WARWICK
THE UNIVERSITY OF WARWICK

WARWICK
THE UNIVERSITY OF WARWICK

# Scalable Deep Feature Learning for Person Re-identification

by

**Shan Lin**

**Thesis**

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

**Doctor of Philosophy**

**Department of Computer Science**

September 2019

# Contents

# List of Tables

# List of Figures

# Abstract

Person Re-identification (Person Re-ID) is one of the fundamental and critical tasks of the video surveillance systems. Given a probe image of a person obtained from one Closed Circuit Television (CCTV) camera, the objective of Person Re-ID is to identify the same person from a large gallery set of images captured by other cameras within the surveillance system. By successfully associating all the pedestrians, we can quickly search, track and even plot a movement trajectory of any person of interest within a CCTV system. Currently, most search and re-identification jobs are still processed manually by police or security officers. It is desirable to automate this process in order to reduce an enormous amount of human labour and increase the pedestrian tracking and retrieval speed. However, Person Re-ID is a challenging problem because of so many uncontrolled properties of a multi-camera surveillance system: cluttered backgrounds, large illumination variations, different human poses and different camera viewing angles.

The main goal of this thesis is to develop deep learning based person re-identification models for real-world deployment in surveillance system. This thesis focuses on learning and extracting robust feature representations of pedestrians. In this thesis, we first proposed two supervised deep neural network architectures. One end-to-end Siamese network is developed for real-time person matching tasks. It focuses on extracting the correspondence feature between two images. For an offline person retrieval application, we follow the commonly used feature extraction with distance metric two-stage pipline and propose a strong feature embedding extraction network. In addition, we surveyed many valuable training techniques proposed recently in the literature to integrate them with our newly proposed NP-Triplet loss to construct a strong Person Re-ID feature extraction model. However, during the deployment of the online matching and offline retrieval system, we realise the poor scalability issue in most supervised models. A model trained from labelled images obtained from one system cannot perform well on other unseen systems. Aiming to make the Person Re-ID models more scalable for different surveillance systems, the third work of this thesis presents cross-Dataset feature transfer method (MMFA). MMFA can train and transfer the model learned from one system to another simultaneously. Our goal to create a more scalable and robust person re-identification system did not stop here. The last work of this thesis, we address the limitation of MMFA structure and proposed a multi-dataset feature generalisation approach (MMFA-AAE), which aims to learn a universal feature representation from multiple labelled datasets. Aiming to facilitate the research towards Person Re-ID applications in more realistic scenarios, a new datasets ROSE-IDENTITY-Outdoor (RE-ID-Outdoor) has been collected and annotated with the largest number of cameras and 40 mid-level attributes.

# Acknowledgments

I would like to express my sincerest gratitude to my supervisor Professor Chang-Tsun Li for his valuable guidance. Thought my PhD at the University of Warwick, I really appreciate his patience support and helps by providing me with many insightful comments and suggestions on my academic research. I would also like to thank Professor Alex Chichung Kot for providing me with excellent research environments and facilities at ROSE lab, Nanyang Technological University during my two year secondment under EU IDENTITY project. I also want to thank all my friends, my colleagues, my team members at the University of Warwick, Hong Kong Baptist University and Nanyang Technological University for their support and encouragement. Finally, I would like to thank my family for their sincere love and support over many years.

# Publications

Parts of this thesis have been previously published by the author in the following papers:

- Shan Lin and Chang-Tsun Li. End-to-End Correspondence and Relationship Learning of Mid-Level Deep Features for Person Re-Identification. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017
- Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex C Kot. Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification. In *British Machine Vision Conference (BMVC)*, 2018
- Shan Lin and Chang-Tsun Li. Person Re-identification with Soft Biometrics through Deep Learning. In *Deep Biometrics*, pages 21–36. Springer, Cham, 2020. ISBN 978-3-030-32582-4

# Sponsorships and Grants

All research works of this thesis are supported by EU Horizon 2020 project, entitled Computer Vision Enable Multimedia Forensics and People Identification (acronym:IDENTITY, Project ID:690907)

Some research outcomes were carried out at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University (NTU), Singapore and supported by a grant from NTUs College of Engineering (M4081746.D90).

# Acronyms

**AAE** Adversarial Auto-Encoder.

**Adam** Adaptive Moment Estimation.

**AP** Average Precision.

**BN** Batch Normalisation.

**CCTV** Closed Circuit Television.

**CMC** Cumulative Matching Characteristic.

**CNN** Convolutional Neural Networks.

**CV** Computer Vision.

**FC** Fully Connected.

**GAN** Generative Adversarial Network.

**GAP** Global Average Pooling.

**GDPR** General Data Protection Regulation.

**GMP** Global Max Pooling.

**GPU** Graphics Processing Units.

**HSV** Hue, Saturation and Value.

**IVSS** Intelligent video surveillance systems.

**KISSME** Keep It Simple and Straight-forward Metric Learning.

**KNN** K-Nearest Neighbour.

**LBP** Local Binary Patterns.

**LFDA** Local Fisher Discriminant Analysis.

**LOMO** Local Maximal Occurrence.

**mAP** Mean Average Precision.

**ML** Machine Learning.

**MMD** Maximum Mean Discrepancy.

**MMFA** Multi-task Mid-level Feature Alignment Network.

**MMFA-AAE** Multi-domain Mid-level Feature Alignment Adversarial Auto-Encoder Network.

**MOT** Multiple Object Tracking.

**NTU** Nanyang Technological University.

**Person Re-ID** Person Re-identification.

**PSD** Positive Semi Definite.

**RBF** Radial Basis Function.

**Re-ID-Outdoor** Rose-Identification-Outdoor Dataset.

**Re-ID** Re-identification.

**RGB** Red, Green, Blue.

**SDALF** Symmetry-Driven Accumulation of Local Features.

**SGD** Stochastic Gradient Descent.

**SIFT** Scale Invariant Feature Transforms.

**SILTP** Scale Invariant Local Ternary Pattern.

**SVM** Support Vector Machine.

**UDA** Unsupervised Domain Adaptation.

**XQDA** Cross-view Quadratic Discriminant Analysis.

# Symbols

| | |
|---|---|
| $*$ | Convolution |
| $\bigodot$ | Element-wise multiplication |
| $x \in \mathbb{R}^D$ | Dimensional vector |
| $(\cdot)^T$ | Matrix or vector transpose |
| $[x\ y]$ | Concatenation of the vectors $x$ and $y$ |
| $\|x\|_2$ | L2 norm of the vector $x$ |
| $\|x - y\|_2$ | Euclidean distance between vectors $x$ and $y$ |

# Chapter 1

# Introduction

Due to the increasing number of terrorist attacks and riots all over the world in recent years, there are strong public calls for greater surveillance systems to thwart acts of terror preemptively. Many governments and agencies are seriously concerned about public security in areas such as airports and shopping malls. To accomplish this goal, Closed Circuit Television (CCTV) is playing a key role in public area surveillance and has become an integral part of national security. Hundreds and thousands of surveillance camera networks have already been deployed in many public places to address various kind of security issues such as forensic investigations, crime preventing, safeguarding the restricted areas, etc. These surveillance cameras generate a large amount of video recordings per day. Currently, these recorded videos have to be analysed manually by the surveillance operators to detect any specific incident or anomaly. In recent years, surveillance camera networks are rapidly deployed all over the world, and the conventional manual analysing is unable to cope with the rapid expansion of camera networks. Intelligent video surveillance systems (IVSS) aim to automate the video monitoring and analysing to assist the surveillance operators quickly extracting relevant information from the recorded videos. Therefore, it is one of the most active and challenging research area in computer vision (CV) and machine learning (ML). This field of research enables various applications such as on-line people/object detection and tracking [24, 93, 94], recognising a suspicious action/behaviour [47, 48]; and off-line suspect images retrieval from video frames [92, 108, 136, 155].

In video surveillance systems, one primary task for operators is to identify and track the same individuals across different cameras, also known as person re-identification (Person Re-ID). To replace the human jobs in video surveillance, an effective IVSS should be able to track the appearance of a person and re-identify the same individual if he or she re-appears in other cameras. Figure 1.1 gives an example of re-identifying people in a non-overlapping multi-camera network. The red dot and the green dot in Figure 1.1 represent two different persons who are assigned ID2 and ID3 by the system. As they move from the location near Camera 2 to a place near Camera 1, the system needs to be able to recognise them as pedestrian 2 (ID2) and 3 (ID3) by using the information obtained previously in Camera 2. By continuously and successfully re-identifying the selected persons across all cameras in the system, the trajectories of the multiple targets can be reconstructed for further investigation. This re-identification process seems simple and intuitive for a human to operate, but it is difficult for CV to accomplish due to numerous open issues. In particular, since each camera captures the people of interest from different angles, distances, and lighting conditions, the same person may look very different in different cameras. Moreover, the occlusion, the colour calibration between cameras and diverse backgrounds will also affect the extraction of the visual features.

Person Re-ID is a challenging problem, but it is a vital part of IVSS. In recent years, Person Re-ID has received a large amount of attention in the computer vision research communities. Fundamentally, the goal of Person Re-ID is to match images of humans from one surveillance camera to the other. Figure1.2 (a) shows example images of three different people captured by a network of 6 cameras. Given a query image of a particular person captured by one camera, the goal of Person Re-ID is to retrieve the images of the same person from the other cameras (Figure 1.2 (c)). By integrating this function into the current surveillance system, it will help to save an enormous amount of manual labour in cross-camera people tracking

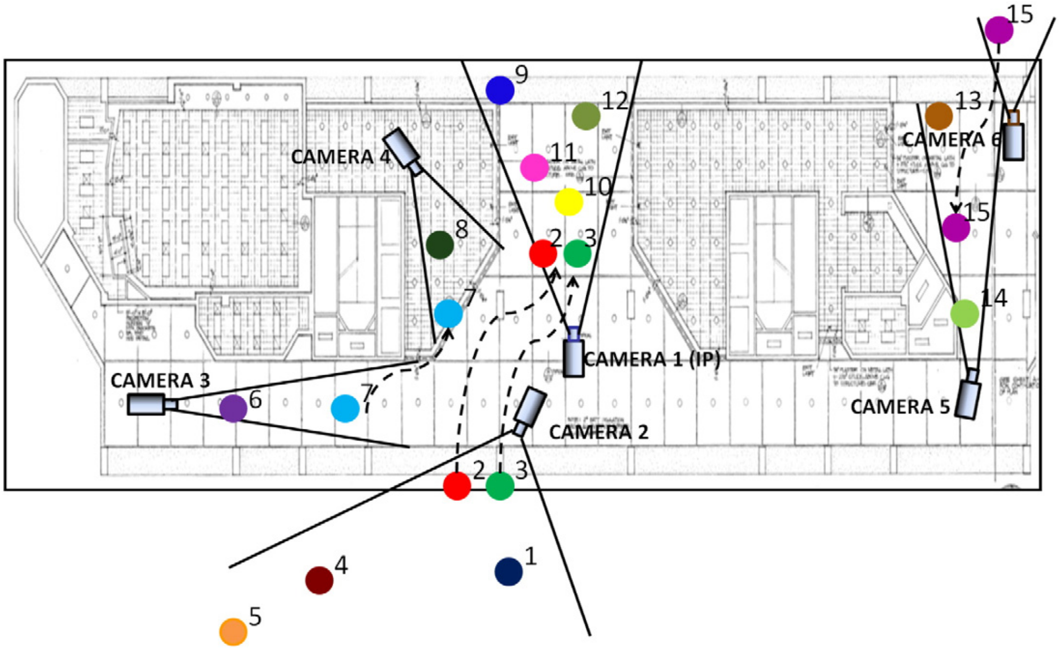


Figure 1.1: Demo of re-identify the same person in a non-overlapping multi-camera network (figure provided in [7])

within a CCTV system. Besides, it can also be extended to other applications such as multi-camera person retrieval, plotting the trajectory of a subject's movement and many other real-time and forensic applications. Current researches rely on several publicly available datasets captured under restricted settings. The largest public dataset consists of images captured from a network of 15 cameras [122]. The images are captured with minimal occlusion, and several junk/distractor images are present due to miss-detections by the pedestrian detector. Different taxonomies are used in various research works for classifying the existing person re-identification methods. Single-shot recognition refers to a one-to-one matching of a pair of images from two cameras. Multi-shot recognition refers to matching two sets of images obtained from two different cameras. The two sets of images are obtained by capturing multiple frames from the motion trajectory of the pedestrians, as shown in Figure1.2 (b). In the multi-shot recognition experiments, multiple images can provide more visual information of a query person compared to a single-image recognition setting. Hence, multiple-images recognition is a relatively less challenging problem and usually yields a higher retrieval rate. However, the multi-shot recognition setting requires multiple images of a person from each camera. This requirement forces the Person Re-ID system to use more processing time in order to capture multiple images of each person. Hence, the Re-ID model developed for multiple-images recognition is difficult to be integrated into the real-time person re-identification applications. As a result, in this thesis, we focus on the more challenging and single-image Person Re-ID problem and conduct experiments mostly in the single-shot setting.

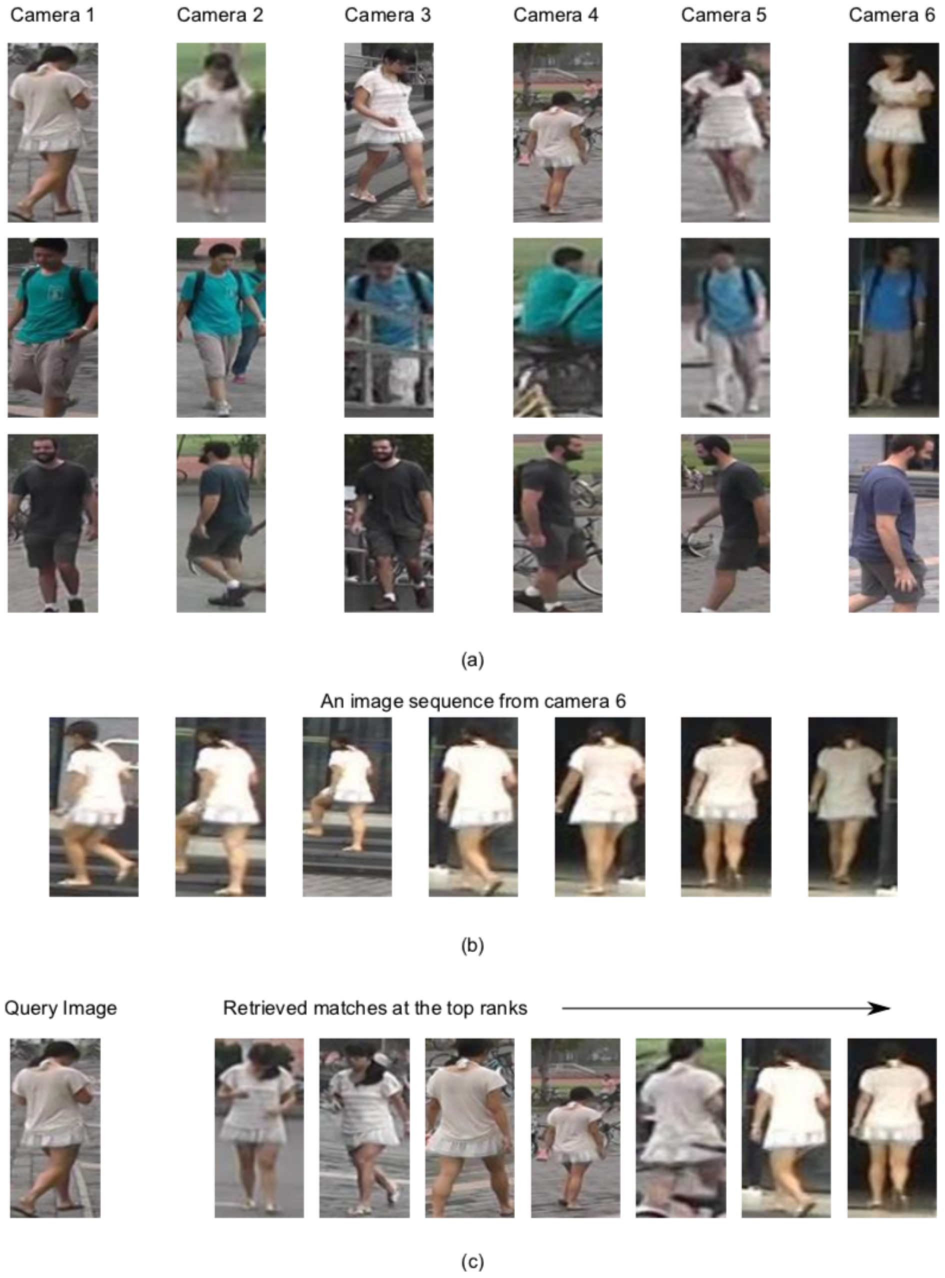


Figure 1.2: (a) Sample images captured by a network of 6 cameras (From Market-1501 dataset [1]), (b) Image sequence of an identity from a single camera (c) Example re-identification scenario with a query image and the retrieved matches.

## 1.1 Challenges and Motivations

Person re-identification is an inherently challenging task due to several reasons. Some of the major challenges are listed below.

1. **Intra-class Variation**: In the typical person re-identification setting, images are captured by different cameras with non-overlapping fields of view. Hence the environmental conditions such as background, illumination, the viewing angle at one location may not be necessarily the same as other cameras at different locations. This will result in a substantial intra-class variations as the appearance of the images in different views may substantially differ from the original appearance. As shown in Figure1.3 (a) and Figure1.3 (b), the images in view 1 and view 2 are significantly different from each other.

2. **Occlusion in Crowded Scenes**: Public places such as railway stations, streets and shopping malls can often be very crowded. Re-identifying people in these environments is extremely challenging to extract the pedestrian bounding box for a subject without any occlusion. Since many public datasets employ automatic pedestrian detectors, this will result in miss detections or wrong detections. All the above conditions make the re-identification very challenging in crowded scenes as the full appearance of the subject is unavailable. Figure1.3 (c) shows some example cases with occlusion.

3. **Generalisation capability and scalability**: Supervised single-dataset person re-identification models often over-fit to the training dataset. Once trained from images of one particular video system in a supervised fashion, most models do not generalise well to another system with different viewing conditions. If we directly deploy a model trained from a public dataset to a new CCTV system, the model usually sufferers from considerable performance degradation. Hence, it requires a large number of labelled matching pairs obtained from the new system for training. Such a setting severely limits their scalability in

real-world applications where annotating every camera systems is a costly and time-consuming job.

4. **Long-term Re-identification**: Long-term person re-identification can pose challenges in two main ways. First, tracking a pedestrian from location 1 to location 2 becomes increasingly difficult as the separation in space and time increases. To search a person in real-time becomes challenging to decide the search space as well as the time frame to conduct the search. Though exhaustive searching across all cameras can be a solution, it becomes tedious considering the scalability of the re-identification algorithms. Second, it poses a higher possibility that there may be some changes in the appearance of the pedestrians (i.e. change of clothes, accessories etc.). Due to these issues, collecting a person re-identification dataset is extremely time-consuming and tedious work.

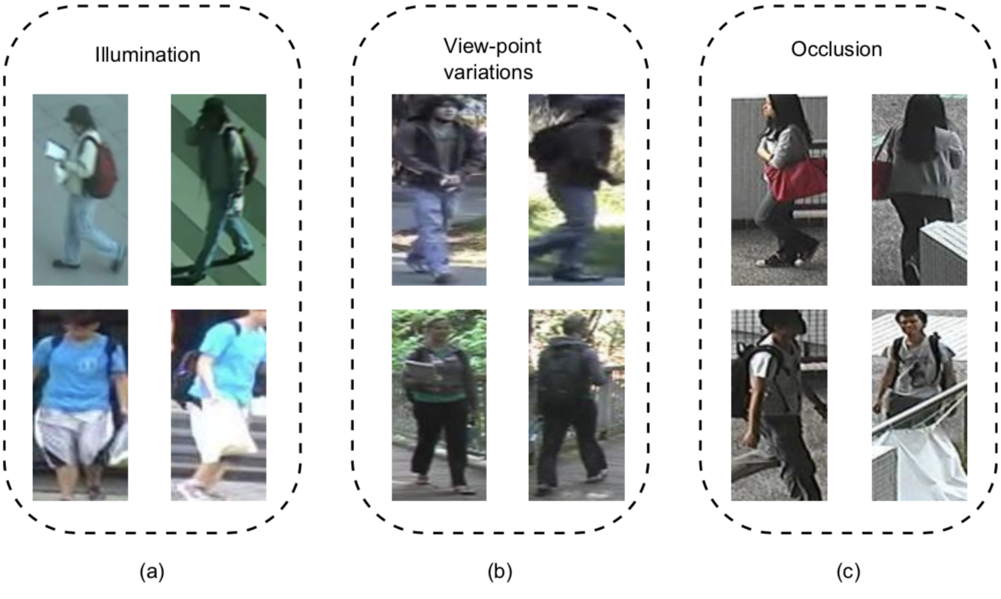


Figure 1.3: Summary of major challenges in Person Re-identification.

This thesis mainly focuses on addressing the first and the third challenges. The first aim is to learn robust representations which are invariant to illumination, pose and viewpoint changes. Instead of addressing each of these problems individually, the focus is to develop a holistic approach that can handle the problem of large intra-class variations. The fundamental idea is to learn the feature representations where the distance between the similar pairs (i.e. images of subjects belonging to the same identity) is lesser compared to the distance between dissimilar pairs (i.e. images of subjects belonging to different identities). The second aim of this thesis is to address the limited scalability issue in many Person Re-ID models. We first proposed a cross-dataset feature transfer method which can transfer from a pre-trained existing model from one system to another. Later, we proposed a multi-dataset feature generalisation model which aim to learn a universal domain invariant feature representation by leveraging the labelled data from multiple available datasets.

## 1.2 Thesis Contributions

Motivated by the ideas mentioned in Section 1.1, this thesis focuses on learning robust and invariant image representations for the real-world of Person Re-ID applications. This thesis proposed four Person Re-ID approaches from a constrained single-dataset setting to a more scalable cross-dataset and multi-dataset setting. The main contributions of the thesis can be summarised as follows.

- We propose an end-to-end deep mid-level feature correspondence network that learns to find the common mid-level salient features of people. As an end-to-end architecture, the output of the network can be used to provide the similarity score between the query and gallery images directly. The similarity score from our model can be used for the online person matching task and real-time cross-camera pedestrian tracking application.

- Person Re-ID model can also be used for person retrieval from a large gallery of human images obtained from several historical video files. For person retrieval application, we proposed a robust and simple feature extraction network based on our novel negative competing triplet loss function (NC-Triplet). Our proposed method also integrate with several data refinements and training techniques proposed in recent years.

- The existing public datasets are collected from a very limited number of cameras (ranging from 6 to 15), compared to hundreds of cameras in a real-world video surveillance system. Most of these datasets are collected without any privacy consideration. Hence, we collected and annotated a new Person Re-ID dataset called Rose-IDentification-Outdoor (Re-ID-Outdoor). Our dataset is collected from 50 real surveillance cameras and come with privacy consideration from all participants. Overall, our Re-ID-Outdoor dataset is currently the most realistic and also the only privacy-aware public dataset for Person Re-ID research.
- To address the scalability problem in these supervised single-dataset Person Re-ID models, we proposed a cross-dataset feature transfer network which utilises the mid-level attributes of the person to bridge the domain gap between two different CCTV system (different datasets). By aligning the distribution of each attribute from the source dataset to the target dataset, the network can simultaneously learn the feature representation from the source dataset and adapt to the target datasets.
- The cross-dataset transfer learning approaches require a large amount of unable data from the target domain. The adaptation process also introduces additional training time before deployment. Therefore, we proposed a novel domain generalisation network. It leverages the images from multiple Person Re-ID datasets to generate a more robust and well-generalised feature representation. This setting simulates the real-world scenario in which a strong feature learner is trained once and deployed to multiple camera networks without further data collection or adaptive training.

## 1.3 Thesis Outline

This thesis is organised as follows.

- **Chapter 2: Literature Review**

  This chapter describes various research studies and works carried out by researchers to tackle the problem of person re-identification. It provides a comprehensive overview of the state-of-the-art Person Re-ID deep learning algorithms from threes different perspectives:

  - Single-Dataset Supervised Learning
  - Cross-Dataset Domain Adaptation
  - Multi-Dataset Domain Generalisation.

  We also include detailed distributions and statistical summary of several popular Person Re-ID datasets used in this thesis and many other research works.

- **Chapter 3: Single-Dataset Supervised Feature Learning (Online Matching)**

  This chapter presents an end-to-end single-dataset supervised mid-level feature correspondence learning network for Person Re-ID. The previous siamese structure deep learning approaches focus only on pair-wise matching between feature maps of two images. They rarely discuss the internal relationship between feature maps. In our method, we considered the latent relationship between different combinations of multiple mid-level features and proposed a network structure to automatically construct the correspondence features from all input features without a pre-defined matching function. The detailed architecture and experimental results are demonstrated in this chapter. The advantage of the end-to-end network structure is suitable for the online person matching task. Base on this model, we have developed a real-time functional person matching application using real-world surveillance cameras.

- **Chapter 4: Single-Dataset Supervised Feature Learning (Offline Retrieval)**

  Chapter 4 focuses on developing an offline person retrieval application. The proposed two-stage framework divides the Person Re-ID process into feature extraction stage and similarity ranking stage. As a result, the feature embedding of the gallery images can be stored and reused for different query images. We train the ResNet50 backbone network based on our novel negative competing triplet loss function (NC-Triplet). By integrating several data refinements and training techniques, we proposed a simple and robust Person Re-ID model for offline person retrieval applications. In addition, due to the recent implementation of General Data Protection Regulation (GDPR) in Europe and some investigation on the DukeMTMC [95, 152] dataset in the US, it has drawn much public attention on the privacy issue in most of the Person Re-ID datasets. Besides, all existing Person Re-ID datasets only contain extremely limited the number of cameras (ranging from 6 to 15), compared to hundreds of cameras in a real-world video surveillance system. To address the privacy concern and collect a new dataset from a real-world size camera network, we proposed a new privacy-aware Person Re-ID dataset collection strategy. By following this strategy, we successfully collected a new large-scale dataset called Re-ID-Outdoor from a total of 50 outdoor surveillance cameras. The new Re-ID-Outdoor dataset is currently the most realistic and most challenging dataset for Person Re-ID.

- **Chapter 5: Cross-Dataset Feature Transfer**

  The most significant drawback of the supervised method is the requirement of a large amount of labelled data for training. In the real world scenarios, collecting images of the same person across hundreds and thousands of CCTV cameras is extremely expensive and time-consuming. Chapter 5 focuses on leveraging the publicly available datasets and proposed a novel domain adaptation framework:

Multi-task Mid-level Feature Alignment (MMFA) network. It can adapt the model from a labelled source dataset to any unlabelled datasets in an unsupervised manner. The proposed MMFA network shows a useful performance improvement compared to the direct model transfer and outperforms most of the state of the art methods.

- **Chapter 6: Multi-Dataset Feature Generalisation**

  Cross-dataset domain adaptation solves many problems of the practical deployment of Person Re-ID models. However, domain adaptation still requires an adaptation process before it can be applied to a new system. We believe the most practical Person Re-ID algorithm should generate a robust model which could perform well on any video surveillance system out of the box. In Chapter 6, we re-think the Person Re-ID algorithm as a multi-dataset feature generalisation problem. We proposed a multi-domain generalisation framework: Multi-domain Mid-level Feature Alignment Adversarial Auto-Encoder Network (MMFA-AAE) which leverages labelled data from multiple datasets and learns a universal feature representation for any unseen system. The detailed experimental results demonstrate the effectiveness of the proposed method.

- **Chapter 7: Conclusion and Future Directions**

  Chapter 7 provides the concluding remarks of this thesis. The limitations and recommendations for future work are discussed to provide opportunities for further research and improvement in the area of Person Re-ID.

# Chapter 2

# Literature Review

In this chapter, a survey of related works in Person Re-ID is presented. Section 2.1 begins by laying out the basic concept of person re-identification and discusses two primary issues in the research of the person re-identification problem. In Section 2.2, related hand-crafted approaches are presented. Section 2.3 first gives an overview of the deep learning approaches for many computer vision tasks and provide a general structure for the convolutional neural network (CNN). Then, we discuss the recent deep learning approaches for Person Re-ID from three different perspectives: single-dataset feature learning, coss-dataset feature transfer and multi-dataset feature generalisation. The single-dataset feature learning methods focus on learning the robust feature representation from one dataset in a fully supervised manner. The section of coss-dataset feature transfer discusses the recent research transition from fully unsupervised person re-identification to cross-dataset transfer learning. The last multi-dataset feature generalisation learning is a large underexposed research domain in Person Re-ID. Section 2.4 provides a list of famous person re-identification benchmark datasets, followed by the concluding remarks in Section 2.5.

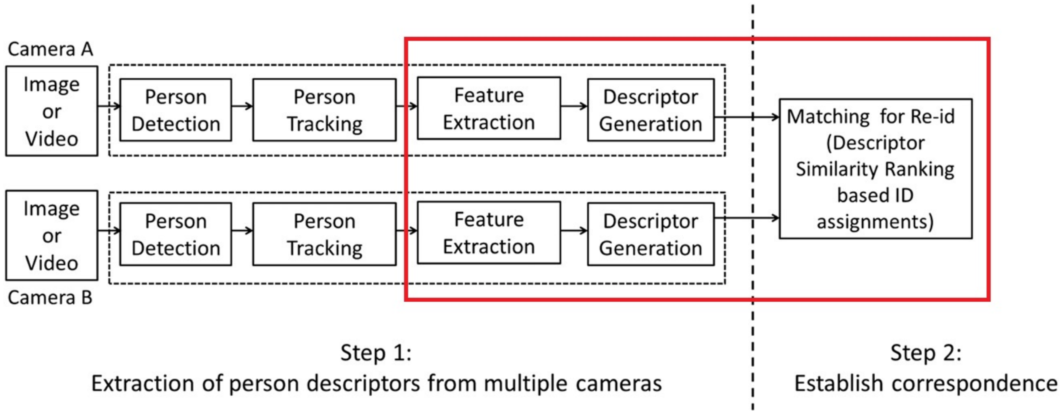


Figure 2.1: The system diagram for a typical person re-identification process

## 2.1 Person Re-identification Overview

With the prevalence of surveillance systems, there has been much research and study on the problems of person detection, person tracking, and the most recent Person Re-ID. The main objective of Person Re-ID is to match pedestrians across multiple CCTV cameras. The general schematic steps of a person re-identification system are demonstrated in Figure 2.1. In a multi-camera surveillance network, the images or videos obtained from each camera need to be analysed for detecting the presence of people. Once a person is detected, person tracking algorithms are used to detect the bounding boxes of every people in each video frame. This step removes the most irrelevant background and reduces the data size for the following processes. Then, imagery features are extracted from each pedestrian. Based on these features, a descriptor is generated for the ensuing metric learning and matching process. The first two steps: person detection and multiple people tracking are challenging problems with their own hurdles. A significant amount of work has gone into addressing issues of person detection over the years [8, 21]. Multiple Object Tracking (MOT) within a single camera has also been widely researched, and many algorithms have been proposed over the past two decades [56, 88]. Although person detection and multiple person tracking have achieved a significant improvement in terms of efficiency, accuracy and robustness in recent years [101, 127], sustained tracking across cameras with varying observation environments remains an open problem. Therefore, the primarily focuses on people re-identification research are the last three steps, highlighted in the red box in Figure 2.1:

1. Finding imagery features that are more robust and concise than raw pixels.
2. Constructing feature descriptors or representations which are capable of both describing and discriminating individuals, yet invariant to illumination, viewpoint and colour calibration.
3. Developing a matching procedure, optimised for the previous features descriptors.

Each step in Person Re-ID entails various requirements on the algorithm and system design. These requirements lead to both the development of new and the exploitation of existing computer vision techniques for addressing the problems of features representation and model matching. The majority of the existing research in human re-identification concentrates on two aspects of the problem: developing a feature representation [12, 19, 29, 44, 58, 67, 75, 121, 134, 145] and learning a distance metric [15, 42, 53, 57, 58, 60, 87, 105, 124, 130, 144].

**Feature Representations:** Contemporary approaches to re-identification typically exploit low-level features such as colour [44, 60, 76, 134, 144, 145, 147], texture [75, 144, 145], spatial structure [6], etc. It is because these features provide a reasonable level of inter-person discrimination together with inter-camera invariance. Such features are further encoded into fixed-length person descriptors, e.g. in the form of histograms [90], covariances [3] or Fisher vectors [74]. In recent deep learning approaches [1, 54, 117, 151], the feature embeddings extracted from the fully connected (FC) layer usually used as the feature descriptors for similarity measurement. This thesis primarily focuses on extracting robust deep feature representations for various Person Re-ID applications. Section 2.2 and Section 2.3 will present a comprehensive survey on various feature learning methods from hand-craft feature engineering to deep learning approaches.

**Similarity Metrics:** Once a suitable representation of features has been obtained, a similarity metric is needed to measure the similarity between two samples. Many prominent similarity learning algorithms [42, 53, 57, 87, 123] have been proposed for person re-identification. Local Fisher Discriminant Analysis (LFDA) was proposed for person re-identification in [87]. The objective is to maximise the inter-class separability and to minimise the within-class variance. To address the non-linearities in feature space, kernel-based dimensionality reduction techniques were proposed in [123]. Support Vector Machine (SVM) learning was proposed in [57], and the idea is to learn decision boundaries that are adaptive to the data samples. In [130], several kernel-based metric learning methods for person re-identification were evaluated, and kernel-based LFDA was found to be the best performing algorithm for several re-identification datasets. Recently, a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) was proposed in [60]. The algorithm is an extension of the Keep It Simple and Straight-forward Metric Learning (KISSME) approach [42] to the cross-view metric learning. Another prominent metric learning algorithm: Positive Semi Definite (PSD) logistic metric learning was introduced in [58]. It uses an efficient asymmetric sample weighting strategy. Most of the metric learning methods mentioned above can also be trained from the features extracted from deep neural networks. However, they introduce an additional training process in the training of the deep learning models. Also, because the recent deep learning Person Re-ID approaches can generate more robust feature representations, direct application of the preliminary distance metrics such as Euclidean distance or Cosine distance on the deep features can achieve excellent performance. This thesis focuses on the Person Re-ID feature extraction from the neural network and utilise the Euclidean distance and Cosine distance for a fair comparison with other deep learning approaches.

## 2.2 Hand-crafted Person Re-ID Features

Most of the Hand-crafted features for person re-identification are focusing on combining the low-level colour [44, 60, 134, 144, 145, 147], texture [75, 144, 145] or interest point detectors [144, 145] information. Features such as Colour Histograms [60, 130, 144, 145], Local Binary Patterns (LBP) [114, 130], Colour Names [134, 147], Scale Invariant Feature Transforms (SIFT) [70, 144, 145], Scale Invariant Local Ternary Pattern (SILTP) [58, 60] are commonly used in Person Re-ID. Gabor-like edges extracted from the Hue, Saturation and Value (HSV) channels for each image have also been used in [75]. In [144, 145], local patches on a dense grid are extracted and 128 dimensional SIFT features are computed for each patch of size $10 \times 10$ at a stride of 5. LBP histogram features on horizontal stripes are extracted for each image in [130]. Image is divided into horizontal stripes to handle the pose changes across views. In [58, 60], SILTP features are used in conjunction with colour histograms. Image is divided into $10 \times 10$ blocks at a stride of 5 and SILTP histograms are extracted at two scales. The maximal occurrence of each pattern in a horizontal stripe is computed to address the viewpoint changes. The resulting histogram features called Local Maximal Occurrence (LOMO) features achieve some invariance to viewpoint changes and demonstrate impressive performance on several benchmark datasets.

Alongside the research progress of the low-level feature analysis, some of the works also focus on removing the irrelevant background. One background removal approach, Symmetry-Driven Accumulation of Local Features (SDALF), is proposed by exploiting human body shape. As the human body is naturally symmetrical, the backgrounds rarely show such coherent and symmetric patterns. SDALF uses this symmetric and asymmetric difference between human and background to extract meaningful body parts, as shown in Figure 2.2. Besides, the SDALF approach gives higher weights to features extracted near the vertical and horizontal axis, which will further reduce the potential background feature contamination [6]. This method

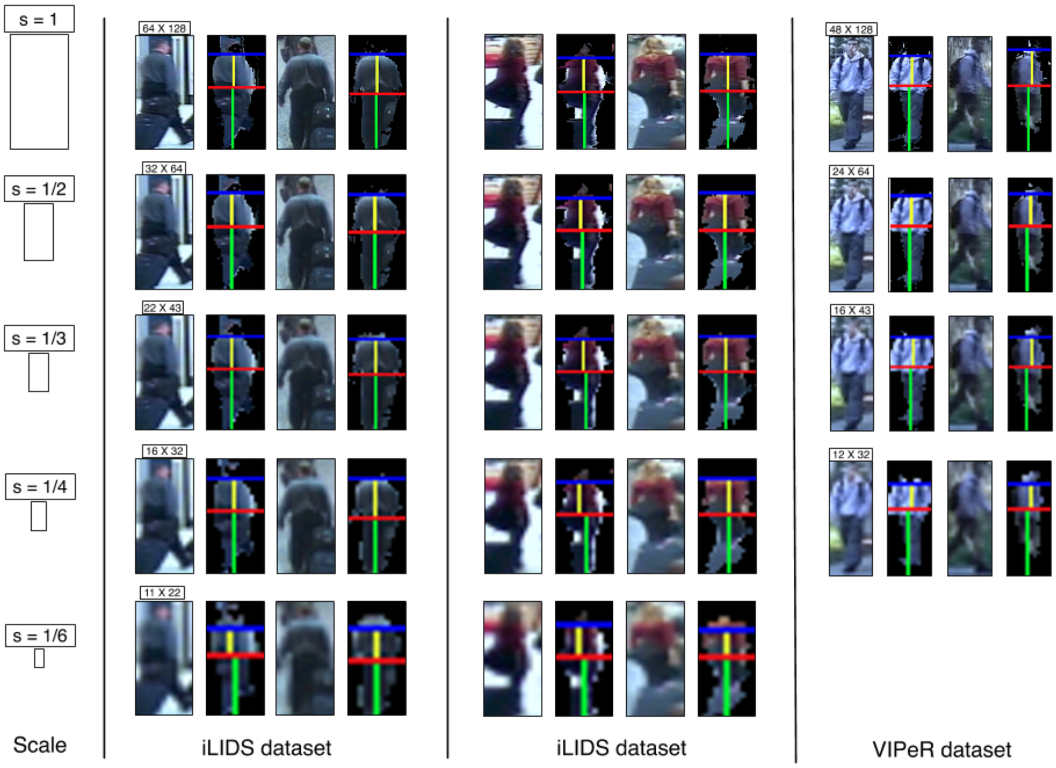


Figure 2.2: Examples of foreground segmentation and symmetry-based partitions (SDALF)

shows excellent robustness when used in conjunction with colour and texture features. The perfromace of SDALF for the CUHK03 dataset and Market-1501 dataset is shown in Table 2.1.

In recent years, some of the hand-crafted feature learning works move from low-level features to mid-level features such as salient regions of the human body. Salient regions are the discriminative areas, which make a person standing out from their companions, as shown in Figure 2.3. These prominent regions provide valuable information for boosting the performance of Person Re-ID models. An innovative approach in this direction is developed by Zhao et al. [145]. Each image is broken down into patches. Features such as Colour Histogram and SIFT are extracted from each patch. These features will be categorised into regular groups and salient groups by the K-Nearest Neighbour(KNN) algorithm. As the salient patches possess an uniqueness property than other regular patches, salient patches can only have a very limited number of neighbours, and each salient patch group is distributed far away from normal patch groups. The distance to the normal patch group will be the score of the salient. The result is shown in Figure 2.4 below. By discovering the salient patches of each individual, more weight will be given to the features extracted from these locations. It improved the robustness of the person re-identification system.

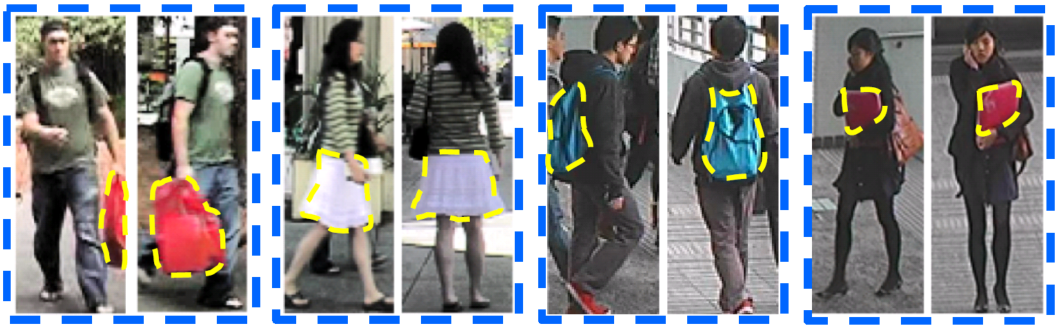


Figure 2.3: A salient region could be a body part or an accessory being carried. Some salient regions of pedestrians are highlighted with yellow dashed boundaries.

This local salient region analysis is based on only one or two features. Similar to other existing techniques [80, 90, 123], those features are pre-defined or globally selected. The weight of each feature is implicitly determined. However, because of different conditions and situations, not all the features are equally useful for person re-identification. Some features such as colour are more discriminative for identity. Features like texture are more tolerant to illumination. The concept of finding the most distinctive region can be further developed for selecting the most salient features of different images at different circumstances. The weighting for each feature for the specific dataset can be learned through boosting [29], ranking [90] or distance metric learning [123]. The top section in Table 2.1 illustrates the Person Re-ID performance for several state-of-the-art hand-crafted feature based methods.

## 2.3 Deep Neural Network Feature

In the previous sections, features extraction methods are designed and hand-crafted by the human. The system designer is telling the system which features the operator

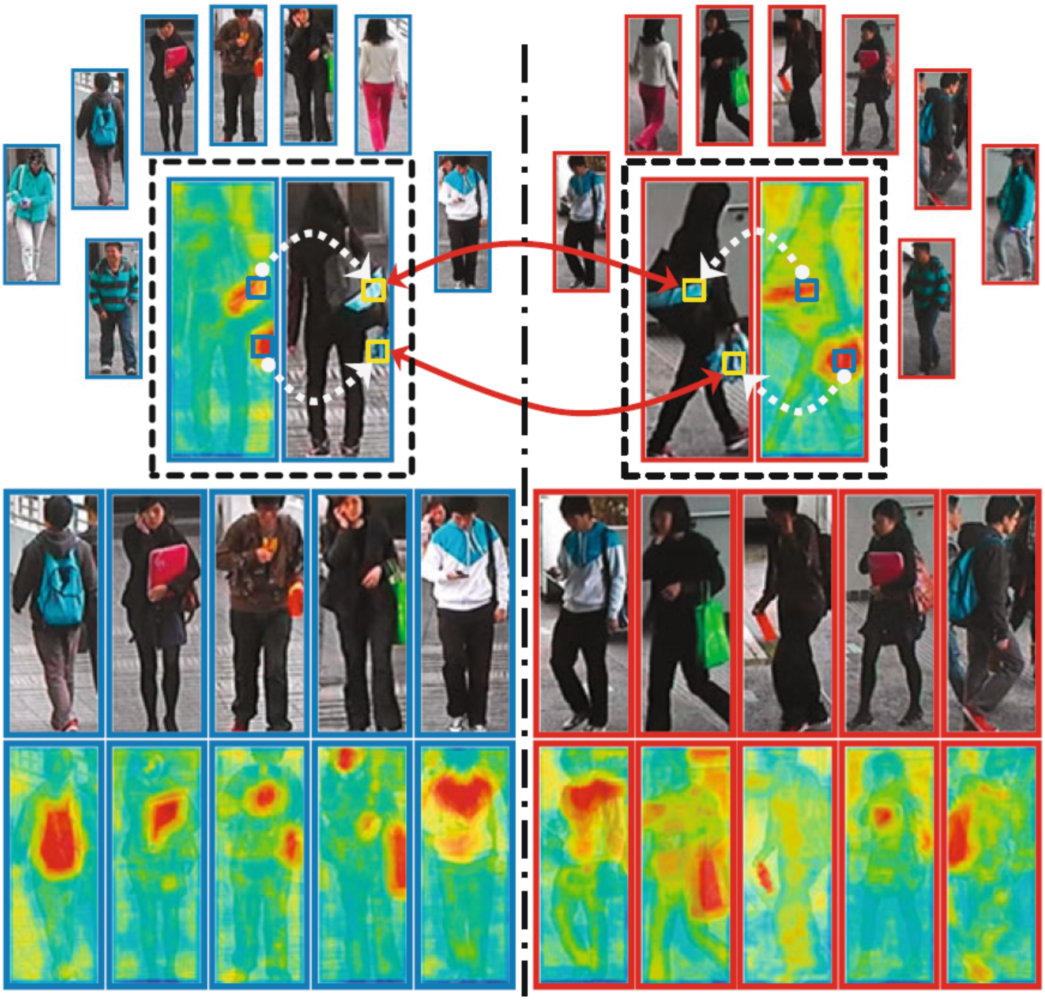


Figure 2.4: A illustration of patch-based person re-identification with salient estimation. The dashed line in the middle divides the images observed in two different camera views. The salience maps of exemplar images are also shown

wants to extract from the pedestrian images. Due to the recent breakthrough in the deep learning area [43], deep neural network structures such as CNN [46] can give the system the ability to learn visual features automatically during the training stage. The emergence of Graphics Processing Units (GPU) and big datasets also help boost the speed and accuracy of the deep learning approaches. In recent years, the deep learning based approaches has been widely used in Person Re-ID area. The bottom section of Table 2.1 demonstrates performance of the Deep Learning based models on the CUHK03 [54] and Mark-1501 dataset [147]. It is clear that the deep

| Methods | | CUHK03 | | | Market-1501 | |
|---|---|---|---|---|---|---|
| | | R-1 | R-5 | R-10 | R-1 | mAP |
| Hand-crafted | SDALF | 4.9 | 21.0 | 31.7 | 20.5 | 8.2 |
| | KISSME | 11.7 | 33.3 | 48.0 | 40.5 | 19.0 |
| | LOMO+XQDA | 46,3 | 79.0 | 88.6 | 43.8 | 22,2 |
| Deep Learning | FPNN | 20.7 | 50.1 | 64.3 | 19.90 | - |
| | IDLA | 45.0 | 75.6 | 83.3 | - | - |
| | PCB | 63.7 | 80.6 | 86.9 | 93.8 | 81.6 |
| | MGN | 66.8 | - | - | 95.7 | 86.9 |

Table 2.1: Performance comparison on CUHK03 and Market-1501 dataset for hand-craft feature approaches and deep feature approaches

learning feature can provide a much better Person Re-ID performance compared to the conventional hand-crafted feature methods

### 2.3.1 Deep Convolutional Neural Network Overview

The advancements in deep learning methods for computer vision tasks have been constructed and imporved with time, primarily over one particular algorithm: the CNN [46]. A CNN model usually consists of one or more convolutional layers (often with a pooling step) followed by one or more fully-connected layers as in a standard multi-layer neural network. The CNN architecture is designed to take advantage of the 2D structure of an input image. It constructs hierarchical connected translation-invariant features directly learn from the training dataset. Besides, CNN models are easier to train and have much fewer parameters compared to fully connected networks with the same number of hidden layers and neurons.

Convolution: The initial layers that receive an input signal are called convolution filters. Instead of assigning different weights per each pixel of an image, some kernel filters that are smaller than the input picture can slide through it. By applying the same set of weights to different parts of the picture (also called weight sharing), the same patterns in different parts of the image can be detected. By connecting multiple convolution layers, the network tries to label the input signal by referring to what it has learned in the past.

Convolution has the nice property of being translation-invariant. Intuitively, this means that each convolution filter represents a feature of interest (e.g from edge detector to eyes and noses). By hierarchically connecting these convolution filters, the CNN algorithm can learn a robust feature combination which can comprise the resulting reference (i.e. face). The output signal strength is not dependent on where the features are located, but on whether the features are present. Hence, a face could be located in different positions, and the CNN algorithm would still be able to recognise it. Moreover, we need to specify other important parameters such as channel depth, stride, and zero-padding. The channel depth corresponds to the number of filters we use for the convolution operation. The more filters we have, the more image features are extracted and the better the network becomes at recognising patterns in unseen images. Stride is the number of pixels (i.e. displacement) by which we slide our filter matrix over the input matrix. When the stride is 1, then we move the filters by one pixel at a time. When the stride is 2, then the filters jump 2 pixels at a time as we slide them around. Having a larger stride will produce smaller feature maps. Sometimes, it is convenient to pad the input matrix with zeros along the border, so that we can apply the filter to bordering elements of our input image matrix. A useful feature of zero padding is that it allows us to control the size of the feature maps.

Pooling: The outputs from the previous convolutional layer need to lower the sensitivity to noise before processing by other operation. A commonly used process in many CNN architectures is pooling (also called sub-sampling). It can be achieved by taking the average or maximum value over a kernel filter. Such spatial pooling reduces the dimensionality of each feature map but retains the most important information. In the case of max-pooling shown in Figure 2.5, we define a spatial neighbourhood with a $2 \times 2$ window and take the largest element from the rectified feature map within that window. Instead of taking the largest element, the average pooling computes the mean value of all elements in that window. In Chapter 4 of this thesis, max-pooling has been shown to obtain better performance.

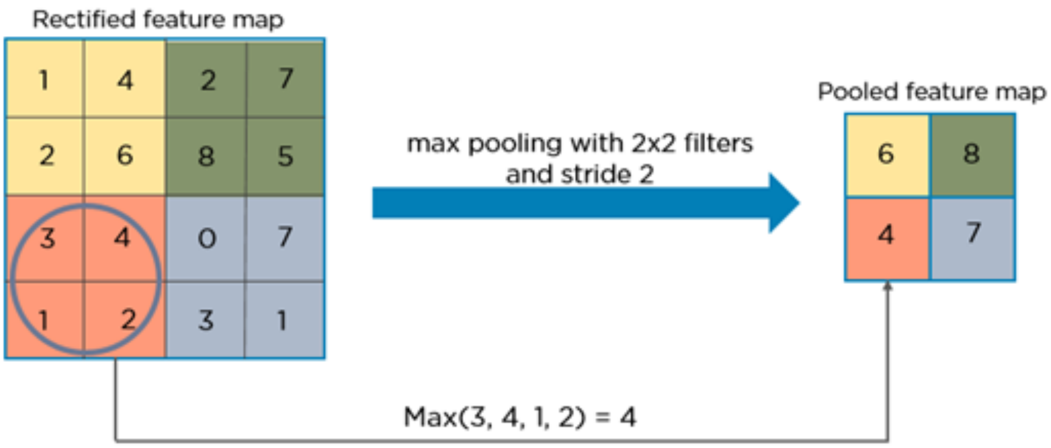


Figure 2.5: The max pooling operation.

Non Linearity Activation: The activation layer controls how the processed signal flows from one layer to the next. It emulates how neurons are activated in the network. The outputs which are strongly associated with past layer would activate more neurons. It enables the signals to be propagated more efficiently for the design task. CNN is compatible with a large variety of complex activation functions to model signal propagation. The most common function is the Rectified Linear Unit (ReLU) [83], which is favoured for its faster training speed. Its output is given by:

$$f(x) = \max(0, x) \tag{2.1}$$

ReLU is an element-wise operation (applied per pixel) and replaces all negative pixel values in the feature map by zero. Figure 2.6 provides a line plot of ReLU for both negative and positive inputs. The purpose of ReLU is to introduce non-linearity in the CNN model since most of the real-world data we would want the network to learn would be non- linear. However, convolution is a linear operation (element-wise matrix multiplication and addition). Hence, we have to account for non-linearity by introducing a non-linear function like ReLU into the network.

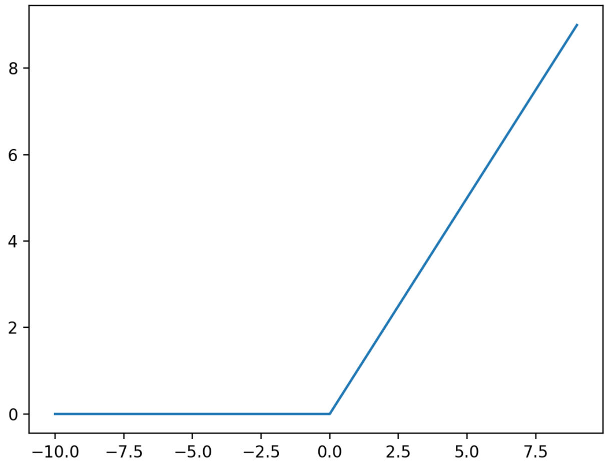
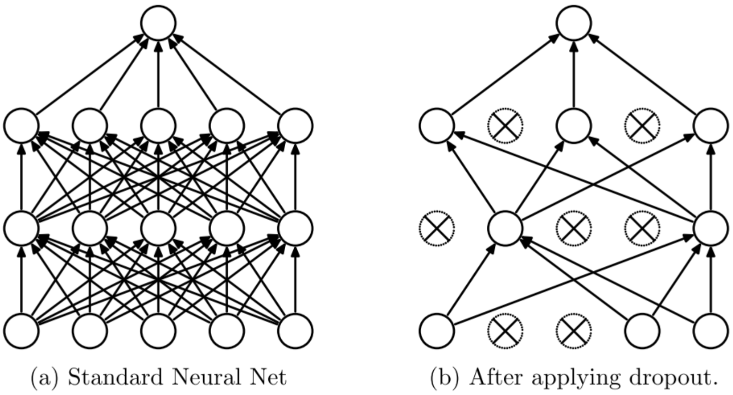


Figure 2.6: A line plot of ReLU for negative and positive inputs

Fully Connected (FC) Layer: The last layers in the network are usually a layer with fully connected neurons. It means that neurons of preceding layers are connected to every neuron in subsequent layers. The outputs from the convolutional and pooling layers represent high-level features of the input images. The purpose of the Fully Connected (FC) layer is to use these features for classifying the input image into various classes. Apart from the classification purpose, adding a fully-connected layer is also an easy method to learn non-linear combinations of these features.

Dropout: Dropout is a popular regularisation technique for neural network models proposed by Srivastava et al. [104]. A fully connected layer makes the neurons co-dependent to each other, which suppresses the individual power of each neuron leading to over-fitting of training data. The dropout mechanism can randomly select neurons and ignore them during the training phase, as shown in Figure 2.7. By shutting down neurons randomly, it can prevent the network over-reliant on a few active neurons. Each neuron has the opportunity to leans a useful feature representation and overall improving the generalisation ability of the network. In many deep learning approaches, dropout is usually applied before the final fully connected layer to alleviate the over-fitting problem. A typical deep convolutional neural network (CNN) architecture usually consists multiple groups of convolution, non-linearity activation and spatial pooling follow by one or two Fully Connected (FC) Layers with Dropout, as shown in Figure 2.8.



(a) Standard Neural Net (b) After applying dropout.

Figure 2.7: Dropout Neural Net Model. Left: A standard neural net with 2 hidden layers. Right: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped. (figure provided in [104])

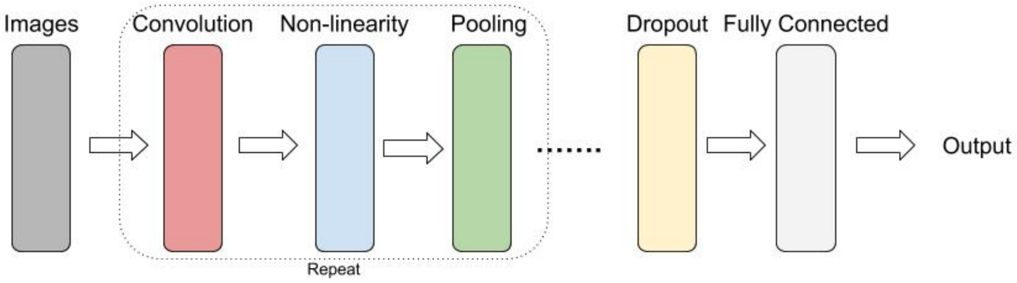


Figure 2.8: Key operations in typical deep CNN architectures

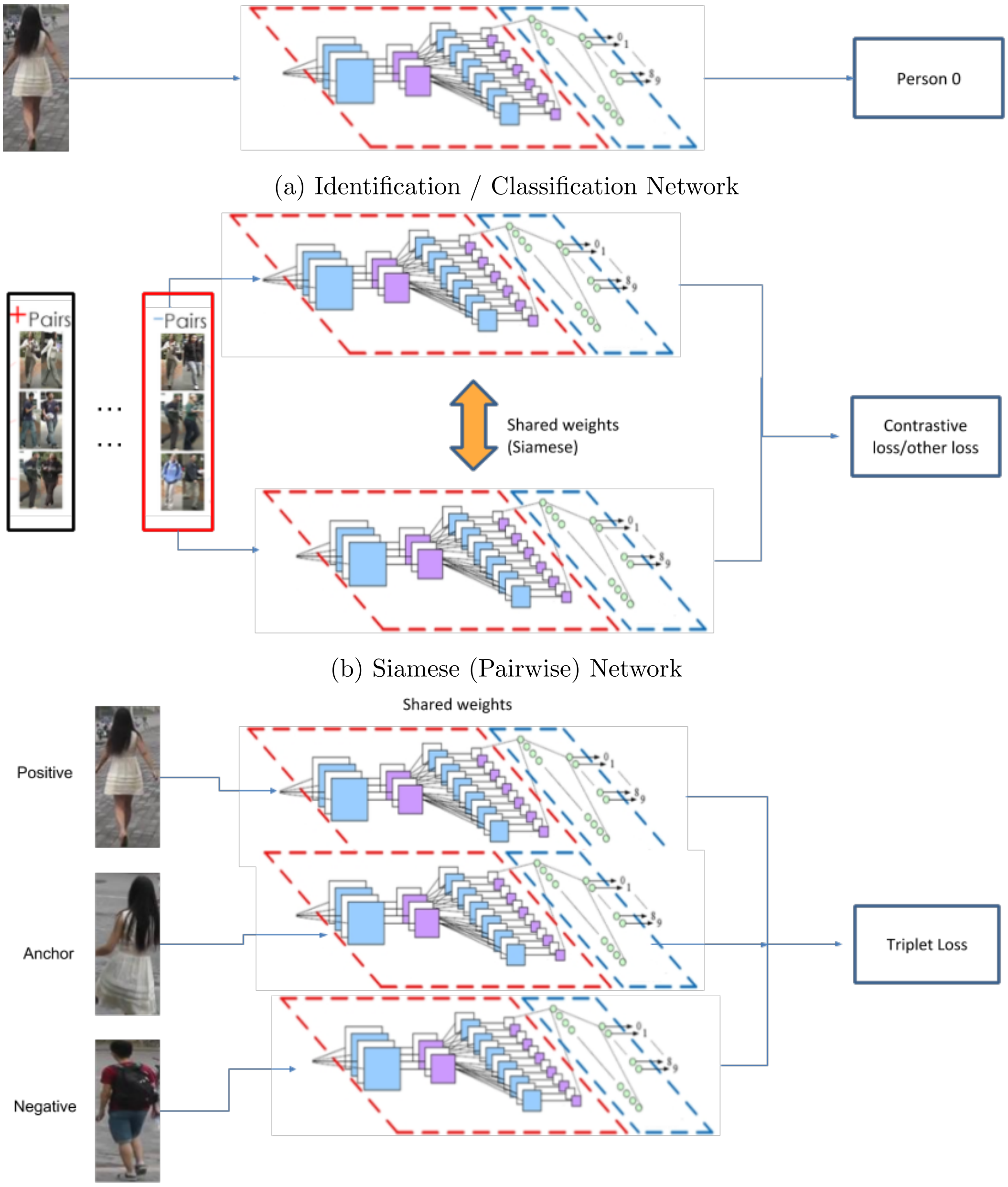
### 2.3.2 Single-Dataset Deep Feature Learning

Since Krizhevsky et al. [43] won ILSVRC12 by using a CNN-based model, deep learning approaches have been widely used in various computer vision tasks. Several deep learning methods such as ”FPNN”[54] and ”IDLA”[1] are developed based on implementing the deep neural network into the person re-identification system for features extraction and matching tasks [54]. Generally speaking, most of supervised deep learning based approaches can be categorised in to two types of CNN structures.

The first is the classification model as used in image classification [43] and object detection [25], as shown in Figure 2.9a. The second type is the Siamese model using image pairs [92] or triplets [96] as inputs, as shown in Figure 2.9b and Figure 2.9c.

### Models based on Siamese Network

In the early stage, Siamese network models have been widely employed due to lack of training instances. Unlike traditional networks, Siamese networks are modelled in a pairwise setting, i.e. inputs are taken as pairs as opposed to single inputs in other conventional networks. The Siamese neural network contains two or more subnetworks which share the same network architecture and the same weight parameters, as shown in Figure 2.9b and Figure 2.9c. The objective of most of the Siamese networks [9, 32] is to learn an embedding such that the same class objects are closer to each other with the different classes objects are far apart. Hence, Hadsell et al. [32] proposed the contrastive loss function to learn an invariant mapping. The objective of this loss function is to separate objects belonging to different classes by a margin distance while keeping the images of objects belonging to the same class as close as possible in the embedding space.



(a) Identification / Classification Network

(b) Siamese (Pairwise) Network

(c) Siamese (Triplet) Network

Figure 2.9: Different types of CNN structure for Person Re-ID

The first Siamese architecture for person re-identification was proposed by Yi et al. [138]. In [138], the network consists of a set of 3 CNNs for different regions of the image and the features are combined by using the cosine similarity as the connection function. Binomial deviance is used as the cost function to optimise the network end-to-end. Local body-part based features and the global features were modelled by using a multi-channel CNN framework in [11]. Deep Filter Pairing Neural Network (FPNN) was introduced in [54] to jointly handle misalignment, photo-metric and geometric transformations, occlusion and cluttered background. Later, the IDLA method Ahmed et al. [1] improved the FPNN by introducing a cross-input neighbourhood difference module to extract the cross-view relationships of the features and have achieved impressive results in several benchmark datasets. [116] also attempts to model the cross-view relationships by jointly learning sub-networks to extract the single image as well as the cross image representations. In [98], a Siamese network takes a CNN learning feature pair and outputs the similarity value between them by applying the cosine and Euclidean distance functions. This CNN framework employed to obtain deep feature of each input image pair, and then, each image is split into three overlapping colour patches. The deep network built in three different branches and each branch takes a single patch as its input. Finally, the three branches are concluded by an FC layer. One recent work named Pyramid Person Matching Network (PPMN) [78] proposed a two-channel convolutional neural network with the new Pyramid Matching Module component. The Pyramid Matching Module aimed to learn the corresponding similarity between semantic features based on multi-scale convolutional layers. In Chapter 3 of this thesis, we also utilise the Siamese network structure to create an end-to-end mid-level deep features correspondence learning, which specially designed for real-time person re-identification and matching [62]

#### Models based on Classification

One biggest drawback of the Siamese model is that it does not make full use of person id annotations. In fact, the Siamese model or triplet model only needs to consider pairwise (or triplet) labels. Telling whether an image pair is similar (belong to the same identity) or not is a relatively weak label for training a deep neural network. Due to the recent release of two large-scale training set: Market-1501 [147] and DukeMTMC-reID [152], the most recent works start to utilise the identity labels and train the Person Re-ID model in the classification/identification setting.

Zheng et al. [149] and [148] directly use conventional fine-tuning approaches on the Image-Net [43] pre-trained classification network. The CNN embeddings from their classification networks can outperform many Siamese structured methods. The PDC method [107] then integrate the identification with pose-estimation in order to handle the misalignment and pose variations of pedestrians images. The APR method [65] integrated the identity information with attributes to obtain more robust feature representations. The Embedding method [151] combined both the identification loss from the person ID classification network and verification loss from the Siamese network. Currently, most of the state of the art methods such as MGN [117] and BFE [14] use both identification loss and verification loss in order to achieve the best performance. In Chapter 4, we developed a two-stage baseline network tailored to the offline person retrieval task. In the proposed model, we utilise both the identification loss and verification loss integrated with various training techniques from multiple state-of-the-art approaches. We also improve the triplet loss function by introducing a batch-based adversarial competing mechanism to enhance the discriminability of the feature embedding. By using our proposed triplet loss function with many training techniques and data refinements, our simple and effective person re-identification model can achieve state-of-the-art performance.

### 2.3.3 Cross-Dataset Feature Transfer Learning

A brief overview of the related works in supervised single-dataset person re-identification is presented in Section 2.3.2. However, in the real-world Person Re-ID system deployment, supervised methods usually suffer from poor scalability due to the lack of training dataset obtained from the new system. Therefore, some unsupervised Person Re-ID methods have been developed based on hand-crafted features with dictionary learning. [40, 41, 118, 145]. Kodirov et al. [40] proposed to formulate unsupervised Person Re-ID as a sparse dictionary learning problem. To regularise the learned dictionary, they utilise graph Laplacian regularisation and iteratively updated the graph Laplacian matrix. Later, they introduced an $l_1$-norm graph

Laplacian to learn the graph and the dictionary jointly [41]. Wang et al. [119] use a kernel subspace learning model to learn cross-view identity-specific information from unlabeled data. Yang et al. [135] propose a weighted linear coding method to learn multi-level descriptors from raw pixel data in an unsupervised manner. These unsupervised methods, due to the absence of the pairwise identity labels, cannot learn robust cross-view discriminative features and usually yield much weaker performance compared to the supervised learning approaches.

Because of the poor person Re-ID performance of the single dataset unsupervised learning, many of recent works are focusing on developing the cross-dataset transfer learning methods [16, 45, 73, 89, 120]. These approaches leverage the pre-trained supervised Re-ID models and adapt these models to the target dataset. Early proposed cross-dataset person Re-ID domain adaptation approaches rely on weak label information in target dataset [45, 73]. Therefore, these methods can only be considered as semi-supervised or weakly-supervised learning. The recent cross-dataset works such as UMDL [89], SPGAN [16] and TJ-AIDL [120] do not require any labelled information from the target dataset and can be considered as fully unsupervised cross-dataset domain adaptation learning. The UMDL method [89] tries to transfer the view-invariant feature representation via multi-task dictionary learning on both source and target datasets. The SPGAN approach [16] uses the generative adversarial network (GAN) to generate new training dataset by transferring the image style from the target dataset to the source dataset while preserving the source identity information. Hence, the supervised training on the new translated dataset can be automatically adapted to the target domain. The TJ-AIDL approach [120] individually trains two models: an identity classification model and an attribute recognition model. The domain adaptation in TJ-AIDL is achieved by minimising the distance between inferred attributes from the identity classification model and the predicted attributes from the attribute recognition model. In [154], Zhong et al. introduced a Hetero-Homogeneous Learning (HHL) method, which aims to improve the generalisation ability of Person Re-ID models on the target set by achieving camera invariance and domain connectedness simultaneously. Compared to the previous single dataset unsupervised approaches, the recent cross-dataset unsupervised domain adaptation methods yield much better performance. However, the performance is still unsatisfactory compared with fully supervised approaches. In Chapter 5 of this thesis, we proposed a Multi-task Mid-level Feature Alignment (MMFA) Network [64]. With an assumption that the source and target datasets share the same set of mid-level semantic attributes, our proposed model can be jointly optimised under the people identity classification and the attribute learning task with a cross-dataset mid-level feature alignment regularisation term. In this way, the learned feature representation can be better generalised from one dataset to another, which further improve the person re-identification accuracy. Experimental results on four benchmark datasets demonstrate that our proposed method outperforms the state-of-the-art baselines.

### 2.3.4 Multi-Datasets Domain Generalisation

Cross-Dataset domain adaptation increases the practicality of any pre-trained Person Re-ID model deployed into any unseen system. However, it requires an additional unsupervised or semi-supervised fine-tuning process on the target datasets. The ultimate goal of an effective Person Re-ID models should generalised to any new unseen dataset/system. It has great value for real-world massive- scale deployment. Specifically, when a customer purchases a Person Re-ID system for a specific camera network, the system is expected to work out-of-the-box, without the need to go through the tedious process of data collection, annotation and model updating or fine-tuning. One possible way to achieve this goal is to utilise multiple existing datasets to generalise a domain invariant feature representation. Surprisingly, there is a very little prior study of this topic. To the best of author's knowledge, only two related works were specifically designed to using multiple datasets for models generalisation purpose.

Xiao et al. [128] proposed learning deep features representations from multiple data sets by using CNNs to discover effective neurons for each training data set. They first produced a strong baseline model that works on multiple data sets simultaneously by combining the data and labels from several Person Re-ID data sets together and trained the CNN with a softmax loss. Next, for each data set, they performed the forward pass on all its samples and computed the average impact of each neuron on the objective function. Then, they replaced the standard Dropout operation with the deterministic Domain Guided Dropout in order to discard useless neurons for each data set, and continue to train the CNN model for several more epochs. The learned generic embedding after domain-guided drop out yields competitive Person Re-ID accuracy. Song et al. [102] on the other hand, proposed a Domain-Invariant Mapping Network (DIMN) which produces a classifier using a single shot from the source dataset. Once learned, for a target dataset, each gallery image is fed into the network to generate the weight vector of a specific linear classifier for the corresponding identity. A probe image is then be matched using the classifier by computing a simple dot product between the weight vector and a deep feature vector extracted from the probe. They follow a meta-learning pipeline and sample a subset of source domain training tasks (identities) during each training episode for the domain-invariant purpose. In Chapter 6, we proposed a new novel domain generalisation structure (MMFA-AAE). The proposed network is based on adversarial auto-encoders to learn a generalised latent feature representation across camera domains with Maximum Mean Discrepancy (MMD) measure to align the distributions cross datasets. Extensive experiments on both single-dataset and cross-dataset demonstrate the effectiveness of the proposed method.

### 2.3.5 Deep Learning Person Re-ID Methods Summary

Table 2.2 shows the performance the state-of-the-art deep learning Person Re-ID methods which discussed in the previous sections.

| Methods | | VIPeR | PRID | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|---|---|---|
| | | R-1 | R-1 | R-1 | mAP | R-1 | mAP |
| Single-Dataset | IDLA | 34.8 | - | - | - | - | - |
| | APR | - | - | 84.3 | 64.7 | | |
| | PDC | 51.2 | - | 84.1 | 63.4 | - | - |
| | Embedding | - | - | 79.5 | 59.9 | - | - |
| | MGN | - | - | 95.7 | 86.9 | 88.7 | 78.4 |
| | BFE | | | 94.4 | 85.0 | 88.8 | 75.8 |
| Cross-Dataset | TJ-AIDL (Market) | 38.5 | 26.8 | - | - | 44.3 | 23.0 |
| | SPGAN (Market) | - | - | - | - | 41.1 | 22.3 |
| | TJ-AIDL (Duke) | 35.1 | 34.8 | 58.2 | 26.5 | - | - |
| | SPGAN (Duke) | - | - | 51.1 | 22.8 | - | - |
| Multi-Dataset | DIMN | 51.2 | 39.2 | - | - | - | - |
| | DualNorm | 53.9 | 60.4 | - | - | - | - |

Table 2.2: Summary on the deep learning Person Re-ID methods

## 2.4 Datasets and Evaluation Protocols

### 2.4.1 Person Re-identification Datasets

This section contains a list of the datasets used for training and testing person re-identification systems. In order to properly evaluate Person Re-ID models, a good person re-identification dataset has to mirror the actual video surveillance setting in a real-world scenario: viewpoint changes, differences in illumination, differences in background and camera characteristics. A realistic dataset should include images taken from different surveillance cameras to capture the same identity from different viewpoint and trajectories, demonstrated in Figure 2.10. However, due to the security concern and privacy issue, most existing person re-identification datasets are collected from point-and-shoot cameras mounted on tripods. In this section, we provide a brief description of 6 famous and commonly used public Person Re-ID datasets.

#### VIPeR

The VIPeR dataset [28] is one of the oldest Person Re-ID dataset. It contains 632 identities taken from two camera views with pose and illumination changes. Due to its low resolution and large variation in illumination and viewpoints, the VIPeR

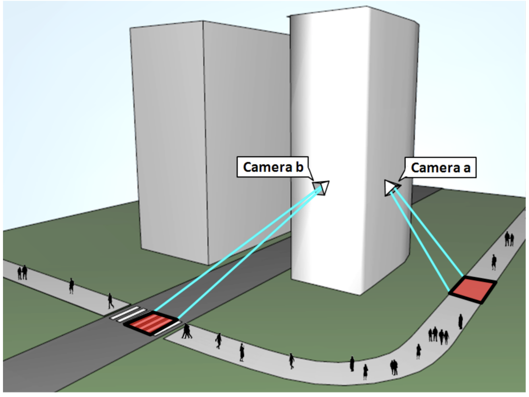


Figure 2.10: Camera setup for a person re-identification dataset construction.

dataset is one of the most challenging and widely used datasets for Person Re-ID model evaluation. The images are all cropped and scaled to be 128 × 48 pixels. Figure 2.11 provide sample images of three different identities in the dataset. Each identity has two images under two camera views.

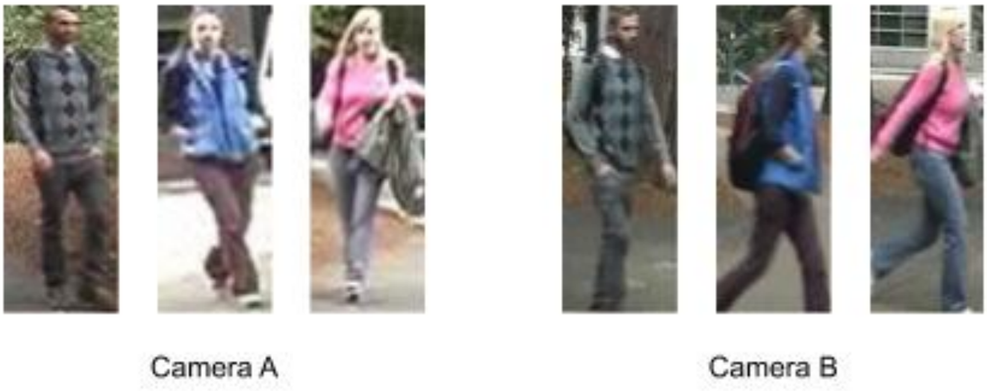


Figure 2.11: Sample images from the VIPeR dataset. Each identity has two images under two camera views.

### PRID

The PRID dataset [35] is specially designed for Person Re-ID focusing on the single-shot scenario. It consists of 934 identities generated from two camera views. There are 385 identities in View A and 749 identities in View B, but only 200 identities appear in both views. The images are cropped and resized to be 128 × 64 pixels.

Figure 2.12 provides some sample images selected from 200 common identities in the PRID dataset. By analysing the image quality in Figure 2.12, the colour profiles of the two capture cameras are very different. Due to the large colour profile difference between cameras, the PRID is also a challenging dataset.

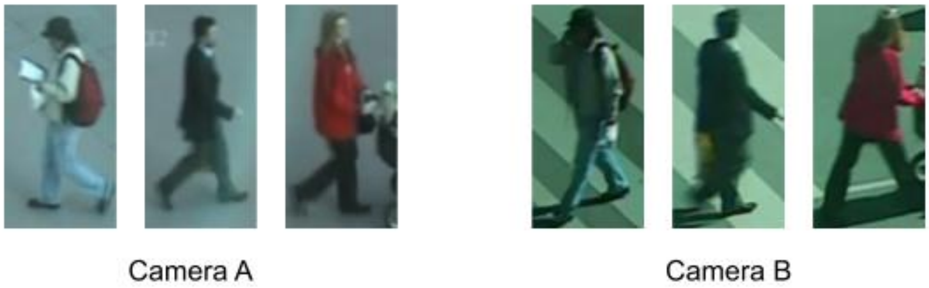


Figure 2.12: Sample images selected from 200 common identities of the PRID dataset. the common identities has images under two camera views.

### CUHK

The CUHK datasets are collected by Chinese University of Hong Kong. It contains three different partitions: CUHK01 [53], CUHK02 [52], and CUHK03 [54]. The CUHK01 dataset includes 1,942 images of 971 pedestrians. It has only two images captured from two disjoint camera views, as shown in Figure 2.13. Camera A mainly used for capturing the frontal view and the back view of each identity. Camera B has more variations of viewpoints and poses. All images are re-sized to 160 × 60 pixels.

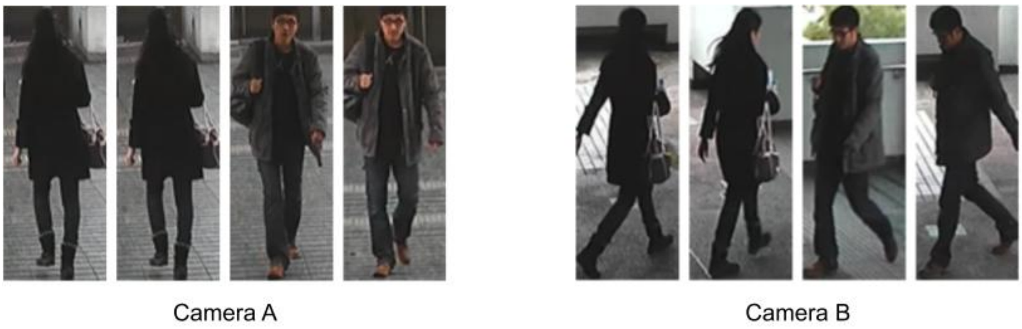


Figure 2.13: Sample images from the CUHK01 dataset.

The CUHK02 dataset contains 1,816 individuals grouped by five pairs of camera views (P1-P5 with ten camera views). Each camera pair includes 971, 306, 107, 193 and 239 individuals, respectively. Each individual has two images in each camera view. Similar to CUHK01, all images in CUHK02 are re-sized to $160 \times 60$ pixels. This dataset is also the first Person Re-ID work, which evaluates the performance when the camera views in the test are different from those in training. Some sample images in each pair of camera views are illustrated in Figure 2.14.

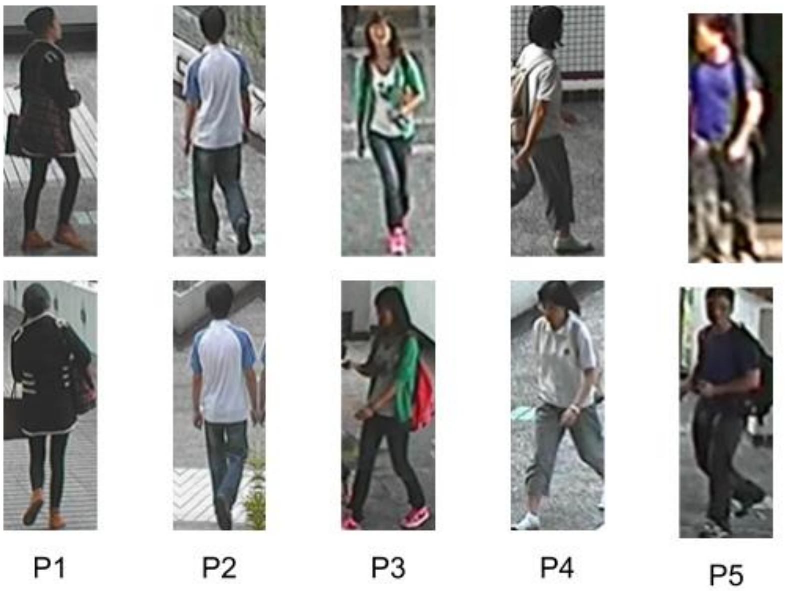


Figure 2.14: Sample images from the CUHK02 dataset from 5 pairs of camera views

Finally, CUHK03 is the first person re-identification dataset that is large enough for training a deep learning model. It includes 13,164 images from 1,360 pedestrians. Each identity is observed by two disjoint camera views and has an average of 4.8 images in each view, as shown in Figure 2.15. Unlike CUHK01 and CUHK02, the CUHK03 dataset contains images with various image sizes.

The CUHK03 is also the first dataset which utilises an automatic person detection algorithm for detecting the bounding boxes of the people. It is composed of two versions with the same identities: *CUHK03-labelled* and *CUHK03-detected*. The *CUHK03-labelled* is composed by bounding boxes manually cropped like other Person Re-ID dataset mentioned above. The bounding boxes in the *CUHK03-detected* are

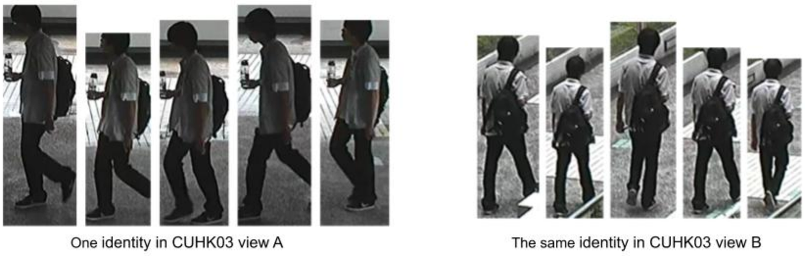

One identity in CUHK03 view A
The same identity in CUHK03 view B

Figure 2.15: Sample images of one identity in the CUHK03 dataset.

detected by using the Deformable Part Models detector (DPM) [20]. The differences are showed in Figure 2.16. Due to the imprecision of the DPM detector with respect to the manually cropping, manually cropped pedestrian images exhibit illumination changes, misalignment, occlusions and body part missing. Hence, *CUHK03-detected* is more challenging compared to *CUHK03-labelled*. Since CUHK02 and CUHK03 are very similar in the data collection setting, most of the Person Re-ID methods only train and evaluated on the more challenging CUHK03 dataset.

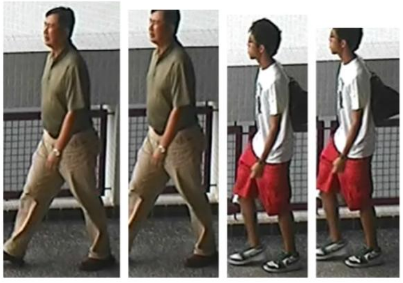


Figure 2.16: In each pair on the left the image manually cropped, on the right the image automatically detected.

### Market-1501

The Market-1501 dataset [147] contains 32,668 images of 1,501 pedestrians. 751 identities are selected for training, and 750 remaining identities are for testing. Each identity was captured by at most 6 non-overlapping cameras. It also uses the Deformable Part Models detector (DPM) to detect the bounding box of person automatically. All images are re-sized to 128 × 64 pixels. Figure 2.17 gives some sample images of the Market1501 dataset.

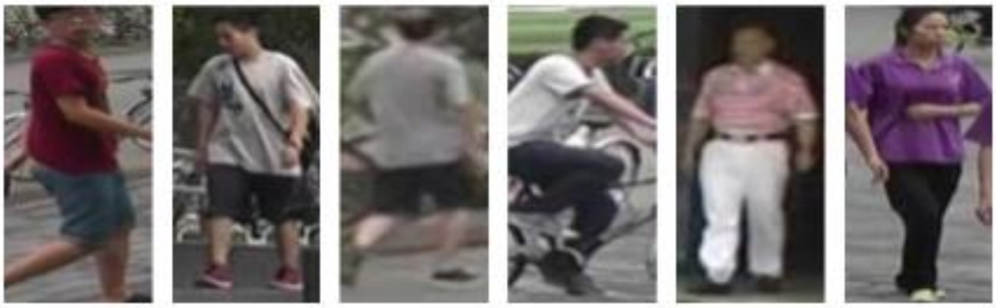


Figure 2.17: Sample images from the Marekt1501 dataset.

**DukeMTMC-reID**

The DukeMTMC-reID dataset [152] is the redesign version of pedestrian tracking dataset DukeMTMC [95] for Person Re-ID task. It is one of the few Re-ID datasets collected from actual surveillance cameras. The DukeMTMC-reID dataset contains 34,183 image of 1,404 pedestrians. 702 identities are used for training, and the remaining 702 are for testing. Each identity was captured from 8 non-overlapping cameras. Figure 2.18 gives some sample images of the DukeMTMC-reID dataset. Due to unconstrained image size, a large number of camera, large illumination changes and occlusions, the DukeMTMC-reID is much more challenging compared to the Matket1501 dataset.

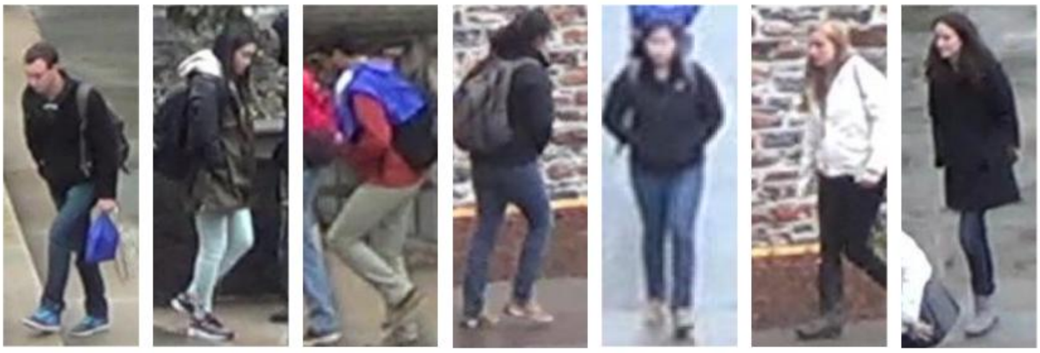


Figure 2.18: Sample images from the DukeMTMC-reID dataset.

**Summary**

Table 1 below provides a statistical information and characteristics summary of each dataset. The number of identities and the total number of images increase significantly over the year. However, the number of cameras did not increase as much. The 6-8 cameras cannot fully represent the actual camera number in the real-world surveillance systems. As a result, we collected a new large-scale Person Re-ID Dataset: Re-ID-Outdoor. This new dataset is collected from a total of 50 cameras cover sophisticated scene transformations, background changes and illumination variance.

| Dataset | Year | # Identities | # Cameras | # Images | Label Method | Crop Size |
|---|---|---|---|---|---|---|
| VIPeR | 2007 | 632 | 2 | 1,264 | Manual | 128X48 |
| PRID | 2011 | 934 | 2 | 1,134 | Manual | 128X64 |
| CUHK01 | 2012 | 971 | 2 | 3,884 | Manual | 160X60 |
| CUHK02 | 2013 | 1,816 | 10 (5 Pairs) | 7,264 | Manual | 160X60 |
| CUHK03 | 2014 | 1,467 | 10 (5 Pairs) | 13,164 | Manual/DPM | Vary |
| Market-1501 | 2015 | 1,501 | 6 | 32,217 | Manual/DPM | 128X64 |
| DukeMTMC-reID | 2017 | 1,812 | 8 | 36,441 | Manual | Vary |
| Re-ID-Outdoor | 2020 | 805 | 50 | 67,050 | YOLO V3 | Vary |

Table 2.3: Summary on benchmark person re-identification datasets

## 2.4.2 Evaluation Protocols

The cumulative matching characteristics (CMC) curve is the most common metric used for evaluating Person Re-ID performance. This metric is adopted since Person Re-ID is intuitively posed as a ranking problem. Each image in the gallery is ranked based on its comparison to the probe. The probability that the correct match in the ranking equal to or less than a particular value is plotted against the size of the gallery set [28]. Due to the slow training time of deep learning models, the CMC curve comparisons for recent deep Re-ID methods are simplified to only comparing Rank 1, Rank 5, Rank 10, Rank 20 retrieval rates. Figure 2.19 illustrate the conversion between CMC curve to Rank 1, 5, 10. Rank 1,5,10 can be considered as a simplified version of the CMC curve.

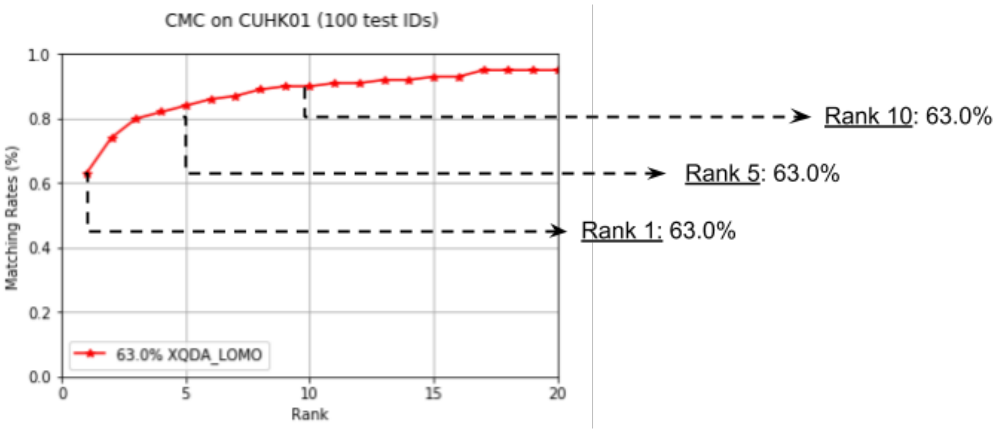


Figure 2.19: CMC curve to Rank 1, 5, 10 conversion

However, the CMC curve evaluation is valid when only one ground truth match for each given query image. The recent datasets such as Market-1501 and DukeMTMC-reID usually contain multiple ground truths for each query images. Therefore, Zheng et al. [147] have proposed the mean average precision (mAP) as a new evaluation metric. For each query image, the average precision (AP) is calculated as the area under its precision-recall curve. The mean value of the average precision (mAP) will reflect the overall recall of the Person Re-ID algorithm. The performances of current Person Re-ID methods are usually examined by combining the CMC curve for retrieval precision evaluation and mAP for recall evaluation.

## 2.5 Concluding Remarks

In this chapter, a brief overview of the related works in person re-identification is presented. It covers from the early hand-crafted feature engineering to the recent deep learning methods. As the deep learning methods show a superior performance compared to hand-crafted feature methods, as shown in Table 2.1. The methods proposed in this thesis are all based on deep convolutional neural network. This thesis focuses on building real-world applications for person re-identification. We specially designed two different models for the real-time person matching application and offline person retrieval application. As most of the existing Person Re-ID models are trained from a single dataset (a single camera system). Most of the existing Person Re-ID models [1, 14, 54, 117] are suffered from dataset over-fitting and show very limited generalisation ability to other camera system. Hence, to address the scalability problem in the existing Person Re-ID models, this these also proposes two methods via a domain adaptation approach and a domain generalisation approach. In addition, all the existing Person Re-ID datasets [54, 122, 147, 152] contains very limited number of cameras, as shown in Table 2.3. We also create a new real-world sized Person Re-ID dataset: Re-ID-Outdoor. Out dataset also address the privacy issue in all existing Person Re-ID dataset by following the European General Data Protection Regulation (GDPR).

# Chapter 3

# Single-Dataset Feature Learning (Online Matching)

## 3.1 Introduction

Person Re-ID applications can be categorised into two types: **online person matching** and **offline person retrieval**. Online person matching across different cameras is the fundamental procedure for the real-time person tracking in a multi-camera CCTV system. Offline person retrieval, on the other hand, does not have the processing time constraints. It is concern more with the feature embedding extraction, storage and ranking, rather than meeting the real-time requirement. This chapter focuses on developing a Person Re-ID model for the real-time person matching applications.

In order to automatically track a person in a video surveillance network, the system should be able to quickly and correctly match the same person across multiple cameras and assign a consistent ID to him/her. Figure 3.1 demonstrates the process of the online person matching. The probe image can be a picture of the suspect uploaded manually by the operator or a bounding box of a person obtained from other surveillance cameras. The online matching Person Re-ID model should generate the similarity scores for every person in the current video frame based on

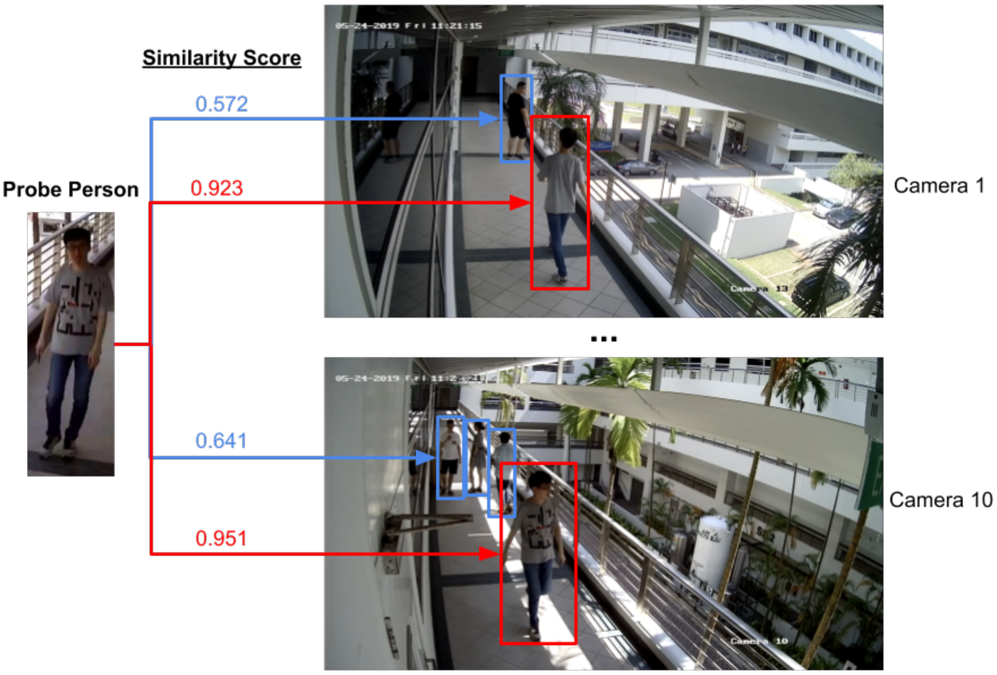
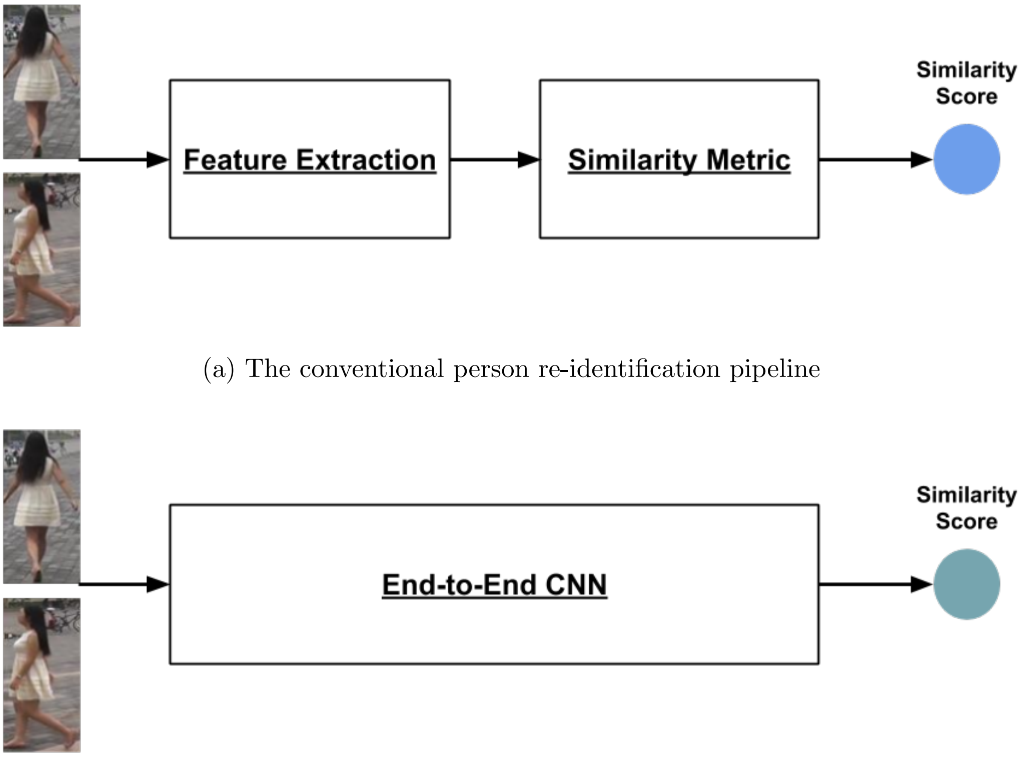


Figure 3.1: This figure shows the online person matching process. The red bounding boxes are the correct matchings with high similarity scores. The blue bounding boxes are different persons with low similarity scores.

the target person's appearance. In Figure 3.1, the red bounding boxes indicate the correct matching persons with over 0.9 similarity score. The remaining pedestrians in the blue bounding boxes only show less than 0.6 similarity score in our model. Base on the similarity score difference, we can easily separate the probe person from other pedestrians.

Many Person Re-ID models [11, 14, 98] use a two-stage pipeline: 1) feature extraction stage and 2) similarity measuring stage, as shown in Figure 3.2a. However, online person re-identification applications require the matching process completed in real-time. To maximise the processing speed, we propose an end-to-end mid-level deep feature correspondence learning network which merges the feature extraction and metric learning stages into one single network. The proposed network can produce the similarity score directly from an image pair, as shown in Figure 3.2b.



(a) The conventional person re-identification pipeline

(b) The end-to-end person re-identification pipeline

Figure 3.2: Different types of person re-identification pipelines

## 3.2 Problem Definition

A practical real-time Person Re-ID model should learn a robust feature representation which needs to be invariant to different camera viewpoints, illumination or human's poses. There are many Person Re-ID models developed by exploiting low-level features such as colour [44, 60, 134, 144, 145, 147], texture [75, 144, 145], salient region [144, 145] or spatial structure [6]. However, these low-level visual features are not robust to variations in illumination, viewpoint, misalignment, *etc.* In human perception, different people can be easily recognised by their mid-level visual features such as gender, hair length, clothing colours or additional accessories. These attributes can represent the mid-level semantics of a person which are more robust to misalignment and camera variation comparing to low-level local features. However, manual annotation of these mid-level semantics features can be very expensive and time-consuming for a large camera network. As a result, it is difficult to acquire enough training data with a large set of attributes.

Our proposed method uses an alternative approach to obtain the mid-level features. In recent years, deep CNN have contributed to a significant improvement in performance in solving many computer vision tasks. There are also many studies analysing the features obtained by CNNs. As the features from a CNN architecture are structured in a hierarchical nature, the lower layer behaves in a manner similar to low-level local feature extractors such as edge or colour filters. At higher layers, the features start showing significant variation and become more class-specific [140]. In our proposed method, we use the feature maps obtained from the mid-layer of the CNN architecture as an alternative to the actual mid-level semantic attributes. By finding the correspondence between the feature maps, the network can be trained to capture the most distinctive features of a person. Many existing approaches focus on constructing the correspondence distributions between each pair of the same feature map between the probe image and gallery images [13, 53]. However, we contend that the mid-level feature correspondences should not be limited to the regional feature map matching. The potential relationship between these mid-level features should also be taken into consideration. In this chapter, we proposed a new strategy for establishing a feature correspondence by considering different combinations of mid-level deep features. In our proposed network, each correspondence feature is not limited to the correlation between feature maps obtained from two images, but from the multiple feature maps of two images.

One example of our system prediction result is shown in Figure 3.3. By using the Inception network [110] as the mid-level feature extractor, the proposed method can adaptively discover the distinctive mid-level deep features. The similarity scores are calculated by analysing the relationship between these correspondence features. These mid-level features and their latent relationships are learned through a data-driven approach. Furthermore, as the parameters are initialised from an ImageNet pre-trained model, the training process of our network can be considered as a fine-tuning process for transforming the deep mid-level features from an object classification problem into a similarity matching task. As a result, it improves the discriminative power and generalisation ability of these features.

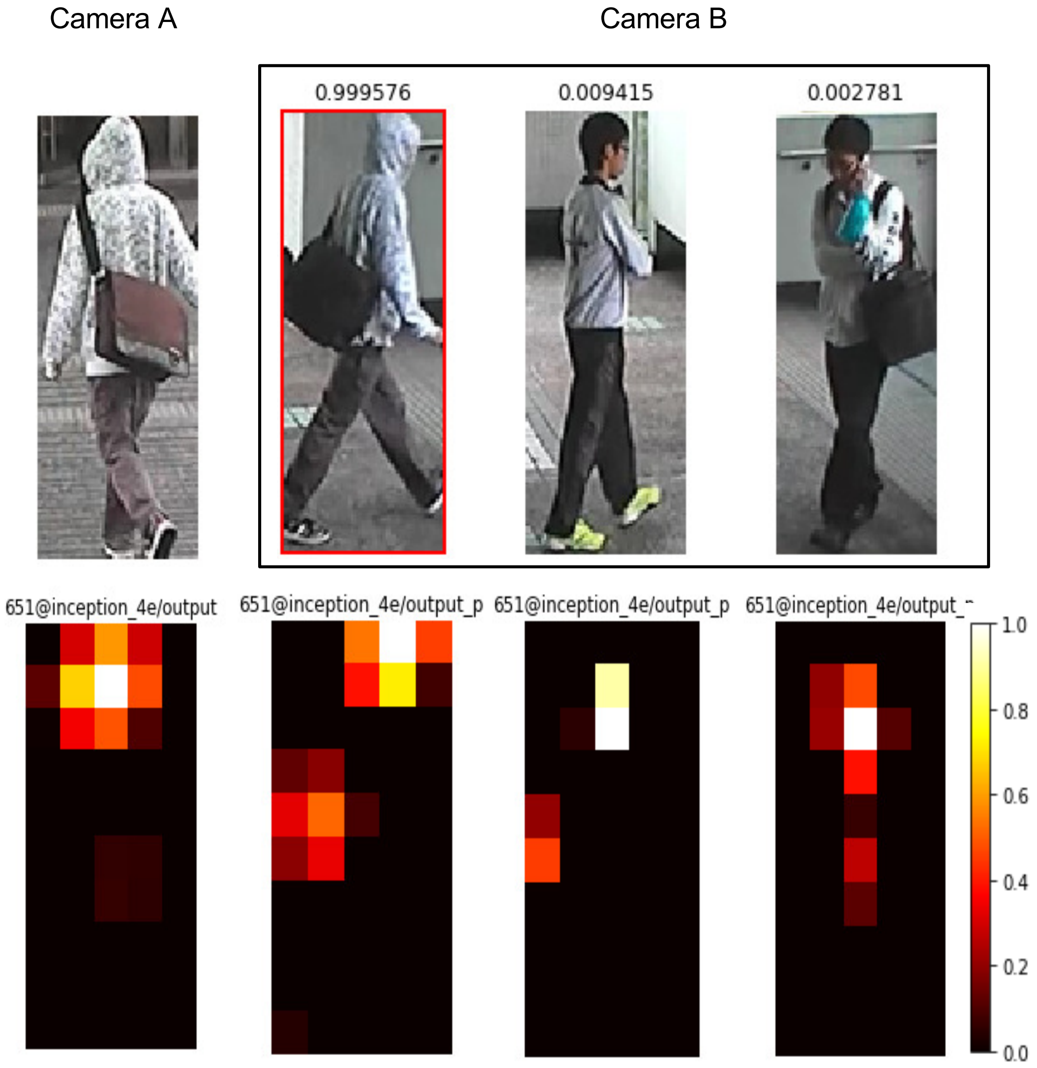


Figure 3.3: This figure shows one of our predicted results in the CUHK01 dataset. The ground-truth images are marked by the red bounding boxes. The second row shows one mid-level feature map with the highest activation value for this person obtained by our network. The highlighted white hoodie hat region is one of the distinctive mid-level features to re-identify the person with the probe image.

## 3.3 Mid-level Deep Features Correspondence and Relationship Learning

### 3.3.1 Network Architecture

The overview of our network architecture is demonstrated in Figure 3.4. The network can be divided into three components:

**Deep Mid-level Feature Extraction:** The feature extraction network is modified from Google's Inception network by removing the last Inception module: *(Inception_5)*. The feature maps extracted from the *(Inception_4)* module are used as mid-level deep features.

**Correspondence and Relationship Learning:** The mid-level feature correlations between two images and the correspondence relationship between related features are learned by using multi-layer convolution operations on the concatenated feature maps.

**Metric for Similarity Measure:** The metric network with three fully connected layers is utilised for computing the similarity score.

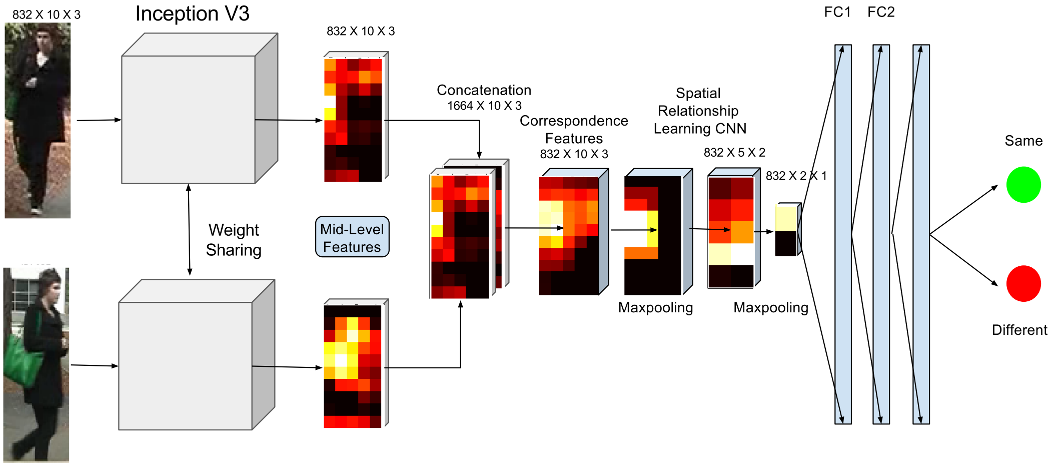


Figure 3.4: The network architecture for correspondence and relationship learning of mid-level deep features

### 3.3.2 Deep Mid-level Feature Extraction

In our proposed network architecture, the Inception network [111] is used as the basic deep neural network architecture for creating the Siamese network structure. Two networks serve as the feature extractors for obtaining the mid-level feature maps of each input image. In order for the features to be comparable, the weights of all convolutional layers for the feature extraction process are shared. The existing person Re-ID datasets are too small to train a well-generalised model. To prevent over-fitting to the specific dataset, we transferred the weight from an ImageNet pre-trained model [43] as a good starting point for the later network training and fine-tuning.

As our training objective differs from the ImageNet classification task, the input image shape is not restricted to the 256 × 256 image shape from the ImageNet. In our architecture, we decide to normalise all the input images to 160 × 80 which is similar to the height-width ratio of the images in many person re-identification datasets [28, 52, 53] and generates less distortion to the original images. In our architecture, we decided to use the mid-level feature maps obtained from the "*Inception_4*" module instead of the last Inception layer (*Inception_5*) as our mid-level feature outputs. There are two reasons:

- With the 160 × 80 input shape, the size of the last convolutional layer outputs will be 5 × 2 which loses too much spatial structure information. The feature maps from the "*Inception_4*" are relatively larger with the 10 × 5 in shape.
- The last convolutional layer of the deep learning model produces high-level features. In the person re-identification situation, mid-level features are more suitable for the task. Therefore, we use the feature maps from a lower level convolutional layer to represent the mid-level deep features.

### 3.3.3 Features Correspondence and Spatial Relationship Learning

Given a probe image from camera A and a gallery image from camera B, each image is represented by 832 feature maps after the mid-level feature extraction process, detailed in the section 3.3.2 above. Let $X_i^A$ and $X_i^B$ represent the $i$th mid-level feature map ($1 \leq i \leq 832$) extracted from two input images. The similarity between the people in the probe and gallery can be learned by analysing the correspondence between $X_i^A$ and $X_i^B$ of the image pair. The previous approaches focus on learning the correspondence features by calculating pair-wise matching probabilities. For example, the first feature maps from probe and gallery images, $X_0^A$ and $X_0^B$ are divided into patches. The correspondence feature of the first feature map is obtained by dense patch matching [97] or local searching in the neighbourhood of the given location [1], as shown in Figure 3.5a. However, with these approaches, each correspondence feature is obtained from only a pair of respective feature maps like [$X_0^A$, $X_0^B$] or [$X_1^A$, $X_1^B$]. They assume the extracted feature maps are independent and fail to address the possible latent relationship among different feature maps. For example, feature maps 1, 3 and 6 can be grouped together to give a better correspondence feature: [$X_{1,3,6}^A$, $X_{1,3,6}^B$].

In our proposed method shown in Figure 3.5b, our correspondence features are obtained by using a convolutional layer:

$$C = f_* \left( \left[ X^A, X^B \right], \Theta \right) \tag{3.1}$$

where $f_*$ denotes the convolution operation. $[X^A, X^B]$ is the concatenation of two mid-level feature maps with shape $1664 \times 10 \times 5$. With kernel size of 3, padding 1 and stride 1, the output feature maps can maintain the shape of $10 \times 5$. The number of output feature maps is set to be the same as the mid-level feature maps, to represent the 832 correspondence features. As the convolution operation is performed on all mid-level feature maps, every kernel filters used for convolutional operation are applied to the feature maps of the two images. The output feature from each kernel can be considered as one possible feature correlation between all feature maps of each image pair. In our proposed method, the correspondence features are not limited to the specific pair of two feature maps but learned from the combinations of many different feature maps. All the weights for combination and convolutional filters are automatically learned in a data-driven manner.

One of the biggest problems in Person Re-ID is person misalignment. Since the convolution operation applies on the entire feature maps, we added another convolutional layer to learn the spatial relationship between all these correspondence features. The input and output shapes are the same ($832 \times 10 \times 5$) with a kernel size of 3, padding 1 and stride 1. To deal with viewpoint variation and misalignment, the max-pooling layer is used to reduce the spatial size of the representation further. With a small representation, correspondence features can represents a large region of a human body. Hence, it eliminate the needs for correspondence alignment between images.

### 3.3.4 Metric Network

Inspired by the MatchNet [33], our similarity metric between features is modelled by using three fully-connected layers with the ReLU non-linearity activation function. The output of the last fully-connected layer will be two values in the range of [0,1]. They can be interpreted as the probability whether the two input images are capturing the same person or not. Besides, we also add a dropout layer after the first and second fully-connected layers. The dropout mechanism can randomly select neurons and ignore them during the training phase. By shutting down neurons randomly, it can prevent the network over-reliant on a few active neurons. Each neuron has the change to learn a useful feature representation and overall improve the generalisation ability of the network and alleviate the over-fitting problem.

### 3.3.5 Loss Function

Our network is trained and optimised by minimising the cross-entropy error of the output labels using stochastic gradient descent (SGD):

$$E = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \cdot \log\left(\hat{y}_n\right) + (1 - y_n) \cdot \log\left(1 - \hat{y}_n\right)\right] \tag{3.2}$$

$N$ refers to the number of image pairs used in a mini-batch during training. Here $y_n$ is the ground truth of image pair $x_n$. $y_n = 1$ indicates the image pair is the same person and $y_n = 0$ means negative matching. $\hat{y}_n$ is the Softmax activation computed based on the output value from the two nodes in the last fully-connected layer $v_0(x_n)$ and $v_1(x_n)$:

$$\hat{y}_n = \frac{e^{v_1(x_n)}}{e^{v_1(x_n)} + e^{v_0(x_n)}} \tag{3.3}$$

In summary, our proposed method can adaptively obtain the mid-level features, automatically construct the correspondence features with their relationship and finally learn the similarity metric. Comparing to previous one-to-one feature map matching approaches, we considered the latent relationship between features when learning the correspondence features. In addition, we eliminated the distance metric stage and proposed an end-to-end similarity measure network which should help reduce the person matching time for the real-time Person Re-ID applications.

## 3.4 Experiments

### 3.4.1 Datasets and Settings

Three publicly available datasets are used to evaluate the performance of the proposed Re-ID network: VIPeR [28], CUHK01[53] and CUHK03[54]. The VIPeR dataset is the oldest and the most tested benchmark for Person Re-ID problem. It contains 632 identities and two images for each identity. Because of the low resolution and large variation in illumination and viewpoints, the VIPeR dataset is a very challenging dataset. The CUHK01 was captured from two camera views as well. It has 971 persons, and each person has two images from camera A, and the other two from camera B. Camera A takes a frontal view and Camera B, the side view. The CUHK03 dataset contains 13,164 images of 1,360 pedestrians, captured by six surveillance cameras. Each identity is observed in two disjoint camera views. On average, there are 4.8 images per identity from each view. The statistics of these datasets are summarised in Table 3.1 below.

Table 3.1: Statistics of each dataset

| Dataset | #ID | #Image | #Camera | label |
|---|---|---|---|---|
| VIPeR | 632 | 1264 | 2 | hand |
| CUHK01 | 971 | 3884 | 2 | hand |
| CUHK03 | 1360 | 13164 | 2 | hand/DPM |

In the training process, the training image pairs are divided into mini-batches of size 96. Therefore, the total number of batches are over one hundred thousand. The stochastic gradient descent (SGD) is used as the optimisation method for minimising the cross-entropy error. The learning rate is set to 0.01 with polynomial decay. The momentum is set to 0.9 with the weight decay of 0.0002.

### 3.4.2 Balancing Training Data

For each person in the datasets, there are only a few positive matching images with a vast amount of negative matching images. Therefore, during the training process, the number of positive image pairs will be much less than negative pairs, which can lead to data imbalance and over-fitting. To reduce the potential over-fitting problem, we also implemented two commonly used pre-processing methods [1]. The data augmentation and hard negative mining as explained below.

#### Data Augmentation

The original training images are reshaped with a random 2D affine transformations around the image center to obtain an additional five augmented images. Then all these images are further augmented by a horizontal flip, which doubles the size of the training sample. This process not only mitigates the data imbalance problem but also generates more training samples.

#### Hard Negative Mining

Data augmentation increases the number of positive pairs, but the training dataset is still imbalanced with many more negatives than positives. If we train the network with this imbalanced dataset, it might learn to predict every pair as negative. Therefore, we randomly down-sample the negative sets to get just twice as many as the positives (after augmentation). So in every batch, there will be 32 positive image pairs and 64 negative image pairs. The converged model obtained is not optimal since it has not seen all possible negatives. We run the pre-trained model to classify all of the negative pairs and choose the highest similarly scored negative pairs (hard negative sample). We then retrain our network with these hard negative samples to boost the robustness of our model.

### 3.4.3 Visualisation of Deep Mid-level Features

Figure 3.6 gives a visualisation of the mid-level feature learned after the training process. They are the highest weighted feature map from the "*Inception_4e/output*" layer when extracting the mid-level features from two images of the same person. The region with very light colour means high activation values. In this case, the most activated region is highlighted around her green handbag. From this experiment, we realised that many mid-level feature maps obtained from our proposed network have semantic meanings which can be very useful for later feature correspondence learning. As a result, it proves that our network can successfully learn useful mid-level features to representation pedestrian's appearance.

### 3.4.4 Performance Evaluation

In this section, the performance of our model is compared with several state-of-the-art methods developed in recent years such as KISSME [42], SalMatch [144], FPNN [54], IDLA [1], LMNN [124], DML [138] and XQDA+LOMO [60]. To evaluate the performance of these Person Re-ID algorithms, the cumulative matching characteristics (CMC) curve is used in our experiment. CMC represents the probability that a query identity appears in a large gallery of images. This metric is adopted since Re-ID is intuitively posed as a ranking problem, where each image in the gallery is ranked based on its comparison to the probe. The probability that the correct match in the ranking is equal to or less than a particular value is plotted against the size of the gallery set [28].

#### Experiments on CUHK01

The CUHK01 dataset contains 3884 images of 971 identities from two different cameras. Previous state-of-the-art approaches normally have two different settings for this dataset: 100 test IDs and 486 test IDs [60, 144]. As the deep learning approaches require a large dataset for training, we did not perform the 486 test IDs experiment. In our experiment, we only focus on 100 randomly selected identities for testing. The remaining identities are used for training. Table 3.2 is the comparison of our proposed method with the recent state-of-the-art results. Our method outperforms the IDLA [1] in this setting by a large margin. The CMC curves of all these methods are shown in Figure 3.7.

Table 3.2: Comparison with the state-of-the-art methods on CUHK01 dataset

| Methods | Rank 1 | Rank 5 | Rank 10 |
|---|---|---|---|
| SDALF | 9.9 | 41.2 | 56.9 |
| LMNN | 21.2 | 48.5 | 62.9 |
| FPNN | 27.9 | 48.5 | 63.0 |
| KISSME | 29.4 | 60.2 | 74.4 |
| SalMatch | 28.5 | 45.0 | 55.0 |
| XQDA+LOMO | 63.2 | 83.9 | 90.0 |
| ImprovedReID | 65.0 | 89.0 | 94.0 |
| Proposed | **81.2** | **95.8** | **97.4** |

**Experiments on CUHK03**

The CUHK03 dataset contains 13164 images of 1360 identities from six different cameras. This dataset has two different pedestrians datasets. One is manually labelled while the other is extracted with the Deformable Parts Model (DPM) human detector [20]. Our model is tested based on the manually labelled dataset. Table 3.3 is the comparison of our methods with the recent state-of-art results. Overall deep learning approaches such as FPNN and IDLA show better results on large datasets when compared to many traditional handcrafted features and learning metrics approaches. Our model still outperforms the IDLA from 55% to 72% in rank-1 accuracy and yields over 90% rank-5 accuracy. The detail CMC performance comparison with other models are shown in Figure 3.8.

Table 3.3: Comparison with the state-of-the-art methods on CUHK03 labelled dataset

| Methods | Rank 1 | Rank 5 | Rank 10 |
|---|---|---|---|
| SDALF | 5.6 | 23.5 | 36.1 |
| LMNN | 7.3 | 19.6 | 30.7 |
| FPNN | 20.6 | 50.9 | 67.1 |
| KISSME | 14.7 | 37.3 | 52.2 |
| XQDA+LOMO | 52.2 | 82.2 | 93.9 |
| IDLA | 54.7 | 86.5 | 93.9 |

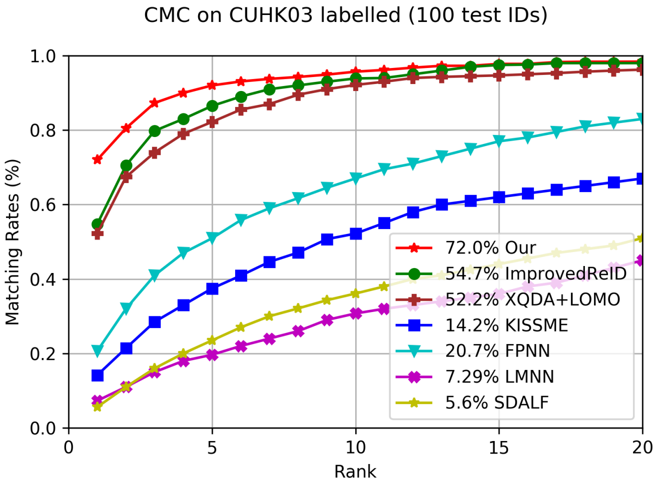


Figure 3.8: CMC curves on CUHK03 labelled dataset

### Experiments on VIPeR

Due to the small number of images in the dataset, VIPeR alone cannot provide enough training data for deep learning methods to coverage well enough. Therefore, the IDLA and our proposed method have to be pre-trained on the combination of the CUHK03 and CUHK01 datasets, then fine-tuned on VIPeR training data. The rest of the traditional approaches such as KISSME and XQDA+LOMO follow the commonly applied 50% training and 10 fold cross-validation evaluation. Table 3.4 below illustrates the overall performance of our model. It outperforms the state-of-art methods significantly even on a small fine-tuned training sample. The detailed CMC performance comparisons with other models are shown in Figure 3.9.

### Cross-dataset Evaluations

As our model can adaptively obtain the excellent correspondence of the mid-level features and learn the relationship between them, we would like to know whether it has the ability to generalise to distinctive features and a similarity metric network for person re-identification tasks in the cross-dataset scenario. In the experiment, our model after training on the full CUHK03 dataset can achieve 64.2% rank-1 accuracy

Table 3.4: Comparison with the state-of-the-art methods on VIPeR dataset

| Methods | Rank 1 | Rank 5 | Rank 10 |
|---|---|---|---|
| SDALF | 19.8 | 39.3 | 49.7 |
| KISSME | 19.6 | 48.0 | 62.2 |
| SalMatch | 30.2 | 52.0 | 65.5 |
| LMNN+LOMO | 29.4 | 59.8 | 73.5 |
| KISSME+LOMO | 34.8 | 60.4 | 77.2 |
| XQDA+LOMO | 40.0 | 68.1 | 80.5 |
| DML | 28.2 | 59.3 | 73.5 |
| IDLA | 34.8 | 63.6 | 75.6 |
| Proposed | **42.5** | **71.4** | **80.6** |

when tested on the full CUHK01 dataset (similar performance to the XQDA+LOMO model on the CUHK01 dataset) and 14.5% rank-1 accuracy on the VIPeR. It gave a comparable performance to the KISSME model on the VIPeR dataset. Although it surpasses many popular methods in a supervised setting, it cannot be directly deployed to a real-world system and requires a lot of optimisation and fine-tuning.

## 3.5 Real-World Implementation

In order to test out our proposed method performance in a real-world CCTV system, we have developed a real-time cross-camera person matching application using the surveillance cameras in the School of Electrical and Electronic Engineering Building (EEE) of Nanyang Technological University (NTU). We selected 12 cameras in the two central corridors of the S1 Building B3 floor, as shown in Figure 3.10.

The detailed system process flow is illustrated in Figure 3.11. In our system, we use YOLOv3 [93] as the person detection to process the real-time video streams from every surveillance cameras. The probe image is the image of a person we want to search on. It can be manually uploaded by the operator or directly selected from one of the video streams. These people detected from surveillance cameras are processed by our end-to-end person re-identification network. Our proposed

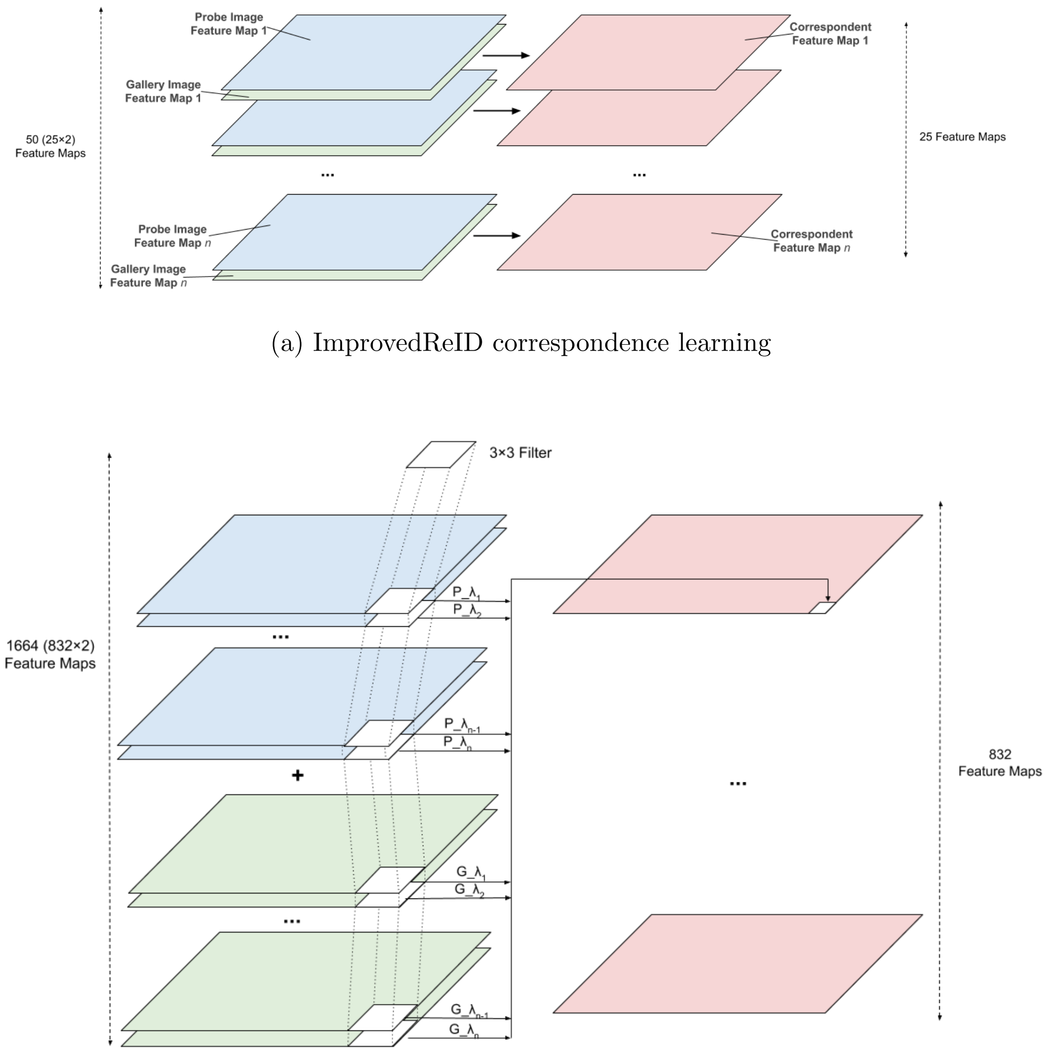
Re-ID model will compute the similarity score for every image pair. During our implementation, we have run several test in the surveillance system and discover that the 0.75 is the best threshold value to achieve a relatively consistent person matching performance. So the person in the video frame with the similarity score above 0.75 will be treated as the same person and highlighted by the red bounding box. Those below 0.75 will be labelled by blue bounding boxes.

The end-to-end Re-ID model used in this system is trained in a fully supervised setting from the ROSE-IDENTITY-Indoor (Re-ID-Indoor) dataset which is collected in the same locations (NTU EEE Build) with 104 cameras in total. During the real-world deployment, we used 4 individual desktop computers with two Nvidia GTX 1070 GPUs installed. One GPU is only used for people detection; another one is dedicated purely for the person re-identification task. This hardware setup can achieve 15 frames per second (fps) processing speed with one full 1080p resolution ($1920 \times 1080$) video streams. In order to monitor the entire 12 cameras, we lower the video resolution to $960 \times 576$. With this setting, each PC can process three real-time video streams simultaneously. In our experiment, our system can achieve nearly 85% matching accuracy overall. However, because of the low threshold value we set, the false detection rate some times can research 30% if other pedestrians are wearing very similar outfits with the probe person.

## 3.6 Conclusion

Our proposed approach can learn and fuse mid-level deep features to handle the misalignment and viewpoint variation problems across two camera views. In contrast to many previous deep learning approaches, our model considers the possible latent relationship between mid-level features when generating the feature correspondences. As an end-to-end network, our network can simultaneously learn the deep mid-level features, feature correspondences and automatically assign the similarity scores from a metric learning network in one single process. We have evaluated the proposed approach on three publicly available person re-identification datasets: VIPeR, CUHK01 and CUHK03 and demonstrated superior performance compared to several state-of-the-art approaches. Benefiting from our latent mid-level feature correspondences learning, the proposed method achieves promising results on assigning feature correspondences score of an image pair. In addition, we also extend the model to the real-world cross-camera matching application and achieve 15 frames per second processing speed with one GTX 1070 GPU.

Our proposed end-to-end deep mid-level feature network can directly assign the similarity score for every image pair. It is extremely efficient in dealing with the real-time cross-cameras person matching. However, when performing the person search from thousands of terabyte (TB) video archives, it needs to perform the person matching scoring pair by pair all over again. For the person retrieval task, the most effective way is to store the pre-processed feature embedding of each image and only perform distance matching when it is needed. So in the next chapter, we will focus on the person retrieval part of the Person Re-ID problems and explore a simple and efficient baseline for person re-identification for the retrieval tasks. In addition, fusing the mid-level deep feature and finding the corresponding matching region provide promising performance in a single-dataset supervised setting. However, the extracted features show poor performance on the cross-dataset scenario. Chapter 5 will address this issue by proposing a cross-dataset Person Re-ID model based on a novel domain adaptation strategy.



(a) ImprovedReID correspondence learning

(b) Proposed method correspondence and relationship learning

Figure 3.5: Correspondence learning difference between IDLA and proposed method

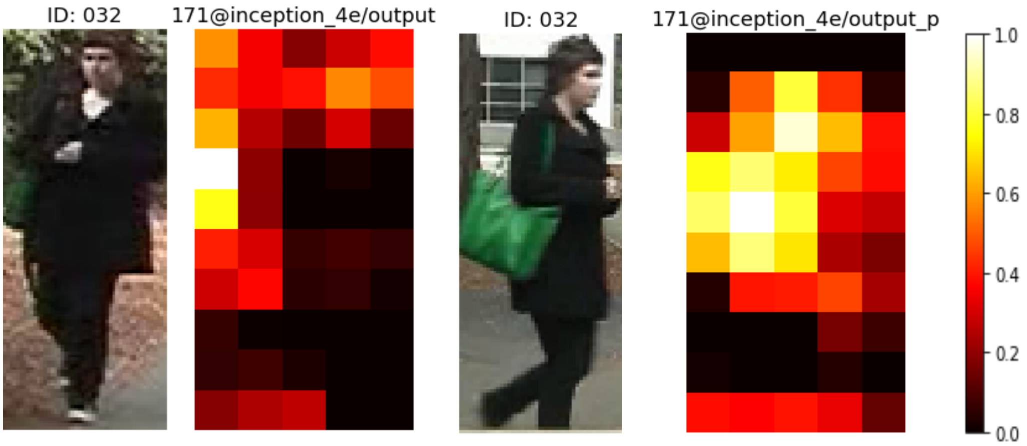


Figure 3.6: The 171th activation feature map from inception_4e/output detecting the handbag

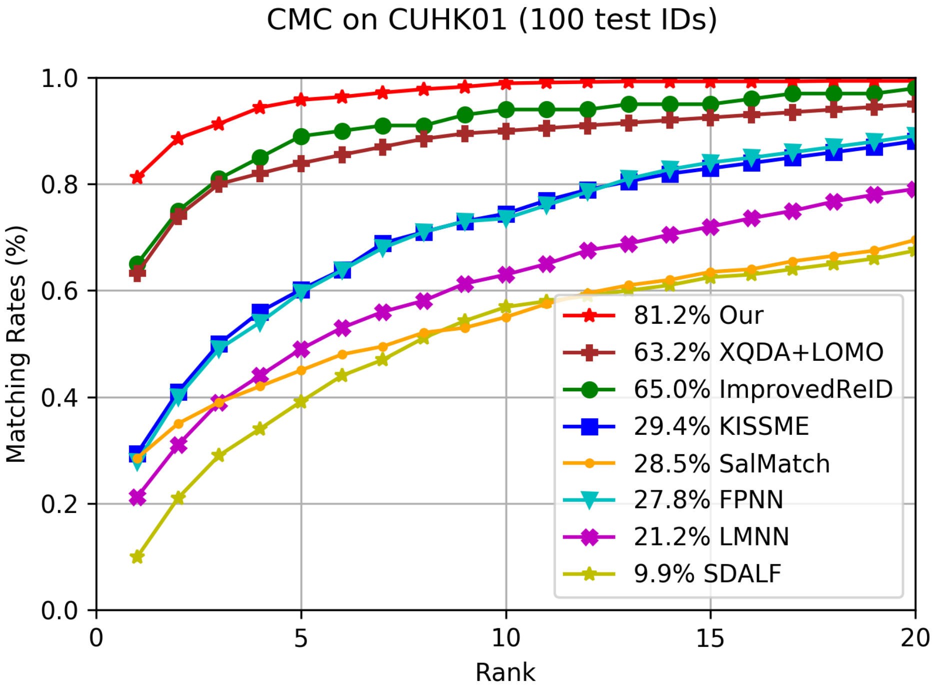


Figure 3.7: CMC curves on CUHK01 dataset

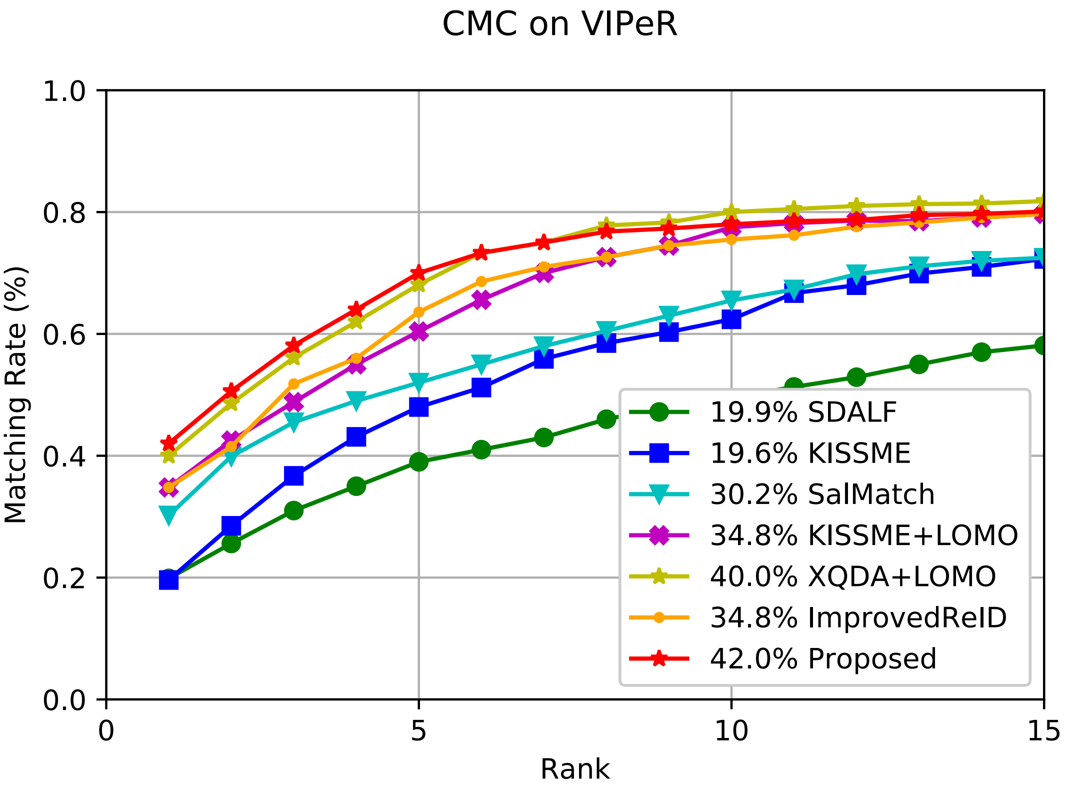


Figure 3.9: CMC curves on VIPeR dataset

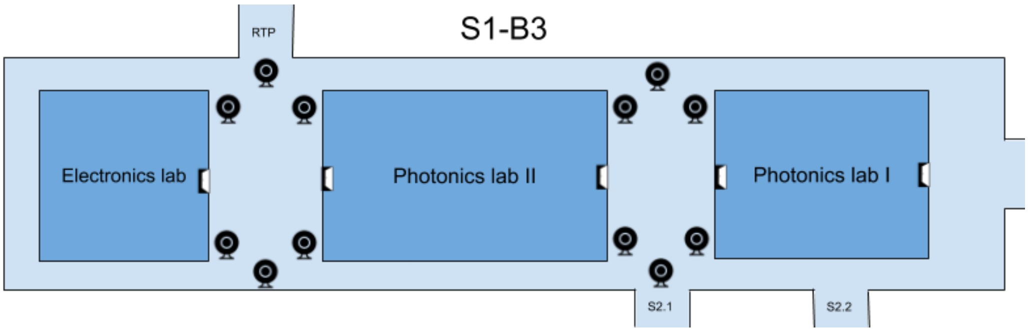


Figure 3.10: 12 Camera Locations in the S1 building B3 floor of EEE, NTU

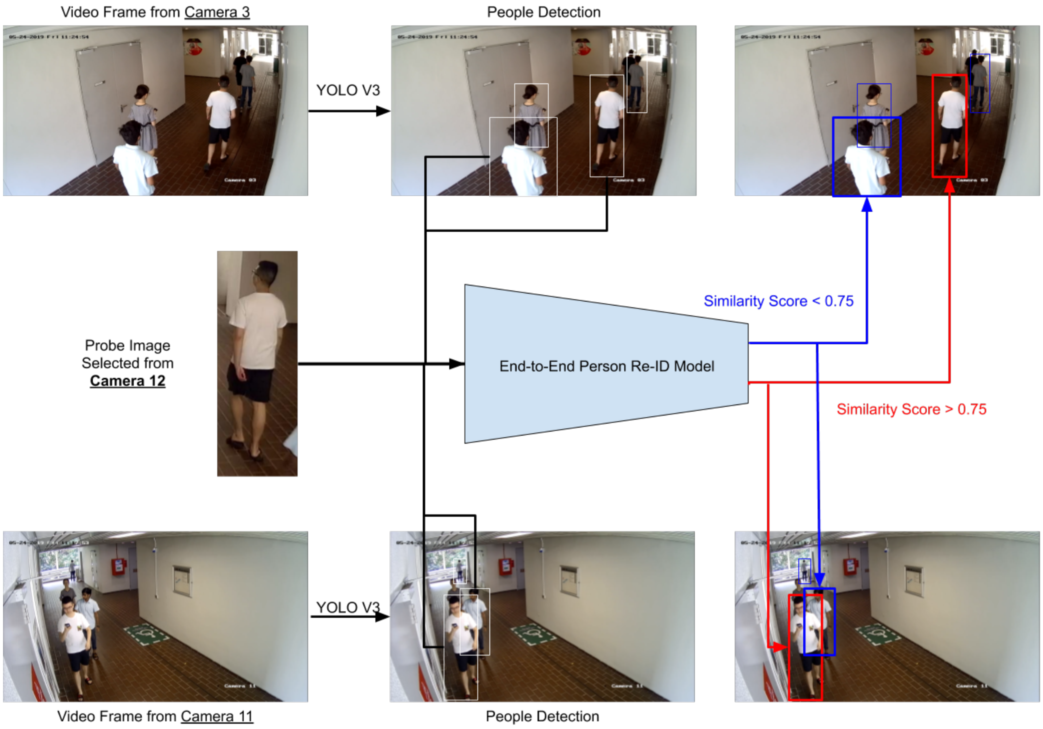


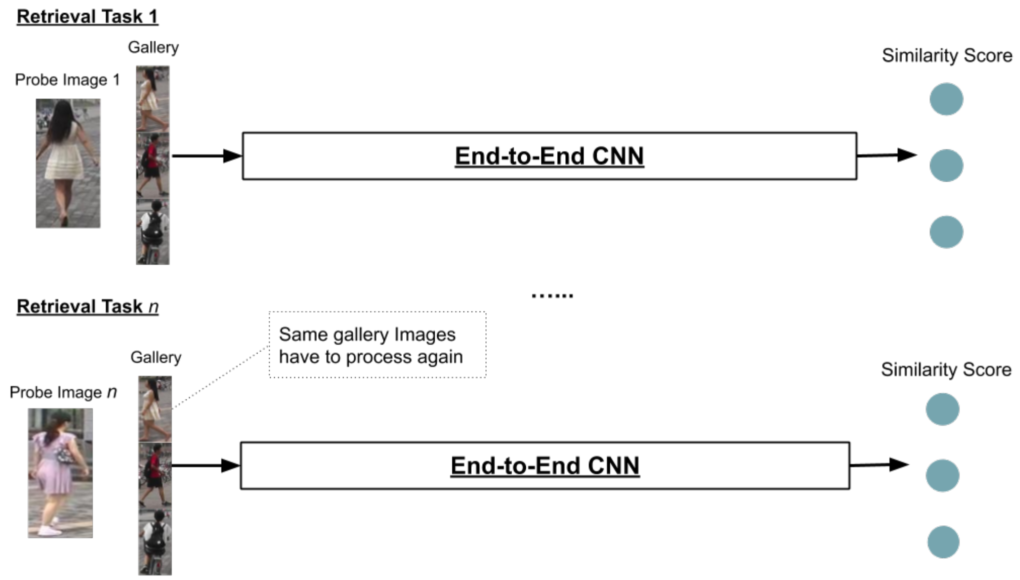
Figure 3.11: The system flow chart of person online matching

# Chapter 4

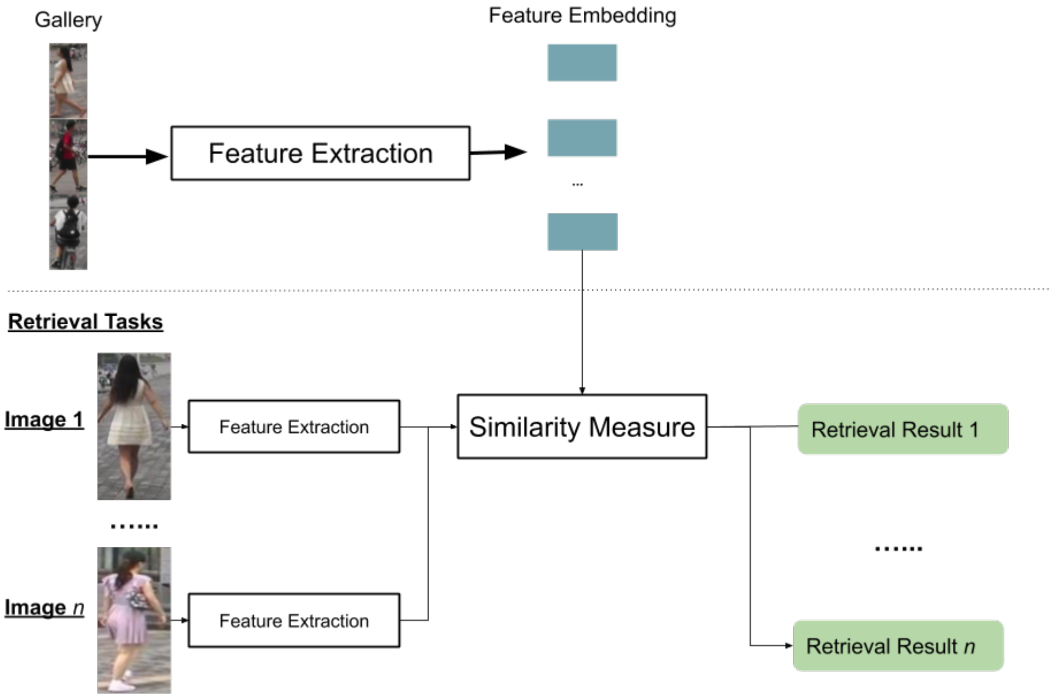
# Single-Dataset Feature Learning (Offline Retrieval)

## 4.1 Introduction

The previous chapter proposed an end-to-end mid-level feature correspondence network for solving the real-time online person matching problem. This chapter focuses on the Person Re-ID problem for offline person retrieval. Offline person retrieval aims at retrieving images of a specified pedestrian from a large gallery of human images obtained from several historical video files. The single-stage end-to-end network proposed in Chapter 3 is suitable for the real-time person matching task because it combines the feature extraction and similarity scoring in one single stage, which reduces the processing time. However, in the offline person retrieval task, the operators usually need to search for multiple different subjects. The end-to-end framework we proposed in Chapter 3 requires both query image and gallery images as the input. For every retrieval request, it needs to re-process the entire gallery images again, as shown in Figure 4.1a. Much time and the computational resource are wasted on re-processing the feature of the gallery images.



(a) How an end-to-end network is used for the person retrieval task. For every new retrieval request, the network needs to process the same gallery images again.



(b) How a multi-stage network is used for the person retrieval task

Figure 4.1: The comparison between the single-stage and the multi-stage framework for the person retrieval task. The multi-stage framework separates the feature extraction stage and distance metric stage. As a result, the gallery images only need to be processed once.

The multi-stage framework, on the other hand, separates the feature extraction step and distance metrics step. As a result, the feature embedding of the gallery images can be reused for different query persons, as shown in Figure 4.1b. In our offline person retrieval application, we store pre-processed feature embeddings of all the gallery images. For every retrieval request, we only need to process the query person's image and match him/her with the stored gallery feature embeddings. It significantly reduces the retrieval time for our application. In this chapter, we focus on building a simple and robust feature extractor base on our novel negative competing triplet loss function (NC-Triplet). Additionally, we provide a comprehensive ablation study of several data refinements and training techniques.

## 4.2 Problem Definition

Person Re-ID with deep neural networks has made progress and achieved high performance in recent years. However, many state-of-the-art methods design complex network structures and concatenate multi-branch features [14, 132, 142]. This chapter explores a simple and efficient Person Re-ID feature extractor trained from our newly proposed negative competing triplet loss function. In addition, we collected and evaluated some effective training techniques or refinements which appeared in several papers published in the past two years. A practical Person Re-ID model needs to be simple and effective rather than concatenating lots of local features into a multifarious output. In pursuit of high accuracy, researchers combine several local features or utilise the semantic information from pose estimation [14, 132] or segmentation models [59, 132]. Such methods involve too many additional computational processes. Also, large feature embeddings greatly reduce the speed of the retrieval process. The overall contribution of this work can be summarised as follows:

1. We proposed a new negative competing triplet loss (NC-Triplet) function, which improves the mean average precision (mAP) performance of the Person Re-ID model.

2. Using our NP-Triplet loss function combined with these training techniques, we established a strong baseline for researchers to achieve higher accuracies in the future Person Re-ID works.

3. Address the limited camera number and lack of privacy concern in the existing Person Re-ID datasets, we collected a more realise and more challenging Re-ID-Outdoor dataset. It is the first Person Re-ID dataset with a privacy declaration form singed by all participants.

## 4.3 Network Architecture

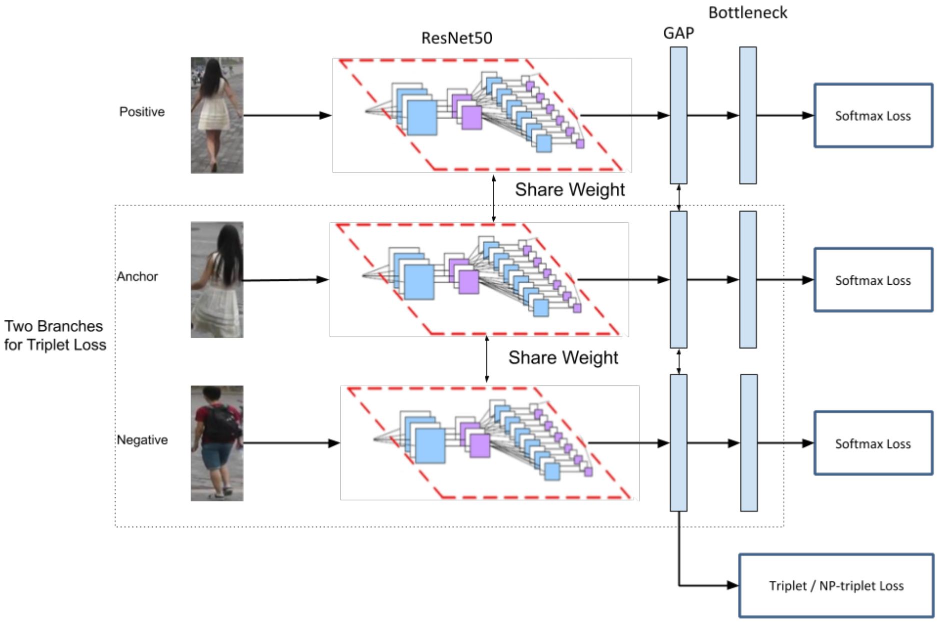


Figure 4.2: Our proposed network architecture

Figure 4.2 shows the network architecture for our Person Re-ID baseline model. We use the ResNet50 [34] as the backbone structure for our feature extractor. The 2048 feature maps obtained from the last residual module will undergo a Global Average Pooling (GAP) or a Global Max Pooling (GMP) layer to form a 2048-size 1-dimensional feature vector. This feature vector is used to compute the triplet loss. The 2048 feature vector of each image will then be passed through a bottleneck layer to compute the softmax loss based on its corresponding person ID label. Section 4.5 below will give detailed explanations of the GAP/GMP layer and the bottleneck layer. Overall, the models used for our experiments follow the pipeline below:

1. The ResNet50 backbone network is initialised with pre-trained parameters on the ImageNet [43] dataset.

2. For the softmax loss based model, we use $B$ number of images in one single batch. For the triplet loss based model, we randomly sample $P$ identities and $K$ images per person to constitute a triplet loss training. The final batch size equals to $B = P \times K$. For example, we set $P = 16$ and $K = 4$. The final batch size $B$ will be 64

3. Unlike Market1051 dataset, the images in the DukeMTMC-reID dataset have various height and width ratio. We re-size all images into $(266 \times 138)$ pixels with 10 pixels padding, then randomly crop them into a $(256 \times 128)$ size. In our experiment, we also test our model with a larger input size of $(384 \times 128)$.

4. Each image can be randomly flipped horizontally with 0.5 probability [43]. This process enlarges the training sample size and makes the model invariant to the horizontal direction changes.

5. The ResNet50 backbone network we used is initialised by training the ImageNet dataset. Hence, we apply the same image pre-processing for the ImageNet training to the Person Re-ID model training which normalises RGB channels of all input images by subtracting 0.485, 0.456, 0.406 and dividing by 0.229, 0.224, 0.225, respectively [43].

6. We change the dimension of the last fully connected layer to $N$ neurons. $N$ denotes the number of human identities in the training dataset. Our model will output the Person Re-ID feature $f$ from the GAP/GMP layer and their ID prediction logit $p$.

7. The Person Re-ID feature $f$ is used to compute the proposed NP-triplet loss. In each batch, we also use the hard-negative mining strategy mentioned in Chapter 3. ID prediction logits $p$ is used to calculate the softmax loss.

8. The optimisation method we adopted for training our model is Adam [39]. For a fair comparison with other state-of-the-art methods [51, 72], we follow the same Adam learning rate setting. The initial learning rate is set to be 0.00035. Then, the learning rate is decreased by 0.1 at the 40th epoch and 70th epoch, respectively. Totally, there are 120 training epochs.

## 4.4 Loss Function

Many recent Person Re-ID models use a weighted sumation of the softmax loss with triplet loss.

### 4.4.1 Softmax Loss (Identification Loss)

In 2016, Zheng et al. [151] proposed the ID Embedding (IDE) network. They considered the training of the person re-identification model as a human id classification task. The objective of the feature embedding learning from the network should be able to successfully map the images to their corresponding identity labels (ID labels). Hence, they adopted the widely used softmax loss in their model. As shown in Figure 4.3, the last layer of IDE is a fully connected (FC) layer with a hidden size equal to the number of persons $N$ in the training set.

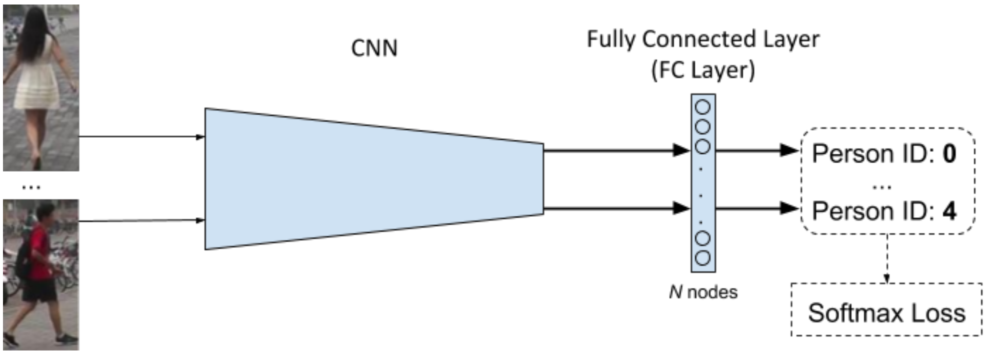


Figure 4.3: The softmax loss in the ID embedding network

Given an image $i$ and $N$ is the number of persons in the training set, we denote the $p_i$ and $q_i$ as the ground truth and the predicted probability. The softmax loss is computed as:

$$L_{Softmax} = \sum_{i=1}^{N} -p_i \log(q_i) \tag{4.1}$$

The softmax loss is suitable for the case that inter-class distance is much larger than intra-class distance, such as the classification task on ImageNet dataset [43]. However, this loss function does not consider the intra-class and inter-class distance. In fact, the appearance of the same individual varies greatly and different people may be similar across views. The softmax loss alone is not suitable for the person re-identification task. Therefore, it needs the help of other loss functions, which considers intra-class and inter-class distance.

### 4.4.2 Triplet Loss (Verification Loss)

Triplet loss is a commonly used loss function which considers intra-class and inter-class distance. In 2005, Schroff and Philbin developed the FaceNet model [96] for face recognition and clustering. They proposed a modified "Large Margin Nearest Neighbor Loss" [124] called "triplet loss". The triplet loss has been adapted in many recent Person Re-ID works [11, 14, 72, 99, 117]. The softmax loss considers the Person Re-ID training as learning an ID label classification model. The triplet loss approaches treat the Person Re-ID training as learning a person verification model.

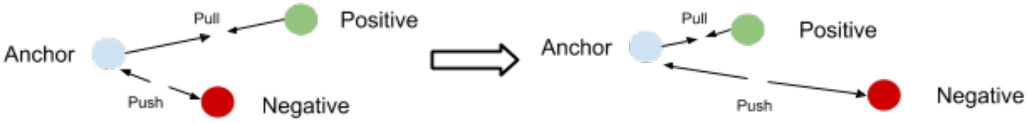


Figure 4.4: Illustration of the distance changes of the positive and the negative feature embedding pairs after training with the triplet loss

To train a triplet loss model, one feature embedding $f_a^i$ of an image of person $i$ is used as an *anchor* of the triplet. $f_p^i$ denotes a feature embedding of the same person $i$ (positive pair to the anchor image). $f_n^j$ denotes a feature embedding of a different person $j$ (negative pair to the anchor image). The training process encourages the model to make the $l_2$ distance between the positive pair $D_{ap} = D(f_a^i, f_p^i)$ smaller than the negative pair $D_{an} = D(f_a^i, f_n^j)$ by a distance margin $\alpha$, as shown in Figure 4.4. The triplet loss function of one triplet can be defined as

$$\begin{aligned} L_{Triplet} &= \max\{0, D_{ap} - D_{an} + \alpha\} \\ &= \max\left\{0, D(f_a^i, f_p^i) - D(f_a^i, f_n^j) + \alpha\right\} \end{aligned} \tag{4.2}$$

where $D_{ap}$ and $D_{an}$ are the $l_2$ distances of the positive pair and the negative pair. $\alpha$ is the margin of the triplet loss. Our experiments follow the same setting of most triplet based models which set the distance margin $\alpha$ to 0.3.

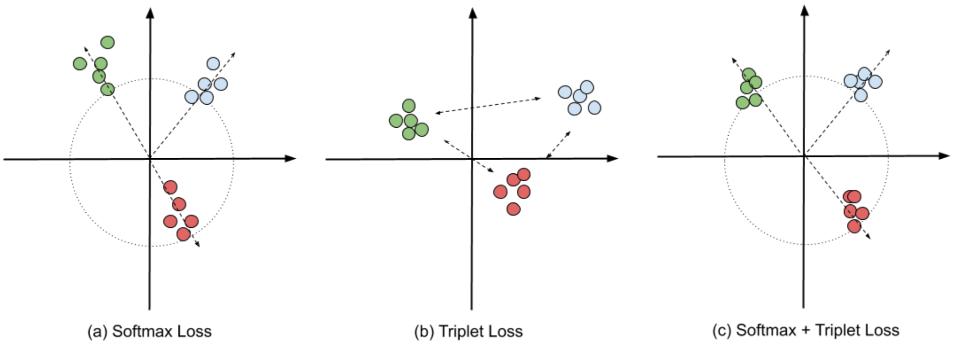


Figure 4.5: Two-dimensional visualisation of sample distribution in the embedding space supervised by (a) Softmax Loss, (b) Triplet Loss, (c) Softmax + Triplet Loss (figure provided in [72])

The softmax loss constructs several hyper-planes to divide the embedding space into different sub-spaces. Hence, it separates the feature embeddings by enlarging the cosine of angles between them. However, without explicit constraints on the feature space distribution, the learned feature embeddings may not be optimal. As shown in Figure 4.5(a), there is no constraint on the distribution in the embedding space, which leads to a general spread. On the other hand, the triplet loss function enhances the intra-class compactness and inter-class separability in the Euclidean space, as shown in Figure 4.5(b). However, the triplet loss function does not have a global optimal constraint. The inter-class feature embedding distance sometimes may be smaller than intra-class distance. Combining the softmax loss and the triplet loss can alleviate each others drawbacks. The softmax loss takes full advantages of labels and optimises the cosine distances of the feature embedding. The triplet loss considers intra-class and inter-class distance and optimises the Euclidean distance. As a result, many of the recent Person Re-ID works [14, 38, 72, 99, 117] uses the weight summations of two losses:

$$L_{Combine} = \lambda L_{Softmax} + (1 - \lambda) L_{Triplet} \tag{4.3}$$

### 4.4.3 Negative Competing Triplet Loss

The original triplet loss pushes the negative feature embedding $f_n^j$ away from the anchor embedding $f_a^i$, shown in Figure 4.4. However, once the negative pair is further away than the positive pair + distance margin $\alpha$, there will be no gain for the Person Re-ID model for any improvement. To alleviate this problem, we proposed a new negative competing triplet loss (NC-Triplet), which further enlarges the distance between the positive and negative embeddings.

Our NC-Triplet is the sum of two different losses: original triplet loss and newly proposed negative competing loss, as shown in Figure 4.6. The negative competing loss ensure the distance of the positive-negative embedding pair $D_{pn} = D(f_p^i, f_n^j)$ is

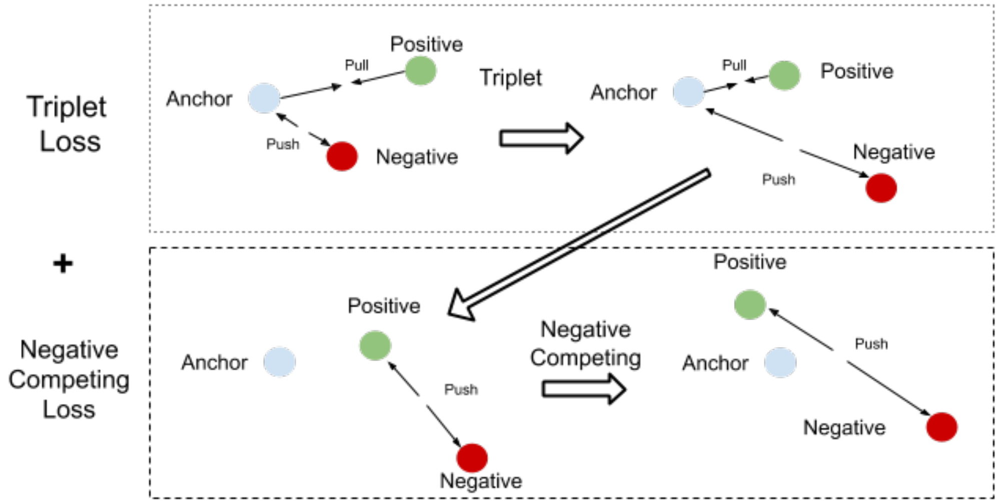


Figure 4.6: NC-Triplet loss combines the original triplet loss and an additional negative competing loss. The negative competing loss pushes the positive embedding further away from the negative one.

larger than the anchor-negative pair,$D_{an} = D(f_a^i, f_n^j)$, as shown in Figure 4.6. As the anchor embedding and the positive embedding are from the same person in the dataset, the negative competing loss further enlarges the embedding distance between different pedestrians.

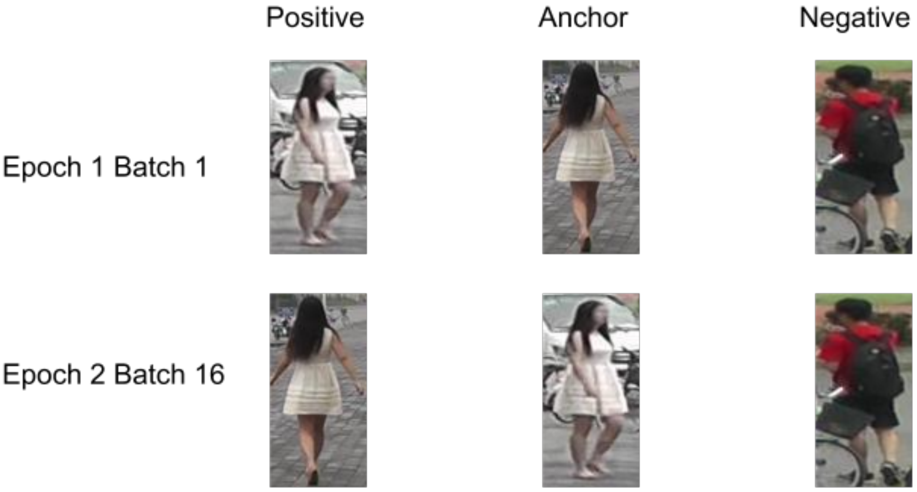


Figure 4.7: One example of the different order of the anchor image and positive image under different training epoch and training batch

In addition, our deep convolutional neural network models are trained with random anchor image selection for each batch. So the anchor image of epoch 1 can also be the positive image in epoch 2 and vice versa. Figure 4.7 demonstrates an example of the two image triplets in two different epochs during the training. The three images are the same for two triplets with a different order of anchor and positive images. The entire training process of the NC-Triplet loss can be considered as the anchor images, and positive images are competing with each other to move away from the negative images. As a result, we name this modified triplet loss function as native competing triplet loss (NC-Triplet loss). Overall, the NC-Triplet loss can be computed as:

$$L_{NC-Triplet} = \max\left\{0, (D_{ap} - D_{an} + \alpha_1) + (D_{pn} - D_{an} + \alpha_2)\right\} \tag{4.4}$$

where $D_{ap}$ and $D_{an}$ are feature distances of the anchor-positive pair and the anchor-negative pair. $D_{pn}$ are feature distances between the positive and the negative feature embeddings. $\alpha_1$ is the distance margin for triplet loss, and $\alpha_2$ is the distance margin for negative competing loss. We set both distance margins $\alpha_1$ and $\alpha_2$ to 0.3 which follow the same setting as triplet loss function used in ImpTripet [11].

### 4.4.4 Center Loss

The original triplet loss and our proposed NC-Triplet loss only consider the difference among $D_{ap}$, $D_{an}$ and $D_{pn}$. They ignore the absolute values of the feature distance. For instance, when $D_{ap} = 0.3$, $D_{an} = 0.5$, the triplet loss will be 0.1. In another case, when $D_{ap} = 1.3$, $D_{an} = 1.5$, the triplet loss also is 0.1. The triplet loss and our NC-Triplet loss are computed from feature embeddings sampled randomly from two different persons. It is difficult to ensure that $D_{ap} < D_{an}$ in the whole training dataset. Center loss proposed by Wen et al. [125] can simultaneously learn a center for deep features of each class and penalises the distances between the deep features and their corresponding class centers. The center loss function is formulated as:

$$\mathcal{L}_{Center} = \frac{1}{2}\sum_{i=1}^{B} \|f_i - c_{y_i}\|_2^2 \tag{4.5}$$

where $y_i$ is the ID label of the $i$th image in one training batch. $c_{y_i}$ denotes the $y_i$th class center of deep features. $B$ is the number of batch size. The objective of this loss function is to reduce the square $l_2$ norm distance between every sample features and their corresponding feature centers. Hence, the formulation effectively characterises the intra-class variations. Finally, we formulate the overall loss function by incorporating the weighted sum of the softmax loss, NC-Triplet loss and center loss:

$$L_{Final} = \frac{\lambda_1}{\lambda} L_{Softmax} + \frac{\lambda_2}{\lambda} L_{NC-Triplet} + \frac{\lambda_3}{\lambda} L_{Center} \text{ ,while } \lambda = \lambda_1 + \lambda_2 + \lambda_3 \tag{4.6}$$

For a fair compassion to the state-of-the-art methods, we follow the same setting of many other methods [38, 72, 117] and empirically fixed the $\lambda_1$, $\lambda_2$ and $\lambda_3$ to 1.

## 4.5 Training Techniques and Refinements

In this section, we will introduce some effective training techniques or refinements used for training our models. These techniques are collected from the mainstream conference proceedings and journal papers in recent years. The detailed ablation studies will be discussed in Section 4.7.

### 4.5.1 Different Input Size

In our model, each image is re-sized to 256 × 128 pixels with an additional 10 pixels padding. We then randomly crop them back to 256 × 128 rectangular shape. Random cropping prevents a neural network from over-fitting to specific features by changing the location of the apparent features in an image [112]. That is due to the fact that the images in the Market1501 dataset are all re-sized with the 2:1 height-width ratio.

Other datasets such as DukeMTMC-reID and MSMT17 have unconstrained image sizes but close to the 3:1 height-width ratio. In our experiments, we have trained two networks with different input sizes: $256 \times 128$ and $384 \times 128$ to determine which size is more suitable for Person Re-ID task. The experimental results will be shown in Section 4.7.

### 4.5.2 Warm-up Learning Rate

Using different learning rates have a great impact on the performance of most deep learning models. Many recent Person Re-ID works [14, 64, 72, 117] use the multi-step learning rate which reduces the learning rate after a certain epoch stage, shown as the blue line in Figure 4.8. Goyal et al. [26] proposed a new warm-up learning rate strategy. It uses a lower learning rate at the start of training and gradually increases the learning rate for the first few epochs. It helps initialise the model well before applying a large learning rate for optimisation. As a result, recent Person Re-ID works use warm-up learning rate to train their models [18, 36, 72]. The red dotted line in Figure 4.8 illustrates how the learning changes during our model training process. The first 10 epochs linearly increase the learning rate from $3.5 \times 10^{-6}$ to $3.5 \times 10^{-4}$. Then, the learning rate is decreased to $3.5 \times 10^{-5}$ and $3.5 \times 10^{-6}$ at 40th and 70th epoch, respectively. In our experiment, the learning rate warm-up strategy can give a 1% increase in both CMC and mAP performance metrics.

### 4.5.3 Random Erasing Augmentation

In Person Re-ID, people in the images are sometimes occluded by other objects. To address the occlusion problem and improve the generalisation ability of the models, Zhong et al. [153] proposed a new data augmentation approach named as Random Erasing Augmentation (REA). For an image $I$ in a mini-batch, the probability of it undergoing Random Erasing Augmentation is $p_e$ and the probability of it being kept unchanged is $(100\% - p_e)$. Then, REA randomly selects a rectangle region $I_e$

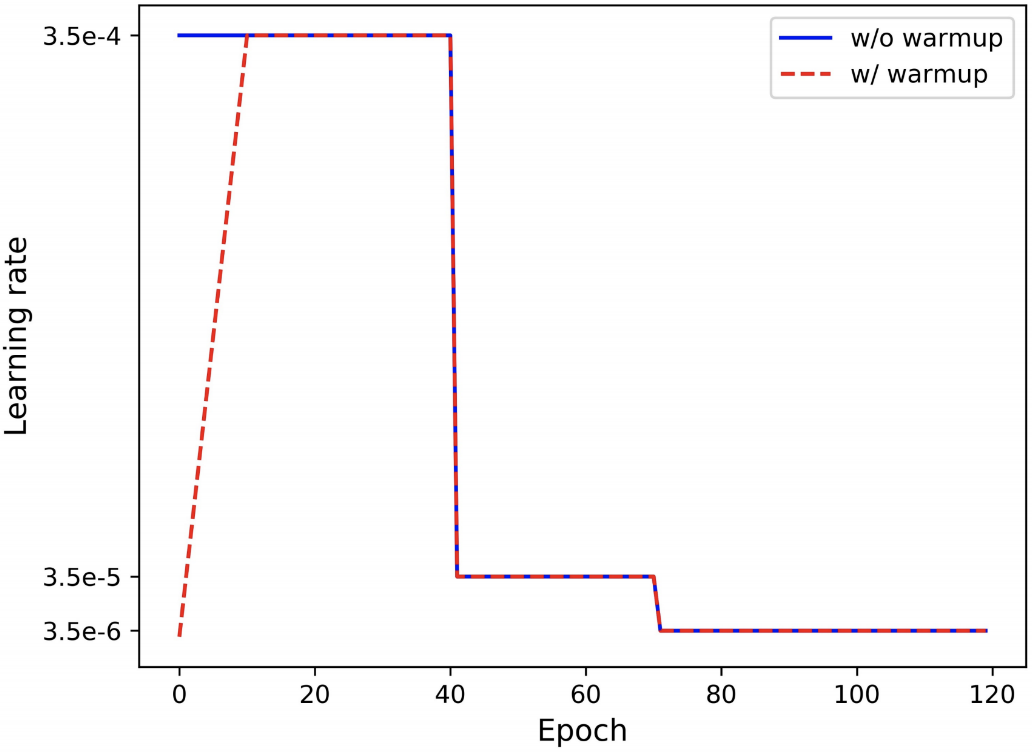


Figure 4.8: Comparison of learning rate schedules. With warm-up strategy, the learning rate is linearly increased in the first 10 epochs.

with size $(W_e, H_e)$ in image $I$ ,and erases its pixels with random values, as shown in Figure 4.9. This augmentation mimics the common Person Re-ID problem: human body occlusion. By training with images containing occlusion in different human body parts, the trained model should be more robust to the occlusion and more sensitive to the local region features.

### 4.5.4 Last Stride

In the previous chapter, we removed the last *Inception* module from the backbone network to increase the size of the feature maps [62]. This process creates near mid-level feature representation and enriches the granularity of the extracted features. Another way to increase the size of the feature map is to remove the last spatial down-sampling operation in the backbone network [109]. The default last spatial down-sampling operation (Last Stride) in the ResNet50 is set to 2. With a $256 \times 128$

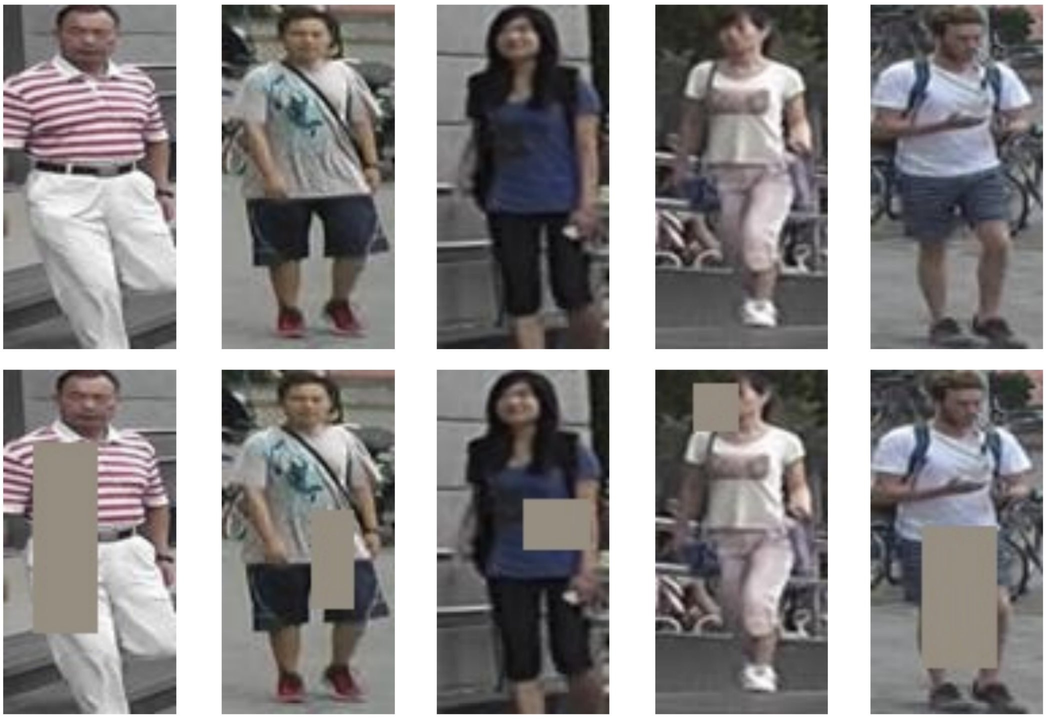


Figure 4.9: Sampled examples of the random erasing augmentation of the Market1501 dataset. The first row shows five original training images. The processed images are presented in the second low

size image as an input, the ResNet50 will output a feature map with the spatial size of 8 × 4. By changing the last stride from 2 to 1, we can obtain a feature map with a larger size of 16 × 8. This manipulation increases very little computation cost and does not involve any extra training parameters. The performance improvement from the higher spatial resolutions will be analysed in Section 4.7 later.

### 4.5.5 GAP and GMP

CNN perform convolution in the lower layers of the network. For classification, the feature maps of the last convolutional layer are vectorised and fed into FC layers followed by a softmax logistic regression layer. However, the fully connected layers are prone to over-fitting, thus hampering the generalisation ability of the overall network. Lin et al. [61] proposed a new strategy called global average pooling (GAP) to generate a one-dimensional feature vector before feed to the traditional fully connected layers in CNN. Instead of adding fully connected layers on top of the feature maps, GAP takes the average of each feature map and produces a one-dimensional feature vector which is fed directly into the softmax layer. One advantage of global average pooling is that it enforces the correspondences between feature maps and labels (person id in Person Re-ID problem). Since there is no parameter to optimise in the global average pooling, the over-filling problem is avoided at this layer. In our baseline model, we also introduce the global max-pooling (GMP) layer which takes the maximum value of each feature map. As the GMP layer takes the maximum value of the feature map, it helps the model emphasise on the semantic regions from the feature maps. In our GAP and GMP comparison experiment, GMP usually outperforms the GAP by a small margin.

### 4.5.6 Bottleneck Layer

Our baseline model uses both the softmax loss and the triplet loss. The softmax loss constructs several hyper-planes to divide the embedding space into different sub-spaces. The features of each class are distributed in different sub-spaces. The softmax loss function separates the feature embeddings by enlarging the cosine of angles between them. On the other hand, the triplet loss enhances the intra-class compactness and inter-class separability in the Euclidean space. The softmax loss mainly optimises the cosine distances while the triplet loss focuses on the Euclidean distances. If we use these two losses to optimise one feature vector simultaneously, their goals may be inconsistent and conflicting with each other. To overcome the aforementioned problem, Luo et al. [72] proposed a bottleneck structure shown in Figure 4.10. The bottleneck layer adds an additional batch normalisation (BN) layer between the GAP/GMP layer and the softmax classifier layer. By adding an extra buffer layer, it reduces the constraint of the triplet loss from the softmax loss which could alleviate the conflict between them.

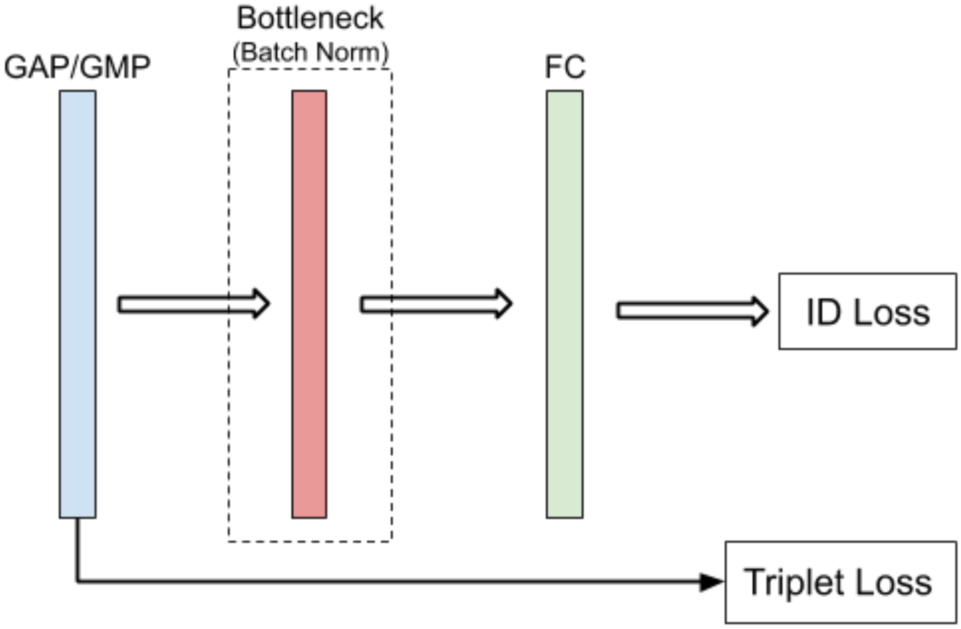


Figure 4.10: The bottleneck layer.

## 4.6 Rose-Identification-Outdoor Dataset

The existing public datasets have three main limitations that need to be addressed:

1. Small camera network size
2. Unrealistic surveillance environment
3. Lack of privacy consideration

Hence, we collected a new large-scale Person Re-ID dataset Rose-Identification-Outdoor (Re-ID-Outdoor) to address these three problems.

### 4.6.1 Increase Camera Network Size

Much attention has been paid in recent years to the problem of Person Re-ID. Most existing deep learning based Person Re-ID algorithms are typically trained and evaluated on three large-scale public datasets: Market1501 [147], DukeMTMC-reID [152] and the most recent MSMT17 [122]. All these datasets are obtained from a very limited number of cameras ranging from 6 to 15, shown in Table 4.1 below.

The small camera number reduces the variation and diversity of a person under different backgrounds, illumination or camera colour profiles. Overall, it makes it too easy for searching and matching people across different cameras. As a result, many recently proposed Person Re-ID algorithms [14, 117] can achieve more than 90% rank-1 accuracy in both Market1501 and DukeMTC-reID. However, real-world surveillance systems usually consist of hundreds of cameras. As a result, the images obtained from them are much more dynamic in nature.

| Datasets | Market1501 | DuketMTMC-reID | MSMT17 | **Re-ID-Outdoor** |
|---|---|---|---|---|
| # Cameras | 6 | 8 | 15 | **50** |

Table 4.1: Number of cameras for recent Person Re-ID datasets and our Re-ID-Outdoor dataset

The Re-ID-Outdoor dataset is collected within the Nanyang Technological University (NTU) campus by using actual surveillance cameras installed on lamp posts. There are a total of 34 camera locations, with each camera location installed with one to four cameras pointing in different directions. The location of all cameras can be found in Figure 4.11. During the data collection period, we selected the 50 cameras from 23 highly active camera locations. Overall, our Re-ID-Outdoor dataset consists of a total of 50 different cameras covering the entire $2km^2$ NTU campus area, which gives the most dynamic changes in the image background. The detailed comparison with other public datasets are shown in Table 4.1.

### 4.6.2 More realistic Surveillance Environment

Another drawback of many existing Person Re-ID datasets such as VIPeR [28], Market1501 [147] and MSMT17[122] are using non-surveillance cameras mounted on tripods for video recording, which result in a near-horizontal point of view of all captured persons, as shown in Figure 4.12(a). However, in actual surveillance systems, cameras with wide-angle lens are mounted on lamp-posts or ceilings, which

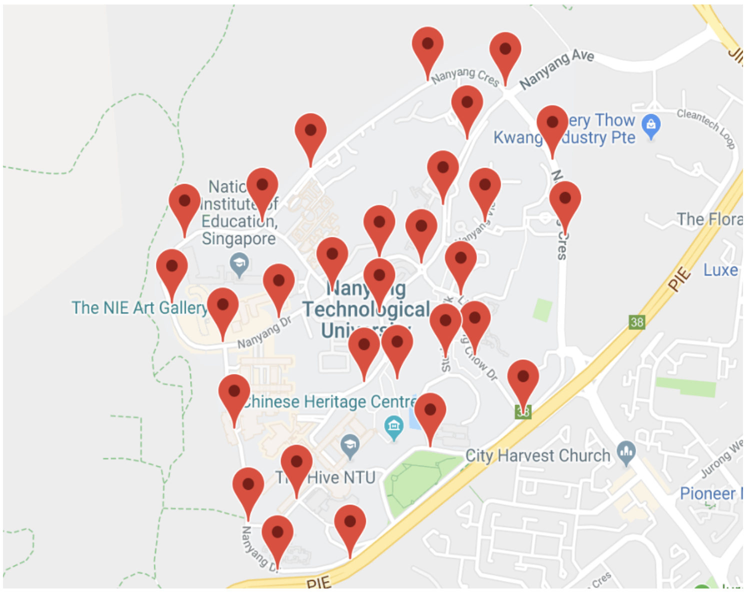
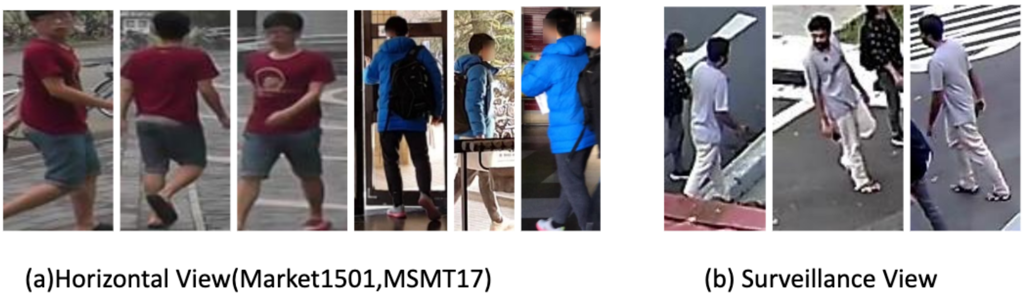


Figure 4.11: An overview of our outdoor surveillance cameras

give unique top-down wide-angle views of passing pedestrians. In Re-ID-Outdoor dataset, all the images are captured from actual surveillance cameras mounted on lamp-posts with distinctive top-down wide-angle viewing, as shown in Figure 4.12(b).


(a)Horizontal View(Market1501,MSMT17)

(b) Surveillance View

Figure 4.12: Viewing angles of cameras used in Market1501, MSMT15 versus actual surveillance cameras in Re-ID-Outdoor

In addition, most of the real-world surveillance systems run 24/7. However, all existing public datasets only use the videos recorded during the daytime, which limited the capability of the Person Re-ID models trained on them. In Re-ID-Outdoor dataset collection, we take into consideration the difference of a person during the daytime and nighttime by running all cameras 24 hours non-stop. In Re-ID-Outdoor, the daytime videos and nighttime videos have very different colour profiles and image quality shown in Figure 4.13. However,

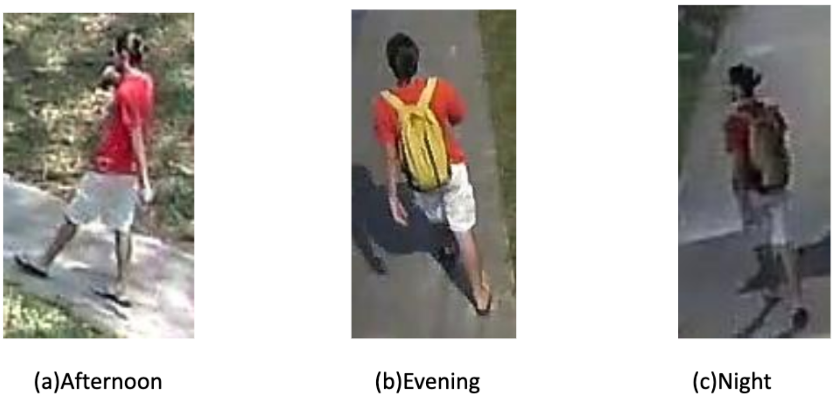


Figure 4.13: Same person in the afternoon, evening and nighttime in our Re-ID-Outdoor dataset

If we consider the period from 7am to 7pm as the daytime, 38.6% of the total images in the Re-ID-Outdoor dataset are captured during the nighttime, shown in Figure 4.14.

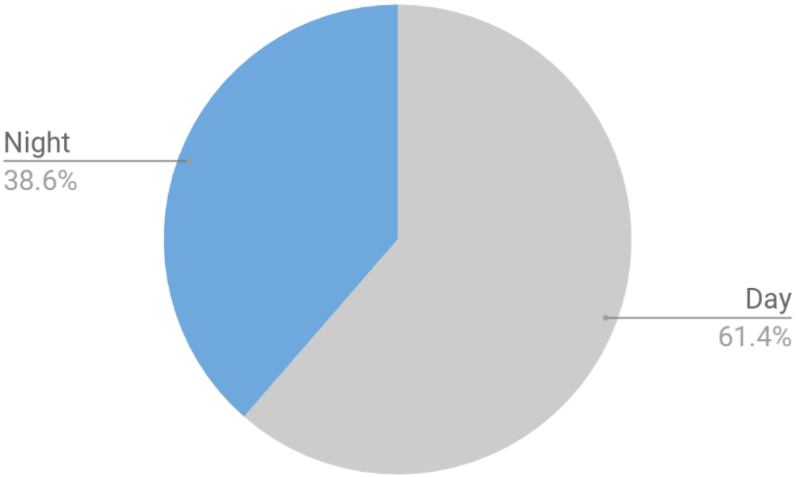


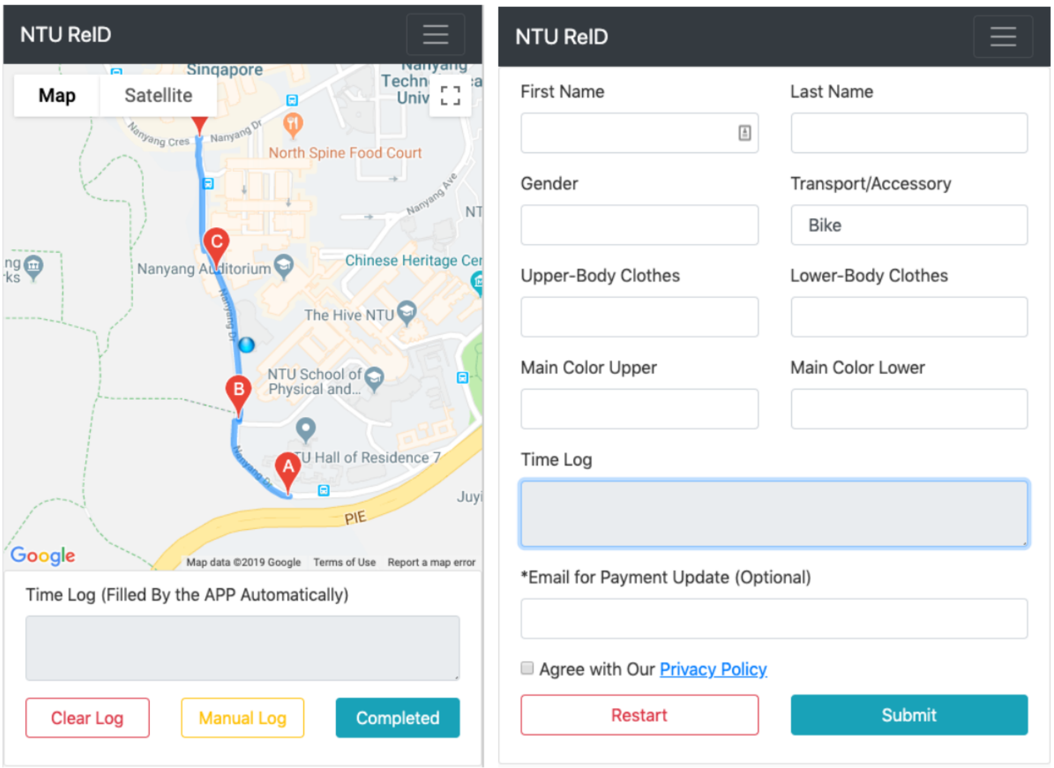
Figure 4.14: Percentage of daytime images and nighttime images in our Re-ID-Outdoor dataset

### 4.6.3 Lack of Privacy Consideration

Most of the Person Re-ID datasets are collected and annotated from hours of video footage recorded from several cameras set up in a public space. For example, the Market1501 dataset [147] was collected in front of a supermarket in Tsinghua University in 2015. DukeMTMC-reID [152] are the subset of a surveillance dataset extracted from video footage taken on Duke University's campus in 2014. They annotated every single pedestrian walk passing the camera without their awareness and consent. Their data collection approaches introduced a lot of controversies, and some datasets are currently under investigation. As a result, Duke University has shutdown the DukeMTMC dataset project on 2nd June 2019 and canceled the computer vision surveillance workshop using the DukeMTMC dataset. DukeMTMC-reID dataset as an extension of DukeMTMC has also been removed from the internet. Currently, Market1501 dataset main page has been shutdown. The MSMT17 dataset has to release a new version to mask up the faces of all pedestrian. There is a huge demand for a privacy-aware dataset for future Person Re-ID research.

To address the privacy issue in the current Person Re-ID datasets, we proposed a new Person Re-ID dataset collection method, and we call it privacy-aware user-driven dataset collection strategy. We developed a mobile web app for the Re-ID-Outdoor dataset collection. The design of our web app is demonstrated in Figure 4.15. Every participant needs to read and agree with our privacy policy before register for our data collection. Hence, they are willing to share their face and appearance information for the research and commercial purpose. In the final public version of the dataset, we removed the names and email addresses of all the participants and assign a random number for them to preserve their anonymity.

By running our web application in the smart-phone, the GPS information of the registered participant was analysed in the cell-phone background. When the participant walk passes our surveillance cameras, The timestamps are automatically recorded in the web app. With the actual time log information, it significantly reduces the searching window for our annotators. In addition, we also ask the participant to self-annotate their appearance attributes. Hence, we don't need another round of attributes annotation, which saves us tremendous amount of time.



Time Log (Filled By the APP Automatically)

First Name

Last Name

Gender

Transport/Accessory: Bike

Upper-Body Clothes

Lower-Body Clothes

Main Color Upper

Main Color Lower

Time Log

*Email for Payment Update (Optional)

□ Agree with Our Privacy Policy

Figure 4.15: Mobile web app for the Re-ID-Outdoor dataset collection

There are three main advantages of our privacy-aware user-driven person Re-ID dataset collection strategy:

1. **Privacy Aware:** We only collect images of the registered participants. All registered participants have to accept our privacy policy which allows us to use their appearances for research and commercial purposes.

2. **User Driven:** With our mobile web application, the participant can actively report the time when they are passing each camera. It significantly reduces the annotation difficulty in the large-scale camera network. In addition, the participants also provide us their accurate appearance attributes.

3. **Long-term Person Re-ID:** By using our collection strategy, we can have different appearances of the same person on different days. A new long-term Person Re-ID dataset could be derived from the Re-ID-Outdoor dataset for long-term Person Re-ID research.

### 4.6.4 Comparison With Other Datasets

In this Re-ID-Outdoor dataset collection, 26,175 three-minutes long videos clips have been extracted from the raw surveillance videos. From those video clips, a total of 45,397 bounding box images of pedestrians have been successfully annotated with 40 additional attribute labels. These images are collected from 805 different appearances on multiple-days over an eight week period.

| Datasets | **Re-ID-Outdoor** | MSMT17 | DuketMTMC-reID | Market1501 |
|---|---|---|---|---|
| Surveillance Camera | **Yes** | No | **Yes** | No |
| # Cameras | **50** | 15 | 8 | 6 |
| Collection Period | **8 Weeks** | 4 Days | Single Day | Singel Day |
| Time Variant | **24 Hours**<br>**Day and Night** | Morning<br>Noon<br>Afternoon | - | - |
| # Identities | 805/278 | **4,101** | 1812 | 1501 |
| # BBoxes | 45,397 | **126.441** | 36,411 | 32,668 |
| # distractors | 0 | 0 | >2000 | 2.793 |
| # Attrribute | **40** | - | 23 | 30 |
| People Dectection | YOLO V3 | Faster RCNN | DPM | DPM |

Table 4.2: Detailed comparison with existing large-scale Person Re-ID datasets

Table 4.2 gives a detailed comparison of our Re-ID-Outdoor Dataset with three recent large-scale Person Re-ID datasets: MSMT17 [122], DukeMTMC-reID [152] and Market1501 [147]. Re-ID-Outdoor has more bounding boxes of human images compared to Market1501 and DukeMTMC. Although the MSMT17 dataset has many more identities and human images for training and testing, the limited camera number and horizontal viewing positions cannot fully represent the real-world outdoor surveillance system. On the other hand, the Re-ID-Outdoor dataset is obtained from 50 real surveillance cameras and contains people with different daytime and nighttime views. Hence, it is the most realistic dataset for Person Re-ID tasks at present.

## 4.7 Experiments

### 4.7.1 Datasets and Evaluation Protocol

We evaluate our models on two newly released large-scale datasets: Market1501, DukeMTMC-reID. The Market1501 dataset [147] contains 32,668 images of 1,501 pedestrians. 751 identities are selected for training and 750 remaining identities are for testing. Each identity was captured by at most 6 non-overlapping cameras. The DukeMTMC-reID dataset [152] is the redesigned version of pedestrian tracking dataset DukeMTMC [95] for Person Re-ID tasks. It contains 34,183 image of 1,404 pedestrians. 702 identities are used for training and the remaining 702 are for testing. Each identity was captured by 8 non-overlapping cameras. In our experiments, we follow the proposed single-query evaluation protocols for Market1501 and DukeMTMC-reID [147, 152].

The re-identification of a query image is achieved by ranking the $l_2$ distance of the 2048-D feature embeddings (after the global max-pooling layer) between query and gallery images. To evaluate the performance of these Person Re-ID algorithms, the cumulative matching characteristics (CMC) curve is used in our experiments. The cumulative matching characteristics (CMC) curve is the most common metric used for evaluating person Re-ID performance. This metric is adopted since Re-ID is intuitively posed as a ranking problem, where each image in the gallery is ranked based on its comparison to the probe. The probability that the correct match in the rankings equal to or less than a particular value is plotted against the size of the gallery set [28]. Due to the slow training time of deep learning models, the CMC curve comparisons for recent deep Re-ID methods are simplified to only comparing

Rank 1, Rank 5, Rank 10, Rank 20 retrieval rates.

However, the CMC curve evaluation is valid when only one ground truth match for each given query image. The recent datasets such as Market-1501 and DukeMTMC-reID usually contain multiple ground truth images for the same person. Therefore, Zheng et al. [147] proposed the mean average precision (mAP) as a new evaluation metric. For each query image, the average precision (AP) is calculated as the area under its precision-recall curve. The mean value of the average precision (mAP) will reflect the overall recall of the person Re-ID algorithm. The performances of our models are examined by combining the Rank-1 accuracy for retrieval precision evaluation and mAP for recall evaluation.

### 4.7.2 Performance Comparison of Triplet Loss and NC-Triplet Loss

The first experiment we conducted is to compare the performance difference between the conventional triplet loss and the proposed NC-Triplet loss. We have two different loss function settings for the comparison:

- softmax+triplet loss **vs** softmax+NC-Triplet loss
- softmax+triplet+center loss **vs** softmax+NC-Triplet+center loss

Except for loss function difference, all models are trained with the same training configuration. The overall performance is demonstrated in Table 4.3. Our proposed NC-Triplet consistently improves the mAP from 1.5 to 2%. It proves that the NC-Triplet loss function can further increase the feature distance between the negative image pairs. It allows our Person Re-ID model to generate more discriminative and robust feature embeddings. As the proposed NC-Triplet loss only helps the negative images more distinctive, it has no effect on improving the Rank-1 accuracy. The center loss, on the other hand, improves the Rank 1 accuracy of the model. Overall, by replacing the triplet loss with our proposed NC-Triplet loss and combining with the center loss, the two loss functions are complementary to each other and help the Person Re-ID and achieve better performance.

| Loss Functions | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank 1 | mAP | Rank 1 | mAP |
| softmax + Triplet | 93.2 | 85.3 | 85.9 | 75.3 |
| softmax + NC-Triplet | 93.2 | **86.8** | **86.0** | **76.7** |
| softmax + Triplet + Center | 94.2 | 85.7 | **86.4** | 76.1 |
| softmax + NC-Triplet + Center | **94.3** | **86.5** | 86.3 | **76.7** |

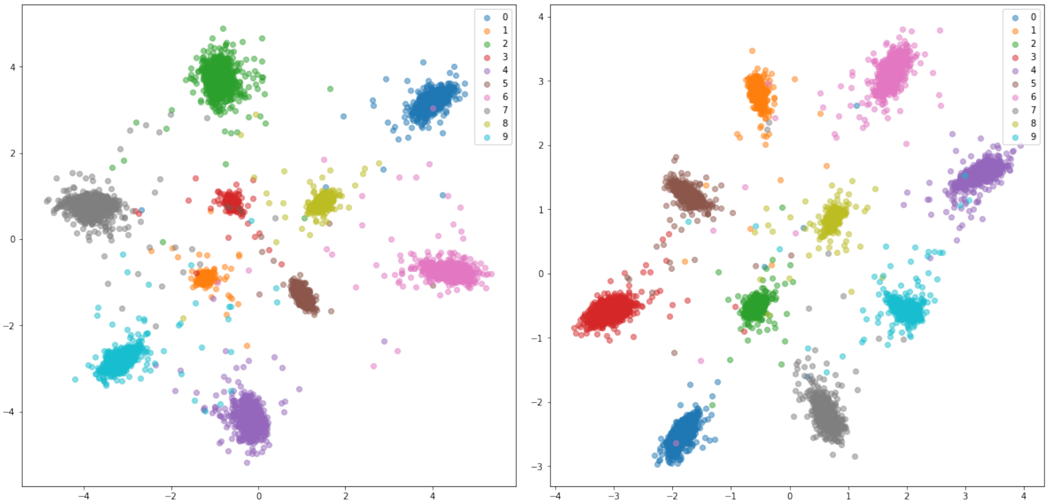
Table 4.3: The performance comparison of triplet loss and NC-Triplet loss.

### 4.7.3 Feature Visualisation: Triplet vs NC-Triplet

To further demonstrate the effectiveness of our proposed NC-triplet loss function, we provide a visualisation of the distributions of the feature embeddings trained with the conventional triplet loss function and our NC-Triplet loss function. Due to a large number of human IDs in the Person Re-ID datasets, it is difficult to visualise hundreds of classes in one t-SNE plot. We use a small MNIST dataset for feature distribution visualisation. The MNIST dataset consists of 60,000 training and 10,000 test images of 10 hand-written digits [43]. We used the same network structure (AlexNet) to learning the feature embedding of the MNIST dataset with the conventional triplet loss function and our NC-Triplet loss function. Figure 4.16 is two t-SNE plots of the feature embeddings after 20 epochs. Compared to the normal triplet loss function, the feature embeddings of the same class learned with the NC-Triplet are more densely compacted. The different classes feature embeddings learned with our NC-Triplet loss function are also more uniformly separated.

### 4.7.4 Comparison with State-of-the-Arts Methods

The results on the Market1501 and DukeMTMC-reID are shown in Table 4.4. All experiments are conducted in a single query setting. The baseline model trained with our proposed loss function outperforms all the classic deep learning approaches such as IDE [149], SVDNet [108] and ImpTripet [11] by a large margin. It is superior to many state-of-the-art methods such as AlignedReID [142], DuATM [99],PCB [126].



(a) Triplet

(b) NC-Triplet

Figure 4.16: The t-SNE visualisation of the feature embeddings from the MNIST dataset trained with (a) triplet loss function and (b) NC-Triplet loss function

It also gives a comparable performance with the most recent BFE method [14] and MGN method [117].

### 4.7.5 Ablation Studies

We also perform extensive experiments on Market-1501 and DukeMTMC-reID datasets to analyse the effectiveness of each training technique used in our model training. Most of the experiments are conducted by using the softmax + triplet loss only.

**Influences of Different Image Sizes**

The first experiment we conducted is to determine the influences of the different input image size. We evaluate our models on both the Market1501 dataset and the DukeMTMC-reID dataset. The Rank 1 accuracy and mean Average Precision (mAP) are reported as evaluation metrics. In this experiment, we did not integrate any other training techniques. The model used for this experiment is trained with only the softmax loss and the original triplet loss with 64 images per batch. As shown in Table 4.5, we can see the performance increases when the image size increases. The

| Method | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank 1 | mAP | Rank 1 | mAP |
| IDE | 79.5 | 59.9 | 67.7 | 47.1 |
| SVDNet | 82.8 | 63.4 | 71.6 | 51.5 |
| ImpTripet | 84.9 | 69.1 | 73.0 | 56.6 |
| AlignedReID | 90.6 | 77.7 | 81.2 | 67.4 |
| DuATM | 91.4 | 76.6 | 81.2 | 62.3 |
| PCB | 93.8 | 81.6 | 83.3 | 69.2 |
| BFE | 94.4 | 85.0 | 88.7 | 75.1 |
| MGN | **95.7** | **86.9** | **88.7** | **78.4** |
| Our | 94.3 | 86.5 | 86.3 | 76.7 |

Table 4.4: Comparison of state-or-the-arts methods.

384 × 128 image size yields the best overall performance. Therefore, we decided to use the 384 × 128 as the default input image size for our model.

| Image Size | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank 1 | mAP | Rank 1 | mAP |
| 256 × 128 | 87.7 | 74.0 | 79.7 | 63.7 |
| 384 × 128 | **88.**1 | **75.4** | **80.2** | **64.1** |

Table 4.5: Performance of our Re-ID models with different image sizes.

### Influences of Different Batch Size

One batch of images for the triplet loss based model includes $B = P \times K$ images. $P$ and $K$ denote the number of different persons and the number of different images per person, respectively. One Nvidia Titan X (12G) GPU used for our experiments cannot contain 128 images per batch. As a result, we limit the maximum number of batch size to 128 (16 × 8). We only test 3 different batch size settings: 32, 64 and 128 with total 5 different configurations, as shown in Table 4.6. In this experiment, we only used the softmax loss with the original triplet loss and did not include any other techniques. The results are presented in Table 4.6. A slight trend we observed is that the larger batch size is beneficial for the model performance. We contend that large $K$ could help to generate hard positive pairs while large $P$ may help to generate hard negative pairs.

| Batch Size | Positive $A\times$ Negative | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|---|
| | | Rank 1 | mAP | Rank 1 | mAP |
| 32 | $8 \times 4$ | 87.5 | 74.3 | 79.9 | 64.2 |
| 64 | $8 \times 8$ | 88.0 | 75.1 | 80.2 | 64.1 |
| 64 | $16 \times 4$ | 88.1 | 75.4 | 80.5 | 65.0 |
| 128 | $16 \times 8$ | 88.7 | 76.1 | 80.7 | 65.4 |
| 128 | $32 \times 4$ | 88.5 | 75.9 | 81.1 | 65.5 |

Table 4.6: Performance of our Re-ID models with different batch sizes.

### Influences of GAP and GMP

Many Person Re-ID models proposed recently use the GAP layer after the CNN backbone. The person re-identification task requires the model to capture the most distinctive feature of a person. We contend that the GMP is more suitable for this problem because it emphasises the semantic regions from the feature maps. In this experiment, we trained two different models. One uses GAP layer, and another one uses GMP layers. By replacing GAP with GMP, we can see an overall 1% performance gains in terms of both Rank 1 accuracy and mAP. As a result, we decided to use the global max-pooling (GMP) for our model.

| Global Pooling | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank 1 | mAP | Rank 1 | mAP |
| GAP | 88.1 | 75.4 | 80.5 | 65.0 |
| GMP | **89.3** | **76.4** | **80.8** | **66.1** |

Table 4.7: Performance comparison of our Re-ID models with GAP layer and GMP layer.

**Influences of Other Training Techniques**

In this experiment, we integrate the warm-up strategy, random erasing augmentation, changing the last stride to 1, adding Bottleneck layer and label smoothing process into our baseline model, one by one. Table 4.8 demonstrates the incremental performance gain for each technique on the Market1501 dataset and the DukeMTMC-reID dataset. Most of the techniques increase the CMC and mAP of our model by 0.5% to 1%.

| techniques | Market1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank 1 | mAP | Rank 1 | mAP |
| Baseline | 89.3 | 76.4 | 80.8 | 66.1 |
| +Warm Up | 89.8 | 79.8 | 81.5 | 67.7 |
| +Random Erasing | 90.2 | 80.3 | 82.2 | 69.0 |
| +Last Stride = 1 | 91.7 | 81.9 | 83.6 | 70.6 |
| +Bottleneck Layer = 1 | **92.4** | **84.9** | **85.1** | **73.1** |

Table 4.8: The performance gain by the techniques.

## 4.8 Performance on Our Re-ID-Outdoor Dataset

We also tested our NC-Triplet model with two state-of-the-art methods (MGN [117] and AANet [113]) on our Re-ID-Outdoor dataset. The performance comparison is shown in Table 4.9.

| Dataset | Re-ID-Outdoor | | MSMT17 | | DukeMTMC-reID | | Market1501 | |
|---|---|---|---|---|---|---|---|---|
| | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP |
| NC-Triplet | **66.0**% | **35.7**% | 74.6% | 47.3% | 86.3% | 76.7% | 94.3% | 86.5% |
| MGN | 63.2% | 26.2% | - | - | 88.7% | 78.4% | 95.7% | 86.9% |
| AANet | 65.4% | 28.5% | - | - | 87.7% | 74.3% | 93.9% | 83.4% |

Table 4.9: Performance comparison of state-of-the-art methods on different datasets

Our NC-Triplet is a relatively simple architecture which yields good performance on all MSMST17, DukeMTMC-reID and Market1501 dataset. Based on the performance of our NC-Triplet model, the new Re-ID-Outdoor give the lowest

Rank 1 and mAP scores compared to the other three datasets. It indicates that the Re-ID-Outdoor dataset is the most challenging Person Re-ID dataset so far. The MGN approach is one of the best Person Re-ID model currently with the highest Rank 1 and mAP scores on the Market1051 dataset and the DukeMTMC-reID dataset. It can only achieve the 63.2% Rank 1 accuracy on the Re-ID-Outdoor dataset. The AANet approach combines the attribute attention mechanism with a part-base pooling technique, which helps to boost the overall generalisation ability of the Person Re-ID model. By integrating with the 40 attributes information from the Re-ID-Outdoor, the AANet model can achieve over 65% Rank 1 accuracy. However, this figure is far lower than what AANet can achieve on the other two datasets. This again suggests the challenging nature of the RE-ID Outdoor dataset. Moreover, only the NC-triplet methods can achieve over 30% mAP scores in the Re-ID-Outdoor dataset.

### 4.8.1 Performance for Cross-dataset Scenario

To further explore the effectiveness of our baseline model, we also conducted a cross-dataset experiment, as shown in Table 4.10. Our model can achieves 27.7% and 47.4% Rank 1 accuracy on the Market1501 to DukeMTMC-reID and the DukeMTMC-reID to Market1501 settings.

| Market1501 → DukeMTMC-reID | | DukeMTMC-reID → Market1501 | |
|---|---|---|---|
| Rank 1 | mAP | Rank 1 | mAP |
| 29.7 | 15.0 | 47.4 | 21.1 |

Table 4.10: The performance of our best model in a cross-dataset scenario. Market1501 → DuketMTMC-reID means that the model was trained on the Market1501 dataset and evaluate on the DukeMTMC-reID dataset, vice versa.

## 4.9 Conclusion

In this chapter, we proposed a novel negative competing triplet loss (NC-Triplet), which helps to discriminate the negative sample pairs further and significantly boost the overall mAP score of many existing models. We also collected a more realistic and challenging Person Re-ID dataset called: Re-ID-Outdoor. It is the first privacy-aware Person Re-ID dataset. We conducted extensive experiments to demonstrate the high performance of our NC-Triplet Person Re-ID models. Finally, only using global features, our model can achieve 94.3% Rank 1 accuracy and 86.3% mAP on Market1501 and yield the best result for our Re-ID-Outdoor dataset. Although our Person Re-ID method can achieve impressive performance in the supervised learning framework, the features extracted from our model still show poor performance on the cross-dataset scenario. In the next chapter, we will alleviate this issue by proposing a cross-dataset Person Re-ID model based on a novel domain adaptation strategy.

# Chapter 5

# Cross-Dataset Feature Adaptation

## 5.1 Introduction

In Chapter 3 and Chapter 4, we focused on creating deep learning models trained from a single dataset in a fully supervised manner. However, similar to many other Person Re-ID approaches [11, 14, 72, 142], the two models proposed also require a large number of manually labelled datasets for learning the view-invariant feature representation or the robust matching function. In the real-world Person Re-ID application, a typical surveillance system usually consists of over one hundred cameras. Manual annotating images from hundreds of cameras is prohibitively expensive. On the other hand, if we directly deploy a model trained from a public dataset to a new system, it usually sufferers from considerable performance degradation. In Chapter 4, our model trained from the Market1501 dataset [147] can achieve 94% Rank 1 retrieval accuracy. However, when we test the same model on the DukeMTMC-reID dataset [152], it can only achieve 37% Rank 1 accuracy. The limited scalability severely hinders the applicability of the single-dataset supervised Person Re-ID approaches in the real-world scenarios. One solution to make a Person Re-ID model scaleable is designed an unsupervised algorithm which can train Person Re-ID models directly from the unlabelled data.

In recent years, some unsupervised methods have been proposed to extract the view-invariant features and measure the similarity of images without label information [40, 118, 119, 139]. Figure 5.1 demonstrates an example of a general clustering-based unsupervised method. It analyses the unlabelled dataset and partitions them into multiple clusters with the corresponding pseudo labels. These unsupervised approaches [40, 118, 119, 139] generally yield poor Person Re-ID performance due to the lack of active supervised tuning and optimisation.

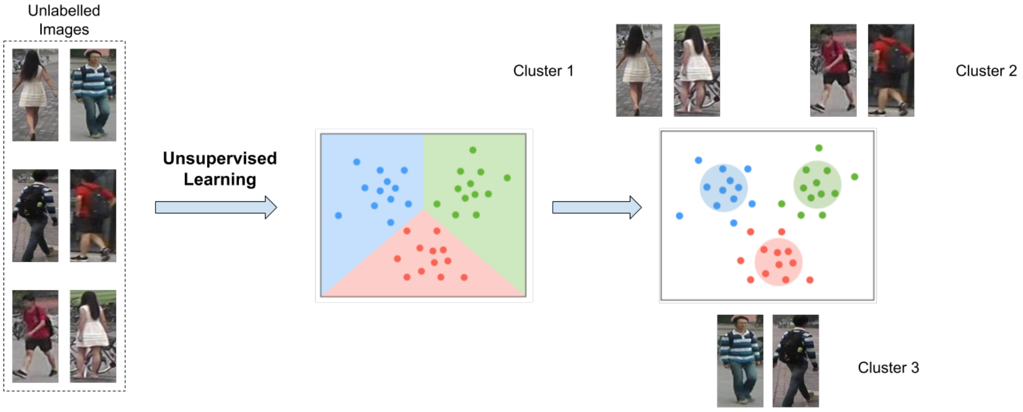


Figure 5.1: Unsupervised feature learning

There are many Person Re-ID datasets available for training, and the unlabelled data can also be easily obtained from a new camera network. In this chapter, we address the scalability issue of Person Re-ID via an unsupervised cross-dataset domain adaptation strategy. Figure 5.2 illustrates our unsupervised cross-dataset domain adaptation framework. We leverage labelled data from an existing dataset (known as the source domain) for training a base model. By analysing the properties of the unlabelled images obtained from a new surveillance system (known as the unlabelled target domain), the model will be modified to adapt to the new system. The labelled source dataset images (source domain) help the model to learn a strong feature representation and provide a foundation for the domain adaptation. The unlabelled target system images (target domain) guide the model to perform the domain level fine-tuning and boost the performance for the cross-dataset scenario.

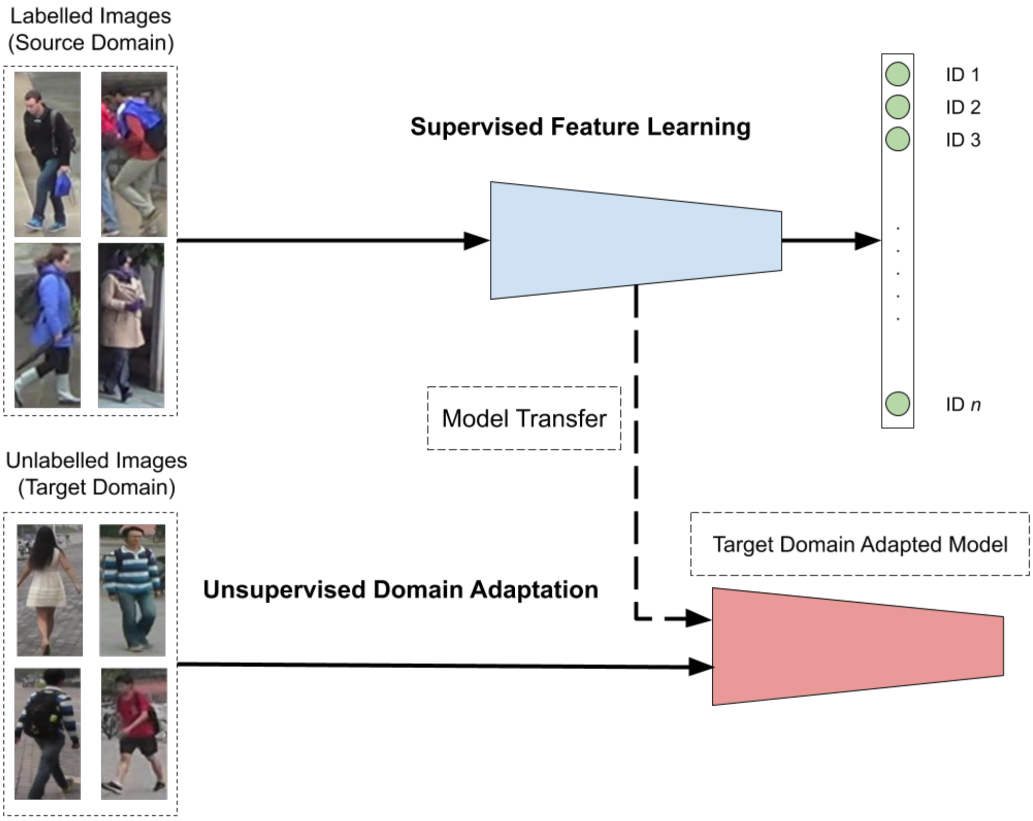


Figure 5.2: Unsupervised domain adaptation

## 5.2 Problem Definition

The images in a Person Re-ID dataset are usually taken under similar conditions, such as camera setting, environment, weather. As a result, these features extracted from one particular dataset tend to form a compact statistical distribution. However, different datasets are collected under different conditions. For example, Market1501 dataset [147] was collected at Tsinghua University, China during the summer. DukeMTMC was collected at Duke University, USA during the winter time. So people's appearances and outfits in these two datasets are very different, as shown in Figure 5.3. As a result, there will be a large distributions difference between the two datasets. One primary goal of domain adaption is to reduce the difference between the distributions of the source and target domain data, as shown in Figure 5.4.

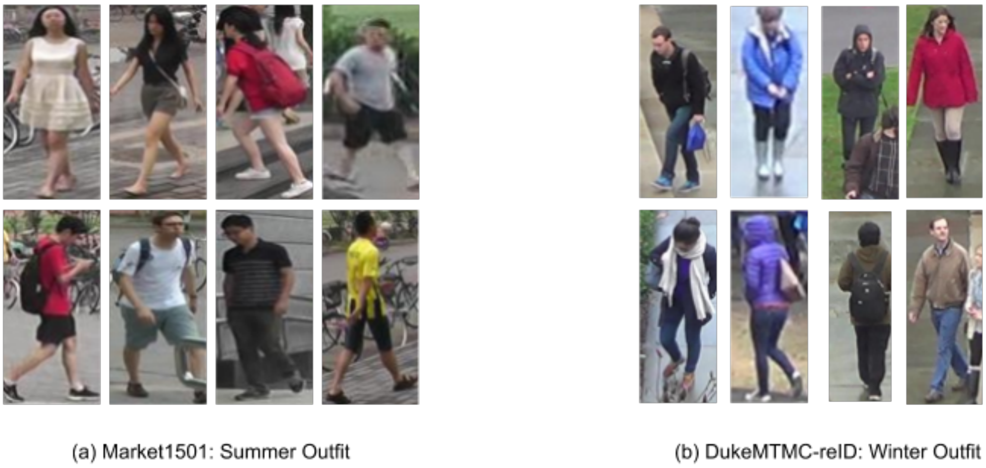

(a) Market1501: Summer Outfit
(b) DukeMTMC-reID: Winter Outfit

Figure 5.3: Majority of the people in the Market1501 dataset are wearing the summer outfits. The DukeMTMC-reID dataset only contains the winter outfit appearances.

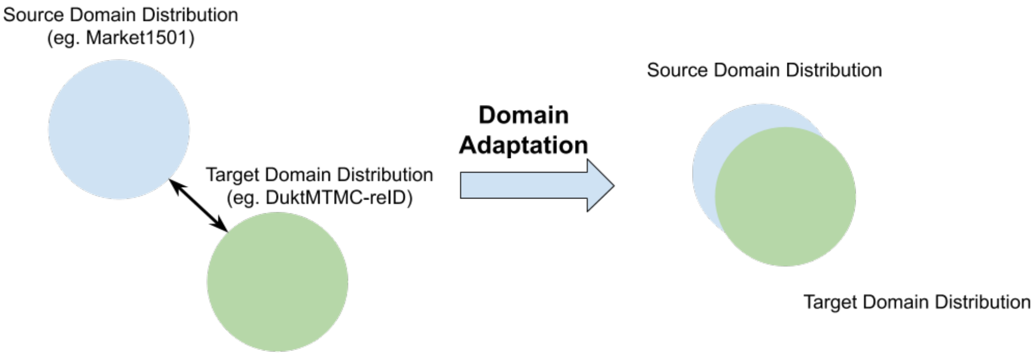

Source Domain Distribution
(eg. Market1501)
Target Domain Distribution
(eg. DuktMTMC-reID)
Domain
Adaptation
Source Domain Distribution
Target Domain Distribution

Figure 5.4: Domain Adaptation aims to reduce the distributions different between the source and the target domain.

Most domain adaptation frameworks [68, 69] assume that the source domain and the target domain contain the same set of class labels. Such an assumption does not hold for person Re-ID because different Re-ID datasets usually contain completely different sets of persons (classes). The pedestrian with ID 1 label in the Market1501 dataset and the person with ID 1 label in the DukeMTMC-reID dataset are two different individuals. Therefore, most of the unsupervised cross-dataset Re-ID methods proposed in recent years [16, 120, 122] did not use conventional domain adaptation mechanisms. For example, [16] uses image-to-image translation to transfer the style of images in the target domain to the source domain images for generating a new training dataset. These newly generated samples which inherit the identity labels from the source domain and the image style of the target domain can be used for supervised Person Re-ID learning. [120] trains two individual models: identity classification and attribute recognition and performs the domain adaptation between two models.

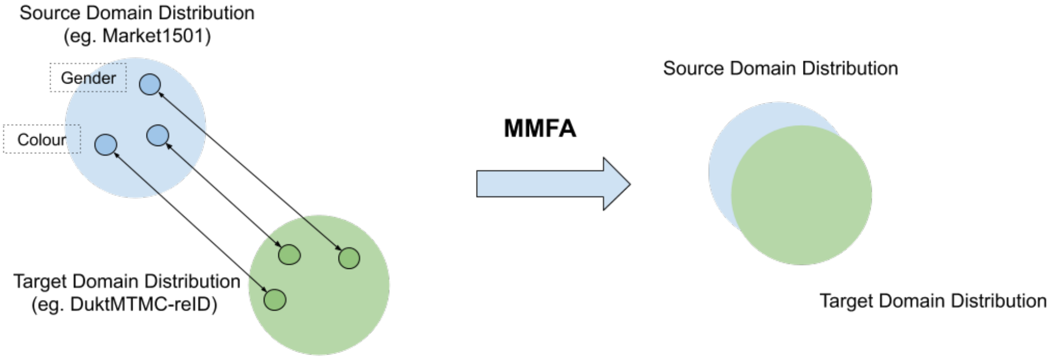

Source Domain Distribution
(eg. Market1501)
Gender
Colour
Target Domain Distribution
(eg. DuktMTMC-reID)
MMFA
Source Domain Distribution
Target Domain Distribution

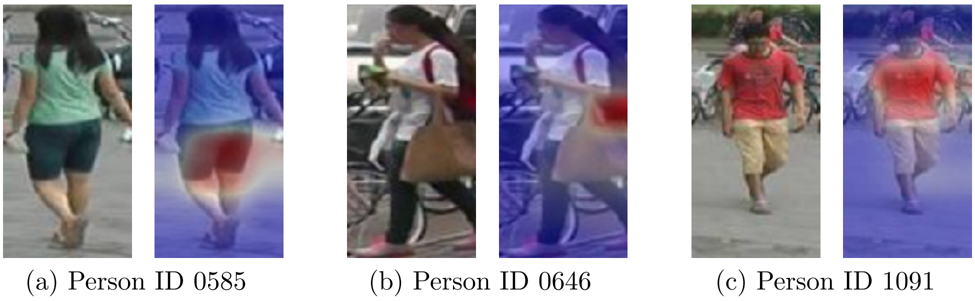
Figure 5.5: MMFA reduces the domain distributions based on the mid-level attributes such as gender and colour of clothing.

In our work, we reformated the assumption made by the unsupervised cross-dataset Re-ID. Although the identity labels of the source and target datasets are non-overlapping, many of the mid-level semantic features of people such as genders, age-groups or colour and texture of the outfits are commonly shared between different people across different datasets. Hence, these mid-level visual attributes of people can be considered as the common labels between different datasets. If we assume these mid-level semantic features are shared between the different domains, we can then treat the unsupervised cross-dataset person Re-ID as a domain adaptation transfer learning based on the mid-level semantic features from the source domain to the target domain, as shown in Figure 5.5. Therefore, we proposed a **M**ulti-task **M**id-level **F**eature **A**lignment network (MMFA) which can simultaneously learn the feature representation from the source dataset and perform domain adaptation to the target dataset via aligning the distributions of the mid-level features. The contributions of our MMFA model are summarised below:

- We propose a novel unsupervised cross-dataset domain adaptation framework for Person Re-ID, which minimises the distribution variation of the source's and the target's mid-level features based on the MMD distance [31]. Due to the low dimensionality of attribute annotations, we also include mid-level feature maps in our deep neural network as additional latent attributes to capture a more completed representation of mid-level features of each domain. In our experiments, the proposed MMFA method surpasses other state-of-the-art unsupervised models on four popular unsupervised benchmarks datasets.
- The existing unsupervised domain adaptation Re-ID approaches based on deep learning [16, 120] require two-stage learning processes: supervised feature learning and unsupervised domain adaptation. Different from those methods, our MMFA model introduces a new jointly training structure which simultaneously learns the feature representation from the source domain and adapts the feature to the target domain in a single training process. Because our model does not require a two-step training procedure, the training time for our method is much less than many other unsupervised deep learning person Re-ID approaches.

## 5.3 The Proposed Methodology

One basic assumption behind domain adaptation is that there exists a feature space which is commonly shared between the source and the target domains. Although high-level information like a person's identity is not shared between different Re-ID datasets, the mid-level features such as visual attributes can be overlapped between datasets. For example, the people in dataset A and dataset B are different people with different ID labels, but some of the mid-level semantic information like genders, age-groups, the colour of clothes or accessories could be similar. Hence, in our



(a) Person ID 0585 (b) Person ID 0646 (c) Person ID 1091

Figure 5.6: In each of these three pairs of images, the one on the left-hand side is randomly selected from the Market1501 dataset while the other one shows the attention regions from highest activated feature maps ($1749_{th}$, $511_{th}$ and $1091_{th}$) of the last convolutional layer. These feature maps highlight distinctive semantic features such as green shorts, a red backpack, a red T-shirt. Best view in colour.

proposed method MMFA, we assume that the source and the target datasets contain the same set of mid-level attribute labels. As a result, the unsupervised cross-dataset person Re-ID can be transformed into an unsupervised domain adaptation problem by regularising the distribution variance of the attribute feature space between the source domain and the target domain.

Currently, there are a few attribute annotations available for some Re-ID datasets. However, the number of these attribute labels are limited. There are 27 attribute labels for the Market1501 dataset and 23 for the DukeMTMC-reID dataset [65]. The features obtained from 27 or 23 user-defined attributes alone cannot give a good representation of the overall mid-level semantic features for both source and target datasets. There may exist many shared mid-level visual clues between domains which cannot be fully captured by those 27/23 user-defined annotations. To obtain more attributes for representing the shared mid-level features, we start to consider the feature-maps generated from the different convolutional layers. In our experiment, we observed that those highly activated feature maps from the last convolutional layer of an attribute-identity multi-task classification model could capture many distinctive semantic features of a person, see Figure 5.6 for example. Hence, we treat those feature maps as the attribute-like mid-level deep features in our proposed MMFA model.

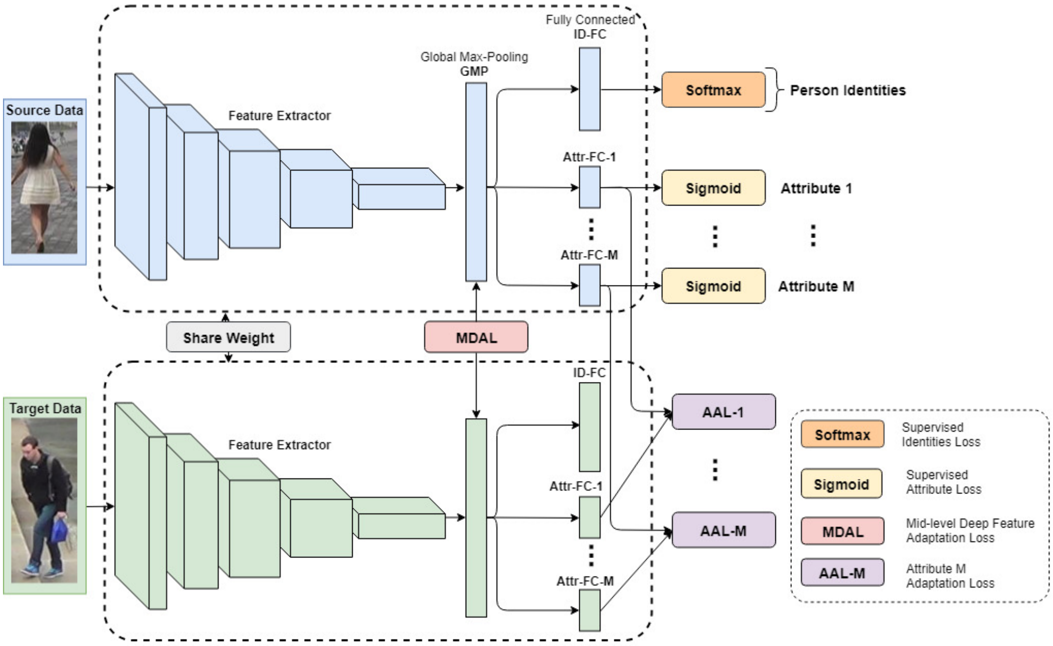


Figure 5.7: The network architecture of the proposed MMFA model

### 5.3.1 Architecture

Our model is optimised by using Adam optimiser on mini-batches [39]. Each mini-batch consists of $n_S$ of labelled images $[\mathbf{I}_{S,1}, \mathbf{I}_{S,2}, ..., \mathbf{I}_{S,n_S}]$ from a source dataset $S$ and $n_T$ unlabelled images $[\mathbf{I}_{T,1}, \mathbf{I}_{T,2}, ..., \mathbf{I}_{T,n_T}]$ from a target dataset $T$. Each labelled image $\mathbf{I}_{S,i}$ is associated with an identity label $y_{S,i}$ and a set of M attributes $\mathbf{A}_{S,i} = [a^1_{S,i}, a^2_{S,i}, ..., a^M_{S,i}]$. Our model consists of one pre-trained ResNet50-based backbone network [34] as the feature extractor with one fully connected layer for identity classification and $M$ individual fully connected layers for single attribute recognition. The overview of our architecture is shown in Figure 5.7. Based on the experimental results in Chapter 4, we change the last average pooling layer from ResNet50 to a global max-pooling (GMP) layer. By taking the maximum value from each feature map, the network can focus these highly activate feature maps.

$\mathbf{H}_S = [\mathbf{h}_{S,1}, \mathbf{h}_{S,2}, ..., \mathbf{h}_{S,n_S}]$ and $\mathbf{H}_T = [\mathbf{h}_{T,1}, \mathbf{h}_{T,2}, ..., \mathbf{h}_{T,n_T}]$ are the mid-level deep features of the inputs from both the source domain and the target domain obtained after the GMP layer, respectively. The identity features $\mathbf{H}^{id}_S =$

$[\mathbf{h}_{S,1}^{id}, \mathbf{h}_{S,2}^{id}, ..., \mathbf{h}_{S_{n,S}}^{id}]$ and $\mathbf{H}_T^{id} = [\mathbf{h}_{T,1}^{id}, \mathbf{h}_{T,2}^{id}, ..., \mathbf{h}_{T,n_T}^{id}]$ are the outputs from the fully connected layer with $\mathbf{H}_S$ and $\mathbf{H}_T$ as inputs for identity classification (shown as *ID-FC* in Figure 5.7). For a specific $m$-th attribute where $m \in M$, the $m$-th attribute features $\mathbf{H}_S^{attr_m} = [\mathbf{h}_{S,1}^{attr_m}, \mathbf{h}_{S,2}^{attr_m}, ..., \mathbf{h}_{S,n_S}^{attr_m}]$, $\mathbf{H}_T^{attr_m} = [\mathbf{h}_{T,1}^{attr_m}, \mathbf{h}_{T,2}^{attr_m}, ..., \mathbf{h}_{T,n_T}^{attr_m}]$ can be obtained from its corresponding fully connected layer with $\mathbf{H}_S$ and $\mathbf{H}_T$ as input (shown as *Attr-FC-m* in Figure 5.7). Our model can be jointly trained in a multi-task manner: two supervised classification losses for identity classification and attribute recognition, one adaptation losses based on the attribute features and another adaptation loss based on the mid-level deep features.

### 5.3.2 Multi-task Supervised Classification for Feature Learning

The view-invariant feature representations are learned from a multi-task identity and attribute classification training. The additional attribute annotations provide further regularisation and additional supervision to the feature learning process.

**Identity Loss:** We denote that $p_{id}(\mathbf{h}_{S,i}^{id}, y_{S,i})$ is the predicted probability on the identity feature $\mathbf{h}_{S,i}^{id}$ with the ground-truth label $y_{S,i}$. The identity loss is computed according to the softmax cross entropy function:

$$L_{id} = -\frac{1}{n_S}\sum_{i=1}^{n_S} log(p_{id}(\mathbf{h}_{S,i}^{id}, y_{S,i})) \tag{5.1}$$

**Attribute Loss:** We denote that $p_{attr}(\mathbf{h}_{S,i}^{attr_m}, m)$ is the predicted probability for the $m$-th attribute feature $\mathbf{h}_{S,i}^{attr_m}$ with ground-truth label $a_{S,i}^{m}$. The overall attributes loss can be expressed as the average of sigmoid cross entropy loss of each attribute:

$$\begin{aligned} L_{attr} = -\frac{1}{M}\frac{1}{n_S}\sum_{m=1}^{M}\sum_{i=1}^{n_S}(a_{S,i}^{m} \cdot log(p_{attr}(\mathbf{h}_{S,i}^{attr_m}, m)) \\ + (1 - a_{S,i}^{m}) \cdot log(1 - p_{attr}(\mathbf{h}_{S,i}^{attr_m}, m))) \end{aligned} \tag{5.2}$$

### 5.3.3 MMD-based Regularisation for Mid-level Feature Alignment

As we make a shared mid-level latent space assumption in our MMFA model, the domain adaptation can be achieved by reducing the distribution distance of attribute features between the source domain and the target domain. Based on the attribute features $\{\mathbf{H}_S^{attr_1}, .., \mathbf{H}_S^{attr_M}\}$ and $\{\mathbf{H}_T^{attr_1}, .., \mathbf{H}_T^{attr_M}\}$ obtained from the supervised classification learning, we use the MMD measure [31] to calculate the feature distribution distance of each attribute. The overall attribute distribution distance is the mean MMD distance of all attributes:

$$\begin{aligned} L_{AAL} &= \frac{1}{M}\sum_{m=1}^{M} MMD(\mathbf{H}_S^{attr_m}, \mathbf{H}_T^{attr_m})^2 \\ &= \frac{1}{M}\sum_{m=1}^{M} \left\| \frac{1}{n_S}\sum_{i=1}^{n_S}\phi(\mathbf{h}_{S,i}^{attr_m}) - \frac{1}{n_T}\sum_{j=1}^{n_T}\phi(\mathbf{h}_{T,j}^{attr_m}) \right\|_{\mathcal{H}}^2 \end{aligned} \tag{5.3}$$

$\phi(\cdot)$ is a map operation which projects the attribute distribution into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [30]. $n_S$ and $n_T$ are the batch sizes of the source domain images and target domain images. The arbitrary distribution of the attribute features can be represented by using the kernel embedding technique [100]. It has been proven that if the kernel $k(\cdot,\cdot)$ is characteristic, then the mapping to the RKHS $\mathcal{H}$ is injective [103]. The injectivity indicates that the arbitrary probability distribution is uniquely represented by an element in RKHS. Therefore, we have a kernel function $k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{T,j}^{attr_m}) = \phi(\mathbf{h}_{S,i}^{attr_m})\phi(\mathbf{h}_{T,j}^{attr_m})^\intercal$ induced by $\phi(\cdot)$. Now, the average MMD distance between the source domain's and the target domain's attribute distributions can be re-expressed as:

$$\begin{aligned} L_{AAL} = \frac{1}{M}\sum_{m=1}^{M} \Big[ &\frac{1}{(n_S)^2}\sum_{i=1}^{n_S}\sum_{i'=1}^{n_S} k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{S,i'}^{attr_m}) \\ &+ \frac{1}{(n_T)^2}\sum_{j=1}^{n_T}\sum_{j'=1}^{n_T} k(\mathbf{h}_{T,j}^{attr_m}, \mathbf{h}_{T,j'}^{attr_m}) \\ &- \frac{2}{n_S \cdot n_T}\sum_{i=1}^{n_S}\sum_{j=1}^{n_T} k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{T,j}^{attr_m}) \Big] \end{aligned}$$

In our MMFA model, the commonly used Radial basis function (RBF) characteristic kernel with bandwidth $\alpha$ is used as the kernel function for computing the MMD distance [50]:

$$k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{T,j}^{attr_m}) = exp(-\frac{1}{2\alpha} \left\| \mathbf{h}_{S,i}^{attr_m} - \mathbf{h}_{T,j}^{attr_m} \right\|^2) \tag{5.4}$$

Due to the limited size of available attribute annotations, these attributes alone cannot give a good representation of all domain-shared mid-level features. By assuming the last feature maps after the feature extractor is attribute-like mid-level features, we introduce the additional mid-level deep feature alignment to our model. The mid-level deep features adaptation loss $L_{MDAL}$ is the MMD distance between the source and the target mid-level deep features $\mathbf{H}_S, \mathbf{H}_T$, similar to our attributes feature adaptation loss:

$$L_{MDAL} = MMD(\mathbf{H}_S, \mathbf{H}_T)^2 = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(\mathbf{h}_{S,i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \phi(\mathbf{h}_{T,j}) \right\|_{\mathcal{H}}^2 \tag{5.5}$$

Finally, we formulate the overall loss function by incorporating the weighted summation of above components $L_{id}$, $L_{attr}$, $L_{AAL}$ and $L_{MDAL}$:

$$L_{all} = L_{id} + \lambda_1 L_{attr} + \lambda_2 L_{AAL} + \lambda_3 L_{MDAL} \tag{5.6}$$

## 5.4 Experiments

### 5.4.1 Datasets and Settings

#### Person Re-ID Datasets

Four widely used person Re-ID benchmarks are chosen for experimental evaluations: Market1501, DukeMTMC-reID, VIPeR and PRID. The Market-1501 dataset [147] contains 32,668 images of 1,501 pedestrians. 751 identities are selected for training and 750 remaining identities are for testing. Each identity was captured by at most

6 non-overlapping cameras. The DukeMTMC-reID dataset [152] is the redesigned version of pedestrian tracking dataset DukeMTMC [95] for person Re-ID task. It contains 34,183 image of 1,404 pedestrians. 702 identities are used for training and the remaining 702 are for testing. Each identity was captured by 8 non-overlapping cameras. The VIPeR dataset [28] is one of the oldest person Re-ID dataset. It contains 632 identities, but only two images for each identity. Due to its low resolution and large variation in illumination and viewpoints, the VIPeR dataset is still a very challenging dataset. The PRID dataset [35] consists of 934 identities from two camera views. There are 385 identities in View A and 749 identities in View B, but only 200 identities appear in both views.

### Person Re-ID Attributes

For Market-1501, there are 27 labelled attributes: gender (male, female), hair length (long, short), sleeve length (long, short),length of lower-body clothing (long, short), type of lower-body clothing (pants, dress), wearing hat (yes, no), carrying bag (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), 8 colours of upper-body clothing (black, white, red, purple, yellow, grey, blue, green), 9 colours of lower-body clothing (black, white, pink, purple, yellow, grey, blue, green, brown) and age (child, teenager, adult, old). For DukeMTMC-reID, 23 labelled attributes are provided: gender (male,female), shoe type (boots, other shoes), wearing hat (yes, no), carrying bag (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), colour of shoes (dark, light), length of upper-body clothing (long, short), 8 colours of upper-body clothing (black, white, red, purple, grey, blue, green, brown) and 7 colours of lower-body clothing (black, white, red, grey, blue, green, brown). Figure 5.8 displayed some sample images and their corresponding attribute labels.

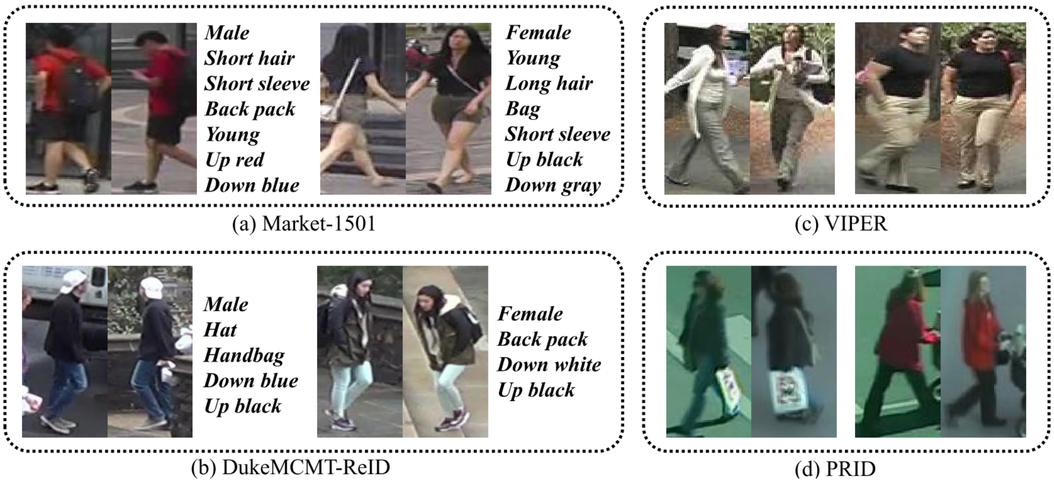


Figure 5.8: Example of person images and attribute labels. Each pair represents two images of the same person.

### Evaluation Protocol

We follow the proposed single-query evaluation protocols for Market1501 and DukeMTMC-reID. For the VIPeR dataset, we randomly half-split the dataset into training and testing sets. The overall performance on VIPeR is the average results from 10 randomly 50/50 split testing. For the PRID dataset evaluation, we follow the same single-shot experiments as [141]. Similar to the VIPeR dataset setting, the final performance is the average of the experimental results based on 10 random split testing. Since the VIPeR and PRID datasets are too small for training the deep learning network, our MMFA model trains on the entire Market1501 or the DukeMTMC-reID datasets. Similar to the experiment setting in Chapter 4, the Rank 1,5,10 retrieval accuracy and mean Average Precision (mAP) is used to evaluate the performance of our MMFA model.

### Implementation Details

The input images are randomly cropped and re-sized to (256,128,3). All the fully-connected layers after global max-pooling layer are equipped with batch normalisation, the dropout rate of 0.5 and the leaky RELU activation function. For all the adaptation losses, we adopted the same mixture kernel strategy proposed by Li et al. [50] by averaging the RBF kernels with the bandwidth $\alpha = 1, 5, 10$. For Adam optimiser, we use the same hyperparameter setting from our best supervised Re-ID model in Chapter 4. The first 10 epochs linearly increase the learning rate from $3.5 \times 10^{-6}$ to $3.5 \times 10^{-4}$. Then, the learning rate is decreased to $3.5 \times 10^{-5}$ and $3.5 \times 10^{-6}$ at 40th and 70th epoch, respectively. Totally, there are 120 training epochs. The person Re-ID evaluation of the target domain is measured by the $l_2$ distance of the 2048-D mid-level deep features $H_T$ after the global max-pooling layer.

### 5.4.2 Parameter Validation

We first conducted several experiments to determine the best combination of parameter $\lambda 1$, $\lambda 2$ and $\lambda 3$ in the final loss function (Equation 5.6). Parameter $\lambda 1$ determines the contribution of the attributes recognition loss. The value of $\lambda 1$ can be set based on the performance of the model training and testing on the Market1501 dataset. The performance variation on the Market1501 testing dataset is illustrated in Figure 5.9. The $\lambda_1 = 0.4$ yields the best performance for both Rank 1 accuracy and mAP scores. Therefore, we fixed $\lambda_1 = 0.4$ for the following experiments.

The $\lambda 2$ and $\lambda 3$ parameters are used for the unsupervised domain adaptation. The following experiments are used for analysis performance variation with different $\lambda 2$ and $\lambda 3$ values on the target datasets. We use the Market1501 as the source dataset and evaluate the performance on the DuketMTMC-reID dataset.

In the first experiment, we fix the $\lambda_3$ value to 1 and only change the $\lambda_2$ value from 0 to 2 with 0.2 incremental steps. There is a slight increasing in performance when $\lambda_2 < 1$. The performance researches the peak when $\lambda_2$ is between 0.8 to 1.2. In the second experiment, we fix the $\lambda_2$ to 1 and only change the $\lambda_3$ value from 0 to 2 with 0.2 increments. We observed a quick increase in both Rank 1 and mAP when $\lambda_3$ is increasing from 0 to 1.2. As a result, $\lambda 1$, $\lambda 2$ and $\lambda 3$ in the final loss function (Equation 5.6) are empirically fixed to $0.4, 1, 1$. Besides, our MMFA model is more sensitive to the value of $\lambda 3$. Since the $\lambda 3$ controls the weight of the mid-level deep feature alignment loss ($L_{MDAL}$), we contend that deep mid-level feature contributes

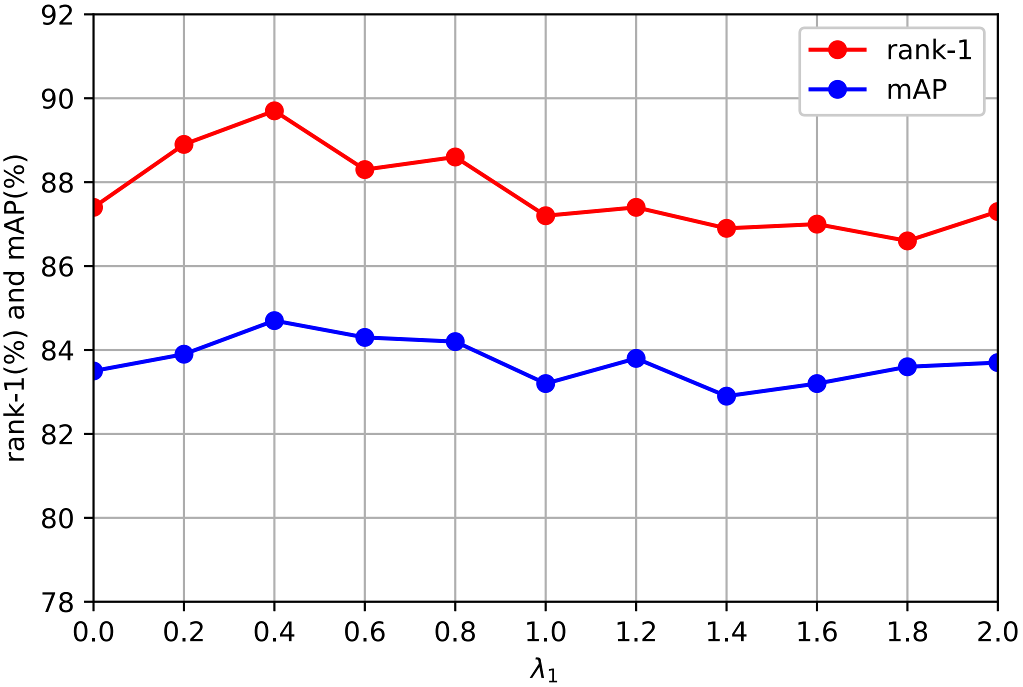


Figure 5.9: The Person Re-ID performance (Rank 1 accuracy and mAP) on the testing set of Market-1501 when parameter $\lambda_1$ varies.

more for aligning the source and the target domain.

### 5.4.3 Comparisons with State-of-the-Art Methods

The performance of our proposed MMFA model is extensively compared with 16 state-of-the-art unsupervised person Re-ID methods as shown in Table 5.1. These methods include: view-invariant feature learning methods SDALF [19] and CPS [12], graph learning method GL [41], sparse ranking method ISR [66], salience learning methods GTS [118] and SDC [146], neighbourhood clustering methods AML [137], UsNCA [91], CAMEL [139] and PUL [17], ranking SVM method AdaRSVM [73], attribute co-training method SSDAL [106], dictionary learning method DLLR [40] and UDML [89], id-to-attribute transfer method TJ-AIDL [120] and image style transfer method SPGAN [16]. These methods can be categorised into three groups:

1. hand-craft features approaches: SDALF,CPS,DLLR,GL,ISR,GTS,SDC

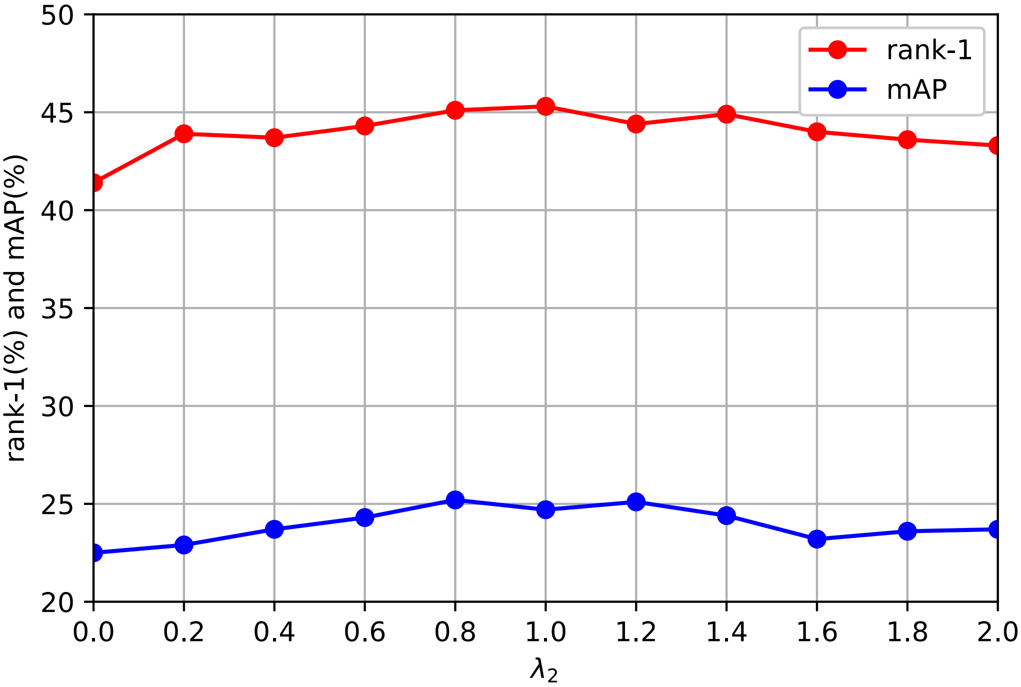


Figure 5.10: The Market1501 trained model performance (Rank 1 accuracy and mAP) on the DukeMTMC dataset with different parameter $\lambda_2$ values ($\lambda_3$ is fixed to 1)

2. clustering approaches: AML, UsNCA, CAMEL, PUL
3. domain adaptation approaches: AdaRSVM, UDML, SSDAL, TJ-AIDL, SP-GAN

Our MMFA method outperforms most existing state-of-the-art models on VIPeR, PRID, Market1501 and DukeMTMC-reID datasets. the Rank 1 accuracy increases from 38.5% to 39.1% in VIPeR, from 34.8% to 35.1% in PRID and from 44.3% to 45.3% in DukeMTMC-reID. The mAP performance of our approach surpasses all existing methods by a good margin from 23.0% to 24.7% and 26.5% to 27.4% in DukeMTMC-reID and Market1501 receptively. Although the Rank-1 accuracy of our MMFA model on the Maket1501 dataset did not surpass the TJ-AIDL method, our mAP score and the overall performance (Rank-5 to Rank-10 accuracy) are better than TJ-AIDL. The complete comparisons with TH-AIDL and SPGAN are shown in Table 5.2. It is worth noting that the performance of our MMFA is achieved in

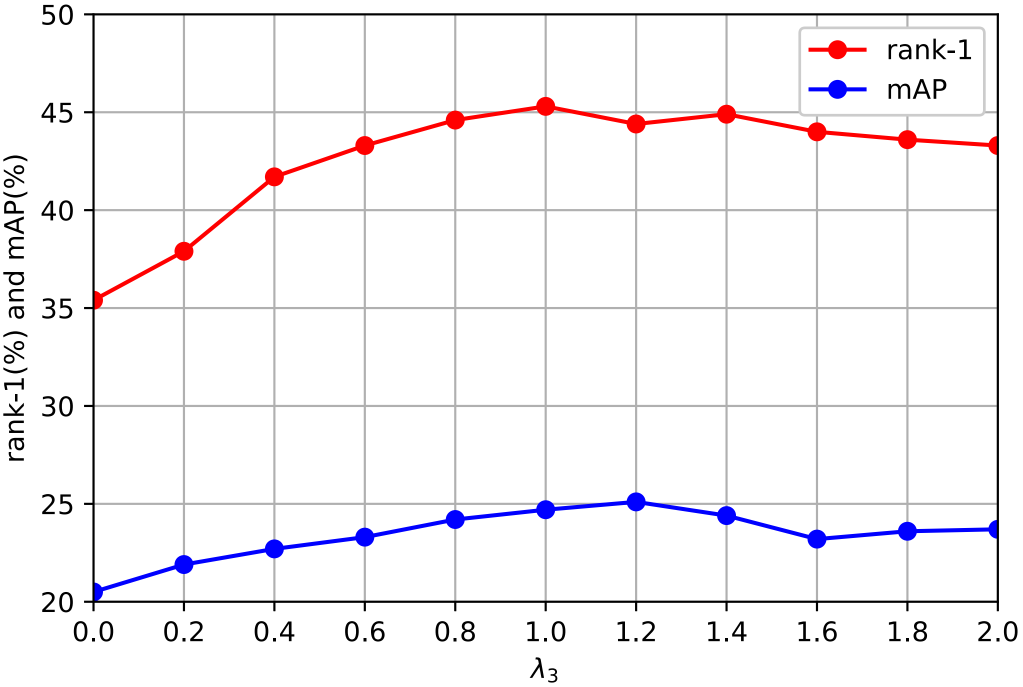


Figure 5.11: The Market1501 trained model performance (Rank 1 accuracy and mAP) on the DukeMTMC dataset with different parameter $\lambda_3$ values ($\lambda_2$ is fixed to 1)

one single end-to-end training session. Our performance can be further improved by implementing any pre- and post-processing techniques such as part-based local max-pooling (LMP), attention mechanisms or re-ranking. For fair comparisons, the performance results shown the Table 5.1 and Table 5.2 are all based on the basic models without any pre or post-processing.

### 5.4.4 Component Analysis and Evaluation

We have also analysed each component of our MMFA model based on their contributions to the cross-domain feature learning. The first set of experiments is the unsupervised performance based on the feature representation learned from the source domain attributes or identities, without any domain adaptation. In the top section of Table 5.3, the attribute annotations alone cannot give a good representation of a person due to its low dimensionality, only 6.4% and 19.2% Rank1 accuracy

| Dataset | VIPeR | PRID | Market1501 | | DukeMCMT-reID | |
|---|---|---|---|---|---|---|
| Metric (%) | Rank 1 | Rank 1 | Rank 1 | mAP | Rank 1 | mAP |
| SDALF [19] | 19.9 | 16.3 | - | - | - | - |
| CPS [12] | 22.0 | - | - | - | - | - |
| DLLR [40] | 29.6 | 21.1 | - | - | - | - |
| GL [41] | 33.5 | 25.0 | - | - | - | - |
| ISR [66] | 27.0 | 17.0 | 40.3 | 14.3 | - | - |
| GTS [118] | 25.2 | - | - | - | - | - |
| SDC [146] | 25.8 | - | - | - | - | - |
| AML [137] | 23.1 | - | 44.7 | 18.4 | - | - |
| UsNCA [91] | 24.3 | - | 45.2 | 18.9 | - | - |
| CAMEL [139] | 30.9 | - | 54.5 | 26.3 | - | - |
| PUL [17] | - | - | 44.7 | 20.1 | 30.4 | 16.4 |
| AdaRSVM [73] | 10.9 | 4.9 | - | - | - | - |
| UDML [89] | 31.5 | 24.2 | - | - | - | - |
| SSDAL [106] | 37.9 | 20.1 | 39.4 | 19.6 | - | - |
| TJ-AIDL$^{Duke}$ [120] | 35.1 | 34.8 | **58.2** | 26.5 | - | - |
| SPGAN$^{Duke}$ [16] | - | - | 51.1 | 22.8 | - | - |
| TJ-AIDL$^{Market}$ [120] | 38.5 | 26.8 | - | - | 44.3 | 23.0 |
| SPGAN$^{Market}$ [16] | - | - | - | - | 41.1 | 22.3 |
| **MMFA$^{\mathbf{Duke}}$** | 36.3 | 34.5 | 56.7 | **27.4** | - | - |
| **MMFA$^{\mathbf{Market}}$** | **39.1** | **35.1** | - | - | **45.3** | **24.7** |

Table 5.1: Performance comparisons with state-of-the-art unsupervised person Re-ID methods.The best and second best results are highlighted by bold and underline receptively. The superscripts: *Duke* and *Market* indicate the source dataset which the model is trained on.

achieved. The features from identity labels, on the other hand, yield much better performance compared to attributes. When attribute and identity information are jointly trained as a multi-objective learning task, the feature representations show a better generalisation-ability. This experiment shows that the attribute annotations do provide extra information to the system which serves as additional supervision for learning more generalised cross-dataset features.

The lower section of Table 5.3 shows the unsupervised re-id performances after aligning the mid-level feature distribution. After aligning the source and target distributions of attributes features, mid-level features or both, we can see a large

| Source → Target | Market1501 → DukeMTMC-reID | | | | DukeMTMC-reID → Market1501 | | | |
|---|---|---|---|---|---|---|---|---|
| Metric (%) | Rank1 | Rank5 | Rank10 | mAP | Rank1 | Rank5 | Rank10 | mAP |
| SPGAN | 41.1 | 56.6 | 63.0 | 22.3 | 51.5 | 70.1 | 76.8 | 22.8 |
| TJ-AIDL | 44.3 | 59.6 | 65.0 | 23.0 | **58.2** | 74.8 | 81.1 | 26.5 |
| **MMFA** | **45.3** | **59.8** | **66.3** | **24.7** | 56.7 | **75.0** | **81.8** | **27.4** |

Table 5.2: Detail Comparison with SPGAN and TJ-AIDL

| Source → Target | Market1501 → DukeMTMC-reID | | | | DukeMTMC-reID → Market1501 | | | |
|---|---|---|---|---|---|---|---|---|
| Metric (%) | Rank1 | Rank5 | Rank10 | mAP | Rank1 | Rank5 | Rank10 | mAP |
| Attribute Only | 6.4 | 14.4 | 18.6 | 2.3 | 19.2 | 34.8 | 45.1 | 6.2 |
| ID Only | 37.6 | 54.9 | 61.6 | 20.6 | 48.2 | 66.1 | 73.3 | 21.6 |
| Attribute+ID Only | 41.7 | 57.5 | 63.6 | 23.3 | 52.2 | 69.1 | 75.7 | 23.5 |
| Attribute with Attribute Feature Adaptation | 15.8 | 26.0 | 48.2 | 5.7 | 35.5 | 55.3 | 64.0 | 12.7 |
| ID with Mid-level Deep Feature Adaptation | 42.1 | 57.7 | 63.9 | 24.3 | 53.4 | 70.2 | 76.4 | 25.2 |
| Mid-level Deep Feature + Attribute Adaptation | **45.3** | **59.8** | **66.3** | **24.7** | **56.7** | **75.0** | **81.8** | **27.4** |

Table 5.3: Adaptation performance on each model components

performance increase when compared with previously non-adapted features. It shows that the proposed mid-level feature distribution alignment strategy is a feasible approach for the unsupervised Person Re-ID task.

## 5.5 Conclusion

This chapter has presented a novel unsupervised cross-dataset feature learning and domain adaptation framework MMFA for Person Re-ID. We utilised the multi-supervision (identity and attributes) classifications to learn a discriminative feature for Person Re-ID on the labelled source dataset. With a shared mid-level feature space assumption, we proposed the mid-level feature alignment domain adaptation strategy to reduce the MMD distance based on the source domain's and the target domain's mid-level feature distributions. In contrast to most existing learn-then-adapt unsupervised cross-dataset approaches, our MMFA is a one-step learn-and-adapt method which can simultaneously learn the feature representation and adapt to the target domain in a single end-to-end training procedure. Meanwhile, our proposed method is still able to outperform a wide range of state-of-the-art unsupervised Re-ID methods. Our MMFA framework improves the scalability of the Person Re-ID models in real-world deployment. However, it needs a vast number of unlabelled images obtained from the new system. It also requires some additional adaptive training to create a bespoke model for the new system. In the next chapter, we aim to develop a robust feature learner which just needed to be trained once and can be deployed to any camera networks without further data collection or adaptive training required. We proposed a domain generalisation model which can leverage the labelled images from multiple datasets to learn a universal representation of people's appearances.

# Chapter 6

# Multi-Datasets Feature Generalisation

In the previous chapter, we have proposed a novel domain adaptation approach to transfer the Person Re-ID model from a labelled dataset to any unlabelled datasets. It alleviates the performance degradation problem of the Person Re-ID model in the cross-dataset scenario. However, in real-world deployment, there are two prominent drawbacks of this approach, which slow down the speed of the Re-ID model.

1. When deploying the model in a new CCTV system, we need to extract a massive amount of unlabelled pedestrian images from the cameras for training our proposed MMFA network [64].

2. We need to train a bespoke model for every new CCTV system. The training of the new models may take from hours to days completely depending on the system scale. Besides, if there is any new camera integrated into the system, the model must be retrained.

In recent years, many large scale Person Re-ID datasets have been collected. In this chapter, we aim to learn an universal domain invariant feature representation by leveraging the labelled data from multiple available datasets. A domain generalisable Person Re-ID model has great value for real-world large- scale deployment. Specifically, when a company or an agency purchases a Person Re-ID system for a specific camera network, the system is expected to work out-of-the-box without the need to go through the tedious process of data collection, annotation and model fine-tuning.

## 6.1 Problem Definition

Conventional supervised single-dataset Person Re-ID models often over-fit to the training dataset, hence they usually suffered from considerable performance degradation on the 'unseen' new cameras or 'untrained' new systems. In the previous chapter, we alleviated the problem by using a cross-dataset domain adaption (DA) model MMFA [64]. However, this approaches requires some unlabelled images from the target domain and introduces additional adaptation process, as shown in Figure 6.1(a). In this chapter, we reformat the Person Re-ID problem as a domain generalisation (DG) problem. Unlike our domain adaptation approach, the domain generalisation model aim to develop a domain generalisation model which can leverage the labelled images from multiple datasets to learn a domain-invariant feature representation, as shown in Figure 6.1(b). Different datasets are often collected in very different visual scenes (e.g., indoors/outdoors, shopping malls, traffic junctions and airports). Each dataset can be considered as a different system representing different domains. Domain generalisation applying on the feature learned from these datasets could help learn a representation which can be relatively well generalised to any unseen surveillance system. This setting simulates the real-world scenario in which a strong feature learner is trained once and deployed to multiple camera networks without further data collection or adaptive training required.

However, there is a very minimal prior study on the domain generalisation for the Person Re-ID task. Some existing Person Re-ID works occasionally evaluate their models cross-dataset generalisation, but no specific design is attempted to make the models more generalisable cross-datasets. Recently, unsupervised domain adaptation (UDA) methods [16, 64, 120] such as our MMFA approach (mentioned in Chapter

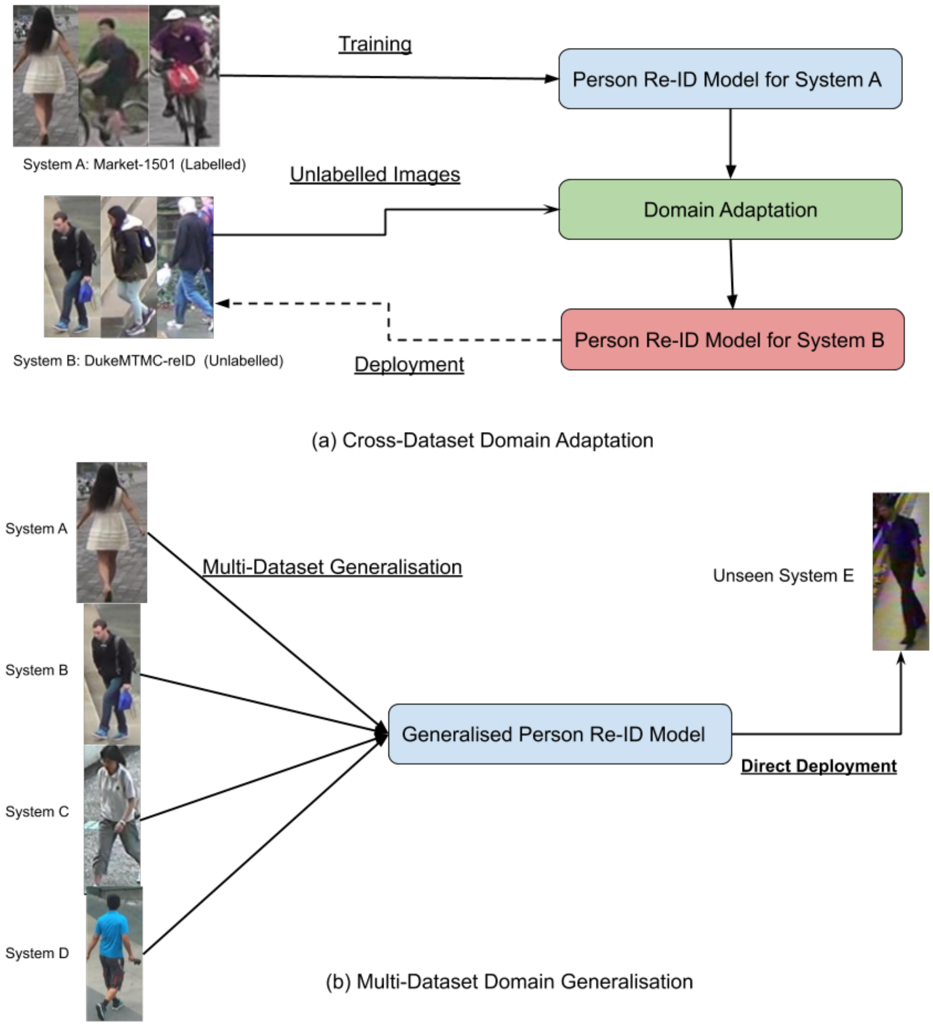

Training
Person Re-ID Model for System A
System A: Market-1501 (Labelled)
Unlabelled Images
Domain Adaptation
System B: DukeMTMC-reID (Unlabelled)
Deployment
Person Re-ID Model for System B
(a) Cross-Dataset Domain Adaptation
System A
Multi-Dataset Generalisation
Unseen System E
System B
Generalised Person Re-ID Model
Direct Deployment
System C
System D
(b) Multi-Dataset Domain Generalisation

Figure 6.1: Difference between cross-dataset domain adaptation and multi-dataset domain generalisation

5) have been studied to adapt a Person Re-ID model from the source to the target domain. However, UDA models update using unlabeled target domain data, so data collection and model update are still required. Beyond Person Re-ID, the problem of domain generalisation (DG) has been investigated in deep learning. Previous works on domain generalisation focused on developing data-driven approaches to learn invariant features among different source domains [49, 81, 82, 131, 133]. However, these methods assume a fixed number of classes for target domains and are trained specifically for that number using source data. They thus have limited efficacy

for Person Re-ID, where the target domain has a different and variable number of identities.

In this chapter, we propose a novel framework for domain generalisation, which aims to learn an universal representation across domains not only by minimising the difference between the multiple seen source domains but also by aligning the distribution of mid-level features between them. In a high level, our proposed framework can be considered as an extension of our proposed MMFA network [64], in the multiple domain learning setting. We develop an algorithm to simultaneously minimise loss of data reconstruction, identification and verification loss and domain difference via adversarial training. In the meanwhile, we also match the distribution of the mid-level features across multiple datasets.

## 6.2 The Proposed Methodology

A basic assumption behind domain generalisation is that there exists a feature space underlying the seen multiple source domains and the unseen target domain, on which a prediction model learned with the training data from the seen source domains can generalise well on the unseen target domain. In order to find this feature space, we extend our previous work MMFA with recently proposed adversarial auto-encoder (AAE) [77] to the multi-domain setting. We call it MMFA with **A**dversarial **A**uto-**E**ncoder (MMFA-AAE). Our proposed method aims to learn a feature space underlying all the seen source domains by minimising the mid-level feature distribution variance among them based on the MMD distance [31]. In this section, we describe how our proposed MMFA-AAE network is designed for domain generation.

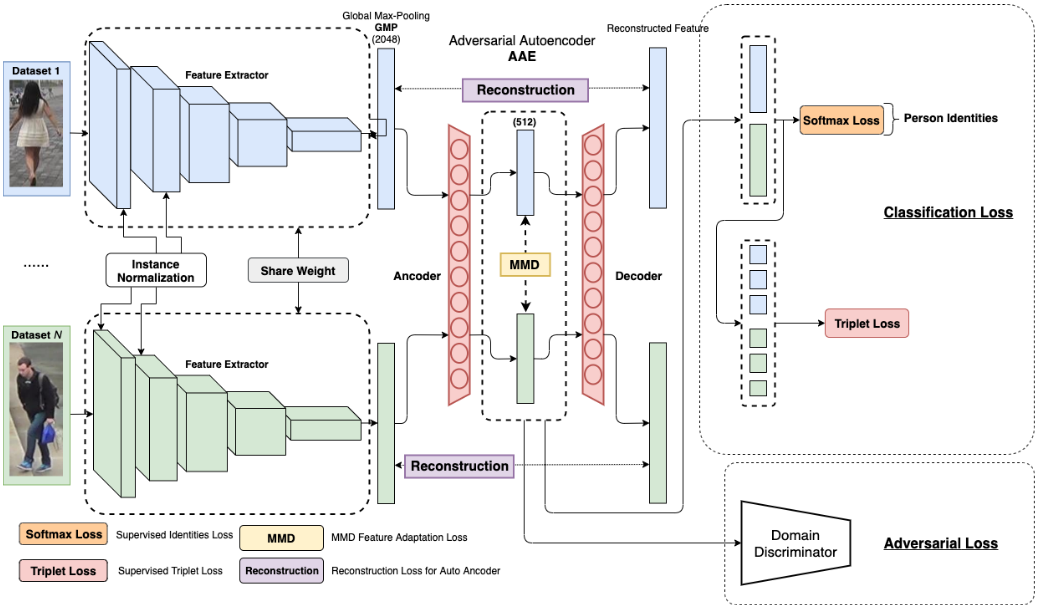


Figure 6.2: An overview of our proposed framework (MMFA-AAE) for Person Re-ID multi-domain generalisation.

### 6.2.1 Architecture

The architecture of the proposed MMFA-AAE network is shown in Figure 6.2. In the MMFA-AAE model, images from multiple domains will be the inputs for the same ResNet50 backbone networks [34] with shared weights. The global max-pooling layer will select the maximum value from every feature map and form a 2048 feature vector. The feature vector will then pass into an adversarial auto-encoder. The adversarial auto-encoder [77] is a probabilistic auto-encoder. It aims to perform variational inference by matching the aggregated posterior of the hidden codes with an arbitrary prior distribution using an adversarial training procedure. The objective of the adversarial auto-encoder in our network is to produce a clean latent space among multiple domains (multiple datasets). The reconstructed feature vectors will then be used for Person Re-ID. In order to further generalise the feature representation across multiple domains, we used MMD [31] regularisation to align the distribution of the mid-level deep features between different domains. In the following section, we will describe how our proposed MMFA-AAE network generalise the feature representations from multiple domains.

### 6.2.2 Instant Normalisation

In the recent studies of the generative adversarial networks (GANs), especially in the style transformation area [84, 86], some image style information could be encoded in the mean and variance of the convolutional feature maps inside the network [84]. Hence, the instance normalisation (IN) [115] which performs the normalisation on a single image across all channels could potentially eliminate the appearance divergence caused by style variances [86]. Hence, Pan et al. [86] proposed the IBN module, which helps to enhance the generalisation capacities of the network for various computer vision tasks. Jia et al. [37] applied this technique to the Person Re-ID problem and yield an impressive Person Re-ID performance boost in the multi-dataset domain generalisation setting. Hence, our MMFA-AAE network follows the same setting in [37] and apply the IN in the first 6 blocks in MobileNetV2 and the fist 4 blocks in ResNet50.

### 6.2.3 Reconstruction Loss

In the domain adversarial auto-encoder of our MMFA-AAE network, we have a feature extractor $Q(x)$ to map the feature embeddings to hidden codes and a decoder $P(h)$ to recover inputs from the hidden codes. The pair of encoder and decoder are shared across all the domains. Let $X = [x_1, ..., x_n]$ be the extracted feature vectors (feature embeddings) from the backbone network. The hidden codes will be $\mathbf{H} = Q(\mathbf{X})$ and the reconstructed feature embedding will be $\hat{\mathbf{X}} = P(\mathbf{H})$, the reconstruction loss of the auto-encoder is defined as follows.

$$\mathcal{L}_{\text{rec}} = \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_2^2 \tag{6.1}$$

### 6.2.4 Adversarial Loss

The hidden nodes can create a common latent feature space for multiple domains. Although, the Instance Normalisation help remove the domain style information. The extracted feature vectors may still contain other kinds of domain-specific knowledge. Hence, it may still exist a risk that certain hidden codes could be over-fitted to the training datasets. Therefore, we impose a domain discriminator $D$. $D$ can classify which dataset the feature vector is drawn from. Suppose, we have $K$ different Person Re-ID datasets in total ($K$ domains). Let $X = [x_1, ..., x_n]$ be the extracted feature vectors with batch size $n$. $Y^d = \left[y_1^d, ..., y_n^d\right], Y^d \in \{1, 2, ..., K\}$ denotes the domain labels of $X$. Thus, the domain discriminator $D$ can be optimised by a standard cross-entropy loss.

$$\mathcal{L}_{\mathrm{D}}(D, Q) = \sum_{l=l}^{n} log(D(Q(x_i), y_i^d)) \tag{6.2}$$

where $D(\cdot)$ denotes the predicted probability that the feature $x_i$ belongs to the domain $y_i^d$. After training a strong domain discriminator, it can capture the hidden domain information which can help the model determine the source domain of the feature vector. We can then eliminate the domain information from the feature vector via adversarial learning using the domain discriminator we trained on. The overall adversarial learning process is a mini-max optimisation problem:

$$arg \min_{Q} \max_{D} \mathcal{L}_{\mathrm{D}}(D, Q) \tag{6.3}$$

$Q$ needs to be minimised for learning a proper person identity mapping of the feature vector. $D$, on the other hand, needs to be maximised to help the network suppress the domain-related features. To simply the training process, we convert the mini-max optimisation problem to a full minimisation optimisation by utilising the gradient reversal layer [22]:

$$\mathcal{L}_{\mathrm{adv}} = -\mathcal{L}_{\mathrm{D}}(D, Q) \tag{6.4}$$

### 6.2.5 MMD-based Regularisation

To further enhance the domain invariant of the hidden code, we follow our previous MMFA architecture to use the Maximum Mean Discrepancy (MMD) [31] regularization to align the distributions among different training datasets. Given a feature embeddings from two domains $\mathbf{H}_l = [\mathbf{h}_{l,1}, \mathbf{h}_{l,2}, ..., \mathbf{h}_{l_{n_l}}]$ and $\mathbf{H}_t = [\mathbf{h}_{t,1}, \mathbf{h}_{t,2}, ..., \mathbf{h}_{t,n_t}]$ with a batch size $n$ and unknown probability distributions. $\phi(\cdot)$ is a mean map operation which projects the distributions into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [30]. Let $n_l$ and $n_t$ are the batch sizes of $\mathbf{H}_l$ and $\mathbf{H}_t$ feature embeddings. The MMD distance between domains $l$ and $t$ can be measured by the following equation.

$$MMD(\mathbf{H}_l, \mathbf{H}_t)^2 = \left\| \frac{1}{n_l} \sum_{i=1}^{n_l} \phi(\mathbf{h}_{l,i}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(\mathbf{h}_{t,j}) \right\|_{\mathcal{H}}^2 \tag{6.5}$$

The arbitrary distribution of the hidden codes of different domains can be represented by using the kernel embedding technique [100]. If the kernel $k(\cdot, \cdot)$ is characteristic, the mapping to the RKHS $\mathcal{H}$ is injective [103]. The injectivity indicates that the arbitrary probability distribution is uniquely represented by an element in RKHS. Therefore, we have a kernel function $k(\mathbf{h}_{l,i}, \mathbf{h}_{t,j}) = \phi(\mathbf{h}_{l,i})\phi(\mathbf{h}_{t,j})^\intercal$ induced by $\phi(\cdot)$.

$$\begin{aligned} MMD(\mathbf{H}_l, \mathbf{H}_t)^2 = & \frac{1}{(n_l)^2} \sum_{i=1}^{n_l} \sum_{i'=1}^{n_l} k(\mathbf{h}_{l,i}, \mathbf{h}_{l,i'}) \\ & + \frac{1}{(n_t)^2} \sum_{j=1}^{n_t} \sum_{j'=1}^{n_t} k(\mathbf{h}_{t,j}, \mathbf{h}_{t,j'}) \\ & - \frac{2}{n_l \cdot n_t} \sum_{i=1}^{n_l} \sum_{j=1}^{n_t} k(\mathbf{h}_{l,i}, \mathbf{h}_{t,j}) \end{aligned} \tag{6.6}$$

We follow the same setting with our previous domain adaptation MMFA model, which uses the RBF characteristic kernel with bandwidth $\alpha = 1, 5, 10$ to computing the MMD distance.

$$k(\mathbf{h}_{l,i}, \mathbf{h}_{t,j}) = exp(-\frac{1}{2\alpha} \left\| \mathbf{h}_{l,i} - \mathbf{h}_{t,j} \right\|^2) \tag{6.7}$$

Since the MMFA-AAE network focuses on the feature generalisation on multiple domains. The overall MMD regularisation term $L_{hidden}$ on the hidden codes is expressed as follows.

$$\mathcal{L}_{\text{mmd}}\left(\mathbf{H}_1, \ldots, \mathbf{H}_K\right) = \frac{1}{K^2} \sum_{1 \leq i,j \leq K} \text{MMD}\left(\mathbf{H}_i, \mathbf{H}_j\right) \tag{6.8}$$

### 6.2.6 Training Procedure

The learning procedure of MMFA-AAE is similar to train an AAE network [77]. Unlike AAE, which only aims to minimise the reconstruction loss, our MMFA-AAE aims to jointly minimise the identification loss, triplet loss, reconstruction loss as well as the MMD regularisation on hidden codes. In our MMFA-AAE, the MMD-based adversarial auto-encoder with the early layer instance normalisation enhances the feature generalisation among different dataset domains. However, in order to learn a robust feature representation for the Person Re-ID task, the network also needs to incorporate the person identity loss and triplet loss. Our MMFA-AAE network uses the same network structure as our baseline method proposed in the earlier section. We use the same equation to compute the cross-entropy identity loss $\mathcal{L}_{\text{id}}$ and the triplet verification loss $\mathcal{L}_{\text{tri}}$. Unlike our baseline method, the MMFA-AAE model introduces three additional loss functions. The reconstruction loss $\mathcal{L}_{\text{rec}}$ is used to preserve the content information of the feature vectors while performing latent space projection during the dimension reduction. The MMD regularisation $\mathcal{L}_{\text{MMD}}$ help align the distribution between different domains. The final feature training loss will be a weight summation of all these losses. The adversarial loss $\mathcal{L}_{\text{adv}}$ is computed from a strong domain discriminator. By maximising the domain classification loss, it helps to guide network focus less on the domain-specific feature.

Similar to training other adversarial learning models, the training procedures for the MMFA-AAE model can be divided into two training phrases:

1. Frozen the feature extractor, use the feature vectors extracted from the network

to train the domain discriminator $D$ by minimising the $\mathcal{L}_{\mathrm{D}}$. The domain discriminator $D$ aims to predict which dataset a feature map is extracted from accurately.

2. Frozen the domain discriminator, training the feature extractor using the identity loss $\mathcal{L}_{\mathrm{id}}$ and triplet loss $\mathcal{L}_{\mathrm{tri}}$ to accurately predict the identity labels and minimise the triplet distance. Meanwhile, update parameters of the network by minimising the reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ and MMD distance $\mathcal{L}_{\mathrm{MMD}}$ between different domain features and adversarial loss $\mathcal{L}_{\mathrm{adc}}$. The overall loss function can be expressed as:

$$\mathcal{L}_{\mathrm{final}} = \mathcal{L}_{\mathrm{id}} + \lambda_1 \mathcal{L}_{\mathrm{triplet}} + \lambda_2 \mathcal{L}_{\mathrm{rec}} + \lambda_3 \mathcal{L}_{\mathrm{mmd}} \tag{6.9}$$

Let $C_{id}$ and $C_{triplet}$ denotes the parameters for ID classifier and triplet classifier. The overall algorithm of MMFA-AAE is illustrated in Algorithm 1.

**Algorithm 1** Training MMFDA-AAE Network

```
Input: Multiple Dataset Domains $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$
Output: Learned parameters $Q^*$, $P^*$ ,$C^*_{id}$ and $D^*$.
for $t = 1$ to max iteration do
    Sample a domain $\mathcal{D}_l \in \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$
    Sample a mini-batch $\mathbf{X}_d$ with the corresponding $\mathbf{ID}_d$ from $\mathbf{X}$,$\mathbf{ID}$. Where
    $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_K\}$, $\mathbf{ID} = \{\mathbf{ID}_1, \ldots, \mathbf{ID}_K\}$ and $\{\mathbf{X}, \mathbf{ID}, \} \in \mathcal{D}_l$
    Sample $\mathbf{h_d}$ from the Laplace distribution.
    Compute the gradient of Eq.(6.9) with respect to $D$ on $\mathbf{X_d}$ and $\mathbf{h_d}$.
    Use the gradient to update $D$ for maximising the objective of Eq.(6.9).
    Compute the gradient of Eq.(6.9) with respect to $Q$, $P$, $C_{id}$ on $\mathbf{X_d}$.
    Use the gradient to update $Q$, $P$, $C_{id}$ for minimising the objective of Eq.(6.9).
end for
```

## 6.3 Experiments

### 6.3.1 Datasets and Settings

#### Person Re-ID Datasets

To evaluate our method, we follow the new experiment settings in DIMN method [102], which was also adopted by the current state-of-the-art method DualNrom [37]. In the settings, multiple large-scale Person Re-ID benchmark datasets are combined to train a model. The small-scale datasets are individually used to evaluate the domain generalization ability of our MMFA-AAE model. The evaluation process simulates a real-world scenario where a Person Re-ID model is trained on multiple public datasets and deploy on an unseen camera system. In our experiments, we select the CUHK02 [52], CUHK03 [54], Market-1501 [147], DukeMTMC-reID [152] and CUHK-SYSU [129]. All these datasets have more than one thousand identities and thousands of images. As the combined training dataset, we use all the images in these datasets to train our model, regardless of their original training/testing splits. The Person Re-ID models are trained with $121,765$ images from $18,530$ identities. The statistics of the training dataset are shown in Table 6.1.

| Dataset | Total IDs | Total Images |
|---|---|---|
| CUHK02 [52] | 1,816 | 7.264 |
| CUHK03 [54] | 1,467 | 14,097 |
| Market-1501 [147] | 1,501 | 29,419 |
| DukeMTMC-reID [152] | 1,812 | 36,411 |
| CUHK-SYSU [129] | 11,934 | 34,574 |
| Total | 18,530 | 121,765 |

Table 6.1: The statistics of the training datasets

The evaluation of our model domain generalisation performance follows is following the same setting in [37, 102] which are tested on the VIPeR dataset [28],

| Dataset | #Test IDs | | # Test Images | |
|---|---|---|---|---|
| | Probe | Gallery | Probe | Gallery |
| VIPeR [28] | 316 | 316 | 316 | 316 |
| PRID [35] | 100 | 649 | 100 | 649 |
| GRID [71] | 125 | 900 | 125 | 1025 |
| i-LIDS [150] | 60 | 60 | 60 | 60 |
| MSMT17 [122] | 3,060 | 3,060 | 9,716 | 82,161 |

Table 6.2: The statistics of testing datasets

the PRID dataset [35], the GRID dataset [71] and the i-LIDS dataset [150]. However, these datasets are relatively small and have no more than one thousand identities. To illustrate the more realistic real-world Person Re-ID performance, we also conducted the test on currently the largest Person Re-ID dataset MSMT17 [122]. The overall statistics of the testing datasets are shown in Table 6.2.

**Evaluation Protocols**

We follow the proposed evaluation protocols for VIPeR [28],PRID [35] GRID [71] and i-LIDS [150]. For the VIPeR dataset, we randomly half-split the dataset into training and testing sets. We only use the testing set for evaluation. The overall performance on VIPeR is the average results from 10 randomly split testing set. For the PRID dataset evaluation, we follow the same single-shot experiments as [141]. Similar to the VIPeR dataset setting, the final performance is the average of the experimental results based on 10 random split testing. Since the VIPeR and PRID datasets contain only two images per person, the mean average precision (mAP) metric cannot be used here. On GRID, we follow the standard testing split recommended in [71]. On i-LIDS, two images per identity are randomly selected as the probe image and the gallery image, respectively. For all the testing datasets, the average results over 10 random splits are reported. For MSMT17 dataset, the dataset has already been split into training, query and gallery set. We follow the single-query retrieval setting for the MSMT17 dataset evaluation.

The cumulative matching characteristics (CMC) curve is used for our performance evaluation, as it is the most common metric used for evaluating person Re-ID performance. This metric is adopted since Re-ID is intuitively posed as a ranking problem, where each image in the gallery is ranked based on its comparison to the probe. The probability that the correct match in the ranking equal to or less than a particular value is plotted against the size of the gallery set [28]. To make the comparison concise, we simplified the CMC curve to only comparing Rank 1, Rank 5, Rank 10 successful retrieval rates. The CMC curve evaluation is valid when only one ground truth match for each given query image. The MSMT17 dataset contains multiple ground truth images for the same person. Therefore, we use the mean average precision (mAP) proposed by [147] as an additional new evaluation metric. For each query image, the average precision (AP) is calculated as the area under its precision-recall curve. The mean value of the average precision (mAP) will reflect the overall recall of the person Re-ID algorithm.

### Implementation Details

For the auto-encoder sub-network, we follow the same setting as that reported in [23], which uses a single hidden layer with a size of 512 neurons. The value of the hidden layer is used as an input for both the adversarial sub-network and the classification sub-network. The adversarial sub-network and the classification sub-network are composed of two fully-connected (FC) layers. One FC layer is set to the same size as the hidden layer; another is set to the same size as the ID labels. The weights for ID loss and triplet loss are set as equal, *i.e*, $\lambda_1 = 1$. Through various testing, we observed that the parameters $\lambda_2 = 1, \lambda_3 = 0.002, \lambda_4 = 0.1$ yield the best performance. The Adam optimiser [39] is used for all our experiments. The initial learning rate is set to 0.00035 with the warm-up training technique [26] and is decreased by 10% at the 40th epoch and 70th epoch, respectively. Totally, there are 120 training epochs with the batch size of 64. We implement our model in PyTorch and train it on a single Titan X GPU. The extracted features are L2 normalised before matching scores are calculated.

### 6.3.2 Comparison against state-of-the-art methods

To demonstrate the superiority of our method, we compare with various state-of-the-art methods under three different experimental conditions: fully supervised, unsupervised domain adaptation and domain generalisation. In Table 6.3, the *DG* methods are the multi-dataset domain generalisation approaches. The AGG methods in the *DG* category are the domain aggregation baselines trained without any domain generalisation layer or sub-network. *S* denotes a fully supervised method trained using images and labels from the corresponding target dataset. The *DA* method means a cross-dataset Person Re-ID approach by utilising unsupervised domain adaptation techniques. It is important to note that the *DA* and *S* methods are not fair competitors in the sense that they use more information about the target domain than ours. We include them not as direct competitors, but to contextualise our results.

#### Comparison with Domain Generalisation Methods

Domain generalisation is the most practical requirement for the Person Re-ID problem. It assumes that a target dataset cannot be seen during training. Because of this challenge, domain generalisation Person Re-ID methods have to learn general feature representation from other datasets. However, there is a little prior study on the domain generalisation for the Person Re-ID task. Only two methods have been proposed [37, 102]. For a fair comparison with these methods, we followed the same evaluation protocol and experiment setting. The lower part of Table 6.3 shows the benchmark results of the methods. Our AGG baseline is slightly higher due to the additional triplet loss during the supervised training. MMFA-AAE network can give a 10% to 30% increase in the Rank 1 retrieval accuracy for all four datasets. Our

| Method | Type | VIPeR | | | PRID | | | GRID | | | i-LIDS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-5 | R-10 | R-1 | R-5 | R-10 | R-1 | R-5 | R-10 | R-1 | R-5 | R-10 |
| Ensemble [85] | S | 45.9 | 77.5 | 88.9 | 17.9 | 40.0 | 50.0 | - | - | - | 50.3 | 72.0 | 82.5 |
| DNS [141] | S | 42.3 | 71.5 | 82.9 | 29.8 | 52.9 | 66.0 | - | - | - | - | - | - |
| ImpTrpLoss [11] | S | 47.8 | 74.4 | 84.8 | 22.0 | - | 47.0 | - | - | - | 60.4 | 82.7 | 90.7 |
| GOG [79] | S | 49.7 | 79.7 | 88.7 | - | - | - | 24.7 | 47.0 | 58.4 | - | - | - |
| MTDnet [10] | S | 47.5 | 73.1 | 82.6 | 32.0 | 51.0 | 62.0 | - | - | - | 58.4 | 80.4 | 87.3 |
| OneShot [4] | S | 34.3 | - | - | 41.4 | - | - | - | - | - | 51.2 | - | - |
| SpindleNet [143] | S | 53.8 | 74.1 | 83.2 | 67.0 | 89.0 | 89.0 | - | - | - | 66.3 | 86.6 | 91.8 |
| SSM [2] | S | 53.7 | - | 91.5 | - | - | - | 27.2 | - | 61.2 | - | - | - |
| JLML [55] | S | 50.2 | 74.2 | 84.3 | - | - | - | 37.5 | 61.4 | 69.4 | - | - | - |
| MMFA(Market-1501) [64] | DA | 39.1 | - | - | 35.1 | - | - | - | - | - | - | - | - |
| MMFA(DukeMTMC-reID) [64] | DA | 36.3 | - | - | 34.5 | - | - | - | - | - | - | - | - |
| TJ-AIDL(Market-1501) [120] | DA | 38.5 | - | - | 26.8 | - | - | - | - | - | - | - | - |
| TJ-AIDL(DukeMTMC-reID) [120] | DA | 35.1 | - | - | 34.8 | - | - | - | - | - | - | - | - |
| SyRI [5] | DA | 43.0 | - | - | 43.0 | - | - | - | - | - | 56.5 | - | - |
| AGG(DIMN) | DG | 42.9 | 61.3 | 68.9 | 38.9 | 63.5 | 75.0 | 29.7 | 51.1 | 60.2 | 69.2 | 84.2 | 88.8 |
| AGG(DualNorm) | DG | 42.1 | - | - | 27.2 | - | - | 28.6 | - | - | 66.3 | - | - |
| AGG(MMFA-AAE) | DG | 48.1 | - | - | 27.7 | - | - | 32.6 | - | - | 67.3 | - | - |
| DIMN [102] | DG | 51.2 | 70.2 | 76.0 | 39.2 | 67.0 | 76.7 | 29.3 | 53.3 | 65.8 | 70.2 | 89.7 | 94.5 |
| DualNorm [37] | DG | 53.9 | - | - | **60.4** | - | - | 41.4 | - | - | 74.8 | - | - |
| MMFA-AAE | DG | **58.4** | - | - | 57.2 | - | - | **47.4** | - | - | **84.8** | - | - |

Table 6.3: Comparison results against state-of-the-art methods. (R: Rank, S: Supervised training with a target dataset, DA: Domain Adaptation, DG: Domain Generalisation, -: No report)

MMFA-AAE method outperforms the DIMN and DualNorm on VIPeR, GRID and i-LIDS by a large margin. MMFA-AAE only fall 3% behind DualNorm in Rank 1 accuracy when testing the PRID dataset but still near 20% higher than the DIMN method.

To further demonstrate our proposed MMFA-AAE's superiority to other methods in the real-world application, we also conduct the experiment on the largest Person Re-ID benchmark at the moment: MSMT17. Table 6.4 provides a performance comparison of our domain aggregation baseline, the current stat-of-the-art DualNorm method and our MMFA-AAE network. All three methods use the ResNet50 backbone to allow a fair comparison. The domain aggregation baseline without any domain generalisation technique can only achieve 14.8% Rank 1 accuracy and 5.9% mAP score. Both DualNorm and our MMFA-AAE can boost the baseline performance by a large margin in both Rank 1 and mAP scores. Our MMFA-AAE consistently surpass the DualNorm by 3 to 4% in Rank 1 accuracy. Overall, our MMFA-AAE can achieve a much better performance out-of-box without any additional data collection and domain adaptation process.

| Model | MSMT17 | | | |
|---|---|---|---|---|
| | Rank 1 | Rank 5 | Rank 10 | mAP |
| ResNet50 Baseline | 14.8 | 27.8 | 37.6 | 5.9 |
| DualNorm (ResNet50) | 42.6 | 55.9 | 61.8 | 19.6 |
| MMFA-AAE (ResNet50) | **45.4** | **59.5** | **64.2** | **20.7** |

Table 6.4: Comparison results of domain aggregation baseline, ResNet50 DualNorm and our MMFA-AAE with ResNet50 backbone on the MSMT17 dataset

### Comparison with Domain Adaptation Methods

We also compare our MMFA-AAE with other unsupervised domain adaptation methods. The multi-dataset domain generalisation approaches focus on learning the universal feature representation from multiple different Person Re-ID datasets and assume the model can learn well-generalised features for any unseen camera network. On the other hand, the domain adaptation approaches focus on analysing the images characterises between the images from label ed public datasets and images obtained from the unseen cameras. Although, the training and experimentation setting is different for DA and DG Person Re-ID models. Our MMFA-AAE model without using any target domain image can still surpass the latest unsupervised domain adaptation approaches such as TJ-AIDL [120], MMFA [64] and SyRI [5]. The performance results are shown in the middle section of Table 6.3. SyRI performs the best among them by utilising a synthetic dataset. The MMFA-AAE outperforms all of them on all the benchmark datasets without using any the images from the target dataset and does not introducing additional adaptation process. This means that our method can competitively use the feature learned from multiple large-scale datasets.

#### Comparison with Supervised Methods

Although many fully supervised Person Re-ID methods are reported to have high performance on the large-scale datasets such as Market-1501 and DukeMTMC-reID, their performance is still low when trained on a small-scale dataset. Many methods have been proposed to address this issue [2, 4, 10, 11, 55, 79, 85, 141, 143]. We have selected several supervised methods with reports on any of the four benchmark datasets (labeled as $S$ in Table 6.3): Ensemble [85], DNS [141], ImpTriplet[11], GOG [79], MTDnet [10], OneShot [4], SpindleNet [143], SSM [2], and JLML [55]. These methods follow conventional single-dataset training and testing procedures. It is not a fair comparison with MMFA-AAE method, which operates under the more challenging cross-dataset generalisation setting. However, we use their results as references to illustrate the generalisation capability of our MMFA-AAE model. Our MMFA-AAE method shows competitive or even better results on all four benchmarks.

Overall, our proposed MMFA-AAE network demonstrates a state-of-the-art Person Re-ID performance not only in the multi-dataset domain generalisation experiments but also in the domain adaptation and supervise settings. It proves that our proposed MMFA-AAE network can effectively reduce the domain specific features via using the adversarial training method and learn a more general feature representation.

### 6.3.3 Ablation study

#### components Analysis

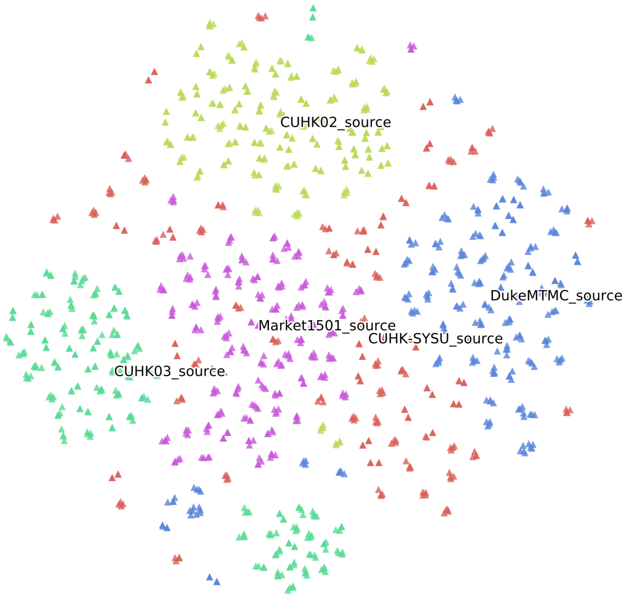
There are four important components in the MMFA-AAE framework: Instance Normalisation, Triplet Loss, Adversarial Auto-Encoder(AAE), and Maximum Mean Discrepancy (MMD). To evaluate the contribution of each component, we incrementally adding one component into our baseline method and compare the performance in Table 6.5. The baseline we use in the experiment is based on a ResNet50 feature extractor with batch normalisation after global average pooling. The baseline is trained with softmax identity loss only first. We then introduce the instance normalisation in the lower convolutional layer the same as DualNorm. The triplet loss will further enhance the performance by 1% to 2% on VIPeR, GRID and i-LIDS. The domain-based adversarial auto-encoder give a large 3% to 8% boost for all the datasets. The final MMD alignment helps the further boosts the overall performance by 1% to 2%.

| Method | VIPeR | PRID | GRID | i-LIDS |
|---|---|---|---|---|
| | R-1 | R-1 | R-1 | R-1 |
| Baseline (ResNet50) | 42.9 | 38.9 | 29.7 | 69.2 |
| Baseline + IN (DualNorm) | 54.4 | 68.6 | 43.7 | 72.2 |
| Baseline + IN + Triplet | 55.9 | 61.6 | 43.0 | 74.8 |
| Baseline + IN + Triplet + AAE | 57 | 67.6 | 46.3 | 82.3 |
| Baseline + IN + Triplet + AAE + MMD (MMFA-AAE) | **58.4** | **65.7** | **47.4** | **84.8** |

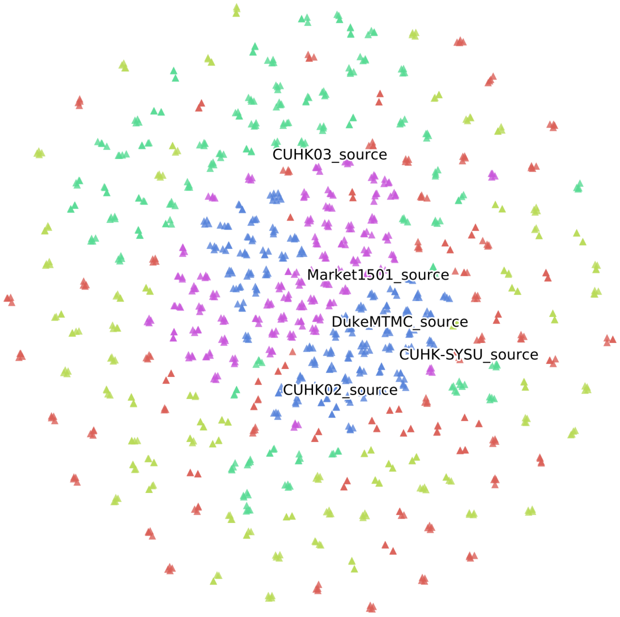
Table 6.5: Ablation study on the impact of different components for MMFA-AAE networks

### t-SNE Visualisation

For completeness, we also visualise the 2D point cloud of the feature vectors extracted from the DualNorm network and our MMFA-AAE method using t-SNE [27], as shown in Figure 6.3. We used a random sample of 6000 images from all five training datasets and perplexity of 5000 for this visualisation. As shown in Figure 6.3 (a), the DualNorm network can merge 5 different datasets well with low domain gaps between different datasets. However, the datasets are still clustered into several groups based on the property of the extracted feature vectors. On the other hand, our MMFA-AAE introduced the additional Adversarial-Auto-encoder(AAE) to mix up the feature vector distribution of different domains and alleviated the domain information for the Person Re-ID task. Figure 6.3 (b) depicts our feature-point clouds extracted from the MMFA-AAE network. We can easily see that the overlap between different feature domains is more prominent in the case of MMFA-AAE network.



(a) (a) DualNorm



(b) MMFA-AAE

Figure 6.3: The t-SNE visualisations of the feature vectors from the DualNorm network and our MMFA-AAE network. Different colour points indicate the training dataset domains.

## 6.4 Conclusion

In this chapter, we propose a novel framework for multi-dataset feature generalisation network MMFA-AAE. Our MMFA-AAE network was proposed to enable a Person Re-ID model to be deployed out-of-the-box for any new camera network. The main objective of our MMFA architecture is to learn a domain invariant feature representation by jointly optimising an adversarial auto-encoder with an MMD distance regularisation. The adversarial auto-encoder is designed to learn a latent feature space among different Person Re-ID datasets by matching the distribution of the hidden codes to an arbitrary prior distribution. The MMD-based regularisation further enhances the domain invariant feature by aligning the distributions among different domains. In this way, the learned feature embedding is supposed to be universal to the seen training datasets and is expected to generalise well on the other unseen datasets because of the introduction of the prior distribution. Extensive experiments demonstrate that our proposed MMFA-AAE is able to learn domain-invariant features, which lead to state-of-the-art performance on many Person Re-ID dataset, which is never seen by the network. The experiments also showed that domain generalisation in Person Re-ID is an extremely challenging problem. Many existing domain generalisation and meta-learning methods failed to beat the strong but naive domain aggregation baseline. In conclusion, our MMFA-AAE approach addresses the scalability issue of many existing Person Re-ID methods by providing the most practical multi-dataset feature generalisation strategy. Given our promising result, our MMFA-AAE approach provides a good starting point for discussion and further research.

# Chapter 7

# Conclusion and Future Works

## 7.1 Conclusion

This thesis focuses on developing camera-invariant feature learning frameworks for person re-identification (Person Re-ID). In Chapter 3 and 4, we proposed two different Person Re-ID applications: online person matching and offline person retrieval. During the deployment of these applications, we have encountered the scalability issue when integrating the existing single-dataset supervised methods into a real-world surveillance system. The model trained on one dataset (one CCTV system) usually sufferers from considerable performance degradation when directly used for a new 'unseen' camera network. Therefore, we have proposed a cross-dataset domain adaptation method (Chapter 5) and a multi-dataset domain generalisation approach (Chapter 6) to strengthening the generalisation capability of the existing feature extraction networks.

The first framework proposed in Chapter 3 is tailored for the online person matching application. In our proposed method, we use the feature maps obtained from the mid-layer of the CNN architecture as an alternative to the actual mid-level semantic attributes. We developed a Siamese structure neural network which is designed to learn the discriminative deep mid-level features of a person and construct the correspondence features between an image pair in a data-driven manner. Unlike other Siamese structures, our proposed network performs the regional feature map matching not only in a pairwise fashion between two images, but also considers many different combinations of multiple feature maps. As an end-to-end network, the model will directly output the similarity score for each image pairs. By integrating the feature extraction and metric learning into one network, the processing time can be reduced, making it suitable for our online person matching applications.

The second framework proposed in Chapter 4 is designed for offline person retrieval applications. In this work, we proposed a novel negative competing triplet loss (NC-Triplet), which helps to discriminate the negative sample pairs further and significantly boost the overall mAP score of many existing models. In addition, we collected a new privacy-aware Person Re-ID dataset called: Re-ID-Outdoor. It not only follows the recent implementation of General Data Protection Regulation (GDPR) in Europe but also address the limitations of existing Person Re-ID datasets, such as small camera number and unrealistic survillance environment. We conducted extensive experiments to demonstrate the state-of-the-art performance of our model in Market-1501, DukeMTMC-reID and the our Re-ID-Outdoor detest.

Although our Person Re-ID method can achieve impressive performance in several benchmark datasets, the features extracted from our model show poor performance on the new camera system. Due to the limited size of the training dataset, the single-dataset supervised models usually over-fit to the specific camera settings and show a strong dataset bias. Hence, these models will suffer significant performance degradation in a new camera system. To improve the existing Person Re-ID model's scalability to different camera networks, we proposed a cross-dataset Person Re-ID model (MMFA). The MMFA network utilises the multi-supervision (identity and attributes) classification to learn a discriminative feature for Person Re-ID on the labelled source dataset. With a shared mid-level feature space assumption, we proposed the mid-level feature alignment domain adaptation strategy to reduce the MMD distance based on the source domains and the target domain's mid-level feature distributions. In contrast to most existing learn-then-adapt unsupervised cross-dataset approaches, our MMFA network is a one-step learn-and-adapt method, which can simultaneously learn the feature representation and adapt to the target domain in a single end-to-end training procedure.

Although, our MMFA framework improves the scalability of the Person Re-ID models in real-world deployment. However, it needs a vast number of unlabelled images obtained from the new system. It also requires some additional adaptive training to create a bespoke model for the new system. In Chapter 6, we aim to develop a robust feature learner that needs to be trained only once and can be deployed out-of-the-box for any new camera network without further data collection or adaptive training. With this motivation, we proposed a domain generalisation model (MMFA-AAE) that can leverage the labelled images from multiple datasets to learn a universal representation of people's appearances. Our MMFA-AAE architecture learns a domain invariant feature representation by jointly optimising an adversarial auto-encoder with the MMD distance regularisation. The adversarial auto-encoder is designed to learn a latent feature space among different Person Re-ID datasets by matching the distribution of the hidden codes to an arbitrary prior distribution. The MMD-based regularisation further enhances the domain invariant features by aligning the distributions among different domains. Extensive experiments demonstrate that our proposed MMFA-AAE is able to learn domain-invariant features, which lead to state-of-the-art performance on many Person Re-ID datasets.

## 7.2 Future Work

### 7.2.1 Rose-Identification-Corridor Dataset

Due to recent privacy and data protection movement over the world, many Person Re-ID datasets have been removed from the internet. There is a huge demand for the new privacy-aware Person Re-ID datasets for Person Re-ID research. After the completion of the outdoor Person Re-ID dataset collection, we are currently working on collecting another Person Re-ID dataset Rose-IDentification-Corridor Dataset

(Re-ID-Corridor) for the indoor environment. There are not many Person Re-ID datasets for the indoor scene. By using 150 indoor surveillance cameras mounted along the corridors in the School of Electrical and Electronic Engineering (EEE) Buildings of Nanyang Technological University (NTU), we hope to contribute a new large-scale dataset specialised for the indoor environment. The camera locations of one floor in the EEE building is shown as Figure 7.1.

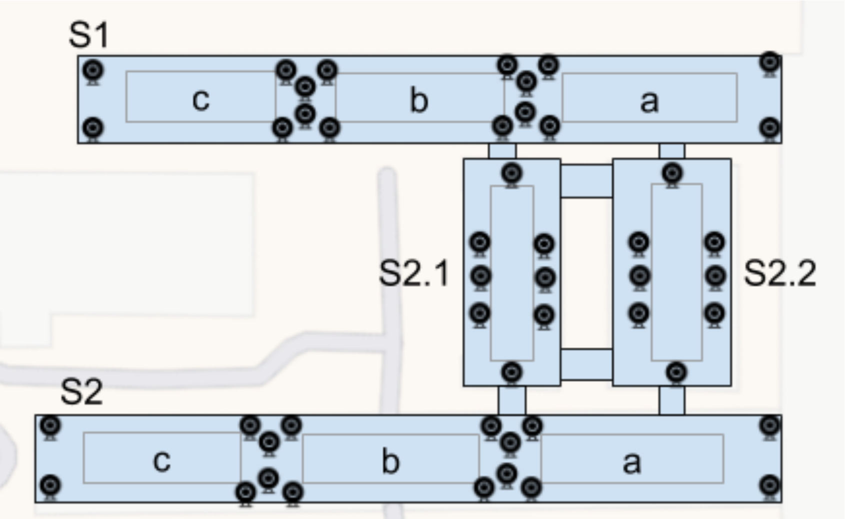
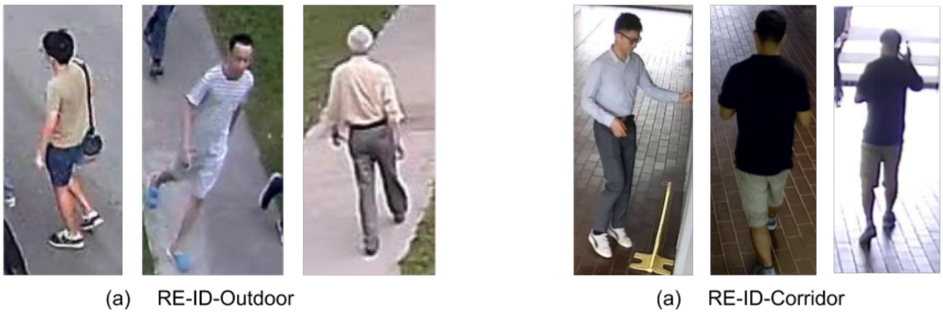


Figure 7.1: One floor of EEE builds with the locations of all surveillance cameras

Unlike the Re-ID-Outdoor date, the indoor cameras used for the new Re-ID-Corridor have higher resolution (1080p). The positions of the cameras on the corridor are much closer the pedestrians. Hence, images captured in the new dataset have better image quality and contain more visual information compared to the outdoor dataset. The comparison between the quality of the images in Re-ID-Outdoor and Re-ID-Corridor is illustrated in Figure 7.2. The higher resolution and better image quality could extend our dataset to other application such as face recognition and gait recognition.

### 7.2.2 Camera-level Model Boosting

Our future works are not only just collecting the new dataset, but also focusing on improving our existing models and developing new evaluation protocol. Real-world



(a) RE-ID-Outdoor

(a) RE-ID-Corridor

Figure 7.2: Sample images from the Re-ID-Outdoor dataset and the new Re-ID-Corridor dataset

video surveillance systems usually consist of hundreds of cameras. Each camera can be considered as an independent module for detecting and extracting the pedestrian images. A competent Person Re-ID system should have good and consistent person re-identification performance across all cameras. Due to the small camera network size of the existing datasets, the conventional evaluation protocol only analyses the overall CMC and mAP metrics of the entire Person Re-ID system. For a massive camera network, it is insufficient to provide a complete picture of system performance without analysing the individual performance of every camera. Some cameras of a surveillance system need to be installed in poor lighting areas or have very different colour profiles with most other cameras. Hence, when we use images from these cameras for person re-identification or perform the people search on these cameras, the retrieval success rate would be much lower than the overall system scores. Hence, in our future work, we would like to propose two additional camera-based evaluation protocols: Camera-Query Evaluation and Camera-Gallery Evaluation. Camera-Query Evaluation uses the images obtained from one specific camera as the query images and tests the Person Re-ID performance on other cameras. Camera-Gallery Evaluation uses images from other cameras to search the person from the image gallery obtained from the specified cameras. These two evaluation protocols will give us a fine-grained performance analysis to the camera level and help us pinpoint the bottleneck of the Person Re-ID system. In order to boost up the model performance on those challenging cameras in the system, we would like to extend our domain adaptation and domain generalisation methods from the system level to the camera level.

# Bibliography

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An Improved Deep Learning Architecture for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Song Bai, Xiang Bai, and Qi Tian. Scalable Person Re-identification on Supervised Smoothed Manifold. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.358.

[3] Sawomir Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In *International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010.

[4] Sawomir Bąk and Peter Carr. One-shot metric learning for person re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Sawomir Bąk, Peter Carr, and Jean Franois Lalonde. Domain Adaptation Through Synthesis for Unsupervised Person Re-identification. In *European Conference on Computer Vision (ECCV)*, 2018.

[6] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-Driven Accumulation of Local Features for Human Characterization and Re-identification. *Computer Vision and Image Understanding*, 2013.

[7] Apurva Bedagkar-Gala and Shishir K. Shah. A Survey of Approaches and Trends in Person Re-identification. *Image and Vision Computing*, 2014.

[8] Lubomir Bourdev and Jitendra Malik. Poselets: Body Part Detectors Trained Using 3D

Human Pose Annotations. In *International Conference on Computer Vision (ICCV)*, 2009.

[9] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature Verification using a Siamese Time Delay Neural Network,. In *Advances in Neural Information Processing Systems (NIPS)*, 1994.

[10] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A Multi-task Deep Network for Person Re-identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[11] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6 2016.

[12] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conference (BMVC)*, 2011.

[13] Yeong-Jun Cho and Kuk-Jin Yoon. Improving Person Re-identification via Pose-Aware Multi-shot Matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6 2016.

[14] Zuozhuo Dai, Mingqiang Chen, Siyu Zhu, and Ping Tan. Batch Feature Erasing for Person Re-identification and Beyond. In *arXiv preprint*, 2018.

[15] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-Theoretic Metric Learning. In *International Conference on Machine Learning (ICML)*, 2007.

[16] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised Person Re-identification: Clustering and Fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018.

[18] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. SphereReID: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation (JVCIR)*, 2019.

[19] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person Re-identification by Symmetry-driven Accumulation of Local Features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[20] P F Felzenszwalb, R B Girshick, D McAllester, and D Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 9 2010.

[21] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research (JMLR)*, 2016. ISSN 21916594. doi: 10.1007/978-3-319-58347-1{_}10.

[23] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain Generalization for Object Recognition with Multi-task Autoencoders, 2015.

[24] Ross Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, 2015.

[25] Ross Girshick, Jeff Donahue, Trevor Darrell, U C Berkeley, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[26] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. In *arXiv preprint*, 2017.

[27] Duncan Graham-Rowe. Visualizing Data Using t-SNE. *Journal of Machine Learning Research (JMLR)*, 2008. ISSN 02624079.

[28] Doug Gray, Shane Brennan, and Hai Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007.

[29] Douglas Gray and Hai Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *European Conference on Computer Vision (ECCV)*, 2008.

[30] Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A Kernel Method for the Two-Sample Problem. *Journal of Machine Learning Research (JMLR)*, 2008.

[31] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K. Sriperumbudur. A Fast, Consistent Kernel Two-Sample Test. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[32] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[33] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C. Berg. MatchNet : Unifying Feature and Metric Learning for Patch-Based Matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person Reidentification by Descriptive and Discriminative Classification. In *Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[36] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. EANet: Enhancing Alignment for Cross-Domain Person Re-identification. In *arXiv preprint*, 2018.

[37] Jieru Jia, Qiuqi Ruan, and Timothy M. Hospedales. Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice. In *British Machine Vision Conference (BMVC)*, 2019.

[38] Minyue Jiang, Yuan Yuan, and Qi Wang. Self-attention Learning for Person Re-identification. In *British Machine Vision Conference (BMVC)*, 2018.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[40] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification. In *British Machine Vision Conference (BMVC)*, 2015.

[41] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person Re-identification by Unsupervised L1 Graph Learning. In *European Conference on Computer Vision (ECCV)*, 2016.

[42] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large Scale Metric Learning from Equivalence Constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[44] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[45] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Domain Transfer for Person Re-identification. In *International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream (ARTEMIS)*, 2013.

[46] Yann LeCun, Patrick Haffner, Lon Bottou, and Yoshua Bengio. Object Recognition with Gradient-Based Learning. In *Shape, Contour and Grouping in Computer Vision*. Springer, 1998.

[47] Roberto Leyva, Victor Sanchez, and Chang-tsun Li. Video Anomaly Detection With Compact Feature Sets for Online Performance. *IEEE Transactions on Image Processing (TIP)*, 7 2017.

[48] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li. Compact and Low-complexity Binary Feature Descriptor and Fisher Vectors for Video Analytics. *IEEE Transactions on Image Processing (TIP)*, 2019.

[49] Da Li, Yongxin Yang, Yi Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *International Conference on Computer Vision (ICCV)*, 2017.

[50] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain Generalization with Adversarial Feature Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[51] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised Person Re-identification by Deep Learning Tracklet Association. In *European Conference on Computer Vision (ECCV)*, 2018.

[52] Wei Li and Xiaogang Wang. Locally Aligned Feature Transforms across Views. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[53] Wei Li, Rui Zhao, and Xiaogang Wang. Human Reidentification with Transferred Metric Learning. In *Asian Conference on Computer Vision (ACCV)*, 2012.

[54] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[55] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[56] Xi Li, Weiming Hu, Chunhua Shen, Zhongfei Zhang, Anthony Dick, and Anton van den Hengel. A Survey of Appearance Models in Visual Object Tracking. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2013.

[57] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith. Learning Locally-Adaptive Decision Functions for Person Verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[58] Shengcai Liao and Stan Z. Li. Efficient PSD Constrained Asymmetric Metric Learning for Person Re-identification. In *International Conference on Computer Vision (ICCV)*, 2015.

[59] Shengcai Liao, Zhipeng Mo, Jianqing Zhu, Yang Hu, and Stan Z. Li. Open-set Person Re-identification. *arXiv preprint*, 2014.

[60] Shengcai Liao, Yang Hu, Xiangyu Zhu, Stan Z. Li, Xiangyu Zhu, and Stan Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[61] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. In *arXiv preprint*, 2013.

[62] Shan Lin and Chang-Tsun Li. End-to-End Correspondence and Relationship Learning of Mid-Level Deep Features for Person Re-Identification. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017.

[63] Shan Lin and Chang-Tsun Li. Person Re-identification with Soft Biometrics through Deep Learning. In *Deep Biometrics*, pages 21–36. Springer, Cham, 2020. ISBN 978-3-030-32582-4.

[64] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex C Kot. Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification. In *British Machine Vision Conference (BMVC)*, 2018.

[65] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving Person Re-identification by Attribute and Identity Learning. In *arXiv preprint*, 2017.

[66] Giuseppe Lisanti, Iacopo Masi, Andrew D. Bagdanov, and Alberto Del Bimbo. Person Re-Identification by IterativeRe-Weighted Sparse Ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[67] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person Re-identification: What Features Are Important? In *European Conference on Computer Vision Workshops (ECCVW)*, 2012.

[68] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. In *International Conference for Machine Learning (ICML)*, 2015.

[69] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep Transfer Learning with Joint Adaptation Networks. In *International Conference for Machine Learning (ICML)*, 2017.

[70] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

[71] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera Activity Correlation Analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[72] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and A Strong Baseline for Deep Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[73] Andy J Ma, Jiawei Li, Pong C Yuen, and Ping Li. Cross-Domain Person Reidentification Using Domain Adaptation Ranking SVMs. *IEEE Transactions on Image Processing (TIP)*, 5 2015.

[74] Bingpeng Ma, Yu Su, and Frdric Jurie. Local Descriptors encoded by Fisher Vectors for Person Re-identification. In *European Conference on Computer Vision Workshops (ECCVW)*, 2012.

[75] Bingpeng Ma, Yu Su, Frdric Jurie, Bingpeng Ma, Yu Su, Frdric Jurie, and Bingpeng Ma. BiCov: a novel image representation for person re-identification and face verification. In *British Machine Vision Conference (BMVC)*, 2012.

[76] Christopher Madden, Eric Dahai Cheng, and Massimo Piccardi. Tracking People across Disjoint Camera Views by an Illumination-Tolerant Appearance Representation. *Machine Vision and Applications*, 5 2007.

[77] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial Autoencoders. In *International Conference on Learning Representations Workshop (ICLRW)*, 2015.

[78] Chaojie Mao, Yingming Li, Zhongfei Zhang, Yaqing Zhang, and Xi Li. Pyramid Person Matching Network for Person. In *Asian Conference on Machine Learning (ACML)*, 2017.

[79] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical Gaussian Descriptor for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[80] Alexis Mignon and Frederic Jurie. PCCA: A New Approach for Distance Learning from Sparse Pairwise Constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2666–2672. IEEE, 6 2012.

[81] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, Gianfranco Doretto, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. Few-Shot Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[82] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation. In *International Conference for Machine Learning (ICML)*, 2013.

[83] V Nair, GE Hinton Proceedings of the 27th international Conference, and Undefined 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference for Machine Learning (ICML)*, 2010.

[84] Hyeonseob Nam and Hyo Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[85] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to Rank in Person Re-identification with Metric Ensembles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[86] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at Once: Enhancing

Learning and Generalization Capacities via IBN-Net. In *European Conference on Computer Vision (ECCV)*, 2018.

[87] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[88] S. Pellegrini, A. Ess, K. Schindler, and L van Gool. Youll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking. In *International Conference on Computer Vision (ICCV)*. IEEE, 9 2009.

[89] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised Cross-Dataset Transfer Learning for Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[90] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person Re-Identification by Support Vector Ranking. In *British Machine Vision Conference (BMVC)*, 2010.

[91] Chen Qin, Shiji Song, Gao Huang, and Lei Zhu. Unsupervised Neighborhood Component Analysis for Clustering. *Neurocomputing*, 168:609–617, 2015.

[92] Filip Radenovi. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In *arXiv preprint*, 2016.

[93] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. In *arXiv preprint*, 2018.

[94] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[95] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set forMulti-Target, Multi-Camera Tracking. In *European Conference on Computer Vision Workshops (ECCVW)*, 2016.

[96] Florian Schroff and James Philbin. FaceNet : A Unified Embedding for Face Recognition and Clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[97] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person Re-Identification with Correspondence Structure Learning. In *International Conference on Computer Vision (ICCV)*, 12 2015.

[98] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, and Zhen Lei. Embedding Deep Metric for Person Re-identification: A Study Against Large Variations. In *European Conference on Computer Vision (ECCV)*, 2016.

[99] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[100] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert Space Embedding for Distributions. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[101] Achmad Solichin, Agus Harjoko, and Agfianto Eko. A Survey of Pedestrian Detection in Video. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2014.

[102] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable Person Re-identification by Domain-Invariant Mapping Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[103] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R. G. Lanckriet, and Bernhard Schölkopf. Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[104] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 2014.

[105] Chi Su, Fan Yang, Shiliang Zhang, and Qi Tian. Multi-Task Learning with Low Rank Attribute Embedding for Person Re-identification. In *International Conference on Computer Vision (ICCV)*, 2015.

[106] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep Attributes Driven Multi-camera Person Re-identification. In *European Conference on Computer Vision (ECCV)*, 2016.

[107] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-Driven Deep Convolutional Model for Person Re-identification. In *International Conference on Computer Vision (ICCV)*, pages 3980–3989, 2017.

[108] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. SVDNet for Pedestrian Retrieval. In *International Conference on Computer Vision (ICCV)*, 2017.

[109] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling. *arXiv preprint*, 2017.

[110] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Chapel Hill, and Ann Arbor. Going Deeper with Convolutions. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[111] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[112] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data Augmentation using Random Image Cropping and Patching for Deep CNNs. *Proceedings of Machine Learning Research (PMLR)*, 2018.

[113] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. AANet: Attribute Attention Network for Person Re-Identifications. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[114] Ojala Timo, Pietikinen Matti, and Menp Topi. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002.

[115] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. In *arXiv preprint*, 2016.

[116] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6 2016.

[117] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACM International Conference on Multimedia (ACM MM)*, 2018.

[118] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised Learning of Generative Topic Saliency for Person Re-identification. In *British Machine Vision Conference (BMVC)*, 2014.

[119] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong. Towards Unsupervised Open-Set Person Re-identification. In *International Conference on Image Processing (ICIP)*, 2016.

[120] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[121] Xiaogang Wang and Rui Zhao. Person Re-identification: System Design and Evaluation Overview. In *Person Re-Identification*. 2014.

[122] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[123] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by Relative Distance Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 3 2013.

[124] Kilian Q Weinberger and Lawrence K Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *The Journal of Machine Learning Research (JMLR)*, 2009.

[125] Yandong Wen, Kaipeng Zhang, Zhifeng Li B, and Yu Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. In *European Conference on Computer Vision (ECCV)*, 2016.

[126] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-velez. Beyond Sparsity : Tree Regularization of Deep Models for Interpretability. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[127] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online Object Tracking: A Benchmark. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6 2013.

[128] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[129] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint Detection and Identification Feature Learning for Person Search. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[130] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person Re-Identification using Kernel-based Metric Learning Methods. In *European Conference on Computer Vision (ECCV)*, 2014.

[131] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting Low-Rank Structure from Latent Domains for Domain Generalization. In *European Conference on Computer Vision (ECCV)*, 2014.

[132] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention Driven Person Re-identification. *Pattern Recognition*, 2019.

[133] Pei Yang and Wei Gao. Multi-view Discriminant Transfer learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[134] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. Salient Color Names for Person Re-identification. In *European Conference on Computer Vision (ECCV)*, 2014.

[135] Yang Yang, Longyin Wen, Siwei Lyu, and Stan Z Li. Unsupervised Learning of Multi-Level Descriptors for Person Re-Identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[136] Hantao Yao, Shiliang Zhang, Dongming Zhang, Yongdong Zhang, Jintao Li, Yu Wang, Qi Tian, Yao Hantao, Zhang Shiliang, Zhang Dongming, Zhang Yongdong, Li Jinta, Wang Yu, and Tian Qi. Large-scale Person Re-identification as Retrieval. In *International Conference on Multimedia and Expo (ICME)*, 2017.

[137] Jieping Ye, Zheng Zhao, and Huan Liu. Adaptive Distance Metric Learning for Clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[138] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep Metric Learning for Person Re-identification. In *International Conference on Pattern Recognition (ICPR)*, 2014.

[139] Hong-Xing Xing Yu, Ancong Wu, and Wei-Shi Shi Zheng. Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification. In *International Conference on Computer Vision (ICCV)*, 2017.

[140] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*, 2014.

[141] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a Discriminative Null Space for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[142] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. In *arXiv preprint*, 2017.

[143] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[144] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person Re-identification by Salience Matching. In *International Conference on Computer Vision (ICCV)*. IEEE, 12 2013.

[145] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised Salience Learning for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[146] Rui Zhao, Wanli Oyang, Xiaogang Wang, Wanli Ouyang, and Xiaogang Wang. Person Re-Identification by Saliency Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 12 2017.

[147] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *International Conference on Computer Vision (ICCV)*, 2015.

[148] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-identification. In *European Conference on Computer Vision (ECCV)*, 2016.

[149] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person Re-identification: Past, Present and Future. *arXiv preprint*, 2016.

[150] Wei Shi Zheng, Shaogang Gong, and Tao Xiang. Associating Groups of People. In *British Machine Vision Conference (BMVC)*, 2009.

[151] Zhedong Zheng, Liang Zheng, and Yi Yang. A Discriminatively Learned CNN Embedding for Person Re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2016.

[152] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *International Conference on Computer Vision (ICCV)*, 2017.

[153] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. In *arXiv preprint*, 2017.

[154] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing A Person Retrieval Model Hetero- and Homogeneously. In *European Conference on Computer Vision (ECCV)*, 2018.

[155] Xiaoke Zhu, Xiao-yuan Jing, Fei Wu, Yunhong Wang, Wangmeng Zuo, and Wei-shi Zheng. Learning Heterogeneous Dictionary Pair with Feature Projection Matrix for

Pedestrian Video Retrieval via Single Query Image. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.