Cross-domain Person Re-identification

by

Raj H. Shah
202011028

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

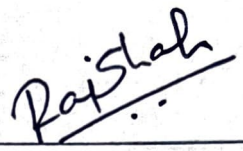
July, 2022

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.



Raj Shah

Certificate

This is to certify that the thesis work entitled Cross-domain Person Re-Identifcation has been carried out by Raj Shah for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

Dr. Shruti Bhilare
Thesis Supervisor

Dr. Avik Hati
Thesis Co-Supervisor

Acknowledgments

I want to take this opportunity and sincerely thank my supervisor, Dr. Shruti Bhilare, and co-supervisor, Dr. Avik Hati, who helped me throughout the thesis with their in-depth knowledge of the domain and constant input in better solving the problem at hand. I am also indebted not just for their guidance but also for the interest they helped me build towards the thesis. I thank them for being only one call away, even during the tough times the world was going through during the course of this work.

It would also be apt to thank Dhirubhai Ambani Institute of Information and Communication Technology for providing all the necessary resources and hardware to implement our work and promote a healthy learning culture.

I also want to thank my parents and relatives, who constantly boosted my morale and supported me in every way possible. Lastly, I would also like to thank my peers and friends who helped me have a pleasant experience here at DAIICT but also helped me academically and professionally throughout our time together.

Contents

Abstract

The problem of person re-identification has been getting much attention in the computer vision community. The task is to recognize pictures of the same individuals in images with different backgrounds taken from multiple cameras. It involves complexities such as different people with similar outfits or the same person with different outfits, differential illuminations, low-resolution images, inaccurate bounding boxes, and occlusion. The increasing progress is due to the increased demand for automated surveillance systems. The problem is often formulated as a retrieval task. When given a query image and a gallery set of images taken from different cameras, possibly at different locations, the system aims to find the pictures with the same person.

In this thesis, we have used ResNet50 and Omni-scale Network (OSNet) for feature extraction and different loss functions such as softmax loss, triplet loss, quadruplet loss, KL-divergence, etc. to train and infer on models for cross-domain person re-identification (re-id). We observe that using a multi-source training strategy boosts the performance of such cross-domain re-id systems. We also show that using re-ranking significantly improves the performance of both same-domain and cross-domain person re-id.

List of Principal Symbols and Acronyms

BIN: Batch-Instance normalization

BN: Batch normalization

CMC: Cumulative Matching Characteristics

CNN: Convolutional Neural Network

GAN: Generative Adversarial Network

IN: Instance normalization

mAP: Mean Average Precison

MSFA: Multi-scale Focusing Attention

OSNet: Omni-scale Network

ResNet: Residual Network

SENet Squeeze-and-Excitation Network

List of Tables

List of Figures

CHAPTER 1
Introduction

1.1 What is Person Re-identification?

Person re-identification (re-id) is a computer vision task that has been formulated and studied widely as a retrieval problem that aims to build a discriminative, identity-preserving person descriptor for a person present in a query image for performing retrieval from a diverse gallery set coming from different cameras with varying scenes and lighting conditions. With increasing impetus given to public safety and with an increasing number of surveillance cameras, person re-id is imperative in intelligent surveillance systems. Typical person re-identification systems have the following flow:

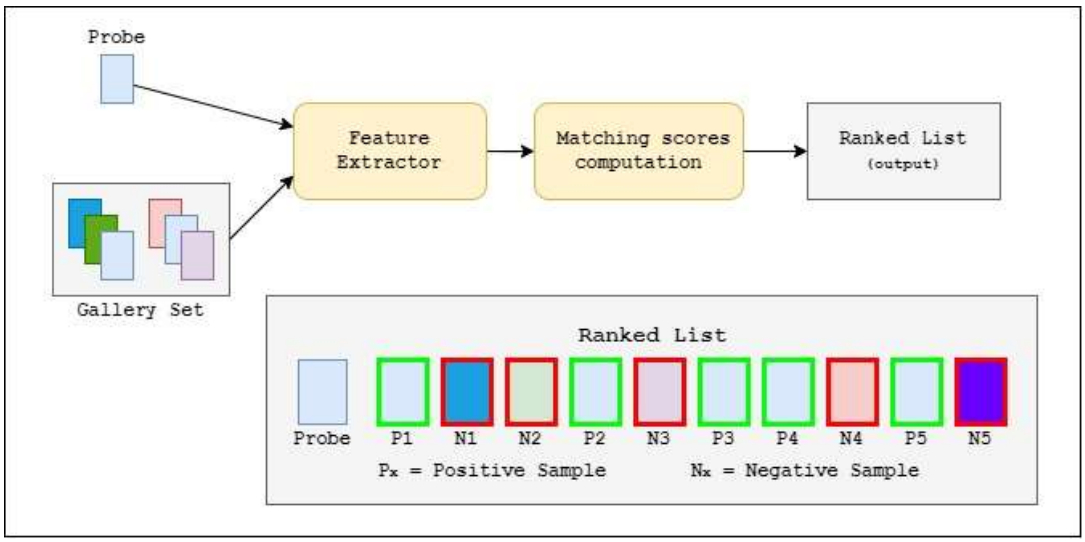


Figure 1.1: Generic person re-id system working pipeline. Green borders represent true positives (correct retrievals) and red borders represent false positives (incorrect retrievals).

1.2 Open-world vs. Closed-world Systems

There are two approaches to person re-identification: *open-world* person re-id and *closed-world* [13] person re-id. For training and testing, closed re-id approaches assume a controlled environment. The data representation is consistent across the dataset, the bounding boxes are as accurate as possible, there are enough training images with correct annotations, and the query identities are present in the gallery image set. On the other hand, open-world person re-id approaches have heterogeneous, noisy, unlabeled, and sometimes even insufficient data, as well as an open gallery set. All approaches discussed in this work are based on closed-world person re-id [6], [20], [18].

1.3 Challenges in Person Re-identification

Person re-identification (re-id) and facial recognition [4] are both used to identify people, but they are fundamentally different. Most facial recognition datasets have well-annotated images with a similar or higher resolution than those from person-id datasets. Low resolution is frequently caused by pre-processing on raw images that are required to make them suitable for training a person re-id model, such as person detection, bounding box generation, and cropping. The presence of different viewpoints, varying low-image resolutions, illumination changes, unconstrained poses, occlusions, heterogeneous modalities, complex camera environments, background clutter, unreliable bounding box generation, and other factors, as illustrated in Figure 1.2, pose challenges in developing a person re-id system.

Another point to note in comparing these two tasks is the number of images in the most popular training datasets. For instance, MSCeleb-1M [4] used for facial recognition contains about 10M images of 1M identities, whereas the largest person re-id dataset - MSMT17, contains 1,26,411 images of 4,101 identities. A few datasets have been used for training and testing a person re-identification system. The more prominent ones include the Market1501, MSMT17, DukeMTMC-Reid, CUHK03, etc. Each of these datasets is from cameras placed in different surrounding settings. For instance, Market1501 has around 32,668 images of 1,501 identities taken using six cameras in a market. Typically, a person appearing in this dataset wears casual clothing such as t-shirts, hoodies, and shorts. In the case of the images in MSMT17, the clothing style is primarily smart casuals. Since persons appearing in different domains (datasets) have a different appearance, fea-

(a) Inaccurate detection (b) Pose misalignment (c) Occlusion

Figure 1.2: Challenges in person re-identification

ture learning and representation for cross-domain training and testing becomes challenging, as illustrated in Figure 1.3.

1.4 Motivation

Public places such as markets, shopping malls, and parking lots usually have a network of cameras with a non-overlapping field of view for video surveillance. Identifying individuals in such a setting can have various applications in security, such as access control, person tracking, etc. As mentioned in Section 1.3, training on a particular domain and testing on another has complexities regarding the different scene settings and clothing style differences. The systems at the site of deployment of a typical re-id system may not always be powerful enough for fine-tuning on new data. Hence, the system needs to be generalizable to different domains for testing and predictions for any practical application of person re-id. To this end, it is vital to explore cross-domain person re-id approaches to generalize over different domains without significantly hampering the system's performance. Application of such a person re-id system can be in the field of video surveillance in areas of interest to ensure no unwanted identities are present at the scene or to track a person of interest to maintain an eye on them for safety

(a) Variation due to camera views

(b) Visual similarity among different IDs

Figure 1.3: Reason for need of generalization in person re-id systems

purposes.

1.5 Problem Statement

Object detection algorithms are used to recognise objects or humans in a scene. However, identifying the same person in two different images taken from different cameras and with different illumination and background settings is a complex task due to the problems mentioned in Section 1.3. Also, the system should perform reliably even in diverse environments from the one in which it was trained. Cross-domain person re-id, thus, is a requirement for any practical application of a person re-id system. This work attempts to perform person re-identification on image datasets. It can be viewed as a retrieval task wherein the system finds the top matching images to a query image. Mathematically, finding the images with minimum distances to the query image can be viewed as a task. The following, when performed k times, gives the top-k retrievals.

$$y = \underset{y_i \in G}{\arg\min} \quad D(q, y_i) \tag{1.1}$$

where y_i is the representation of the i^{th} image in the gallery set G and q is that of the query. D is the chosen distance metric (e.g. Euclidean distance).

1.6 Contributions

We have explored a few different approaches for the task of person re-id in this work, and they are as follows:

- The use of omni-scale network-based methods represents the images as a linear combination of features of different 'scales'. By modifying the OSNet architecture, we have incorporated an attention mechanism as a SENet block after each 3×3 lite convolution layer.
- We have also explored a Generative Adversarial Network (GAN)-based method that performs image generative and identity discriminative tasks in the same network by using structure and appearance encoders of input images.
- Triplet loss-based methods are also used, wherein we also experiment with cosine distance instead of Euclidean distance while calculating the loss.
- We also point out the advantages of multi-source training and post-processing after retrieval (re-ranking) to help boost the system's results.

The rest of this work is arranged as follows: Chapter 2 contains the literature review and discussion of the existing methods deployed for the task of person re-id. Chapter 3 discusses the methods we adopted, followed by Chapter 4, which describes the experiments conducted along with their results. Chapter 5 consists of the conclusion and the future scope of this work.

CHAPTER 2
Literature Survey

In this section, we discuss the literature review in the field of same-domain and cross-domain person re-id. Omni-scale (OSNet)-based methods have been adopted for both same-domain and cross-domain person re-id. OSNet is a computationally lightweight network due to the use of depthwise separable convolutions. Studies also used ResNet50 - a significantly bigger and computationally expensive network for feature extraction. Attention mechanism was used to improve feature extraction in some studies, as mentioned in the upcoming sections. A few other approaches involving aligning two images for distance calculation during literature and ones based on using a Generative Adversarial Network were also studied.

2.1 OSNet-based Feature Extraction

Omni Scale Network or OSNet [20], aims at learning discriminative features that are a weighted combination of multiple scale features. It captures homogeneous features that consist of features of only one scale and heterogeneous features that are composed of multiple scale features. A homogeneous feature contains features of the same scale. For instance, a feature learnt with a bigger receptive field (e.g., a person in white tee and denim). On the other hand, heterogeneous features can consist of a medium scale feature (e.g., a white t-shirt) and a small scale feature (e.g., a logo on the t-shirt). A semantic understanding of combinations of different scale features can help the model identify the imposter. These 'weights' are dynamically calculated for each image using a unified aggregation gate depending on its composition.

Batch normalization (BN) is used in the 3×3 lite to normalise the samples based on parameters of the entire mini-batch to learn the discriminative features among the examples in the mini-batch. Zhou et al. [21] tweak the OSNet by incorporating instance normalization (IN) to potentially eliminate the instance-specific

cues and thus helping generalisation of the model to the differential background and illumination scenarios. However, determining the optimal position of IN in the OSNet architecture is not a straightforward task. To this end, the authors have used neural architecture search to determine the best possible network configuration involving IN. It is worth noting that this tweak boosted both the same-domain and cross-domain scores (mAP and R_1 accuracy) over the original OSNet architecture.

2.2 Attention-based Approaches

Liu et al. [10] propose model adaptation and model generalization by introducing an attention mechanism in the backbone network to learn the discriminative features for every domain. The attention results are incorporated in the output using skip connections to improve features of high and medium levels of semantic information. Prior information about other domains is not required as it is learnt using an attention mechanism.

The architecture consists of three components - the backbone network, the attention module, and the skip connection, as shown in figure 3. ResNet-50 is the backbone network in this work, consisting of four stages. Attention modules are incorporated after each of these stages for experimentation. The authors explored the use of spatial attention, channel attention, and hybrid attention - a combination of both. Spatial attention captures the position of different body parts and partial body features. High-level features of every channel correspond to different body parts as a response for each part is different in different channels. Channel attention aims to capture this information to learn features better and adaptatively. Two types of attention - long-range dependency-based and direct generation-based were studied, and incorporating them was explored in various permutations and combinations.

Wang et al. [12] contribute a fully attentional block that can plug into any CNN for performing better on misaligned images. ResNet-50 [5] is used as the backbone network. A fully attentional block (FAB) based on the SENet is used for the attention mechanism. The only change compared to the SE block is that instead of an FC layer at the end, the authors use a 1×1 convolution to keep the size of the feature map the same as the input to the attentional block. This is why the FAB can be used with any backbone. The attention features are added element-wise to the original feature map, and attention loss is calculated.

OSNet [20] is used as a backbone instead of ResNet50 in [8]. The number of

convolutional and max pool layers was minimized, and the architecture consisted of multi-scale focusing attention (MSFA) blocks. These blocks were arranged in stacks of different quantities with a different number of channels in each stack. Each of these MSFA blocks comprises fully attentional (FA) blocks. Using these fully-attentional blocks consisting of a 3×3 lite layer (using depthwise separable convolution, as in OSNet) followed by a Squeeze-and-Excitation (SE) block, multi-scale feature attention is captured and used for cross-domain person re-id.

Alongside spatial attention that has been explored in various computer vision tasks, channel attention has also garnered research interest. Squeeze-and-Excitation Net (SENet) [7] is one such network that tries to capture channel attention. Each channel contains some feature information. Modern deep networks consist of a large number of channels, each of which may not contain considerable or valuable feature information. SENet captures the channel attention by first 'squeezing' (or reducing) the number of channels by a reduction ratio r (generally, the value of r is set to 16), followed by 'excitation' (or expansion) to get back the original number of channels after going through ReLU activation.

Residual Network (ResNet) [5] architecture-based networks have been used extensively for feature extraction of images. ResNet-50, ResNet-101 and ResNet-152 are used for feature extraction in person re-id systems. The number in these names refers to the number of convolution layers. Residual networks have helped the deep network perform feature extraction and improved the performance of most computer vision tasks.

2.3 Triplet Loss as Loss Function

Triplet loss is a loss function that involves a triplet including an anchor, a positive sample, and a negative sample. This loss function aims to pull the anchor and the positive sample close to each other while pushing the negative sample away from the anchor. For triplet loss to be effective, the sampling strategy for the positive and negative samples is critical. Usually, authors go with hard sample mining. Concretely, the hardest positive and the hardest negative samples are chosen for each anchor sample from each mini-batch. The hardest positive sample refers to the sample of the same class as the anchor but with the least similarity. The hardest negative sample is from any other class that resembles the anchor the most. Other sampling methods have also been explored, such as cumulative sampling, wherein the model is fed with easier samples initially, and the difficult ones are progressively fed. Another work [14] explores triplet non-local loss wherein the

positive sample in the triplet is also considered as the anchor to push the negative sample away from the positive sample along with the anchor.

2.4 Domain Generalization using Meta-Learning

Batch normalization (BN) aims at generalising the person re-id model to get good results on cross-domain test sets. Instance normalization (IN) , on the other hand, helps in distinguishing samples of different classes. Hence, over-style normalization where the IN removes even the identity discriminative information or under-style normalization due to the BN trained models failing to perform on cross-domain datasets may occur. The key idea is to generalize Batch-Instance Normalization (BIN) layers by simulating the cases that are likely to occur in a cross-domain retrieval setting in the meta-learning pipeline. By overcoming the harsh differences caused due to different domain settings, the model prevents overfitting. The parameter $\rho \in [0, 1]$ defines the weight given to IN and BN. Cho et al. [1] separate the feature learning from the meta-training. Meta training deals with the simulation of the inter-domain variations that the model has to address. All the parameters but the weight parameter ρ are updated (i.e., θ_f - feature extractor and ψ - classifier), and the balancing parameters θ_ρ are updated in the meta-training phase.

$$y = \rho(\gamma_B \cdot \hat{x}_B + \beta_B) + (1 - \rho)(\gamma_I \cdot \hat{x}_I + \beta_I) \tag{2.1}$$

where γ, $\beta \in R^C$ are affine transformation parameters with C being the number of channels of x , $\hat{x}$ is the normalized response, and $\rho \in [0, 1]$ is a learnable parameter that decides the weightage given to batch normalization and instance normalization.

In the feature learning phase, the cross-entropy loss is used for ID-discriminative learning and triplet loss for similarity learning. Label smoothening is also applied since many labels exist in a multi-source domain dataset. The mini-batch of data from source domains is divided into meta-train and meta-test sets in the meta-training phase. The balancing parameter (ρ) biases towards IN or BN to simulate the domain shift due to the other. Two losses are used to counter over-style normalization - scatter loss and shuffle loss. Scatter loss enhances the intra-domain diversity, and shuffle loss pulls the sample from the same class and pushes the inter-class samples away. Triplet loss is used to enhance intra-class compactness irrespective of the style difference. The total meta-training loss is the sum of these three. θ_p are the parameters updated with β. After the meta test phase, θ_p is

updated.

2.5 Aligned Re-id

Aligned Re-ID [15] involves learning global and local features. Global features include the structure and pose information, while local features can be understood as the position of body parts in the pictures. Global features are calculated by performing pooling over all the channels of the feature map extracted from a network such as ResNet50 (say CxHxW), thus giving C dimensional vector. Local features are obtained by pooling over the H direction and then taking a 1x1 convolution to reduce it to c-dimension. Hence, the local features are $H \times c$-dimensional. Distance between global features is calculated using the L_2 distance formula, and the local distances are learnt as shown in Figure 2.1. Concretely, to compare two images, the aim is to find at what position each body part is compared to in the other image. In Figure 2.1, the first stripe in the left image is mapped to the fourth stripe, the second to the fifth and sixth stripe, and so on. After mapping horizontal stripes, local feature distances and total distances are then calculated. Training is done with metric learning using the TriHard loss function (i.e., mining the hardest positive and negative sample for each anchor).

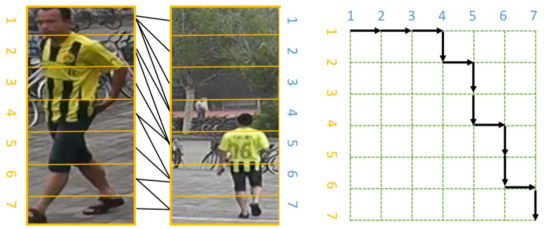


Figure 2.1: Local distance computation between features

Mutual learning is used for training the model. A large teacher network learns the features of images, and knowledge distillation is used to transfer them to smaller student networks (with parameters θ_1 and θ_2) that are trained simultaneously. Each of these networks uses classification loss and metric learning loss. The corresponding losses are combined to form classification mutual loss and metric mutual loss that consider only the global features for both networks. In the inference stage, using only global features learnt via mutual learning can give good results.

CHAPTER 3
Methodology

In this chapter, we will discuss our methodologies in this work. We use Omni-scale feature learning, Generative Adversarial Network (GAN)-based joint discriminative and generative learning, using ResNet50 as feature extractor with triplet loss and quadruplet loss function. We highlight the benefits of using multi-source training for generalizing a cross-domain person re-id model. We also use re-ranking in order to boost retrieval scores.

3.1 Using OSNet Backbone with Softmax Loss

In this approach, we make use of the Omni-scale network (OSNet) architecture for the task of feature extraction. OSNet computes features for images by dynamically fusing multi-scale features. Specifically, it fuses a small-scale feature with a smaller receptive field (e.g., a logo on a t-shirt) and medium to large-scale features (e.g., the outfits' color and the person's structure in the image, respectively). Such feature detection helps to differentiate an imposter during the retrieval task, as shown in Figure 3.1 by detecting the logo on the white t-shirt.

Figure 3.1: Detection of imposter using omni-scale features

OSNet is computationally efficient as it only consists of around 2.25M parameters compared to approximately 25M parameters in the case of ResNet50. This

is because it uses depthwise separable convolutions [2] where pointwise convolution is followed by depthwise convolution, which reduces the number of computations and the number of parameters by a factor of c(no. of channels in the input). The smaller number of parameters makes the model less prone to overfitting. Due to its compactness, these backbone components of OSNet are termed as 3×3 lite layers (shown in Figure 3.2) as the receptive field is of size 3×3. In general, when t such blocks are stacked together, the receptive field becomes $2t + 1$ $\times$ $2t + 1$ ($t = [2, 3, 4]$). Each block calculates $\tilde{x}$ given x where x is passed through t 3×3 lite layers as given in Equation (3.1).

$$\tilde{x} = \sum_{t=1}^{T} F^t(x), \quad s.t. \quad T \geq 1 \tag{3.1}$$

where F is the feature extraction function and t ($t = 1, 2, ..., T$) is an added dimension which gives the number of scales the learnt feature is composed of. We set $T = 4$.

Each stack of 3×3 lite layers outputs homogeneous scale features that are dynamically fused to learn the best discriminative features for the input image. A shared and unified aggregation gate is used to calculate the output of the aggregation gate. It dynamically learns the importance of features captured at each scale, and the weights for each are chosen dynamically. Learning the fusion in such a way helps ensure that the scales of learnt features are not general but different for each input image. The shared aggregation gate also has the desired property that all streams are combined to guide the learning of G.

$$\tilde{x} = \sum_{t=1}^{T} G(x^t) \odot x^t, \quad s.t. \quad T \geq 1 \tag{3.2}$$

where G is the feature extractor and $G(x^t)$ is a vector with length spanning the entire channel dimension of x^t and $\odot$ is the Hadamard product.

Batch normalization (BN) is used in the 3×3 lite to normalise the samples based on parameters of the entire mini-batch to learn the discriminative features among the examples in the mini-batch. We tweak the OSNet by incorporating instance normalization (IN), similar to [21], to potentially eliminate the instance-specific cues and thus help generalisation of the model to the differential background and illumination scenarios. However, determining the optimal position of IN in the OSNet architecture is not a straightforward task. To this end, the authors have used neural architecture search to determine the best possible network configuration involving IN. It is worth noting that this tweak boosted both the

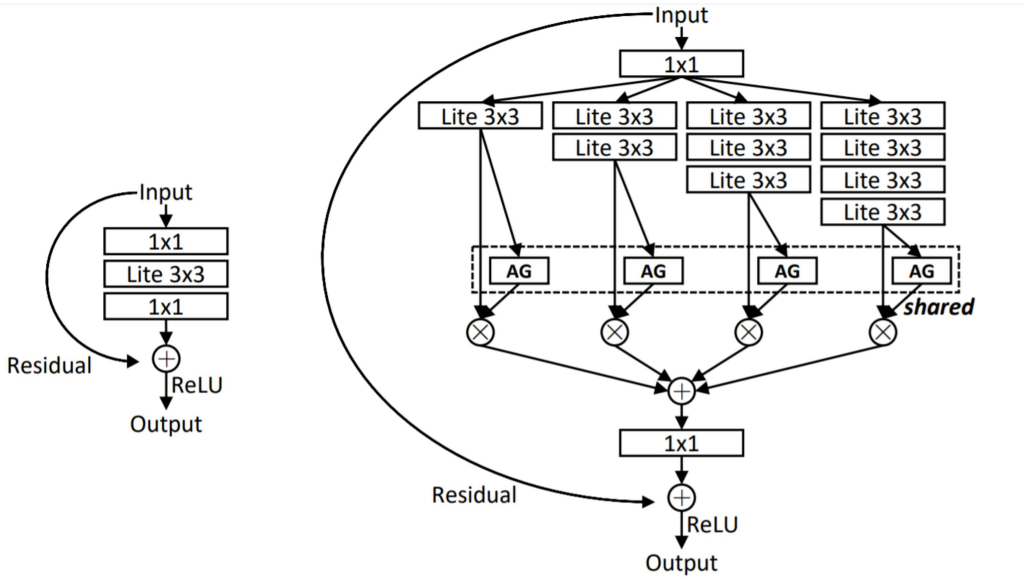


Figure 3.2: Bottleneck network (described on the left) alongside the shared aggregation gate (AG).

same-domain and cross-domain scores (mAP and R_1 accuracy) over the original OSNet architecture.

We have used Squeeze-and-Excitation Network (SENet) as an attention mechanism plugged into the OSNet architecture as another experiment. By fusing the attentions in this way, [8] built their fully-attentional (FA) block. Different from their work, we use the entire OSNet architecture as opposed to only the first convolution and pooling layer in their work. The results of all the changes in the OSNet and the experiments performed are presented in Section 4.1.

3.2 Using a GAN-based Method

We take motivation from [18] and consequently use a network that can generate images and perform identity discrimination in the same framework. It uses a specialized Generative Adversarial Network (GAN) that uses the structure code consisting of the pose and the appearance code (a_i) consisting of outfit features, accessories, etc., for each image. Using these two codes, the GAN generates N^2 images from N images that could consist of images that consist of the structure (s_i) and appearance (a_j) of the same identity where $i = j$ or cross-combination of two separate identities such that where $i \neq j$ as shown in Figure 3.3. We can reconstruct the images from these codes and vice versa. We consider E_a and E_s as the appearance and structure encoders respectively and G as the decoder. An encoder can output the respective code when provided with an input. Inversely, the

decoder outputs an image using the structure and appearance codes. The image and code reconstruction losses are given as follows in Equation (3.3), (3.4), (3.5). $code_1$ and $code_2$ stand for appearance code and structure code in Equation (3.4) and (3.5), respectively and x_i, s_i and a_i represent the i^{th} image, its structure code and appearance code, respectively. In Equation (3.3), we use two different images of the same identity to ensure that the generator should be able to reconstruct image x_i from x_j.

$$L_{recon}^{img} = \sum_{x_i, x_j \in I, i \neq j} [||x_i - G(a_j, s_i)||_1] \tag{3.3}$$

where the classes (identities) y_i and y_j of both images x_i and x_j are the same.

$$L_{recon}^{code_1} = \sum_{i,j \in I, i \neq j} [||a_i - E_a(G(a_i, s_j))||_1] \tag{3.4}$$

$$L_{recon}^{code_2} = \sum_{i,j \in I, i \neq j} [||s_j - E_s(G(a_i, s_j))||_1] \tag{3.5}$$

where I in Equation (3.3), (3.4) and (3.5) denotes the set of all images, x_i represents the i^{th} image, s_i and a_i represents the structure and appearance code for the i^{th} image.

For self-identity generated images that are generated using two different images belonging to the same class, and for cross-identity generation, the identity discriminative loss is given as L_{id}^{s} and L_{id}^{c} in Equation (3.6) and (3.7), respectively. Here $p(y_i|x_i)$ is the probability of image x_i belonging to class y_i. x_i^j denotes an image composed of the structure code of the i^{th} image and the appearance code of the j^{th} image.

$$L_{id}^{s} = \sum_{i \in I} -log(p(y_i|x_i)) \tag{3.6}$$

where log is the natural logarithm and the same notation has been used throughout the thesis.

$$L_{id}^{c} = \sum_{i,j \in I, i \neq j} -log(p(y_i|x_i^j)) \tag{3.7}$$

The generated images are then used for primary and fine-grained feature learning. The feature learning and identity discrimination are done using a teacher-student model wherein the teacher predicts "dynamic soft labels" based on the structure and appearance composition of the images. The student model aims to minimise the KL-divergence between the teacher labels' probability distribu-

tion and the learning task's discriminative model. Unlike their work that reports results for same-domain person re-id, we experiment by eliminating the fine-grained feature learning. Loss functions for primary feature learning and fine-grained feature learning are mentioned as Equation (3.8) and (3.9), respectively.

$$L_{prim} = -\sum_{k=1}^{K} q(k|x_i^j) log(\frac{p(k|x_i^j)}{q(k|x_i^j)}) \tag{3.8}$$

$$L_{fine} = \sum_{i,j \in I, i \neq j} -log(p(y_j|x_j^i)) \tag{3.9}$$

where K is the number of identities and each k denotes an identity (classes).

The adversarial loss (L_{adv}) ensures matching distributions between generated and real data. It is given as follows:

$$L_{adv} = \sum_{i,j \in I, i \neq j} logD(x_i) + log(1 - D(G(a_i, s_j))) \tag{3.10}$$

The overall loss function, using all these losses, is given as follows in Equation (3.11). The scores are reported in detail in Section 4.3.

$$\begin{aligned} L_{total} = \lambda_{recon}^{img} L_{recon}^{img} + \lambda_{recon}^{code} L_{recon}^{code} + \lambda_{id}^{s} L_{id}^{s} + \lambda_{id}^{c} L_{id}^{c} \\ + \lambda_{adv} L_{adv} + \lambda_{prim} L_{prim} + \lambda_{fine} L_{fine} \end{aligned} \tag{3.11}$$

where $\lambda_{recon}^{img}, \lambda_{recon}^{code}, \lambda_{id}^{s}, \lambda_{id}^{c}, \lambda_{adv}, \lambda_{prim}, \lambda_{fine}$ are hyperparameters to control weights of the related loss terms, L_{recon}^{img}, L_{recon}^{code}, L_{id}^{s}, L_{id}^{c}, L_{adv}, L_{prim} and L_{fine} are the image reconstruction loss, code reconstruction loss, self-identity generation loss and cross-identity generation loss and adversarial loss, the primary and fine-grained feature learning losses, respectively as specified above.

The feature learning for the discriminative part of the joint network performs learning smartly to utilize better how the generative module generates the images. Concretely, the learning is performed in two parts - primary feature learning and fine-grained feature learning.

- Primary feature learning: Primary features focus on structure-invariant clothing information. It uses a teacher-student model where the teacher is simply a baseline CNN that has information about the identities of the constituent images that contribute to the structure and appearance codes, while the student model deals with each image without this information. The teacher model gives "soft-labels" i.e., a probability distribution for each class. The

Figure 3.3: Image generation using structure and appearance codes

student model, too, predicts the probability distribution. Finally, KL-divergence is calculated between these two distributions.

- Fine-grained feature learning: Fine-grained features focus on appearance-invariant structure information. Since, in the generation phase, we generate images that simulate the same person wearing different clothing, the model is forced to learn fine-grained id-related attributes such as hair, hat, bag, body size, etc.

Experiments have been performed by considering and disregarding the fine-grained features, and the results are discussed in Section 4.2.2.

3.3 Using ResNet Backbone with Triplet Loss

As the name suggests, triplet loss involves a triplet including an anchor, a positive sample, and a negative sample. This loss function aims to pull the anchor and the positive sample close to each other while pushing the negative sample away from the anchor (illustrated in Figure 3.4). The equation for triplet loss is given as Equation (3.12). For triplet loss to be effective, the sampling strategy for the positive and negative samples is critical. Usually, authors go with hard sample mining. Concretely, the hardest positive and the hardest negative samples are chosen for each anchor sample from each mini-batch. The hardest positive sample refers to the sample of the same class as the anchor but with the least similarity.

The hardest negative sample is from any other class that resembles the anchor the most.

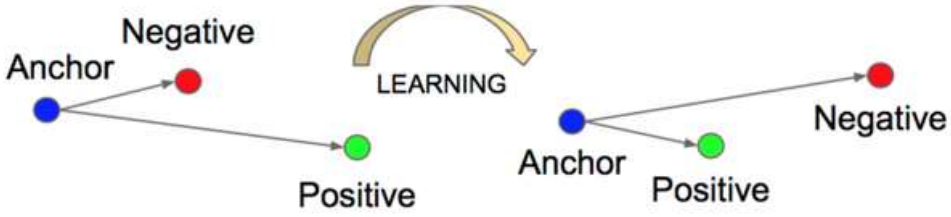


Figure 3.4: Learning using triplet loss. Positive sampled pulled closed to the anchor and negative sample pushed away after training.

$$L_{tr}(a,p,n) = \sum_{a,p,n} \max[D(a,p) - D(a,n) + \alpha, 0], \quad s.t. \quad y_a = y_p \neq y_n \tag{3.12}$$

where $D(x_1, x_2)$ stands for for the distance measure used to calculate the distance between x_1 and x_2, y_x is the class to which x belongs, α is the margin and a, p and n stand for the anchor, positive and negative sample, respectively.

Our work uses the ResNet-50 backbone for feature extraction and triplet loss as the loss function. In Natural Language Processing tasks, it has been found that using the cosine distance between the representations of different documents in the vector space gives good results for the retrieval task. With this idea, we have experimented with the distance metric used in the triplet loss formula given in Equation (3.12). We replaced the Euclidean distance with cosine distance. We experiment using multi-source training as it alleviates the challenges posed by lesser training data and provides generalization to the model as the model is not confined to learning domain-specific cues like single-source training. Post-processing the results after the retrieval are complete can help in enhancing the results. Subsequently, we use re-ranking to enhance the results of our model.

3.4 Using ResNet Backbone with Quadruplet Loss

Quadruplet is similar in concept to the previously discussed triplet loss. Unlike in triplet loss, we consider four samples in quadruplet loss - an anchor, a positive sample belonging to the same class as the anchor, and two negative samples, each belonging to different classes compared to each other, the anchor and positive

sample. Equation (3.13) shows the formula to calculate quadruplet loss. The first term in this equation is the same as in triplet loss, and the second term considers all the four samples in the quadruplet.

$$
\begin{aligned}
L_{quad}(a,p,n_1,n_2) = &\sum_{a,p,n_1} \max[D(a,p) - D(a,n_1) + \alpha_1, 0] \\
&+ \sum_{a,p,n_1,n_2} \max[D(a,p) - D(n_1,n_2) + \alpha_2, 0] \\
\text{s.t.} \quad & y_a = y_p, y_a \neq y_{n_1}, y_a \neq y_{n_2}, y_{n_1} \neq y_{n_2}.
\end{aligned}
\tag{3.13}
$$

where $D(x_1, x_2)$ stands for the distance measure used to calculate the distance between x_1 and x_2, y_x is the class to which x belongs, α_1 and α_2 are the margins and a, p, n_1 and n_2 stand for the anchor, positive and the two negative samples, respectively.

The first term is equivalent to Equation (3.12). It focuses on the relative distances between positive and negative pairs of probe images. The second term is a new constraint that takes into account the sequences of positive and negative pairs with different probe images. This constraint requires that the minimum inter-class distance be greater than the maximum intra-class distance, regardless of whether the pairs contain the same probe.

As previously stated, the first term aims to obtain the correct orders in training data using the same probe. The second term is useful from the standpoint of orders with different probe images. It can broaden the inter-class variations and improve performance on testing data. Though it is a useful auxiliary term, it should not be used to lead the training phase and is not as important as the first term. As a result, in Equation (3.13), we treat the two terms differently. Instead of using weights, we use margin thresholds to determine the balance of two terms in our loss. We require that the margin between pairs with the same probe be large enough to keep the main constraint in place.

3.4.1 Re-ranking

Re-ranking [19] is a low-cost post-processing step used to boost the performance of a person re-id model. The initial ranked list $\mathcal{L}(q, \mathcal{G}) = \{g_1^0, g_2^0, \ldots g_N^0\}$, where q is the probe image, $\mathcal{G}$ is the gallery set, is obtained using the sorting the pairwise distances between the probe image and the gallery images in ascending or-

der. The k-nearest neighbours $N(q,k)$ of the probe image q, i.e., the top-k nearest neighbours of the probe image in the retrieval can be given as:

$$N(q,k) = \left\{g_1^0, g_2^0, \ldots, g_k^0\right\}, |N(q,k)| = k \tag{3.14}$$

where $|.|$ denotes the number of candidates in the set.

Two images are k-reciprocal nearest neighbours of each other if they are both present in the k-nearest neighbour set of the other, as can be seen in Figure 3.5. This provides a much more stringent policy for ranking retrieved images and thus improves the retrieval scores. In Figure 3.5, green borders represent the true positives. Notice that the query image (blue border) is present in all of their ranking lists. The k-reciprocal nearest neighbours $\mathcal{R}(q,k)$ are represented as:

$$\mathcal{R}(q,k) = \{g_i \mid (g_i \in N(q,k)) \wedge (q \in N(g_i,k))\} \tag{3.15}$$

Next, the $\frac{k}{2}$-reciprocal nearest neighbours are incrementally added to the more robust set $\mathcal{R}^*(q,k)$ as:

$$\begin{aligned} &\mathcal{R}^*(q,k) \leftarrow \mathcal{R}(q,k) \cup \mathcal{R}\left(z, \tfrac{1}{2}k\right) \\ &\text{s.t. } \left|\mathcal{R}(q,k) \cap \mathcal{R}\left(z, \tfrac{1}{2}k\right)\right| \geq \tfrac{2}{3}\left|\mathcal{R}\left(z, \tfrac{1}{2}k\right)\right| \ \forall z \in \mathcal{R}(q,k) \end{aligned} \tag{3.16}$$

In this subsection, we re-calculate the pairwise distance between the probe q and the gallery g_i by comparing their k-reciprocal nearest neighbor set. As described earlier we believe that if two images are similar, their k-reciprocal nearest neighbor sets overlap, i.e., there are some duplicate samples in the sets. The more duplicate samples, the more similar the two images are. The new distance d_J between q and g_i can be calculated by the Jaccard metric of their k-reciprocal sets as:

$$d_J(q,g_i) = 1 - \frac{|\mathcal{R}(q,k) \cap \mathcal{R}(g_i,k)|}{|\mathcal{R}(q,k) \cup \mathcal{R}(g_i,k)|} \tag{3.17}$$

Lastly, weights $\mathcal{V}q, g_i$ are given to each of the retrieved images g_i according to the original distance between the probe q and its neighbour as follows:

$$\mathcal{V}q, g_i = \begin{cases} \mathrm{e}^{-d(q,g_i)} & \text{if } g_i \in \mathcal{R}^*(q,k) \\ 0 & \text{otherwise} \end{cases} \tag{3.18}$$

Using this simple and efficient technique we were able to boost our system's performance for both same-domain person re-id and cross-domain person re-id.

The discussion of the same is done in Section 4.2.2.

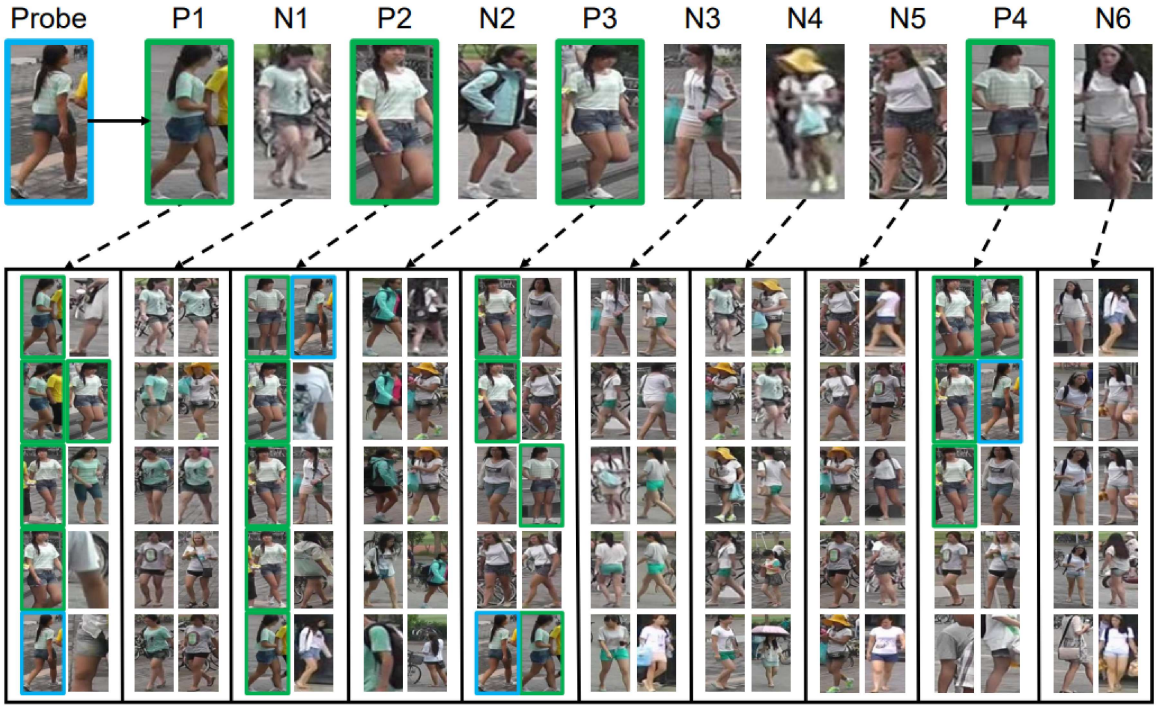


Figure 3.5: Demonstration of *k*-reciprocal nearest neighbours. Green borders represent true positive retrievals and blue border represents the query image.

CHAPTER 4
Experiments and Results

4.1 Experiments

In this work, we have performed several experiments on same-domain and cross-domain person re-id using ResNet50 and OSNet as backbone networks. We experimented with softmax and triplet loss as loss functions and implemented re-ranking as a post-processing step to improve our model's scores. We explain the experiments in detail in this section.

4.1.1 Datasets

The experiments for this work were performed on the following publicly available datasets:

Market1501: Market1501 [17] is a dataset collected in front of a supermarket using six cameras, out of which five are high resolution, and one is lower resolution. The dataset contains 32668 images with 1501 identities. Each person's image has been taken by at least two cameras and at most by six.

DukeMTMC-ReID: DukeMTMC-ReID [3] is a subset of the DukeMTMC dataset that includes 85 minutes of high-resolution video from eight separate cameras. The split recommended by the contributors has 16,522 training photos from 702 people, 2,228 query images from another 702 people, and a search gallery of 17,661 images.

CUHK03: CUHK03 [9] is collected from five different cameras. It contains 13164 images of 1360 identities.

4.1.2 Evaluation Metrics

Mean Average Precision (mAP): Mean Average Precision is a metric often used for retrieval tasks in natural language processing as well as other fields. It is the mean of the average precision for several queries. It is calculated as follows:

$$mAP = \frac{1}{|Q|} \sum_{q \in Q} AP_q \tag{4.1}$$

where AP_q is the average precision for query q in the query set Q which has $|Q|$ queries in total. All calculated values of mAP are in percentage (%) in this thesis.

Cumulative Matching Characteristics (CMC): Cumulative Matching Characteristics (CMC) are the most popular evaluation metrics for person re-identification methods and other retrieval tasks. Consider a simple single-gallery-shot scenario in which a single instance represents each gallery identity. For each query, an algorithm ranks all gallery samples by their distances from the query and the top-k scores (R_1, R_5, R_{10}, and so on) are produced.

$$R_k = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{j=1}^{k} F_q(x_j)}{k} \tag{4.2}$$

where $F_q(x_j) = 1$ if the retrieved image x_j and the query image q belong to the same class and $F_q(x_j) = 0$ otherwise. All calculated values of CMC (R_1, R_5, R_{10}) are in percentage (%) in this thesis.

4.2 Results with Different Approaches

We have conducted experiments using the three publicly available datasets mentioned in Section 4.1.1 for same-domain and cross-domain person re-identification.

4.2.1 Same-Domain Person Re-identification

We used the omni-scale network (OSNet) with adaptive instance normalization (AIN) for training and testing on the Market1501 dataset and DukeMTMC-ReID in two separate experiments. We also used a ResNet50 backbone and softmax loss, where we experimented with an Adam optimiser and an AMSGrad optimizer on the Market1501 dataset. Furthermore, we performed various experiments with ResNet50 backbone and triplet loss with Adam optimizer on all three datasets.

Here, we also experimented with Euclidean distance and cosine distance as distance measures in triplet loss. Lastly, we also report results on Market1501 using the GAN-based methodology.

Table 4.1: Results using different approaches for same-domain person re-id.

Market1501 → Market1501					
Backbone	Loss Function	mAP (%)	$R_1(\%)$	$R_5(\%)$	$R_{10}(\%)$
ResNet50	Softmax	69.00	85.70	94.20	96.30
ResNet50 (w. re-ranking)	Triplet	79.41	85.90	92.80	94.40
ResNet50 (w. re-ranking)	Quadruplet	80.52	86.20	93.00	94.60
OSNet-AIN	Softmax	82.80	93.30	98.60	99.10
JDGL	Combined*	55.34	77.29	91.12	94.00
DukeMTMC-ReID → DukeMTMC-ReID					
Backbone	Loss Function	mAP (%)	$R_1(\%)$	$R_5(\%)$	$R_{10}(\%)$
OSNet-AIN	Softmax	72.10	86.10	92.80	94.70
ResNet50 (w. re-ranking)	Triplet	55.70	67.60	77.60	81.40
ResNet50 (w. re-ranking)	Quadruplet	62.60	73.00	81.50	86.70
CUHK03 → CUHK03					
Backbone	Loss Function	mAP (%)	$R_1(\%)$	$R_5(\%)$	$R_{10}(\%)$
ResNet50 (w. re-ranking)	Triplet	44.88	40.86	55.5	65.36
ResNet50 (w. re-ranking)	Quadruplet	48.00	43.90	59.10	69.10

We conducted experiments to obtain preliminary results for same-domain results on three publicly available datasets, viz. Market150, DukeMTMC-ReID and CUHK03. Using our ResNet50 model with triplet loss and re-ranking, with only a ten epochs, we were able to get good results in the context of the results studied in our literature review. The results are compared in detail with other approaches in Table 4.1.

4.2.2 Cross-Domain Person Re-identification

Using OSNet as the backbone for feature extraction, we experimented with some variations in the architecture - one as mentioned in the above section (OSNet-AIN) and another with Squeeze-and-Excitation network (SENet) incorporated in the network. We tried two such configurations. We added a SENet block after every bottleneck layer in the first configuration. We then used the idea of a fully attentional (FA) block from the work by Huang et al. [8]. The SE blocks are added after every 3×3 lite layer, and their outputs are added.

We also use the idea of joint discriminative and generative learning from Zheng et al. [18] to generate images using structure and appearance codes. Unlike their work, we use only primary features during learning using the teacher-student

model. Furthermore, we report scores for cross-domain person re-id using this method by training on Market1501 and testing on DukeMTMC-ReID. We set the values of hyperparameters to $\lambda_{recon}^{img} = 0.5$, $\lambda_{recon}^{code} = 1$, $\lambda_{id}^{s} = 1$, $\lambda_{id}^{c} = 5$, $\lambda_{adv} = 1$, $\lambda_{prim} = 0.8$. We set $\lambda_{fine} = 0.2$ when considering fine-grained feature learning loss.

- Without fine-grained features: mAP = 9.80% R_1 score = 21.05%
- With fine-grained features: mAP = 8.72% R_1 score = 19.75%

Note that these are preliminary results for 5000 iterations. In our experiment, we disregard the computation of the fine-grained feature learning loss.

Lastly, we use the ResNet50 backbone for feature extraction and triplet loss. We train the model for Market1501 to DukeMTMC-ReID and vice versa and report the results. The margin α was set to α = 0.25. We also use quadruplet loss in place of triplet loss where the two margin parameters - α_1 and α_2 are set to α_1 = 0.8 and α_2 = 0.25 respectively. Further, we experiment by using two source domains during training. This serves two purposes - firstly, it provides the model with more data as compared to single-source training, and secondly, it also helps the model to generalise better as the model is fed with data from two different domains where the images have variations in lighting, clothing sense of individuals, etc. as mentioned in Section 1.3. We conclude that for cross-domain person re-id, using multiple source datasets helps the generalising ability of the model and boosts the cross-domain scores obtained from single-source training. We also make use of a post-processing technique in order to boost the retrieval results - re-ranking. The boost in results obtained by these techniques is tabulated in Figure 4.1 and Table 4.4, respectively.

In Table 4.1 and Table 4.2, for joint discriminative and generative learning (JDGL), we use a combined loss function that consists of softmax loss for identity discrimination, image and code reconstruction loss for primary and fine-grained feature learning and adversarial loss that governs the quality of generated images. We train the model only for 10000 iterations during our experiments.

Similar to same-domain re-id systems discussed in Section 4.2.1, we also performed experiments with various methods for cross-domain re-id. Table 4.2 discusses the results using single-source training for OSNet-based and ResNet50-based triplet loss methods with re-ranking. Like in same-domain systems, ResNet50 with triplet loss is able to perform better than OSNet-based methods for both the cases we experimented on - Market1501 $\rightarrow$ DukeMTMC-ReID and DukeMTMC-ReID $\rightarrow$ Market1501.

Table 4.2: Results using different approaches for cross-domain person re-id.

Market1501 → DukeMTMC-ReID					
Backbone	Loss Function	mAP (%)	$R_1(\%)$	$R_5(\%)$	$R_{10}(\%)$
OSNet - IBN	Softmax	21.80	39.60	55.90	61.30
OSNet-AIN	Softmax	22.30	39.20	56.00	62.10
OSNet with SENet	Softmax	9.60	19.90	31.50	38.00
ResNet50	Triplet	23.59	30.48	42.95	48.83
ResNet50	Quadruplet	21.51	29.94	40.75	46.23
JDGL	Combination*	12.23	24.64	37.93	44.08
DukeMTMC-ReID → Market1501					
Backbone	Loss	mAP (%)	$R_1(\%)$	$R_5(\%)$	$R_{10}(\%)$
ResNet50	Triplet	31.45	50.24	63.45	68.85
ResNet50	Quadruplet	32.00	52.08	64.55	68.71

Table 4.3: Comparison of Euclidean and cosine distance

Source	Target	Loss Function	mAP (%)	$R_1(\%)$
M	M	Euclidean	79.41	85.9
M	M	Cosine	65.27	76.84
D	D	Euclidean	55.66	67.64
D	D	Cosine	50.39	61.18
C	C	Euclidean	44.88	40.86
C	C	Cosine	35.4	32.79
M	D	Euclidean	23.59	30.48
M	D	Cosine	18.96	26.48
D	M	Euclidean	31.45	50.24
D	M	Cosine	25.15	44.12

We experimented with Euclidean distance and cosine distance measure in the case of the triplet loss function to calculate the distance between the images consisting of the triplet. Using Euclidean distance, the results were better in terms of mAP and CMC scores than cosine distance. The results are tabulated in Table 4.3.

As discussed in Section 3.3.1, we perform re-ranking as a post-processing step to enhance the results of our system. Table 4.4 below tabulates the results without and with re-ranking and the gain in mean average precision (mAP). We then show the gains in performance in cross-domain person re-id while using ResNet50 as the backbone with triplet loss with multiple source datasets in training in Table 4.5.

Lastly, we compare our results obtained with ResNet50 feature extractor and triplet loss as the loss function and re-ranking for performance boost with the results reported in [10]. We also mention the state-of-the-art results for cross-domain person re-identification [16] wherein a heterogeneous convolution net-

Table 4.4: Gain in results using re-ranking

Source	Target	mAP (%) (w/o. re-ranking)	mAP (%) (w. re-ranking)	Gain (% points)
M	M	52.73	79.41	26.68
D	D	39.53	55.66	16.13
C	C	30.98	44.88	13.90
M	D	14.24	23.59	9.35
D	M	25.66	31.45	5.79

work (HCN) is used to learn the correlation between pedestrian images from the training data. From Table 4.5, it can be seen that using our multi-source and re-ranking approach, we outperform the mean average precision (mAP) scores for cross-domain person re-id reported in [10].

Table 4.5: Comparison of results using multi-source and single-source training data

Title	Source	Target	mAP (%)	$R_1(\%)$	$R_5(\%)$	$R_{10}(\%)$
Zhang et al.[16]	M	D	57.30	78.90	-	-
Luo et al.[11]	M	D	55.20	68.80	-	-
Liu et al.[10]	M	D	37.40	56.80	-	-
ResNet50 (Triplet)	M	D	23.59	30.48	42.95	48.83
ResNet50 (Quadruplet)	M	D	21.51	29.94	40.75	46.23
ResNet50 (Triplet)	M + C	D	34.55	44.97	58.53	64.18
ResNet50 (Quadruplet)	M + C	D	**37.97**	49.46	80.48	84.43
Zhang et al.[16]	D	M	70.20	90.20	-	-
Luo et al.[11]	D	M	61.70	82.10	-	-
Liu et al.[10]	D	M	36.30	67.20	-	-
ResNet50 (Triplet)	D	M	31.45	50.24	63.45	68.85
ResNet50 (Quadruplet)	D	M	32.00	52.80	64.55	68.71
ResNet50 (Triplet)	D + C	M	**36.30**	47.70	58.53	64.18
ResNet50 (Quadruplet)	D + C	M	**41.76**	60.51	64.55	68.71

In Table 4.3, Table 4.4 and Table 4.5, it should be noted that in the source and target columns, "M" stands for Market1501, "D" stands for DukeMTMC-ReID and "C" stands for CUHK03 dataset, respectively.

For cross-domain person re-id, using a ResNet50 feature extractor and triplet loss along with multi-source training and re-ranking, we obtained a mean average precision (mAP) score comparable to the one reported in [10] for the same target dataset. Using quadruplet loss, we beat the mAP reported in the same work for both the target datasets (viz. Market1501 and DukeMTMC-ReID). The values in boldface in Table 4.5 represent the mAP scores obtained using our approaches that outperform the mAP reported in [10]. The gains obtained in the system for

cross-domain person re-id using multi-source training strategy can be illustrated using the bar graph below.

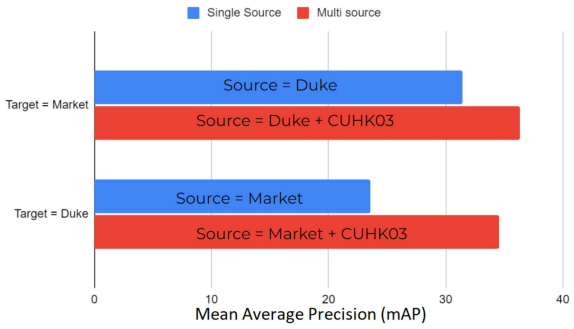


Figure 4.1: Gains obtained using multi-source training

Figure 4.2 and Figure 4.3 below show a visualization of the ranking list generated by the model as output. The leftmost image is the query image and the following ten images are retrieved from the gallery set. Retrieved images with a green box represent a true positive or a "hit" whereas red borders represent a false positive. In Figure 4.2, from top to bottom, we see two examples of retrieval each for same-domain, single-source cross-domain and multi-source cross-domain, respectively when triplet loss was used as the loss function whereas Figure 4.3 shows the same for quadruplet loss. It should be noted that all results are post applying re-ranking. The true positive retrievals for same-domain examples are very high while it performs much worse on the single-source cross-domain examples. Hence, we exhibit the complexity involved in cross-domain re-id. We use multi-source training to improve the retrievals in comparison to single-source training. Figure 4.3 has more retrievals with green borders for all cases, suggesting quadruplet loss performs better than triplet loss.

(a) Market1501 ⟶ Market1501

(b) DukeMTMC-ReID ⟶ Market1501

(c) DukeMTMC-ReID + CUHK03 ⟶ Market1501

Figure 4.2: Visualization of ranking list for probe images in Market1501 dataset for triplet loss in case of (a) same domain, (b) single-source cross-domain and (c) multi-source cross-domain person re-identification.

(a) Market1501 ⟶ Market1501

(b) DukeMTMC-ReID ⟶ Market1501

(c) DukeMTMC-ReID + CUHK03 ⟶ Market1501

Figure 4.3: Visualization of ranking list for probe images in Market1501 dataset for quadruplet loss in case of (a) same domain, (b) single-source cross-domain and (c) multi-source cross-domain person re-identification.

CHAPTER 5
Conclusions

5.1 Conclusion

The main challenge in the person re-identification task is for the model to generalize well on data from different domains. One way to solve this problem of generalization is to feed the model with diverse data to ensure that the model does not learn domain-related cues during training. To this end, we have experimented with feature extractors like ResNet50 and OSNet, as well as a variety of loss functions, including triplet loss and softmax loss. We also experimented using a specialised Generative Adversarial Network (GAN) that performs generative and identity discriminative modeling in the same network. Using multi-source training and re-ranking, we found an improvement in performance with triplet and quadruplet loss as loss functions. Exhaustive experiments have been conducted on three publicly available person re-id datasets viz. Market1501, DukeMTMC-ReID and CUHK03.

5.2 Future Work

We aim to fine-tune the extraction process to calculate better features that can better represent the discriminative identity features of the images. Using lightweight, compact models such as OSNet and improving their feature extraction process will be a direction to focus on going forward.

We also aim to experiment with different attention mechanisms that are well-adapted to the task of person re-identification and provide a boost to the current results.

References

[1] S. Choi, T. Kim, M. Jeong, H. Park, and C. Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2021.

[2] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[3] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–19, 2017.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 87–102, Cham, 2016. Springer International Publishing.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[7] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020.

[8] W. Huang, Y. Li, K. Zhang, X. Hou, J. Xu, R. Su, and H. Xu. An efficient multi-scale focusing attention network for person re-identification. *Applied Sciences*, 11(5):2010, 2021.

[9] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[10] H. Liu, J. Cheng, S. Wang, and W. Wang. Attention: A big surprise for cross-domain person re-identification. *arXiv preprint arXiv:1905.12830*, 2019.

[11] X. Luo, Z. Ouyang, N. Du, J. Song, and Q. Wei. Cross-domain person re-identification based on feature fusion. *IEEE Access*, 9:98327–98336, 2021.

[12] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018.

[13] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

[14] M. Zhang, Q. Cheng, F. Luo, and L. Ye. A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2711–2723, 2021.

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

[16] Z. Zhang, Y. Wang, S. Liu, B. Xiao, and T. S. Durrani. Cross-domain person re-identification using heterogeneous convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1160–1171, 2022.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[18] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[19] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. pages 3652–3661, 07 2017.

[20] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712.

[21] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Learning generalisable omni-scale representations for person re-identification. *CoRR*, abs/1910.06827, 2019.