with the regular DBSCAN in Figure 3.3, where different colors represent different clusters (black represents outliers), and the circles highlight important regions to be analyzed. In the ground truth, the pink and red circles present noisy data with two identities mixed, which would be hard to

split with a single ε value. With $\varepsilon = 0.5$, DBSCAN yields the highest true positive rate, but it cannot group sparser identities such as the one within the green circle, which is subdivided into two clusters (blue and yellow). The same happens within the blue circle: most samples are from the same identity, but DBSCAN subdivides it into five clusters. In turn, our method is able to group results from lower ε values and create more robust clusters, as shown in Figure 3.3, $\varepsilon = 0.7$. We employed T-SNE [176] just for feature visualization in the lower dimensional space. The proposed solution works directly in the distances between data points after distance ensembling (Section 3.2.1), without any kind of dimensionality reduction.

One optimization is the reduction of DBSCAN runs. One must remember that if a point is detected as an outlier for $\varepsilon = 0.5$, it is kept as an outlier in subsequent runs with larger values. The same happens if a sample is an inlier for $\varepsilon = 0.5$; it will be an inlier in further runs as the neighborhood always increases. However, inlier samples assigned to different clusters can be put together in the same cluster due to the increasing neighborhood. For this reason, we say that the results for the j-th run contain the results for the i-th run, with $i \leq j$, as samples that are inliers or outliers keep their status in subsequent runs. It becomes clear that it is enough to run DBSCAN with $\varepsilon = 0.5$ and $\varepsilon = 0.7$ only, as the last run implicitly combines all intermediate results.

The main reason for choosing DBSCAN as the clustering algorithm is that it has the same or better resource consumption compared to other methods that have been used in cluster-based Unsupervised Person Re-Identification. Some previous works have employed Agglomerative Clustering, such as BUC [104], MV-ReID [216], and HCT [223]. Other works considered K-Means, such as PUL [42], MMT [47] MEB-Net [226]. We justify our choice in terms of three aspects: memory consumption, time complexity, and performance.

Memory consumption: Since our method relies on the Re-Ranking technique to refine the distance between samples based on neighborhood similarity, it is necessary to store a $N \times N$ distance matrix, where N is the number of data samples. This has memory complexity of $\mathcal{O}(N^2)$. All clustering methods require the full distance matrix, so all of them are upper-bounded by the same complexity.

Time Complexity: The Agglomerative Clustering [195] with the single-linkage variant has a time complexity of $\mathcal{O}(N^2)$. The K-Means clustering is well-known by the time complexity of $\mathcal{O}(NKT)$, where K is the pre-defined number of clusters and T is the number of iterations. As reported by [138], $T \propto N$ so its effective time complexity is $\mathcal{O}(N^2)$. The DBSCAN, on the other hand, as reported in the original paper [41], has a complexity of $\mathcal{O}(NlogN)$ in an optimal implementation, which makes it more efficient than Agglomerative Clustering and K-Means.

Performance: Agglomerative Clustering performance depends on the choice of the linkage criteria to link the closest clusters. It is also sensible to outliers because a single point can bridge two unrelated clusters and create large clusters with uncorrelated points. K-Means does not detect outliers since it enforces each point to belong to a cluster. It is also highly sensitive to the pre-definition of the number of clusters K and the selection of the initial clusters centers [138]. DBSCAN can detect outliers automatically. It creates clusters based on the reachability of the points, not depending on the number of clusters

a priori. It is only sensitive to the selection of the ε hyper-parameter, which is the main motivation behind the proposed ensemble-based clustering.

3.2.3 Step 3: Learning rate update

To update the learning rate, we consider a warmup strategy [124], which is effective mainly in the first training iterations, where the number of samples is lower than in further iterations. In the first iterations, there is still little data available for training, as the outliers detected by the clustering algorithm are in greater quantity and are discarded as noisy samples. In this context, the model is more prone to overfitting, and a lower learning rate can aid the training in such cases.

The warmup strategy consists of starting the training process with a small learning rate value and gradually increasing it along the first iterations. Based on [124], we define the learning rate at iteration t as

$$lr_t = \begin{cases} lr_{base} * \frac{t}{10}, & t \leq 10, \\ lr_{base}, & 10 < t \leq K_1, \end{cases} \tag{3.2}$$

where lr_{base} is a base value for the learning rate usually set to $3.5e^{-4}$, and $K_1 = 30$. The learning rate linearly increases in the first ten iterations and is constant for the remaining ones.

3.2.4 Step 4: Proxy selection

Once pseudo-labels are assigned to unlabeled data samples based on the clustering results, and the learning rate is adjusted, the backbones can be updated.

This process starts in Step 4, with the selection of cluster proxies, which are prototypes that represent the clusters. For each cluster, its proxy is the feature vector of a randomly selected sample within that cluster.

Although clusters tend to become more reliable as more iterations are performed, there is still a chance they might contain false positive samples. Assuming that the majority of samples are true positives, we hypothesize that a random selection is more likely to return a true-positive proxy than, for instance, computing the mean vector from all samples (true and false positives). We verify the impact of selecting a random or mean proxy in Section 3.4.

3.2.5 Step 5: Batch creation

The next step is batch creation. We consider the PK strategy [64], where we randomly select P out of C clusters generated in the current iteration, and K samples per cluster, creating batches of size $P * K$.

It is important that a cluster appears only once per epoch, exposing the optimization to more diversity. For this, we create $\lfloor \frac{C}{P} \rfloor$ batches, which can leave some clusters out of the current epoch because of the rounding. However, as we perform K_2 epochs per

iteration, the clusters not used in the current epoch will likely be selected in the next epochs.

3.2.6 Step 6: Optimization

The created batches are forwarded to the backbones for optimization. The loss function L to be minimized is composed of two other loss functions: L_{proxy} and L_{hard}.

For the m-th backbone, loss function L_{proxy} is based on cluster proxies p_j^m, $1 \leq j \leq C$ and is defined as

$$L_{proxy}(B;\theta_m) = -\frac{1}{|B|}\sum_{i=1}^{|B|} log\left[\frac{exp(f_i^m.p_+^m/\tau)}{\sum_{j=1}^{C} exp(f_i^m.p_j^m/\tau)}\right], \tag{3.3}$$

where B is the batch, $|B|$ is the batch size, θ_m are the weights of the m-th backbone in the current iteration, f_i^m is the feature vector of the i-th sample in B extracted by the m-th backbone, p_+^m is the proxy of the same cluster as the i-th sample in B, and τ is a temperature hyper-parameter to regulate the sharpness of the distribution of distances from the i-th sample to all proxies. The rationale is to approximate each sample in the batch from its respective proxy and keep it apart from the other proxies.

As hard sample mining has shown promising results in prior art [18, 17], we consider it by utilizing a hard instance-based softmax-triplet loss defined as

$$\begin{aligned} &L_{hard}(B;\theta_m) = \\ &-\frac{1}{|B|}\sum_{i=1}^{|B|} log\left[\frac{exp(f_i^m.f_+^m/\tau)}{exp(f_i^m.f_+^m/\tau) + exp(f_i^m.f_-^m/\tau)}\right], \end{aligned} \tag{3.4}$$

where f_+^m is the hardest positive sample in comparison to the i-th sample in B, i.e., it is the most distant feature from f_i^m within the same cluster in the current batch. Analogously, f_-^m is the hardest negative sample in comparison to f_i^m in the current batch, i.e., it is the sample closest to f_i^m but from another cluster.

The L_{hard} loss provides a local view as it only considers samples in the current batch, while L_{proxy} is a global loss since it considers all proxies from all clusters. The final loss combines them into a single function:

$$L(B;\theta_m) = L_{proxy}(B;\theta_m) + \lambda L_{hard}(B;\theta_m), \tag{3.5}$$

where λ is a hyper-parameter to control the impact of L_{hard}. We provide the sensitivity of our method to τ and λ values in Section 3.4.

Each backbone is trained independently but with the same set of pseudo-labels and learning rate obtained in previous steps. That is, the green flow in Figure 3.1 is performed for each of the M backbones, once at a time.

3.2.7 Step 7: Mean Teacher average

In this step, we leverage the Mean Teacher strategy [173], which averages model weights over training steps to produce a final, more accurate model. It computes a teacher model as the average of consecutive student models.

For each backbone (student model), we keep a teacher (or momentum) model with the same architecture. After each optimization step, the weights of a backbone are used to update the respective momentum model by Exponential Moving Average:

$$\Theta_m^{(t)} := \beta\Theta_m^{(t-1)} + (1-\beta)\theta_m^{(t)}, \tag{3.6}$$

where β controls the inertia of the momentum weights over training, t represents the iteration, and Θ_m and θ_m are the weights of the momentum and student models, respectively, that correspond to the m-th backbone.

3.2.8 Inference

After the training pipeline, inference is done by ranking all gallery samples based on the distance to a query sample. We extract feature vectors for all gallery and query sets using the momentum models from each backbone, which we denote as F_q^m and F_g^m, respectively, with $1 \leq m \leq M$. For each m, we calculate pairwise distances between samples of F_q^m and F_g^m, resulting in a distance matrix $D_{q2g}^m \in \mathbb{R}^{|Q|\times|G|}$, where $|Q|$ and $|G|$ are the number of samples in the query and gallery sets.

A final distance matrix $\overline{\mathrm{D}}_{q2g}$ is obtained by averaging all matrices element-wise:

$$\overline{\mathrm{D}}_{q2g} = \frac{1}{M}\sum_{m=1}^{M} D_{q2g}^m. \tag{3.7}$$

Each row of $\overline{\mathrm{D}}_{q2g}$ holds the distances from a query to the gallery samples. We sort these distances to infer the closest class to the query sample.

3.3 Experiments

We perform experiments to validate our self-supervised learning pipeline, considering two applications: Person ReID and Text Authorship Attribution from short text messages.

3.3.1 Datasets

For **Person ReID**, we use three well-known large-scale datasets: `Market1501`, `DukeMTMC-ReID`, and `MSMT17`, which are described in Appendix B. As done in previous ReID works, we remove images from the gallery with the same identity and camera of the query to assess performance in a true cross-camera scenario. For evaluation, we calculate the Cumulative Matching Curve (CMC), from which we report Rank-1 (R1), Rank-5 (R5), Rank-10 (R10), and mean Average Precision (mAP).

For **Authorship Attribution**, we adopt two subsets of a dataset of tweets [174], which is also described in Appendix B. Following the setup of the ReID validation, we keep disjoint authors for training and testing to verify the generalization capacity of the model. For evaluation, we compute mAP, R1, R5, and R10.

3.3.2 Implementation Details

We adopt $M = 3$ backbones. For person ReID, the backbones are well-known Deep Convolutional Neural Network (DCNN) architectures: ResNet50 [58], OSNet [251], and DenseNet121 [70], all of them previously trained over the ImageNet dataset [31]. For Authorship Attribution, we consider BERT [34], BERTweet [131], and T5 [149] architectures.

For optimization, we consider the Adam [88] optimizer with weight decay 0.00035. The learning rate is set following the behavior in Equation 3.2 with $lr_{base} = 0.00035$. We implement the neighborhood-based distance with re-ranking [245], which relies on two parameters: k_1 which defines the k-reciprocal neighborhood size, and k_2, which defines the neighborhood size to average the distance representation. Following prior art, we set them to $k_1 = 30$ and $k_2 = 6$. For batch creation using the PK technique, we set $P = 16$ and $K = 12$, totaling 192 samples per batch.

The values for ε in the proposed ensemble-based clustering are 0.5, 0.55, 0.6, 0.65, and 0.7. We keep $minPts = 4$ in all DBSCAN runs as done in the prior art. For the loss function, we set $\tau = 0.04$ and $\lambda = 0.5$ in Equations 3.3, 3.4, and 3.5. We analyze the sensitivity of our method to τ and λ in Section 3.4.

The pipeline (blue flow in Figure 3.1) is executed for $K_1 = 30$ iterations and the green flow is executed for $K_2 = 7$ epochs, for each set of proposed clusters, and $\beta = 0.999$ in Equation 3.6.

The training pipeline is implemented using PyTorch [142]. The evaluation part and the OSNet backbone are implemented on Torchreid [250]. We perform all experiments on three RTX5000 GPUs, each with 16 GB of RAM. One is used to perform re-ranking while the other is used to execute the whole training pipeline. The code is available at `https://github.com/Gabrielcb/Leveraging_ensembles_and_self_supervised_fully_ReID`.

3.3.3 Person ReID

We compare our pipeline applied to the U-PReID problem with relevant methods in the literature. The results are shown in Table 3.2, and the main pros and cons for each method are presented in Table 3.1.

We outperform all methods in the fully-unsupervised setup in the most challenging datasets, `Duke` and `MSMT17`, and obtain the second-best result in `Market` dataset, regarding mAP and R1. More specifically, we outperform the recent CACL method by 3.1 and 1.3 percentage points (p.p.) in mAP and R1, respectively, in the `Duke` dataset; and outperform ISE by 7.6 and 3.5 p.p. on `MSMT17`, the most challenging unsupervised ReID dataset. In the `Market` dataset, our results are the second best, considering mAP and R1. As we designed our method to tackle general and complex fully-unlabeled scenarios

Table 3.2: Comparison with relevant Person ReID methods considering three setups: camera-based, tracklet-based, and fully-unsupervised methods. Our work fits in the last category, which is the most challenging. We highlight the three best results in the fully unsupervised scenario: the best one in **blue**, the second best in **green**, and the third in **orange**. For the other categories, we only highlight the best result. Best viewed in color.

		Market				**Duke**				**MSMT17**			
Method	Reference	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Camera-based													
SSL [106]	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9	-	-	-	-
CCSE [105]	TIP'20	38.0	73.7	84.0	87.9	30.6	56.1	66.7	71.5	9.9	31.4	41.4	45.7
MPRD [76]	ICCV'21	51.1	83.0	91.3	93.6	43.7	67.4	78.7	81.8	14.6	37.7	51.3	57.1
DSCE [208]	CVPR'21	61.7	83.9	92.3	-	53.8	73.8	84.2	-	15.5	35.2	48.3	-
JVTC [95]	ECCV'20	47.5	79.5	89.2	91.9	50.7	74.6	82.9	85.3	17.3	43.1	53.8	59.4
JGCL [19]	CVPR'21	66.8	87.3	93.5	95.5	62.8	82.9	87.1	88.5	21.3	45.7	58.6	64.5
IICS [204]	CVPR'21	72.9	89.5	95.2	97.0	64.4	80.0	89.0	91.6	26.9	56.4	68.8	73.4
CAP [185]	AAAI'21	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8	36.9	67.4	78.0	81.4
CCTSE [7]	TIFS'21	67.7	89.5	94.8	96.5	68.8	82.4	90.6	92.5	-	-	-	-
ICE [17]	ICCV'21	82.3	93.8	97.6	98.4	**69.9**	**83.3**	**91.5**	**94.1**	38.9	70.2	80.5	84.4
PPLR [26]	CVPR'22	**84.4**	**94.3**	**97.8**	**98.6**	-	-	-	-	**42.2**	**73.3**	**83.5**	**86.5**
Tracklet-based													
S-Dac [154]	PR'21	33.9	67.0	80.6	84.9	31.6	56.4	72.1	76.5	-	-	-	-
TAUDL [97]	ECCV'18	41.2	63.7	-	-	43.5	61.7	-	-	-	-	-	-
TSSL [196]	AAAI'20	43.3	71.2	-	-	38.5	62.2	-	-	-	-	-	-
UTAL [98]	TPAMI'20	46.2	69.2	-	-	44.6	62.3	-	-	13.1	31.4	-	-
CycAs [193]	ECCV'20	64.8	84.8	-	-	**60.1**	**77.9**	-	-	**26.7**	**50.1**	-	-
UGA [197]	ICCV'19	**70.3**	**87.2**	-	-	53.3	75.0	-	-	21.7	49.5	-	-
Fully Unsupervised													
		Market				**Duke**				**MSMT17**			
Method	Reference	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
BUC [104]	AAAI'19	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4	-	-	-	-
GPUFL [167]	ICIP'21	42.3	69.6	-	-	37.7	57.4	-	-	-	-	-	-
MV [216]	SPL'21	45.6	73.3	85.3	89.1	31.7	54.5	67.5	72.1	-	-	-	-
MMCL [179]	CVPR'20	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0	11.2	35.4	44.8	49.8
HCT [223]	CVPR'20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	**87.4**	-	-	-	-
ABMT [18]	WACV'20	65.1	82.6	-	-	63.1	77.7	-	-	-	-	-	-
SpCL [48]	NeurIPS'20	73.1	88.1	95.1	97.0	-	-	-	-	19.1	42.3	55.6	61.2
RLCC [230]	CVPR'21	77.7	90.8	96.3	97.5	**69.2**	**83.2**	**91.6**	**93.8**	27.9	56.5	68.4	73.1
ICE [17]	ICCV'21	79.5	92.0	97.0	**98.1**	67.2	81.3	90.1	**93.0**	29.8	59.0	71.7	77.0
CACL [96]	TIP'22	80.9	92.7	**97.4**	**98.5**	**69.6**	**82.6**	**91.2**	**93.8**	23.0	48.9	61.2	66.4
PPLR [26]	CVPR'22	**81.5**	**92.8**	**97.1**	**98.1**	-	-	-	-	**31.4**	**61.1**	**73.4**	**77.8**
ISE [234]	CVPR'22	**84.7**	**94.0**	**97.8**	**98.8**	-	-	-	-	**35.0**	**64.7**	**75.5**	**79.4**
Ours		**83.4**	**92.9**	**97.1**	97.8	**72.7**	**83.9**	**91.0**	**93.0**	**42.6**	**68.2**	**77.9**	**81.4**

for multi-modal tasks, it achieves the best results in the most challenging datasets. The other methods were designed specifically for Person ReID and, for this reason, they are usually better in less complex datasets such as `Market`. This shows the effectiveness of our method in the fully unsupervised scenario.

Other works assume metadata, such as camera labels and tracklets, but no identity information. In Table 3.2, we can see that the most helpful metadata is camera information. Person ReID is naturally a cross-camera retrieval task: a method must be able to retrieve (from the gallery) images of the same identity used as a query but seen from other cameras. In this sense, camera information provides a significant impact if it is leveraged during training. This is evinced when we compare our results with the camera-based method PPLR. It considers camera proxies per cluster and pulls images from different cameras closer to overcome differences brought by different points of view. This seems especially beneficial in more complex datasets (`MSMT17`), where PPLR outperforms our method by 5.1 p.p. in R1. However, we still attain the best performance in mAP for this dataset among all methods in any category. The same conclusion is drawn for the `Duke` dataset where our method outperforms all prior art in all scenarios in terms of mAP and R1.

Tracklet-based methods explore temporal information during training but are not as competitive as camera-based ones. We outperform the state-of-the-art methods in all datasets by a large margin.

Finally, our method is able to place most true positive samples from the gallery closer to the query for most of the datasets, which is represented by the highest mAP for the two most difficult datasets, `Duke` and `MSMT17`. Our method is thus able to achieve the best performance among all methods even over those assuming **strong camera information as metadata.** One should also note that, in general, our assumptions are even more relaxed than other methods. Our clustering algorithm, for instance, does not require hyper-parameter tuning, while ICE explicitly fixes a value for ε depending on the target dataset.

In the following section, we also compare our pipeline to UDA methods that require a source domain to provide initial task-related knowledge, which is the same setup as our first solution. We outperform all these methods considering mAP, and obtain the best or second-best ranking values considering other metrics without source domain and any kind of labels.

Comparison to Unsupervised Domain Adaptation models

Now we extend our comparison to prior art considering the same experiment setup of the first solution proposed in this Ph.D. and presented in chapter 2. Those works employ supervised pre-training on a source domain and then leverage their pipelines on the target domain. Our pipeline, as explained in this chapter, does not require pre-training on task-related source datasets and is applied directly over the target domain from weights initialized over ImageNet. Therefore, our model operates under a more challenging scenario with fewer constraints. The results are shown in Tables 3.3 and 3.4. Both tables are subsets of Tables 2.2 and 2.3 presented in the previous chapter, and highlight the best

prior art models compared to our solution.

Table 3.3: Results on Unsupervised Domain Adaptation. Our models **do not** require pre-training on the source domain. Here we present `Market1501` to `DukeMTMC-ReID` and `DukeMTMC-ReID` to `Market1501` adaptation scenarios. We highlight the three best models with **blue**, **green**, and **orange**, respectively. UST [7] is our first solution proposed in Chapter 2

		Duke → Market				Market → Duke			
Method	reference	mAP	R1	R5	R10	mAP	R1	R5	R10
MMT [47]	ICLR'20	71.2	87.7	94.9	**96.9**	65.1	78.0	**88.8**	92.5
DG-Net++ [258]	ECCV'20	61.7	82.1	90.2	92.7	63.8	78.9	87.8	90.4
MEB-Net [226]	ECCV'20	**76.0**	**89.9**	**96.0**	**97.5**	66.1	79.6	88.3	**92.2**
ABMT [18]	WACV'20	**80.4**	**93.0**	-	-	**70.8**	**83.3**	-	-
UST [7]	TIFS'21	78.4	**92.9**	**96.9**	**97.8**	**72.6**	**85.0**	**92.1**	**93.9**
Ours	**This work**	**83.4**	**92.9**	**97.1**	**97.8**	**72.7**	**83.9**	**91.0**	**93.0**

Despite not training on the source domain, we can see in Table 3.3 that our pipeline yields the best mAP on both scenarios and the first or at least the second best value for all other metrics. On the challenging `MSMT17` (Table 3.4), we present an even better performance by obtaining the best values for all metrics, surpassing the second-best method by a margin in all metrics. Our first solution presented in chapter 2 is denoted by "UST". We can see we have a competitive performance in `Market` and `Duke` dataset (Table 3.3), and we outperform ourselves in all metrics considering the `MSMT17` as the target.

This shows the strong ability of our model to learn in a fully unsupervised scenario without requiring any meta-information, source domain, or manual definition of hyper-parameter clustering.

Table 3.4: Results on Unsupervised Domain Adaptation. We present `Market1501` to `MSMT17` and `DukeMTMCRe-ID` to `MSMT17` adaptation scenarios. Our models **do not** require pre-training on the source domain so we replicate our results for both considered scenarios. We highlight the three best models with **blue**, **green**, and **orange**, respectively. UST [7] is our first solution proposed in Chapter 2

		Duke → MSMT17				Market → MSMT17			
Method	reference	mAP	R1	R5	R10	mAP	R1	R5	R10
DG-Net++ [258]	ECCV'20	22.1	48.8	**60.9**	**65.9**	22.1	48.4	60.9	66.1
ABMT [18]	WACV'20	**33.0**	**61.8**	-	-	27.8	55.5	-	-
SpCL [48]	NeurIPS'20	-	-	-	-	**31.0**	**58.1**	**69.6**	**74.1**
UST [7]	TIFS'21	**34.5**	**63.9**	**75.3**	**79.6**	**33.2**	**62.3**	**74.1**	**78.5**
Ours	**This work**	**42.6**	**68.2**	**77.9**	**81.4**	**42.6**	**68.2**	**77.9**	**81.4**

Training time

We also analyze our pipeline in terms of execution time for training (Table 3.5). Clustering (Step 2) time is negligible due to its optimized implementation. The fine-tuning process (Steps 3 to 7) takes longer as it requires optimization using the generated clusters for $K_2 = 7$ epochs. The total time for the pipeline is in the order of a few hours, and it depends on the size of the dataset. For `MSMT17`, the largest one with $32,621$ images in the training set and 1,041 identities (greater than the number of classes on ImageNet), the method presents a reasonable time of around 19 hours.

Table 3.5: Execution time. We report the average time taken by some steps and the total time to execute the pipeline, for the three ReID datasets. The time format is HH:MM:SS.

Dataset	**Step 1**	**Step 2**	**Steps 3 to 7**	**Total Time**	**Inference Time (ms)**
Market	00:01:29	00:00:07	00:08:27	05:10:35	41.4
Duke	00:01:53	00:00:10	00:15:13	08:47:37	59.1
MSMT17	00:04:25	00:00:38	00:30:12	19:06:05	80.2

Inference time also increases with the size of the gallery sets, as there are more samples to compare to the query. Inference for a query on `Market`, the smallest one, takes 41.4 milliseconds, while for `MSMT17`, the largest dataset, it takes 80.2 milliseconds. All scenarios present reasonable inference time under one second.

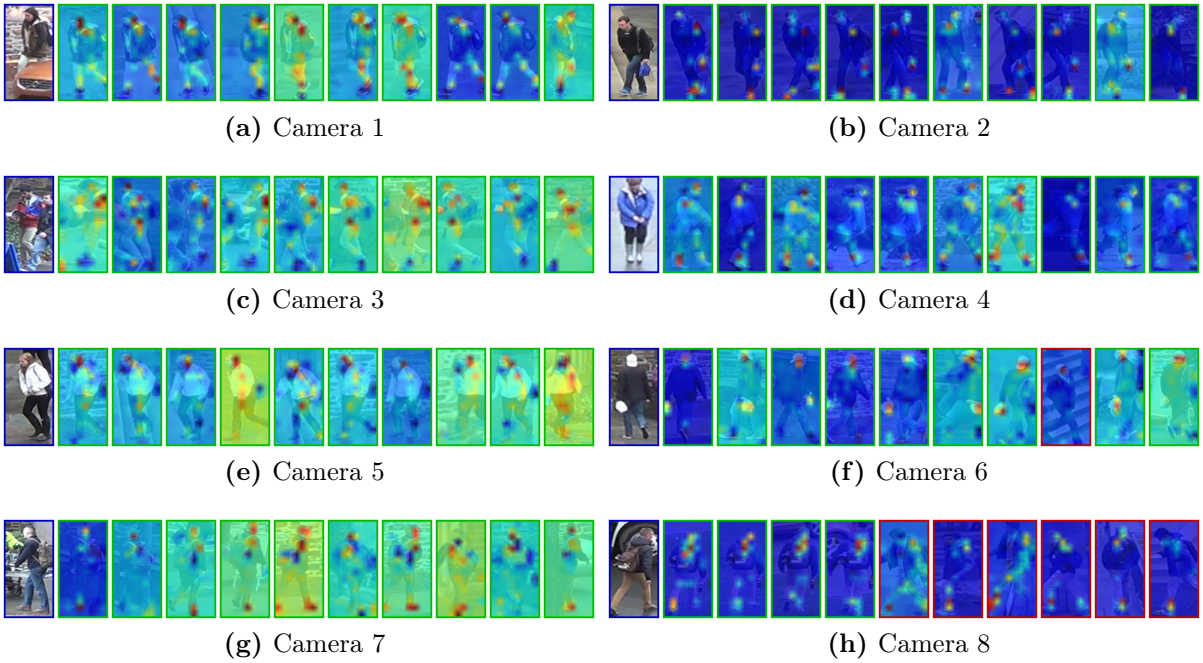
Qualitative Analysis

Similar to the previous chapter, we show the activated regions in the gallery images given a query from each camera in each dataset, considering the ResNet50 backbone. We provide some examples of success and some of failure, for `DukeMTMCReID` (Figures 3.4 and 3.5 respectively) and `Market1501` (Figures 3.6 and 3.7 respectively) datasets. We do not show examples for `MSMT17` as reproducing this dataset's images is not allowed in any format. Blue frames indicate the query image, green indicates the true positive samples, and red the false positive samples. We adapted the method from [163] to visualize the results.

Figure 3.4 shows one success case for each one of the eight cameras present on the `Duke` dataset. We see that the model is able to mine fine-grained details on all images, regardless of the camera, and retrieve true positive samples in the top 10 images. The only exception is for camera 8 (Figure 3.4h), where the last six images are false positive samples. However, for all samples, the model still focuses on fine-grained details, and the errors occur due to similar clothing of the identities. The same happens in the examples in Figure 3.5. Despite being error cases, the model can focus on fine-grained details. The errors occur when there are two or more identities in the query image (Figure 3.5f), the background color is similar to an object held by the identity (in Figure 3.5c the backpack is the same color of the background), or due to very similar clothing of different identities. Since `Duke` was recorded at a university campus during winter, it is natural to find individuals wearing similar (and dark) clothing.

Besides that, in Figure 3.5c, there are two people in the same bounding box, which is ambiguous. The top seven images retrieved present another person walking with the query identity, and the sixth and eighth images show the query identity. So despite being a failure case, the model could help in cases where the goal is to find people walking together or in groups. This can help in investigations and monitoring to locate where two people were together in the environment.

Figure 3.6 shows one success case for each camera on the `Market` dataset. Following the same previous conclusions, the model is able to identify fine-grained features in the image to correctly retrieve the true positive samples among the top 10 images. The failure



(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

(e) Camera 5 **(f)** Camera 6

(g) Camera 7 **(h)** Camera 8

Figure 3.4: Success cases considering one query from each camera on `Duke`. These results are obtained with the ResNet50 backbone after training on `Duke`.

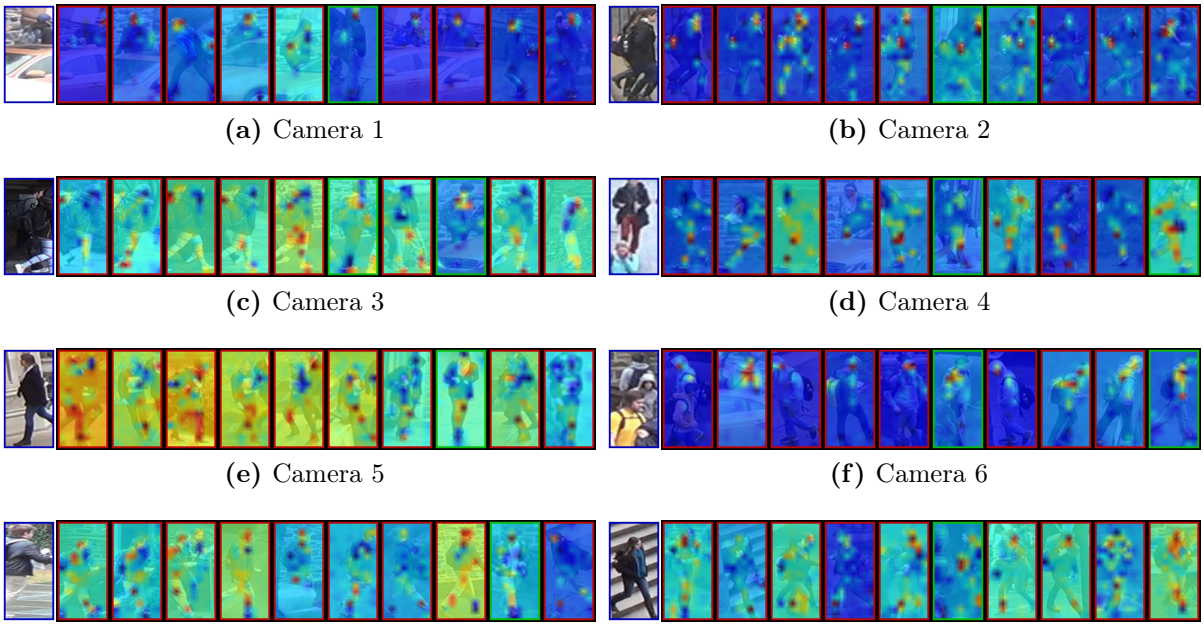
cases (Figure 3.7) are mainly due to the identities being similar (similar T-shirts, hair, and skin tone), which is the case in Figures 3.7b, 3.7d, and 3.7f.

From the successful examples (Figures 3.4 and 3.6), we observe that our method can identify several fine-grained details throughout the image in order to retrieve the correct identity. This means it overcomes differences in point-of-view, pose, illumination, and background.

3.3.4 Authorship Attribution

We now consider a second task — Text Authorship Attribution — with minor adjustments mainly related to the nature of the problem. For the backbones, we consider BERT [34], BERTweet [131], and T5 [149] as they were developed to deal with text. We apply augmentation on tweets with more than 5 tokens by masking from 10% to 20% of the tokens with a "mask" token on BERT and BERTweet, and with an "unknown" token on T5. Even the base version of BERT is too complex for short signals (text tweets), making the training more prone to overfitting. Hence, we freeze the first ten attention blocks of BERT and BERTweet, leaving only the 11^{th} block to be updated. For the same reason, we set $K_1 = 15$ as the number of iterations — half of the value used for the ReID experiments — to alleviate the impact of over-training. We set $P = 8$ and $K = 8$ for batch creation. All other parameters are the same as those used for ReID. The only difference in the pipeline is that we do not apply the optimization shown in Figure 3.2, as it is only for image representations.

We run comparative experiments considering the AdHominem method [9]. It employs an attention-based model for Authorship Attribution on social media text. AdHominem performs supervised training using Siamese networks to answer whether or not the same



(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

(e) Camera 5 **(f)** Camera 6

(g) Camera 7 **(h)** Camera 8

Figure 3.5: Failure cases considering one query from each camera on Duke. These results are obtained with the ResNet50 backbone after training on Duke.

author wrote two tweets. To perform the ranking task, we take the Euclidean distance between two tweets returned by the model and rank the gallery tweets given a query. The results are shown in Table 3.6.

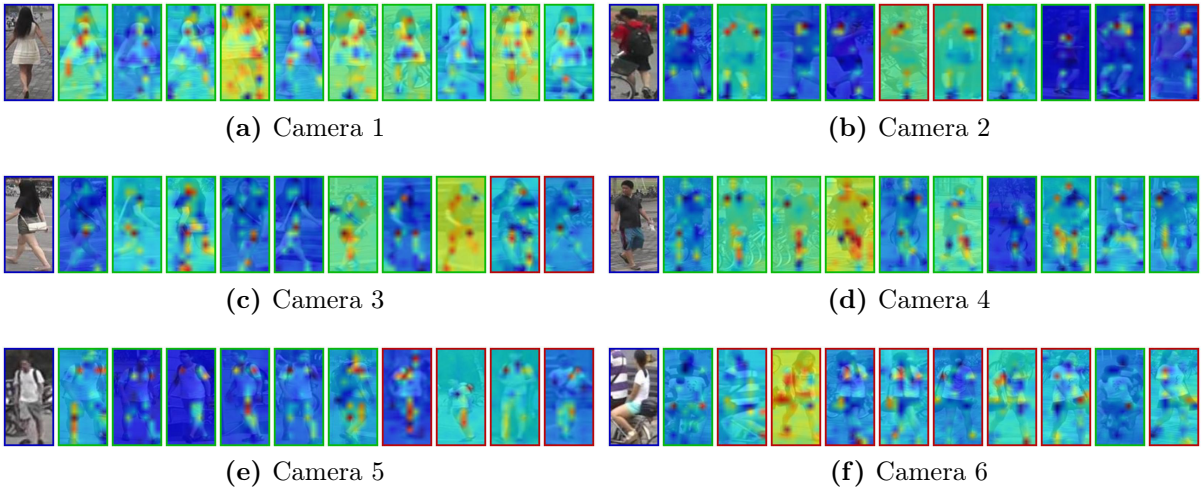
Table 3.6: Results for the Authorship Attribution task for two subsets of tweets: one with 50 authors on training and 50 authors on the test, and the other with 500 authors on training and 500 authors on the test. The best results are in **blue**.

	1st subset (50 authors)				**2nd subset** (500 authors)			
Model	mAP	R1	R5	R10	mAP	R1	R5	R10
AdHominem [9]	7.3	25.5	50.9	61.5	2.4	10.8	23.5	31.6
Ours	**14.3**	**50.0**	**73.0**	**80.3**	**5.0**	**22.5**	**37.0**	**44.8**

Our method outperforms AdHominem in both subsets. More specifically, we outperform AdHominem by 7.0 and 24.5 p.p. in mAP and R1, respectively, in the first subset. In the second one, we outperform it by 2.6 and 11.7 p.p.

One must note that AdHominem is trained in a **supervised** manner considering the identity of each tweet, i.e., the method knows *a priori* "who" wrote the tweets to supervise the training. However, our method relaxes this constraint by taking only the raw tweet text **without any labeling.** Moreover, our training and test data are disjoint on the identities, and since AdHominem is trained for a closed set of authors, it generalizes poorly to unseen authors.

Although our method utilizes pre-initialized weights, these weights were trained for other tasks, such as question answering, next sequence prediction, predicting missing words, and so on, instead of Authorship Attribution.

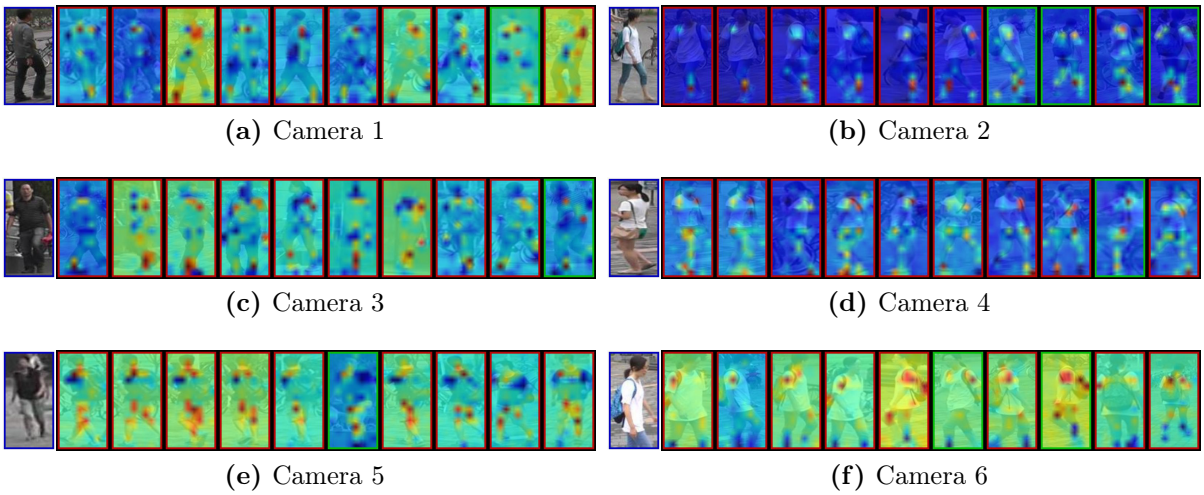


(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

(e) Camera 5 **(f)** Camera 6

Figure 3.6: Success cases considering one query from each camera on `Market`. These results are obtained with the ResNet50 backbone after training on `Market`.



(a) Camera 1 **(b)** Camera 2

(c) Camera 3 **(d)** Camera 4

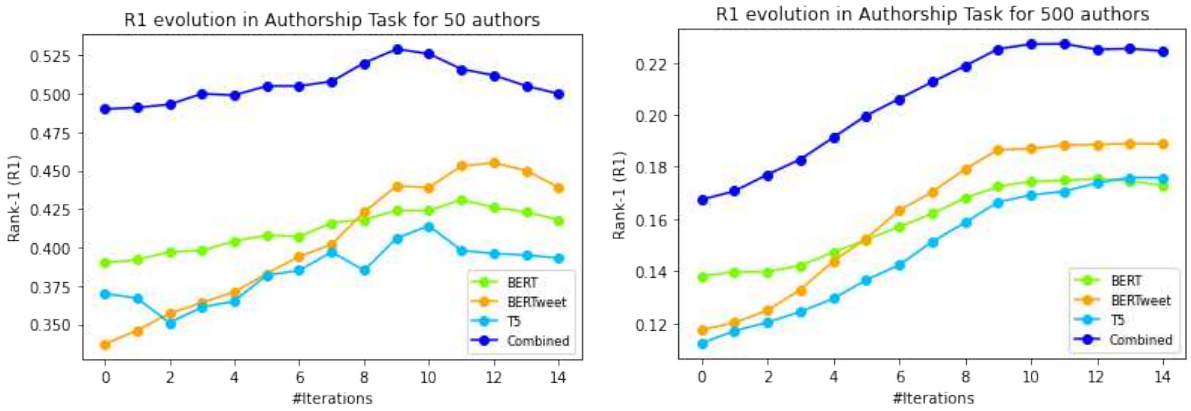
(e) Camera 5 **(f)** Camera 6

Figure 3.7: Failure cases considering one query from each camera on `Market`. These results are obtained with the ResNet50 backbone after training on `Market`.

In Figure 3.8, we provide the evolution of Rank-1 during training to show the merits of employing our pipeline. In the first subset (50 authors), we verify a Rank-1 oscillation after the 10^{th} iteration and a slight decrease until the last iteration. BERTweet provides the greatest gain, followed by BERT and T5. In the second subset (500 authors), the training process is more stable though it is numerically inferior as there are more authors. The feature space is denser, securing more stability for the convergence of the models.

3.3.5 Further Analysis

Our applications are Unsupervised Person Re-Identification and Unsupervised Text Authorship Attribution, which are related to event analysis in a forensic context. Thus, it is important to understand how the performance of our approach is impacted when we consider common forensics features used for other tasks, such as image manipulation detection [5], or pre-processed inputs used in the GAN-generated image detection task [3].



(a) 50 authors. **(b)** 500 authors.

Figure 3.8: Rank-1 evolution over training iterations for the Authorship Attribution task considering a test set with (a) 50 authors and (b) 500 authors. "Combined" is obtained by averaging the final distances acquired with each backbone.

To test this, we apply the method proposed in [5] for manipulation detection to the Unsupervised Person Re-Identification problem. It is an ensemble-based algorithm with forensics features. We employ the same features—SPAM [146] and CRSPAM1372 [4]—, which have also shown to achieve top-tier performance on the manipulation detection task.

However, since Unsupervised Person Re-Identification differs in nature from manipulation detection, SPAM and CRSPAM1372 features induce a huge performance drop when employed in our setup. These hand-crafted features have not been designed to overcome the non-linearity in Unsupervised Person Re-Identification problems caused by the difference in illumination, resolution, pose, background, and occlusion.

We also consider pre-processing techniques employed in the GAN-generated image detection task. We take the method proposed in [3], whose main goal is distinguishing between GAN-generated and real images. The authors argue that it is harder for GANs to reconstruct consistent relationships among the color bands, and they propose an approach to detect these inconsistencies.

For a fair comparison, we keep our whole pipeline and change only the input of our backbones to match the ones from [3]. Their method induces a huge performance drop compared to ours with standard RGB images. Our performance is better than theirs in all metrics in all evaluated datasets. Their method requires a signal analysis across color bands to reconstruct inconsistencies and to detect the neighboring inconsistencies and noise introduced by GAN-based models. However, this process might destroy the semantics present in the images, which is fundamental in Unsupervised Person Re-Identification. Our models, conversely, learn high-level semantic and camera-invariant features to match the same person seen from different camera views and distinguish from other people by looking at discriminant parts, such as clothes, shoes, bags, and faces (Figures 3.4 and 3.6). In summary, the method proposed in [3] looks for fine-grained details at the noise level, while ours looks for fine-grained details at the semantic level.

3.4 Ablation Study

We validate each part of our pipeline by checking how these influence the performance, considering the `Market` and `Duke` datasets. If not specified, we assume ensemble-based clustering, and $\tau = 0.04$ and $\lambda = 0.5$ in Equations 3.3, 3.4, and 3.5.

3.4.1 Step 1: impact of distance averaging

In the first step of the pipeline, we average the distance matrices obtained with each backbone separately, computing a combined distance matrix $\overline{\mathrm{D}}$ (Equation 3.1). This allows complementary knowledge to be grouped together for training. To measure how this impacts the final performance, we train each backbone separately, as expected, but feed each distance matrix directly to Step 2, instead of averaging them together. We present the results in Table 3.7, showing the impact of averaging the distances in Step 1 for each backbone separately and for the combined result (considering Equation 3.7).

Table 3.7: Impact when we remove the proposed backbones knowledge combination (Equation 3.1). Results with (*) mean we do not apply our proposed fusion. The best results are in **blue**.

	Market				**Duke**			
Model	mAP	R1	R5	R10	mAP	R1	R5	R10
ResNet50*	77.8	90.7	96.2	97.7	65.6	80.0	89.2	91.3
ResNet50	**80.4**	**91.1**	**96.8**	**98.0**	**69.4**	**82.0**	**90.4**	**92.9**
OSNet*	73.2	87.8	95.2	96.8	67.5	81.3	90.0	**92.6**
OSNet	**78.6**	**90.5**	**96.1**	**97.2**	**68.9**	**82.9**	**90.2**	92.0
DenseNet121*	73.2	87.5	94.9	96.6	63.8	79.2	87.8	90.4
DenseNet121	**78.9**	**90.3**	**95.9**	**97.4**	**67.7**	**82.1**	**90.0**	**92.2**
Ours*	81.0	91.8	96.9	**97.9**	70.6	82.3	90.2	92.5
Ours	**83.4**	**92.9**	**97.1**	97.8	**72.7**	**83.9**	**91.0**	**93.0**

Distance averaging in Step 1 is important for training, and it positively impacts each backbone individually, as well as their final combination, which allows for better grouping in the clustering step. Considering the combination of backbones (Equation 3.7), the gains are also considerable for both datasets. In `Market`, we achieve an improvement of 2.4 p.p. and 1.1 p.p., in mAP and R1, respectively, and 2.1 p.p. and 1.6 p.p. on `Duke`. These results show the effectiveness of our proposed approach without requiring mutual training [47, 226] or co-teaching [207], which in turn promotes simpler training. Moreover, from Table 3.7, it is possible to conclude that different backbones provide complementary information, as there is an improvement for all metrics when they are combined (Equation 3.7), for both setups ("Ours(*)" and "Ours").

3.4.2 Step 2: impact of the ensemble-based clustering

We verify the effectiveness of our proposed ensemble-based clustering method. We replace it with the standard DBSCAN algorithm in the second step of our pipeline, and keep the

remaining parts unchanged. As we combine the results of DBSCAN runs with different ε values into a single result, in Table 3.8, we present the separate results for each ε value.

Table 3.8: Impact of the ensemble-based clustering. We replace it with the standard DBSCAN by fixing five ε values. The best result for each metric is highlighted in **blue**.

	Market				**Duke**			
DBSCAN ε	mAP	R1	R5	R10	mAP	R1	R5	R10
0.50	81.6	90.7	95.4	96.4	60.9	76.4	82.7	84.7
0.55	82.9	92.3	96.4	97.3	62.8	77.2	83.7	85.8
0.60	82.3	91.6	96.2	97.4	67.5	79.6	87.7	89.9
0.65	82.7	92.3	96.7	97.7	69.1	81.9	88.9	91.5
0.70	81.8	91.8	96.8	**97.8**	70.3	82.8	90.0	92.1
Ours	**83.4**	**92.9**	**97.1**	**97.8**	**72.7**	**83.9**	**91.0**	**93.0**

This experiment shows that the proposed ensemble-based clustering effectively combines DBSCAN intermediate results into a final one, suggesting more robust clusters, and outperforming all results in both datasets. Note that, as motivated in Section 3.2.2, if a single ε is employed for clustering, each dataset has different optimal values. In Table 3.8, we see that the model achieves the best performance in mAP and R1 with $\varepsilon = 0.55$ for Market and $\varepsilon = 0.7$ for Duke, which shows that the optimal values can change significantly from a dataset to another. In contrast, our proposed ensemble-based clustering strategy obtains the best performance by grouping DBSCAN results using lower ε values (denser clusters, lower false positive rate) and higher ε values (more diverse clusters, lower false negative rate), alleviating the burden of choosing a proper unique value for this hyper-parameter.

3.4.3 Step 3: impact of proxy selection

In Step 4, we select a random sample per cluster as a proxy to aid the optimization as in [29]. We validate this choice by replacing the random selection with the mean feature vector of each cluster. The results are shown in Table 3.9.

Table 3.9: Impact of the proxy selection. We replace the random selection of samples to serve as cluster proxies by the mean feature vector of the cluster. The best results are in **blue**.

	Market				**Duke**			
	mAP	R1	R5	R10	mAP	R1	R5	R10
Mean	82.0	91.8	96.2	97.5	70.7	81.8	90.0	92.2
Ours (random)	**83.4**	**92.9**	**97.1**	**97.8**	**72.7**	**83.9**	**91.0**	**93.0**

Random selection improves the performance for all metrics, but mainly for mAP and R1. More specifically, we obtain a gain of 1.4 and 1.1 p.p. in mAP and R1, respectively, in the `Market` dataset, and 2.0 and 2.1 p.p. in the `Duke` dataset. This validates our assumption that using a mean vector as a proxy hinders cluster representation and further training, as it is affected by the false positive samples of the cluster.

3.4.4 Step 4: impact of loss function hyper-parameters

We vary hyper-parameters τ and λ in the loss functions (Equations 3.3, 3.4, and 3.5) to check how they impact the pipeline in both image- and text-based applications.

The τ parameter, proposed in [65], is used to control the sharpness of the distribution related to the distance of a sample to each cluster proxy in the current iteration. The smaller the value, the greater the density towards the most confident value; while the larger the value, the closer the distribution is to the Uniform. Prior Unsupervised Person Re-Identification works usually tune it to control the gradients for a stable convergence. Results with varying τ are shown in Figures 3.9a and 3.9b for ReID, and in Figures 3.9c and 3.9d for Authorship Attribution. For both datasets in ReID, the best value is 0.04 as it provides the best mAP and R1, and top results for R5 and R10. For Authorship Attribution, the results have a marginal improvement for $\tau = 0.03$ in both 50 and 500 authors datasets, but it provides slightly lower performance for ReID. The greater values ($\tau = 0.07$) increase the loss and the gradient magnitude, which leads to suboptimal optimization for all datasets, mainly for ReID. Therefore, we set $\tau = 0.04$ for all experiments in both applications.

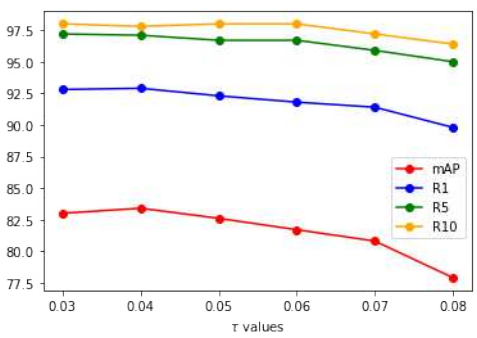
Results with varying λ (Equation 3.5) are shown in Figures 3.9e and 3.9f for ReID, and in Figures 3.9g and 3.9h for Authorship Attribution. This hyper-parameter regulates the influence of the hard instance-based softmax-triplet loss (L_{hard}) on the final loss. This loss brings a local view by considering hard-positive and hard-negative mining triplets at the batch level. When $\lambda = 0.0$ (i.e., L_{hard} is not considered for optimization), we obtain one of the worst performances for all datasets in both modalities, showing that a more local view of the data is also important for model training. As we increase its value, the performance increases. But for higher values (from 0.75 to 1.0), we verify a performance drop as the local view starts dominating the global view, resulting in convergence issues. We keep $\lambda = 0.5$ in all experiments, as it yields the best or second-best mAP and R1 results for all datasets in both modalities.

3.5 Application with supervised learning: the BENTO algorithm

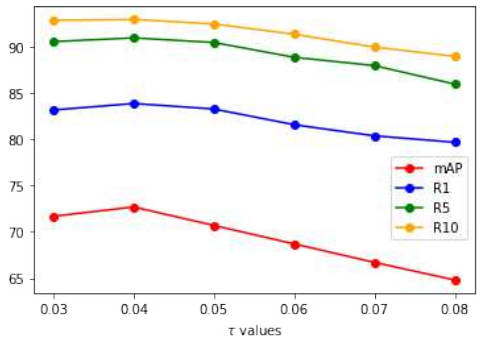
In this section, we show a direct application of the solution presented in this chapter in a context in which we have labeled and unlabeled data. We leverage our solution as part of a method designed for the AGReID2023 competition during the IJCB 2023[3]. The goal of the competition was to perform the whole-body identity matching from an image of the person of interest taken by a UAV (aerial device) to an image of the same person recorded from a CCTV or wearable camera (ground devices). The organizers called Aerial-to-Ground Person Re-Identification. One of the main challenges in this scenario is the strong variation in the point of view.

Our approach named Biometric Ensemble Network Technique Optimization (BENTO – Figure 3.10) is inspired by different strategies present in the literature [7, 186, 17],

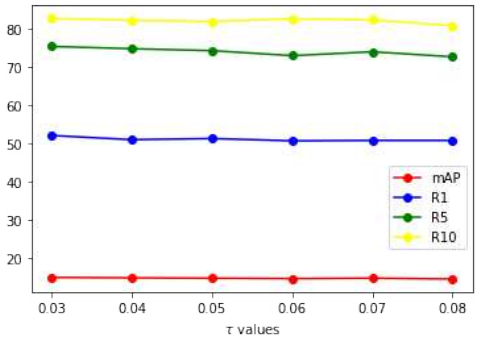
[3]https://www.kaggle.com/competitions/ag-reid2023/leaderboard



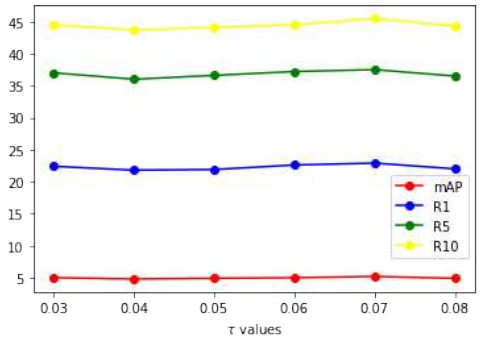
(a) Different τ values on `Market`.



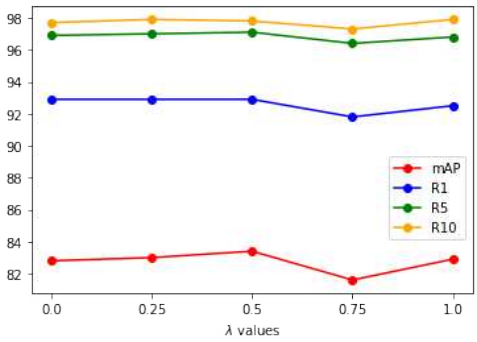
(b) Different τ values on `Duke`.



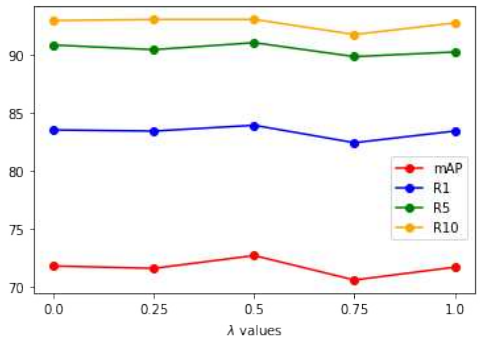
(c) Different τ values on 50 authors.



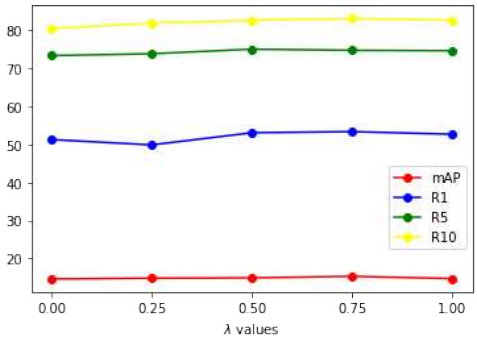
(d) Different τ values on 500 authors.



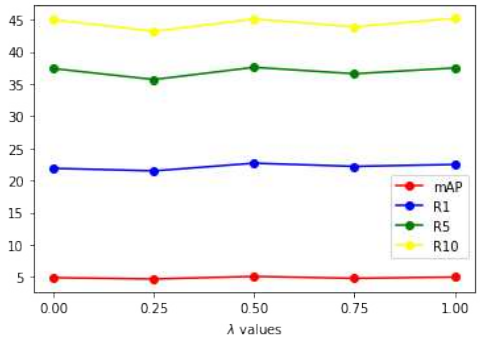
(e) Different λ values on `Market`.



(f) Different λ values on `Duke`.



(g) Different λ values on 50 authors.



(h) Different λ values on 500 authors.

Figure 3.9: Impact of different τ and λ values on `Market`, `Duke`, 50 authors, and 500 authors datasets, considering mAP, R1, R5, and R10.

among them we highlight our first solution [7] presented in chapter 2, and the solution proposed in this chapter. During training before each epoch, we extract all training features and calculate the average feature vector per identity as a class-level proxy (cross symbols in Figure 3.10), and average the feature vector per camera per identity as the camera-level proxies (points with darker border). Let denote the class-level proxies as $P = \{p_{i=1}^{N_c}\}$ where N_c is the number of training identities, and the camera-level proxies as $C = \{(c_a^i, c_w^i, c_{cctv}^i)\}_{i=1}^{N_c}$, where c_a^i, c_w^i and c_{cctv}^i are the average feature vector of the aerial, wearable and CCTV cameras respectively for the i^{th} identity. Given a batch B of images, the losses are the proxy loss (Eq. 3.3), the batch-hard loss (Eq. 3.4), and the camera-proxy loss calculate as follows:

$$L_{cp} = -\frac{1}{|B|}\sum_{b=1}^{|B|} log\frac{exp(f_b^e.c_g^b/\tau)}{exp(f_b^e.c_g^b/\tau) + exp(f_b^e.c_e^-/\tau)} \tag{3.8}$$

where f_b^e is the feature of the b^{th} sample from aerial/ground camera e, c_g^b is the camera-proxy from ground/aerial camera g from the b^{th} sample class ($e \neq g$), and c_e^- is the hardest negative proxy from the same camera of the b^{th} sample but from another class. The rationale of L_{cp} is to encourage the model to have camera-invariant features by enforcing the sample from the aerial device to be close to the ground camera proxies and vice-versa. The final loss function is defined as

$$L_{AG} = L_{proxy} + \lambda_{cp}L_{cp} + \lambda_{hard}L_{hard} \tag{3.9}$$

Considering a real-world application where we can have labeled and unlabeled data, we leverage our second solution proposed in this chapter for a self-supervised fine-tuning on query and gallery images **without any ground-truth labels from them** after training with the labeled data. In other words, we first train the models in a supervised manner using training data, then we leverage our second solution proposed in this chapter on query and gallery data, where we extract features from query and gallery, cluster them, and use the pseudo-label to finetuning for one to three epochs.

On evaluation, we extract all feature vectors from the gallery set and average the ones from the same identity (similar to the training). We use them to perform the matching to query features and Re-Ranking [245]. We do that with four different backbones (ResNet50-IBN, ResNet101-IBN, OSNet [251], and TransReID [61]) and average them to get the rank-1 prediction employing Eq. 3.7.

The results comparing our solution to other competitors' solutions are shown in Table 3.10. We can see that with our solution we obtained the **third place** in the competition being one of the highlights ranked among the top-3 performers. For fairness, it is important to note that the first place team also employed a similar strategy to us of pseudo-labeling finetuning over the query and gallery data along with Re-Ranking [245].

A joint paper[4] with the organizers and other competitors was published and presented as part of the competitions section during the IJCB 2023.

[4]https://ijcb2023.ieee-biometrics.org/accepted-papers/

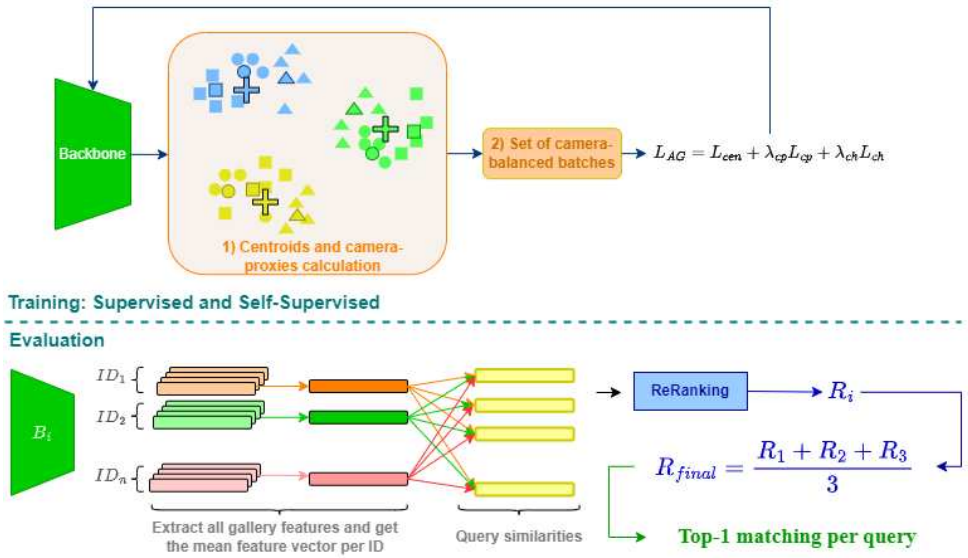


Figure 3.10: Pipeline overview. We first perform a supervised training and then a self-supervised training with query and gallery images **(no ground-truth labels are considered)** with the solution proposed in this chapter. The evaluation ensembles all backbones to retrieve Top-1 matching.

3.6 Final Remarks

In this chapter, we presented the second solution for Unsupervised Person Re-Identification designed in this Ph.D. research. We proposed a novel self-supervised learning pipeline for scenarios with high intra-class semantic disparity and inter-class similarity. General methods do not account for this problem as they are usually devised for less complex datasets, such as Imagenet.

Our pipeline starts from a common concept — clustering steps to propose pseudo-labels for unlabeled samples and optimization steps to update backbones supervised by the pseudo-labels —, but we incorporate novel techniques to address more critical tasks effectively.

We propose the use of a neighborhood-based distance refinement followed by distance averaging to amalgamate complementary knowledge learned by different backbones. We showed that this is highly effective when compared to using distances obtained from each backbone directly. We provide a better distance measurement between samples, even without task-related initialization, due to the joint contribution of neighborhood-based distances and distance matrices ensemble.

Our second contribution is an ensemble-based clustering algorithm to provide pseudo-labels for optimization. The advantages are twofold: our solution creates dense but diverse clusters and does not need clustering hyperparameter tuning.

To show the generalizing ability of our pipeline, we applied it to two highly different Multimedia Forensics tasks: Person Re-Identification and Authorship Attribution from short text messages. To the best of our knowledge, this is the first self-supervised learning method that can be applied to different forensics modalities with only minor adjustments.

For Person ReID, our method yields state-of-the-art performance in terms of mAP and Rank-1 in the most challenging datasets. For Authorship Attribution, we obtained competitive results when compared to a prominent method that was trained in a supervised

Table 3.10: Final teams ranking in AGReID2023 competition held during the IJCB 2023. The methods without reference were proposed during the competition.

Method	R1
LENS-AG-Net	97.73
CentroidNet	94.28
BENTO	**92.95**
MFE	92.80
Swin [116]	83.45
SMTL	81.29
SwinV2 [115]	80.25
MGN (R50) [181]	79.28
BoT (R50) [122]	78.42
HRNet-18 [183]	78.38
SBS (R50) [60]	74.94

manner. Therefore ensemble-based clustering has a strong potential to find satisfactory clusters for model training on the fully-unsupervised scenario. Our self-supervised technique can considerably help the task of Textual Authorship Attribution in this forensic scenario since it opens the possibility of using a massive amount of unlabeled data to foster the results.

We conclude that learning from complex fully-unlabeled data in different modalities is possible. Still, the model requires a robust distance measurement (brought by the ensemble of distance matrices) and a clustering strategy that tackles the unknown feature distribution from different datasets. When both strategies are put together, the method finds robust clusters for optimization.

One important aspect of the method that still needs optimization is memory usage. Currently, it requires quadratic memory $O(N^2)$, where N is the total number of samples available for training, due to the pairwise distance matrices. Nonetheless, all prior art also faces the same issue. Besides, the Re-Ranking technique [245] employed in this solution and widely used in the Unsupervised Person Re-Identification community has a time complexity of $\mathcal{O}(N^3)$, which turns more and more expensive as we increase N.

Moreover, despite alleviating the requirements of selecting a specific clustering hyper-parameter for each dataset, our method still works in a small subset of five possible clustering hyper-parameter values, and it does not account for the change in the distribution of the data points in the feature space. At the beginning of the training, the discrimination power of the models in the target dataset is weak since they do not have any target-related knowledge to start the training. Moreover, the Person ReID problem, as mentioned in this chapter, faces high intra-class disparity and inter-class similarity. Consequently, features of different identities are close, and features from the same identity are farther apart, mainly at the beginning of the training. As the training progresses, their discrimination ability increases which means that the feature distribution changes by setting samples of the same identity closer, and samples from different identities apart.

In this context, it might be more interesting to consider a clustering hyper-parameter setting that considers this dynamic behavior of the feature space, instead of keeping the same set of clustering hyper-parameter values during the whole training.

Motivated by these challenges and considering the large-scale dataset scenarios that likely will appear in real-world applications, we propose our third solution, which is presented in the next chapter. This solution addresses large-scale scenarios, it considers the dynamics of the feature space to define the clustering hyper-parameters, and it is tested in Person and Vehicle Re-Identification tasks, opening the path for general application in Re-Identification and beyond.

Chapter 4

Large-scale Fully-Unsupervised Re-Identification

As mentioned in previous chapters, clustering-based self-supervised learning for fully unsupervised re-identification has attracted attention in the past years due to the capability of learning from unlabeled datasets by performing clustering and finetuning with pseudo-labels. However, one obstacle to the deployment of fully unsupervised ReID state-of-art methods is that most of them have reached top-tier performance in small datasets. The largest well-known Person Re-Identification benchmarks `Market1501` [242], `DukeMTMC-ReID` [151][1], and `MSMT17` [194] have 12,936, 16,522, and 32,621 training images, respectively. For Vehicle Re-Identification, the `Veri` [112] dataset has 37,778 training images. Since the training sets are rather small, researchers could employ effective but costly techniques to achieve top-tier performance, such as Re-Ranking [245] and Co-Training [47, 226, 225].

The Re-Ranking technique [245] has memory complexity of $\mathcal{O}(n^2)$ (where n is the training set size) and time complexity of $\mathcal{O}(n^3 + nk^3)$ (where k is the number of reciprocal neighbors – see Section 4.2.3 for further details). Co-Training usually involves cross-supervision where the confidence level of the samples from one model is used to weigh the loss functions from the other models in the ensemble, and all loss functions are optimized at once with many hyperparameters to control the contribution for each term. Furthermore, these hyperparameter values can be hard to tune, mainly in large-scale fully unsupervised datasets without a validation set.

Aiming to deploy scalable and affordable solutions for large-scale Unsupervised Person/Vehicle Re-Identification, we design a novel pipeline to alleviate the discussed challenges. We propose a novel Re-Ranking method that leverages the previously proposed k-Reciprocal Encoding [245], but considering just a local neighborhood to calculate the Jaccard distance, without relying on their set expansion nor in local query expansion. This allows us to reduce the time complexity and keep the memory complexity in $\mathcal{O}(kn)$

[1] `DukeMTMC-ReID` has been discontinued, and it must not be used for evaluation and benchmarking anymore. For this reason and following the recent literature, we **do not** use it for evaluation. More details in *https://www.dukechronicle.com/article/2019/06/duke-university-facial- recognition-dataset-study-surveillance-video-students-china-uyghur*. When this fact became known to us, we had already proposed the first two solutions (Chapters 2 and 3), and, for this reason, we were still providing results considering this dataset.

where ($k \ll n$).

Moreover, we also propose a sampling method based on the local neighborhood for a randomly chosen point. We have different local neighborhoods in each epoch, which effectively reduces the training set size without violating the neighborhood properties of the selected point. This preserves the hard-positive and hard-negative samples for effective training and, at the same time, reduces memory and time complexities.

Since co-training has shown impressive performance in dealing with noisy data, we also propose a simple co-training strategy based on the co-training theory but not requiring any human supervision or hyperparameter tuning. During training, we generate pseudo-labels for unlabeled training data using different convolutional neural network backbones. Each backbone generates its own feature space and then its own set of pseudo-labels. We propose switching the pseudo-labels among the backbones; thus, one backbone supervises the other through pseudo-label predictions.

Beyond the discussed large-scale challenges, a fundamental aspect has been overlooked: the choice of the clustering hyperparameter. Most methods employ DBSCAN [41], which is controlled by the density parameter ε, and some works consider an optimal value for each dataset. For instance, AdaMG [144] sets $\varepsilon = 0.5$ for `Market1501` and $\varepsilon = 0.7$ for `MSMT17`. When they fix the same value for both datasets, the performance drops. The recent ISE [234] and RTMem [215] also set different hyperparameters per dataset to achieve their best performance. However, since there is no validation set in a fully unsupervised scenario, we argue it is unrealistic to select a specific value per dataset.

Our first attempt to reduce the dependency on the clustering hyperparameter was presented in the previous chapter, where we combined the clustering results for five different hyperparameter values in a final ensembled clustering result. We have shown, in Table 3.8, that our proposed clustering combination yields better performance than using each individual clustering result for each hyperparameter value. However, we always keep the same hyperparameters' values throughout the whole training without considering the dynamic nature of the feature space. Since the backbones evolve their knowledge, getting more discriminative over time, it is expected that the features change their distribution in the feature space, so each epoch might require a different hyperparameter value. Moreover, even with some scheduling scheme to assign a clustering hyperparameter for each training epoch, this scheme should be invariant to the dataset, since we are targeting fully unsupervised scenarios. We must avoid specific choices per dataset.

To alleviate the burden of selecting an optimal hyperparameter per dataset, we analyze the clustering problem from the perspective of noisy-labeling-robust learning. During training, feature vectors are extracted and clustered, and pseudo-labels are assigned to them, which are then used for finetuning. In the first training epochs, the backbones have little knowledge about the dataset, so it is expected that the features are not too discriminative. Therefore, there is more noisy pseudo labeling in the first iterations, which can require a tighter density parameter ε in DBSCAN. As the model gets more robust during training, generating better features, it allows the loosening of the density parameter to include more hard-positive samples in the clusters. However, if we keep it loose until the end of the training, it might include too many non-matching samples, so we should decrease ε after the feature space is reasonably tuned to allow final feature

learning. Finally, in order to define a stopping criterion, it might be interesting to hold the parameter constant at the end. This motivated us to design a new scheduling scheme for ε, which is used for all datasets without the need for hyperparameter tuning. We show that it reaches state-of-the-art performance in all evaluation scenarios, even outperforming the ones that select a dataset-specific hyperparameter for optimal performance.

To verify the performance of the proposed methods, we evaluate them not just in the well-known benchmarks but also in the large-scale `VehicleID` [108] and `Veri-Wild` [119] datasets, which are 3× and 7.35× larger, respectively, than the Veri dataset (the largest among well-known benchmarks).

Similar to the solution presented in the previous chapter, this third solution **does not rely on camera labels or side information**. It requires only that the target object (e.g., people or vehicles) be already detected in the images, which is defined by their bounding boxes.

The key contributions of our work are:

- A neighborhood-based sampling method to decrease the dataset size in each epoch. By preserving the neighborhood, it is able to keep hard-positive and hard-negative samples for model learning.
- A Re-Ranking method that considers just the top-k nearest neighbors and does not need set expansion nor local query expansion as previous methods. In this way, we effectively reduce the time and memory complexities.
- A density parameter scheduling to deal with noisy data during training and, at the same time, bring diverse samples together. We keep the scheduling scheme for all datasets, avoiding hyperparameter tuning.
- A co-training method where we switch the predicted labels among the involved backbones. This allows us to consider co-training without human intervention and parameter tuning.
- We consider the large-scale `VehicleID` and `Veri-Wild datasets`, infrequently used in the prior art, to verify the model's performance in truly large-scale scenarios.

This third solution has been already presented in the InterForensics 2023[2], the largest Forensics conference in Latin America, and it is currently under review in the top-tier IEEE Transactions on Image Processing (IEEE TIP).

4.1 Related Work

4.1.1 Re-Ranking-based approaches

Most Re-Ranking techniques are designed to address rank retrieval during the evaluation phase. Given a query image and the set of gallery images, the initial ranking list is

[2]https://interforensics.com/site/interforensics2023/apresentacao

obtained. Usually, methods take this initial list and rank the samples again based on some strategy to enhance the retrieval.

In [46], the authors define the content and context sets. Given a probe image, they calculate its ranking list with the gallery images. The set with the closest images to the probe is called the content set, which is used with the original ranking list to create the context set and improve the ranking performance.

In [2], the authors use a graph-based on-the-fly affinity learning considering the labeled training, gallery, and probe sets to refine the ranking of the probe to the gallery set. In [155], the authors employ the probe-to-neighbor and neighbor-to-neighbor distances to refine the ranking list by considering an expanded neighborhood from each gallery sample retrieved in the top matches to the probe.

The authors in [121] propose the Spectral Feature Transformation (SFT), where they optimize the model to generate feature representation that optimizes the Min-Cut problem considering the labeled samples at batch level. During the evaluation, given a query sample, they perform SFT in the top-ranked gallery features to improve the retrieval performance.

The most well-known Re-Ranking method is the k-Reciprocal Encoding [245] which has been extensively used for unsupervised re-identification. It improves the feature distances during training and generates clusters with high diversity and true-positive rate. The k-Reciprocal nearest neighbors are calculated for each training sample and respective expanded set. Then Jaccard distances are computed between training points considering each expanded set, and these distances are averaged in the Local Query Expansion step. However, it is time- and memory-consuming [75, 121].

In this context, we propose Local Re-Ranking, which redesigns the neighborhood-based distance calculation to decrease time complexity and keep state-of-the-art performance.

4.1.2 Noise-robust Feature Learning

Large-scale datasets are prone to annotation error due to their size or to the complexity in identifying the positive samples even with human supervision [44]. This introduces noise in the learning process, hindering generalization.

Several works have tackled model learning with noisy labels [44, 54, 53, 161]. They usually estimate the Transition Matrix, which reflects the probability of the samples from one class being misclassified as other classes present in the dataset. With the estimated Transition Matrix, some works [164] propose to change the probability distribution on the final softmax layer or to reweight the final loss function [143] to alleviate the influence of noisy samples. Based on the memorization effect in which Deep Neural Networks first learn from clean samples and then from noisy samples [54, 221, 188], some works have proposed robust regularization through implicit or explicit regularization [49, 200] and co-training methods [54, 188]. The fundamental idea of co-training in noisy-label scenarios is to have two (or more) backbones and select the small-loss samples from one peer backbone to train the other. Usually, the same batch of images is fed to both peers, and just the top-r% small-loss samples are kept from each one. Then one peer optimizes its weights with the small-loss samples from another. Rate r is decreased along the training to keep

fewer samples to train the peers as they start learning from noisy samples due to the memorization effect.

Based on the memorization effect, we propose to control the tightness of the generated clusters by changing the density criteria in a novel manner. But instead of selecting small-loss samples as done previously, we consider the presence of noisy labeling in the feature space.

4.1.3 Co-training for Person Re-Identification

Despite the progress in noise-robust feature learning, most methods assume data is annotated. Instead, we consider unlabeled data and propose a clustering-based solution to generate pseudo-labels for model fine-tuning. Since the backbones have not been pre-trained in any other ReID-related dataset (just in ImageNet), the features are naturally noisy, which results in imperfect clustering and noisy labels.

As already mentioned in chapter 2, prior Unsupervised Person ReID (UPReID) art also faces the same problem, and some works apply co-training to deal with noisy labeling. MMT [47], MEB-Net [226], and PEG [225] share the same principle of multiple models learning from each other's hard and soft labels at the batch level to encourage knowledge decoupling and robustness to noise. These methods employ complex co-training strategies with a lot of terms and hyperparameters in their loss functions or in the selection of the peer networks, which might be challenging to tune and deploy in large-scale unlabeled scenarios.

We propose a simpler co-training strategy, by permuting the predicted labels among the peers instead of performing soft/hard cross-supervision. We take advantage of co-training, but with a hyperparameter-free strategy that does not require any manual or grid-searching-based parameter selection.

4.1.4 Unsupervised Re-Identification

We have already reviewed some prior art in Chapter 3 for fully unsupervised ReID. Here we recap some of them and include new methods that, by the time the second solution was proposed, had not been proposed.

To tackle unsupervised re-identification, some methods rely on pre-training using a source ReID dataset [18, 7, 47, 226] to acquire prior knowledge. Other methods, like IICS [204], CAP [186], CASTOR [202], and PPSL [198], rely on other information, such as camera labels. Since our method is fully unsupervised, i.e., it **does not** rely on camera labels, we focus on methods operating in the same setup. We present in Appendix C a comparison table to methods considering camera labels or any other side information.

ICE [17] has two versions: camera-aware and camera-agnostic. The first considers camera proxies for each cluster. The second considers only a cluster proxy, which is a feature average regardless of camera labels. They use a proxy-based loss, and hard- and soft-instance losses. SpCL [48] uses a self-paced strategy that introduces some metrics to measure cluster reliability: cluster independence and compactness. If both are higher than predefined thresholds, the cluster is kept within the feature space. RLCC [230] refines

clusters by a consensus among iterations. Pseudo-labels are created by considering the ones generated in previous iterations, keeping the training stable. CACL [96] proposes to suppress the dominant colors in images, providing a more robust feature description, and a novel pseudo-label refinement method. CCons [30] uses contrastive learning, with which they select the hardest cluster sample in the batch to update the cluster centroid. ISE [234] synthesizes novel feature examples from real ones to refine sample distribution, aiming to generate clusters with a higher true positive rate, as well as avoiding subdivision of the samples from the same class in different clusters. PPLR [26] employs a part-based model that creates feature spaces from different parts of the feature map. In each one, they calculate the nearest neighbors of the samples and propose a cross-agreement metric to refine the proposed pseudo-labels. GRACL [229] keeps two memory banks where one holds the sample, from each cluster, most dissimilar from other clusters, and the other holds the sample most dissimilar from the positive samples. This approach encourages compactness and separability. AdaMG [144] performs clustering with different density parameters to generate multiple pseudo-labels. Then a teacher-student model is trained considering each set of pseudo-labels, and the cluster feature memory bank is updated based on sample reliability, alleviating noisy labeling impact.

For Unsupervised Vehicle Re-Identification (UVReID), MLPL [62] adopts a multi-level feature description by extracting a global feature and four local features for clustering and feature learning. They also propose converging and promoting stages to learn from global and local features separately and jointly, based on the consistency score of the global- and local-based clustering results. RLCC [230] and PPLR [26] are also evaluated under the UVReID scenario. Our method differs from the others by proposing a regime to adapt a noise-robust density parameter over time, alleviating the burden of choosing optimal hyperparameters, and a co-training strategy to avoid error amplification by the backbones.

4.2 Proposed Method

In this section, we provide the rationale for each part of the proposed pipeline (Figure 4.1) to tackle large-scale unlabeled Re-Identification under noisy labeling. We assume we have $m = 3$ backbones, but it can be extended to any $m \geq 2$.

4.2.1 Self-supervised Initialization

Usually, previous methods adopt ResNet50 [58] pre-trained on ImageNet [31] for model initialization. Despite ImageNet being a general dataset for classification, the pre-trained backbone might not produce discriminant features for clustering in the UPReID and UVReID scenarios. More recently, self-supervised pretext tasks have been widely explored for initialization [13, 57, 22, 14, 222] and further application in downstream tasks. In this context, we propose pre-training the model on the target unlabeled dataset before applying our pipeline. Since we assume a fully unlabeled scenario, we leverage a self-supervised pre-training based on Barlow Twins [222]. We perform this initialization step for all backbones.

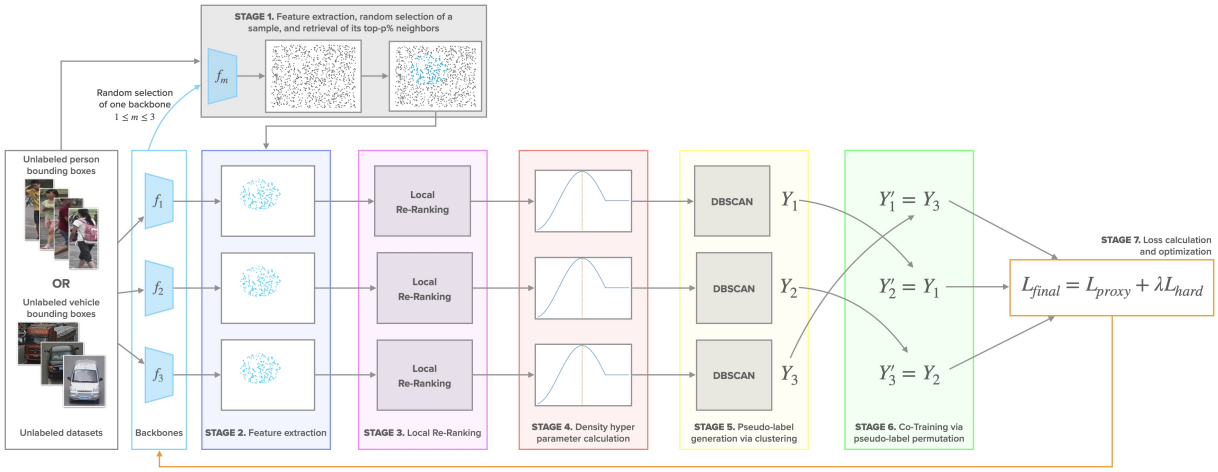


Figure 4.1: Overview of our solution. In Stage 1, we first extract features for the entire unlabeled training set utilizing one randomly selected backbone. Then we randomly select a sample and obtain the top-p% closest neighbors to define the local neighborhood. In Stage 2, features are extracted for the selected samples, with all backbones. In Stage 3, we employ Local Re-Ranking to refine the distances based on the local neighborhood of the samples, keeping a low memory and time footprint. In Stage 4, we select the current density parameter ε based on the proposed noise-robust density scheduling scheme. Then clustering is performed in Stage 5. Once we have the pseudo-labels predicted by each backbone in the pipeline, we permute the pseudo-labels set among the backbones to allow cross-supervision (Stage 6). The illustrated permutation ($Y_1^{'} = Y_3$, $Y_2^{'} = Y_1$, $Y_3^{'} = Y_2$) is an example, and other permutations are possible as long as a backbone is not supervised by its own pseudo-labels. Finally, in Stage 7, the loss function is optimized. Best viewed in color.

Given a batch of randomly selected images $B \in X$, the algorithm generates two augmentation views for each image through some common image transformations, resulting in two new augmented batches B^1 and B^2. In our case, some image transformations might degrade Re-Identification performance, so we consider just random cropping, horizontal flipping, and shifts in brightness, contrast, and saturation. After that, we feed both batches to the backbone f and get the feature representations $Z^1, Z^2 \in R^{N \times d}$, respectively, for each batch, where d is the feature dimension. The features are normalized by mean subtraction and diving by the standard deviation per dimension. After that, we train the model to achieve maximum decorrelation between the dimensions to encourage the model to learn complementary features and, at the same time, be robust among different augmentations. To do so, the algorithm performs the multiplication of batch features, $C = (Z^1)^T Z^2 \in R^{d \times d}$, to calculate the cross-correlation among the different dimensions of the feature vectors. The next step is to maximize the agreement between features in the same dimension (represented by values in the main diagonal of C) and minimize it between features from different dimensions through the following loss function:

$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda_{BT} \sum_i \sum_{j \neq i} C_{ij}^2, \tag{4.1}$$

where the first term aims to increase the invariance representation between different augmentation views of the same image, and the second decouples feature representation considering the dimensions; and λ_{BT} weights the contribution of the second term in the loss. For further details, we refer to the original article [222].

4.2.2 Local Neighborhood Sampling (LNS)

It is a common requirement that the entire distance matrix, with all training feature vectors, be loaded into memory. However, as the number of considered samples grows, memory and time complexity increase which might lead to a costly deployment in a large-scale scenario. For instance, in [69], the authors reported that a distance matrix of size $n \times n$, with $n = 10^6$, takes up 7450.58 GB of memory. The largest datasets used in prior Unsupervised ReID works are `MSMT17` for Person and `Veri` for Vehicle Re-Identification. Since they have fewer than $40K$ training samples, previous methods might not have faced significant memory and time issues. However, there are several practical scenarios in which larger datasets are required [137]. In these cases, a more efficient solution is required.

To tackle this, given an unlabeled dataset $X = \{x_i^n\}$ with n images, we propose a novel neighborhood-based approach for sampling a subset $X_s \subset X$ (Stage 1, Figure 4.1). More specifically, we first select a backbone f_m ($1 \leq m \leq 3$) randomly from the available backbones to perform feature extraction and obtain the feature vector set F_m (blue points in Stage 1, Figure 4.1). Then we randomly select a feature vector $v \in F_m$ and calculate its top-p% ($0 \leq p \leq 100$) nearest neighbors set in F_m, which defines $X_s \subset X$ (blue points in Stage 1, Figure 4.1). That is, X_s comprises the elements of the whole dataset that are closer to v. The computational cost is related to computing the distances from v to every other vector in F_m and sorting the distance vector. Since cosine distance is used, the complexity is $\mathcal{O}(nd + nlgn)$, where d is the feature dimension. In a practical scenario, we usually have $d \ll n$, and the distance computation is performed in parallel, so the actual complexity is $\mathcal{O}(nlgn)$. Once the set X_s is defined, it is used by all backbones to perform feature extraction and Local Re-Ranking in Stages 2 and 3, respectively (blue points in Figure 4.1). The set X_s is redefined every three pipeline iterations by another randomly selected backbone, which brings diversity to the training.

It is important to note that the nearest neighbors calculation is applied in the feature space for each backbone independently, in different rounds. Therefore, the solution can employ backbones that output features in varied dimensionality, allowing the use of any set of backbones without any alignment. In Section 4.3.2, we show the impact of p employed in the calculation of the top-p% nearest neighbors in terms of performance and speedup.

4.2.3 Local Re-Ranking

Prior art usually employs the k-Reciprocal Encoding algorithm [245] (Full Re-Ranking – FRR) to account for the context (neighborhood) of each sample. This helps to compute a more robust distance measure and allows hard positive samples to be closer in the feature space. Despite its effectiveness, we argue it might not be efficient, in terms of memory and space, when considering large-scale scenarios.

To improve efficiency when dealing with large-scale datasets while keeping the advantages of neighborhood-based distance refinement, we propose a new Local Re-Ranking (LRR) algorithm. At a given step, it only considers the local neighborhood of two samples. The idea is to consider samples in common in given neighborhoods, avoiding the full comparison between all samples in the training set. Without loss of generality, consider a feature vector set F_m ($1 \leq m \leq 3$) created after feature extraction by one of the back-

bones in Stage 2. We first calculate the k-Nearest Neighbors set $N(x_i, k)$ for each sample $x_i \in F_m$, and the local distance matrix $D_{loc} \in R^{n \times k}$, where the i-th line is the distance between x_i and each element in $N(x_i, k)$ after applying the exponential decay function. That is, $D_{loc}(i, j) = e^{-d(x_i, x_j)}$, where $x_j \in N(x_i, k)$ and $d(.,.)$ is the Euclidean distance. The lower (or greater) the distance, the closer to one (or zero) they are after the exponential decay transformation. After that, we employ our neighborhood calculation based on the Jaccard distance to refine the distance between each x_i and its nearest neighbors. For each sample $x_j \in N(x_i, k)$ we calculate the following sets:

$$I(x_i, x_j) = \{p | p \in N(x_i, k) \wedge p \in N(x_j, k)\}, \tag{4.2}$$

$$E(x_i, x_j) = N(x_i, k) \backslash I(x_i, x_j). \tag{4.3}$$

$I(x_i, x_j)$ is the Inclusion set, which contains the common neighbor elements, and $E(x_i)$ is the Exclusion set, which contains elements in $N(x_i, k)$ but not in $N(x_j, k)$. Following [245], we assume the greater the cardinality of $I(x_i, x_j)$, the more likely x_i and x_j are samples from the same class. In light of this, we propose the following to refine the distance:

$$s_{min} = \sum_{p \in I(x_i, x_j)} min(D_{loc}(i, p), D_{loc}(j, p)), \tag{4.4}$$

$$s_{max} = \sum_{p \in I(x_i, x_j)} max(D_{loc}(i, p), D_{loc}(j, p)), \tag{4.5}$$

$$s(x_i, x_j) = s_{i,j} = \sum_{p \in E(x_i, x_j)} D_{loc}(i, p), \tag{4.6}$$

$$s(x_j, x_i) = s_{j,i} = \sum_{p \in E(x_j, x_i)} D_{loc}(j, p), \tag{4.7}$$

$$D_{IoU}(x_i, x_j) = \frac{s_{min}}{s_{max} + s_{i,j} + s_{j,i}}, \tag{4.8}$$

where s_{min} is the sum of the minimum values when comparing the distances of x_i and x_j to a common neighbor p (the same for s_{max} but considering maximum values), and $s_{i,j}$ (or $s_{j,i}$) is the sum of the distances between sample x_i (or x_j) and its neighbors that are not in the intersection $I(x_i, x_j)$.

In extreme cases, when x_i and x_j have all their neighbors in common, we have $I(x_i, x_j) = N(x_i, k) = N(x_j, k)$, $E(x_i, x_j) = E(x_j, x_i) = \emptyset$, $s_{i,j} = 0$ and $s_{j,i} = 0$, and Equation 4.8 becomes $0 \leq D_{IoU}(x_i, x_j) = s_{min}/s_{max} \leq 1$. Conversely, when they do not have any neighbors in common, we have $I(x_i, x_j) = \emptyset$, $E(x_i, x_j) = N(x_i, k)$, $E(x_j, x_i) = N(x_j, k)$, $s_{min} = 0$, $s_{max} = 0$, $s_{i,j} \neq 0$ and $s_{j,i} \neq 0$, then $D_{IoU}(x_i, x_j) = 0$. Therefore, we see that $0 \leq D_{IoU}(x_i, x_j) \leq 1$, and the closer it is to 1, the more likely it is for x_i and x_j to be from the same class. Finally, to convert D_{IoU} into a distance measure, we define the refined distance matrix R as

$$R(x_i, x_j) = \begin{cases} 1 - D_{IoU}(x_i, x_j), & \text{if } x_i \in N(x_j, k) \\ 1, & \text{otherwise,} \end{cases} \tag{4.9}$$

for each $x_j \in N(x_i, k)$. Note that $R \in \mathbb{R}^{n \times k}$ is used to perform clustering in further steps.

In terms of complexity, we compare LRR and FRR theoretically. In FRR, the authors first compute the k-Nearest Neighbors set for each training sample in $\mathcal{O}(n^2 \log n)$. Then, they calculate the k-reciprocal nearest neighbor set in $\mathcal{O}(nk^2)$, and the incremental set in $\mathcal{O}(nk^3/2)$. The full distance matrix is created based on the incremental set in $\mathcal{O}(n^2)$, and the final Jaccard distances are computed in $\mathcal{O}(n^3/2)$. Finally, they perform the local query expansion in $\mathcal{O}(n^2k)$. Therefore, FRR's complexity is $\mathcal{O}(n^3 + n^2 \log n + n^2k + n^2 + nk^3 + nk^2) = \mathcal{O}(n^3)$. For LRR, we also calculate the k-Nearest Neighbors in $\mathcal{O}(n^2 \log n)$ and distance matrix D_{loc} in $\mathcal{O}(nk)$. The intersection set $I(x_i, x_j)$ and the sets $E(x_i)$ and $E(x_j)$ are computed in $\mathcal{O}(3k^2)$. Equations 4.4 and 4.5 can be calculated in a single pass in $\mathcal{O}(k)$. Equations 4.6 and 4.7 are computed in $\mathcal{O}(k)$, so the final complexity is $\mathcal{O}(3k^2 + 3k)$ for a single pair x_i and x_j. Considering all possible pairs in D_{loc} and the whole training set, the complexity is $\mathcal{O}(3nk^3 + 3nk^2)$. Therefore, LRR's total complexity is $\mathcal{O}(n^2 \log n + nk^3 + nk^2 + nk) = \mathcal{O}(n^2 \log n)$ with $k \ll n$. Our model is more efficient and applicable to large-scale datasets, which is also corroborated by the practical time analysis (Section 2.4).

4.2.4 Noise-Robust Density Scheduling

Previous clustering-based methods often assume strong assumptions. For instance, when using DBSCAN, it is common to use some side information to define ε, such as holding it fixed or calculating it using extra information. As the backbone's weights are constantly changing, creating different feature spaces, a fixed clustering hyperparameter is likely suboptimal. In early iterations, the backbones are strongly biased by the camera view, i.e., feature vectors from images from the same identity and the same camera tend to be grouped together. Conversely, images from the same person from different cameras are farther away than images from different identities but with the same camera. So, if ε is too high, clusters might include false-positive samples from the same camera. As the training progresses, the backbones start to learn camera-invariant representations, and more cross-view images from the same class are included in the clusters, but ε might not be large enough to include diverse examples from different cameras. Hence, methods that consider a fixed ε face a noise/diversity trade-off.

Other unsupervised methods set an optimal ε per dataset [234, 144, 215] because each dataset has a different complexity level. For instance, previous research has reported a lower ε for the `Market1501` dataset, as it has less variability in terms of cameras and identities, and a greater ε for `MSMT17` due to a higher camera diversity. However, this selection is unrealistic since the main assumption is the full absence of labels or any other side information. It is not trivial to find an optimal value when assuming no prior knowledge about the dataset complexity [66, 157]. Aiming to propose a more general approach, closer to a real deployment scenario, we introduce an ε scheduling scheme to address the diversity/noise trade-off during training, without requiring per-dataset tuning.

We employ DBSCAN for clustering samples given their distances in matrix R (Equation 4.9). As distances are normalized between 0.0 and 1.0, the ε value must be selected within this range. We propose a novel ε scheduling scheme, in which it starts from a

low value and gradually increases, following a cosine scheduling, until half of the training epochs, as shown in Figure 4.2. We call this first phase the *warmup*. As ε is progressively increased, we allow the clustering algorithm to consider more cross-view images in a smoother manner. As previously reported by noise-robust feature learning methods [54, 221, 161], the method first learns from clean examples before starting to learn from the noisy ones. Since we start with high true-positive rate clusters in early iterations and smoothly increase the margin, the effect of noisy samples is alleviated even if the noise ratio starts to increase.

If we keep ε in its highest value until the end of the training, the noisy samples are overemphasized, and the backbone overfits [54, 221, 53, 161]. To tackle this, we gradually decrease ε to avoid adding too many noisy samples to the clusters. We call this the *annealing* phase. After the model reaches 75% of training epochs, ε is kept in a plateau until the end of the training, in a *steady state* phase. As the backbone learns to group the cross-view images in the first half of the training, the cross-view image representations tend to be closer at this stage. Thus, the backbone can keep learning from diverse samples as ε is gradually decreased, alleviating the impact of the noisy samples. While our experiments use a fixed number of epochs, the scheduling would also allow a temporal stopping criterion. We consider 40 training epochs for fair comparison to prior works that usually employ around this number of epochs for training.

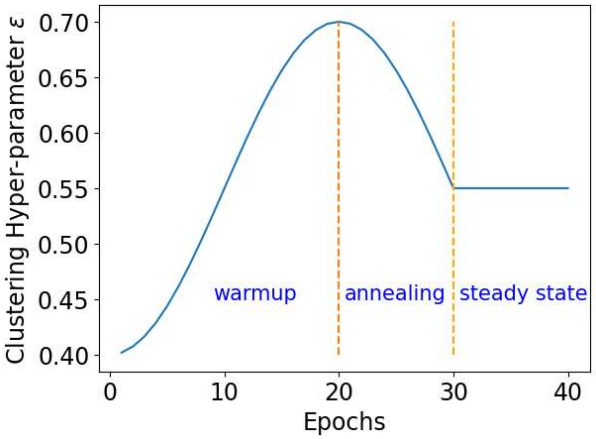


Figure 4.2: Noise-Robust Density Scheduling during training for the ε parameter. First, a *warmup* is performed to address the diversity/noise trade-off. In the *annealing* phase, we decrease ε to reduce the influence of noisy data while keeping the diversity within the clusters. In the last phase, the *steady state*, we keep the same ε value until the end to encourage a stable behavior.

4.2.5 Co-training

After defining ε for the current epoch and performing clustering for each of the M backbones, we have a sequence Y of M pseudo-label sets. It is defined as $Y = (Y_m)_{m=1}^{M}$, where Y_m is the set of pseudo-labels generated by clustering the features extracted by the m-th backbone.

In previous Person Re-Identification works [47, 226, 225], co-training is considered but with a mechanism involving hard and soft supervision among the backbones, often

leading to a complex loss function and optimization process. In our case, we propose a parameter-free co-training strategy by permuting the generated pseudo-labels among the backbones. Formally, we generate a random permutation $Y^{'}$ of the sequence Y, giving that $Y_m^{'} \neq Y_m$, i.e., a backbone should never be supervised by its own pseudo-label set.

After that, each backbone f_m is trained with a permuted pseudo-label set $Y_m^{'}$ (Stage 6, Figure 4.1). As an illustrative example, let $Y_1 = (-1, 0, 1)$ be the pseudo-labels generated by f_1 and $Y_2 = (0, 0, 1)$ by f_2. If they are permuted, then $Y_1^{'} = (0, 0, 1)$ and $Y_2^{'} = (-1, 0, 1)$ are the sets carried on to the next stage for f_1 and f_2, respectively. This encourages complementary knowledge sharing among the backbones and alleviates error amplification since a backbone is supervised by one of other $M - 1$ pseudo-labels sets. In this manner, one model has the chance to learn from each other's knowledge. This solution outperforms all co-training techniques used in PReID methods, in the most complex scenarios. The permuted labeling is employed in the next stage to optimize the loss functions.

4.2.6 Optimization and self-ensembling

The loss function (Stage 7, Figure 4.1) comprises two terms: the proxy loss L_{proxy} and the hard loss L_{hard}. For one backbone f_m, we first randomly select one sample per cluster in $Y_m^{'}$ to be the proxy of the cluster. As $Y_m^{'}$ contains samples from c_m different clusters, we have the set $P_m = \{p_m^1, ..., p_m^{c_m}\}$ with c_m proxies. Given a batch B of norm-1 image feature vectors, we calculate the proxy loss as follows:

$$L_{proxy} = -\frac{1}{|B|} \sum_{v \in B} log \frac{exp(v \cdot p_m^+ / \tau)}{\sum_{q \in P_m} exp(v \cdot q / \tau)}, \tag{4.10}$$

where p_m^+ is the proxy of the cluster of feature vector v, $v \cdot q$ is the dot product between v and q, and τ is a temperature hyperparameter to control the shape of the distribution.

Moreover, we consider the hard-positive and hard-negative samples in the batch [64] to increase feature representation:

$$L_{hard} = -\frac{1}{|B|} \sum_{v \in B} log \frac{exp(v \cdot v_{pos} / \tau)}{exp(v \cdot v_{pos} / \tau) + exp(v \cdot v_{neg} / \tau)}, \tag{4.11}$$

where v_{pos} is the furthest sample in v's cluster and in B, and v_{neg} is the closest sample to v in B from a different cluster.

L_{proxy} is a more global loss because it considers all proxies in the calculation, while L_{hard} enforces a more local view by considering just in-batch samples. The final loss function is

$$L_{final} = L_{proxy} + \lambda L_{hard}, \tag{4.12}$$

where λ controls the contributions of L_{hard}. Our solution's sensibility to parameters τ and λ is shown in Section 4.4.5.

To avoid noise amplification during training, we also perform the self-ensembling of each model's weights as usually done in previous methods [57, 14] and in our second solution. For each backbone f_m, we keep a self-ensembled model with parameters Θ_m^t. They are updated as $\Theta_m^t := \beta \Theta_m^{t-1} + (1 - \beta)\theta_m^t$, where θ_m^t are the parameters of backbone

f_m optimized by Equation 4.12, and β is an inertia hyperparameter set to 0.999 as in [57].

4.2.7 Inference

After the training pipeline, the inference is made by ranking all gallery samples based on the distance to a query sample. We extract feature vectors for all *query* and *gallery* sets using the self-ensembled models. These sets are denoted by F_q^m and F_g^m, respectively, with $1 \leq m \leq M$. We calculate pairwise distances between samples of F_q^m and F_g^m, resulting in a distance matrix $D_{q2g}^m \in \mathbb{R}^{|F_q^m| \times |F_g^m|}$.

A final distance matrix $\overline{\mathrm{D}}_{q2g}$ is obtained by averaging all matrices element-wise:

$$\overline{\mathrm{D}}_{q2g} = \frac{1}{M} \sum_{m=1}^{M} D_{q2g}^m. \tag{4.13}$$

Each row of $\overline{\mathrm{D}}_{q2g}$ holds the distances from a query to the gallery samples. We sort these distances to get the closest class to the query sample. As done in previous works, we remove gallery images with the same class and camera of the query to assess performance in a true cross-camera scenario.

4.3 Experiments and Results

In this section, we present the datasets, metrics, and results of our proposed pipeline compared to the prior art in the fully unsupervised Person and Vehicle ReID problem.

4.3.1 Datasets and Implementation Details

We evaluate our method in two Person ReID datasets, `Market1501` [242] and `MSMT17` [194], and three Vehicle ReID datasets, `Veri` [112], `VehicleID` [108], and `Veri-Wild` [119], which are described in Appendix B. The last two are large-scale datasets with more than $100K$ images in the training set and with three evaluation scenarios. `Veri-Wild` [119] is the most challenging one compressing 174 cameras while `Market1501`, `MSMT17`, and `Veri` compress six, fifiteen and twenty cameras respectively. The number of cameras in `VehicleID` is not informed.

For evaluation, we calculate the Cumulative Matching Curve (CMC), from which we report Rank-1 (R1), Rank-5 (R5), Rank-10 (R10), and mean Average Precision (mAP).

We adopt ResNet50 [58], DenseNet121 [70], and OSNet [251] pre-trained in ImageNet [31] as our backbones. We use Pytorch [142], Torchreid [250], and FAISS [84] as supporting libraries. We first pre-train all backbones in each Person/Vehicle ReID dataset using the self-supervised Barlow Twins strategy. We randomly select one backbone to extract features and obtain the top-p% (p can be 25%, 50%, 75% or 100%) samples in LNS, every three epochs. We linearly warm up the learning rate from $3.5e^{-5}$ to $3.5e^{-4}$ in the first 10 epochs and keep it fixed for the remaining training in a total of 40 epochs. We use the Adam [88] optimizer with weight decay set to $5e^{-4}$. The batches are created by randomly sampling 16 pseudo-identities (clusters) and 12 images from each. We sample

batches until all pseudo-identities are covered, then we repeat this process five times before going to the next epoch. Before each epoch, we renew the class proxies. The parameter τ is set to 0.04 and λ to 0.5 in all experiments.

We trained the models in five TITAN RTX GPUs with 24GB of memory each. One GPU is left just for Local Re-Ranking and clustering, and the others are used for training.

We perform the self-supervised initialization with Barlow Twins for three epochs. The setup for each experiment may vary due to the availability of GPUs and memory constraints, as the datasets have different sizes. For `MSMT17`, `Veri776`, and `Veri-Wild`, we used five NVIDIA RTX A6000 with 49GB of RAM, and a batch size of 1024. For `Market` and `VehicleID`, we keep the same setup, except that for `Market`, we train OSNet in three Quadro RTX 8000 with a batch size of 768, and for `VehicleID`, we set the batch size to 768 for OSNet and DenseNet121. The Adam [88] optimizer was employed with a learning rate set to $3.5e-5$ for the Person ReID datasets and $3.5e-6$ for Vehicle ReID datasets. The weight decay was set to $1.5e-6$, and the λ_{BT} in Equation 4.1 was set to $5e-3$ for all datasets following the original implementation [222].

4.3.2 Comparison to State-of-The-Art methods

Table 4.1 shows our method compared to prior fully-unsupervised Person ReID models that, **like ours, do not consider camera labels**. Results for other methods that use side information are in Appendix C. We flag the methods that tune the clustering parameter per dataset because they cannot be used in a realistic fully unsupervised scenario (column *CPD*).

AdaMG [144] achieves good results in `Market` and `MSMT17` by setting $\varepsilon = 0.5$ and $\varepsilon = 0.7$, respectively. But for the same ε for both datasets, they face a huge performance drop [144]. With our ε scheduling scheme, we outperformed AdaMG by 1.2 and 0.1 percentage points (p.p.) in mAP and R1, respectively, in `Market`, and by 5.2 and 4.6 p.p. in the challenging `MSMT17`. Regarding the other metrics, we rank in the third place in R5 being marginally below AdaMG by 0.2 p.p., and by 0.4 p.p. in R10. However, AdaMG adopts an unrealistic scenario where the clustering hyperparameter is tuned per dataset. Moreover, AdaMG adopts a more memory- and time-complex Re-Ranking strategy.

Due to the proposed Local Re-Ranking, our model has a lower Re-Ranking memory footprint ($\mathcal{O}(kN)$ with $k \ll N$) compared to the best methods HHCL, GRACL, and AdaMG ($\mathcal{O}(N^2)$). Our method also outperforms all other methods [133, 244, 37, 167, 216, 139, 179, 134, 220, 170, 223, 171, 172, 238, 214, 158, 165, 99, 140, 231, 80, 232, 189] by a large margin. Therefore, with a less complex method and without requiring any per-dataset hyperparameter tuning, we achieve state-of-the-art performance in `Market` in mAP and R1, rank in the top-3 best performances in R5 and R10, and outperform prior art in all metrics in `MSMT17` with 100% and 75% of the data.

Since our method is based on three different architectures, we also compare it to prior ensemble-based methods (Table 4.2). We obtain the second-best performance in `Market`. However, the best method, PEG [225], utilizes 8 backbones, which include ResNet50 and DenseNet121, like ours. Our third backbone, OSNet, has a lower memory footprint compared to the other backbones. They also employ a complex evolutionary-based strat-

Table 4.1: Comparison with relevant fully-unsupervised Person ReID methods. The best result is highlighted in **blue**, the second best in **green**, and the third in **orange**. RRMC means Re-Ranking Memory Complexity and CPD (Cluster Parameter per Dataset) indicates if the method relies on specific clustering parameters per dataset. (p%) means that p% of all data points are sampled in the Local Neighborhood Sampling and used in the current epoch.

				Market				**MSMT17**			
Method	Reference	RRMC	CPD	mAP	R1	R5	R10	mAP	R1	R5	R10
ABMT [18]	WACV'20	$\mathcal{O}(N^2)$	**No**	65.1	82.6	-	-	-	-	-	-
SpCL [48]	NeurIPS'20	$\mathcal{O}(N^2)$	**No**	73.1	88.1	95.1	97.0	19.1	42.3	55.6	61.2
GCL+ [20]	TPAMI'22	$\mathcal{O}(N^2)$	**No**	69.3	89.0	94.6	96.0	22.0	47.9	61.3	67.1
GSam [56]	TIP'22	$\mathcal{O}(N^2)$	**No**	79.2	92.3	96.6	97.8	24.6	56.2	67.3	71.5
RLCC [230]	CVPR'21	$\mathcal{O}(N^2)$	**No**	77.7	90.8	96.3	97.5	27.9	56.5	68.4	73.1
HCLP [243]	ICCV'21	-	**No**	78.1	91.1	96.4	97.7	26.9	53.7	65.3	70.2
ICE [17]	ICCV'21	$\mathcal{O}(N^2)$	**No**	79.5	92.0	97.0	98.1	29.8	59.0	71.7	77.0
CACL [96]	TIP'22	-	**No**	80.9	92.7	97.4	98.5	23.0	48.9	61.2	66.4
HDCRL [24]	TIP'22	-	**No**	81.7	92.4	97.4	98.1	24.6	50.2	61.4	65.7
PPLR [26]	CVPR'22	$\mathcal{O}(N^2)$	**No**	81.5	92.8	97.1	98.1	31.4	61.1	73.4	77.8
RTMem [215]	TIP'23	-	Yes	83.0	92.8	97.4	98.3	32.8	57.1	70.0	74.9
CCons [30]	ACCV'22	$\mathcal{O}(N^2)$	**No**	83.0	92.9	97.2	98.0	33.0	62.0	71.8	76.7
ISE [234]	CVPR'22	-	Yes	**84.7**	**94.0**	**97.8**	**98.8**	35.0	64.7	75.5	79.4
HHCL [68]	NIDC'21	$\mathcal{O}(N^2)$	**No**	84.2	**93.4**	**97.7**	98.5	-	-	-	-
GRACL [229]	TCSVT'22	$\mathcal{O}(N^2)$	**No**	83.7	93.2	97.6	**98.6**	34.6	64.0	75.0	79.3
AdaMG [144]	TCSVT'23	$\mathcal{O}(N^2)$	Yes	**84.6**	**93.9**	**97.9**	**98.9**	**38.0**	**66.3**	**76.9**	**80.6**
Ours (25%)		$\mathcal{O}(kN)$	**No**	-	-	-	-	32.0	60.5	71.0	75.2
Ours (50%)		$\mathcal{O}(kN)$	**No**	-	-	-	-	24.3	50.4	60.6	65.4
Ours (75%)		$\mathcal{O}(kN)$	**No**	82.9	92.6	97.0	97.8	**39.3**	**67.3**	**77.3**	**80.8**
Ours (100%)		$\mathcal{O}(kN)$	**No**	**85.8**	**94.0**	**97.7**	98.5	**43.2**	**70.9**	**80.8**	**84.2**

egy and co-training where the backbones are selected at different moments with different losses supervising each other. Our results were achieved with just three backbones and a simpler and parameter-free co-training. In `MSMT17`, which is more complex than `Market`, we outperform PEG by 1.4 and 1.8 p.p. in mAP and R1, respectively. Therefore, with a simpler strategy, we can still take advantage of the complementary knowledge from different backbones and outperform prior art in more complex scenarios.

In the `Veri` dataset (Table 4.3), we have the best R1 among all methods. Although we obtained the second-best result considering mAP, MSCL has a much lower R1 score, and they present a higher memory footprint for Re-Ranking. We are also a margin better than BUC [104], MMCL [179], and SSML [219].

In the large-scale and challenging `VehicleID` (Table 4.4), we scored first in R1 in the most difficult scenario (TS = 2400) along with CCons [30]. However, CCons leverages a more memory-complex Re-Ranking. In the other metrics and scenarios, we scored first or, at least, second place.

The `Veri-Wild` dataset has been less employed in the fully unsupervised scenario. Usually, prior art utilizes camera labels, as shown in Appendix C. Considering the few methods that, like ours, do not consider camera labels, we outperform them in the three evaluation scenarios in all metrics (Table 4.5). More specifically, in the most challenging setup (`VW-Large`), we outperform prior art by 4.5, 1.7, and 0.6 p.p., in mAP, R1, and R5 respectively, with 75% of the data. Furthermore, we provide the second- and third-best results with 100% and 50% of the data, respectively for all metrics and evaluation

Table 4.2: Comparison with relevant ensemble-based Person ReID methods. "# BB" shows the number of backbones used in training or evaluation, "Src" means if a labeled dataset has been used to initialize the model before training: "M" for `Market1501` and "D" for `DukeMTMC-ReID`. The best result is highlighted in **blue** and the second best in **green**.

			Market			MSMT17		
Method	Reference	# BB	Src	mAP	R1	Src	mAP	R1
ACT [207]	AAAI'20	2	D	60.6	80.5	-	-	-
MMT [47]	ICLR'20	2	D	71.2	87.7	M	22.9	49.2
MEB [226]	ECCV'20	3	D	76.0	89.9	-	-	-
UST [7]	TIFS'21	3	D	78.4	92.9	M	33.2	62.3
ESSL [8]	TIFS'23	3	-	83.4	92.9	-	42.6	68.2
PEG [225]	IJCV'22	8	-	**87.1**	**94.6**	-	**41.8**	**69.1**
Ours		3	-	**85.8**	**94.0**	-	**43.2**	**70.9**

scenarios.

Table 4.3: Comparison with relevant fully-unsupervised Vehicle ReID methods in `Veri776`. The best result is highlighted in **blue**, the second best in **green**, and the third in **orange**. RRMC means Re-Ranking Memory Complexity and CPD (Cluster Parameter per Dataset) indicates if the method relies on specific clustering parameters per dataset. (p%) means that p% of all data points are sampled in LNS at each epoch.

				Veri		
Method	Reference	RRMC	CPD	mAP	R1	R5
SpCL [48]	NeurIPS'20	$\mathcal{O}(N^2)$	**No**	36.9	79.9	86.8
GRACL [229]	TCSVT'22	$\mathcal{O}(N^2)$	**No**	39.4	82.9	-
RLCC [230]	CVPR'21	$\mathcal{O}(N^2)$	**No**	39.6	83.4	88.8
CCons [30]	ACCV'22	$\mathcal{O}(N^2)$	**No**	40.8	**86.2**	**90.5**
AdaMG [144]	TCSVT'23	$\mathcal{O}(N^2)$	Yes	41.0	**86.2**	**90.6**
RTMem [215]	TIP'23	-	Yes	**41.8**	81.6	87.0
MSCL [191]	SVIP'22	$\mathcal{O}(N^2)$	**No**	**45.9**	81.2	-
Ours (25%)		$\mathcal{O}(kN)$	**No**	28.0	66.8	72.2
Ours (50%)		$\mathcal{O}(kN)$	**No**	40.8	**84.5**	88.6
Ours (75%)		$\mathcal{O}(kN)$	**No**	**41.7**	**86.2**	**89.9**
Ours (100%)		$\mathcal{O}(kN)$	**No**	41.3	**86.3**	**89.9**

4.3.3 Results and Speedup with LNS

We evaluate our proposed LNS, for which we can take different percentages from the whole data at each iteration. In Table 4.1, we show our results for fully-unsupervised Person ReID with different amounts of data.

With 75% of the data, we obtain the second-best result for `MSMT17` against other methods that use the whole dataset. When we reduce it to 50% we face a natural performance

Table 4.4: Comparison with relevant fully-unsupervised Vehicle ReID methods in `VehicleID`. The best result is highlighted in **blue**, the second best in **green**, and the third in **orange**. RRMC means Re-Ranking Memory Complexity and CPD (Cluster Parameter per Dataset) indicates if the method relies on specific clustering parameters per dataset. Methods with * were reproduced from [62], which seems to follow the same evaluation protocol as ours. (p%) means that p% of all data points are sampled in the Local Neighborhood Sampling and used in the current epoch.

				Test size = 800			**Test size = 1600**			**Test size = 2400**		
Method	Reference	RRMC	CPD	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5
BUC [104]*	AAAI'19	-	**No**	51.8	49.5	62.6	46.2	45.9	59.8	42.4	39.7	57.3
MAC [256]	KBS'22	-	**No**	56.2	54.3	71.1	51.9	47.5	66.8	47.4	44.4	**65.9**
SpCL [48]*	NeurIPS'20	$\mathcal{O}(N^2)$	**No**	60.2	55.4	67.5	58.7	53.1	67.1	54.3	**48.9**	64.8
CCons [30]*	ACCV'22	$\mathcal{O}(N^2)$	**No**	**62.6**	**57.7**	**68.0**	**60.3**	**54.0**	**67.9**	**57.1**	**50.1**	**65.9**
	Speedup											
Ours (25%)	4.97×	$\mathcal{O}(kN)$	**No**	61.0	55.1	68.0	59.0	52.3	67.2	55.6	48.5	64.3
Ours (50%)	2.44×	$\mathcal{O}(kN)$	**No**	61.0	55.0	**68.1**	**59.2**	52.8	67.1	56.0	**48.9**	64.5
Ours (75%)	1.31×	$\mathcal{O}(kN)$	**No**	**61.6**	**55.7**	**68.6**	**59.7**	**53.3**	**67.6**	**56.6**	**49.6**	**65.0**
Ours (100%)	1.00×	$\mathcal{O}(kN)$	**No**	**61.7**	**56.0**	**68.6**	**59.7**	**53.4**	**67.7**	**56.9**	**50.1**	**65.2**

Table 4.5: Comparison with relevant fully-unsupervised Vehicle ReID methods in `Veri-Wild`. The best result is highlighted in **blue**, the second best in **green**, and the third in **orange**. RRMC means Re-Ranking Memory Complexity and CPD (Cluster Parameter per Dataset) indicates if the method relies on specific clustering parameters per dataset. Speedup values are measured in comparison to the version with 100% of the data. (p%) means that p% of all data points are sampled in the Local Neighborhood Sampling and used in the current epoch.

				Veri-Wild (Small)			**Veri-Wild (Medium)**			**Veri-Wild (Large)**		
Method	Reference	RR-MC	CPD	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5
BUC [104]	AAAI'19	-	**No**	15.2	37.5	53.0	14.8	33.8	51.1	9.2	25.2	41.6
MMCL [179]	CVPR'20	-	**No**	15.9	40.1	63.5	19.2	39.1	60.4	14.1	33.1	50.4
SSML [219]	IROS'21	-	**No**	23.7	49.6	71.0	20.4	43.9	64.9	15.8	34.7	55.4
	Speedup											
Ours (25%)	7.60×	$\mathcal{O}(kN)$	**No**	28.0	50.4	74.4	23.6	42.2	66.5	18.0	32.2	55.1
Ours (50%)	2.62×	$\mathcal{O}(kN)$	**No**	**29.8**	**53.5**	**76.4**	**25.6**	**45.7**	**69.2**	**19.8**	**35.5**	**58.6**
Ours (75%)	1.48×	$\mathcal{O}(kN)$	**No**	**30.2**	**54.6**	**77.1**	**26.0**	**46.8**	**70.0**	**20.3**	**36.4**	**59.2**
Ours (100%)	1.00×	$\mathcal{O}(kN)$	**No**	**29.9**	**54.1**	**76.6**	**25.8**	**46.4**	**69.2**	**20.0**	**36.2**	**59.0**

drop, but even with 25% of the data, we are still on par or even better than top-tier methods. This shows our method can mine useful patterns even with a reduced amount of data.

In `Veri` (Table 4.3), the performance drops when considering fewer data, but the difference is small even with 50% of the data. We still achieve competitive performance in comparison to previous methods using the whole dataset, and we use a more memory-efficient Re-Ranking.

In `VehicleID`, performance is stable across all percentages (Table 4.4). In the most challenging evaluation scenario (TS = 2400), the drop in mAP and R1 is 1.3 and 1.6 p.p., respectively, when going from 100% to 25% of the data. This is significantly lower than we observe in smaller datasets, such as `Veri` and `MSMT17`. We expect this to happen since our method was tailored to large-scale scenarios.

In `Veri-Wild` (Table 4.5), our method outperforms the prior art in most metrics with different amounts of data. With 50%, we perform better in all scenarios. With only 25%, we still outperform prior art by 2.2 p.p. in mAP, in a harder scenario (Large), and in all metrics in the other scenarios. With this reduced amount of data, the speedup is 7.6× in comparison to training with the whole dataset. With such a high speedup, it can be

a good trade-off for fast training and deployment. This shows our Local Neighborhood Sampling can effectively reduce the dataset size while keeping superior results in large-scale datasets.

4.3.4 Visualization of Results

Figures 4.3 and 4.4 depict the activated regions for the top-5 retrieved images from the gallery given a query image in `Veri` and `Market` datasets, respectively. In the correct matches, we can see that our model is able to learn fine-grained and point-of-view invariant features. In all images for both Vehicle and Person Re-Identification, just a few regions of the image are strongly activated over the identity (redder regions of the activation maps), showing that our model focuses on specific discriminant parts and is invariant to the background (no activation in any background region). In the failure matches, we see that our model retrieves images with high visual similarity to the query. It would be hard even for a human to tell they are from different identities.

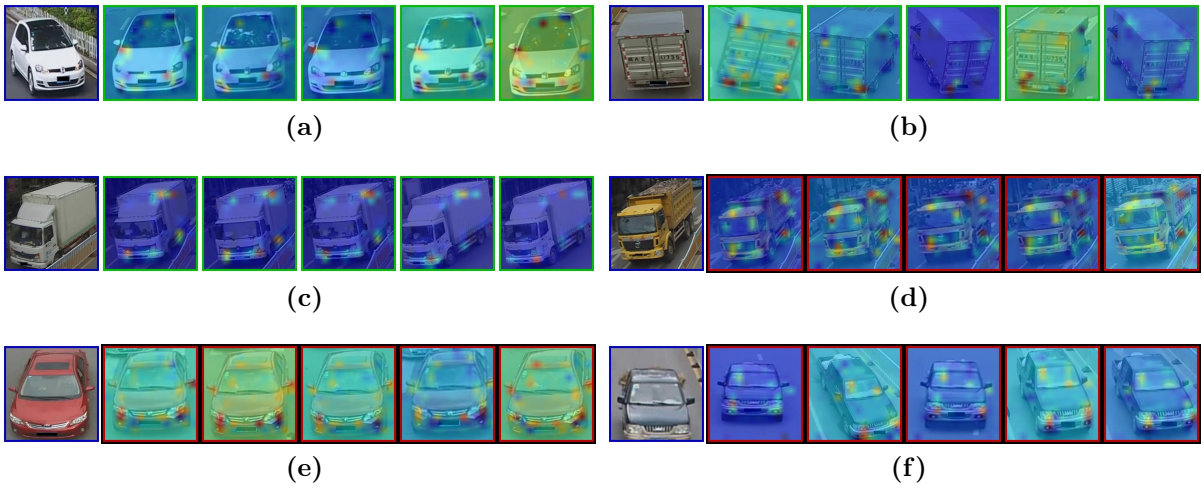


Figure 4.3: Activation maps for the top-5 images retrieved from the gallery, given a query image (blue border) in the `Veri` dataset. Images (a), (b) and (c) show successful cases, and images (d), (e) and (f) show failure cases. The visualizations were generated considering the ResNet50 backbone.

4.4 Ablation Study

We evaluate the impact of each contribution. For more controlled experimentation, when we change one factor of the method, the others remain unchanged.

4.4.1 Impact of Pre-training with Barlow Twins

The impact of employing the self-supervised pre-training with Barlow Twins (BT) is shown in Table 4.6. We compare this initialization against simple pre-training with ImageNet. Pre-training with BT has proved crucial. For `Market`, our method produces state-of-the-art results when the self-supervised initialization is used, but yields degenerated clusters in

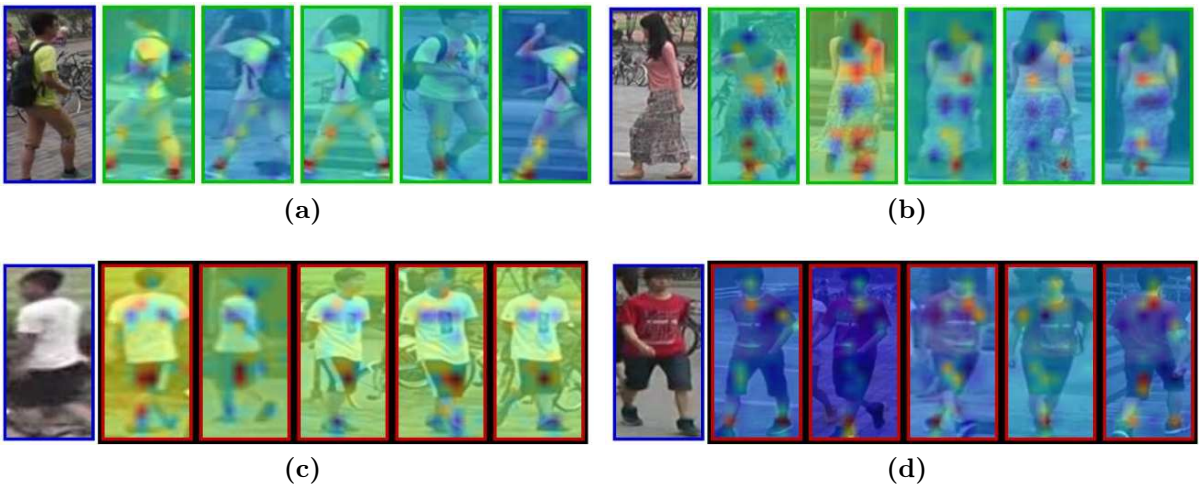


Figure 4.4: Activation maps for the top-5 images retrieved from the gallery, given a query image (blue border) in the `Market` dataset. Images (a) and (b) show successful cases, and images (c) and (d) show failure cases. The visualizations were generated considering the ResNet50 backbone.

the first epochs when BT is not used. For `Veri776`, results are similar in both scenarios; however, for `MSMT17`, mAP and R1 increase by 4.7 and 4.2, respectively, when BT is considered.

Table 4.6: Comparison between our model with and without pre-training with Barlow Twins (BT). The best results are highlighted in **blue**.

	Market		**MSMT17**		**Veri776**	
BT	mAP	R1	mAP	R1	mAP	R1
	-	-	38.5	66.7	**41.3**	86.8
✓	**85.8**	**94.0**	**43.2**	**70.9**	**41.3**	**86.3**

4.4.2 Comparison between Local and Full Re-Ranking

We compare our proposed LRR to FRR in terms of accuracy and time when applied in our proposed pipeline. For fairness, we run both methods on the same machine.

Table 4.7 shows the comparison between LRR and FRR when our contributions are considered: Noise-Robust Density Scheduling (NR-ε), and Co-Training (CT). Our LRR is even more advantageous as the complexity of the dataset increases. When NR-ε and CT are not used (Line #1 vs. line #4 in Table 4.7), we assume we know the best ε per dataset. In this case, for `Market` and `MSMT17`, our LRR is below FRR just by a small margin and surpasses it in the complex `Veri776` dataset. Considering all contributions (line #3 vs. line #6), our LRR surpasses FRR in `MSMT17` and `Veri776`. This shows that we keep a faster and better re-ranking strategy with marginal looses in the `Market` dataset (the smallest one), but better performance for larger datasets such as `MSMT17` and `Veri776`.

We also compare re-ranking execution time when considering LRR vs. FRR (Figure 4.5a). Corroborated by the theoretical time analysis presented in Section 4.2.3, our

Table 4.7: Ablation study. We compare the proposed Local Re-Ranking (LRR) with the Full Re-Ranking (FRR) as well as the influence of the noise-robust ε scheduling (NR-ε) and Co-Training (CT). The underlined results assume an oracle where the best ε for DBSCAN is selected per dataset.

				Market		**MSMT17**		**Veri776**	
		NR-ε	CT	mAP	R1	mAP	R1	mAP	R1
FRR	#1			86.8	94.4	42.7	69.7	39.7	84.6
	#2	✓		87.2	94.4	34.4	64.9	36.5	81.1
	#3	✓	✓	87.8	94.8	39.1	68.3	39.5	84.7
LRR	#4			86.4	94.1	42.5	69.3	41.6	85.1
	#5	✓		84.8	93.3	38.8	67.1	39.8	82.8
	#6	✓	✓	85.8	94.0	43.2	70.9	41.3	86.3

LRR is faster in all datasets, which becomes more evident in the large-scale datasets `VehicleID` and `Veri-Wild`, with a Python implementation (LRR-P). LRR-P is approximately 3.14× and 40× faster in `VehicleID` and `Veri-Wild`, respectively.

Targeting practical scenarios, we also developed a Cython implementation of LRR (LRR-C), which boosts, even more, the time efficiency, as shown in Figure 4.5b. This implementation is an improvement over our own Python implementation of LRR and is 25× and 257× faster than FRR in `VehicleID` and `Veri-Wild`, respectively. This analysis shows LRR is more time-efficient, theoretically and in practice, which is more suitable for large-scale scenarios.

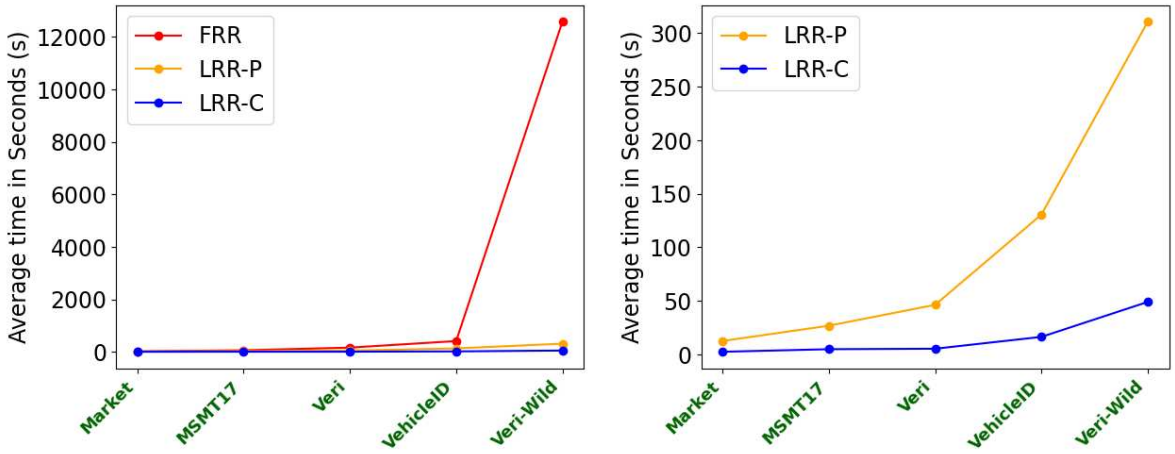
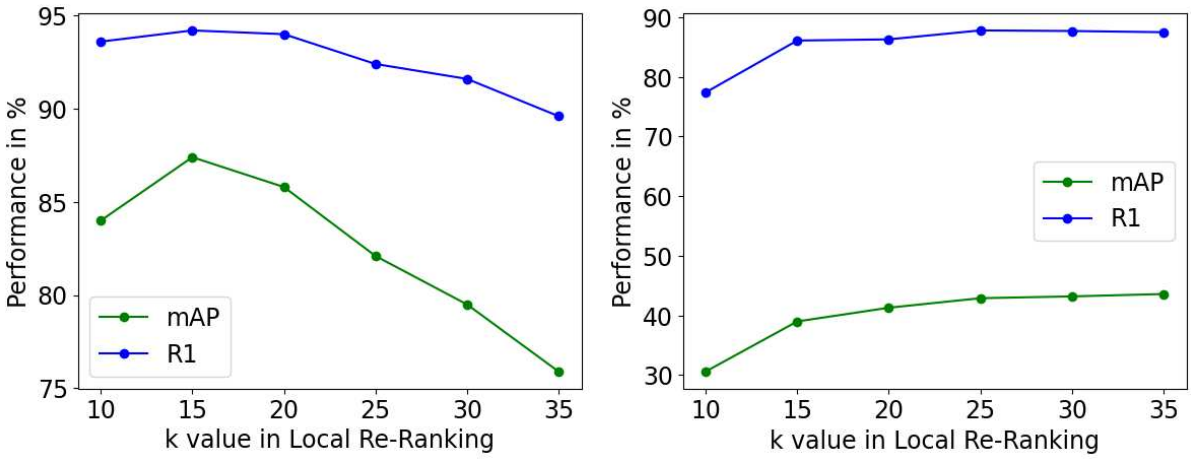


Figure 4.5: (a) Time comparison between Full Re-Ranking (FRR) and the proposed Local Re-Ranking (LRR) in datasets with increasing complexities. LRR is implemented in Python (LRR-P) and Cython (LRR-C). The time is an average of three runs over the three backbones. For better visualization, (b) shows only LRR-P and LRR-C.

Finally, in Figure 4.6, we show the impact of the nearest neighbors parameter k in `Market` and `Veri`, which display contrasting behaviors when k is changed. For `Market`, performance decreases for $k > 20$ but, in `Veri`, performance increases with k. Since our method is fully unsupervised, we keep $k = 20$ for all other experiments and datasets.

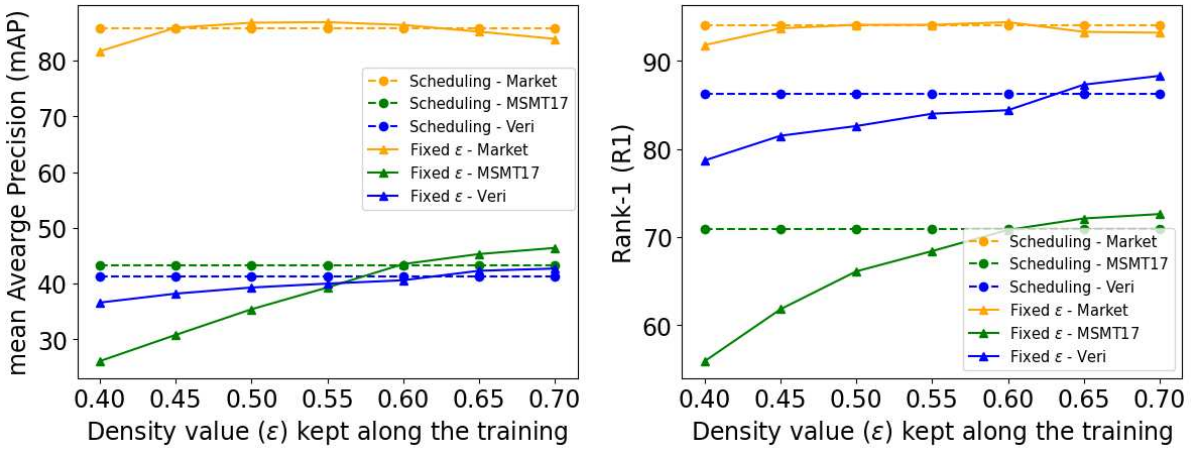


(a) `Market` (b) `Veri`

Figure 4.6: Sensibility of our model to parameter k in the proposed Local Re-Ranking, for datasets (a) `Market` and (b) `Veri`.

4.4.3 Impact of the Noise-Robust Density Scheduling

We verify the impact of the proposed Noise-Robust Density Scheduling. Figure 4.7 shows mAP and R1 in three datasets when using our scheme and with a fixed ε during training.

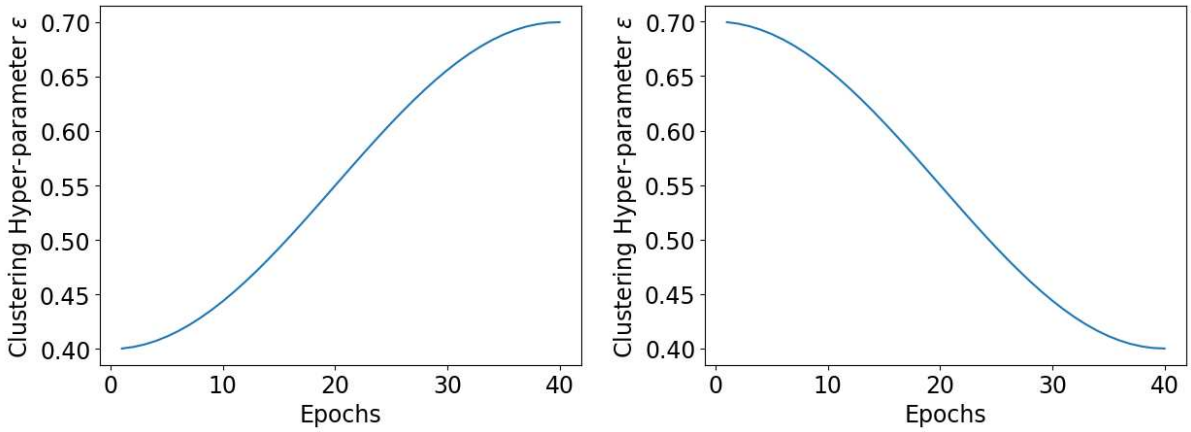


(a) (b)

Figure 4.7: Our method's performance for three datasets, with the Noise-Robust Density Scheduling (Scheduling) and when a fixed density parameter (Fixed ε) is used during training, considering two metrics: (a) mAP and (b) R1.

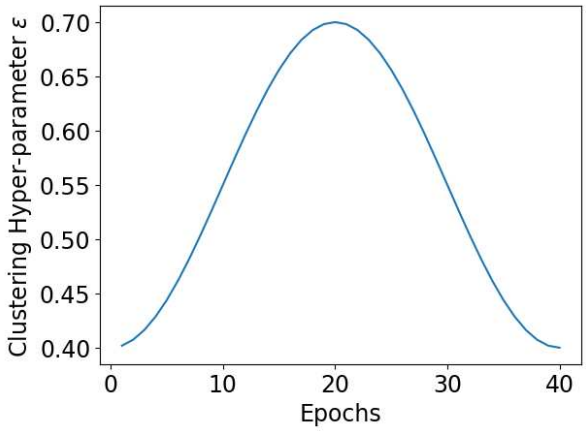
The performance of our proposed scheduling scheme is better than considering fixed ε values for the majority of the tested values. For MSMT17 and Veri, $\varepsilon = 0.65$ or $\varepsilon = 0.7$ are optimal values; however, directly setting those values is unrealistic in fully unsupervised scenarios. Indeed, when we compare our pipeline with and without the Noise-Robust Density Scheduling (line #1 vs. line #2, and line #4 vs. line #5 in Table 4.7), there is a small performance drop since our method does not have oracle knowledge about the optimal clustering parameter for each dataset. Even so, our proposed scheme does not require any grid-searching or manual selection of hyperparameters.

To compare our Noise-Robust Density Scheduling with other possible schemes, we test four alternatives: only the warm-up stage (Figure 4.8a); only the annealing stage (Figure 4.8b); and two adaptations of the cosine learning rate scheduling [118] but applied to the density parameter (Figures 4.8c and 4.8d). The results are reported in Table 4.8.

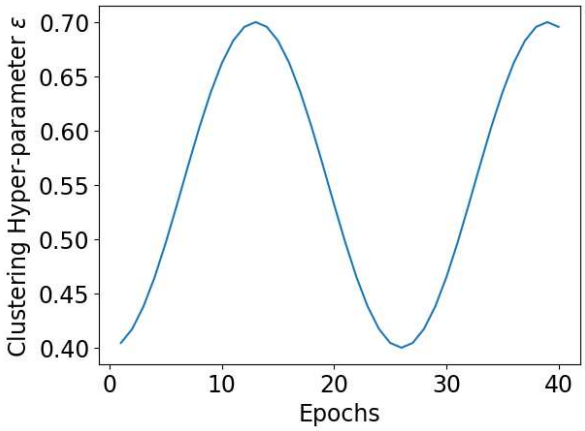


(a) Only warmup stage

(b) Only annealing stage



(c) One cosine cycle



(d) One + half cosine cycle

Figure 4.8: Four alternative ε scheduling schemes: (a) only the warmup phase, (b) only the annealing phase, (c) one cycle of the cosine scheduling, and (d) one and a half cycle of the cosine scheduling.

Despite the fourth alternative (One + half cosine cycle) being better in `Market` and `Veri`, it is possible to verify that the cosine scheduling is sensible to the number of cycles. Our proposed scheduling outperforms the one-cycle cosine scheduling in all metrics in `MSMT17` and `Veri776`, and it is marginally inferior in `Market`. The first two strategies (Figures 4.8a and 4.8b) can be interpreted as part of the cosine scheduling with just a single half cycle (only *warmup* or only *annealing*). Our strategy performs better in most metrics in comparison to them. Therefore, to achieve the best performance with cosine scheduling, tuning the number of cycles is necessary, which is hard to do in a large-scale fully unsupervised scenario. Our scheduling, on the other hand, does not require the number of cycles to be set. Based on the noise-robust learning theory, we designed our scheduling to be directly employed in large-scale fully unsupervised setups.

Table 4.8: Our method's performance with different ε scheduling schemes. The best results are in **blue**.

	Market		**MSMT17**		**Veri776**	
	mAP	R1	mAP	R1	mAP	R1
Only warm up (Fig. 4.8a)	85.3	93.7	41.8	70.0	**42.0**	86.0
Only annealing (Fig. 4.8b)	86.5	94.2	35.9	66.1	39.1	83.7
One cosine cycle (Fig. 4.8c)	86.0	94.1	40.0	69.3	40.2	85.5
One + half cosine cycle (Fig. 4.8d)	**86.8**	**94.4**	43.1	70.7	**42.0**	**87.0**
Ours	85.8	94.0	**43.2**	**70.9**	41.3	86.3

4.4.4 Impact of Co-Training

The impact of the proposed Co-Training strategy is shown in Table 4.7. When we remove the Co-Training strategy (lines #2 and #5) all backbones are supervised by their own generated pseudo-labels, and there is no knowledge sharing among them. We see a performance drop for all metrics in all datasets. The main reason is that a backbone does not have a chance to correct itself based on the knowledge of others. In this case, any noise is propagated and amplified during training without the possibility of recovering. When co-training is in place, this problem is mitigated as discussed in Sections 4.1.3 and 4.2.5. Therefore, our proposed Co-Training yields performance gains for both Re-Ranking strategies without any hyperparameter tuning or human intervention.

4.4.5 Impact of loss hyperparameters

In this section, we verify the impact of the hyperparameters τ and λ in the loss function and its terms (Equations 4.10, 4.11, and 4.12).

The τ parameter has the goal of changing the distribution of the scores, which allows smoother gradients to aid the optimization. Its impact is verified in Figure 4.9 for the `Market` and `Veri776` datasets. We see that mAP and R1 reach their maximum when $\tau = 0.04$. After that, performance starts to decrease rapidly, for both datasets. This happens because the gradients increase together with τ, causing instability during training. Therefore, we set $\tau = 0.04$ for all datasets.

The λ value weights the contribution of the batch-hard triplet loss (L_{hard}) in the final loss function. While L_{proxy} is a more global loss term since it considers all class proxies for optimization, L_{hard} enforces a more local view since the hard triplets are mined at the batch level. That is, $\lambda = 0.0$ means that there is no local contribution, while a too-large value might make L_{hard} dominant over L_{proxy} and hinder model optimization. The impact of the λ value is shown in Figure 4.10.

When $\lambda = 0.0$ (i.e., no L_{hard}), the performance is among the worst for both datasets, which shows the importance of having a local view during optimization. Higher values negatively impact the performance in the `Market` dataset, but it does not affect results in `Veri776`. To achieve a trade-off we keep $\lambda = 0.5$ for all datasets.

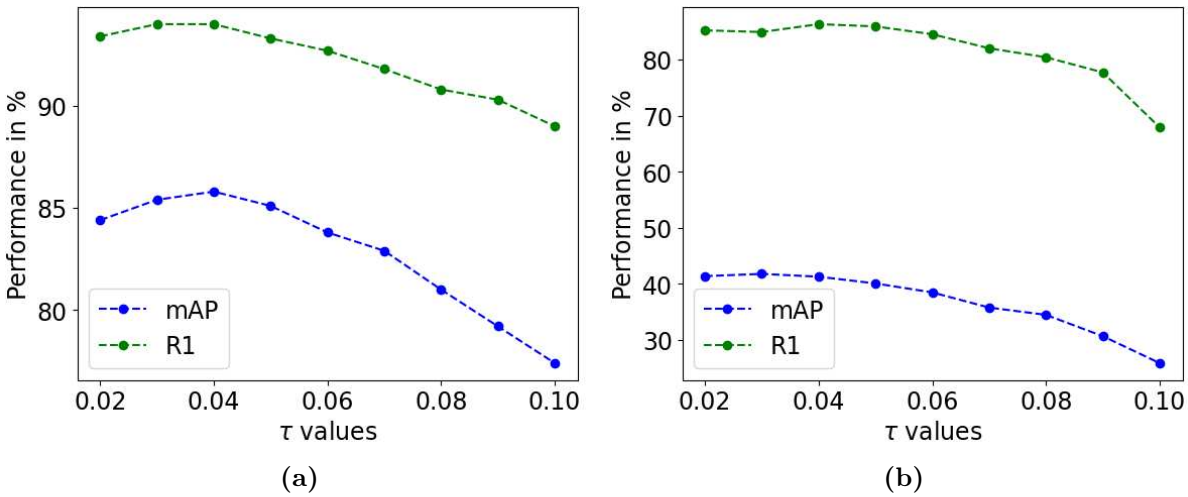


Figure 4.9: mAP and R1 variation for different values of the τ parameter in the loss function, considering the datasets (a) `Market` and (b) `Veri776`.

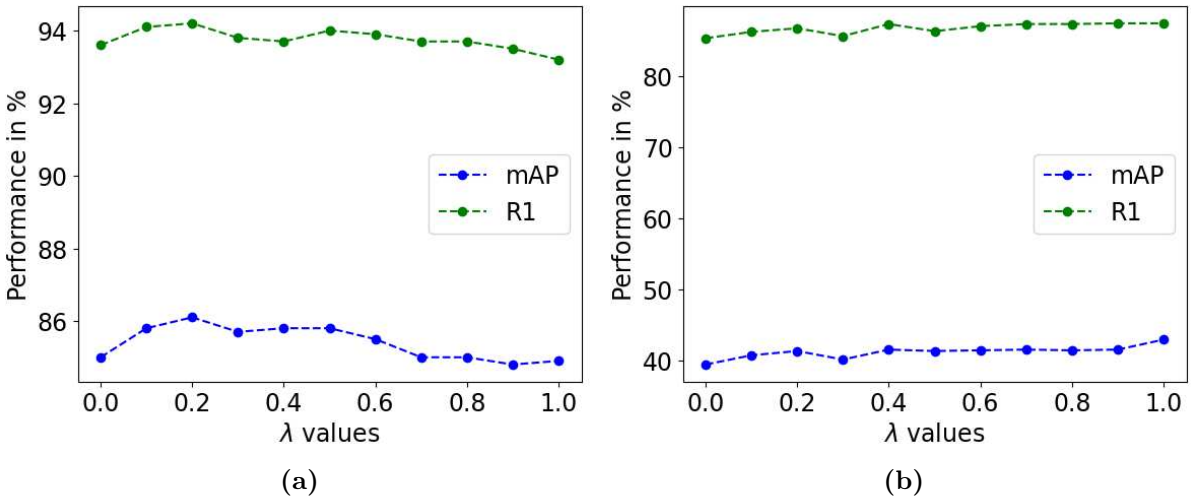


Figure 4.10: mAP and R1 variation for different values of the λ parameter in the loss function, considering the datasets (a) `Market` and (b) `Veri776`.

4.5 Final Remarks

We presented a novel method for fully unsupervised Person and Vehicle Re-identification in large-scale scenarios. Most prior works rely upon costly techniques or consider unrealistic assumptions, making them infeasible for real-world deployment. For instance, they might select dataset-specific hyperparameters, use re-ranking techniques that scale cubically, or adopt ensemble methods that harden the training.

We provide contributions that enable the deployment of re-identification models in large-scale real-world applications, without the necessity for label or dataset-specific information: Local Neighborhood Sampling, Local Re-Ranking, Noise-Robust Density Scheduling, and simple Co-Training.

LNS selects a neighborhood around a random point, reducing the dataset size at each iteration. Local Re-Ranking reduces the memory and time complexities of re-ranking,

by decreasing the amount of data necessary at each step, while still producing superior results. Noise-robust density Scheduling provides parameter-free clustering, taking into consideration the evolution of the backbones during training, which positively impacts the ability of the clustering step to deal with noisy labeling. Finally, our Co-Training technique enables inexpensive knowledge sharing among the backbones.

Our experiments consider datasets that are often not used by other works due to their high complexity. We provide an extensive ablation study showing that our third solution oftentimes provides the best results with fewer assumptions. It also offers a good trade-off between execution time, memory consumption, and accuracy, especially in large-scale datasets.

We apply our method to two different domains—person and vehicle re-identification—which are distinct in terms of the target objects and prominent features. This indicates that our contributions can be applied to other domains with similar characteristics, broadening their application.

One aspect that was only superficially explored is the number of backbones and the considered architectures. Our model relies on ResNet50 and DenseNet121 which could be replaced by lighter models, bringing even more gains in terms of memory and execution time. Another interesting exploration relates to knowledge-distilling techniques, aiming to transfer the knowledge from the ensemble to a single backbone.

As a more advanced future study, to employ our solution in event understanding, leveraging its power also to produce contextual information. This can be done by correlating different types of targets, like people, vehicles, and even places. As these targets are re-identified within an event, we can understand how they are correlated to each other and in time. This is important, for instance, in forensic investigations and in some biometrics applications.

This third solution concludes our trilogy of contributions to the fully unsupervised re-identification task. In the subsequent chapter, we explore a related line of research and delve into the Person Re-Identification task within the expanding domain of long-range recognition in the biometric field.

Chapter 5

DaliID: Distortion-Adaptive Learned Invariance for Identification

Humans can recognize faces or objects before and after considerable distortions. Consider Dali's renowned works *Persistence of Memory* and *Lincoln in Dalivision* shown in part I of 5.1 where the reader will have no trouble recognizing multiple clocks or Lincoln, despite the distorted presentation. Comparatively, neural networks are brittle when presented with even mildly distorted images. Within the field of biometrics, the tasks of Face Recognition and Person Re-Identification can be subject to distortions at inference time, such as atmospheric turbulence, motion blur, and artifacts from upsampling. Such distortions are common in security-sensitive settings such as energy infrastructure security, surveillance systems, or counter-terrorism [74]. Thus, there is a significant social need for models that are robust in these conditions.

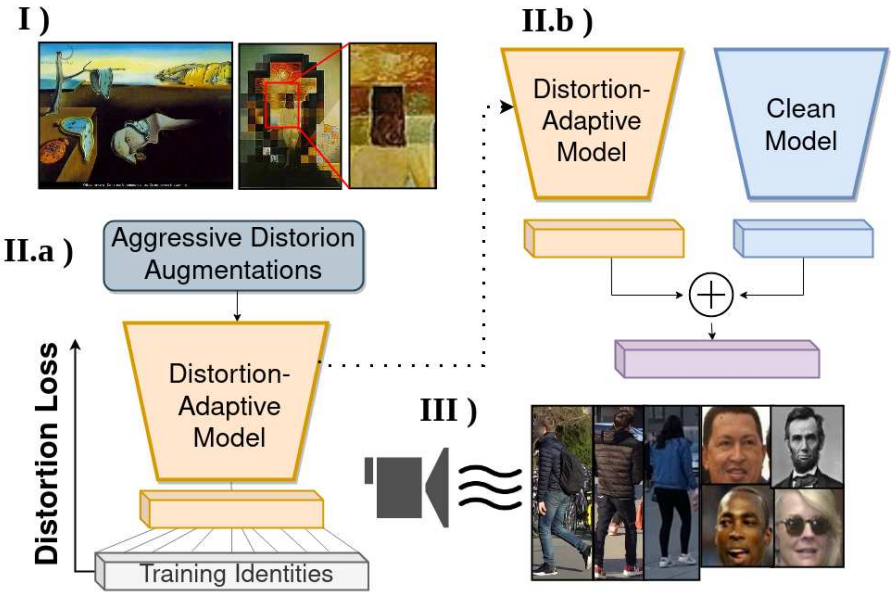


Figure 5.1: To overcome realistic distortions encountered by biometric models operating in unconstrained scenarios, we propose **II.a)** a novel training procedure for distortion-robust models and **II.b)** magnitude-weighted feature-fusion from high- and low-quality training domains. To supplement evaluations on realistic distortions, **III)** we collect and provide an IRB-approved academic-use dataset at a range of 750+ meters.

In this chapter, we presented a co-authored solution to deal with distortions mainly caused by atmospheric turbulence. The work has practical novel updates to training and inference to improve model performance in challenging test-time scenarios. Additionally, to aid evaluation in such scenarios, we collect and provide an IRB (Institutional Review Board)-approved long-distance recognition dataset from over 750+ meters (Appendix D). To demonstrate the generality of the proposed method, we perform experiments with benchmarks for Face Recognition and Person Re-Identification. However, since Face Recognition is not the scope of this thesis, we present the method, discussions, and experiments focusing just on Person Re-Identification. We refer to our article [152] for the full analysis also considering Face Recognition.

The first contribution of this work is a novel *distortion augmentation*, which combines spatial distortion and blur. While prior work [87] has used augmentations such as cropping and down-sampling for Face Recognition and Person Re-Identification, our augmentation contains a more complex transformation that is more closely matched to scenarios with motion blur, atmospheric turbulence, and even upsampling artifacts. Thus, by introducing our distortion-based augmentation in training, test-time domain shift is decreased in challenging scenarios. Our augmentation is performed by leveraging the atmospheric turbulence image simulator proposed in [125]. It is important to note that the authors of the simulator proposed an algorithm to generate simulated data under different levels of atmospheric turbulence, however, they have not employed it as augmentation in any model training. Conversely, we propose to use the simulated atmospheric turbulence data for training employing our designed architecture.

To integrate the augmentation during training, we propose an adaptive weighting mechanism that trains the model in an easy-to-hard manner. Each sample in every batch is reweighted as a function of the training iteration number and the strength of the augmentation. The augmentation's strength (severity) on any given image is sampled from an empirically tuned distribution. In early training iterations, images with higher distortion are assigned lower weighting, and images with lower or no distortion are assigned greater weighting. The weighting of distorted samples is increased throughout training such that by the end of training, all samples have equal weighting. We show that the proposed weighting strategy is highly effective in combination with our distortion augmentation. We refer to a backbone trained with the distortion augmentation and the adaptive weighting schedule as a *distortion-adaptive* model. For Person Re-Identification, we additionally propose to use class centers and multiple class proxies that allow the model to better adapt to training distortions. The corresponding proxy loss (see Section 5.2.2) also follows the adaptive weighting schedule.

To further improve robustness at inference, two backbones are run in parallel: a distortion-adaptive backbone and a standard (or 'clean') backbone. The clean one is trained with images without atmospheric-turbulence simulated data. The final distance between samples for open-set evaluations is calculated with a magnitude-weighted combination of feature distances from each backbone, respectively. Feature magnitude is used since it reflects the response of the learned features at the final layer, which is known to be correlated with sample quality [87, 129, 35]. Maybe surprisingly, this fusion approach is more robust than more complicated learned fusions such as an attention layer or full trans-

former encoder. Relative to a single distortion-adaptive backbone, the parallel backbone fusion improves performance on all person re-identification benchmarks used for evaluation. The final result is a method that is highly robust across evaluation scenarios for both face recognition and Person Re-Identification. We refer to the entirety of our proposed strategy as **DaliID**: **D**istortion-**A**daptive **L**earned **I**nvariance for **Id**entification. It has the Face Recognition version (DaliFace) and the Person Re-Identification version (DaliReID). Since the focus of the research is PReID, we present the method, experiments, and discussions based on DaliReID. The effectiveness of DaliReID is demonstrated empirically, showing it achieves the best performances compared to prior works on two ReID benchmarks already presented in previous chapters (`Market1501` and `MSMT17`), and another ReID benchmark (`DeepChange` [203]) that considers clothing changing in the identities, i.e., the same person can wear different clothes when recorded by different cameras or in different moments in time.

The last contribution of this work is the recapture of face recognition data over long distances with high-end imaging equipment and displays. At 750+ meters, our proposed datasets have the longest range of any academic-use dataset available. We will not go into deeper details since most of the experiments with the recaptured data were conducted for Face Recognition. We refer to our article [152] for further details about the experiments and discussions, and to Appendix D for details about the capturing setup.

In summary, the contribution of this work includes:

- Distortion augmentation, which contains physically realistic spatial distortion and blur.
- Novel distortion-adaptive training strategy in which we leverage the construction of distortion augmentation for an easy-to-hard weighting scheme.
- Novel weighted combination strategy based on the feature magnitudes from both backbones from the training phase, allowing us to exploit complementary knowledge and reach the best performances compared to prior works across evaluation scenarios.
- Identification datasets for long-distance (750+ meters) face recognition, to provide an assessment of the impact of significant atmospheric turbulence (detailed in our published paper [152] and Appendix D).

The method was designed during the international internship of the Ph.D. candidate at the University of Colorado Colorado Springs (UCCS), USA, under the supervision of Prof. Dr. Terrance E. Boult and in partnership with his former master's student (and my friend) Wes Robbins. In this internship, the Ph.D. candidate was a member of the Biometric Recognition and Identification at Altitude and Range (BRIAR) program[1], a United States Government-supported project devoted to counterterrorism, protection of critical infrastructure, and transportation facilities, military force protection, and border security. The article regarding the solution proposed in this chapter has been published in IEEE Access [152], and is also part of an end-to-end identification method that has been published in a joint paper with other BRIAR members in the IEEE IJCB 2023 [39].

[1]https://www.iarpa.gov/research-programs/briar

5.1 Related Work

So far we have reviewed and compared our designed solutions to the Unsupervised Person Re-Identification models. In this chapter, we briefly review the supervised Person Re-Identification works most related to our evaluation scenario, works that have studied image quality assessment for model designing, and the atmospheric turbulence effects in recognition models.

Most of image quality-aware models have been proposed for Face Recognition. CurriculumFace [73] changes the margin of the loss throughout training, and MagFace [129] and AdaFace [87] loss functions use adaptive margins that are a function of feature magnitude, which is a proxy for quality. Controllable Face Synthesis Model (CFSM) [107] is a method that learns the style of a test environment and uses a latent style model to modify training samples.

In [153], the effects of atmospheric turbulence on face recognition are studied, where atmospheric distortions are found to significantly affect face recognition performance. Other works have developed upstream image restoration for atmospheric turbulence [212, 211, 92]. Image restoration methods focus on image-based metrics such as PNSR, not recognition. Besides, image restoration tends to insert artifacts [210, 130, 183] and are challenging to train [50, 210, 25, 183] which can hinder the recognition performance. Inspired by the studies of quality based on feature magnitude, we are the first to employ it for model ensembling for both Face Recognition and Person Re-Identification tasks (Section 5.2.3). Given that we do not know *a prior* the quality level of the images in real-world applications, we use the magnitudes of the features output by each model to weigh the distance calculation between images in query and gallery sets. As each model is trained with images in different qualities they tend to yield different features for the same input image.

For Person Re-Identification (PReID), CBDB-Net [168] proposes the Batch DropBlock to encourage the model to focus on complementary parts of the input image. CDNet [94] improves architecture search for PReID. FIDI [206] proposes a novel loss function to give different penalizations based on distances between images to encourage fine-grained feature learning. To deal with clothes-changing, CAL [51] regularizes the model learning with respect to the clothes labels to learn clothes-invariant features. There are many other prior art that leverage attention models [21, 236, 43, 213, 67, 15, 150, 237, 235, 72, 182, 252, 101, 38] neighborhood-based analysis [182], auxiliary data [61, 77], segmentation-based [85], semantics-based [82, 160] and part-based learning [166, 241, 180, 181, 192, 254, 236, 213, 255, 254, 160]. To directly deal with different resolutions and points of view, some work leverages the camera information associated with each identity [257], super-resolution strategies [79, 25], and attention and multi-level mechanisms for cross-resolution feature alignment [227, 130]. There is insufficient space to compare orthogonally to all combinations of the described methods above for PReID. We limit our scope comparison to the global feature representation learning models as described in the taxonomy of the recent survey from Ye et al. [213] mostly focusing on supervised Person Re-Identification, in which we just perform global pooling operations over the last feature map of a CNN without further mechanisms. The core contributions

of this solution are focused on learning distortion-invariant feature spaces and a methodology for dealing with distortion, which is demonstrated to apply to both face recognition and person re-identification.

5.2 Approach

We propose DaliID for learning models robust to realistic test-time distortions such as motion blur, upsampling artifacts, and atmospheric turbulence. We use strong levels of distortion augmentation (5.2.1), which serves the purpose of supervising the model to learn a feature space that is invariant to distortions that have been shown to considerably degrade model performance [211, 153]. To allow the model to adapt to strong levels of augmentation, we propose an adaptive-weighting distortion-aware strategy (5.2.2) where we dynamically change the weights of different distortion levels throughout training. To get the highest performance across the range of evaluation scenarios, we train two models in parallel: one with clean images and the other with clean and distorted images (5.2.3). Then, we perform a weighted combination of the feature spaces from both models based on the magnitude of the feature vectors from each, which yields the highest performance. DaliID methodology is designed for general identification scenarios such as face recognition and person re-identification tasks. Figure 5.2 shows an overview of the approach.

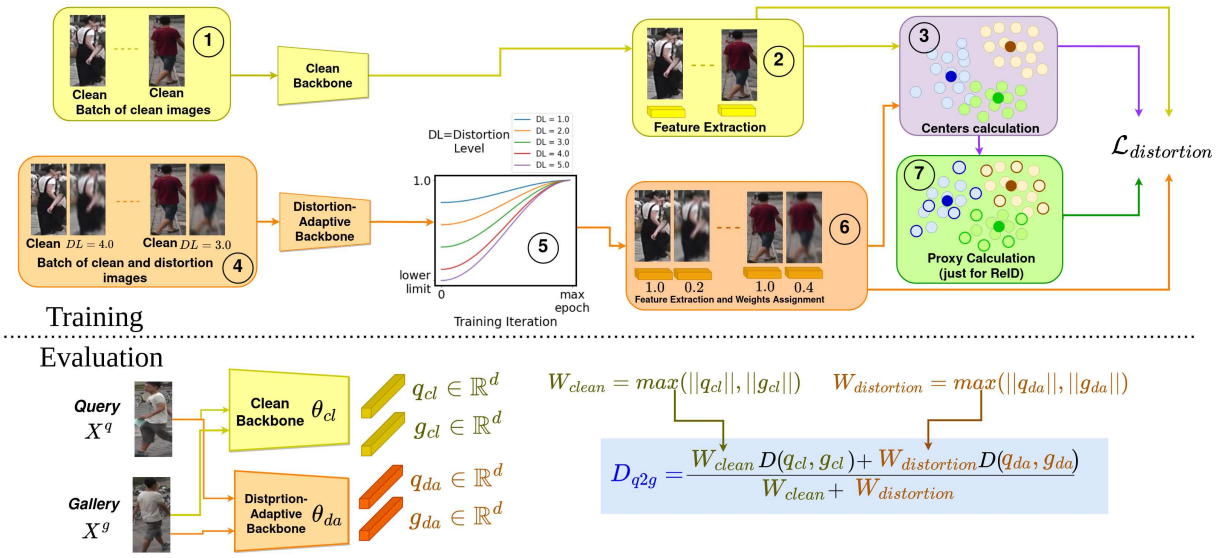


Figure 5.2: An overview of the DaliID pipeline for face recognition (DaliFace) and person re-identification (DaliReID). Steps 1,2 and 3 are performed for training without distorted images, while steps 3,4,5,6 are distortion-adaptive training. In Step 4, we create a batch of clean and distorted images, then a dynamically varied weight is assigned as a function of the distortion level (DL) (Step 5). Then we extract the features and optimize the distortion loss ($\mathcal{L}_{distortion}$). Step 7 is applied just for DaliReID training due to the high intra-class variation faced in the whole-body recognition task. On evaluation, both clean and distortion-adaptive backbone decisions are weighted and combined based on the magnitudes of the query and gallery feature vectors to obtain the final decision (distance) for retrieval.

5.2.1 Distortion Augmentations

Image augmentations allow better generalization by adding variance to training data. A vast space of augmentations can be performed on an image; many have been successful for computer vision tasks. However, there is a bias-variance trade-off. In this work, we leverage a new augmentation for PReID training (and Face Recognition) based on atmospheric turbulence to generate the different distortion levels for the images. Atmospheric turbulence contains random temporally and spatially variable distortions, which are absent in Gaussian blur or down-sampling augmentations. Atmospheric turbulence simulation code [125] is used to implement the augmentation, which generates physically realistic distortions. It is important to note that the author of [125] proposed an algorithm to generate simulated data under different levels of atmospheric turbulence, however, they have not employed it in any training or evaluation. Conversely, we propose to use the simulated atmospheric turbulence data for training in order to achieve distortion-invariant feature representation, which has not been done before by any prior work. Our approach of simulated distortions is of practical interest because it is not tractable to collect real labeled data through atmospherics at a scale suitable for training deep learning models. Experimentally, we find training with our distortion augmentation yields the best performances compared to prior works on long-distance and low-resolution test sets. Distortion levels used herein are based on different atmospheric turbulence conditions to train our models.

5.2.2 Adaptive Weighting

Different levels of distortion compress different degrees of difficulty during training. Randomly sampling images from different distortion levels can result in sub-optimal performance since higher distortion levels (i.e., lower-quality samples) dominate the gradient during training. In other words, the mere use of atmospheric turbulence data as an augmentation might deteriorate the performance in standard (high-quality) datasets and cross-quality datasets (Table 5.3). Therefore a strategy needs to be designed to effectively employ the distorted data based on simulated atmospheric turbulence. In this context, we propose an easy-to-hard training regime in which we start by assigning higher weights for lower levels of distortion and lower weights for higher levels of distortion. *Different than prior works, we directly leverage the construction of the augmentation to assign weights.* Weighting the loss as a function of the distortion level allows the model to focus on easier examples (by giving them higher weights). By lowering the weighting of high-distortion samples, the model becomes distortion-aware without allowing them to dominate the loss in early epochs. As the training progresses, the weights for all distortion levels increase according to a cosine schedule. Step 5 of Figure 5.2 illustrates the weighting for each distortion level and is formally described below.

The distortion-aware training considers a batch of images $B = \{X^i\}_{i=1}^{N_b}$ that is composed by a mix of clean images X_{cl}^i and distorted images X_{dl}^i with distortion level randomly sampled from five possible values ($dl \in \{1, 2, 3, 4, 5\}$), where N_b is the batch size. A higher dl value indicates a stronger distortion. We keep the same number of clean and distorted images in the batch. Then features f_t^i, with $t \in \{cl, dl\}$ are extracted from the backbone

(θ_{da}). During the loss calculation, the respective weight w_t^i is assigned to each image according to the cosine weighting schedule. These steps are shown in Steps 4, 5, and 6 in Figure 5.2. For the same distortion level, the weights increase along the training following a cosine schedule (Step 5). Then, with the centers obtained for each class (Step 3), and if we are performing PReID training we also take the classes' proxies in Step 7, the distortion loss is calculated as follows:

$$\mathcal{L}_{ce}(f,q,P) = -log[\frac{e^{cos((\omega_{fq}+m_1)/\tau)+m_2}}{e^{cos((\omega_{fq}+m_1)/\tau)+m_2} + \sum\limits_{p\in P, p\neq q} e^{(cos\omega_{fp})/\tau}}] \tag{5.1}$$

$$\mathcal{L}_{distortion} = \frac{1}{W}\sum_{i=1}^{N_b}\sum_{t\in\{cl,dl\}} w_t^i \mathcal{L}_{ce}(f_t^i, p_+, P) \tag{5.2}$$

where p_+ is the positive class-center (i.e., proxy) of the feature f_t^i, P is the set of all class-centers, ω_{fq} is the angle between vectors f and q (same definition for ω_{fp}), and $W = \sum_{i=1}^{|B|}\sum_{t\in\{cl,dl\}} w_t^i$. For hyperparameters, τ is the temperature to regulate the probability distribution, m_1 is the angular margin, and m_2 is the additive margin. For PReID, $m_1, m_2 = 0$ and $\tau = 0.05$.

The class proxies

In this subsection, we present how we calculate the class proxies. To better adapt to distortions, we extend the use of multiples proxies[186] to the supervised case. This is necessary due to limited training samples and high intra-class variance, which occurs since the whole-body images are captured from different cameras resulting in views of the same person in different poses, illumination conditions, backgrounds, occlusions, and resolutions leading to high intra-class variance and low inter-class distances [180, 213] as already explained in this thesis. Step 7 of Figure 5.2 shows the multiple proxies with the circles with dark outlines.

Without loss of generality, consider a class $C = \{c_1, ..., c_{N_C}\}$ in the dataset with N_C examples. To calculate the proxies set, we start by randomly selecting a sample $c_i \in C$ ($1 \leq i \leq N_C$) to be the first proxy, and we calculate the distance between c_i and each element in C and store these distances in a cumulative vector $V_C \in R^{N_C}$. We call the first proxy as $p_C^1 = c_i$ .To calculate the second proxy, we consider the element with the furthest distance to the first proxy (the sample with maximum distance value in V_C). Formally,

$$p_C^2 := \arg\max V_C. \tag{5.3}$$

After that, we calculate the distance of p_C^2 to all samples in C to obtain the distance vector $D(p_C^2) \in R^{N_c}$. Then we update V_C considering its current values (the distances of the class samples to the first proxy) and $D(p_C^2)$ (the distance of the class samples to the second proxy) following the formulation:

$$V_C := min(V_C, D(p_C^2)), \tag{5.4}$$

where $min(.,.)$ is the element-wise minimum operation between two vectors. More specif-

ically, the j^{th} position of V_C will hold the minimum distance of the sample $c_j \in C$ considering the first and second proxies. So the j^{th} position holds the distance of c_j to the closest proxy, and the maximum value in V_C is from the sample most apart from both proxies. We consider this sample as the next proxy p_C^3. To obtain p_C^3, we apply again Eq. 5.3 but considering the updated V_C calculated from Eq. 5.4, and repeat the whole process for the new proxy. We write both equations in their general formats:

$$p_C^t := \arg\max V_C^{t-1}. \tag{5.5}$$

$$V_C^t := min(V_C^{t-1}, D(p_C^t)). \tag{5.6}$$

As explained before, we initialize $V_C^1 := D(p_C^1)$ where p_C^1 has been randomly selected from C to be the first proxy. We keep alternating between Equations 5.5 and 5.6 until $t = 5$ to get five proxies per class. During training, for a sample $X_i \in B$ (where B is the batch), we call P_i by the proxies set of its class and N_i by the set of the top-50 closest negative proxies and use them to calculate $\mathcal{L}_{proxy}$ in Eq. 5.7.

$$\mathcal{L}_{proxy} = \frac{1}{W}\sum_{i=1}^{N_b}\sum_{t\in\{cl,dl\}} w_t^i \frac{1}{|P_i|}\sum_{q\in P_i} L_{ce}(f_t^i, q, P_i \cup N_i). \tag{5.7}$$

After that, $\mathcal{L}_{proxy}$ loss is added in Eq. 5.2, obtaining the final loss function in Eq 5.8 for PReID:

$$\mathcal{L}_{distortion} = \frac{1}{W}\sum_{i=1}^{N_b}\sum_{t\in\{cl,dl\}} w_t^i \mathcal{L}_{ce}(f_t^i, p_+, P) + \lambda\mathcal{L}_{proxy}, \tag{5.8}$$

where λ controls the contribution of $\mathcal{L}_{proxy}$ to the final loss. $\mathcal{L}_{distortion}$ is applied for both distortion-adaptive and clean backbones training. To train the clean backbone, we have $w_i = 1$ for all samples because no distortion augmentations are applied. The class proxy calculation is used just for PReID training.

To improve the performance, we adopt the Mean-Teacher [173] to self-ensemble the weights of the backbones along the training. Considering both Clean and Domain-Adaptive backbones with parameters θ_{cl} and θ_{da} (which are initialized with weights pre-trained on Imagenet), respectively, we keep another backbone for each one with parameters Θ_{cl} and Θ_{da} with the same architecture to self-ensemble their weights along training through the following formula:

$$\Theta_s^{t+1} := \beta\Theta_s^t + (1-\beta)\theta_s^t \tag{5.9}$$

where $s \in \{cl, da\}$, β is a hyper-parameter to control the inertia of the weights, and t is the instant of time. We set $\beta = 0.999$ for all models following prior PReID works [47, 226]. We use the backbones Θ_{cl} and Θ_{dl} for the final evaluation.

5.2.3 Cross-Domain Fusion

After training, we have the self-ensembled weights for the distortion-adaptive and clean backbones, Θ_{da} and Θ_{cl}, respectively. Since the first is trained with clean and distorted images, it is more invariant to distortions than the second model trained just with clean images. To leverage and combine knowledge from both backbones, we apply magnitude-weighted fusion between the backbones as shown in Figure 5.2. This idea is inspired by recent studies in magnitude-based training [87] and the effects of atmospheric turbulence in Face Recognition [153], where usually features with higher magnitudes mean an input image in higher resolution.

The main rationale is that since we do not know the distortion level of images in the testing scenario, we can use the magnitude of the feature vectors generated by each backbone as a proxy for it. In other words, the stronger the distortion we expect a lower magnitude for the output feature vector from the clean model (since it has not been trained with distorted data), and the higher will be the magnitude of the output feature vector from the distortion-adaptive backbone. Conversely, the higher the resolution of the input image, we expect a higher magnitude for the feature vector output by the clean model, and a lower one for the feature vector output by the distortion-adaptive backbone. Then we use the magnitude of the feature vectors to weigh the decision from each backbone. The advantage of this approach is evident in 5.3. At inference, for a query and gallery image pair, we extract both feature vectors $q_{cl} = \Theta_{cl}(X^q)$ and $g_{cl} = \Theta_{cl}(X^g)$ from the query and gallery images pair considering the clean model with parameters Θ_{cl}, and the feature vectors $q_{da} = \Theta_{da}(X^q)$ and $g_{da} = \Theta_{da}(X^g)$ from the distortion-adaptive backbone. We calculate the distance between the query and gallery considering each backbone to obtain distances $D(q_{cl}, g_{cl})$ and $D(q_{da}, g_{da})$, which are weighted combined considering the maximum feature magnitude for each pair before L2 normalization as shown in the equation on the lower half of Figure 5.2. We use that distance to rank all gallery images given a query and calculate the metrics.

5.3 Experiments and Results

In our paper [152], our experiments are performed on face recognition and person re-identification tasks with an emphasis on low-image-quality scenarios. Common training and evaluation procedures are followed for each task, respectively. Since Face Recognition is out of the scope of this research, we focus our analysis just on Person Re-Identification.

5.3.1 Datasets

For Person Re-Identification, we used two same-clothes datasets: `Market1501` and `MSMT17`, and one clothes-changing dataset: `DeepChange` (described in Appendix B). For evaluation, following prior work, experiments are run with predefined train-test splits, and mAP and CMC metrics are reported.

5.3.2 Implementation details

For fair comparison to the prior person re-identification work, we adopt the ResNet50 [58] as the model backbone. Following previous works [124, 130], we change the stride of the last residual block to 1 to increase the feature map size. Then we insert a global average pooling and global max pooling layer after the last feature map and sum their outputs element-wise [130]. After that, we add batch normalization and perform the L2-normalization to project them to the unit hyper-sphere. To train the clean and distortion models, we employ the Adam [88] optimizer with weight decay of $5e^{-4}$ and initial learning rate of $3.5e^{-4}$. We train both models for 250 epochs and divide the learning rate by 10 every 100 epochs. As explained, the number of proxies per class is fixed in 5 (i.e., $\forall_i |P_i| = 5$) for all datasets. To create the batch to optimize the clean model, we adopt a similar approach to the PK batch strategy [64] in which we randomly choose P identities and, for each identity, K clean images (without distortion). To train the distortion model, we sample K clean images and K distorted images randomly sampled from five different levels of distortion strength. We also apply Random Crop, Random Horizontal Flipping, Random Erasing, and random changes in brightness, contrast, and saturation as data augmentation.

5.3.3 Comparison to the state of the art

DaliReID is compared with state-of-the-art methods in PReID for both the same-clothes scenario and the clothes-changing scenario. For the same-clothes scenario, results are reported in Table 5.1. Our method is orthogonal to the backbone, and we show results with two backbones used in prior works: ResNet50 and OSNet [251]. DALIReID achieves the highest performance on the `Market1501` dataset, outperforming FIDI [206] by 0.8 in mAP, and the second position (along with FIDI) with R1 = 94.5. In `MSMT17`, the most challenging PReID benchmark, we reach the best performance by outperforming CDNet by a margin of 5.9 and 3.2 in mAP and R1, respectively, with ResNet50. With OSNet, we achieve the best performance in both datasets for both metrics. Our method is able to rank ground-truth gallery images closer to the query and outperforms prior art in mAP in all setups.

To show our model generalization ability, we trained DaliReID for `DeepChange`, in which subjects' clothes differ among views, and the results are shown in Table 5.2. We outperformed the recent CAL [51] by 2.9 and 6.8 in mAP and R1, respectively. Besides the clothes changing, `DeepChange` has more distortions and low-quality data than `Market` and `MSMT17`. We obtain the highest gain on it for R1 and the second highest gain for mAP (after `MSMT17`), showing our method can better improve performance in low-quality datasets. For fair comparison, we do not employ any kind of part-based, alignment, segmentation mask, or pose variation strategies, in order to verify the performance improvement brought just by our DaliReID model. For this reason, prior methods that employ one of those strategies are not mentioned in the table.

		Market		MSMT17	
Method	Venue	mAP	R1	mAP	R1
OSNet-based models					
OSNet [251]	ICCV'19	84.9	94.8	52.9	78.7
DaliReID (OSNet)	**This work**	**87.2**	**95.0**	**59.5**	**82.6**
ResNet50-based models					
GCS [16]	CVPR'18	81.6	93.5	-	-
SFT [121]	ICCV'19	82.7	93.4	47.6	73.6
CBN [257]	ECCV'20	83.6	94.3	-	-
STNReID [123]	TMM'20	84.9	93.8	-	-
CBDB-Net [168]	TCSVT'21	85.0	94.4	-	-
BAT-Net [43]	ICCV'19	85.5	94.1	50.4	74.1
CDNet(*) [94]	CVPR'21	86.0	**95.1**	**54.7**	**78.9**
FIDI [206]	TMM'21	**86.8**	**94.5**	-	-
DaliReID (R50)	**This work**	**87.6**	**94.5**	**60.6**	**82.1**

Table 5.1: Comparison to the state-of-the-art models in same-clothes Person Re-Identification setup. **Blue** and **Green** indicate the best and second-best values. *CD-Net is not based on ResNet50, but the authors of that paper mostly compared to ResNet50-based models, so we leave it here for a fair comparison.

		DeepChange	
Method	Venue	mAP	R1
ReIDCaps [71]	TCSVT20	11.3	39.5
ViT [203]	ArXiv20	15.0	49.8
ViT (with Grayscale) [203]	ArXiv20	15.2	48.0
CAL [51]	CVPR22	**19.0**	**54.0**
DaliReID (R50)	**This work**	**21.9**	**60.8**

Table 5.2: Comparison to the state-of-the-art models in clothes-changing person re-identification setup. **Blue** and **Green** indicate the best second-best values. All methods, except ViT, consider ResNet50 (R50) as the backbone.

5.4 Ablation Study

We perform a set of ablation studies over the PReID datasets to measure the impact of different components. In Table 5.3, we ablate the different components of DaliReID. When we use distorted images as augmentations without our adaptive-weighting strategy (second line), we see a performance drop in 2.9 p.p. and 2.8 p.p. for mAP and R1 respectively in `MSMT17` compared to our proposed Distortion-adaptive strategy (third line). For `DeepChange` we also see a suitable performance drop of 0.5 p.p. and 0.6 in mAP and R1 respectively. For `Market`, R1 is slightly better, but the mAP is below 1.3 p.p. The main reason is that just using the distortion augmentation as a regular augmentation without any further treatment, the varied distortions images have the same importance for training, then the model does not effectively learn from distorted data. It is interesting to see that, for `MSMT17` and `DeepChange`, the results are also worse than the baseline (first line)

	Market		MSMT17		DeepChange	
ReID Ablation	mAP	R1	mAP	R1	mAP	R1
Clean (Θ_{cl})	86.6	94.2	57.6	80.3	20.5	59.3
Distortion Aug	86.3	**94.7**	55.4	78.5	20.2	58.6
Distortion-Adaptive (Θ_{da})	86.6	94.3	58.3	81.3	20.7	59.2
Distortion-Adaptive w/o $\mathcal{L}_{proxy}$	82.4	92.9	47.9	72.9	19.2	55.6
DaliReID	**87.6**	94.5	**60.6**	**82.1**	**21.9**	**60.8**

Table 5.3: Ablation study for PReID. The first line shows the performance of the clean model (trained without simulated distorted data). The second and third lines are for backbones trained with distortion as augmentation and our adaptive weighting strategy, respectively. The fourth line ablates the proxy loss, and the final line is the proposed DaliReID model.

supporting that just employing distortion as augmentations in fact might hinder model performance. This shows the potential of the distortion-adaptive strategy in effectively learning useful patterns from distorted data and improving the model's robustness to atmospheric turbulence and distortions, which increases the metrics in standard evaluation scenarios.

When we check the impact of the proxy loss we see a performance dropping when we take it out ($\lambda = 0$ in Eq. 5.8), as shown in the fourth line of Table 5.3. Indeed the results show that when we remove our proxy loss we obtain the worst results in all metrics and datasets, showing its importance for effective learning in PReID. Our final DaliReID model (last line) combines both clean and distortion-adaptive backbones (first and fourth lines), which leads to the best performance for `MSMT17` (an increase of 5.2 and 3.6 p.p. for mAP and R1, respectively, compared to just employing the Distortion Augmentation in the second line) and `DeepChange` (an increase of 1.7 and 2.2 p.p. for mAP and R1, respectively, compared to the second line). For `Market` we see a slight performance drop in R1, but an increase of 1.3 p.p. in mAP. It is important to recall that `MSMT17` and `DeepChange` are more challenging than `Market`, since they have more cameras in different moments of the day and of the year. In the case of `DeepChange` the identities still have different clothing. This shows that DaliReID can effectively combine knowledge from both backbones and improve the model's performance with distorted data.

Impact of the distortion augmentation

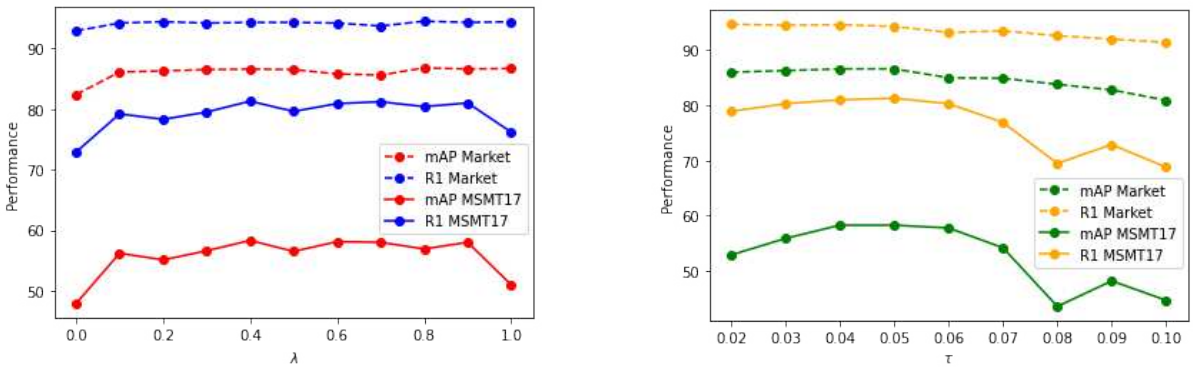
A key contribution of this solution is the use of distortion augmentation inspired by atmospheric turbulence. So one natural question would be "Why not use Gaussian Blur and Down-sampling operations to create distortions?". In this ablation study, we show a comparison of our model trained with the proposed atmospheric turbulence simulated data to those well-known image processing operations. Gaussian blur and down-sampling are applied at equally challenging levels as atmospheric turbulence distortion augmentation. In 5.4, it can be seen that distortion augmentation performs better than data augmentation Gaussian blur and down-sampling on all PReID benchmarks (first and second lines). But DaliReID still retains the best performances.

	Market		MSMT17		DeepChange	
	mAP	R1	mAP	R1	mAP	R1
DS+GB	78.0	91.2	44.7	69.5	16.2	51.5
Distortion Aug (ours)	86.3	**94.7**	55.4	78.5	20.2	58.6
DaliReID (ours)	**87.6**	94.5	**60.6**	**82.1**	**21.9**	**60.8**

Table 5.4: A comparison between training augmentations. The distortion augmentation performs better than using Gaussian blur and down-sampling. "DS" holds for Down-Sampling, and "GB" for Gaussian Blur.

Parameter Analysis

There are two hyper-parameters on the final loss function (Eq. 5.8): τ value to control the sharpening of the probability distribution in its both terms, and λ value to weight the contribution of $\mathcal{L}_{proxy}$ term. The impact of these parameters on the performance of the Distortion-Adaptive Backbone is shown in Figure 5.3.



(a) Impact of different λ values

(b) Impact of different τ values

Figure 5.3: Analysis of the impact of the parameters τ and λ on the final loss function considering the training of the Distortion-Adaptive Backbone for PReID.

For λ in Figure 5.3a, we see stable performance for `Market` along different values after $\lambda = 0.1$, while for `MSMT17` we see a peak at $\lambda = 0.4$, then a suitable decrease after this value. For both datasets, we see a performance drop for $\lambda = 0.0$ (no $\mathcal{L}_{proxy}$), showing again the proxy-based loss term has a positive impact on training. In contrast, an equal contribution of both terms $\lambda = 1.0$ hurts the performance mainly for `MSMT17`. Since `MSMT17` is more challenging, we select $\lambda = 0.4$ as the operational value. Further analysis of the impact of $\mathcal{L}_{proxy}$ is presented in Table 5.3.

The impact of τ is shown on Figure 5.3b. The performance drops when τ is lower than 0.04 for `MSMT17` but a stable behavior for `Market`, while values greater than 0.06 deteriorate the performance for both datasets. To achieve a good trade-off considering the dataset complexities, we choose $\tau = 0.05$.

Impact of pooling operations

As shown in Fig. 5.2, the inference is performed by a weighted combination of the decisions from Clean and Distortion-Adaptive backbones. The weights W_{clean} and $W_{distortion}$ are the

maximum magnitudes of the feature vectors for each query and gallery image pair for each backbone. Among the different pooling strategies to get the final feature representation, we choose Global Average Pooling (GAP), Global Max Pooling (GMP), and a combination of both (GAP+GMP) to check the impact on final performance. The performances are reported in Table 5.5. Note that in this case, the pooling operations are **just to calculate the magnitudes**, since the final representation is always obtained by the element-wise sum of the output of the GAP and GMP layers for PReID.

	Market		MSMT17		DeepChange	
Setup	mAP	R1	mAP	R1	mAP	R1
GMP	87.6	94.4	60.5	82.1	21.9	60.7
GMP+GAP	87.6	94.4	60.6	82.1	21.8	60.8
DaliReID (GAP)	**87.6**	**94.5**	**60.6**	**82.1**	**21.9**	**60.8**

Table 5.5: Ablation of the pooling operation to calculate the magnitudes for fusion in PReID.

We see among GAP, GMP, and GAP+GMP, we have a similar performance in evaluation, with a slighter improvement for GAP. All of them have similar performances over the final result showing our proposed fusion strategy is robust to different pooling operations.

5.5 BRIAR Results

The BRIAR dataset has been proposed to tackle the long-range recognition problem. It has images of the same identity recorded in controlled and field conditions. In the controlled conditions, the person walks in a well-illuminated indoor environment close to the cameras, resulting in high-resolution data, and there are still images of the person in different poses and angles with arms up and down. The field condition is cameras capturing the identities from 100m to 1,000m, including a UAV which brings changes in the pitch angle, yielding to an increased impact of the atmospheric turbulence effect. Moreover, each identity has two different outfits (called "set1" and "set2"), so the clothing-change challenge is also present. In the evaluation setup, the gallery is composed of just controlled videos and images with the identities wearing their respective "set2" clothes, and the probe set is in-field videos of the identities wearing the "set1" clothes. Therefore, the BRIAR data comprises both long-range and clothing change challenges. An example is shown in Figure 5.4.

The BRIAR training set has around 14 million whole-body bounding boxes from 577 identities in controlled and field conditions across different cameras and long-range distances. In the training set both "set1" and "set2" of clothing are considered to encourage the model to learn clothing-invariant features.

The BRIAR evaluation protocol considers two gallery sets, Gallery 1 (G1) and Gallery 2 (G2), and two probe sets, Face Included (FI) and Face Restricted (FR).

G1 has 485 subjects where 351 are distractors with a total of 43,728 images and 4,197 videos. The recordings sum up to 23.7 hours. G2 has 481 subjects where 351 are distractors with a total of 43,242 images and 4,171 videos. The recordings sum up to 23.5

Figure 5.4: Example of different recording conditions for an identity in the BRIAR dataset. All images are from the same identity. The first and second rows represent clothing from sets 1 and 2 respectively. Permission granted by subjects for use of imagery in public presentations (G00430).

hours. Both galleries were recorded under indoor controlled conditions. All media in both galleries belong to the "set2" of clothing of the identities.

FI has 5,435 videos of 260 subjects in the field conditions with face-visible views. It sums up to 11.6 hours of video. FR has 2,078 videos of the same 260 subjects in the field conditions, however, with low-resolution or unusable faces. It sums up to 4.3 hours of video. All media in both probe sets belong to the "set1" of clothing of the identities. The rationale is to check model robustness to atmospheric turbulence and clothing.

To address this task we first trained our backbones with the solution presented in this chapter in a supervised manner. Then, in evaluation, we employ our proposed unsupervised method. More specifically, for each identity in the gallery set, we extract features from all of their still images and video frames and perform clustering with the hyper-parameter-free clustering approach PEACH [93]. PEACH is based on the Extreme Value Theory (EVT) to design a hyper-parameter clustering value definition for the agglomerative clustering [91]. To illustrate the results, four sampled clusters are depicted in Figure 5.5.

For each cluster, we take the average feature vector as a representative (proxy) for it. So, suppose that N_c^i clusters are found for the i^{th} gallery identity, we will have a set F_c^i with N_c^i feature vectors. This is done for all gallery identities, then we will have the set $F_c = \{F_c^i\}_{i=1}^{N_g}$ with N_g set of features where N_g is the number of identities in the gallery.

The same process is done for the query input video. If N_c^p clusters are found in the probe video p we will have the set F_c^p with N_c^p feature vectors. All the feature vectors are normalized to have norm one. After that, for each $F_c^i \in F_c$ we calculate the similarities between all features in F_c^p and F_c^i, which will result in a similarity matrix $S_{pi} \in \mathbb{R}^{N_c^p \times N_c^i}$.

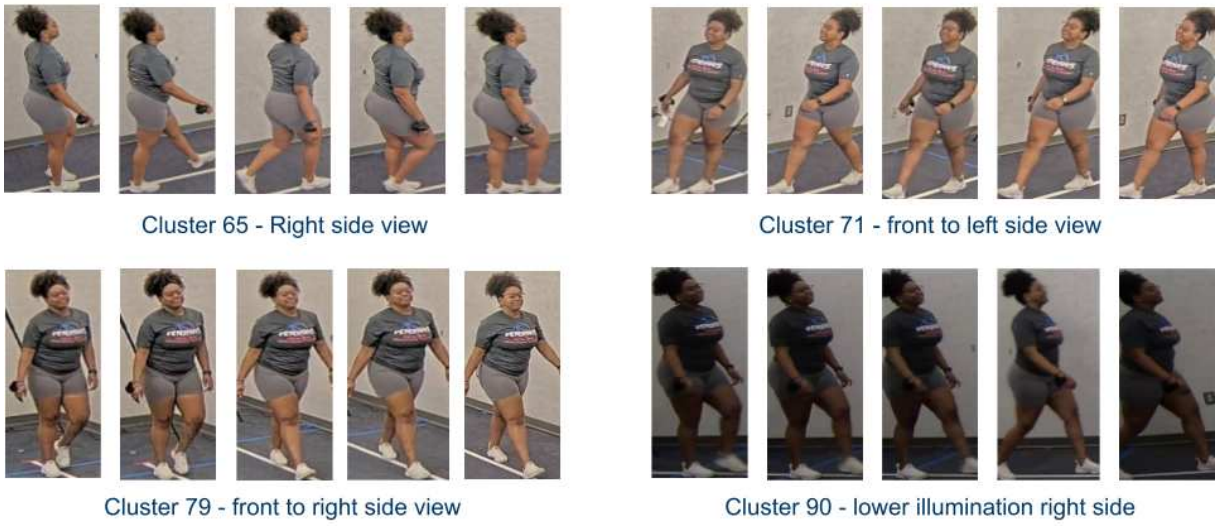


Figure 5.5: Subset of the clusters obtained after clustering all features from still images and video frames from identity G02078 in the BRIAR dataset. We can see that each cluster effectively captures different aspects of the capturing conditions of the identity. Permission granted by subjects for use of imagery in public presentations (G02078).

To avoid the impact of outliers in out-of-distribution similarities, we take the top 10 highest similarities and average them to get the final score between the input probe video and the i^{th} identity in the gallery. We evaluated with different numbers (top-5, top-10, top-15, etc.) and also with the average of all similarities, and we saw that top-10 resulted in the best performance. After that, we sorted all the results and calculated the average Rank-20 (R20) across all probes.

As one of the metrics of the BRIAR program, we also calculate the $TPR@FPR = 1\%$ considering each probe-gallery-ID pair a binary output (same or different class) and average them over all probes. This solution has been initially proposed for Phase 1 and its results are shown in the first line of Table 5.6. We see that the model has a better performance when evaluated in Face Included (where faces usually are visible) than in Face Restricted (where the face is in low resolution or not visible).

Table 5.6: Results on BRIAR dataset. "R50", "OSN" and "DEN" are for ResNet50, OSNet and DenseNet121. "FI", "FR", "G1" and "G2" are for Face Included, Face Restricted, Gallery 1 and Gallery 2. "1:1" means TPR@FPR = 1%. "CL" and "DA" hold for Clean and Distortion-Adaptive backbones.

		FI to G1		FR to G1		FI to G2		FR to G2	
	Backbones	R20	1:1	R20	1:1	R20	1:1	R20	1:1
Phase 1	R50/R50 (CL/DA)	56.29	35.29	53.98	34.30	50.15	26.58	46.57	24.18
	TransReID/R50 (CL/DA)	60.33	40.00	58.71	39.89	57.91	32.85	54.78	31.61
Phase 2	R50/DEN (DA/DA)	80.98	46.32	**80.89**	44.77	82.90	46.53	81.03	38.34
	OSN/DEN (DA/DA)	**82.03**	**48.06**	80.80	**45.91**	**86.91**	**48.07**	**84.19**	**40.02**

We extend our solution to consider a second setup where the clean and the distortion-adaptive backbones have different architectures, and the results are shown in the second line of Table 5.6. In this case, the clean model is a Transformer model [61]. We see improvement in more than 4.0 p.p. in all metrics, with the highest gains in the Face Restricted scenario, which follows previous conclusions in this thesis that diversity brings

gain in the performance with ensemble.

Finally, Phase 2 started in June of 2023 and the DaliReID has been extended and modified to increase its performance. The main modifications are the inclusion of the center loss [124] and other triplet-based loss functions to consider the clothing changing and the atmospheric turbulence. Moreover, we have also evaluated models where both backbones are distortion-adaptive. Phase 2 is still ongoing, so preliminary results are added in the third and fourth lines of Table 5.6, showing promising performance with high gains in all metrics, mainly for R20, across different evaluation scenarios.

5.6 Final Remarks

Image distortions are frequently encountered in real-world unconstrained forensics and biometrics scenarios. Among them, atmospheric turbulence is the most often in surveillance systems and border security [74], which can deploy long-range cameras to capture possible people around facilities from long distances.

In this chapter, the DaliID is presented as a methodology for improving robustness to such distortions mainly targeting atmospheric turbulence. The proposed components include distortion augmentation, distortion-adaptive weighting, and a parallel-backbone magnitude-weighted feature fusion. While face recognition and person re-identification have considerable differences, DaliID is applicable in both tasks with state-of-the-art performance on benchmarks in both tasks. The proposed LD datasets (Appendix D), captured over the longest distance of any academic dataset, allow for further evaluation of realistic distortions.

Another common challenge the researchers find in Person Re-Identification is the clothing changing of the identities when the analysis spans over hours, days, or months. To show that the designed solution also has potential for this scenario, we evaluate our solution in the `DeepChange` dataset which compresses people changing their clothes since it was created by matching the same person seen in different months of the year. Therefore we show our designed solution can tackle one of the most challenging scenarios in PReID where both atmospheric turbulence distortion and clothing changing of the identities happen.

We also point out that the proposed weighting strategy could be redesigned and extend to consider a selective criteria, where, instead of keeping all samples with varying weights, some of them could be discarded, for instance, the ones with low resolution and unusable biometric features. This could make the training more stable. Furthermore, the model could be extended to also consider the loss function values for each sample as a way to measure hardness. With the loss values, a selective or weighting criteria, besides the adaptive weighting, could be employed to boost performance in a data-driven manner.

Differently from previous chapters, we have not considered a fully unsupervised scenario. However, in the context of the BRIAR program, which has its own dataset for benchmarking, the solution has been extended and a fully unsupervised methodology was proposed to create clusters in the gallery set and in the query input video to improve the matching performance. The solution helped our team to reach the top positions compared

to other BRIAR teams considering the program metrics.

This research is also based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. We also thank UCCS VAST Lab for the hardware infrastructure support.

Chapter 6

Conclusions

In this research, we present three fully unsupervised methods and one supervised method, which target a common task: Person Re-Identification (PReID). We have designed the first three solutions to deal with the fully unsupervised re-identification task, where the primary goal is to learn how to find the same identity in a camera system with varying recording conditions. Some of these conditions can be changes in illumination, resolution, identity pose, background, and occlusions. We also explore an even harder setup where we assume fully unlabeled data, i.e., we do not have the identity of the people in the images. In this case, the solutions should be able to overcome the challenges present in different recording conditions without any identity labeling.

The first solution (Chapter 2) addresses the Unsupervised Person Re-Identification task under two assumptions: we assume we know the camera label, that is, we know which camera recorded each image in the dataset, and we assume all of our backbones have been initialized through supervised training in a source dataset. For this reason, it is an Unsupervised Domain Adaptation (UDA) model. The solution achieves state-of-the-art performance and the qualitative analysis shows that our model addresses the main challenges to match the same identity under different recording conditions. This answers our first research question: "Which constraints or meta-information can improve the self-supervised learning performance on this kind of task?", wherein we have found that the proper creation of triplets based on camera labels, ensembling on evaluation, and supervised source dataset pre-training can be helpful and a promising path to increase performance. Moreover, the proposed self-ensembling solution was designed considering an unsupervised scenario without validation set. In this context, other strategies from different research fields could be explored to evaluate the training quality for selection and combination of checkpoints. For instance, in Reinforcement Learning there are strategies to evaluate safe behaviors of the agent during the learning process [126, 147]. This can be an interesting research direction to explore in future works.

The second solution (Chapter 3) extends the applicability and is more general than the first one. Different from the first one, we assume a more realistic setup: no side information (such as the camera label) and no pre-training in the labeled source dataset. This method is fully unsupervised because it relies solely on detected bounding boxes without further annotation or side information. Our solution presents two novel ensembling-based approaches to amalgamate knowledge from different backbones and obtain a clustering

result derived from clustering runs with different hyperparameters. There is no need to tune the cluster hyperparameter for each dataset. The obtained results outperform prior art and are on pair to UDA models that require a source dataset for model initialization. Since our solution does not rely upon task-related side information or annotation, it can be deployed in tasks beyond U-PReID. We validate this hypothesis by employing the model in an NLP task: Text Authorship Attribution (TAA) for social media short messages. We assume all social media texts (posts from X - former Twitter) do not have the authorship information, i.e., we do not know "who" wrote that tweet. Our solution provides better results than a supervised method tailored to this task, and the ablation studies show that, indeed, the model can provide increased performance in fully unlabeled social media posts. This answers our second research question: "How to design a general self-supervised learning algorithm to deal with problems with fully unlabeled data comprising high intra-class dissimilarity and inter-class similarity with a set of classes not seen on training during the evaluation phase?". Our findings show that the ensembling of knowledge from different architectures, and the combination of different clustering results allow the model to tackle challenging scenarios with high intra-class variation and inter-class similarity.

Our second solution was also extended to be part of a model that combines both supervised and self-supervised training. This model was designed to address the aerial-to-ground PReID task in the AGReID2023 competition in IEEE IJCB 2023, obtaining the **third place** in the competition, being one of its highlights.

The third solution (Chapter 4) operates under the same conditions as the second one: it relies only on people bounding boxes without any further side information. However, it employs new strategies we designed to address large-scale scenarios. With a novel Local Neighborhood-based sampling, a Local Re-Ranking, new scheduling schemes for clustering hyperparameters, and simple co-training, we reach the state-of-the-art performance in standard benchmarks in PReID and large-scale datasets in Vehicle Re-Identification (VReID). The qualitative results also show our model can address the cross-recording condition in the camera environment to match the same person/vehicle from different points of view. The solution opens the path for studies in further re-identification tasks (e.g., places) that can be employed to help in investigations and surveillance systems. This helps to answer the third research question: "How to scale those solutions to handle thousands of data samples?". Our proposed Local Neighborhood Sampling and Local Re-Ranking directly address this question by reducing the number of samples to train and by proposing a lower time complexity upper bound, which is more suitable for large-scale scenarios.

The third solution was also presented during the research and training consultancy to the Police of Dubai in December 2023. It was considered as part of a set of methods to help with multimedia investigations in suspect searching and event understanding.

The last solution (Chapter 5), differently from the previous ones, targets a supervised scenario within the context of an evolving interest setup: long-range recognition. It allows the learned model to overcome image distortions mainly caused by atmospheric turbulence in an easy-to-hard manner. It also allows the backbones to effectively learn discriminative features from distorted images, achieving state-of-the-art performance in

cross-resolution scenarios. Besides that, we consider another challenge in PReID: the clothing-change identities. In one of the evaluated benchmarks and the BRIAR dataset, the identities change their clothes from one camera to the other, hardening the task as the models cannot rely upon clothing features anymore. Even so, our method achieves the best performance in this scenario compared to prior art and methods proposed for the BRIAR program. Finally, this answers the fourth research question: "Which strategies can potentially help to perform long-range Person Re-Identification?". We show that the proposed techniques — easy-to-hard scheme to progressively consider distorted images and magnitude-based ensembling — effectively allow the model to mine useful patterns and increase the performance in long-range and cross-resolution benchmarks.

In summary, our methods address important challenges in re-identification considering fully unsupervised scenarios under varied constraints. Besides, we also present an approach to learn from distorted data caused by atmospheric turbulence with cloth-changing identities. The list of all published and under-review papers produced along this Ph.D. is shown in Appendix A. We argue that our solutions can be employed and extended in further re-identification scenarios and further forensics and biometrics tasks, with the proper modifications. One of those extensions can be the joint analysis of the main pieces of an event during an investigation. More specifically, the solutions can be incorporated into a broader pipeline to mine and group people, vehicles (or other objects), and places in an event. This would allow the proposing of possible relationships between people and objects in a scene, helping investigations and event understanding.

For instance, suppose the illustrative case depicted in Figure 6.1. After employing our pipeline, the red and blue lines, indicating images from different classes and the same class, respectively, could be discovered. Then a later step is taken to find possible relations between the elements (people and vehicles). Suppose that the model found the green line, that is, a positive relation between that person and the car in the event. Based on this, could we infer possible relations between the other people and cars (represented by the dashed yellow lines with question marks)? The authors in [253] propose a similar idea, however considering just a group of people spatially close to each other, and not accounting for other target objects (e.g., vehicles). This shows that joint analysis has not been deeply studied yet being a promising future research step, and can help to perform event understanding.

Moreover, recent advances in Large Language Models and Large Vision Models based on Transformers [177, 55] have shown a promising research direction by employing advanced models for feature extraction based on attention. Given that they are usually trained in large-scale datasets, they provide useful image feature descriptions that can be employed for downstream tasks. These features can be explored to extend the solutions proposed in this Ph.D. research, enhancing feature representation and extension to further re-identification tasks. Additionally, attention mechanisms in large models could also be employed to explain the cluster structures and relations between their elements, as proposed in [132].

However, employing large models can also bring more challenges related to the required time and memory to work with them. In this way, our first solution (Chapter 2), for instance, would face limitations since it requires that all checkpoints are saved during the

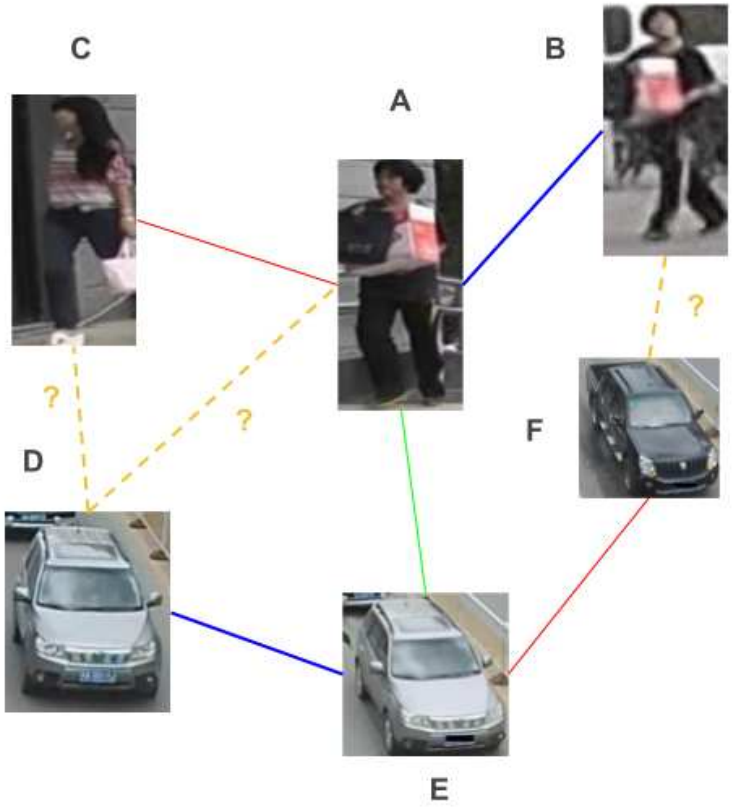


Figure 6.1: Illustrative example of joint analysis. Blue lines mean that the connected images are from the same class, red lines mean they are from different classes, the green line means that there is a relation between the person and the vehicle (that could be discovered by a model or inserted by the expert analyzing the event). The dashed yellow lines with question marks show possible questions to the model which might help find other relations between identities and vehicles.

training for late self-ensembling. In this case, we suggest keeping most of the large model weights frozen, fine-tuning just part of its weights, and saving only these learnable weights instead of the whole model. This could allow us to leverage the power of large models and limit the number of saved parameters, saving time and memory for self-ensembling (Chapter 2), and momentum ensembling (Chapters 3 and 4).

Additionally, for the first, second and third proposed solutions (Chapters 2, 3, 4, respectively), we show the XAI activation maps for some images to highlight which parts the models mostly focus on to rank the images. The activation maps have been utilized throughout this research solely for visual and qualitative analysis, without employing any feedback to enhance the models. Moreover, we suggest, as possible future work, the design of a feedback strategy considering which parts of the identities have been (or not) activated by the models, aiming to improve feature learning [136]. This process could involve training a model focusing on other parts of the images to learn from complementary features, or human intervention by manually removing noisy images and images with non-biometric activations (e.g., in the background).

Another limitation that has not been deeply studied is the camera failures. Most outdoor cameras, which are present in all datasets employed in this research, are exposed to weather and environmental conditions. This exposes them to various failure scenarios, such as broken lens, fog, malfunctions, interferences, etc. A deeper study about how to deal with these challenges could be valuable for effective learning and deployment in unconstrained real-world scenarios.

We expect that the proposed methods can aid future advancements in the forensic and biometrics research domains. Moreover, these innovations have the potential to assist authorities in conducting thorough investigations, gaining deeper insights into various events, and fostering enhancements in surveillance system technologies.

Bibliography

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60, 1999.

[2] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *Conf. Comput. Vis. Pattern Recog.*, pages 2530–2539, 2017.

[3] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. Cnn detection of gan-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020.

[4] Mauro Barni, Ehsan Nowroozi, and Benedetta Tondi. Detection of adaptive histogram equalization robust against jpeg compression. In *2018 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–8. IEEE, 2018.

[5] Mauro Barni, Ehsan Nowroozi, and Benedetta Tondi. Improving the security of image manipulation detection through one-and-a-half-class multiple classification. *Multimedia Tools and Applications*, 79(3):2383–2408, 2020.

[6] Raja Muhammad Saad Bashir, Muhammad Shahzad, and MM Fraz. VR-PROUD: Vehicle re-identification using progressive unsupervised deep architecture. *Pattern Recog.*, 90:52–65, 2019.

[7] Gabriel C. Bertocco, Fernanda Andaló, and Anderson Rocha. Unsupervised and self-adaptative techniques for cross-domain person re-identification. *IEEE Trans. Inf. Forensics Security*, 16:4419–4434, 2021.

[8] Gabriel Capiteli Bertocco, Antonio Theophilo, Fernanda Andaló, and Anderson De Rezende Rocha. Leveraging ensembles and self-supervised learning for fully-unsupervised person re-identification and text authorship attribution. *IEEE Transactions on Information Forensics and Security*, 18:3876–3890, 2023.

[9] Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. Explainable authorship verification in social media via attention-based similarity learning. In *Int. Conf. Big Data*, pages 36–45, 2019.

[10] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conf. Knowl. Discovery Data Mining*, pages 160–172, 2013.

[11] João Phillipe Cardenuto, Jing Yang, Rafael Padilha, Renjie Wan, Daniel Moreira, Haoliang Li, Shiqi Wang, Fernanda Andaló, Sébastien Marcel, and Anderson Rocha. The age of synthetic realities: Challenges and opportunities. *arXiv preprint arXiv:2306.11503*, 2023.

[12] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Eur. Conf. Comput. Vis.*, pages 132–149, 2018.

[13] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint*, arXiv:2006.09882, 2020.

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint*, arXiv:2104.14294, 2021.

[15] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 371–381, 2019.

[16] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8649–8658, 2018.

[17] Hao Chen, Benoit Lagadec, and François Bremond. ICE: Inter-instance contrastive encoding for unsupervised person re-identification. In *Int. Conf. Comput. Vis.*, pages 14960–14969, 2021.

[18] Hao Chen, Benoit Lagadec, and Francois Bremond. Enhancing diversity in teacher-student networks via asymmetric branches for unsupervised person re-identification. In *Winter Conf. Appl. Comput. Vis.*, pages 1–10, 2020.

[19] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 2004–2013, 2021.

[20] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Learning invariance from generated variance for unsupervised person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6), 2022.

[21] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8351–8361, 2019.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*, pages 1597–1607, 2020.

[23] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *Int. Conf. Comput. Vis.*, pages 232–242, 2019.

[24] De Cheng, Jingyu Zhou, Nannan Wang, and Xinbo Gao. Hybrid dynamic contrast and probability distillation for unsupervised person re-id. *IEEE Trans. Image Process.*, 31:3334–3346, 2022.

[25] Zhiyi Cheng, Qi Dong, Shaogang Gong, and Xiatian Zhu. Inter-task association critic for cross-resolution person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 2605–2615, 2020.

[26] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 7308–7318, 2022.

[27] Anna Gabriela Costa. How Artificial Intelligence helped the Federal Police of Brazil to identify coupers in January 8. `http://surl.li/swsqr`, 2023. [Online; accessed 10-January-2024].

[28] Yongxing Dai, Jun Liu, Yan Bai, Zekun Tong, and Ling-Yu Duan. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE Transactions on Image Processing*, 30:7815–7829, 2021.

[29] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. *arXiv preprint arXiv:2103.11568*, 2021.

[30] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Asian Conf. Comput. Vis.*, pages 1142–1160, 2022.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

[32] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[33] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 994–1003, 2018.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805, 2018.

[35] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

[36] Peter H Diamandis and Steven Kotler. *The future is faster than you think: How converging technologies are transforming business, industries, and our lives*. Simon & Schuster, 2020.

[37] Guodong Ding, Salman H Khan, Zhenmin Tang, Jian Zhang, and Fatih Porikli. Dispersion based clustering for unsupervised person re-identification. In *British Mach. Vis. Conf.*, page 264, 2019.

[38] Husheng Dong, Yuanfeng Yang, Xun Sun, Liang Zhang, and Ligang Fang. Cascaded attention-guided multi-granularity feature learning for person re-identification. *Machine Vision and Applications*, 34(1):1–16, 2023.

[39] Dawei Du, Cole Hill, Gabriel Bertocco, Mauricio Pamplona Segundo, Wes Robbins, Brandon RichardWebster, Roderic Collins, Sudeep Sarkar, Terrance Boult, and Scott McCloskey. Doers: Distant observation enhancement and recognition system. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2023.

[40] Fabian Dubourvieux, Romaric Audigier, Angelique Loesch, Samia Ainouz, and Stephane Canu. Unsupervised domain adaptation for person re-identification through source-guided pseudo-labeling. *arXiv preprint*, arXiv:2009.09445, 2020.

[41] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Int. Conf. Knowl. Discovery Data Mining*, pages 226–231, 1996.

[42] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimedia Comput., Commun., Appl.*, 14(4):1–18, 2018.

[43] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8030–8039, 2019.

[44] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Netw. Learn. Syst.*, 25(5):845–869, 2013.

[45] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Int. Conf. Comput. Vis.*, pages 6112–6121, 2019.

[46] Jorge Garcia, Niki Martinel, Alfredo Gardel, Ignacio Bravo, Gian Luca Foresti, and Christian Micheloni. Discriminant context information analysis for post-ranking person re-identification. *IEEE Trans. Image Process.*, 26(4):1650–1665, 2017.

[47] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint*, arXiv:2001.01526, 2020.

[48] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *arXiv preprint*, arXiv:2006.02713, 2020.

[49] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint*, arXiv:1412.6572, 2014.

[50] Klemen Grm, Walter J Scheirer, and Vitomir Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29:2150–2165, 2019.

[51] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[52] Manuel Günther, Peiyun Hu, Christian Herrmann, Chi-Ho Chan, Min Jiang, Shufan Yang, Akshay Raj Dhamija, Deva Ramanan, Jürgen Beyerer, Josef Kittler, et al. Unconstrained face detection and open-set face recognition challenge. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 697–706. IEEE, 2017.

[53] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *arXiv preprint*, arXiv:2011.04406, 2020.

[54] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-Teaching: Robust training of deep neural networks with extremely noisy labels. In *Adv. Neural Inf. Process. Syst.*, page 8536–8546, 2018.

[55] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2023.

[56] Xumeng Han, Xuehui Yu, Guorong Li, Jian Zhao, Gang Pan, Qixiang Ye, Jianbin Jiao, and Zhenjun Han. Rethinking sampling strategies for unsupervised person re-identification. *IEEE Trans. Image Process.*, 32:29–42, 2022.

[57] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[59] Linbing He, Haishun Du, Yiming Fu, and Yanfang Ye. Multiple camera styles learning for unsupervised person re-identification. *Optik*, 277:170718, 2023.

[60] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9664–9667, 2023.

[61] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[62] Zhijun He, Hongbo Zhao, Jianrong Wang, and Wenquan Feng. Multi-level progressive learning for unsupervised vehicle re-identification. *IEEE Trans. Veh. Technol.*, 72(4):4357–4371, 2022.

[63] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

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

[65] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint*, arXiv:1503.02531, 2015.

[66] Jian Hou, Huijun Gao, and Xuelong Li. DSets-DBSCAN: A parameter-free clustering algorithm. *IEEE Trans. Image Process.*, 25(7):3182–3193, 2016.

[67] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019.

[68] Zheng Hu, Chuang Zhu, and Gang He. Hard-sample guided hybrid contrast learning for unsupervised person re-identification. In *IEEE Int. Conf. Netw. Intell. Digit. Content*, pages 91–95, 2021.

[69] Dong Huang, Chang-Dong Wang, Jian-Sheng Wu, Jian-Huang Lai, and Chee-Keong Kwoh. Ultra-scalable spectral clustering and ensemble clustering. *IEEE Trans. Knowl. Data Eng.*, 32(6):1212–1226, 2019.

[70] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Conf. Comput. Vis. Pattern Recog.*, pages 4700–4708, 2017.

[71] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3459–3471, 2019.

[72] Yewen Huang, Sicheng Lian, and Haifeng Hu. Avpl: Augmented visual perception learning for person re-identification and beyond. *Pattern Recognition*, 129:108736, 2022.

[73] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[74] IARPA. Biometric recognition and identification at altitude and range (briar) program. *IARPA Broad Agency Announcement: IARPA-BAA-20-04*, 2020.

[75] Raymond Austin Jarvis and Edward A Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.*, 100(11):1025–1034, 1973.

[76] Haoxuanye Ji, Le Wang, Sanping Zhou, Wei Tang, Nanning Zheng, and Gang Hua. Meta pairwise relationship distillation for unsupervised person re-identification. In *Int. Conf. Comput. Vis.*, pages 3661–3670, 2021.

[77] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, 2022.

[78] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. *Adv. Neural Inf. Process. Syst.*, 27:2078–2086, 2014.

[79] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[80] Xin Jin, Tianyu He, Xu Shen, Tongliang Liu, Xinchao Wang, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Meta clustering learning for large-scale unsupervised person re-identification. In *ACM Int. Conf. Multimedia*, pages 2163–2172, 2022.

[81] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 3143–3152, 2020.

[82] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180, 2020.

[83] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[84] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 7(3):535–547, 2019.

[85] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1062–1071, 2018.

[86] Nathan D Kalka, Brianna Maze, James A Duncan, Kevin O'Connor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K Jain. Ijb–s: Iarpa janus surveillance video benchmark. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–9. IEEE, 2018.

[87] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, June 2022.

[88] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980, 2014.

[89] David D. Kirkpatrick. Who is behind qanon? linguistic detectives find fingerprints. `https://www.nytimes.com/2022/02/19/technology/qanon-messages-authors.html`, 2022. [Online; accessed on January 22nd, 2023].

[90] Long Lan, Xiao Teng, Jing Zhang, Xiang Zhang, and Dacheng Tao. Learning to purification for unsupervised person re-identification. *IEEE Trans. Image Process.*, 32, 2023.

[91] Godfrey N Lance and William Thomas Williams. A general theory of classificatory sorting strategies: 1. hierarchical systems. *The computer journal*, 9(4):373–380, 1967.

[92] Chun Pong Lau, Hossein Souri, and Rama Chellappa. Atfacegan: Single face image restoration and recognition from atmospheric turbulence. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 32–39. IEEE, 2020.

[93] Chunchun Li, Manuel Günther, Akshay Raj Dhamija, Steve Cruz, Mohsen Jafarzadeh, Touqeer Ahmad, and Terrance E Boult. Agglomerative clustering with threshold optimization via extreme value theory. *Algorithms*, 15(5):170, 2022.

[94] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6729–6738, 2021.

[95] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Eur. Conf. Comput. Vis.*, pages 483–499, 2020.

[96] Mingkun Li, Chun-Guang Li, and Jun Guo. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE Trans. Image Process.*, 31:3606–3617, 2022.

[97] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Eur. Conf. Comput. Vis.*, pages 737–753, 2018.

[98] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(7):1770–1782, 2019.

[99] Xiaobao Li, Qingyong Li, Fengjiao Liang, and Wen Wang. Multi-granularity pseudo-label collaboration for unsupervised person re-identification. *Comput. Vis. Image Understanding*, 227:103616, 2023.

[100] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Int. Conf. Comput. Vis.*, pages 7919–7929, 2019.

[101] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

[102] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Conf. Comput. Vis. Pattern Recog.*, pages 2197–2206, 2015.

[103] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint*, arXiv:1807.01440, 2018.

[104] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Conf. Artif. Intell.*, volume 33, pages 8738–8745, 2019.

[105] Yutian Lin, Yu Wu, Chenggang Yan, Mingliang Xu, and Yi Yang. Unsupervised person re-identification via cross-camera similarity exploration. *IEEE Trans. Image Process.*, 29:5481–5490, 2020.

[106] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *Conf. Comput. Vis. Pattern Recog.*, pages 3390–3399, 2020.

[107] Feng Liu, Minchul Kim, Anil Jain, and Xiaoming Liu. Controllable and guided face synthesis for unconstrained face recognition. In *European Conference on Computer Vision*, pages 701–719. Springer, 2022.

[108] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Conf. Comput. Vis. Pattern Recog.*, pages 2167–2175, 2016.

[109] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 7202–7211, 2019.

[110] Tianyang Liu, Yutian Lin, and Bo Du. Unsupervised person re-identification with stochastic training strategy. *IEEE Trans. Image Process.*, 31:4240–4250, 2022.

[111] Weidong Liu, Shibo Nie, Junhui Yin, Rui Wang, Donghui Gao, and Ling Jin. Sskd: Self-supervised knowledge distillation for cross domain adaptive person re-identification. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 81–85. IEEE, 2021.

[112] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *IEEE Int. Conf. Multimedia Expo*, pages 1–6, 2016.

[113] Yuxuan Liu, Hongwei Ge, Liang Sun, and Yaqing Hou. Camera-aware progressive learning for unsupervised person re-identification. *Neural Comput. Appl.*, 35(15):11359–11371, 2023.

[114] Yuxuan Liu, Hongwei Ge, Zhen Wang, Yaqing Hou, and Mingde Zhao. Discriminative identity-feature exploring and differential aware learning for unsupervised person re-identification. *IEEE Trans. Multimedia*, pages 1–14, 2023.

[115] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[116] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[117] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982.

[118] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint*, arXiv:1608.03983, 2016.

[119] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Conf. Comput. Vis. Pattern Recog.*, pages 3235–3243, 2019.

[120] Zefeng Lu, Ronghao Lin, Qiaolin He, and Haifeng Hu. Mask-aware pseudo label denoising for unsupervised vehicle re-identification. *IEEE Trans. Intell. Transp. Syst.*, 24(4), 2023.

[121] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *Int. Conf. Comput. Vis.*, pages 4976–4985, 2019.

[122] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[123] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia*, 22(11):2905–2913, 2020.

[124] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimedia*, 22(10):2597–2609, 2019.

[125] Zhiyuan Mao, Nicholas Chimitt, and Stanley H. Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14759–14768, October 2021.

[126] Enrico Marchesini, Davide Corsi, and Alessandro Farinelli. Exploring safer behaviors for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7701–7709, 2022.

[127] Tara McKelvey. Capitol riots: The hunt to identify and arrest the rioters. `https://www.bbc.com/news/world-us-canada-55578092`, 2021. [Online; accessed 11-June-2021].

[128] Danilo Melo, Sávyo Toledo, Fernando Mourão, Rafael Sachetto, Guilherme Andrade, Renato Ferreira, Srinivasan Parthasarathy, and Leonardo Rocha. Hierarchical density-based clustering based on gpu accelerated data indexing strategy. *Procedia Comput. Sci.*, 80:951–961, 2016.

[129] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[130] Asad Munir, Chengjin Lyu, Bart Goossens, Wilfried Philips, and Christian Micheloni. Resolution based feature distillation for cross resolution person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 281–289, 2021.

[131] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for english tweets. *arXiv preprint*, arXiv:2005.10200, 2020.

[132] Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D Bui, and Khoa Luu. Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10847–10856, 2021.

[133] Kshitij Nikhal and Benjamin S Riggan. Unsupervised attention based instance discriminative learning for person re-identification. In *Winter Conf. Appl. Comput. Vis.*, pages 2422–2431, 2021.

[134] Kshitij Nikhal and Benjamin S Riggan. Multi-context grouped attention for unsupervised person re-identification. *IEEE Trans. Biom., Behav., Id. Sci.*, 5, 2023.

[135] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[136] Rafael Padilha, Fernanda A Andaló, and Anderson Rocha. Improving the chronological sorting of images through occlusion: A study on the notre-dame cathedral fire. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2972–2976. IEEE, 2020.

[137] Rafael Padilha, Caroline Mazini Rodrigues, Fernanda Andalo, Gabriel Bertocco, Zanoni Dias, and Anderson Rocha. Forensic event analysis: From seemingly unrelated data to understanding. *IEEE Security Privacy*, 18(6):23–32, 2020.

[138] Malay K. Pakhira. A linear time-complexity k-means algorithm using cluster shifting. In *2014 International Conference on Computational Intelligence and Communication Networks*, pages 1047–1051, 2014.

[139] Lin Pan, Gege Qi, Biao Guo, and Yuesheng Zhu. Unsupervised person re-identification using multi-branch feature compensation network and link-based cluster dissimilarity metric. In *IEEE Int. Conf. on Acoust., Speech Signal Process.*, pages 4302–4306, 2020.

[140] Zhiqi Pang, Chunyu Wang, Junjie Wang, and Lingling Zhao. Reliability modeling and contrastive learning for unsupervised person re-identification. *Knowl.-Based Syst.*, 263:110263, 2023.

[141] Zhiqi Pang, Lingling Zhao, Qiuyang Liu, and Chunyu Wang. Camera invariant feature learning for unsupervised person re-identification. *IEEE Trans. Multimedia*, pages 1–12, 2022.

[142] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, et al. PyTorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inf. Process. Syst.*, pages 8024–8035, 2019.

[143] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Conf. Comput. Vis. Pattern Recog.*, pages 1944–1952, 2017.

[144] Jinjia Peng, Guangqi Jiang, and Huibing Wang. Adaptive memorization with group labels for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.*, pages 1–1, 2023.

[145] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 1306–1315, 2016.

[146] Tomáš Pevny, Patrick Bas, and Jessica Fridrich. Steganalysis by subtractive pixel adjacency matrix. *IEEE Transactions on Information Forensics and Security*, 5(2):215–224, 2010.

[147] Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–0, 2023.

[148] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Int. Conf. Comput. Vis.*, pages 8080–8089, 2019.

[149] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint*, arXiv:1910.10683, 2019.

[150] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034, 2021.

[151] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Eur. Conf. Comput. Vis.*, pages 17–35, 2016.

[152] Wes Robbins, Gabriel Bertocco, and Terrance E. Boult. Daliid: Distortion-adaptive learned invariance for identification – a robust technique for face recognition and person re-identification. *IEEE Access*, pages 1–1, 2024.

[153] Wes Robbins and Terrance E. Boult. On the effect of atmospheric turbulence in the feature space of deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1618–1626, June 2022.

[154] Sridhar Raj S, Munaga V.N.K. Prasad, and Ramadoss Balakrishnan. Spatio-temporal association rule based deep annotation-free clustering (STAR-DAC) for unsupervised person re-identification. *Pattern Recog.*, 122:108287, 2022.

[155] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Conf. Comput. Vis. Pattern Recog.*, pages 420–429, 2018.

[156] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Conf. Comput. Vis. Pattern Recog.*, pages 815–823, 2015.

[157] Ankush Sharma and Amit Sharma. KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density based clustering. In *Int. Conf. Intell. Comput., Instrum. Control Technol.*, pages 787–792, 2017.

[158] Tongzhen Si, Fazhi He, Penglei Li, Yupeng Song, and Linkun Fan. Diversity feature constraint based on heterogeneous data for unsupervised person re-identification. *Inf. Process. Manag.*, 60(3):103304, 2023.

[159] Daniel Silveira. Fire that destroyed National Museum started on the air conditioning of auditorium as told on Federal Police report (in Portuguese). `http://surl.li/swsrl`, 2019. [Online; accessed 11-June-2021].

[160] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1613–1623, 2023.

[161] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–19, 2022.

[162] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recog.*, 102:107173, 2020.

[163] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *Winter Conf. Appl. Comput. Vis.*, pages 2029–2037, 2019.

[164] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint*, arXiv1406.2080, 2014.

[165] He Sun, Mingkun Li, and Chun-Guang Li. Hybrid contrastive learning with cluster ensemble for unsupervised person re-identification. In *Asian Conf. Pattern Recog.*, pages 532–546, 2022.

[166] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[167] Zongzhe Sun, Feng Zhao, and Feng Wu. Unsupervised person re-identification via global-level and patch-level discriminative feature learning. In *IEEE Int. Conf. Image Process.*, pages 2363–2367, 2021.

[168] Hongchen Tan, Xiuping Liu, Yuhao Bian, Huasheng Wang, and Baocai Yin. Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):160–171, 2021.

[169] Haotian Tang, Yiru Zhao, and Hongtao Lu. Unsupervised person re-identification with iterative self-supervised domain adaptation. In *Conf. Comput. Vis. Pattern Recog. Workshops*, pages 1536–1543, 2019.

[170] Qing Tang, Ge Cao, and Kang-Hyun Jo. Fully unsupervised person re-identification via multiple pseudo labels joint training. *IEEE Access*, 9:165120–165131, 2021.

[171] Qing Tang and Kang-Hyun Jo. Unsupervised person re-identification via nearest neighbor collaborative training strategy. In *IEEE Int. Conf. Image Process.*, pages 1139–1143, 2021.

[172] Qing Tang and Kang-Hyun Jo. Unsupervised person re-identification via mining label homogeneity. In *IEEE Int. Conf. Ind. Technology*, pages 1–6, 2022.

[173] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint*, arXiv:1703.01780, 2017.

[174] Antonio Theophilo, Romain Giot, and Anderson Rocha. Authorship attribution of social media messages. *IEEE Trans. Comput. Social Syst.*, 2021.

[175] Antonio Theophilo, Luís AM Pereira, and Anderson Rocha. A needle in a haystack? harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages. In *IEEE Int. Conf. on Acoust., Speech Signal Process.*, pages 2692–2696. IEEE, 2019.

[176] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal Mach. Lear. Res.*, 9(11):2579–2605, 2008.

[177] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, pages 5998–6008, 2017.

[178] Jane Wakefield. Can technology help avoid stampedes? `https://www.bbc.com/news/technology-24463736`, 2013. [Online; accessed 11-June-2021].

[179] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Conf. Comput. Vis. Pattern Recog.*, pages 10981–10990, 2020.

[180] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020.

[181] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[182] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7307, 2022.

[183] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[184] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 2275–2284, 2018.

[185] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. *arXiv preprint*, arXiv:2012.10674, 2020.

[186] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *Conf. Artif. Intell.*, volume 35, pages 2764–2772, 2021.

[187] Menglin Wang, Jiachen Li, Baisheng Lai, Xiaojin Gong, and Xian-Sheng Hua. Offline-online associated camera-aware proxies for unsupervised person re-identification. *arXiv preprint*, arXiv:2201.05820, 2022.

[188] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *Int. Conf. Comput. Vis.*, pages 9358–9367, 2019.

[189] Xueping Wang, Min Liu, Fei Wang, Jianhua Dai, Anan Liu, and Yaonan Wang. Relation-preserving feature embedding for unsupervised person re-identification. *IEEE Trans. Multimedia*, pages 1–10, 2023.

[190] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Conf. Comput. Vis. Pattern Recog.*, pages 5022–5030, 2019.

[191] Yuefeng Wang, Ying Wei, Ruipeng Ma, Lin Wang, and Cuyuan Wang. Unsupervised vehicle re-identification based on mixed sample contrastive learning. *Signal, Image Video Process.*, 16:2083–2091, 2022.

[192] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2022.

[193] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. CycAs: Self-supervised cycle association for learning re-identifiable descriptions. In *Eur. Conf. Comput. Vis.*, pages 72–88, 2020.

[194] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 79–88, 2018.

[195] Hannes Whittingham and Stephanie Kay Ashenden. Chapter 5 - hit discovery. In Stephanie Kay Ashenden, editor, *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, pages 81–102. Academic Press, 2021.

[196] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *Conf. Artif. Intell.*, volume 34, pages 12362–12369, 2020.

[197] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z Li. Unsupervised graph association for person re-identification. In *Int. Conf. Comput. Vis.*, pages 8321–8330, 2019.

[198] Lin Wu, Deyin Liu, Wenying Zhang, Dapeng Chen, Zongyuan Ge, Farid Boussaid, Mohammed Bennamoun, and Jialie Shen. Pseudo-pair based self-similarity learning for unsupervised person re-identification. *IEEE Trans. Image Process.*, 31:4803–4816, 2022.

[199] Yiming Wu, Xintian Wu, Xi Li, and Jian Tian. MGH: Metadata guided hypergraph modeling for unsupervised person re-identification. In *ACM Int. Conf. Multimedia*, pages 1571–1580, 2021.

[200] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *Int. Conf. Learn. Representations*, 2020.

[201] Kun Xie, You Wu, Jing Xiao, Jingjing Li, Guohui Xiao, and Yang Cao. Unsupervised person re-identification via k-reciprocal encoding and style transfer. *Int. J. Mach. Learn. Cybern.*, 12(10):2899–2916, 2021.

[202] Mingyuan Xu, Haiyun Guo, Yuheng Jia, Zhitao Dai, and Jinqiao Wang. Pseudo label rectification with joint camera shift adaptation and outlier progressive recycling for unsupervised person re-identification. *IEEE Trans. Intell. Transp. Syst.*, 24(3), 2022.

[203] Peng Xu and Xiatian Zhu. Deepchange: A long-term person re-identification benchmark with clothes change. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11196–11205, 2023.

[204] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 11926–11935, 2021.

[205] Shiyu Xuan and Shiliang Zhang. Intra-inter domain similarity for unsupervised person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2022.

[206] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Xin Ning, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia*, 24:1665–1677, 2021.

[207] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *Conf. Artif. Intell.*, pages 12597–12604, 2020.

[208] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 4855–4864, 2021.

[209] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 3633–3642, 2019.

[210] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019.

[211] Rajeev Yasarla and Vishal M. Patel. Learning to restore images degraded by atmospheric turbulence using uncertainty. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1694–1698, 2021.

[212] Rajeev Yasarla and Vishal M. Patel. Cnn-based restoration of a single face image degraded by atmospheric turbulence. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):222–233, 2022.

[213] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[214] Junhui Yin, Siqing Zhang, Jiyang Xie, Zhanyu Ma, and Jun Guo. Unsupervised person re-identification via simultaneous clustering and mask prediction. *Pattern Recog.*, 126:108568, 2022.

[215] Junhui Yin, Xinyu Zhang, Zhanyu Ma, Jun Guo, and Yifan Liu. A real-time memory updating strategy for unsupervised person re-identification. *IEEE Trans. Image Process.*, 32:2309–2321, 2023.

[216] Qingze Yin, Guan'an Wang, Guodong Ding, Shaogang Gong, and Zhenmin Tang. Multi-view label prediction for unsupervised learning person re-identification. *IEEE Signal Process. Lett.*, 28:1390–1394, 2021.

[217] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *Conf. Comput. Vis. Pattern Recog.*, pages 2148–2157, 2019.

[218] Jongmin Yu, Junsik Kim, Minkyung Kim, and Hyeontaek Oh. Camera-tracklet-aware contrastive learning for unsupervised vehicle re-identification. In *IEEE Int. Conf. Robot. Autom.*, pages 905–911, 2022.

[219] Jongmin Yu and Hyeontaek Oh. Unsupervised vehicle re-identification via self-supervised metric learning using feature dictionary. In *Int. Conf. Intell. Robots Syst.*, pages 3806–3813, 2021.

[220] Jongmin Yu and Hyeontaek Oh. Graph-structure based multi-label prediction and classification for unsupervised person re-identification. *Appl. Intell.*, 52:14281–14293, 2022.

[221] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Int. Conf. Mach. Learn.*, pages 7164–7173, 2019.

[222] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Int. Conf. Mach. Learn.*, pages 12310–12320, 2021.

[223] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 13657–13665, 2020.

[224] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 9021–9030, 2020.

[225] Yunpeng Zhai, Peixi Peng, Mengxi Jia, Shiyong Li, Weiqiang Chen, Xuesong Gao, and Yonghong Tian. Population-based evolutionary gaming for unsupervised person re-identification. *Int. J. Comput. Vis.*, 131:1–25, 2023.

[226] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. *arXiv preprint*, arXiv:2007.01546, 2020.

[227] Guoqing Zhang, Yu Ge, Zhicheng Dong, Hao Wang, Yuhui Zheng, and Shengyong Chen. Deep high-resolution representation learning for cross-resolution person re-identification. *IEEE Transactions on Image Processing*, 30:8913–8925, 2021.

[228] Guoqing Zhang, Hongwei Zhang, Weisi Lin, Arun Kumar Chandran, and Xuan Jing. Camera contrast learning for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.*, pages 1–1, 2023.

[229] Hongwei Zhang, Guoqing Zhang, Yuhao Chen, and Yuhui Zheng. Global relation-aware contrast learning for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.*, 32(12):8599–8610, 2022.

[230] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 3436–3445, 2021.

[231] Xin Zhang and Ziliang Feng. Unsupervised person reidentification via quantitative random selection for cluster centroid. *Appl. Intell.*, 53:10726—–10733, 2022.

[232] Xin Zhang, Keren Fu, and Yanci Zhang. Graph correlation-refined centroids for unsupervised person re-identification. *Signal, Image Video Process.*, 17:1457–1464, 2023.

[233] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Int. Conf. Comput. Vis.*, pages 8222–8231, 2019.

[234] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 7369–7378, 2022.

[235] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020.

[236] Zhong Zhang, Haijia Zhang, and Shuang Liu. Person re-identification using heterogeneous local graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12136–12145, 2021.

[237] Shizhen Zhao, Changxin Gao, Jun Zhang, Hao Cheng, Chuchu Han, Xinyang Jiang, Xiaowei Guo, Wei-Shi Zheng, Nong Sang, and Xing Sun. Do not disturb me: Person re-identification under the interference of other pedestrians. In *European Conference on Computer Vision*, pages 647–663. Springer, 2020.

[238] Yu Zhao and Qiaoyuan Shu. Unsupervised person re-identification by patch-based nearest neighbor mining. *Digit. Signal Process.*, 133:103885, 2022.

[239] Aihua Zheng, Xia Sun, Chenglong Li, and Jin Tang. Aware progressive clustering for unsupervised vehicle re-identification. *IEEE Trans. Intell. Transp. Syst.*, 23(8):11422–11435, 2021.

[240] Dingyuan Zheng, Jimin Xiao, Mingjie Sun, Huihui Bai, and Junhui Hou. Plausible proxy mining with credibility for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.*, pages 1–1, 2022.

[241] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8522, 2019.

[242] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, pages 1116–1124, 2015.

[243] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Int. Conf. Comput. Vis.*, pages 8371–8381, 2021.

[244] Yi Zheng, Yong Zhou, Jiaqi Zhao, Ying Chen, Rui Yao, Bing Liu, and Abdulmotaleb El Saddik. Clustering matters: Sphere feature for fully unsupervised person re-identification. *ACM Trans. Multimedia Comput., Commun., Appl.*, 18(4):1–18, 2022.

[245] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Conf. Comput. Vis. Pattern Recog.*, pages 1318–1327, 2017.

[246] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Eur. Conf. Comput. Vis.*, pages 172–188, 2018.

[247] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Conf. Comput. Vis. Pattern Recog.*, pages 598–607, 2019.

[248] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(8):2723–2738, 2021.

[249] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Trans. Image Process.*, 28(3):1176–1190, 2018.

[250] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in Pytorch. *arXiv preprint*, arXiv:1910.10093, 2019.

[251] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Int. Conf. Comput. Vis.*, pages 3702–3712, 2019.

[252] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2022.

[253] Ji Zhu, Hua Yang, Weiyao Lin, Nian Liu, Jia Wang, and Wenjun Zhang. Group re-identification with group context graph neural networks. *IEEE Transactions on Multimedia*, 23:2614–2626, 2020.

[254] Kuan Zhu, Haiyun Guo, Songyan Liu, Jinqiao Wang, and Ming Tang. Learning semantics-consistent stripes with self-refinement for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[255] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020.

[256] Wenjie Zhu and Bo Peng. Manifold-based aggregation clustering for unsupervised vehicle re-identification. *Knowl.-Based Syst.*, 235:107624, 2022.

[257] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *European Conference on Computer Vision*, pages 140–157. Springer, 2020.

[258] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. *arXiv preprint*, arXiv:2007.10315, 2020.

Appendix A

Published and under-review papers

In this appendix, we list all the literature produced during this Ph.D. research. They are listed below in chronological order and by Impact Factor (IF).

- **Gabriel C. Bertocco**, Antonio Theophilo, Fernanda Andaló, and Anderson De Rezende Rocha. Leveraging ensembles and self-supervised learning for fully-unsupervised person re-identification and text authorship attribution. *IEEE Transactions on Information Forensics and Security*, 18:3876–3890, 2023.

- W. Robbins, **G. Bertocco** and T. E. Boult. DaliID: Distortion-Adaptive Learned Invariance for Identification – a Robust Technique for Face Recognition and Person Re-Identification.". *IEEE Access*, 2024.

- Nguyen, Kien*; ... **Bertocco, Gabriel**; Boult, Terrance; Andaló, Fernanda; Rocha, Anderson; ... . AG-ReID 2023: Aerial-Ground Person Re-identification Challenge Results. *IEEE International Joint Conference on Biometrics*, 2023.

- Du, Dawei; Hill, Cole; **Bertocco, Gabriel Capiteli**; Pamplona Segundo, Mauricio ; Robbins, Wes J; RichardWebster, Brandon; Collins, Roderic; Sarkar, Sudeep; Boult, Terrance E; McCloskey, Scott. DOERS: Distant Observation Enhancement and Recognition System. *IEEE International Joint Conference on Biometrics*, 2023.

- **G. C. Bertocco**, F. Andaló and A. Rocha. Unsupervised and Self-Adaptative Techniques for Cross-Domain Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4419-4434, 2021.

- Rafael Padilha, Antônio Theóphilo, Fernanda A. Andaló, Didier A. Vega-Oliveros, Joao P. Cardenuto, **Gabriel Bertocco**, Jose Nascimento, Jing Yang, and Anderson Rocha. The Artificial Intelligence and the challenges on the Digital Forensics Science in the XXI century (in Portuguese). *USP Advanced Studies*, 2021.

- Rafael Padilha, Caroline Mazini Rodrigues, Fernanda Alcantara Andaló, **Gabriel Bertocco**, Zanoni Dias, and Anderson Rocha. Forensic event analysis: From seemingly unrelated data to understanding. *IEEE Security and Privacy*, 18(6):23–32, 2020.

- **Gabriel Bertocco**, Fernanda Andaló, Terrance E Boult, and Anderson Rocha. Large-scale fully-unsupervised re-identification. arXiv preprint arXiv:2307.14278, 2023. Under review in IEEE Transactions on Image Processing.

Appendix B

Datasets

In this appendix, we present all datasets employed in this research to evaluate our models. They are described below and summarized in Table B.1. We compare our methods to prior work under the same evaluation setup, i.e., same query and gallery sets described in this section. Furthermore, all results were directly taken from the respective papers, or from other accepted papers. All backbones employed along this thesis are fully finetuned without layer freezing.

`Market1501` [242] has 12,936 images of 751 identities in the training set and 19,732 images in the test set. The test set is divided into 3,368 images for the query set and 15,913 images for the gallery set. We removed "junk" images from the gallery set as done by all previous works, so 451 images were discarded. It has six non-overlapping cameras; each identity is captured by at least two.

`DukeMTMC-ReID` [151] has 16,522 images of 702 identities in the training set and 19,889 images in the test set. The test set is divided into 2,228 query images and 17,661 gallery images of 702 other identities plus 408 distractors. It has eight cameras, and each identity is captured by at least two. This dataset has been discontinued, and it must not be used for evaluation and benchmarking anymore. When this fact became known to us, we had already proposed the first two solutions (Chapters 2 and 3), and, for this reason, we are still providing results considering this dataset. However, following recent literature, we **do not** use it for evaluation in the third and fourth solutions (Chapters 4 and 5), which are the most recently designed in the research. More details in *https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur*.

`MSMT17` [194] has 32,621 images of 1,401 identities in the training set and 93,820 images of 3,060 identities in the test set. The test set is divided into 11,659 images for the query set and 82,161 images for the gallery. It has 15 cameras recording three-day periods (morning, afternoon, and night) on four different days. Out of the 15 cameras, 12 are outdoor, and 3 are indoor. Each identity is captured by at least two cameras. It is one of the most challenging datasets.

`DeepChange` [203] has 75,083 images of 450 identities in the training set. It has a validation and a test set. There are 150 identities on the validation set, which is divided into 4,976 images for the query set and 17,865 images for the gallery. For the test set, there are 521 identities, which are divided into 17,527 images for the query set and 62,956

images for the gallery. It has a total of 17 cameras and each identity is captured by at least two cameras. Differently from previous ReID datasets, `DeepChange` is a clothing-changing dataset, which means that a person can change their clothes when recorded in different locations and moments in time. The dataset has the same person appearing in different moments of the year in different seasons varying their clothes and in, some cases, hairstyles.

`Veri776` [112] has 37,778 images of 576 identities in the training set and 13,257 images of 200 identities in the test set. The test set is divided into 1,678 images for the query set and 11,579 images for the gallery. It has 20 cameras.

`VehicleID` [108] has 113,346 images of 13,164 identities in the training set. It has three evaluation scenarios. The smallest has 800 identities on the test set, which is divided into 5,693 images for the query set and 800 images for the gallery. The middle has 1,600 identities, which is divided into 11,777 images for the query set and 1,600 for the gallery. The largest has 2,400 identities, which is divided into 17,377 images for the query set and 2,400 images to the gallery. The number of cameras is not specified, but all the vehicles have two views: the front and back views.

`Veri-Wild` [119] has 277,797 images of 30,671 identities in the training set. It has three evaluation scenarios. The smallest has 3,000 identities on the test set, which is divided into 3,000 images for the query set and 38,861 images for the gallery. The middle has 5,000 identities, which is divided into 5,000 images for the query set and 64,389 for the gallery. The largest has 10,000 identities, which is divided into 10,000 images for the query set and 128,517 images to the gallery. It has 174 cameras being one of the most challenging re-identification datasets.

`Subset of tweets` are two subsets of the dataset of tweets [174]. The first subset contains messages from 100 authors, which we randomly split into two sets: 50 authors for training and 50 for testing. In the training set, there are 400 tweets per author. We divide the test set into query and gallery sets, with 20 tweets per author in the query set and 300 tweets per author in the gallery set. There are 1000 authors in the second subset, with 500 for training and another 500 for testing. We select 70 tweets per author for the training and gallery sets and 20 tweets per author for the query. The rationale is to verify the capacity for generalization in two scenarios with different complexities. The smallest subset comprises fewer identities and more tweets per author; the other has 10 times more authors and fewer tweets per author.

Table B.1: Information about the used datasets.

	train		**gallery**		**query**	
	#IDs	#Images	#IDs	#Images	#IDs	#Images
`Market1501` [242]	751	12,936	751	15,913	750	3,368
`DukeMTMC-ReID` [151]	702	16,522	1,110	17,661	702	2,228
`MSMT17` [194]	1,041	32,621	3,060	82,161	3,060	11,659
`DeepChange` [203]	450	75,083	150	17,865	150	4,976
			521	62,956	521	17,527
`Veri776` [112]	576	37,778	200	11,579	200	1,678
`VehicleID` [108]	13,164	113,346	800	800	800	5,693
			1,600	1,600	1,600	11,777
			2,400	2,400	2,400	17,377
`Veri-Wild` [119]	30,671	277,797	3,000	38,861	3,000	3,000
			5,000	64,389	5,000	5,000
			10,000	128,517	10,000	10,000
`Subset of tweets` [174]	50	20,000	50	15,000	50	1,000
	500	35,000	500	35,000	500	10,000

Appendix C

Comparison of the third solution to models considering meta information

We compare, in Table C.1, our third proposed method to not-fully unsupervised re-identification methods, i.e., the ones that leverage camera labels or tracklets to help the optimization. Particularly, the camera labels provide strong regularization since they enable the model to train in the same scenario of the cross-camera evaluation. For this reason, methods that consider camera labels have usually higher performance. However, they are **not** fully unsupervised.

Even though our model does not use camera labels, we still have competitive performance, and the best mAP for `Market`. Considering the tracklet-based models, we outperform them in all metrics. Despite the tracklets being a meta-information about the pedestrian motion, which enables the possibility to leverage temporal information, our method better mines the discriminant information just relying upon people's still images.

Similar conclusions can be reached for the `Veri` dataset, as shown in Table C.2. Considering mAP, the best method is MLPL [62] which relies on multi-part analysis that improves feature description but adds complexity to the training process. Our method relies solely on the feature map extracted from the bounding boxes, without any kind of sub-part analysis. The same conclusions can be drawn from Table C.3 where we compare our method to MLPL in `VehicleID` dataset. Other methods employ camera or viewpoint labeling, making the task easier. Indeed, the best R1 is achieved by DiDAL [114], which employs camera labels for optimization. Some methods employ segmentation (MAPLD [120]) or color information (VRPRD [6]) to help optimization, but our model performs better than both without any kind of supervision or side information.

In `Veri-Wild` (Table C.4), following previous conclusions, the inclusion of camera and viewpoint labeling demonstrates a robust potential to enhance model learning capabilities. Different from `Veri`, the segmentation-based model MAPLD [120] achieves higher performance than ours. Since `Veri-Wild` is the most challenging dataset, any side information has the potential to aid model learning. As our model operates in the fully unsupervised scenario, it sometimes faces performance drops compared to other methods that rely upon meta-information or labeling.

Table C.1: Comparison with relevant Person ReID methods considering some meta-information or camera/viewpoint labeling. The best result is highlighted in **blue**.

		Market				**MSMT17**			
Method	Reference	mAP	R1	R5	R10	mAP	R1	R5	R10
Camera-based									
SSL [106]	CVPR'20	37.8	71.7	83.8	87.4	-	-	-	-
CCSE [105]	TIP'20	38.0	73.7	84.0	87.9	9.9	31.4	41.4	45.7
MPRD [76](*)	ICCV'21	51.1	83.0	91.3	93.6	14.6	37.7	51.3	57.1
Xie *et. al.* [201]	IJMLC'21	54.1	82.6	91.3	94.5	13.4	37.5	48.5	52.0
DSCE-MC [208]	CVPR'21	61.7	83.9	92.3	-	15.5	35.2	48.3	-
JVTC [95]	ECCV'20	47.5	79.5	89.2	91.9	17.3	43.1	53.8	59.4
JGCL [19]	CVPR'21	66.8	87.3	93.5	95.5	21.3	45.7	58.6	64.5
IICS [204]	CVPR'21	72.9	89.5	95.2	97.0	26.9	56.4	68.8	73.4
IIDS [205]	TPAMI'22	78.0	91.2	96.2	97.7	35.1	64.4	76.2	80.5
CAP [186]	AAAI'21	79.2	91.4	96.3	97.7	36.9	67.4	78.0	81.4
CCTSE [7]	TIFS'21	67.7	89.5	94.8	96.5	-	-	-	-
CAPL [113]	NCA'23	80.4	92.8	97.3	-	40.7	71.2	81.4	-
MGH [199]	ICM'21	81.7	93.2	96.8	98.1	40.6	70.2	81.2	84.5
ICE [17]	ICCV'21	82.3	93.8	97.6	98.4	38.9	70.2	80.5	84.4
O2CAP [187]	TIP'22	82.7	92.5	96.9	98.0	42.4	72.0	81.9	85.4
O2CAP-IBN [187]	TIP'22	83.7	93.1	97.4	98.1	**46.9**	**75.5**	**84.8**	**87.7**
CASTOR-ICE [202]	TITS'22	82.8	93.6	97.5	98.5	41.7	72.3	82.3	85.8
CASTOR-CCL [202]	TITS'22	84.5	93.0	97.8	98.6	33.2	61.9	74.0	78.2
Liu *et. al.* [110]	TIP'22	82.4	93.0	-	-	38.4	68.6	-	-
Liu *et. al.*-IBN [110]	TIP'22	82.0	92.8	-	-	42.4	71.6	-	-
CIFL [141]	TMM'22	82.4	93.9	97.9	98.1	38.8	70.1	80.7	83.9
RTMem [215]	TIP'23	83.1	93.9	97.7	98.4	40.8	72.0	81.5	84.6
MCSL [59]	OPTIK'23	83.5	93.7	97.5	98.4	38.7	71.1	80.8	84.3
PPLR [26]	CVPR'22	84.4	94.3	97.8	98.6	42.2	73.3	83.5	86.5
PPSL [198]	TIP'22	68.7	88.6	95.2	96.6	40.9	71.1	83.3	87.0
PPSL(Concat) [198]	TIP'22	82.3	94.1	97.4	**98.8**	43.1	73.2	89.4	90.8
PEG [225]	IJCV'22	84.5	94.3	98.0	98.5	44.9	73.9	83.2	86.3
PPCL+CAP [240]	TCSVT'22	82.4	94.0	98.1	-	37.8	70.8	80.7	-
PPCL+ICE [240]	TCSVT'22	82.8	93.9	97.6	-	39.8	70.8	81.2	-
DiDAL [114]	TMM'23	84.8	94.2	**98.2**	-	45.4	74.0	84.3	-
CCL [228]	TCSVT'23	85.3	94.1	97.8	**98.8**	41.8	71.4	-	-
Multi-part based models									
PPLR [26]	CVPR'22	81.5	92.8	97.1	98.1	31.4	61.1	73.4	77.8
LPur [90]	TIP'23	**85.8**	**94.5**	97.8	98.7	39.5	67.9	78.0	81.6
Tracklet-based									
Star-Dac [154]	PR'21	33.9	67.0	80.6	84.9	-	-	-	-
TSSL [196]	AAAI'20	43.3	71.2	-	-	-	-	-	-
UTAL [98]	TPAMI'20	46.2	69.2	-	-	13.1	31.4	-	-
CycAs [193]	ECCV'20	64.8	84.8	-	-	26.7	50.1	-	-
UGA [197]	ICCV'19	70.3	87.2	-	-	21.7	49.5	-	-
Fully-Unsupervised									
Ours		**85.8**	94.0	97.7	98.5	43.2	70.9	80.8	84.2

Table C.2: Comparison with relevant Vehicle ReID methods in the `Veri776` dataset considering some meta-information or camera/viewpoint labeling. The best results are highlighted in **blue**.

		Veri		
Method	Reference	mAP	R1	R5
Camera/Viewpoint-based				
SSL [106]	CVPR'20	23.8	69.3	72.1
VAPC [239]	TITS'21	30.4	76.2	81.2
CAPL [113]	NCA'23	41.1	87.3	91.3
O2CAP [187]	TIP'22	41.9	87.5	92.7
O2CAP-IBN [187]	TIP'22	42.4	89.6	**93.5**
CCL [228]	TCSVT'23	42.6	87.0	-
Liu *et. al.* [110]	TIP'22	43.2	87.0	-
Liu *et. al.*-IBN [110]	TIP'22	43.9	88.9	-
PPLR [26]	CVPR'22	43.5	88.3	92.7
DiDAL [114]	TMM'23	43.5	**89.0**	93.5
CTACL [218]	ICRA'22	44.2	81.6	89.5
Segmentation-based				
MAPLD [120]	TITS'23	33.4	78.7	83.5
Multi-part-based models				
PPLR [26]	CVPR'22	41.6	85.6	91.1
MLPL [62]	TVT'22	**45.1**	88.3	91.1
Attribute-based Models				
Method	Reference	mAP	R1	R5
VRPRD [6]	PR'19	40.1	83.2	91.1
Fully-Unsupervised				
Ours		41.3	86.3	89.9

Table C.3: Comparison with relevant fully-unsupervised Vehicle ReID methods in `VehicleID`. The best result is highlighted in **blue**.

		TS = 800		**TS = 1600**		**TS = 2400**	
Part-based Models							
Method	Reference	mAP	R1	mAP	R1	mAP	R1
MLPL [62]	TVT'22	**65.3**	**61.1**	**62.7**	**57.3**	**59.6**	**52.4**
Fully Unsupervised							
Method	Reference	mAP	R1	mAP	R1	mAP	R1
Ours		61.7	56.0	59.7	53.4	56.9	50.1

Table C.4: Comparison with relevant Vehicle ReID methods in `Veri-Wild` dataset considering some meta-information or camera/viewpoint labeling. The best result is highlighted in **blue**.

		Veri-Wild (Small)			**Veri-Wild (Medium)**			**Veri-Wild (Large)**		
Method	Reference	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5
Camera/Viewpoint-based										
SSL [106]	CVPR'20	16.1	38.5	58.1	17.9	36.4	56.0	13.6	32.7	48.2
VAPC [239]	TITS'21	33.0	**72.1**	**87.7**	28.1	64.3	83.0	22.6	55.9	75.9
CTACL [218]	ICRA'22	**58.2**	71.1	86.6	**49.2**	**69.2**	83.7	**41.2**	**60.1**	**81.5**
Segmentation-based										
MAPLD [120]	TITS'23	36.6	**72.1**	87.6	33.4	66.2	**84.5**	27.7	55.9	77.3
Fully-Unsupervised										
Ours		29.9	54.1	76.6	25.8	46.4	69.2	20.0	36.2	59.0

Appendix D

Long Distance Recapture Data

Long-range recognition is relevant in many applications. However, the collection of biometric data is extremely expensive and time-consuming. Currently, the most related dataset, IJB-S [86], is not available for common academic use, and an earlier dataset at 100M [52] was withdrawn from public use. Furthermore, IJB-S is not a strictly long-range dataset. To overcome the lack of available long-range data, some prior works have used simulated atmospheric turbulence as a proxy for real data [212, 211, 153]. However, the effectiveness of simulated atmospherics for face recognition has not been validated because, as mentioned before, there is no real data for validation.

We recapture datasets through the atmosphere to facilitate academic research on biometric recognition over long distances. To perform the capture, we use three 4k outdoor televisions, a 4k Basler camera, and an 800mm lens with a 1.4x adapter. Custom capture and display software are developed for the collection, and custom mounting hardware is built for stable capture. The displays are mounted to avoid direct sunlight on the screens. The camera is directed at the displays from a structure at a distance of 770 meters, and videos of the displays are captured at 30 frames per second. A video of the displays running is provided on the GitHub site of the data, where considerable atmospheric effects can be seen. Our collection setup yields significant atmospheric distortions, which can even be noticed between sequential frames. D.1 shows two examples. The data collection process went through IRB approval and is being distributed for non-commercial use.

We refer to our recapture datasets as the original dataset name followed by “-LD" (“the LD datasets"), where LD stands for long-distance. The evaluation datasets provided are LFW-LD and CFP-LD. Twelve recaptured samples are provided for each image in the original dataset because atmospheric turbulence is temporally variable. For CFP-LD and LFW-LD, two protocols are proposed: clean-to-long-distance (C-to-LD) and long-distance-to-long-distance (LD-to-LD). C-to-LD uses verification pairs where one image is standard (and thus higher quality). For LD-to-LD, all samples are recaptured over long distances. The LD datasets expand the evaluation of our methods in the following section, where evaluations are made with a single frame for each image. However, future work should consider new protocols allowing frame fusion or frame selection across frames. Previous unconstrained evaluation datasets (e.g., IJB-S) have been distributed over a terabyte of raw video, which is burdensome to process. In contrast, the LD datasets are pre-processed and pre-aligned in the same format as the original datasets, which

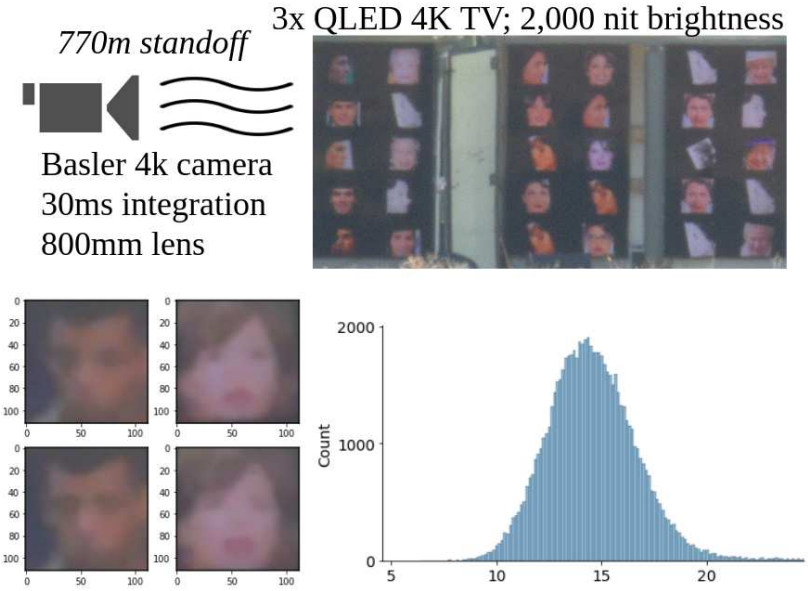


Figure D.1: Top. Recapture specifications and a raw frame from our recollection. **Lower Left.** Two consecutive frames (33.3ms apart) for two different identities from LFW-LD. Differences can be observed between sequential frames, such as around the eyes or face outline. **Lower Right.** Distribution of feature distances in degrees between sequential frames of the same display image from LFW-LD and CFP-LD. Surprisingly, the distances are not 0 – the effects of atmospherics from frame to frame are considerable!

streamlines evaluation and comparison. The final release will include the recapture of person re-identification datasets, plus a WebFace4M recapture for training.

The collection setup went through IRB approval, and both the LFW and CFP dataset licenses allow redistribution. Specifications of imaging equipment and collection conditions are shown in Table D.1. Figure D.2 shows the display and Figure D.3 shows the camera used for recapture. The LD datasets contain twelve recaptured face chips for each original face as the capture occurs continuously over time, and atmospheric turbulence is temporally variable (atmospheric effects are shown in Figure D.4 and at `https://youtu.be/cBcik5U7kfM`). The nature of the data allows for research uses such as frame selection, frame aggregation, distortion robustness, quality prediction, and direct feature comparisons to the same image with and without real atmospheric turbulence.

To post-process the images, fixed regions from the screens are cropped, and then RetinaFace [32] face detector is used to detect landmarks and re-align the images. Non-local mean denoising algorithm is used to reduce noise in the recaptured images. D.4 shows samples from the LD datasets.

Parameter	Value
Camera	Basler acA2440-35uc
Lens focal length	800mm +1.4x Extender
Capture distance	770 meters
Integration time	30μs
Capture rate	30 fps
Wind speed	5-15mph
Temperature	15$^{\circ}C$

Table D.1: camera, weather condition, and display settings for the collection of LFW-LDand CFP-LD.

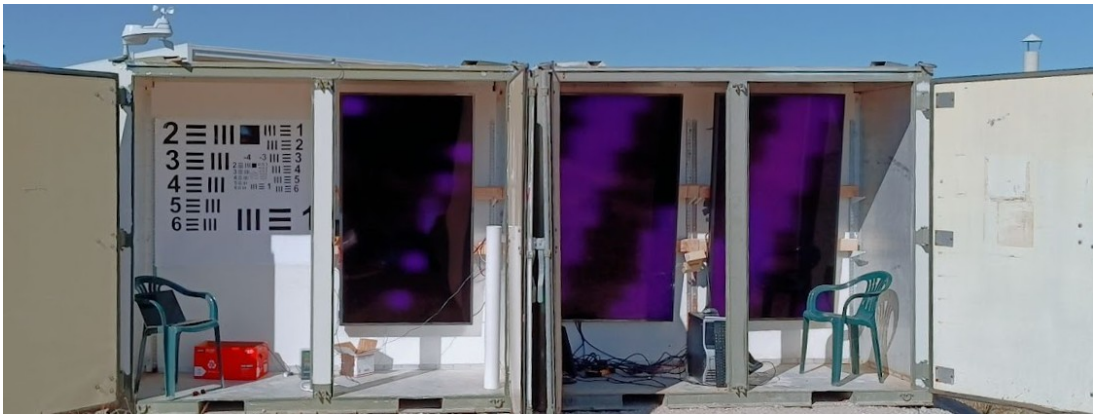


Figure D.2: 3x 75" 4k OLED 2,000 nit outdoor displays mounted in containers for recapture.

Figure D.3: Lens and camera with custom mounting hardware for recapture.

Figure D.4: Sample images from the LD datasets. It can be seen that our recapture setup yielded significant atmospheric turbulence effects (*also see video at* `https://youtu.be/cBcik5U7kfM`). These datasets can facilitate research into 1) quality/confidence-aware models, 2) models that are robust to face-feature distortion, and 3) frame aggregation under atmospheric turbulence (12 frames are provided per display image).