An Investigation into the Use of ConvNext within IICS/IIDS Framework for Person Re-ID

by

Roya Dehghani

A thesis submitted to the
School of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Masters of Science in **Computer Science**

Faculty of Science
University of Ontario Institute of Technology (Ontario Tech University)
Oshawa, Ontario, Canada
May 2023

THESIS EXAMINATION INFORMATION

Submitted by: **Roya Dehghani**

Master of Science in Computer Science

Thesis Title: An Investigation into the use of ConvNeXt within IICS/IIDS Framework for Person Re-ID

An oral defense of this thesis took place on September 25th, 2023 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Patrick Hung
Research Supervisor	Dr. Faisal Qureshi
Examining Committee Member	Dr. Heidar (Kourosh) Davoudi
Thesis Examiner	Dr. Miguel Vargas Martin

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

In this thesis, we explore the integration of ConvNeXt, a CNN-based network inspired by vision transformers, into the Intra and Inter Camera Similarity (IICS) and Intra and Inter Domain Similarity (IIDS) frameworks for unsupervised person Re-ID. Building upon IICS/IIDS framework that generates pseudo labels through intra and inter stages and utilizing techniques such as Adaptive Instance and Batch Normalization (AIBN) and Transform Normalization (TNorm) to minimize intra-camera and inter-camera variations respectively, our work emphasizes the application of ConvNeXt as a feature extractor. ConvNeXt gets higher mAP and CMC on the Market1501 and MSMT17 datasets than most unsupervised learning methods. Furthermore, we explored the effect of AIBN and TNorm normalization techniques in ConvNeXt. We showed their effectiveness in reducing intra-camera and inter-camera variations if AIBN is inserted in the final stages (Stage 3 and stage 4) and TNorm layers are included after stage 1, stage 2, and stage 3. We also examined the effects of four ConvNeXt variants within the IICS/IIDS framework, emphasizing the advantages of using larger variants of ConvNeXt as a feature extractor for person Re-ID.

Keywords: Pedestrian Identification; Unsupervised Learning, Pseudo Labels, Computer Vision; Deep Learning; Surveillance; Camera Networks

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I authorize the Ontario Tech University to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the Ontario Tech University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Roya Dehghani

Statement of Contributions

I hereby certify that I am the sole author of this thesis, and I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

Acknowledgements

I want to sincerely thank my academic supervisor, Dr. Faisal Qureshi, for his advice and support during my research endeavors. His knowledge, perceptive criticism, and guidance have all been crucial in determining the focus and success of my work. I sincerely appreciate his dedication to my academic advancement.

In addition, I want to express my sincere thanks to the Vector Institute and Ontario Tech University for the funding package throughout my study. Their kindness has made it possible for me to devote myself fully to my studies.

Finally, I want to thank my colleagues, family members, and friends. Throughout my academic career, their unconditional love, support, and encouragement have served as a consistent source of inspiration and drive. My ability to overcome obstacles and pursue success is a result of their confidence in my talents and their ongoing support in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Person Re-Identification (ReID) involves matching a person seen in a camera to individuals seen previously in this or other cameras. The problem of person Re-ID naturally arises in wide-area surveillance settings where person Re-ID enables a myriad of downstream tasks, e.g., those related to security, search and rescue, and smart spaces. Consider, for example, a shopping mall that has installed a collection of cameras to boost its security and increase the sense of safety for its visitors. These cameras record images of individuals who visit this mall. Person Re-ID technology, for example, will enable the camera operators to tie visitors of this mall to their previous visits. Similarly, person Re-ID can also help track the motion of individuals as they meander through the mall, moving in and out of the field-of-view of different cameras. Person Re-ID is a long-standing problem in the video surveillance community and there is a large body of work in this area, including a number of schemes developed over the last decade that leverage deep learning (Ye et al., 2021).

Despite the volume of work in this area and the substantial interest in this capability by security agencies, police, and military, the problem of person Re-ID is far from solved. Matching and identifying individuals in images taken under a variety of viewing conditions—e.g., camera types and viewpoints, lighting, occlusions, pose variations, clothing changes, multiple individuals wearing similar clothes, etc.—is a challenging task. If we consider also that the images are taken at different times, say days or weeks apart, the problem of person Re-ID becomes even more intractable. As mentioned earlier, the problem of person Re-ID typically takes place in video surveillance scenarios where oftentimes it is not possible to acquire a high-quality photograph of an individual. This suggests that person Re-ID cannot use biometric information for the purposes of identification.

Within this context, this thesis aims to study the problem of person Re-ID. For the purposes of this work, we assume that the persons are photographed by multiple cameras with possibly non-overlapping fields-of-view. Furthermore, there are no clothing changes between images captured by different cameras. Our reasons for these assumptions are practical: we had access to three person Re-ID datasets—Market1501 (Zheng et al., 2015), DukeMTMC-ReID (Ristani et al., 2016), and Multi Scene Multi Time (MSMT17) (Wei et al., 2018b)—and these datasets contain images from multiple cameras that are taken around the same time. Therefore, each person is shown wearing the same clothes in the images where this individual is visible. The imaging assumptions that we have made in this work are similar to those made by other recent approaches for person Re-ID.

In order to match individuals across multiple cameras, it is important to develop techniques that are able to compute camera-invariant features. Specifically, these features need to capture identify-related information, ignoring confounding factors, e.g., pose, background, partial visibility, camera viewpoints, and color shifts, etc. Over the last few years, researchers have investigated deep learning approaches for the purposes of constructing features that are well-suited for person Re-ID applications. Here, CNN backbones that are pre-trained on imagenet type data are employed as feature extractors and the computed features are matched using nearest-neighbor-inspired approaches or calculate the distance between these features and ranke them from the most matched features to the least ones. More recently, metric learning and domain adaptation techniques have been used to improve the robustness and generalizability of these features.

The purpose of metric learning is to push features that are of the same individual closer to each other while at the same time pulling apart the features of different persons. Domain adaptation, on the other hand, explicitly models the differences in images (of the same person) captured by different cameras. These differences arise due to imaging characteristics of various cameras, e.g., viewpoint, lighting, color shifts, etc. Both of these techniques assume that the images are labeled, i.e., each image is labeled with the "identity" of the individual seen in this image. Specifically, it means that 1) only one individual is visible in an image and 2) if two images show the same person then both images have the same label. Of course, surveillance cameras capture images that show multiple individuals; however, it is still fair to assume (1). It is relatively straightforward to perform person detection followed by a tight crop to construct an image where only a single person is visible. Item (2), i.e., labeling person Re-ID datasets is tedious, consequently, the current trend is to develop person Re-ID techniques that eschew labeled data. One commonly used approach is to assign pseudo-labels and use these to train models capable of constructing features suitable for person Re-ID applications. These methods combine feature extraction and label assignment within a single loop: extracted features are clustered to assign labels that are used to refine the feature extractor, and so on.

Feature similarity computation is the fundamental operation underlying any person Re-ID scheme. The idea is simple. Features corresponding to the same individual are "closer" to each other; whereas, those that belong to different persons sit "farther" away from each other. Within person Re-ID, however, we need to deal with camera/domain gaps that arise when the same person is captured by multiple cameras.

Two recent methods called IICS and IIDS for unsupervised person Re-ID generates pseudo-labels in two stages, 1) intra and 2) inter (Xuan & Zhang, 2021; Xuan & Zhang, 2022). Furthermore, to reduce the effect of intra-camera and inter-camera variations on the accuracy of pseudo-labels, Adaptive Instance and Batch Normalisation (AIBN) and Transform Normalization (TNorm) techniques are added to the feature extractor (ResNet50). Our work closely follows IICS/IIDS approach. We employ a CNN-based network, ConvNeXt, designed by the principles of vision transformers as a feature extractor into the IICS/IIDS framework.

Furthermore, we study the effect of adapting AIBN and TNorm techniques to ConvNeXt. AIBN reduces the impact of intra-camera variations caused by different poses, viewpoints, occlusion, and other identity-related factors. Our result shows that AIBN can be effective if inserted in the final stages of ConvNeXt (stage 3 and stage 4). Also, to mitigate the effect of inter-camera variations, TNorm layers are added after stage 1, stage 2, and stage 3 of ConvNeXt to convert the style of images taken by different cameras. The result shows that TNorm is effective and is it highly related to its insertion location.

Finally, we investigate four variants of ConvNeXt in IICS/IIDS Framework, and the result shows that we reach higher mAP and CMC when we use a larger variant of ConvNeXt as a feature extractor. Fig. 5.5 compares the different states of our method with IICS and IIDS methods on the MSMT benchmark dataset, which is the most challenging among Market1501 and Duke. ConvNeXt-s (isotropic) shows the smallest variant of ConvNeXt without adopting AIBN and TNorm normalization techniques. ConvNeXt-s (AIBN) shows AIBN is included in ConvNeXt. Also, ConvNeXt-s (AIBN, TNorm) is when we insert AIBN and TNorm techniques into ConvNeXt. Also, ConvNeXt-B (AIBN, TNorm) shows a larger ConvNeXt variant, including AIBN and TNorm, showing the highest accuracy among other states of our method. Also, Fig. 5.6 compares our approach to IICS and IIDS methods on Market1501.

1.1 Social Impact

According to the Almasawa 2019 Survey, video surveillance installations have an immediate impact on crime prevention, suspect identification, and maintaining public safety in congested regions, transit hubs, and other security-sensitive sectors (Almasawa et al., 2019). Law enforcement uses person Re-ID routinely in crime investigations. Manual person Re-ID where a person has to go through troves of image data to identify the culprit is tedious, sometimes impractical, and often extremely time intensive. In order to improve efficiency and speed up the identification process, law enforcement organizations also use (automated) person re-identification procedures. By eliminating the need for human annotation and utilizing unsupervised learning techniques, the advances in this thesis provide workable solutions to the resource-intensive issue of identity annotation. Law enforcement authorities may strengthen their investigation abilities, speed up suspect identification, and increase their overall efficacy by using this development (Tahboub, 2017).

It is important to bear in mind that the widespread use of person Re-ID technologies has serious social consequences. This technology erodes individual privacy. Additionally, it can be easily misused by authoritarian regimes to quell dissent and control behavior.

Consequently, It is important to understand where and when it is appropriate to use this technology. The society, as a whole, has to balance its needs for security and a sense of safety to the very real and important issue of personal privacy and freedom of thought, movement, and action.

1.2 Thesis Outline

In chapter 2, we start by assessing supervised and unsupervied methods in person Re-ID. In chapter 3, we explain IICS/IIDS framework, ConvNeXt, and AIBN and TNorm normalization techniques. Then, we discuss the formulation of our methodology in chapter

4. This involves using a different network as a feature extractor. Chapter 5 shows the performance of our method, providing a thorough assessment of its effectiveness in comparison to other approaches. Chapter 6 concludes with a summary of our contributions, findings, and recommendations for further research in person Re-ID.

1.3 Source Code

The Python implementation of our model can be accessed through the following link.

https://github.com/faisalqureshi/roya-dehghani-msc-person-reid

Chapter 2

Related Works

This chapter provides a review of the related research in person Re-ID. We start with an introduction about person Re-ID and then we categorize methods into supervised and unsupervised learning to investigate the various techniques researchers employ for person Re-ID. We also identify the commonly used datasets in person Re-ID based on our review of the previous works. Finally, we discuss our contribution which is based on the literature review.

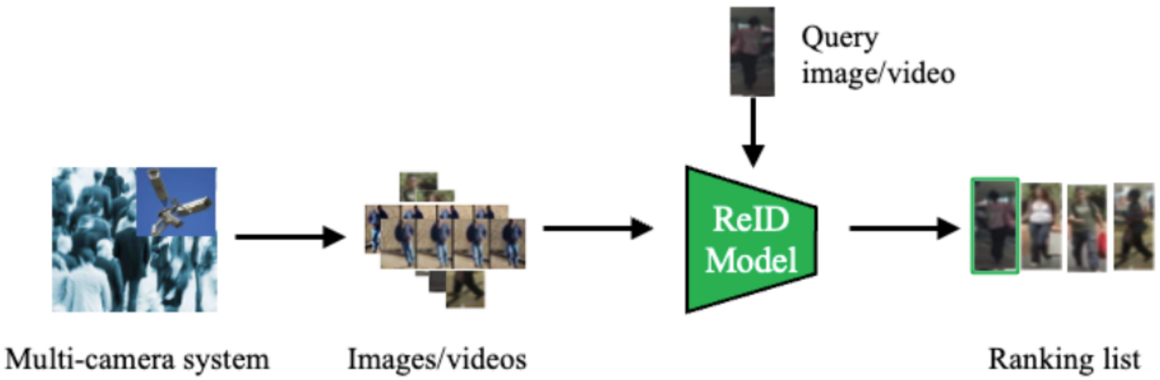


Figure 2.1: A common system for person Re-ID.

2.1 Supervised Person ReID

Person Re-ID may be used for various practical purposes, such as criminal investigations, multi-camera tracking, and missing person searches. Fig. 2.1 shows the process of person Re-ID where person Re-ID's main goal is to identify distinguishing characteristics of people to match a given query person with the gallery's most comparable persons in pictures or videos. Early person Re-ID models often relied on hand-crafted attributes, such as colors and textures, to capture the visual similarities between different images (Liu et al., 2018; Cheng et al., 2017; Liu et al., 2017). However, these hand-crafted features are not sufficient for large-scale applications since they often cannot capture visual aspects found in large galleries. Starting in 2014, Convolutional Neural Network, CNN-based deep learning replaced the role of the hand-crafted feature technique, and it emerged as the dominant technique in computer vision research (Li et al., 2014).

Thanks to CNN-based deep learning, supervised person Re-ID has achieved impressive accuracy on commonly used Re-ID benchmark datasets. For example, in the Market-1501 dataset (Zheng et al., 2015), the Rank-1 benchmark for single query search has increased from 44.4% (Zheng et al., 2015) at the time of its release to 98.5% (Liu et al., 2020) in 2020. Similarly, the Rank-1 evaluation result for the DukeMTMC-Re-ID dataset has shown substantial improvement, rising from only 30.8% (Zheng et al., 2017)in 2017 to over 95% in 2020.

In exploring supervised learning for person ReID, AlignedReID (Zhang et al., 2017) has been examined. AlignedReID is a technique for person re-identification that combines global and local feature learning processes. To extract global and local properties, it uses a two-branch network, the global branch, and the local branch. The global branch's convolutional network analyses the input picture and produces feature maps. The most important or noticeable characteristics are then preserved while the spatial dimensions are decreased by utilizing a global pooling operation to compress feature maps. However, the local branch generates a feature map using the same convolutional network as the global branch. Next, one (1 × 1) kernel-sized convolutional layer and horizontal pooling are used. The network is trained using triplet hard loss, which chooses triplet samples based on global distances among global features.

Furthermore, (Sun et al., 2018) utilizes a Part-based Convolutional Baseline (PCB) in conjunction with Refined Part Pooling (RPP). PCB separates the picture into horizontal parts instead of processing it as a whole and learns unique attributes for each of these parts. One image is first put through a series of convolutional layers, often using the backbone network from a pre-trained model like ResNet50. This process produces a 3-dimensional tensor. Next, the tensor is divided into some column vectors, followed by (1 × 1) kernel-sized convolutional layer reduces the dimension of each column vector. Then, a classifier receives each dimension-reduced column vector, in turn, to create the final descriptor of the input picture by concatenating all generated column vectors.

On the other hand, (Luo et al., 2019) concentrated on training methodologies that are covered in other articles or source codes to create a strong baseline for person Re-ID. The recommended baseline model is based on ResNet50 and was trained using a combination of classification loss and triplet loss. The authors also offer a novel neck structure called BNNeck to split metric and classification losses into two different feature spaces. For this purpose, after the global pooling layer, the batch normalization layer is added before the fully connected layer, which is used for classifying identities. Separation of feature spaces for triplet loss and classification loss improves the performance of person Re-ID.

However, (Liu et al., 2019) employ hash code to provide individual photos for quick indexing and retrieval on huge datasets. The suggested method uses the Deeper Cut method, which determines the locations of 14 critical points for each person's picture, to break down human photos into bodily parts including head, arm, upper body, and lower body. Then these pieces are combined with other photographs to produce new positive and negative training examples. This decomposition procedure makes it easier to create useful training examples, which enhances feature learning. The technique is known as “self-guided” since it creates training samples repeatedly on its own, which results in discriminative hash codes that may be effectively learned even with a little amount of labeled data.

In contrast to the previous method which is focused on binary hash code, VP-ReID employs modern techniques such as Deep Convolutional Neural Networks (DCNNs), and efficient off-line indexing (Wei et al., 2018a). The technique uses Deepercut to distinguish between various human body parts, particularly the head, upper body, and lower body. Then, to obtain both global and regional properties, a Convolutional Neural Network (CNN) made up of four sub-networks is utilized. Additionally, an offline stage of the retrieval procedure is used to arrange pictures using the hierarchical organization technique known as Temporally Dependent Clustering (TDC). Images that belong to the same person will be grouped together in TDC, resulting in an index that will speed up the retrieval process later on.

For clothing-changing person re-identification (re-id) in RGB photos and videos, (Gu et al., 2022) suggests a unique Clothes-based Adversarial Loss (CAL). The proposed strategy penalizes the re-id model’s clothing-specific prediction ability to separate clothes-relevant characteristics from clothes-irrelevant ones. A classifier for clothing follows the re-id model’s core, and CAL is defined as a classification loss with many positive classes. The authors additionally create a new dataset called Clothes-Changing Video person re-ID (CCVID) using the raw data of a gait recognition dataset to serve as a publicly accessible benchmark for clothes-changing video person Re-ID. The suggested technique is a supervised learning technique that trains users using identification and clothing information.

However, (Somers et al., 2023) proposed a model for occluded person Re-ID called BPBreID that uses body part representations to overcome occlusions. The parts of BPBreID are a global-local representation learning module that generates body part-based features of the Re-ID target, a body part attention module that predicts attention maps highlighting the body parts of the Re-ID target, and a novel training method called GiLt that is resistant to occlusions and non-discriminative local appearance. The studies on well-known holistic and occluded datasets revealed that BPBreID beat state-of-the-art algorithms on the challenging Occluded-Duke dataset by 0.7% mAP and 5.6% rank-1 accuracy. The study concludes that part-based approaches are advantageous for occluded person ReID because they provide fine-grained data and are well-suited to represent human bodies that are only partially visible.

Similarly, another innovative method for dealing with occlusion in person re-identification is the Dynamic Prototype Mask (DPM) technique. In order to transfer the alignment from occluded retrieval to the subspace selection job, it makes use of prototype classification. This method eliminates the additional pre-trained networks that were used to deliver body cues while also keeping the knowledge from the global wisdom and achieving automatic alignment. To fully utilize the potential of DPM, the DPM technique also makes use of a Hierarchical Mask Generator (HMG) and a Head Enrich Module (HEM). The DPM technique can be very helpful in real-world situations, as in surveillance systems, where occlusion is a frequent problem with human re-identification. It can significantly increase the effectiveness and precision of individual re-identification in such circumstances (Tan et al., 2022).

Table 2.1 shows the summary of explained supervised person ReID methods. Deep learning-based techniques depend on a lot of labeled data, which takes a lot of time and it is almost infeasible on a large scale to annotate. Due to the substantial costs involved in data labeling, supervised person Re-identification (Re-ID) techniques have difficulty scaling up to huge datasets.

Table 2.1: Related Works Summary of Supervised Person ReID

Type of method	Summary	Ref.
Local and Global Features	Combination of Global and Local Feature Learning	(Zhang et al., 2017)
Part-based Features	Part-based Convolutional Baseline	(Sun et al., 2018)
Useful Techniques	Re-ID strong Baseline	(Luo et al., 2019)
Hash code	Binary Hash Codes	(Liu et al., 2019)
DCNNs	DCNNs, Off-line Indexing, and Distance Metric Optimization	(Wei et al., 2018a)
Clothes-based Adversarial Loss (CAL)	Clothes-based Adversarial Loss for Clothes-changing Person Re-ID	(Gu et al., 2022)
Part-based ReID model	Body Part Representations	(Somers et al., 2023)
Occlusion Re-ID Approach	Dynamic Prototype Mask (DPM)	(Tan et al., 2022)

2.2 Unsupervised Person ReID

The objective of unsupervised learning methods is to train Re-ID models utilizing unlabeled data, decreasing the reliance on labeled data. We categorized unsupervised learning methods into three groups based on the techniques they use.

2.2.1 Distribution Alignment

Feature distribution alignment of pictures recorded by multiple cameras can be a way to overcome camera/domain gaps. This alignment strategy is utilized to bridge the camera gap between cameras. (Sun & Saenko, 2016) focus on the situation where the target domain lacks labeled data, requiring unsupervised adaptation. The CORAL method aims to address the camera shift problem by aligning the second-order statistics of the source and target distributions using a linear transformation. The method follows a three-step process: feature extraction, transformation application, and training of an SVM classifier. Deep CORAL proposes a direct integration of the CORAL technique into deep networks. This is done by creating the CORAL loss, a differentiable loss function that reduces the difference between the correlations of the source and target domains. A deep neural network that has already been trained can adapt its features to a new target domain with the help of Deep CORAL loss, which makes it easier to learn a non-linear transformation. A pre-trained model is used to set the basic network parameters, and the labeled source data is then used to fine-tune them.

Similarly, (Wu et al., 2019b) introduced a strategy to address the camera gap in feature spaces due to variations in inter-camera scenes like illumination and viewpoint, leading to conflicting pairwise similarity distributions and impacting matching performance. The proposed solution employs camera-aware similarity consistency learning in a two-step coarse-to-fine methodology. A key component, the Camera-Aware Similarity Consistency Loss, preserves understanding of intra-camera similarities while learning both local and global similarity consistency. This loss function also explores the relationship between intra-camera and inter-camera matching and uses a coarse-to-fine consistency learning technique, with global and local phases, to enhance similarity learning. The approach helps in retrieving the right top-ranked samples for person re-identification and depicts the camera-aware similarity inconsistency. Fig. 2.2 shows that ResNet50 model (He et al., 2015), pre-trained on the MSMT17 dataset (Wei et al., 2018b), was used to match samples in two cameras (denoted by Cam 1 and Cam 2) on the DukeMTMC dataset (Ristani et al., 2016). In intra-camera matching or inter-camera matching, pairwise similarities are computed between each pair of samples, and the distributions are shown on the left. The top-8 cosine similarity matches are given on the right, with the accurate matches indicated by green bounding boxes. The variance in the feature space caused by the inter-camera scene results in inconsistent pairwise similarity distributions, which degrades matching efficiency (Wu et al., 2019b)

Likewise, (Lin et al., 2018) put out the Multi-task Mid-level Feature Alignment (MMFA) network, a unique unsupervised methodology. With the use of a mid-level feature alignment regularisation term, the model attempts to improve the tasks of classi-

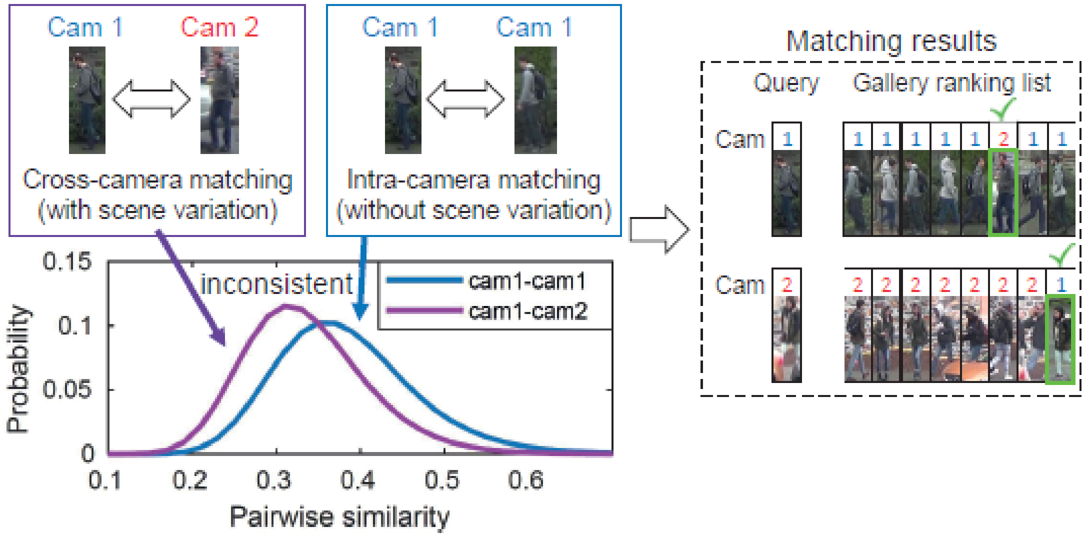


Figure 2.2: The camera-aware similarity inconsistency problem illustration.

fying people's identities as well as learning their attributes. By treating the final feature mappings of the feature extractor as attribute-like mid-level features, the MMFA network enables mid-level deep feature alignment. As a result, mid-level properties between the source and target domains can be aligned. The suggested technique uses a domain adaptation strategy for mid-level feature alignment to minimize the Maximum Mean Discrepancy (MMD) between the source and target domains' mid-level feature distributions.

In order to overcome the noisy distillation problem and preserve feature space structure during evolution, (Lu et al., 2022) suggests a novel data-free incremental person ReID framework dubbed AGD that makes use of geometric distillation and dreaming memory. To enhance the quality of the distilled information, AGD augments distillation in a pairwise and cross-wise manner over several perspectives of memory. In order to avoid exemplars from arbitrarily "roiling" the space structure, AGD preserves connections between exemplars as representations drift. This allows it to modify the feature space for new knowledge while keeping rich prior information for retrieval.

2.2.2 GANs (Generative Adversarial Networks)

In other studies, GANs (Generative Adversarial Networks) are used to translate the style of photographs taken by one camera to images taken by another camera. This method seeks to close the feature distribution camera gap between pictures captured by various cameras.

To tackle the challenge of image style variations caused by different cameras, CycleGAN is proposed (Zhong et al., 2018b). Based on a training image taken by a particular camera, the approach may create equivalent images that appear to have been shot by various cameras by utilizing the learned CycleGAN models. The procedure mitigates camera style discrepancies and lowers the danger of convolutional neural network (CNN) overfitting. It may be used as a data augmentation tool as well. The technique makes learning pedestrian descriptors with a camera-invariant attribute easier by adding camera information, enabling more reliable feature extraction.

Similar to the previous approach, GANs are used for the distribution of translated images to be identical to the target domain (Zhu et al., 2017). In the absence of matched instances, the authors offer a method for learning to translate an image from a source domain to a target domain by introducing a cycle consistency loss, which assures that translating an image from to and vice versa with functions and produces an image that is close to the original input.

Previously studied methods often adopt a representation space containing id-related and unrelated features, limiting efficiency. (Zou et al., 2020a) addressed this issue by using a disentangling module to separate id-related and irrelevant features and an adaptation module to work on alignment and self-training in a shared appearance space. These co-developed modules offer mutual benefits. The disentangling module simplifies adaptation by focusing only on id-related data. In contrast, the adaption module bridges the distribution gap between cameras/domains and assists in separating appearance and structure characteristics. Self-training further supports disentangling by encouraging unique appearance features. The joint learning framework solves the mismatch between training and testing data, bridging camera gaps and enhancing performance in new cameras or environments.

Similarly, to close the gap between cameras and lessen the need for annotating new instances of training, the Person Transfer Generative Adversarial Network (PTGAN) approach is suggested (Wei et al., 2018b). Style transfer and the preservation of individual identification are two key restrictions that PTGAN is made to operate by. For transferred person photos to have styles comparable to those in the target dataset, style transfer seeks to learn style mapping functions between various person datasets. The person identity preservation goal assures that each person's identity remains unchanged after the transfer, which is critical for person re-identification training. The person re-identification datasets lack matched person pictures (shots of the same person from different datasets), which makes the style transfer task an unpaired image-to-image translation challenge. The last category under unsupervised learning methods in person Re-ID is pseudo-label-based methods that we study in the next section.

2.2.3 Generating Pseudo Labels

Pseudo-label-based methods in unsupervised learning initially generate pseudo-labels by applying predefined rules based on sample similarity. These pseudo-labels are then utilized to train the Re-ID model. The accuracy and reliability of the computed pseudo-labels play a crucial role in determining the performance of these methods. High quality pseudo-labels that accurately capture the underlying patterns and semantics of the data lead to improved performance, as they effectively guide the model training process. Conversely, low-quality or erroneous pseudo-labels can negatively impact the performance of the Re-ID model, potentially leading to suboptimal results.

Most pseudo-label-based methods use clustering algorithms based on distance or similarity criteria among extracted features of images (Chen et al., 2020; Zhang et al., 2019;

Wang & Zhang, 2020). (Fan et al., 2018) introduced progressive unsupervised learning (PUL) using k-means clustering to generate pseudo-labels and transfer pre-trained deep representations to unknown environments. The method consists of an iterative process of pedestrian grouping and CNN fine-tuning. Initially, when the model is weak, the CNN is fine-tuned using a carefully selected set of trustworthy examples near cluster centroids. As the model improves, more images are progressively chosen for training, allowing the model to learn from a broader and more varied dataset. Both pedestrian clustering and the CNN model are enhanced simultaneously, naturally following the principles of self-paced learning, where the task gets more challenging as the model improves. Through the iterative approach, PUL improves the original model's performance, adapting to the target camera and yielding better results in unsupervised learning scenarios.

Similarly, (Lin et al., 2019) proposes a novel Bottom-Up Clustering (BUC) approach. The suggested technique optimizes the interaction between various samples and a convolutional neural network (CNN) model. In the first stages of training, each distinct picture is treated as a distinct identity. To extract feature embeddings from the pictures, the CNN model is used. The number of unique classes or identities is then decreased using a bottom-up clustering approach on the feature embeddings. This clustering step allows for discovering relationships and similarities among the samples in an unsupervised manner. The CNN model continuously learns from various unlabeled photos, gradually exploiting the dataset's similarities.

Similar work in pseudo-label-based methods adds a memory bank to the network architecture to eliminate the need to re-initialize the classifier at each epoch and propose memory-based multi-label classification loss (MMCL) (Wang & Zhang, 2020). Every image in the suggested method receives a single-class label before moving on to multi-label classification utilizing the modified Re-ID model for label prediction. The label prediction approach uses cycle consistency and similarity computing to guarantee the accuracy of the predicted labels. To increase the Re-ID model's training effectiveness in multi-label classification, the research develops the memory-based multi-label classification loss (MMCL). The suggested Re-ID approach significantly improves due to MMCL operations and iterative label prediction.

Other researchers focus on improving the quality of pseudo-labels. To reduce the influence of low-quality pseudo-labels, NRMT (Zhao et al., 2020), MMT (Ge et al., 2020a) and MEB-Net (Zhai et al., 2020b) used mutual-training. (Zhao et al., 2020) introduced a noise-resistant mutual-training architecture to address the problem of label noise for person Re-ID. This approach aims to generalize Re-ID models from a labeled source domain to an unlabeled target domain where the data distribution may differ significantly. The framework consists of two main components: mutual training and noise resistance. Mutual training involves two Re-ID models, a source model, and a target model, alternately updated using each other's pseudo-labels, helping them learn and improve in the target domain. Noise resistibility strategies, such as label smoothing and a noise-resistible loss function, are employed to reduce the impact of noisy pseudo-labels that could harm model performance. Overall, the framework enhances the models' adaptability to the target domain by making them more resilient to label noise.

Moreover, (Ge et al., 2020a) proposed a method called Mutual Mean-Teaching (MMT) to decrease the negative effect of noisy pseudo labels. To reduce the effects of label noise, the proposed MMT architecture has a two-step pseudo label refinement process. The refining procedure uses offline and online refined hard and soft pseudo labels. By using updated labels, the model improves its feature representations and better reflects the characteristics of the target domain. The study also provides a novel soft softmax-triplet loss to handle the gently refined labels.

Similarly, (Zhai et al., 2020b) employs ensemble learning techniques to propose Multiple Expert Brainstorming Network (MEB-Net). In MEB-Net, several networks with various architectures are pre-trained as expert models inside a source domain using a mutual learning approach. Each expert model has unique abilities and information.

Through a brainstorming process that involves mutual learning among the experts, the objective is to adapt these expert models to the target domain. In each iterative epoch, clustering algorithms forecast pseudo-labels for the target data. The expert networks are then adjusted using these pseudo-labels through mutual learning. Mutual learning makes it easier for experts to share information, which boosts performance in the target domain.

In addition, (Zhang et al., 2018) describes a deep mutual learning (DML) technique that extends the notion of model distillation by allowing numerous student networks to collaborate and educate each other throughout the training process. Unlike the traditional model distillation, which distributes knowledge from a static instructor to a student, DML allows for bidirectional knowledge transmission among student networks. The DML approach exemplifies how mutual training among fundamental student networks may produce noteworthy outcomes for category and instance recognition tasks. The experiments demonstrate that existing strong teacher networks are not necessary for excellent performance. The shared learning of student networks outperforms traditional distillation methods that rely on a more strong but static instructor.

(Zhu et al., 2022) suggests a pre-training technique for person re-identification (ReID) named Part-Aware Self-Supervised Pre-Training (PASS). The technique, which is based on the Transformer architecture, is made to draw out specific information from photos to enhance Re-ID performance. PASS creates part-level features by segmenting pictures into several local regions and giving each part a unique learnable token. The strategy is assessed and found to perform better on several benchmark datasets than current state-of-the-art approaches. The study also contains visualization experiments to demonstrate the model's focus regions and its capability to handle occlusion conditions.

However, The Intra and Inter Camera Similarity (IICS) approach proposed by (Xuan & Zhang, 2021) offers a solution for acquiring accurate pseudo-labels for training the Re-ID network. The method divides the computation of sample similarity into two steps. The first stage is Intra-camera computing, which uses CNN features directly, and similarity calculations are made within each camera. This step generates the pseudo-label by clustering samples and giving them the same label within each cluster. The training of the Re-ID model is conducted using a multi-branch network, where each branch is optimized for a different classification task within the same camera, and multiple tasks optimize the common backbone. The second step uses both CNN feature similarity and Jaccard similarity among classification score vectors in inter-camera computation. The Adaptive Instance and Batch Normalisation (AIBN) technique is also introduced to improve classifier generalization without diminishing discriminative performance.

In addition to AIBN, Transform Normalization (TNorm) and knowledge distillation were added to the IICS method and proposed Intra and Inter Domain Similarity (IIDS) method (Xuan & Zhang, 2022). TNorm is a normalization technique that reduces inter-camera variations caused by camera-related factors. It essentially transfers the style of images taken by one camera to those taken by another. This transformation is achieved by altering the camera-related feature statistics, effectively converting the image style. The work also involves self-knowledge distillation, where information is transferred between the original features and TNorm-simulated features. Knowledge distillation is commonly used to transmit information from a “teacher” model to a “student” model, and it can also facilitate information sharing among multiple samples bearing the same label. Further details on TNorm and AIBN are provided in the subsequent chapter.

Table 2.2 shows a summary of methods in unsupervised learning methods. In unsupervised learning, no labeled data is used. Some research tries to align the feature distributions of images captured by different cameras. These methods seek a transformation or mapping function that aligns the feature distributions without the need for labeled data by learning from the innate structure of the data. Another group of researchers uses GANs to transfer the style of images taken by one camera to the images taken by a different camera. By doing so, the researchers could reduce the camera gap between feature distribution of images taken by different cameras. In contrast to other methods, other researchers generate pseudo labels to supervise the training model. It is interesting to note that those methods that use pseudo labels perform better than other methods in unsupervised learning.

2.3 Dataset

There are three commonly used datasets that most researchers use to test the performance of person Re-ID models, namely Market1501 (Zheng et al., 2015), DukeMTMC-ReID (Ristani et al., 2016), and MSMT17 (Wei et al., 2018b).

Market1501 dataset contains 32,668 photos with 1,501 distinct IDs. 6 cameras were placed in front of a campus grocery to collect these photographs. The Deformable Part Model (DPM) (Felzenszwalb et al., 2009) was used to achieve precise identification and cropping. The training set is used for model training and contains 12,936 photos exhibiting 751 identities. The gallery portion, which contains 19,732 photos depicting 750 identities, serves as a reference for comparison during evaluation. Also, the query subset, which contains 3,368 hand-drawn photos corresponding to the same 750 gallery IDs, is used to test the model's performance. The photos in the Market-1501 dataset are all 128 by 64 pixels in size. To make this dataset's reference easier to understand throughout this thesis, we shall use the word "Market". Fig. 2.3 illustrates some sample images from the Market1501 dataset.

DukeMTMC-reID dataset is derived from the larger DukeMTMC dataset (Ristani et al., 2016), which is primarily concerned with pedestrian tracking. The DukeMTMC collection contains 36,411 photos with 1,812 unique IDs recorded from 8 distinct cameras. Similarly to Market-1501, DukeMTMC-reID is broken into three major components. The training subset includes 16,522 photos representing 702 identities, whereas the gallery part includes 17,661 images representing 1,110 identities. In addition, the gallery includes

Table 2.2: Related Works Summary of Unsupervised Person Re-ID

Type of method	Summary	Ref.
Distribution Alignment	Distributions Alignment of Source and Target Data	(Sun & Saenko, 2016)
	The Camera-Aware Similarity Consistency Loss	(Wu et al., 2019b)
	Optimizing Identity Classification, Feature Alignment	(Lin et al., 2018)
	Incremental Person Re-Identification	(Lu et al., 2022)
GANs	Cycle Consistency Loss	(Zhu et al., 2017)
	Introduction of CycleGAN	(Zhong et al., 2018b)
	Disentangling Id-related and Id-unrelated Features	(Zou et al., 2020a)
	MSMT17 Dataset Introduction, Person Transfer GAN	(Wei et al., 2018b)
	Separation of Positive and Negative Samples	(Jin et al., 2020)
Pseudo label Generation	K-means Clustering	(Fan et al., 2018)
	Bottom-up Clustering	(Lin et al., 2019)
	Memory-based Multi-label Classification Loss	(Wang & Zhang, 2020)
	Mutual-training Framework, domain shift and Label Noise	(Zhao et al., 2020)
	Mutual Mean-Teaching (MMT) Framework, Refining Pseudo Labels	(Ge et al., 2020a)
	Multiple Expert Brainstorming Network (MEB-Net)	(Zhai et al., 2020b)
	Unsupervised ReID Pre-Training	(Zhu et al., 2022)
	Two-stage Generation Pseudo Label Approach, AIBN Inclusion to Feature Extractor	(Xuan & Zhang, 2021)
	Two-stage Generation Pseudo Label Approach, AIBN and TNorm Inclusion to Feature Extractor, Self-Knowledge Distillation Between TNorm-simulated Features and Original Features	(Xuan & Zhang, 2022)

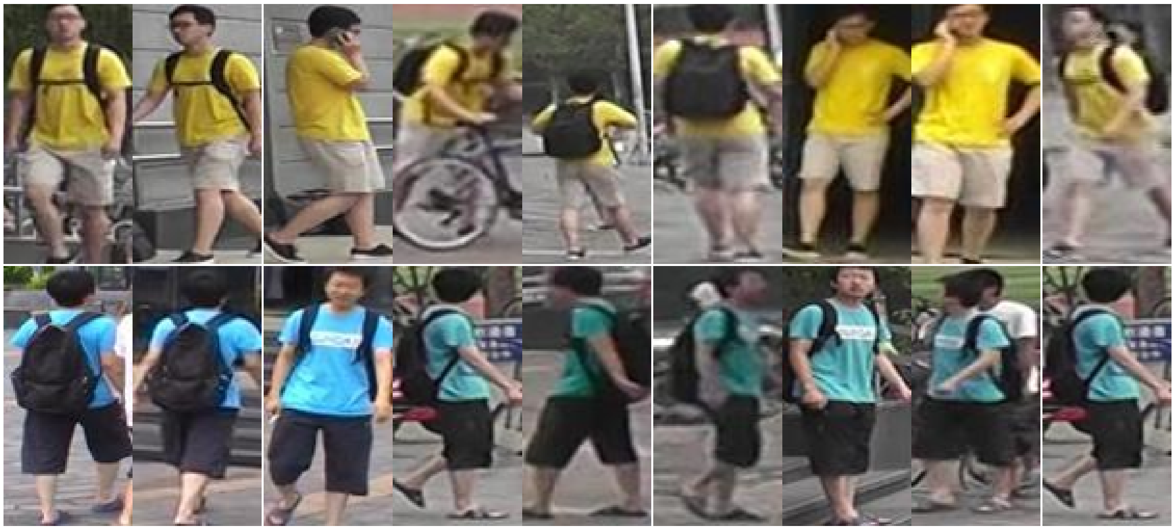


Figure 2.3: Market1501 sample images.

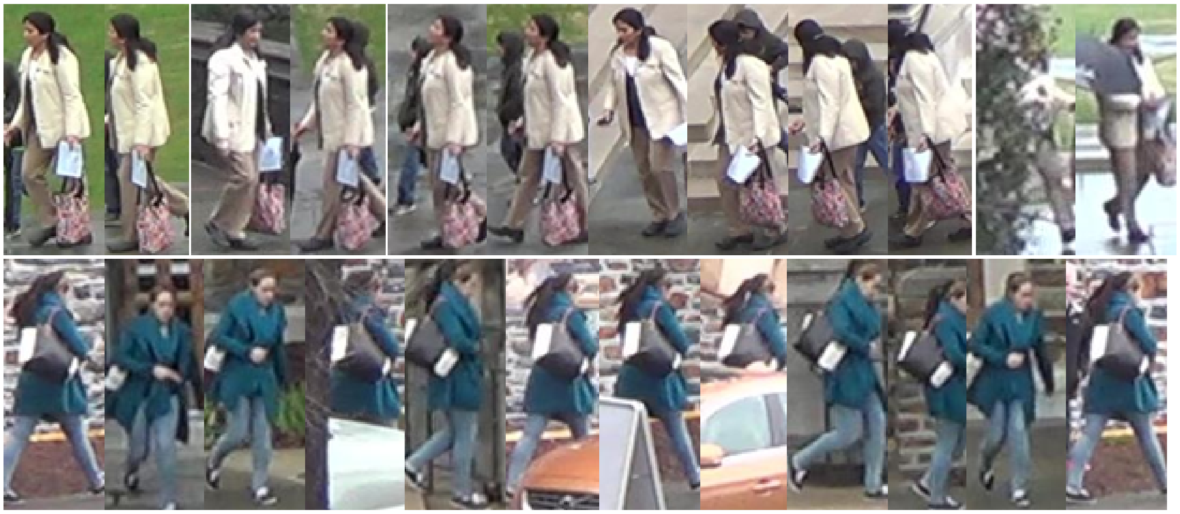


Figure 2.4: Duke sample images.

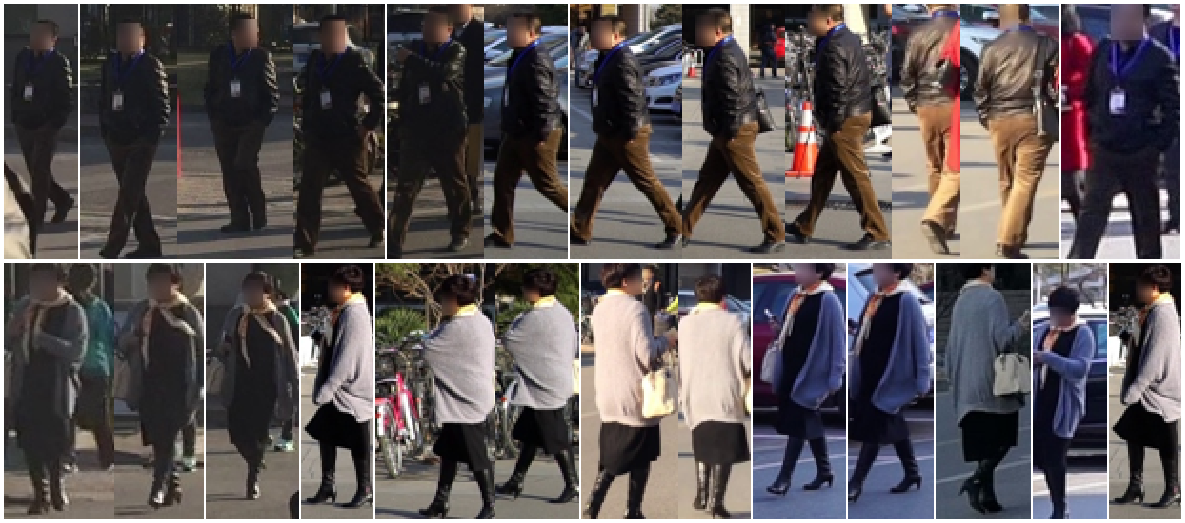


Figure 2.5: MSMT17 sample images.

an extra 2,228 photos with 702 IDs for searching purposes. It is worth mentioning that the photos in this collection vary in size. We shall use the word "Duke" to simplify references to this dataset throughout this thesis report (Fan et al., 2018). Fig. 2.5 depicts Some sample images. It should be noted that the Duke dataset was retracted due to social concern.

MSMT17 is a recent dataset for person Re-ID (Wei et al., 2018b). It consists of 126,441 pictures from 4,101 different people that were taken using 15 different cameras. There are 32,621 photos from 1,041 identities for training and 93,820 images from 3,060 identities for testing. The dataset is more difficult to analyze than the DukeMTMC-ReID and Market1501 datasets because of the variety of scene alterations, the long time range,— day or night time—, and the high number of unique IDs. Fig. 2.5 shows some sample images from this dataset.

2.4 Summary of Person Re-ID Methods

In the previous sections, we examined different methods for person Re-ID, both with and without annotated data. Although supervised learning with deep neural networks has shown impressive results, it is problematic for large-scale datasets since it needs labeling. However, unsupervised learning approaches, which do not require labels for training Re-ID models, have been investigated by researchers as a solution to this problem.

Regarding supervised learning person Re-ID, we assessed a few methods. Some researchers focus on both local and global features for person Re-ID. Others focus on deep learning hash code. On the other hand, in unsupervised person Re-ID, we can categorize methods into three groups. Some methods try to align the distribution of images, and others utilize Generative Adversarial Networks (GANs) to convert the style of images taken by one camera to the ones taken by another camera. However, some generate pseudo labels to supervise model training.

2.5 Our Contributions

Based on our literature review, we can conclude that the unsupervised learning approach is more practical in the real world than supervised learning where labeled data is needed. Since the person Re-ID model's primary goal is to extract discriminative features, we examine a CNN network that has been designed based on a vision transformer as the feature extractor into the IICS/IIDS approach. To this end, we employ ConvNeXt, a CNN-based network as a feature extractor (Liu et al., 2022).

The reason to choose ConvNeXt is that IICS and IIDS methods examined VGGNet-19 (Simonyan & Zisserman, 2014) and GoogLeNet (Szegedy et al., 2015). We decided to examine a recent CNN-based architecture whose performance is better than the vision transformer (Liu et al., 2022). We did not use the vision transformer because our data is images, not sequential data such as Videos. Our contributions can be summarized as follows:

- **ConvNeXt-based Feature Extraction in IICS/IIDS:** One of the major strides we take is to extend the IICS/IIDS framework by incorporating a ConvNeXt as a feature extractor.

- **Exploration of ConvNeXt Variants:** We conduct a study of four different ConvNeXt variants to unravel their suitability and effectiveness within the context of person Re-ID, specifically within the IICS/IIDS framework.

- **Incorporation of AIBN and TNorm with ConvNeXt:** We investigate the impact of integrating Adaptive Instance Batch Normalization (AIBN) and Transform Normalization (TNorm) within the ConvNeXt framework.

- **Intra and Inter-Stage Analysis:** IICS/IIDS divided the problem of training feature extractors for person Re-ID into two steps: intra-camera and inter-camera. In this thesis we analyzied the roles played by these two stages and we confirm the findings in IICS/IIDS that both stages lead to better feature extractors, leading to better performance on the task of person Re-ID.

Chapter 3

Background Knowledge

This chapter begins with an explanation of the IICS/IIDS framework, followed by a detailed explanation of the ConvNext structure, comparing it to Residual Network (ResNet) and Swin Transformer (Swin-T) for the purpose of person Re-ID. We also discuss other variants of ConvNeXt. Finally, we will delve into the AIBN and TNorm normalization techniques and explain how we can integrate them into ConvNeXt.

3.1 IICS/IIDS Framework

IICS/IIDS Framework is an unsupervised learning framework where no labels are used. This approach uses pseudo labels to supervise the training process (Xuan & Zhang, 2021; Xuan & Zhang, 2022). Fig. 3.1 depicts that the IICS/IIDS framework has two different stages for training the person Re-ID model. In the first stage, intra-stage, there is a multi-branch network where a feature extractor is shared among all cameras, and for each camera, there is a different classifier. Images taken by each camera are given to the feature extractor and clustering is used to generate pseudo labels based on the Euclidean similarity among CNN features. On the other hand, in inter stage, the network is the same feature extractor used in the intra-stage and only one classifier. The weights of the feature extractor are directly copied from the intra-stage. At this stage, images from all cameras are given to the feature extractor to extract features and do clustering based on CNN-feature similarity and Jaccard similarity on classification score vectors from classifiers that exist in intra stage. In the following sections, each stage will be explained in more detail.

3.1.1 Intra-camera training

In the Intra stage, the dataset is divided based on the index of cameras. This separation helps to concentrate on capturing the distinct visual characteristics present in each camera's field of view. Fig. 3.3 shows that the images taken by camera 1 are given to the feature extractor to extract CNN feature vectors. Based on Euclidean similarity on CNN feature, agglomerative clustering generates pseudo labels. The same process is done for other cameras' images as well. Generated pseudo labels are then used to train the feature extractor and classifiers. The network at this stage can be considered a multi-branch network where a feature extractor is shared among cameras, and each branch (classifier) is responsible for each camera. Training Intra network helps to reduce the variations between different cameras. During the Re-ID process, the shared embedding space enables direct comparisons between images taken by various cameras. Fig. 3.3 illustrates that for each camera, a different classifier is provided for each branch separately. The outputs of classifiers are compared to the clustering-derived pseudo-labels to compute the softmax cross-entropy loss. All branches are trained with their own losses and add all losses together and update the weights of the network.

3.1.2 Inter-camera training

At this stage, as Fig. 3.4 shows, the network here is a shared feature extractor and only one classifier. The probability that two samples from different cameras belong to the same identity is estimated by the Jaccard similarity between classification score vectors. A camera-independent feature should be used to reflect the similarity between

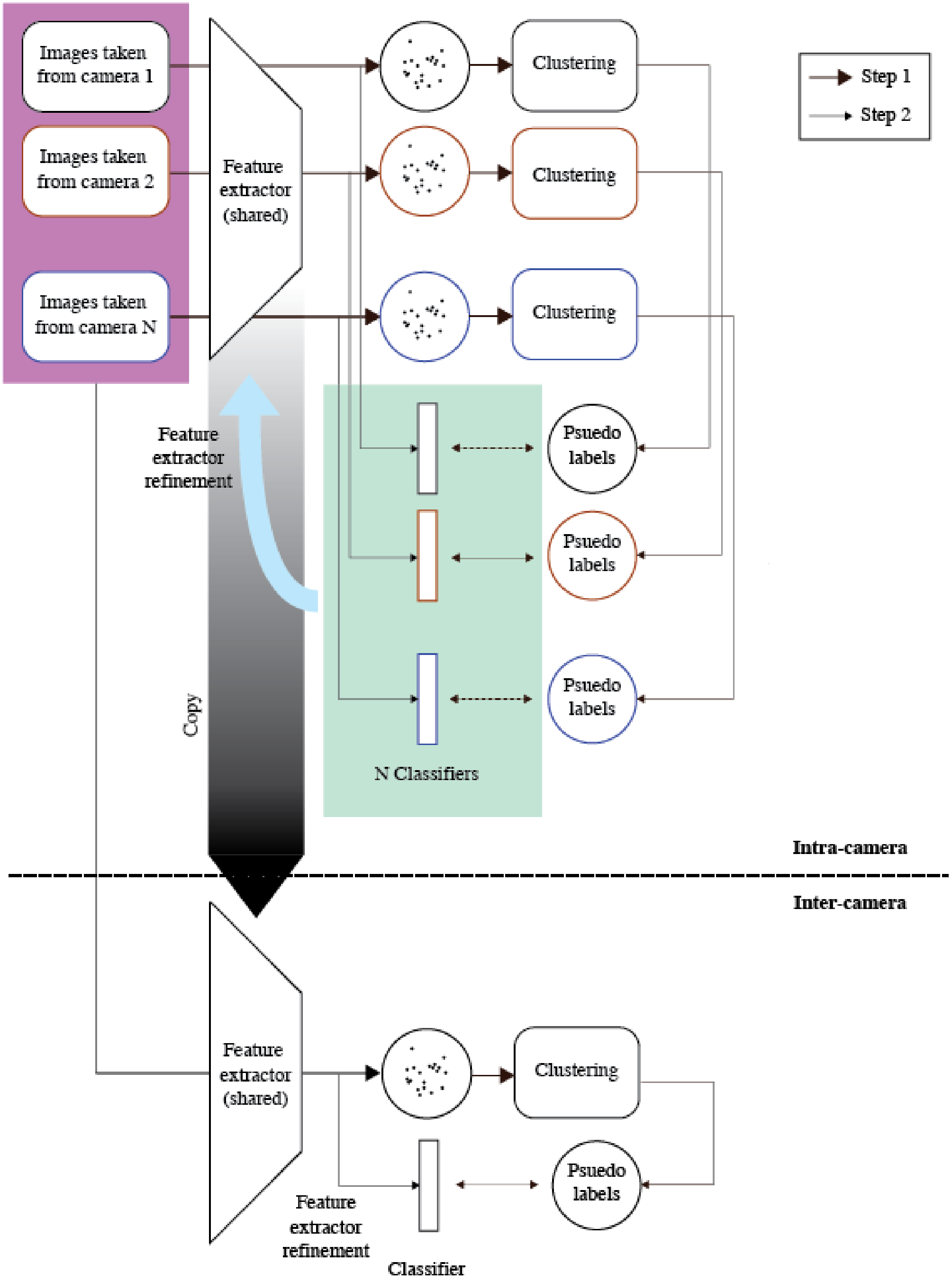

Images taken from camera 1
Images taken from camera 2
Images taken from camera N
Feature extractor (shared)
Clustering
Clustering
Clustering
Step 1
Step 2
Feature extractor refinement
Psuedo labels
Psuedo labels
Psuedo labels
Copy
N Classifiers
Intra-camera
Inter-camera
Feature extractor (shared)
Clustering
Psuedo labels
Feature extractor refinement
Classifier

Figure 3.1: IICS/IIDS structure. (Xuan & Zhang, 2021; Xuan & Zhang, 2022).

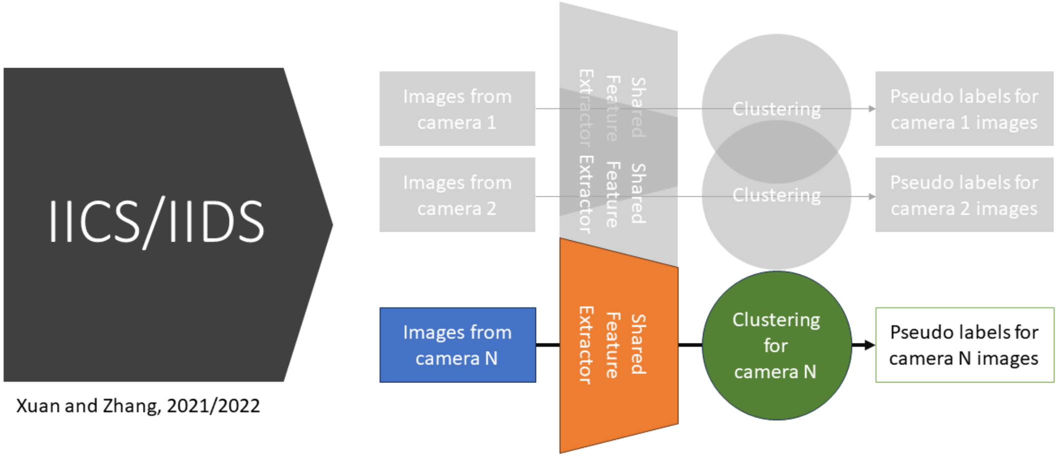


Figure 3.2: Intra-camera training stage. Feature extraction and clustering process is done for each camera separately.

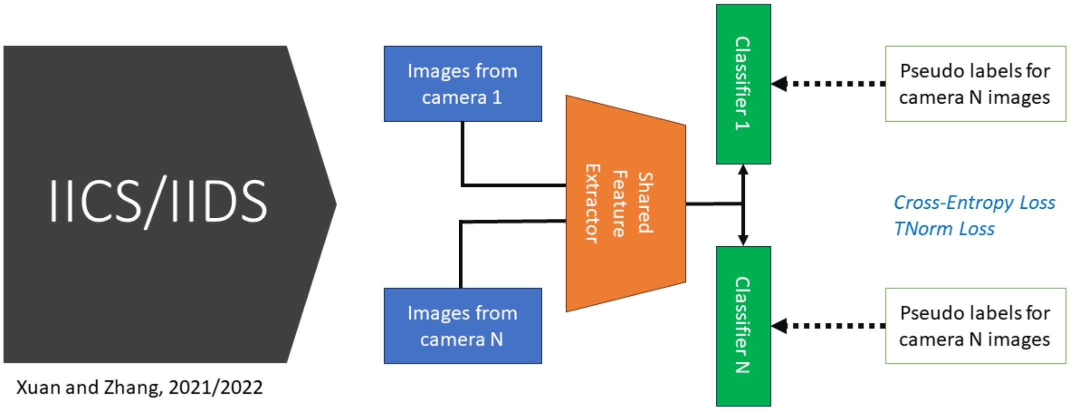


Figure 3.3: Intra-camera training stage. It shows the network at the Intra stage and how pseudo labels are used to train the network.

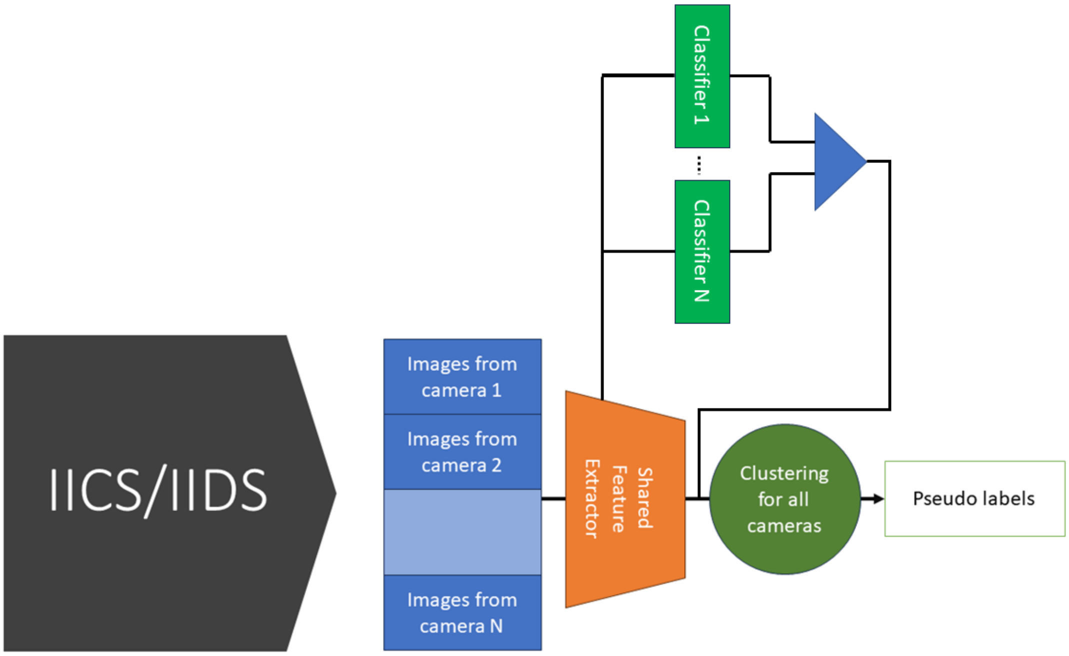


Figure 3.4: Inter-camera training stage. It shows the process of generating pseudo labels. Clustering at this stage incorporates not only Euclidean CNN similarity but also the Jaccard similarity between classification score vectors. This Jaccard similarity is computed by concatenating the classification scores of classifiers from the Intra-camera stage.

the classification probabilities that each classifier in the Intra stage produces for a given image. The CNN feature of each image is passed through all classifiers that exist in the Intra stage, then concatenates all these probabilities as a vector (classification score vector). The Jaccard similarity is used to indicate the probability that the two images belong to the same identity. At this stage, in addition to CNN feature similarity, Jaccard similarity is considered to do clustering to generate pseudo-labels across cameras. Fig. 3.5 depicts that generated pseudo-labels are then used to train the network using a combination of softmax cross entropy and triplet losses. The softmax cross entropy loss ensures that the model accurately classifies each image, while the triplet loss is used to learn discriminative features that can separate different identities.

In the IICS method, AIBN normalization is integrated into the feature extractor to reduce intra-camera variations. However, in the IIDS method, AIBN and TNorm are inserted into the feature extractor to reduce Inter-camera variations. TNorm is used to

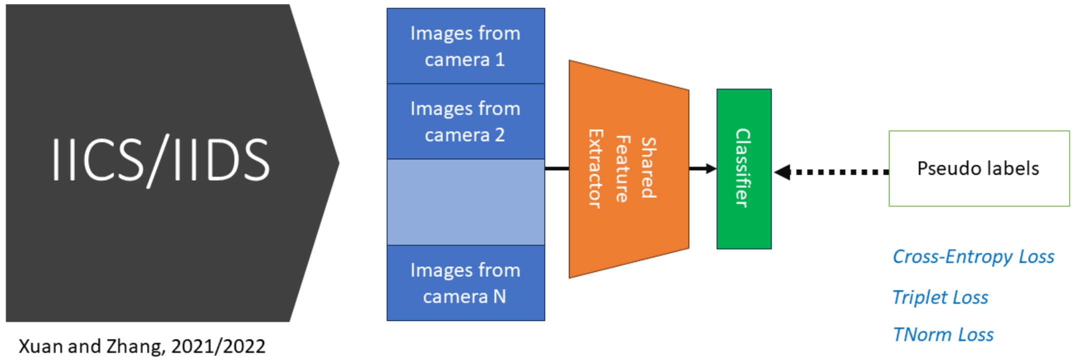


Figure 3.5: Inter-camera training stage. It depicts how generated pseudo labels are used to train the network.

generate the simulated features to increase training samples. As a result, TNorm loss is added to intra-camera and inter-camera training. In IICS/IIDS framework, ResNet50 included by AIBN and TNorm is used as a feature extractor. However, our method was built upon IICS/IIDS methods, with an investigation into employing ConvNeXt as a feature extractor. We will explain the details of ConvNeXt and its differences with ResNet architecture.

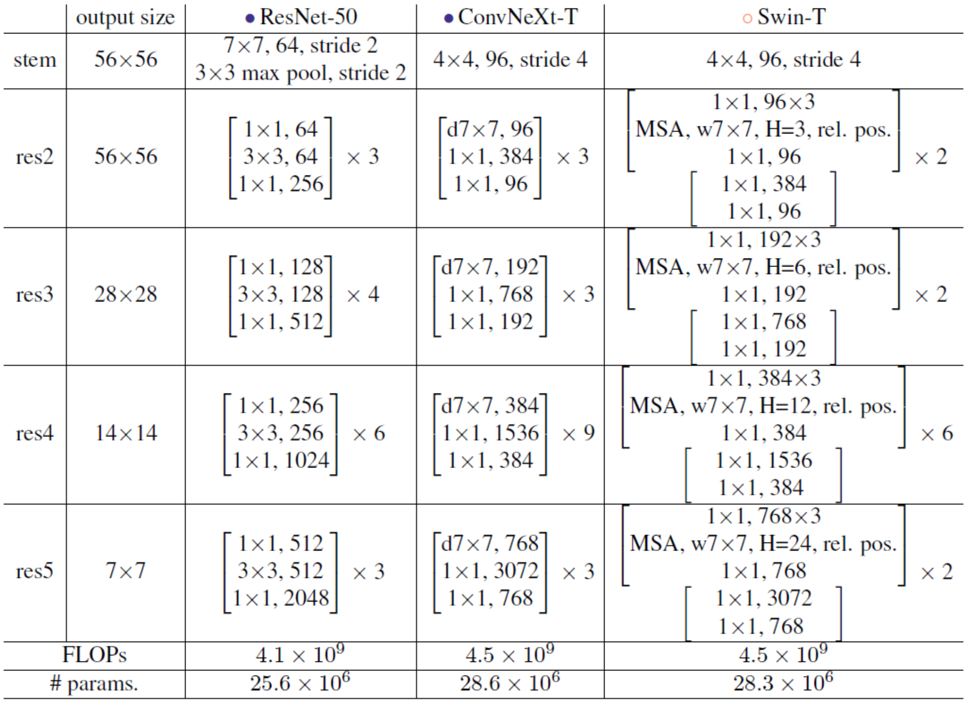
3.2 ConvNeXt

ConvNeXt is a CNN-based network that is designed by the principles of transformers (Liu et al., 2022). This network was designed based on standard ResNet to resemble the design of a vision transformer and apply 1) macro design, 2) ResNeXt, 3) inverted bottleneck, 4) large kernel size, and 5) various layer-wise micro designs. Regarding macro design concept, it should be noted that in ConvNeXt has 4 stages. The number of blocks in each stage is $[3, 3, 9, 3]$, based on the stage compute ratio in Swin transformer (Swin-T) which is $[1 : 1 : 3 : 1]$. Similar to vision transformers, the stem cell which is responsible for downsampling includes "patchify" strategy. The stem cell includes a kernel size of (4×4) with a stride of 4. Furthermore, in terms of ResNeXtify, it is said that depthwise convolution and the weighted sum operation in self-attention that exist in transformers are comparable, which results in information being mixed in the spatial dimension only. Similarly, ConvNeXt separates spatial and channel mixing through the use of depth-wise and point-wise (1×1) convolutions. Similar to transformers where each block has an inverted bottleneck, ConvNeXt includes a similar bottleneck. This means that the MLP (Multi-Layer Perceptron) block's hidden dimension is four times bigger than its input dimension in transformers and it is true in ConvNeXt as well.

In addition to these characteristics, the depth-wise kernel size in ConvNeXt is (7×7) to have a global receptive field that acts similar to one of the most distinguishing aspects of non-local self-attention in vision transformers. In ConvNeXt, depth-wise convolution layer relocate to the beginning of stage 1. This is because after downsampling, the number of channels of input is reduced before applying a large kernel size of (7×7). Fig. 3.7 shows that in the ConvNeXt block, there is only one Gelu activation function (Hendrycks & Gimpel, 2016), which is a smoother variation of ReLU. Also, similar to transformer blocks, ConvNeXt has only one normalization layer in each block which is Layer Normalization (LN) (Ba et al., 2016). In the next section, we compare the differences between ResNet50 and ConvNeXt which shows the better performance of ConvNeXt in vision tasks.

3.3 ConvNeXt (isotropic)

There are different variants of ConvNeXt, including ConvNeXt (isotropic), whose architecture has minor differences from ConvNeXt. The isotropic version of ConvNeXt lacks the notion of many stages and downsample layers. In the ConvNeXt (isotropic) model, we employ ConvNet blocks and a single stem directly, without any downsampling layers, to reduce the input image's spatial resolution by a factor of 16. Also, the dimension or number of channels remains constant throughout the model. ConvNeXt



	output size	• ResNet-50	• ConvNeXt-T	○ Swin-T
stem	56×56	7×7, 64, stride 2 3×3 max pool, stride 2	4×4, 96, stride 4	4×4, 96, stride 4
res2	56×56	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{d}7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 96\times3 \\ \text{MSA, w}7\times7, \text{H=3, rel. pos.} \\ 1\times1, 96 \\ \begin{bmatrix} 1\times1, 384 \\ 1\times1, 96 \end{bmatrix} \end{bmatrix} \times 2$
res3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{d}7\times7, 192 \\ 1\times1, 768 \\ 1\times1, 192 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 192\times3 \\ \text{MSA, w}7\times7, \text{H=6, rel. pos.} \\ 1\times1, 192 \\ \begin{bmatrix} 1\times1, 768 \\ 1\times1, 192 \end{bmatrix} \end{bmatrix} \times 2$
res4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{d}7\times7, 384 \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix} \times 9$	$\begin{bmatrix} 1\times1, 384\times3 \\ \text{MSA, w}7\times7, \text{H=12, rel. pos.} \\ 1\times1, 384 \\ \begin{bmatrix} 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix} \end{bmatrix} \times 6$
res5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{d}7\times7, 768 \\ 1\times1, 3072 \\ 1\times1, 768 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 768\times3 \\ \text{MSA, w}7\times7, \text{H=24, rel. pos.} \\ 1\times1, 768 \\ \begin{bmatrix} 1\times1, 3072 \\ 1\times1, 768 \end{bmatrix} \end{bmatrix} \times 2$
FLOPs		4.1×10^9	4.5×10^9	4.5×10^9
# params.		25.6×10^6	28.6×10^6	28.3×10^6

Figure 3.6: Resnet, ConvNeXt and Swin transformer architectures (Liu et al., 2022).

(isotropic) matches the architecture of vision transformer ViT-S/B/L with (14 × 14) feature resolution throughout. The feature dimension and depth at ConvNeXt-S (isotropic), ConvNeXt-B (isotropic), and ConvNeXt-L (isotropic) is 384, 768, and 1024 with the depth of 18, 18, and 36, respectively.

3.4 Different Variants of ConvNeXt

There are different variants of ConvNeXt. The difference between different variants is in the number of blocks in each stage and the dimension. Table. 3.1 shows different variants of ConvNeXt. For example, ConvNeXt-T has four stages and the dimension of each stage is 96, 192, 384, and 768, respectively. Also, the number of blocks in each stage is 3, 3, 9, and 3 in a row. However, ConvNeXt-S (isotropic) shows that there are 18 blocks, and the dimension is 384.

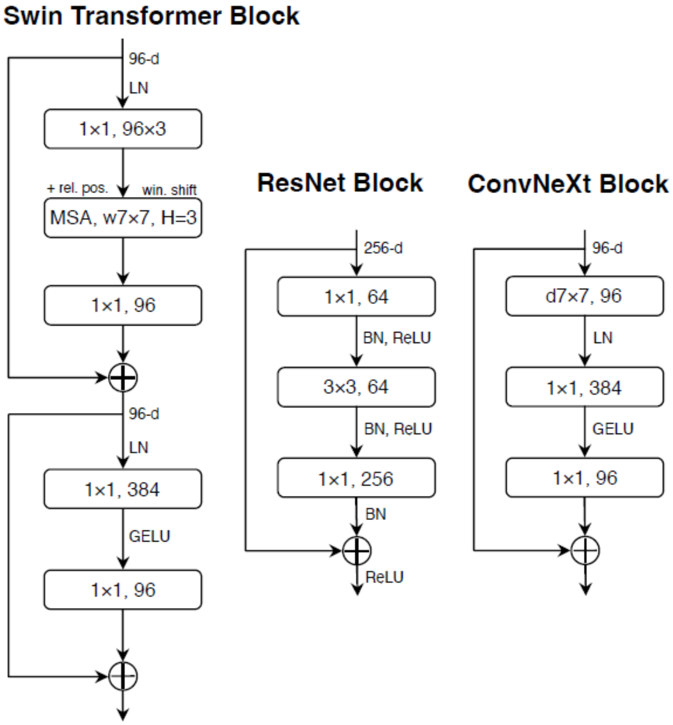


Figure 3.7: Block designs for ResNet, ConvNeXt and Swin-T. In ResNet, we have a Normalization layer after each layer. However, in ConvNeXt, just there is one normalization layer before the (1×1) convolutional layer. Also, the activation function used in ConvNeXt is Gelu, while in ResNet, it is Relu. Another difference is in the size of the kernel which is larger in ConvNeXt than it is in ResNet50. The depth-wise kernel size in ConvNeXt is (7×7) to have a global receptive field that acts similar to one of the most distinguishing aspects of non-local self-attention in vision transformers.

Table 3.1: Different variants of ConvNeXt.

Variants	**Channels**	**Blocks**	**Params**
ConvNeXt-T	(96, 192, 384, 768)	(3, 3, 9, 3)	29M
ConvNeXt-S	(96, 192, 384, 768)	(3, 3, 27, 3)	50M
ConvNeXt-B	(128, 256, 512, 1024)	(3, 3, 27, 3)	89M
ConvNeXt-L	(192, 384, 768, 1536)	(3, 3, 27, 3)	198M
ConvNeXt-XL	(256, 512, 1024, 2048)	(3, 3, 27, 3)	350M
ConvNeXt-S (isotropic)	384	18	22M
ConvNeXt-B (isotropic)	768	18	87M
ConvNeXt-L (isotropic)	1024	36	306M

3.5 Adaptive Instance and Batch Normalisation (AIBN)

The purpose of AIBN (Adaptive Instance and Batch Normalisation) is to mitigate the effects of intra-camera variations caused by different poses and people's appearances, making person Re-ID challenging. Instance-specific style characteristics may be normalized by instance normalization (IN), making the network more resistant to identity-related changes (Ulyanov et al., 2017). However, using IN alone may eliminate person identity-related clues, which might be harmful to feature discriminatory power of the network. On the other hand, Batch Normalisation (BN) speeds up convergence while keeping individual variance, protecting the discriminative ability (Ioffe & Szegedy, 2015). As a result, BN and IN both have advantages and are complementary to each other.

To learn the IN and BN combination in an adaptive manner, AIBN was suggested (Xuan & Zhang, 2021). AIBN is a technique that linearly blends the statistics, such as the mean and variance, obtained from Instance Normalization (IN) and Batch Normalization (BN). For a given feature map $x \in \mathbb{R}^{H\times W\times N}$, AIBN transforms it into $\hat{x}$, i.e.,

$$\hat{x}[i,j,n] = \alpha\frac{x[i,j,n] - (\lambda\mu_{\text{bn}} + (1-\lambda)\mu_{\text{in}})}{\sqrt{\lambda\sigma^2_{\text{bn}} + (1-\lambda)\sigma^2_{\text{in}} + \epsilon}} + \beta, \tag{3.1}$$

where $\mu_{\text{bn}}, \sigma^2_{\text{bn}}$ and $\mu_{\text{in}}, \sigma^2_{\text{in}}$ are the mean and variance calculated by BN and IN, respectively. The symbols β and α are affine parameters, with λ being the learnable mixture weight. To prevent negative values, the mixture weight for each layer is confined to the range of [0, 1] during network forward inference, and it is optimized using back-propagation. Details of IN and BN can be found in previous works (Ulyanov et al., 2017; Ioffe & Szegedy, 2015). Their parameters are computed as,

$$\mu_{\text{bn}} = \frac{\sum_n \sum_{i,j} x[i,j,n]}{N\cdot H\cdot W}, \tag{3.2}$$

$$\sigma^2_{\text{bn}} = \frac{\sum_n \sum_{i,j} (x[i,j,n] - \mu_{\text{bn}})^2}{N\cdot H\cdot W}, \tag{3.3}$$

$$\mu_{\text{in}} = \frac{\sum_{i,j} x[i,j,n]}{H \cdot W}, \tag{3.4}$$

$$\sigma^2_{\text{in}} = \frac{\sum_{i,j}(x[i,j,n] - \mu_{\text{in}})^2}{H \cdot W}, \tag{3.5}$$

where, μ_{bn} and σ^2_{bn} represent the mean and variance of BN, while μ_{in} and σ^2_{in} represent the mean and variance of IN. The symbols N, H, and W denote the dimensions of the feature map x. To integrate AIBN techniques in ConvNeXt, we replace Layer Normalization (LN) in ConvNeXt Block with AIBN in the final stages (Stage 3 and Stage 4).

3.6 Transform Normalization (TNorm)

According to style transfer studies (Dumoulin et al., 2016; Huang & Belongie, 2017), feature statistics can record and store the style information of pictures. To reduce the effect of the camera-related factors in person ReID, IIDS (Xuan & Zhang, 2022) proposed Transform Normalization (TNorm) to improve the performance of Re-ID model by reducing the effect of the camera-related factors on deep feature representations of images. In other words, TNorm normalizes the feature vector of each image based on the statistics of all images captured by the same camera, mitigating variations in the camera's setting, including lighting conditions, illumination levels, and image resolutions.

Computing the camera-related feature statistics on all training images of each camera is time-consuming. For a more efficient implementation, feature statistics within a mini-batch of each camera are computed. Given a mini-batch of features $x_c \in \mathbb{R}^{H \times W \times N}$ from camera c, the feature statistics for camera c are computed as

$$\mu_c = \frac{\sum_n \sum_{i,j} x_c[i,j,n]}{N \cdot H \cdot W}, and \tag{3.6}$$

$$\sigma_c = \sqrt{\frac{\sum_n \sum_{i,j}(x_c[i,j,n] - \mu_c)^2}{N \cdot H \cdot W}}, \tag{3.7}$$

we compute the moving average during training in order to pursue more accurate feature statistics. The resulting $\tilde{\mu}_c$ and $\tilde{\sigma}_c$ for camera c is computed with a momentum parameter η as,

$$\tilde{\mu}_c = (1-\eta)\mu_c + \eta\tilde{\mu}_c, and \tag{3.8}$$

$$\tilde{\sigma}_c = (1-\eta)\sigma_c + \eta\tilde{\sigma}_c, \tag{3.9}$$

given a feature x_c from camera c and feature statistics of camera $d; d \neq c$, TNorm first normalizes x_c, then uses feature statistics on camera d to convert the image style. This leads to another feature $\hat{x}_d$, which preserves identity-realted cues in x_c and presents the style in camera d, i.e.,

$$\hat{x}_d = \frac{\tilde{\sigma}_d(x_c - \tilde{\mu}_c)}{\tilde{\sigma}_c} + \tilde{\mu}_d. \tag{3.10}$$

TNorm can be used for data augmentation by randomly selecting the target camera d. To study the effect of TNorm in the ConvNeXt network, we study different positions of TNorm in ConvNeXt, and we show that to get better performance, the TNorm layers should be added after stage 1, stage 2, and stage 3 into ConvNeXt architecture.

Chapter 4

Method

In this chapter we describe our method for person Re-ID. We begin our discussion by providing a formal definition of the person Re-ID problem.

4.1 A Formal Definition of Person-ReID

Say we have a gallery $\mathcal{G}$ of images of P individuals recorded by a collection of cameras $\mathcal{C}$. Specifically, $\mathcal{G} = \bigcup_{c \in [1,C]} \mathcal{I}_c$, where $\mathcal{I}_c$ represents the set of images recorded by camera c. C is the short-hand for $|\mathcal{C}|$, i.e., the number of cameras. Furthermore, $\forall_{i \neq j} \mathcal{I}_i \cap \mathcal{I}_j = \phi$.

Given a query image $\mathbf{I}_q$, the goal of the person Re-ID problem is to find the "closest match" in $\mathcal{G}$. We can represent this mathematically as

$$g = \underset{g \in [1,G]}{\arg\max}\, \mathrm{sim}(\mathbf{I}_q, \mathbf{I}_g), \tag{4.1}$$

where $\mathbf{I}_g \in \mathcal{G}$ and $G = |\mathcal{G}|$.

For our purposes, finding the matching images in $\mathcal{G}$ is sufficient. In many practical scenarios gallery images are labelled with the identity of the individuals seen in these images. Consider, for example, the passport photos database. Person Re-ID as defined in Eq.4.1 is useful even when images in the gallery do not contain person identity information. In such conditions, person Re-ID allows us to track and monitor individuals over large areas and over extended time periods.

The performance of a person Re-ID system is closely tied to its ability at computing image similarity, i.e., the images of the same individual taken under various viewing conditions should be more similar to each other than the images of different individuals. This is accomplished by constructing feature extractors, which compute image features that latch on to person identities and that are robust to the usual confounding factors, such as lighting, pose, the degree of occlusion, camera characteristics, clothing, etc. Image matching is then performed in the feature space, where the features constructed from the images of the same individual sit closer to each other than the features constructed from the images of different individuals. We can re-write Eq. 4.1 as follows to capture this intuition:

$$g = \underset{g \in [1,G]}{\arg\min} \|\mathbf{x}_q - \mathbf{x}_g\|^2, \tag{4.2}$$

where $\mathbf{x}_q$ and $\mathbf{x}_g$ denote features for images $\mathbf{I}_q$ and $\mathbf{I}_g$, respectively. Specifically, $\mathbf{x}_q = \mathcal{F}(\mathbf{I}_q; \Theta_e)$ and $\mathbf{x}_g = \mathcal{F}(\mathbf{I}_g; \Theta_e)$, where $\mathcal{F} : \mathbf{I} \mapsto \mathbf{x} \in \mathbb{R}^D$ denotes the feature extractor, parameterized by Θ_e and D is the feature dimension.

Our goal then is to construct, or rather learn in the parlance of deep learning, a feature extractor that captures person identity information and is robust to confounding factors. Furthermore, we aim to learn this feature extractor in the absence of labelled data.

4.2 On Intra and Inter Camera Similarity

Let's consider images captured by a single camera for a moment. These images exhibit differences due to multiple factors, including, person identities, poses, orientation, clothing, etc. Images from multiple cameras, on the other hand, also exhibit differences due to camera-related artifacts, such as color response, placement, etc. Work by (Xuan &

Zhang, 2021; Xuan & Zhang, 2022) cogently argues that both intra and inter camera variations must be taken into account when constructing a feature extractor for the purposes of person Re-ID. They propose a general framework that neatly separates the feature extractor learning procedure into two stages, aptly named intra and inter. Furthermore, their method does not assume labelled training data, which makes it well-suited for applications in the real-world where it is often infeasible, if not outright impossible, to collect labelled datasets at the appropriate scale.

Since we follow the feature extractor learning strategy introduced in (Xuan & Zhang, 2021; Xuan & Zhang, 2022), below we formally describe the intra and inter stages and how these two are related to each other. Let's consider a training set of unlabelled images $\mathcal{T}$, which contains images recorded by C cameras. Specifically, $\mathcal{T} = \cup_{c \in [1,C]} \mathcal{I}_c$, where $\mathcal{I}_c$ is the set of images from camera c and $\forall_{i \neq j} \mathcal{I}_i \cap \mathcal{I}_j = \phi$. Recall that our goal is to *learn* a feature extractor $\mathcal{F}$ that is well-suited to the problem of person Re-ID.

4.2.1 Intra Stage

Let's begin with the intra stage. We divide this stage into two phases: 1) psuedo-label generation and 2) feature extractor refinement.

Camera-Specific Psuedo-Labelling (A)

We discuss the pseudo-label generation procedure first. We assume that we have an initial feature extractor $\mathcal{F}$. It is a common practice to use a pretrained model, say a ResNet50 trained on the Imagenet dataset, as the initial feature extractor model. Now, for each camera c, use this feature extractor to construct $\mathcal{X}_c = \{\mathbf{x} \mid \mathbf{x} = \mathcal{F}(\mathbf{I}, \Theta_e) \text{ and } \mathbf{I} \in \mathcal{I}_c\}$. Next, use agglomerative clustering with *average* linkages to partition $\mathcal{X}_c$ into $\{\mathcal{P}_c^k \mid k \in [1, K_c]\}$ sets. Here $\mathcal{X}_c = \cup_{k \in [1,C]} \mathcal{P}_c^k$ and $\mathcal{P}_c^i \cap \mathcal{P}_c^j = \phi$ for all $i \neq j$. The clustering uses pair-wise Euclidean distance $\|\mathbf{x}_l - \mathbf{x}_m\|^2$ between features as distance metric. Clustering information is used to assign pseudo-labels to images in $\mathcal{I}_c$ as follows: assign label k to image $\mathbf{I} \in \mathcal{I}_c$ if $\mathbf{x} \in \mathcal{P}_c^k$, where $\mathbf{x} = \mathcal{F}(\mathbf{I}, \Theta_e)$. It is important to remember that psuedo-labels are camera specific. This suggests that even if images from different cameras have the same pseudo-label, it does not mean that these images represent the same "identity." We will return to this issue in the inter stage.

Feature Extractor Refinement (B)

The pseudo-labels are used to fine-tune the feature extractor $\mathcal{F}$ in a supervised-learning settings as follows. First, setup a K_c-way classifier $\mathcal{K}_c$ for each camera c. Specifically, $\mathcal{K}_c : \mathbf{x} \in \mathcal{X}_c \mapsto \mathbb{R}^{K_c}$. Say $\mathcal{K}_c$ is parameterized by Θ_c then the parameters $\{\Theta_e, \Theta_1, \cdots, \Theta_c\}$ are updated using gradient-descent on the loss defined below:

$$l_{\text{intra}} = \sum_{\mathbf{I} \in \mathcal{T}} \nVdash_{\mathcal{I}_c}(\mathbf{I}) \text{ cross-entropy} \left(\hat{\mathbf{p}}, \mathbf{p} \right),$$

where p is one-hot-encoded psuedo-label for $\mathbf{I}$ and $\hat{p} = \mathcal{K}(\mathcal{F}(\mathbf{I};\ \Theta_e);\ \Theta_c)$. It follows that both $\hat{\mathbf{p}}$ and $\mathbf{p}$ are K_c-dimensional vectors when $\mathbf{I} \in \mathcal{I}_c$. $\nVdash$ denotes an indicator variable defined as follows:

$$\nVdash_{\mathbf{I}_c}(\mathbf{I}) = \begin{cases} 1 & \text{if } \mathbf{I} \in \mathbf{I}_c \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

and

$$\text{cross-entropy}(\hat{\mathbf{p}}, \mathbf{p}) = -\sum_i \mathbf{p}_i \log \hat{\mathbf{p}}_i.$$

Recall that $\hat{\mathbf{p}}$, and by extension $\mathbf{p}$, are only defined within the context of $\mathbf{I} \in \mathcal{I}_c$. The above procedure not only learns camera-specific classifiers, it also learns a shared feature extractor $\mathcal{F}$. Psuedo-labelling plus feature extractor refinement steps can be performed multiple times, in principle improving the feature extractor at each iteration.

4.2.2 Inter Stage

Next, we discuss the inter stage. The process is similar to intra stage. It also comprises two steps: 1) psuedo-label generation and 2) feature extractor refinement.

Psuedo-Label Generation (C)

As before, features $\mathcal{X}$ computed from $\mathcal{I} \in \mathcal{T}$ are clustered into K partitions $\{\mathcal{P}^i \mid i \in [1, K]\}$. Agglomerative heirarchical clustering with average linkage is used for this purpose. In addition to using pair-wise Euclidean distance (in the feature space) as distance metric, Jaccard similarity between $\mathcal{K}_1, \cdots, \mathcal{K}_C$ outputs are also used. Specifically, for a given image $\mathbf{I} \in \mathcal{T}$, each classifier c outputs a K_c-dimensional vector that represents the probability over all labels in camera c. These outputs are concatenated to construct a $(K_1 + K_2 + K_3 + \cdots + K_C)$-dimensional vector, which is normalized to construct a probability distribution that we refer to as $\mathbf{q}$. Then the pair-wise Jaccard similarity is

$$\Delta(\mathbf{I}_l, \mathbf{I}_m) = \frac{\mathbf{q}_l \cap \mathbf{q}_m}{\mathbf{q}_l \cup \mathbf{q}_m}.$$

Next, image $\mathbf{I}$ is assigned psuedo-label k if its feature $\mathbf{x} \in \mathcal{P}^k$. This process assigns psuedo-labels to images across cameras. Now it is possible for two images captured from different cameras to have same label.

Feature Extractor Refinement (D)

As before, the algorithm uses these labels to update the feature extractor as follows: construct K-way classifier $\mathcal{K} : \mathbf{x} \mapsto \mathbb{R}^K$. Say, classifier $\mathcal{K}$ is parameterized by Θ then parameters Θ_e and Θ are updated via gradient descent using the following loss:

$$l_{\text{inter}} = \sum_{\mathbf{I} \in \mathcal{T}} \text{cross-entropy}\left(\hat{\mathbf{p}}, \mathbf{p}\right),$$

where $\mathbf{p}$ is the one-hot-encoded psuedo-label assigned to $\mathbf{I}$ and $\hat{\mathbf{p}} = \mathcal{K}(\mathcal{F}(\mathbf{I};\ \Theta_e);\ \Theta)$ is the predicted label. Similar to intra stage, psuedo-label generation and feature extractor refinements steps can be performed multiple times.

4.2.3 Putting It All Together

The overall scheme followed in (Xuan & Zhang, 2021; Xuan & Zhang, 2022) can be described as a series of intra and inter stages. Specifically in their work uses the following regime:

$$[(A, B) \times 3 \text{ followed by } (C, D) \times 2] \times 40.$$

In the previous chapter 3, we show that both intra and inter stages have a role to play and that either stage on its own leads to poor performance.

4.2.4 Feature Extractor

The above scheme learns a feature extractor that accounts for image variations related to person identities while ignoring variations due to other factors, including camera characteristics and placements. The work by (Xuan & Zhang, 2021; Xuan & Zhang, 2022) proposes to use a ResNet50 feature extractor, which is pre-trained on ImageNet. Their experiments suggested that vanilla ResNet50 does not do well for the problem of person Re-ID. Consequently, they modified ResNet50 by adding normalization layers. Details about their feature extractor model are discussed in 3.

Within the realm of person Re-ID systems, the feature extractor holds a pivotal role in determining performance. Therefore, this thesis is dedicated to the exploration of ConvNext as the feature extractor within the IICS/IIDS framework. In the preceding chapter (3), we elaborated on the ConvNeXt architecture. Additionally, we conducted an analysis of various ConvNeXt iterations, differing in block count and dimensions. To delve into the impact of AIBN and TNorm, we seamlessly integrated them into ConvNeXt.

In the subsequent chapter, we will present our findings that the inclusion of AIBN blocks in ConvNeXt's third and fourth stages yields improved performance. Furthermore, our study reveals that optimal placement of TNorm is achieved after the first, second, and third stages.

Chapter 5

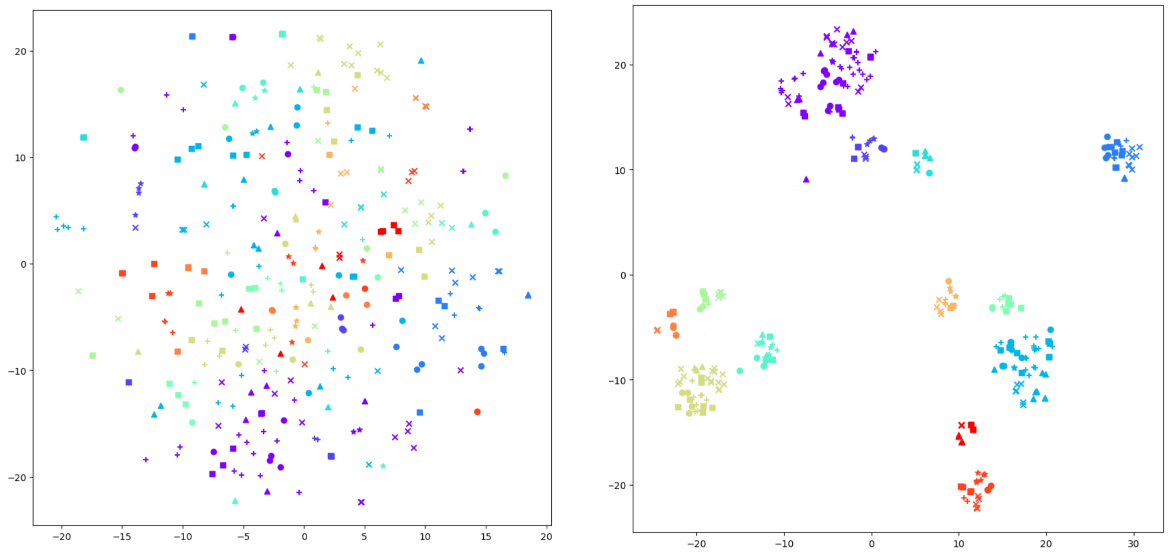
Experiments and Results

In this chapter, we outline the role of Intra-camera and Inter-camera training in person Re-ID. We assess the effect of only focusing on the Intra-camera or Inter-camera stages to show both Intra and Inter stages are complementary to get a better result in person Re-ID. The outcomes of incorporating AIBN and TNorm into ConvNeXt are then discussed. Also, we look at the role of placement of AIBN and TNorm within ConvNeXt. Finally, we demonstrate the outcome of employing different ConvNeXt variants within the IICS/IIDS framework.

5.1 The Role of Intra-camera and Inter-camera training in IICS/IIDS framework

Differences in feature distribution across various cameras pose significant challenges in person ReID tasks. When a person is captured across multiple cameras, their visual appearance can substantially change due to variations in factors such as lighting conditions, camera angles, picture quality, and camera settings. These differences are referred to as inter-camera variations. Fig. 5.1a depicts a t-SNE plot showcasing the features of various identities captured by different cameras. Different markers demonstrate different cameras. The plot illustrates a tendency for features from the same camera, shown in the same marker, to cluster together, suggesting that images taken by a particular camera exhibit a similar distribution. However, Fig. 5.1b demonstrates the effectiveness of the trained person Re-ID model in mitigating camera-related factors on feature distribution. The figure highlights that, following the training of the person Re-ID model, images of the same individual captured by different cameras shown in different markers can cluster together.

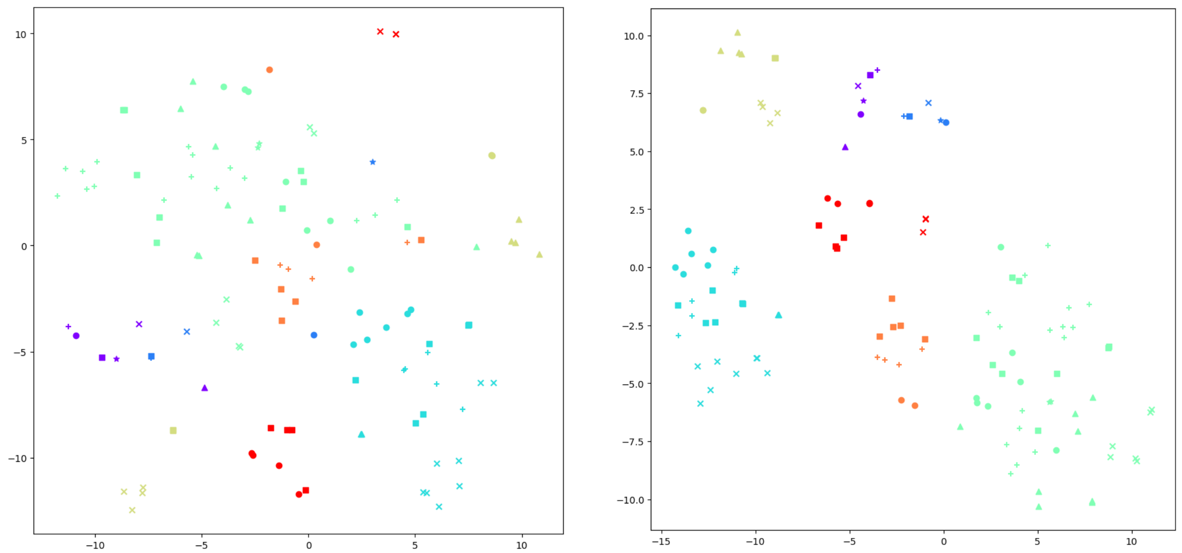
Besides inter-camera variations, intra-camera variations also play a significant role. Fig. 5.1a illustrates also how features of the same identity fail to group together due to intra-camera variations resulting from disparities in pose, appearance, and other identity-related factors. In contrast, Fig. 5.1b demonstrates that after training the Re-ID model, features of the same identity shown in the same color can successfully form cohesive groups.



(a) Before training

(b) After training

Figure 5.1: Intra-camera and Inter-camera variations.



(a) Before training

(b) After training

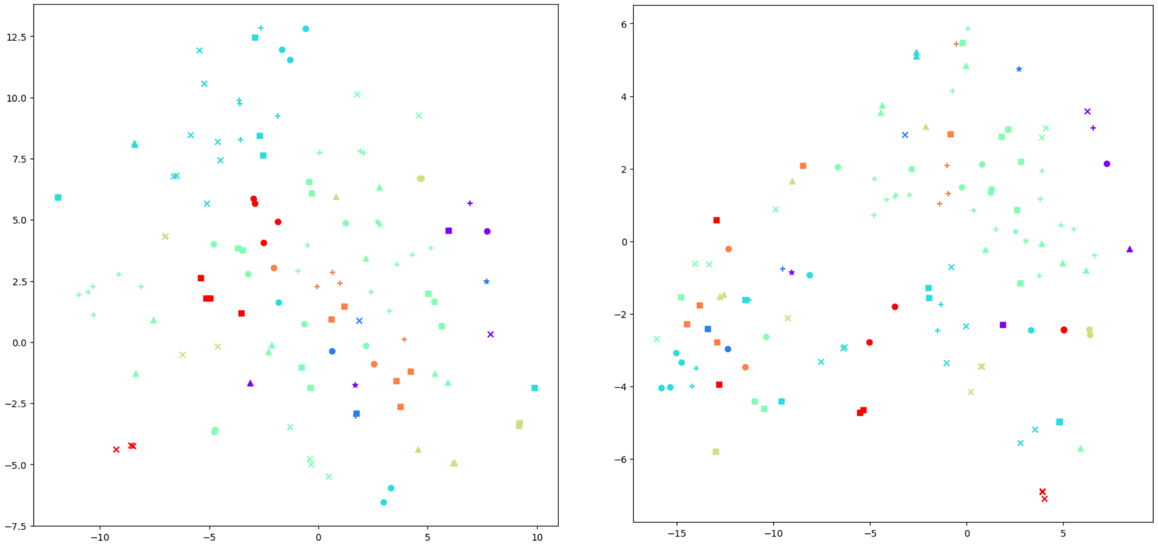
Figure 5.2: The result of intra-stage training. Different identities are shown in different colors and different markers denoted by different cameras.

5.2 Intra-camera training

We investigate the effectiveness of intra-stage. Fig. 5.2 demonstrates that intra-stage training contributes to bringing features from the same identity closer together to a certain extent, regardless of the source cameras of the images. Specifically, Fig. 5.2a reveals that features of the same identity are initially distant. Subsequently, following intra-camera training, Fig. 5.2b illustrates the moderately successful outcome of bringing the features of the same identity into closer proximity. Overall, only the Intra-camera stage is not group features of the same images together.

5.3 Inter-camera training

We examine the effect of only Inter-camera training. In this experiment, we cannot utilize Jaccard similarity due to the absence of an Intra-camera training stage. In other words, our inter-stage training solely relies on CNN feature similarity to generate pseudo-labels for network training. Fig. 5.3a visually portrays the feature distribution of images from different identities, which is scattered and widely dispersed before inter-camera training. Correspondingly, Fig. 5.3b demonstrates that inter-camera training does not effectively cluster features from the same identity in close proximity. Consequently, exclusive inter-camera training does not facilitate the aggregation of features from the same identity. As a result, relying only on Inter-camera training cannot improve the grouping features of the same images. Intra and Inter stages complement each other to perform well in person Re-ID.



(a) Before training

(b) After training

Figure 5.3: The outcome of the inter-stage training is revealed, with different identities represented by varying colors, and different markers signifying different cameras.

5.4 Implementation Details

We use a tiny variant of ConvNeXt, ConvNext-S (isotropic) (Liu et al., 2022) as the feature extractor without AIBN and TNorm. For other experiments, we used other variants of ConvNext with AIBN and with both AIBN and TNorm. We used a pre-trained model on ImageNet to initialize the training of the feature extractor with these pre-trained weights. During the training phase, the input picture is scaled to a fixed size of height 256 and width 128 pixels. Various picture augmentation techniques, such as random flipping (horizontal flipping) and random erasing, are used to improve the model's robustness and generalization. Two phases are completed successively throughout each training round: intra-camera and inter-camera.

5.5 Training Procedure and Hyperparameter Justification

Our training procedure consists of 40 clustering rounds, designed to ensure a fair comparison with prior works such as IICS and IIDS. Within each cluster round, the training process is divided into two stages: intra-network and inter-network.

The choice of hyperparameters resulted from a combination of trial and error as well as insights drawn from the methods used in IICS and IIDS methods.

5.5.1 Intra Network Training

In the intra-camera training stage, the batch size for each camera is set to 8. We utilize the Stochastic Gradient Descent (SGD) optimizer. The learning rate for ConvNeXt layers is set at 0.0005, while for the fully connected layers, it is set at 0.005. Intra-network training involves 3 epochs.

5.5.2 Inter Network Training

In the inter-camera training stage, a mini-batch in the inter stage consists of 32 images, drawn from 8 randomly selected clusters (4 images per cluster). The SGD optimization technique is employed for updating model parameters. The learning rate for ConvNeXt base layers is set to 0.001, while other layers for classification have a learning rate of 0.01. Inter-network training involves 2 epochs.

5.5.3 Loss Functions and Margins

The relative relevance of the cross-entropy and triplet loss terms is set to 1 in the inter stage. The triplet loss margin is set to 0.3, defining the minimum required distance between anchor-positive and anchor-negative pairs.

5.5.4 Jaccard Similarity Parameters

The initial value for parameters demonstrating the relative importance of Jaccard similarity in calculating similarity in the inter stage is set to 0.02.

5.5.5 Clustering Stopping Criterion

We establish a similarity threshold that dynamically determines the number of clusters. This threshold is derived by selecting the similarity value at the 0.2% quantile after arranging the sample similarities in descending order. Also, after each training epoch, it is decreased gradually by 0.001.

5.6 Dataset

We train and test our approaches on three frequently used person Re-ID datasets including DukeMTMC-ReID (Ristani et al., 2016), Market1501 (Zheng et al., 2015) and MSMST17 (Wei et al., 2018b). You can find the description of each dataset at the end of the related works Chapter 2.3.

5.7 Performance Metrics

During the training phase, we only use images and camera ids from the training set of each dataset, with no annotation information. During the inference time, the Re-ID model should extract features from a query image and gallery images, and then the similarity between these feature vectors should be measured. The result would be a ranked list showing gallery set images from the most similar image to the least one. We measure the mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) to evaluate the performance of Our method to other methods.

mean Average Precision (mAP)

The mean Average Precision (mAP) is a common metric used in person ReID to evaluate the performance across all queries. A higher mAP value indicates better overall performance. It can be formally defined as follows:

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{R_q}\sum_{k=1}^{R_q} P(q,k)\cdot \text{rel}(q,k), \tag{5.1}$$

where:

- Q represents the total number of queries.
- R_q is the number of retrieved items for query q.
- $P(q,k)$ denotes the precision at cut-off k in the list for query q.
- $\text{rel}(q,k)$ is an indicator function equal to 1 if the item at rank k is a relevant item for query q, and 0 otherwise.

Cumulative Matching Characteristics (CMC)

One often used performance metric in person ReID to assess the rank-based retrieval performance is Cumulative Matching Characteristics (CMC). The CMC curve demonstrates the probability that the correct match appears within the top-k ranks of the retrieved list. The CMC curve is defined as:

$$\text{CMC}(k) = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{1}\left(\min(\text{rank}_q) \leq k\right), \tag{5.2}$$

Table 5.1: Ablation study on intra and inter stages. Pretrain denotes directly using ImageNet pre-trained model without training. "Intra" and "Inter (w/o Jaccard)" denote intra-camera and inter-camera training stages respectively. "Intra + Inter" displays both intra-camera and inter-camera, where ConvNeXt is used without adaptation of AIBN and TNorm.

Dataset	**Market**		**Duke**	
	mAP	Rank1	mAP	Rank1
Pretrain	5.7	16.5	4.8	13.2
Intra	46.3	69.9	28.1	45.8
Inter (w/o Jaccard)	27.2	48.8	8.2	17.1
Intra + Inter	72.7	89.8	48.2	67.1

where:

- Q represents the total number of queries.
- rank_q is the rank of the correct match for query q within the retrieved list.
- $\mathbf{1}(x)$ is the indicator function, equal to 1 if the condition x holds, and 0 otherwise.

5.8 Ablation study

Impact of solely Intra-stage or Inter-stage: We examine the impact of Intra-camer and Inter-camera training. For the first experiment, we use a pre-trained ConvNeXt model without fine-tuning. Another experiment is when we rely on only Intra or Inter stage training. During the Inter-camera training, we do not consider Jaccard similarity; instead, just CNN feature similarity is calculated. Lastly, we use intra-camera and inter-camera training, with ConvNeXt as the feature extractor without AIBN or TNorm techniques. As Table 5.1 shows, the best performance is when both Intra-camera and Inter-camera training is done.

Effectiveness of AIBN and TNorm and Different inserting locations: To mitigate intra-camera and inter-camera variations in feature extraction, we incorporate two normalization techniques, AIBN and TNorm, into the ConvNeXt network. AIBN

Table 5.2: Ablation study on different locations of AIBN in ConvNeXt architecture. For instance, "AIBN (Block 2-3-4)" denotes replacing the LN in 1st to 4th stage blocks with AIBN. The best state is when we add AIBN in the final stages (stage 3 and stage 4).

Dataset	**Market**		**Duke**	
Setup	mAP	Rank1	mAP	Rank1
Baseline	72.7	89.8	48.2	67.1
All	50.3	69.9	28.1	45.8
AIBN (Block 1-2)	69.2	85.3	35.6	58.2
AIBN (Block 4)	73.2	90	51.8	71.1
AIBN (Block 2-3-4)	74.2	91.2	53.9	72.3
AIBN (Block 3-4)	**74.8**	**91.4**	**54.4**	**72.8**

Table 5.3: Ablation study on different locations of TNorm. "TNorm (stage 1-2-3-4)" shows inserted TNorm after the 1st to 4th stages in ConvNeXt architecture. The highest value is reached in the state "TNorm (stage 1-2-3)".

Dataset	**Market**		**Duke**	
Setup	mAP	Rank1	mAP	Rank1
Baseline	72.7	89.8	48.2	67.1
TNorm (Stage 1)	74.3	92.1	50.9	69.2
TNorm (Stage 1-2)	74.9	92.7	51.8	71.1
TNorm (Stage 1-2-3)	**75.4**	**92.4**	**57.8**	**76.3**
TNorm (Satge 1-2-3-4)	74.2	92	50.2	70.1

mitigates intra-camera variations caused by differences in poses, appearances, and other identity-related factors, while TNorm addresses inter-camera variations due to different camera configurations like color shifts. Our experimentation reveals that substituting all LN layers with AIBN reduces performance; however, employing AIBN instead of LN specifically in stages 3 and 4 of ConvNeXt results in an improvement. The precise impact of these adjustments at various locations within ConvNeXt can be found in Table 5.2. Furthermore, to reduce the effect of inter-camera variations on extracted features of images, we inserted the TNorm layer after each layer in ConvNext. Table 5.3 shows the highest performance is reached when we add them after stage 1-2-3 in ConvNeXt.

5.9 Comparison with State-of-the-art Methods

We compared our method to other unsupervised and transfer learning approaches on three common datasets, Market1501, Duke, and MSMT17. Tables 5.4, 5.5 and 5.6 show the comparison of the different states of our method with other existing approaches on Market1501, Duke, and MSMT17, respectively. The accuracy value of our model is an average obtained from training with three different random seeds. The last four rows of Tables present our method in different states. Iso-ConvNeXt-S is the smallest variant of ConvNeXt (isotropic), which we use as a feature extractor in the IICS/IIDS framework. Iso-ConvNeXt-S (AIBN) illustrates the scenario when ConvNeXt (isotropic) is adopted with the AIBN technique. Iso-ConvNeXt-S (AIBN, TNorm) shows the combination where ConvNeXt (isotropic) is implemented with both AIBN and TNorm. We reach the highest accuracy with ConvNeXt-B (AIBN, TNorm), which is the larger variant of ConvNeXt, used as a feature extractor with both AIBN and TNorm techniques inserted into it. We compared our method to other methods, notably GAN-based methods like PTGAN (Wei et al., 2018b), distribution alignment-based methods like TJ-AIDL (Wang et al., 2018), and pseudo-labels-based methods like MAR (Yu et al., 2019). Pseudo-labels-based techniques regularly outperformed other types of strategies. Fig. 5.5 shows the accuracy of our method is higher than IICS and IIDS methods on MSMT17 datasets. Also, Fig. 5.6 illustrates the higher performance of our method on Market15 as well. However, Fig. 5.7 illustrates the accuracy of our method is less than the accuracy of IICS and IIDS on the Duke benchmark dataset. Overall, our method has higher accuracy than the other two recent methods (Xuan & Zhang, 2021) and (Xuan & Zhang, 2022) on Market1501 and MSMT17 datasets. We report the accuracy of our model as an average obtained from training with three different random seeds, ensuring robustness in the results. The superior performance of ConvNeXt over standard ResNet, which originally was used in IICS/IIDS in person Re-ID tasks, can be attributed to the combination of depthwise and point-wise convolution. This allows a separation of spatial and chan-

Table 5.4: Comparison performance between our method to other methods on Market1501. Pseudo Label* shows that the methods are pseudo-label-based, but they initialize the weights from a person-reid pre-trained model. "NTB" stands for "Not To Be Published".

type	**Method (Reference)**	**Venue**	**Market1501**				
			Source	mAP	Rank-1	Rank-5	Rank-10
GANTs	PTGAN (Wei et al., 2018b)	CVPR18	Duke	-	38.6	-	66.1
	HHL (Zhong et al., 2018a)	ECCV18	Duke	31.4	62.2	78.8	84.0
	DG-Net++ (Zou et al., 2020b)	ECCV20	Duke	**61.7**	**82.1**	**90.2**	**92.7**
Distribution Alignment	TJ-AIDL (Wang et al., 2018)	CVPR18	Duke	26.5	58.2	74.8	81.8
	MMFA (Lin et al., 2018)	BMVC18	Duke	27.4	56.7	75.0	81.8
	CSCL (Wu et al., 2019a)	ICCV19	Duke	**35.6**	**64.7**	**80.2**	**85.6**
Pseudo Label*	MAR (Yu et al., 2019)	CVPR19	MSMT17	40.0	67.7	81.9	-
	AD-Cluster (Zhai et al., 2020a)	CVPR20	Duke	68.3	86.7	94.4	96.5
	NRMT (Zhao et al., 2020)	ECCV20	Duke	71.7	87.8	94.6	96.5
	MMT-500 (Ge et al., 2020a)	ICLR20	Duke	71.2	87.7	94.9	**96.9**
	MEB-Net* (Zhai et al., 2020b)	ECCV20	Duke	**71.9**	**87.5**	**95.2**	96.8
Pseudo Label	LOMO (Liao et al., 2015)	CVPR15	None	8.0	27.2	41.6	49.1
	BOW (Zheng et al., 2015)	ICCV15	None	14.8	35.8	52.4	60.3
	BUC (Lin et al., 2019)	AAAI19	None	29.6	61.9	73.5	78.2
	HCT (Zeng et al., 2020)	CVPR20	None	56.4	80.0	91.6	95.2
	MMCL (Wang & Zhang, 2020)	CVPR20	None	45.5	80.3	89.4	92.3
	JVTC+ (Li & Zhang, 2020)	ECCV20	None	47.5	79.5	89.2	91.9
	IICS (Xuan & Zhang, 2021)	CVPR21	None	72.1	88.8	95.3	96.9
	IIDS (Xuan & Zhang, 2022)	CVPR22	None	**78.3**	**91.2**	**96.2**	**97.7**
Our Method	Iso-ConvNeXt-S	TBD	None	72.7	89.8	95.4	97.2
	Iso-ConvNeXt-S (AIBN)	TBD	None	74.8	91.4	97.2	98.0
	Iso-ConvNeXt-S (AIBN, TNorm)	TBD	None	79.7	94.6	98.1	98.7
	ConvNeXt-B (AIBN, TNorm)	NTB	None	**83.1**	**97**	**99.2**	**99.6**

nel mixing, efficiently processing spatial information (like body structure) independent of channel information (such as appearance). This separation enables the learning of discriminative characteristics in each dimension, enhancing the overall functionality of Re-ID systems and improving their ability to differentiate between individuals.

While ConvNeXt-B with AIBN and TNorm normalizations would offer superior performance, the trade-off between performance gains and environmental costs should be carefully considered.

Table 5.5: Comparison performance between our method to other methods on DukeMTMC-ReID. Pseudo Label* shows that the methods are pseudo-label-based, but they initialize the weights from a person-reid pre-trained model. “TBD” stands for “To Be Determined”.

type	**Method (Reference)**	**Venue**	**DukeMTMC-ReID**				
			Source	mAP	Rank-1	Rank-5	Rank-10
GANs	PTGAN (Wei et al., 2018b)	CVPR18	Market	-	27.4	-	50.7
	HHL (Zhong et al., 2018a)	ECCV18	Market	27.2	46.9	61.0	66.7
	DG-Net++ (Zou et al., 2020b)	ECCV20	Market	**63.8**	**78.9**	**87.8**	**90.4**
Distribution Alignment	TJ-AIDL (Wang et al., 2018)	CVPR18	Market	23.0	44.3	59.6	65.0
	MMFA (Lin et al., 2018)	BMVC18	Market	24.7	45.3	59.8	66.3
	CSCL (Wu et al., 2019a)	ICCV19	Market	**30.5**	**51.5**	**66.7**	**71.7**
Pseudo Label*	MAR (Yu et al., 2019)	CVPR19	MSMT17	48.0	67.1	79.8	-
	AD-Cluster (Zhai et al., 2020a)	CVPR20	Market	54.1	72.6	82.5	85.5
	NRMT (Zhao et al., 2020)	ECCV20	Market	62.2	77.8	86.9	89.5
	MMT-500 (Ge et al., 2020a)	ICLR20	Market	63.1	76.8	**88.0**	**92.2**
	MEB-Net* (Zhai et al., 2020b)	ECCV20	Market	**63.5**	**77.2**	87.9	91.3
Pseudo Label	LOMO (Liao et al., 2015)	CVPR15	None	4.8	12.3	21.3	26.6
	BOW (Zheng et al., 2015)	ICCV15	None	8.3	17.1	28.8	34.9
	BUC (Lin et al., 2019)	AAAI19	None	22.1	40.4	52.5	58.2
	HCT (Zeng et al., 2020)	CVPR20	None	50.7	69.6	83.4	87.4
	MMCL (Wang & Zhang, 2020)	CVPR20	None	40.2	65.2	75.9	80.0
	JVTC+ (Li & Zhang, 2020)	ECCV20	None	50.7	74.6	82.9	85.3
	IICS (Xuan & Zhang, 2021)	CVPR21	None	59.1	76.9	86.1	89.8
	IIDS (Xuan & Zhang, 2022)	CVPR22	None	**68.7**	**82.1**	**90.8**	**93.7**
Our Method	Iso-ConvNeXt-S	TBD	None	48.2	67.1	77.3	80.6
	Iso-ConvNeXt-S (AIBN)	TBD	None	54.3	72.8	81.3	84.6
	Iso-ConvNeXt-S (AIBN, TNorm)	TBD	None	60.8	78.3	85.5	89.8
	ConvNeXt-B (AIBN, TNorm)	TBD	None	**65.2**	**80.3**	**83.4**	**87.6**

Table 5.6: Comparison of our method with other methods on MSMT17 datasets. Pseudo Label* shows that the methods are pseudo-label-based, but they initialize the weights from a person-reid pre-trained model. “TBD” stands for “To Be Determined”.

Type	**Method (Reference)**	**Venue**	**MSMT17**				
			Source	mAP	Rank-1	Rank-5	Rank-10
GANTs	PTGAN (Wei et al., 2018b)	CVPR18	Market	2.9	10.2	-	24.4
	ECN(Zhong et al., 2019)	CVPR 2019	Market	8.5	25.3	36.3	42.1
	SSG(Fu et al., 2019)	ICCV19	Market	**13.2**	**31.6**	-	**49.6**
Distribution Alignment	NRMT(Zhao et al., 2020)	ECCV20	Market	19.8	43.7	56.5	62.2
	DG-Net++(Zou et al., 2020b)	ECCV20	Market	22.1	48.4	60.9	66.1
	MMT-1500(Ge et al., 2020a)	ICLR20	Market	**22.9**	**49.2**	**63.1**	**68.8**
Pseudo Label*	PTGAN (Wei et al., 2018b)	CVPR18	Duke	3.3	11.8	-	27.4
	ECN (Zhong et al., 2019)	CVPR 2019	Duke	10.2	30.2	41.5	46.8
	SSG (Fu et al., 2019)	ICCV19	Duke	13.3	32.2	-	51.2
	NRMT (Zhao et al., 2020)	ECCV20	Duke	20.6	45.2	57.8	63.3
	DG-Net++(Zou et al., 2020b)	ECCV20	Duke	22.1	48.8	60.9	65.9
	MMT-1500 (Ge et al., 2020a)	ICLR20	Duke	**23.3**	**50.1**	**63.9**	**69.8**
Pseudo Label	MMCL (Wang & Zhang, 2020)	CVPR20	None	11.2	35.4	44.8	49.8
	JVTC+ (Zhang et al., 2021)	ECCV20	None	17.3	43.1	53.8	59.4
	SpCL (Ge et al., 2020b)	NeurIPS20	None	19.1	42.3	55.6	61.2
	IICS (Xuan & Zhang, 2021)	CVPR21	None	26.9	56.4	68.8	73.4
	IIDS (Xuan & Zhang, 2022)	CVPR22	None	**35.1**	**64.4**	**76.2**	**80.5**
Our Method	Iso-ConvNeXt-S	TBD	None	27.5	57.3	69.1	74.5
	Iso-ConvNeXt-S (AIBN)	TBD	None	29.6	60.0	72.4	77.9
	Iso-ConvNeXt-S (AIBN, TNorm)	TBD	None	36.4	65.1	77.8	82.6
	ConvNeXt-B (AIBN, TNorm)	TBD	None	**40.2**	**71.3**	**82.0**	**86.3**

Table 5.7: Comparison between different variants of ConvNeXt in IICS and IIDS frameworks.

Dataset		**Market**		**Duke**		**MSMT**	
Set up		mAP	Rank1	mAP	Rank1	mAP	Rank1
Iso-ConvNeXt-S	AIBN	74.8	91.4	54.3	72.8	29.6	60.0
	AIBN+TNorm	79.7	94.6	60.8	78.3	36.4	65.1
ConvNeXt-T	AIBN	75.2	92.3	55.1	73.5	30.2	61.4
	AIBN+TNorm	81.5	95.6	62.0	91.3	38.6	68.3
ConvNeXt-S	AIBN	75.9	93.0	55.7	74.1	32.5	62.7
	AIBN+TNorm	82.3	96.1	62.5	91.9	39.1	69.4
ConvNeXt-B	AIBN	76.5	93.8	56.5	75.3	34.0	63.5
	AIBN+TNorm	**83.1**	**97**	**65.2**	**80.3**	**40.2**	**71.3**

5.10 Different Variants of ConvNeXt

Within the IICS/IIDS framework, we explore other variants of ConvNeXt. Table 5.7 compares different ConvNeXt variants in IICS/IIDS framework. We initialize the weights of the feature extractor by ImageNet-1 K-trained models. Furthermore, Fig. 5.9 and Fig. 5.8 illustrate that larger ConvNeXt variants get better performance.

5.11 Complexity

Floating Point Operations Per Second (FLOPs) is a common unit of measurement for assessing how computationally intensive a certain model or technique is. Typically, a model with a greater FLOPs number is more sophisticated and demands a larger computational budget (Liu et al., 2022). Table 5.8 shows the complexity of ResNet50 and other variants of ConvNeXt. It shows that larger variants of ConvNeXt have higher complexity. Also, Fig. 5.4 demonstrates the same point that a larger variant of ConvNeXt exhibits higher complexity, despite achieving better accuracy. In our approach, we also employ the smallest variant of ConvNeXt (isotropic) to maintain a network complexity comparable to that of ResNet50.

5.12 Discussion

It is worth noting that the number of parameters in a neural network often directly impacts its capacity to learn and represent complex patterns. In our study, we observed that an increase in the number of parameters led to an improvement in performance metrics. The larger variant of ConvNeXt has better performance than the smaller one. This can be attributed to the enhanced ability of the model to capture intricate features and nuances within the data, ultimately resulting in better generalization. it's crucial to consider the trade-off between performance gains and computational efficiency.

In conclusion, our findings underline the significance of parameterization in model performance, emphasizing the need for a judicious choice in tailoring the neural network architecture to the specific problem domain.

Table 5.8: This table presents the complexities and parameter counts of various architectures. 'M' denotes million, and 'G' indicates gigabyte (Liu et al., 2022).

Architecture	#Param	FLOPs
ResNet-50	25.6M	4.1G
Iso-ConvNeXt-S	22M	4.3G
ConvNeXt-T	28.6M	4.5G
ConvNeXt-S	49.6M	8.7G
ConvNeXt-B	88.6M	15.4G

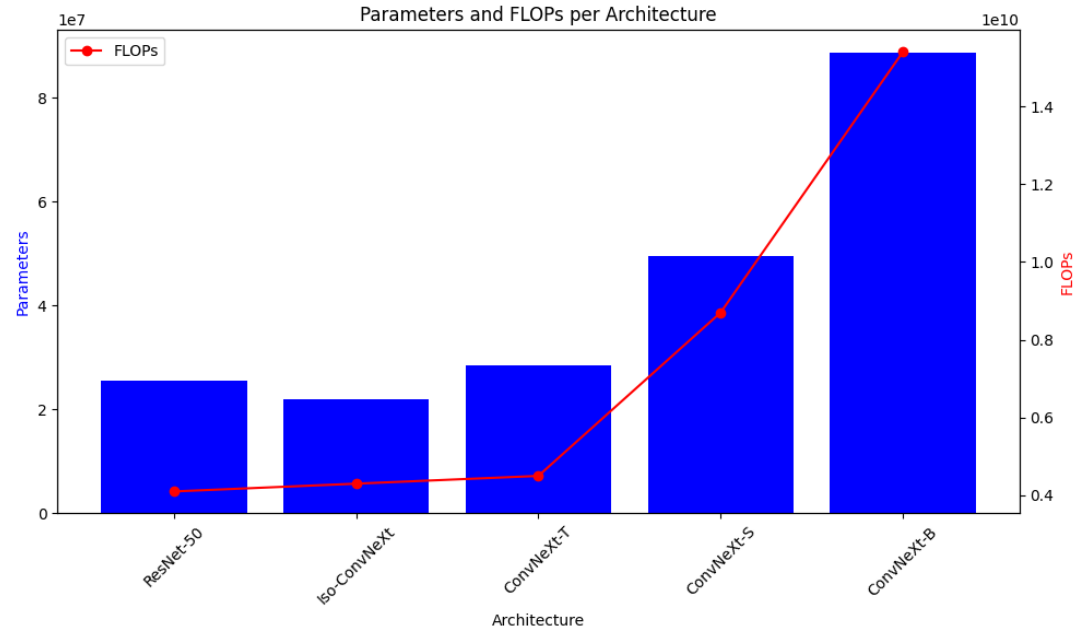


Figure 5.4: Complexity comparison among different architecture.

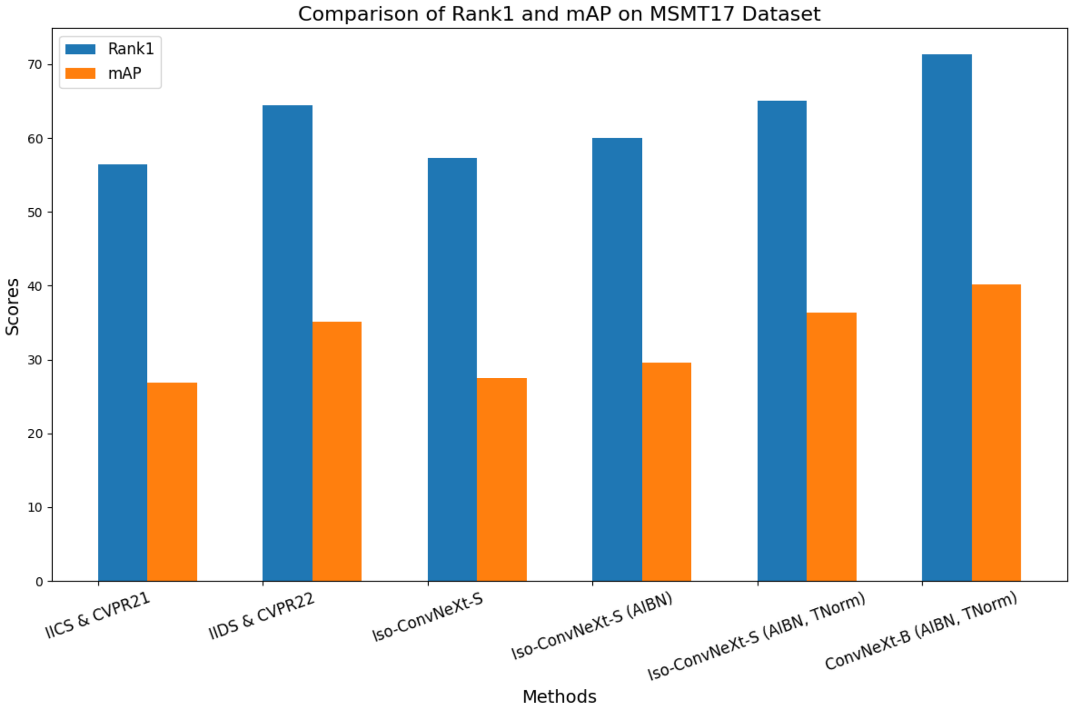


Figure 5.5: Comparison of our method to IICS and IIDS method, in which ResNet50 was originally used as a feature extractor on MSMT dataset.

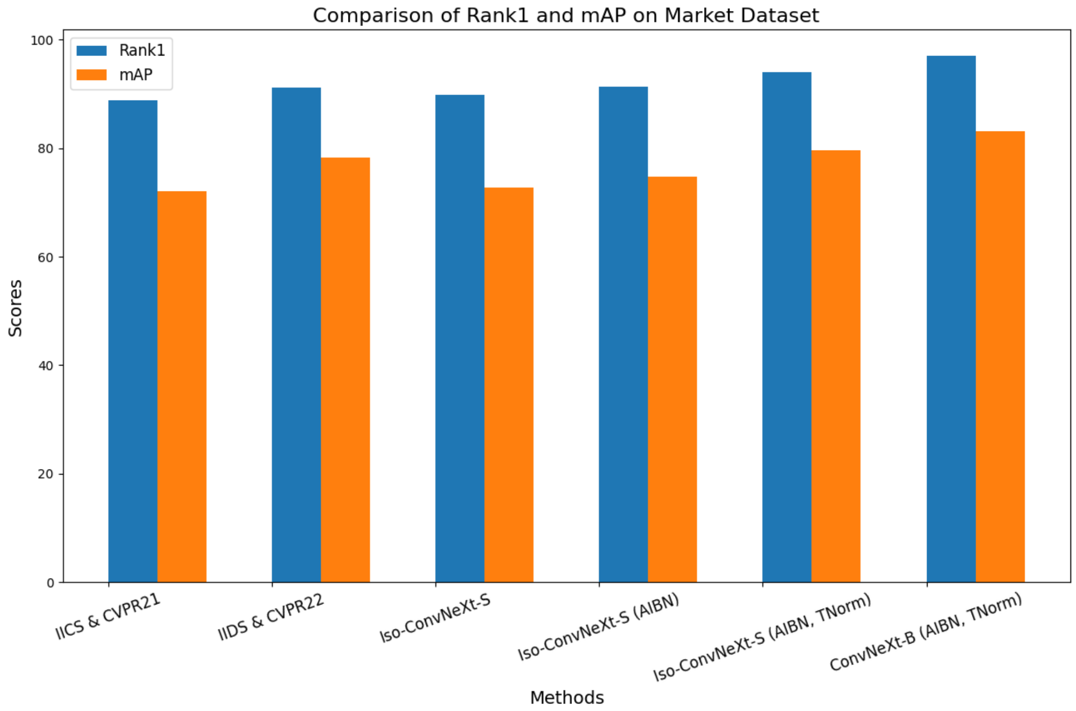


Figure 5.6: Comparison of our method with IICS and IIDS methods on Market1501 dataset.

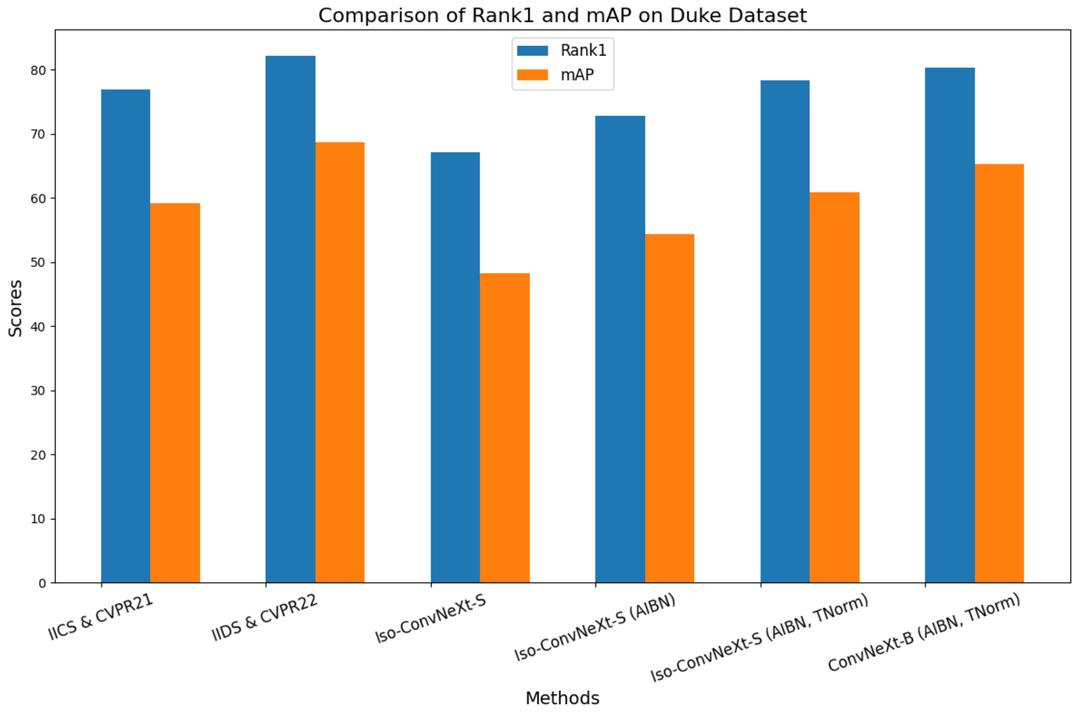


Figure 5.7: Comparison of our method with IICS and IIDS methods on Duke dataset.

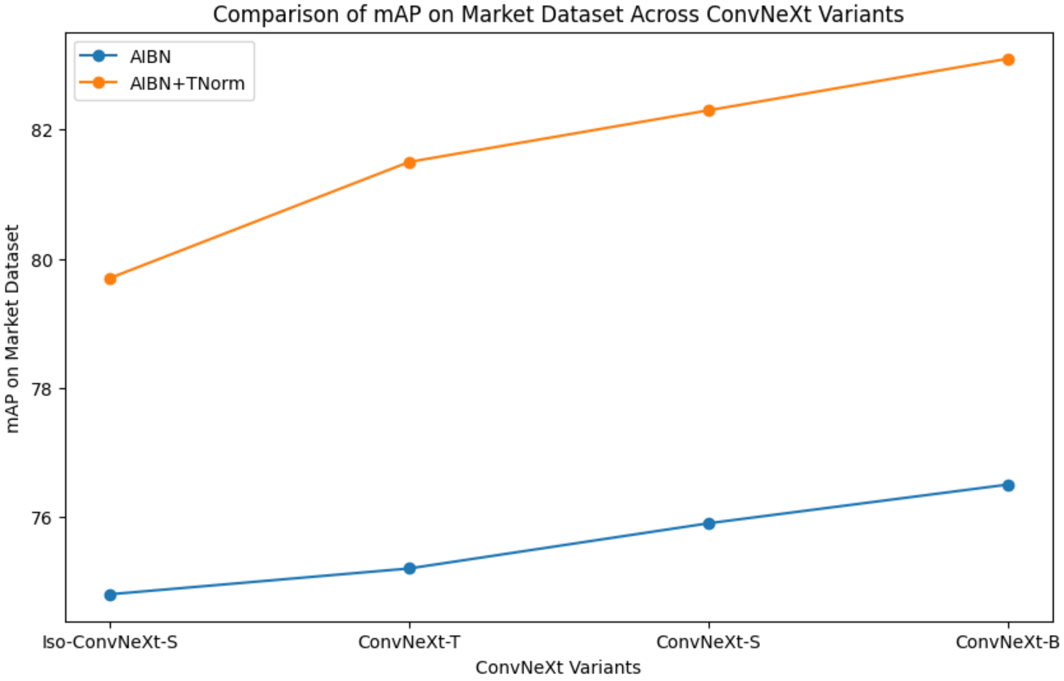


Figure 5.8: mAP comparison of different variants of ConvNeXt on Market dataset.

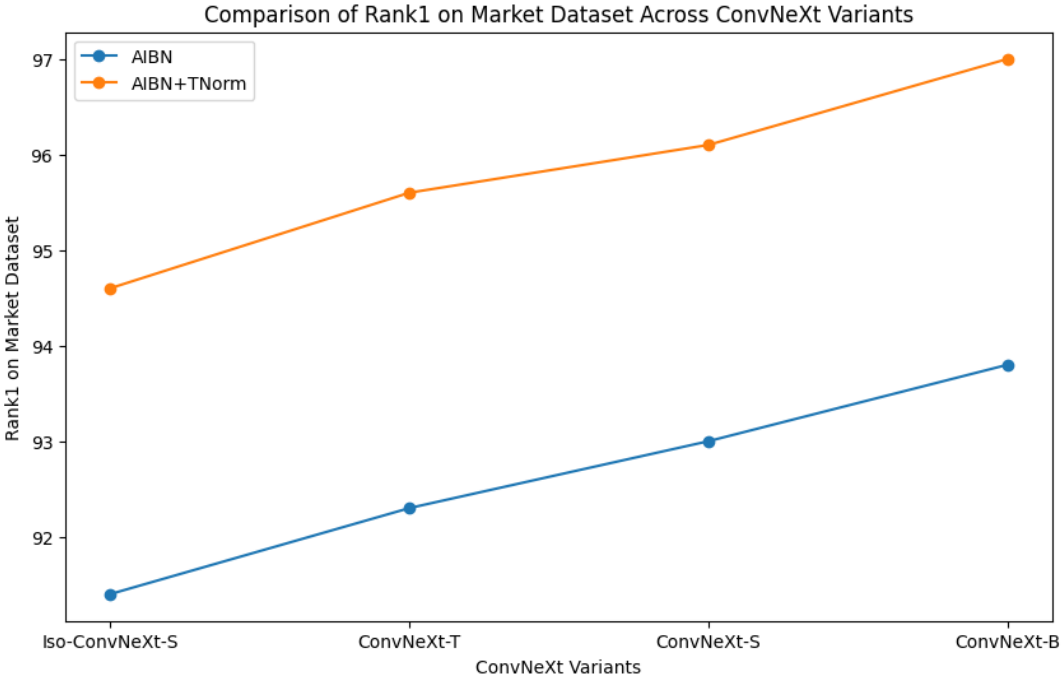


Figure 5.9: CMC comparison of different variants of ConvNeXt on Market dataset.

Chapter 6

Conclusion

In this chapter, we go over our contributions. Also, we mention the limitations of our method and the suggested ideas for future research.

6.1 Summary of Our Conclusion and Contribution in Person Re-ID

In our thesis, we have made some advancements in person Re-ID, mainly focusing on the Intra and Inter Camera Similarity (IICS) and Intra and Inter Domain Similarity (IIDS) paradigms. Our contributions are as follows:

- **Use of ConvNeXt for person Re-ID**: We extended IICS/IIDS approaches by using ConvNeXt-based feature extractor.

- **Detailed Analysis of ConvNeXt Variants**: We explored other variants of ConvNeXt within the IICS/IIDS framework, and our findings highlight that employing larger ConvNeXt models leads to better accuracy within IICS/IIDS framework.

- **Integration AIBN and TNorm into ConvNeXt**: We examined the suitability of AIBN and TNorm techniques in ConvNeXt. Our results revealed that the point of insertion within the network significantly impacts the performance of the AIBN and TNorm techniques. Optimal performance was achieved when we integrated AIBN into the block in the last two stages of ConvNeXt. As for TNorm, our analysis suggests it should be embedded after stages 1-2-3 of ConvNeXt. The adaptation of both these techniques proved effective when inserted at suitable network locations.

- **Significance of Both Intra and Inter Stages**: We studied the effects of intra and inter stages, proposed by IICS/IIDS, and our results confirm that both stages are necessary to construct feature extractors that are well-suited to the problem of person Re-ID as set up in this thesis.

- **Benchmarking Excellence**: We evaluated our method against three commonly used benchmarks: Market1501, DukeMTMC, and MSMT17. The results show that our method has better accuracy on Market1501 and MSMT17 than most unsupervised learning methods, including IICS/IIDS framework.

6.2 Limitations

6.3 Limitations

- **Scalability Concerns:** We consider the potential limitations and challenges associated with the application of our method in practical scenarios. One important consideration lies in the scalability of our method when implemented in real-time scenarios with a substantial number of cameras. The challenge arises from the significant increase in processing time, which can hinder its applicability in situations where timely responses are crucial. As the number of cameras grows, the computational demands placed on the system can lead to delays, potentially limiting the effectiveness of our approach in such high-demand settings.

- **Dataset Biases and Generalization:** Another aspect requiring attention pertains to the dataset employed for training our model. Notably, the dataset predominantly consists of photos capturing individuals in Asia. This composition introduces a potential concern about generalizing our model's performance across diverse demographics. Given the inherent variations in appearance, attire, and features among different groups of people, our model's accuracy and effectiveness might vary when extended to populations outside the dataset's original context. Ensuring robust performance across broader demographics remains a challenge.

- **Challenges with Cluttered Images:** It is also important to think about how our model performs while dealing with congested pictures. It might be challenging for the model to understand situations where numerous people are seen in a single frame. The presence of several people may add complexity that makes it difficult for the model to discriminate between and reliably identify different persons. Additionally, difficulties might occur when it's necessary to use a close crop to isolate certain elements of crowded situations. As a result, it is important to carefully consider the trade-offs between precision and computing efficiency. This may have an influence on the quality of the results acquired.

- **Occlusion Handling and Future Prospects:** It is crucial to note that our current method faces limitations in coping with occlusion. While we incorporate erasing techniques during preprocessing to simulate occlusion in images, the approach is not a comprehensive solution. Instances where individuals are partially or entirely covered, such as a person behind a car, can lead to failures in our model's identification capabilities.

- **Clothing Changes:** Additionally, it is important to recognize that our model assumes no clothing changes between images captured by different cameras. While this assumption is made for practical reasons based on the datasets available to us, it may not reflect real-world scenarios where individuals often change their clothing. The impact of clothing variations on person Re-ID is a recognized challenge, and this limitation suggests that our findings should be interpreted within the context of these assumptions. Future research may explore methods to address the impact of clothing changes on person Re-ID systems.

6.4 Future Work

To cope with the limitations of our method, we can consider the following potential research.

- **Diverse Dataset Collection:** Future work should concentrate on the acquisition of more extensive and diverse datasets. We can increase the robustness and applicability of our model by carefully selecting datasets that represent a variety of populations and environmental circumstances. The desired result is a more robust and adaptive strategy that not only works well across different racial and ethnic groupings but also flourishes in a variety of climatic environments to improve robustness and generalizability.

- **Improving Resistance to Occlusion:** As our method currently struggles with cluttered images and occlusions, future work could look into integrating techniques to better handle occlusions. This can be achieved through the integration of advanced techniques tailored to overcome occlusion-related issues. Additionally, leveraging datasets containing occluded identities can facilitate the development of a more adept model, capable of effectively navigating real-world clutter and occlusion scenarios. This enhancement aims to elevate the method's practicality and broaden its scope within contexts characterized by visual obstructions.

- **Mitigating Privacy Implications in Datasets:** Future studies must reduce privacy implications using datasets like the Duke benchmark utilized in our research. The potential sensitivity of the data raises ethical concerns as we investigate person Re-ID. To address this, we can explore methods such as data masking, synthetic data creation for producing statistically similar data without compromising privacy, and anonymization technique for obscuring personally identifiable information to protect people's privacy while retaining study effectiveness. Close collaboration with legal and ethical experts will be required to ensure compliance with legislation and preserve the highest ethical standards in data usage.

Bibliography

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. H. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. Scalable person re-identification: A benchmark. In: In *Proceedings of the ieee international conference on computer vision (iccv)*. Santiago, Chile, 2015, 1116–1124.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In: In *Computer vision–eccv 2016 workshops: Amsterdam, the netherlands, october 8-10 and 15-16, 2016, proceedings, part ii*. Springer. 2016, 17–35.

Wei, L., Zhang, S., Gao, W., & Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In: In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*. Salt Lake City, UT, USA, 2018, 79–88.

Xuan, S., & Zhang, S. Intra-inter camera similarity for unsupervised person re-identification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2021, 11926–11935.

Xuan, S., & Zhang, S. (2022). Intra-inter domain similarity for unsupervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Almasawa, M. O., Elrefaei, L. A., & Moria, K. (2019). A survey on deep learning-based person re-identification systems. *IEEE Access*, *7*, 175228–175247.

Tahboub, K. (2017). *Person re-identification and intelligent crowdsourcing with applications in public safety* (Doctoral dissertation). Purdue University.

Liu, W., Chang, X., Chen, L., & Yang, Y. Semi-supervised bayesian attribute learning for person re-identification. In: In *Proceedings of the aaai conference on artificial intelligence*. 2018.

Cheng, D., Chang, X., Liu, L., Hauptmann, A. G., Gong, Y., & Zheng, N. Discriminative dictionary learning with ranking metric embedded for person re-identification. In: In *Proceedings of the 26th international joint conference on artificial intelligence*. Melbourne, Australia, 2017, 964–970.

Liu, W., Chang, X., Chen, L., & Yang, Y. Early active learning with pairwise constraint for person re-identification. In: In *Machine learning and knowledge discovery in databases: European conference, ecml pkdd 2017, skopje, macedonia, september 18–22, 2017, proceedings, part i 10*. Springer. 2017, 103–118.

Li, W., Zhao, R., Xiao, T., & Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In: In *Cvpr, columbus, ohio, usa*. 2014.

Liu, C., Chang, X., & Shen, Y.-D. Unity style transfer for person re-identification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, 6887–6896.

Zheng, Z., Zheng, L., & Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: In *Proceedings of the ieee international conference on computer vision (iccv)*. Venice, Italy, 2017, 3754–3762.

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., & Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: In *Proceedings* *of the european conference on computer vision (eccv)*. Munich, Germany, 2018, 480–496.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., & Gu, J. (2019). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, *22*(10), 2597–2609.

Liu, X., Zhang, S., & Yang, M. Self-guided hash coding for large-scale person re-identification. In: In *2019 ieee conference on multimedia information processing and retrieval (mipr)*. IEEE. 2019, 246–251.

Wei, L., Liu, X., Li, J., & Zhang, S. Vp-reid: Vehicle and person re-identification system. In: In *Proceedings of the 2018 acm on international conference on multimedia retrieval*. 2018, 501–504.

Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., & Chen, X. Clothes-changing person re-identification with rgb modality only. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*. IEEE/CVF. New Orleans, LA, USA, 2022, 1060–1069.

Somers, V., De Vleeschouwer, C., & Alahi, A. Body part-based representation learning for occluded person re-identification. In: In *Proceedings of the 2023 ieee/cvf winter conference on applications of computer vision workshops (wacvw)*. IEEE/CVF. Waikoloa, HI, USA, 2023, 1613–1623.

Tan, L., Dai, P., Ji, R., & Wu, Y. Dynamic prototype mask for occluded person re-identification. In: In *Proceedings of the 30th acm international conference on multimedia*. ACM. Lisbon, Portugal, 2022, 531–540.

Sun, B., & Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, *abs/1607.01719*. http://arxiv.org/abs/1607.01719

Wu, A., Zheng, W.-S., & Lai, J. Unsupervised person re-identification by camera-aware similarity consistency learning. In: In *2019 ieee/cvf international conference on* *computer vision (iccv)*. 2019, 6921–6930. https://doi.org/10.1109/ICCV.2019.00702.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. http://arxiv.org/abs/1512.03385

Lin, S., Li, H., Li, C.-T., & Kot, A. C. (2018). Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*.

Lu, Y., Wang, M., & Deng, W. Augmented geometric distillation for data-free incremental person reid. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*. IEEE/CVF. New Orleans, LA, USA, 2022, 7329–7338.

Zhong, Z., Zheng, L., Zheng, Z., Li, S., & Yang, Y. Camera style adaptation for person re-identification. In: In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*. Salt Lake City, UT, USA, 2018, 5157–5166.

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, *abs/1703.10593*. http://arxiv.org/abs/1703.10593

Zou, Y., Yang, X., Yu, Z., Kumar, B. V. K. V., & Kautz, J. (2020a). Joint disentangling and adaptation for cross-domain person re-identification. *CoRR*, *abs/2007.10315*. https://arxiv.org/abs/2007.10315

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. Uniter: Universal image-text representation learning. In: In *European conference on computer vision*. Springer. Glasgow, UK, 2020, 104–120.

Zhang, X., Cao, J., Shen, C., & You, M. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: In *Proceedings of the ieee/cvf international conference on computer vision (iccv)*. Seoul, South Korea, 2019, 8222–8231.

Wang, D., & Zhang, S. Unsupervised person re-identification via multi-label classification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, 10981–10990.

Fan, H., Zheng, L., Yan, C., & Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(4), 1–18.

Lin, Y., Dong, X., Zheng, L., Yan, Y., & Yang, Y. A bottom-up clustering approach to unsupervised person re-identification. In: In *Proceedings of the aaai conference on artificial intelligence*. *33*. (01). 2019, 8738–8745.

Zhao, F., Liao, S., Xie, G.-S., Zhao, J., Zhang, K., & Shao, L. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In: In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xi 16*. Springer. 2020, 526–544.

Ge, Y., Chen, D., & Li, H. (2020a). Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Zhai, Y., Ye, Q., Lu, S., Jia, M., Ji, R., & Tian, Y. Multiple expert brainstorming for domain adaptive person re-identification. In: In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part vii*. Springer. 2020, 594–611.

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. Deep mutual learning. In: In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*. Salt Lake City, UT, USA, 2018, 4320–4328.

Zhu, K., Guo, H., Yan, T., Zhu, Y., Wang, J., & Tang, M. Pass: Part-aware self-supervised pre-training for person re-identification. In: In *European conference on computer vision*. Springer. Tel Aviv, Israel, 2022, 198–214.

Jin, X., Lan, C., Zeng, W., & Chen, Z. (2020). Global distance-distributions separation for unsupervised person re-identification. *CoRR*, *abs/2006.00752*. https://arxiv.org/abs/2006.00752

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, *32*(9), 1627–1645.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. A convnet for the 2020s. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2022, 11976–11986.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. Going deeper with convolutions. In: In *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015, 1–9.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*. Honolulu, HI, USA, 2017, 6924–6932.

Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: In *International conference on machine learning*. PMLR. Lille, France, 2015, 448–456.

Dumoulin, V., Shlens, J., & Kudlur, M. (2016). A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

Huang, X., & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In: In *Proceedings of the ieee international conference on computer vision*. Venice, Italy, 2017, 1501–1510.

Wang, J., Zhu, X., Gong, S., & Li, W. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*. Salt Lake City, UT, USA, 2018, 2275–2284.

Yu, H.-X., Zheng, W.-S., Wu, A., Guo, X., Gong, S., & Lai, J.-H. Unsupervised person re-identification by soft multilabel learning. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2019, 2148–2157.

Zhong, Z., Zheng, L., Li, S., & Yang, Y. Generalizing a person retrieval model hetero-and homogeneously. In: In *Proceedings of the european conference on computer vision (eccv)*. Munich, Germany, 2018, 172–188.

Zou, Y., Yang, X., Yu, Z., Kumar, B. V., & Kautz, J. Joint disentangling and adaptation for cross-domain person re-identification. In: In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part ii 16*. Springer. 2020, 87–104.

Wu, A., Zheng, W.-S., & Lai, J.-H. Unsupervised person re-identification by camera-aware similarity consistency learning. In: In *Proceedings of the ieee/cvf international conference on computer vision*. 2019, 6922–6931.

Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., & Tian, Y. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, 9021–9030.

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. Person re-identification by local maximal occurrence representation and metric learning. In: In *Proceedings of the ieee conference on* *computer vision and pattern recognition (cvpr)*. Boston, MA, USA, 2015, 2197–2206.

Zeng, K., Ning, M., Wang, Y., & Guo, Y. Hierarchical clustering with hard-batch triplet loss for person re-identification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*. Virtual Event, 2020, 13657–13665.

Li, J., & Zhang, S. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In: In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xxiv 16*. Springer. 2020, 483–499.

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*. Long Beach, CA, USA, 2019, 598–607.

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., & Huang, T. S. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: In *Proceedings of the ieee/cvf international conference on computer vision*. Seoul, South Korea, 2019, 6112–6121.

Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., & Tian, Q. Unrealperson: An adaptive pipeline towards costless person re-identification. In: In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*. Nashville, TN, USA, 2021, 11506–11515.

Ge, Y., Zhu, F., Chen, D., Zhao, R., et al. (2020b). Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in neural information processing systems*, *33*, 11309–11321.