UTS UNIVERSITY OF TECHNOLOGY SYDNEY

PERSON RE-IDENTIFICATION IN THE WILD: FROM SHORT-TERM TO LONG-TERM

by Yan Huang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Qiang Wu

University of Technology Sydney
Faculty of Engineering and Information Technology

05/2021

Certificate of Original Authorship Template

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Yan Huang* declare that this thesis, is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *School of Electrical and Data Engineering/Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 21/05/2021

ABSTRACT

PERSON RE-IDENTIFICATION IN THE WILD: FROM SHORT-TERM TO LONG-TERM

by

YAN HUANG

The task of person re-identification (re-ID) is to confirm the identity of a person in visual traces captured by different cameras. Person re-ID "in the wild" is a highly demanded technology and quite challenging due to the lack of data diversity, dramatic background variation between different domains, uncontrollable clothing change, and influence caused by the shortage of lighting. According to the scale of time gaps when footages are captured, these challenges are exposed into two different scenarios: 1) Short-Term person re-ID (ST-reID), and 2) Long-Term person re-ID (LT-reID). ST-reID addresses the time gap of several minutes. This scenario is mainly to deal with variations of illumination, viewpoint, and pose. Thus, data diversity is the primary concern. In addition, when footages are taken in different environments, ST-reID often encounters a large background shift issue. On the contrary, the time gaps between two footages in LT-reID can be several hours or even longer. Thus, a person has a great chance to change clothing. As another widely seen case, after a long-time gap, in order to take footages to re-identify a person when s/he reappears at night, infrared cameras are required.

For ST-reID in the wild, the diversity of training data is essential to ensure a re-ID system can tolerate variations of illumination, viewpoint, and pose, *etc.* In addition, a model trained on one domain can lack certain generalization when it is applied to a new domain. This thesis will deeply study the two challenges exposed in ST-reID. Corresponding solutions are provided by using generated data to compensate for the limited data diversity. Also, a background shift suppression

model is proposed to deal with the background shift issue for cross-domain ST-reID.

For LT-reID in the wild, it is worth investigating approaches to tackle the clothing change issue. This thesis will introduce new clothing change datasets to the community. Corresponding solutions are given to tackle the clothing change issue. In addition, under tight security surveillance in LT-ReID, how to recognize the same person who appears under a RGB camera (in the daytime) and an infrared camera (reappear at night) is an immediate problem. A modality-biased training issue is unveiled for the infrared-visible LT-reID task, and corresponding solutions are given. This thesis will provide useful insights into diverse person re-ID issues in the wild from the short-term scenario to the long-term scenario to support practical usages in the real world.

Dissertation directed by Associate Professor Qiang Wu
School of Electrical and Data Engineering

Acknowledgements

Throughout the past few years during my PhD study I have received a great deal of support and assistance from my supervisor, my colleagues, and my family.

First, I would like to thank my principle supervisor A/Prof. Qiang Wu whose expertise was invaluable in formulating the research questions and methodology. His insightful feedback on my researches helped me to sharpen my thinking and brought my work to a higher level. During my PhD study, A/Prof. Qiang devoted lots of efforts to better my researches on computer vision. My first paper was accepted by a top-ranked journal (TIP, IF:9.34) under the patient guidance of A/Prof. Qiang. I learned a lot of skills about academic writing in my first paper from my supervisor. My second paper was accepted by one of the most challenging computer vision conference ICCV2019, and its extension version has been accepted by IJCV which is one of the top journal in computer vision field. I can still clearly remember how A/Prof. Qiang shared his valuable insight to this paper. His scientific and academic rigour impresses me much more than the outcomes. In addition, A/Prof. Qiang provided lots of support in data acquisition for my research. I am also very fortunate to have chance to work with other researchers and PhD students introduced by A/Prof. Qiang, which influenced my a lot to my research career. In the end, I appreciate A/Prof. Qiang's strong support to my scholarship (AGRTP and IRS) applications, making my lives much easier in Sydney.

Second, I would like to thank my co-supervisor A/Prof. Jian Zhang and Dr. Jingsong Xu. Jian organized many seminars that provided me opportunities to communicate with other researchers. Jian also organized the VCIP2019 conference at UTS. As a volunteer of the conference, I broaden my mind and learn something new from industries and researchers all over the world. Jingsong gave me a lot of constructive suggestions to my researches and always encourages me to focus on the

cutting-edge techniques of computer vision.

Third, I would like to thank all my dear co-authors and colleagues during my doctoral study. Thanks to Prof. Zhaoxiang Zhang (from NLPR), Asst/Prof. Yi Zhong (from BIT), Dr. Peng Zhang (from UTS), and Dr. Zhedong Zheng (from UTS) for their strong support to my publications. Thanks to A/Prof. Deyang Liu and Junyi Wu for their collaborative papers. Thanks to A/Prof. Yifan Zuo who introduced me to my supervisor. Thanks to Dr. Zongjian Zhang, Dr. Lu Zhang, Prof. Yazhou Yao, Dr. Xiaoshui Huang, Dr. Zhibin li, Dr. Huaxi Huang, Dr. Muming Zhao, Dr. Xunxiang Yao, Dr. Junjie Zhang, Dr. Lina Li, Dr. Yongshun Gong, Dr. Lingxiang Yao, Dr. Qian Li, Dr. Anan Du, Jialiang Shen and all other colleagues. With them, my lives in Sydney were rich and colourful.

In the end, I would like to thank my parents for their support, trust, encouragement, and love throughout my doctoral studies these years.

Yan Huang

Sydney, Australia, 2021.

List of Publications

Journal Papers

J-1. **Y. Huang**, Q. Wu, J. Xu, Y. Zhong and Z Zhang. "Unsupervised Domain Adaptation with Background Shift Mitigating for Person Re-Identification," *Springer International Journal of Computer Vision (IJCV)*, 2021.

J-2. **Y. Huang**, Q. Wu, J. Xu, Y. Zhong, P. Zhang and Z. Zhang. "Alleviating Modality Bias Training for Infrared-Visible Person Re-Identification," *IEEE Transactions on Multimedia (TMM)*, accepted, 2021.

J-3. **Y. Huang**, J. Xu, Q. Wu, Y. Zhong, P. Zhang and Z. Zhang, "Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 30, no. 10, pp. 3459-3471, 2020.

J-4. **Y. Huang**, J. Xu, Q. Wu, Z. Zheng, Z. Zhang and J. Zhang, "Multi-Pseudo Regularized Label for Generated Data in Person Re-Identification," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 3, pp. 1391-1403, 2019.

J-5. **Y. Huang**, Y. Zhong, Q. Wu, E. Dutkiewicz and T. Jiang, "Cost-effective foliage penetration human detection under severe weather conditions based on auto-encoder/decoder neural network," *IEEE Internet of Things Journal (IoTJ)*, vol. 6, no. 4, pp. 6190-6200, 2019.

J-6. P. Zhang, J. Xu, Q. Wu, **Y. Huang** and X. Ben, "Learning Spatial-temporal Representations over Walking Tracklet for Long-term Person Re-Identification in The Wild," *IEEE Transactions on Multimedia (TMM)*, 2020.

J-7. L. Huang, Q. Yang, J. Wu, **Y. Huang**, Q. Wu and J. Xu, "Generated Data With Sparse Regularized Multi-Pseudo Label for Person Re-Identification,"

IEEE Signal Processing Letters (SPL), vol. 27, pp. 391-395, 2020.

J-8. P. Zhang, J. Xu, Q. Wu, **Y. Huang** and J. Zhang, "Top-Push Constrained Modality-Adaptive Dictionary Learning for Cross-Modality Person Re-Identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2020.

J-9. Y. Zhong, Y. Yang, X. Zhu, **Y. Huang***, E. Dutkiewicz, Z. Zhou and T. Jiang, "Impact of Seasonal Variations on Foliage Penetration Experiment: A WSN-Based Device-Free Sensing Approach," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 56, no. 9, pp. 5035-5045, 2018.

J-10. D. Liu, **Y. Huang**, Q. Wu, R. Ma and P. An. "Multi-Angular Epipolar Geometry based Light Field Angular Reconstruction Network," *IEEE Transactions on Computational Image (TCI)*, 2020.

J-11. Y. Zhong, S. Wu, J. Wang, T. Jiang, **Y. Huang*** and Q. Wu. "Multi-Location Human Activity Recognition via MIMO-OFDM Based Wireless Networks: An IoT-Inspired Device-Free Sensing Approach," *IEEE Internet of Things Journal (IoTJ)*, 2020.

J-12. J. Wu, **Y. Huang***, Q. Wu, Z. Gao, J. Zhao and L. Huang. "Dual-Stream Guided-Learning Via Priori Optimization for Person Re-identification," *Springer Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, accepted, 2020.

Conference Papers

C-1. **Y. Huang**, Q. Wu, J. Xu and Y. Zhong, "SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification," *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 9526-9535, Oct, 2019.

C-2. **Y. Huang**, Q. Wu, J. Xu and Y. Zhong, "Celebrities-ReID: A Benchmark for Clothes Variation in Long-Term Person Re-Identification," *Proc. IEEE International Joint Conference on Neural Network (IJCNN)*, pp. 1-8, July, 2019.

C-3. J. Wu, L. Yao, **Y. Huang**, J. Xu, Q. Wu and L. Huang, "Improving Person Re-Identification Performance Using Body Mask Via Cross-Learning Strategy," *Proc. IEEE Visual Communications and Image Processing (VCIP)*, Dec, 2019.

Submitted Papers

J-1. **Y. Huang**, Q. Wu, J. Xu, Y. Zhong, P. Zhang and Z Zhang. "Learning Directly from Style Invariant Component for Infrared-Visible Person Re-Identification," *IEEE Transactions on Cybernetics*, 2020 (Under review).

J-2. X. Ding, T. Jiang, J. Yang, Y. Zhong, **Y. Huang** and S. Wu. "WiLISensing: Device-free Location-independent Human Activity Recognition via Wi-Fi," *IEEE Internet of Things Journal*, 2020 (Under review).

Contents

List of Figures

List of Tables

Abbreviation

Person re-ID - Person re-IDentification

ST-reID - Short-Term person re-IDentification

LT-reID - Long-Term person re-IDentification

GAN - Generative Adversarial Network

MBT - Modality Bias Training

DLS - Dual-level Learning Strategy

CNN - Convolutional Neural Network

SBSGAN - Suppression of Background Shift Generative Adversarial Network

MpRL - Multi-pseudo Regularized Label

DA-2S - Densely Associated 2-Stream

SN - Scalar Neuron

VN - Vector Neuron

IDE - ID-discriminative Embedding

HOF - Histograms of Optical Flow

MBH - Motion Boundary Histogram

LSRO - Label Smooth Regularization for Outliers

LSR - Label Smoothing Regularization

sMpRL - static Multi-pseudo Regularized Label

dMpRL-I - dynamic Multi-pseudo Regularized Label-I

dMpRL-II - dynamic Multi-pseudo Regularized Label-II

DPM - Deformable Part Model

ISDC - Inter-Stream Densely Connection

DCCV - Dynamic Clustering Confidence Value

GAP - Global Average Pooling

FC - Fully-Connected layer

CE - Cross-Entropy loss

BN - Batch Normalization

mAP - mean Average Precision

SEA - Soft Embedding Attention

FSR - Feature Sparse Representation

ReIDCaps - Re-IDentification Capsule network

P-Caps - Primary Capsules

C-Caps - Classification Capsules

R-by-A - Routing-by-Agreement

SE - Squeeze-and-Excitation

TS1 - Training Strategy 1

TS2 - Training Strategy 2

RS - Robustness Score

dIDeS - dynamic ID-exclusive Smooth

CMC - Cumulated Matching Characteristics

SGD - Stochastic Gradient Descent

WRT - Weighted Regularization Triplet

Chapter 1

Introduction

1.1 Background

With the growing concern of public security, many surveillance systems are deployed in public areas. However, considering the span of these systems, disjoint views commonly exist in public areas. Therefore, person re-identification (re-ID), which captures one specific target under these disjoint camera views, becomes a crucial technique in video surveillance applications and has gained significant attention in recent years. Fig. 1.1 illustrates a typical application of the person re-id system in the wild, *i.e.*, matching pedestrians under different camera views. In Fig. 1.1, two disjoint places are monitored by two cameras with non-overlapped surveillance views. When a person walking from camera A to B, the person re-ID system will bridge the 'blind spots' between these two views.

Ideally, a sophisticated person re-ID system should be able to deal with the re-ID task in the wild. For person re-ID in the wild, several typical issues should be ad-

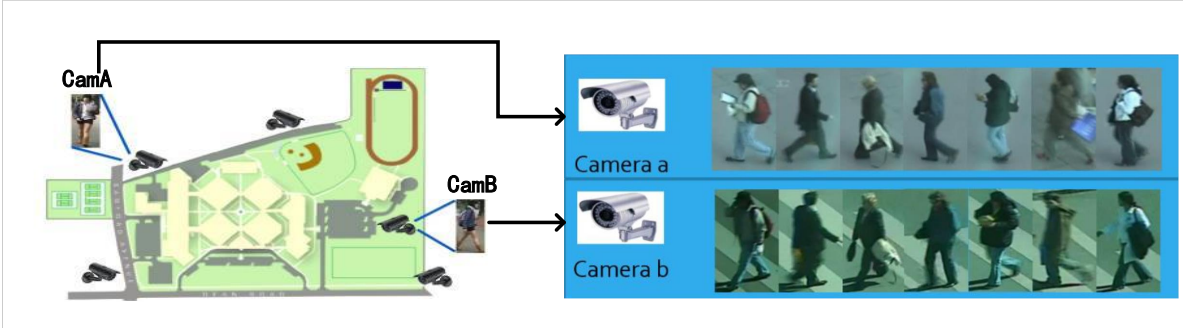


Figure 1.1 : An example of a person re-ID application in the wild. Two disjoint public places are simultaneously monitored by camera A and B, respectively.

dressed. First, to deal with variations of illumination, viewpoint, and pose between footages taken by different cameras, increasing the diversity of data can be regarded as the main challenge. Second, when footage is taken in different environments, the large background shift between different domains can lead to catastrophic performance degradation when no training data can cover the environment presented in testing data. Based on the scale of time gaps when footages are taken, the above mentioned two issues are exposed in Short-Term person re-ID (ST-reID) in the wild. This scenario normally addresses the time gap of several minutes. ST-reID is investigated first by the community, and many state-of-the-art works with encouraging performance have been proposed (Huang et al., 2015, 2019c; Sheng et al., 2016, 2015; Huang et al., 2016; Zheng et al., 2017b; Huang et al., 2017b, 2018a; Zhang et al., 2020c; Wu et al., 2019b; Huang et al., 2020; Zhang et al., 2020b; Huang et al., 2021b).

Unlike ST-reID, some studies have focused on the second scenario in recent years, *i.e.*, Long-Term person re-ID (LT-reID). In this scenario, the time gaps between two footages taken can be several days or even longer. As a typical case can be seen in LT-reID in the wild, the clothing change of each individual is an important issue that should be considered. In addition, in some special wild environments that need tight security surveillance, a common RGB camera may not fit the requirements well. For example, surveillance in a dark environment is a widely seen case. In order to confirm the identity of a person who appears under an RGB camera in the daytime and an infrared camera at night, infrared-visible LT-reID is an immediate problem. For LT-reID in the wild, appearance changes (either caused by clothing change or image modality difference) dramatically constrain the performance of methods designed for the ST-reID scenario which highly rely on the appearance of a person.

This thesis will deeply investigate person re-ID issues according to the two dif-

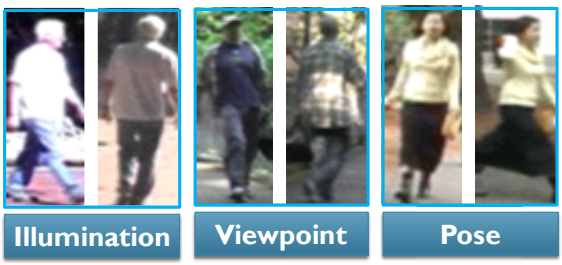


Figure 1.2 : Three persons under different camera views. Large variations such as illumination, viewpoint, and pose can be observed.

ferent scenarios (*i.e.*, ST-reID and LT-reID).

For ST-reID in the wild, two typical challenges are discussed:

- Challenges on variations of illumination, viewpoint, and pose.
- Challenges on footages taken in different environments.

For LT-reID in the wild, another two typical challenges are studied:

- Challenges on clothing change when a person being identified.
- Challenges on footages taken by different styles of cameras.

1.1.1 Challenges on Variations of Illumination, Viewpoint, And Pose in ST-reID

The key issue in single-domain ST-reID is dealing with disturbances caused by variations of illumination, viewpoint, pose, *etc.* Fig. 1.2 illustrates three examples of the same person captured by different camera views. It can be observed that even a human is hard to distinguish the same person in such cases. In order to deal with the variations in single-domain ST-reID, existing approaches can be categorized into three types:

1. **Traditional approach** (extracting robust person descriptors and designing metric learning algorithms to minimize/maximum the intra/inter-class variations using the extracted descriptors) (Liao et al., 2015; Zheng et al., 2015, 2011; Liao and Li, 2015; Yu et al., 2017a).

2. **Deep learning approach** (designing neural network architecture to learn ID-discriminative features for each person) (Qian et al., 2017; Lin et al., 2017; Geng et al., 2016; Zheng et al., 2018, 2016a,b).

3. **Boosting trick** (data augmentation or post-processing trick to improve the performance of off-the-shelf methods) (Huang et al., 2016; Bai et al., 2017; Zhong et al., 2017; Garcia et al., 2015; Ali et al., 2010; Wang et al., 2016b; Liu et al., 2013).

The traditional approaches are manually designed to handle the single-domain ST-reID task. In contrast, deep learning approaches discover more implicit information in matching persons and can significantly outperform the performance of traditional approaches. After that, the performance-boosting trick can be regarded as a promising active research line for the single-domain ST-reID. In order to deal with the variations exposed in single-domain ST-reID, increasing the diversity of data is essential. However, due to the expensive cost of data acquisition that needs to manually find corresponding labels of pedestrians who appear under different camera views, the feasibility of this solution is low in practice. Without sufficient training data, a model is hard to distinguish different identities. Therefore, how to increase/enrich the scale of training data for the single-domain ST-reID is still an open question.

In order to deal with the insufficient training data issue, in 2014s, Generative Adversarial Network (GAN) was proposed to generate data (images) with perceptual quality (Goodfellow et al., 2014). Since then, several improved approaches (Radford

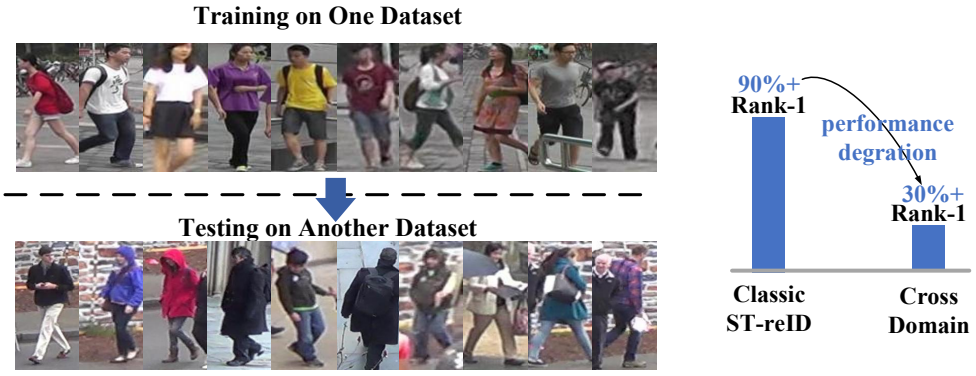


Figure 1.3 : Schematic diagram of cross-domain ST-reID. It can be seen that compared with the single-domain ST-reID issue in which the training and testing datasets are the same, the performance is dropped a lot when testing is conducted on another dataset.

et al., 2015; Arjovsky et al., 2017; Gulrajani et al., 2017) are presented to improve the quality of generated data further. Using generated data to solve the problem of limited training data is a promising solution. In all existing methods by using GAN, there are two main challenging points to assure a better performance: **1)** high-quality data generated by GAN (Radford et al., 2015; Arjovsky et al., 2017; Gulrajani et al., 2017), **2)** a better strategy to use the generated data into the training model (Zhedong et al., 2017). Many works focus on the first point. This thesis will focus on the second point that follows the same pipeline in (Zhedong et al., 2017) to incorporate generated data with real data to train deep models in a semi-supervised learning fashion in order to deal with the challenges for ST-reID in the wild.

1.1.2 Challenges on Footages Taken by Different Environments in ST-reID

The single-domain ST-reID assumes that the training and testing data are captured from the same environment. That is, footages are taken in the same environment (*e.g.*, a campus) by N cameras. Both training and testing images are captured by the N cameras. The only difference is that any pedestrian presented in training data will not appear in testing data. However, this assumption cannot be guaranteed in some applications. For instance, data collected from two different environments (*e.g.*, two different campuses) have distinct backgrounds. In this situation, in addition to non-overlapped identities, backgrounds in testing images captured by N_1 cameras in one environment cannot be seen by models trained with images captured by N_2 cameras in another environment. The change of backgrounds between two environments can jeopardize the re-ID performance. As shown in Fig. 1.3, directly training a classifier from one dataset collected from an environment (*i.e.*, a source domain) often produces a degraded performance when testing is conducted on another dataset collected from a different environment (*i.e.*, a target domain). Therefore, it is important to investigate solutions for such a cross-domain issue. For ST-reID in the wild, the cross-domain ST-reID solutions have drawn attention in recent years. Existing cross-domain ST-reID approaches can be categorized into two types:

1. **General inter-domain style transfer** (Using GAN model to transfer the style of source domain data to target domain) (Bak et al., 2018; Deng et al., 2018; Wei et al., 2018a; Zhong et al., 2018b).

2. **Clustering-based approaches** (Directly estimating labels of target domain images via appearance clustering result) (Fan et al., 2018; Song et al., 2020; Fu et al., 2019; Zhang et al., 2019b).

Note that, as mentioned in Sec. 1.1.1, GAN is used to enrich the scale of training data for single-domain ST-reID by generating new images. When GAN is adopted for the cross-domain issue, it concentrates on image style adaptation between data belonging to different domains. This is achieved by transferring the style of images from a source domain to a target domain to handle the domain shift issue. However, although these cross-domain ST-reID approaches may perform well in certain cases, *i.e.*, domain style changes or camera style changes. These approaches cannot well deal with the catastrophic background shift issue. This is because, for cross-domain ST-reID in the wild, the environment (or background) change is the key reason that incurs inevitable domain shift. For instance, when a network is trained based on limited background information presented in a source domain, such a network may not distinguish essential pedestrian features against dramatic variations caused by background changes in a target domain. Unfortunately, backgrounds in the target domain are normally very different from the source domain. This thesis formulates this problem as a background shift problem that degrades the overall performance of cross-domain person re-ID.

One possible solution to sort out background shift is to remove backgrounds using foreground masks in a hard manner directly (*i.e.*, applying the binary masks on original images) (Farenzena et al., 2010; Huang et al., 2016; Song et al., 2018; Tian et al., 2018). However, it is observed that methods specifically designed for removing background may damage the foreground information. By simply removing backgrounds, this hard manner solution does improve the performance of cross-domain ST-reID in the wild. At the same time, it can be seen that this is still an open problem: Is there a way to suppress background shift better in order to improve cross-domain ST-reID performance? This thesis will make the first effort to generate images while background being suppressed moderately instead of completely removing backgrounds in a hard manner for the cross-domain ST-reID issue

Figure 1.4 : Comparison between samples in ST-reID dataset (*i.e.*, Market-1501 (Zheng et al., 2015)) and clothing change LT-reID dataset. For each dataset, only one ID is shown in the figure.

in the wild.

1.1.3 Challenges on Clothing Change When A Person Being Identified in LT-reID

As a widely seen case, uncontrollable clothing change should be an important research issue that needs to be deeply investigated for LT-reID in the wild. Apparently, significant appearance change can be involved (see Fig. 1.4) when a person wears different clothes, resulting in dramatic performance degradation in the re-ID system. An ideal re-ID system should have the capability to tackle such a challenge in the real world (Gong et al., 2014). In order to deal with the clothing change LT-reID issue, various approaches are reported, although they are still on a preliminary level. Existing approaches for clothing change LT-reID in the wild can be categorized into two types:

1. **Biometrics based approaches** (Exploring clothing-irrelevant biometric features such as gait, body contour/shape, and face) (Zhang et al., 2018b; Yang et al., 2019a; Qian et al., 2020; Li et al., 2020a; Yu et al., 2020; Wan et al.,

2020b).

2. **Depth information compensation approaches** (Providing another source of helpful depth information taken by RGB-D cameras) (Barbosa et al., 2012a; Munaro et al., 2014; Haque et al., 2016).

The biometric information has been used for the clothing change LT-reID issue when the clothing information is no longer reliable. However, it heavily relies on high-quality footage. For example, in order to obtain people gait features, it needs to successfully extract the human body from cluttering backgrounds and track the human body throughout the entire video period. Due to the limitations of image segmentation, tracking, and body part occlusion, it is unlikely to guarantee the reliability of gait features from the footage. Another type of depth information approach introduces additional depth information taken by RGB-D camera. The depth information does provide another source of helpful information for clothing change LT-reID in the wild. However, it introduces extra complexity to the camera setup. Besides, due to its limited sensing distance, it is still far from real practice.

In order to handle the clothing change LT-reID issue, an open question is: Is there a method that does not need to rely on the unreliable biometric information heavily? It is also difficult to build a clothing change LT-reID dataset that is suitable for the study of LT-reID in the wild with a reasonable data scale. This is because based on the experiences of existing person re-ID datasets, the dataset should have: 1) sufficient large number of people IDs, 2) dynamic shooting with true environments, 3) various clothing on each person. This thesis will introduce a new clothing change LT-reID dataset to the community for experiment evaluations. Based on this dataset, a new approach to tackle clothing change LT-reID is presented.

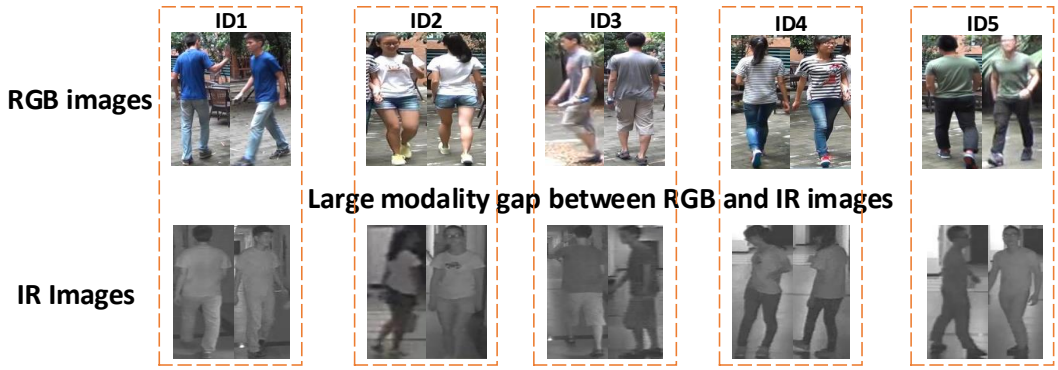


Figure 1.5 : Examples of infrared-visible LT-reID. Even human performance is hard to distinguish the ID when using RGB images as query and IR images as gallery.

1.1.4 Challenges on Footages Taken by Different Styles of Cameras in LT-reID

As a typical and widely seen case, infrared-visible LT-reID comes into sight and gradually attracts more attention Huang et al. (2021a). Unlike traditional person re-ID that all person images are captured by RGB cameras in the daytime under well-lighted conditions, the well-lighted condition cannot be guaranteed in the night time. Therefore, in order to handle re-ID using infrared-visible LT-reID, two different types of sensing cameras are adopted under different lighting environments (*i.e.*, using RGB (IR) cameras to capture images under well-lighted (poor-lighted) conditions). The re-ID matching is carried out between RGB and IR images (as shown in Fig. 1.5). A sophisticated person re-ID system should have the capability to handle such a cross-modality retrieval issue (*i.e.*, using RGB image to retrieve IR image and vice versa) for person re-ID in the wild.

To deal with the infrared-visible LT-reID issue, a model should be able to tolerate discrepancies between the modalities of RGB and IR (Wu et al., 2017). Existing infrared-visible LT-reID approaches can be categorized into two types:

1. **Feature alignment**(Adopting feature-level constraint to learn modality shared features) (Wu et al., 2017; Ye et al., 2018a; Dai et al., 2018; Ye et al., 2018b, 2019).

2. **Feature alignment + data alignment**(Integrating camera style transfer and feature-level constraint) (Wang et al., 2019c,a, 2020; Choi et al., 2020).

Amongst existing approaches, an open question does not get much attention. That is, existing approaches normally adopt ImageNet-trained ResNet50 (He et al., 2016) to learn shared features across two modalities. The ResNet50 is pretrained based on the larger scale ImageNet database (only contains RGB images) which is not compatible with the relevantly small-scale IR data in many practices. Thus, information of the IR modality is inclined to be overwhelmed by RGB information during training, which causes of the shared featured learned from RestNet50 to be compromised.

In order to deal with infrared-visible LT-reID in the wild, this thesis will unveil the Modality Bias Training (MBT) problem and proposes a Dual-level Learning Strategy (DLS) to alleviate this problem. By alleviating the MBT issue, the proposed DLS can be used to improve the performance of existing infrared-visible LT-reID methods.

1.2 Research Problems

The aims of the project are as follows:

i. The situations of person re-ID in the wild are very diverse (containing dramatic variations of illumination, viewpoint, and pose). Basically, the problem can be sorted out in theory by feeding sufficient training data in supervised training, which can represent real diverse situations. However, it is inevitable to lack of data. Thus, ultimately the training data is still not enough. Generating data

to compensate for the limited training data is an obvious solution, which will be introduced in the thesis.

ii. Due to the diversity of situations in the real world, the training data cannot guarantee to cover all possible cases in practical situations. That is, when a trained model is applied to actual cases, it is possible to see some scenarios that are not seen in training data. In other words, the domain where the model is trained is different from the domain where the model is applied. Ideally, if the model can tolerate the difference across domains, the model is able to sort out person re-ID in the wild. Thus, this thesis proposes to mitigate the influence of dramatic background change for the cross-domain domain person re-ID in the wild.

iii. Although clothing change can be regarded as a case of person re-ID in the wild, such unique challenges cannot be well sorted out using genetic solutions (*i.e.*, methods designed for the ST-reID scenario). In the meantime, clothing change is one of the widely seen cases in LT-reID in the wild so it is worth tackling this typical challenge.This thesis will also introduce new clothing change LT-reID datasets for evaluations to compensate for the lack of data to this area.

iv. In some special wild environments that need tight security surveillance, a common RGB camera may not fit the requirements well. For example, surveillance in a dark environment is a widely seen case. Using the infrared camera is efficient to sort out such challenging problems, although the cost may be expensive. However, there is an immediate problem: how it can recognize the same person who appears under an RGB camera and an infrared camera. In order to tackle such a challenge, this thesis will provide new insights into this special but indispensable situation to ensure that re-ID systems can tolerate diverse person re-ID cases in the wild.

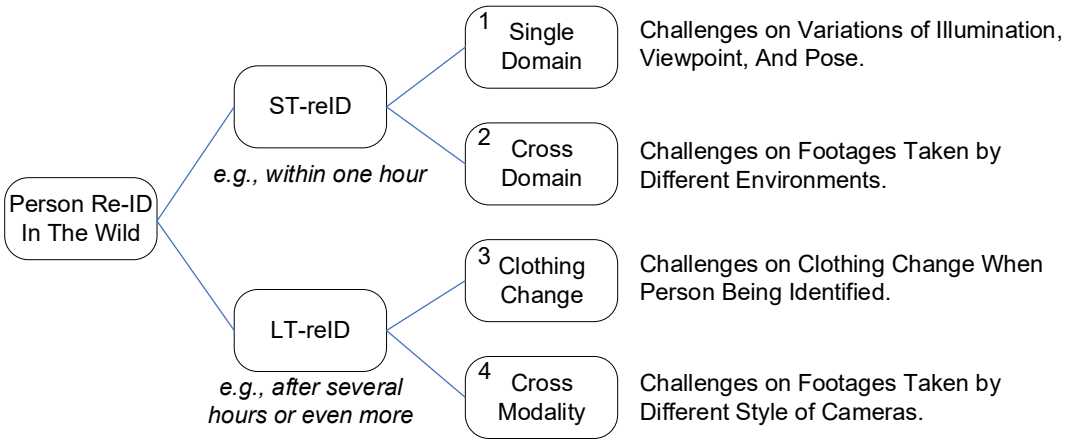


Figure 1.6 : Two main scenarios in person re-ID. Categorizing according to the time gaps of a person reappears under different cameras. For each category, two important challenges need to be addressed for person re-ID in the wild.

To sum up, as shown in Fig. 1.6, person re-ID in the wild, it is categorized into two main scenarios according to different time gaps (ST-reID and LT-reID). For each scenario, two critical issues need to be addressed in order to achieve a sophisticated re-ID system used for the real world. This thesis will deeply investigate issues presented in different re-ID scenarios and provide deep insights and solutions to the field of study.

1.3 Thesis Contribution

The contribution of this thesis are as follows:

Tackle Challenges on Variations of Illumination Viewpoint, And Pose:

This thesis attempts to use unlabeled generated data by GAN for single-domain ST-reID training data augmentation. The training data consists of real data (with true ID labels) and fake data (with proposed virtual labels). To use these generated data, this thesis proposes a Multi-pseudo Regularized Label (MpRL) and assigns

these virtual labels to the generated data when they are involved in training. Unlike the traditional label, which usually is a single integral number, the virtual label proposed in this thesis is a set of weight-based values, each individual of which is a number in (0,1] called multi-pseudo label and reflects the degree of relation between each generated data to every predefined class of real data. A comprehensive evaluation is carried out by adopting two state-of-the-art convolutional neural networks (CNNs) in experiments to verify the effectiveness of MpRL. (Chapter 3)

Tackle Challenges on Footages Taken by Different Environments:

This thesis observes that if backgrounds in the training and testing datasets are very different, it dramatically introduces difficulties in extracting robust pedestrian features, thus compromising the cross-domain person re-ID performance. This thesis formulates such problems as a background shift problem. A Suppression of Background Shift Generative Adversarial Network (SBSGAN) is proposed to generate images with suppressed backgrounds to reduce the domain shift between the training data and testing data. Unlike existing methods that simply remove backgrounds using binary masks (zero out pixels regarded as background region), SBSGAN allows the generator to decide whether pixels should be preserved or suppressed to reduce segmentation errors caused by noisy foreground masks.

In addition to the SBSGAN module, this thesis introduces a Densely Associated 2-Stream (DA-2S) network with an update strategy to best learn discriminative ID features from generated data that consider both human body information and certain useful ID-related cues in the environment. The built re-ID model is further updated using target domain data with corresponding virtual labels calculated by an unsupervised clustering algorithm. Compared with existing methods that heavily rely on the quality of the data transformation model, the proposed method can utilize knowledge from the target domain. In addition, exiting clustering-based

approaches do not consider the raw information gap between data of two domains. The proposed method also considers the domain shift on the data level by mitigating background shifts between different domains. (Chapter 4)

Tackle Challenges on Clothing Change When A Person Being Identified:

Existing LT-reID datasets usually contain a small number of IDs/images and are collected under constrained environments. This thesis introduces a clothing change LT-reID dataset using the street snapshot of celebrities crawled from the Internet (*e.g.*, google images). This new dataset has sufficient large number of identities, dynamic shooting with true environment, and various clothing on each person. Based on this dataset, a new approach to tackle clothing change LT-reID is presented. Compared with existing approaches that use unreliable biometric information, this new approach does not rely on the biometric information. In addition, instead of using traditional scalar neurons (SN) to design the network, the new method attempts to adopt vector-neuron (VN) capsules. Compared with SN, one extra-dimensional information in VN is able be aware of the change of clothing for a person. Finally, a ReIDCaps network with capsule neurons is introduced to deal with the clothing change in LT-reID. (Chapter 5)

Tackle Challenges on Infrared-Visible LT-reID:

This thesis observes existing cutting-edge approaches having a Modality Bias Training (MBT) problem and proposes a Dual-level Learning Strategy (DLS) to alleviate this problem. The MBT issue is caused by using the ImageNet-trained model as the backbone for model design since the ImageNet-trained model is pretrained with RGB images only. Thus, the information of the IR modality is inclined to be overwhelmed by RGB information during training, which causes the performance to be compromised. The proposed DLS is able to alleviate the MBT issue on both

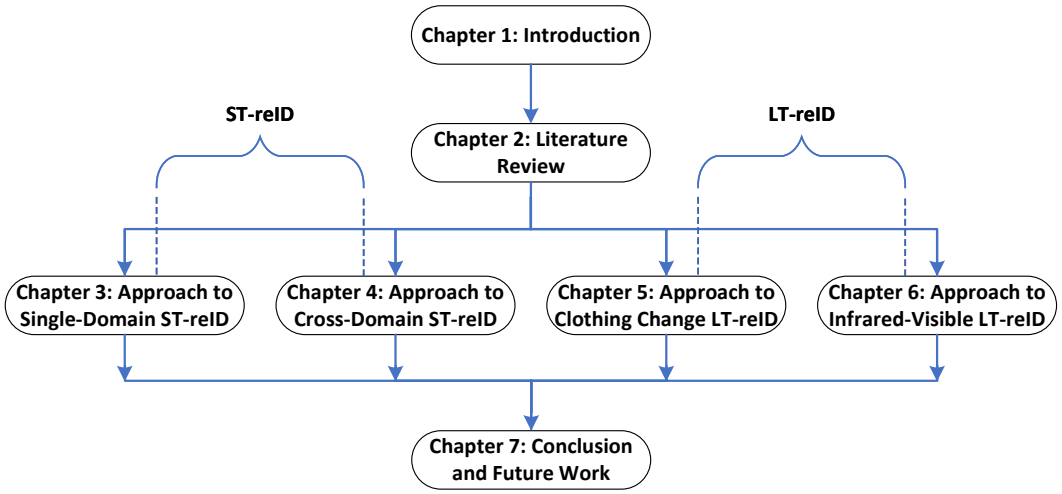


Figure 1.7 : Thesis structure.

feature level (focusing on ID-exclusive input image pair) and data level (introducing generated neutral data between RGB and IR).

1.4 Thesis Organization

This thesis is organised as follows:

- *Chapter 2*: This chapter presents a survey of person re-ID methods
- *Chapter 3*: This chapter presents method of labelling generated data for single-domain ST-reID
- *Chapter 4*: This chapter presents the method for cross-domain ST-reID by exploring the impact caused by background shift between training and testing data.
- *Chapter 5*: This chapter presents the method for clothing change LT-reID. New benchmark datasets and approaches are introduced.

- *Chapter 6*: This chapter presents the method for infrared-visible LT-reID. The MBT issue is unveiled for this cross-modality retrieve issue and corresponding solution is given to tackle MBT issue.

- *Chapter 7*: A brief summary of the thesis contents and its contributions are given in the final chapter. Recommendation for future works is given as well.

The thesis structure is also shown in Fig. 1.7.

Chapter 2

Literature Review

2.1 Review of Single-Domain ST-reID

As mentioned in Chapter 1, the main challenge in single-domain ST-reID in the wild is dealing with variations of illumination, viewpoint, and pose. It has drawn growing interest from research to practical applications (Gong et al., 2014). In the past few years, three mainstream approaches, including non-deep learning approaches, deep learning approaches, and boosting tricks are presented in existing works.

2.1.1 Non-Deep Learning Approaches

In non-deep learning approaches, there are two important modules: feature extraction module and metric learning module. The former one is used to extract robust person descriptions to represent different person IDs. The latter one is used to distinguish different IDs using the output of the feature extraction module. The two processes (*i.e.*, feature extraction and metric learning) are normally conducted independently (Zhong et al., 2018a; Zhang et al., 2018a; Li et al., 2020b, 2018c, 2019b).

In feature extraction, (Liao et al., 2015) propose the local maximal occurrence feature against viewpoint changes and handle illumination variations. (Chen et al., 2015) introduce a mirror representation to alleviate the view-specific feature distortion problem. (Zheng et al., 2015) present a bag-of-words descriptor that describes each person by a visual word histogram.

In metric learning, (Zheng et al., 2011) use a relative distance comparison method to minimize the probability of a negative person image pair with a larger distance than a positive pair. (Liao and Li, 2015) propose logistic metric learning via an asymmetric sample weighting strategy. (Li et al., 2013) employ a locally-adaptive decision function that integrates traditional metric learning with a local decision rule. (Yu et al., 2017a) learn an asymmetric metric that projects each view in an unsupervised learning fashion.

2.1.2 Deep Learning Approaches

Unlike the above non-deep learning methods that are manually designed to handle the single-domain ST-reID task, deep learning discovers more implicit information in matching persons and achieves many state-of-the-art results in a wide range of research areas (Huang et al., 2018b; Zhong et al., 2020; Liu et al., 2020; Zhang et al., 2020a).

To distinguish the appearance of a person at the right spatial locations and scales, (Qian et al., 2017) propose a multi-scale deep learning model to learn discriminative features. (Lin et al., 2017) introduce a consistent-aware deep learning approach which seeks the globally optimal matching. Also, deep features over the full body and body parts are captured from local context knowledge by stacking multi-scale convolutions in (Li et al., 2017a). Two-stream network (Geng et al., 2016; Zheng et al., 2018), triplet loss network (Ding et al., 2015; Cheng et al., 2016) and quadruplet network (Chen et al., 2017a) have been designed for single-domain ST-reID. (Zheng et al., 2016a,b) propose an identification (Identif) CNN. This network takes person re-ID as a multi-classification task, and a CNN embedding is learned to discriminate different identities in training. Beyond that, (Zheng et al., 2018) propose a two-stream deep neural network. A verification function that separates two input images belonging to the same or different identities is considered to

improve the performance of the Identif network further. In testing, the above two networks extract CNN embeddings in the last convolutional layer to compare the similarity between two inputs using squared Euclidean distance.

2.1.3 Boosting Tricks

Deep learning approaches have already achieved encouraging performance for single-domain ST-reID. Thus, another active research line attempts to further improve the performance of deep learning approaches as boosting tricks without changing the network architecture. (Huang et al., 2016) formulate the single-domain ST-reID task as a tree matching problem, and a complete bipartite graph matching is presented to refine the final matching result at the top layer of the tree. To study single-domain ST-reID with the manifold-based affinity learning, (Bai et al., 2017) introduce a manifold-preserving algorithm plunging into existing re-ID algorithms to enhance the performance. As a post-processing method, re-ranking (Zhong et al., 2017) exploits the relationships amongst initial ranking list to improve the performance of re-ID system. Finally, human feedback in-the-loop is adopted that provides an instant improvement to re-ID ranking on-the-fly (Ali et al., 2010; Wang et al., 2016b; Liu et al., 2013).

For person re-ID in the wild, acquiring enough labeled data is expensive for single-domain ST-reID to against variations of illumination, viewpoint, and pose. In order to remedy the lack of data issue, (Zheng et al., 2017b) combine the generated fake data by GAN with real data in network training. This method assigns virtual labels to generated data with a uniform label distribution over all the predefined training classes. However, there are two issues being ill-considered: 1) When mapping the fake data to the real data domain, the weights over all predefined training classes are regarded identical. 2) In the generated data domain, all data share the same virtual label. Therefore, how to utilize the generated fake data to boost

the performance of the model in an effective way needs to be considered.

2.2 Review of Cross-Domain ST-reID

In order to tolerate the difference between different domains, existing cross-domain ST-reID approaches can be categorized into two types: 1) general inter-domain style transfer, 2) clustering-based approaches.

2.2.1 General Inter-Domain Style Transfer for Cross-Domain ST-reID

Recently, following image-to-image translation approaches (*e.g.*, CycleGAN (Zhu et al., 2017) and StarGAN (Choi et al., 2018)), some studies focus on inter-domain style transfer to mitigate the domain shift for the cross-domain ST-reID. (Deng et al., 2018) propose SPGAN to transfer general image style between domains. (Wei et al., 2018a) introduce PTGAN to transfer body pixel values and generate new backgrounds with similar statistic distribution of the target domain. Unlike SP-GAN, PTGAN explicitly considers the background shift problem between domains. However, PTGAN overlooks the fact that backgrounds should be suppressed rather than retained, because the background shift may degrade the Unsupervised Domain Adaptation (UDA) re-ID performance. (Liu et al., 2019) present an ATNet that can decompose cross-domain style transfer into a set of factor-wise sub-transfers (*e.g.*, illumination, resolution, *etc.*). These factor-wise transfers can then be fused by an ensemble strategy to magnitude different sub-factors in cross-domain image generation. In addition to the inter-domain style transfer, (Zhong et al., 2018b) propose transferring the style of images between cameras to reduce the domain shift by using StarGAN (Choi et al., 2018). (Qi et al., 2019) take both cross-domain and cross-camera issues into consideration to alleviate discrepancy between domains and different cameras. (Li et al., 2019a) add a pose disentanglement scheme to improve the cross-domain image transfer process. (Chen et al., 2019c) transfer images from a

source domain to a target domain with diverse target domain contexts. A synthetic dataset is proposed to generalize illumination between different light conditions for cross-domain ST-reID in (Bak et al., 2018). Cycle-consistency translation of GAN is employed to retain identities of the synthetic dataset. However, all these methods do not consider the dramatic background shift between the training and testing domains, which can be regarded as the essential reason that jeopardizes the final performance.

In order to deal with the background shift problem, one possible solution is to completely remove backgrounds using binary body masks obtained by semantic segmentation or human parsing methods. Currently, methods such as Mask-RCNN (He et al., 2017) and JPPNet (Liang et al., 2018) can obtain body masks with pretrained model on large-scale datasets, *e.g.*, MS COCO (Lin et al., 2014) and LIP (Liang et al., 2018). However, masks obtained by these methods often contain errors due to low-resolution person images, and highly dynamic person poses. Directly using noisy masks may further jeopardize the performance of cross-domain ST-reID. Therefore, how to better suppress background shift to improve cross-domain ST-reID performance is still an open question.

2.2.2 Clustering-Based Cross-Domain ST-reID

So far, several works attempt to use clustering-based methods to explore the natural characteristics of images in the target domain for cross-domain ST-reID (Fan et al., 2018; Song et al., 2020; Fu et al., 2019; Zhang et al., 2019b). These works use raw source domain images with ground-truth ID labels to pretrain a re-ID model. Then, this pretrained model is used to extract features of training images in the target domain. An unsupervised clustering approach (*e.g.*, k-means++ (Arthur and Vassilvitskii, 2006), DBSCAN (Ester et al., 1996), or HDBSCAN (Campello et al., 2013)) is adopted to classify such features into different clusters. Each image is

assigned a virtual label according to the clustering results. The images, along with the corresponding virtual labels, are used to update the pretrained CNN model. Both (Fan et al., 2018; Song et al., 2020) employ the classical ID-discriminative Embedding (IDE) network (Zheng et al., 2016a; Deng et al., 2018; Zheng et al., 2017b) as the pretrained CNN model on the source domain. Based on the IDE network, (Fu et al., 2019) promote the performance of clustering-based solution by adopting body part partition. (Zhang et al., 2019b) integrate multiple losses in a conservative stage and a promoting stage to enrich the discriminative ability of features for the cross-domain ST-reID. However, existing clustering-based methods in fact mix target domain data with virtual labels and source domain data with true annotation information. Such a simple mixture does not consider the raw information shift between data distributions of two domains. In the cross-domain person re-ID issue, this shift can be largely contributed by the background differences between different domains.

2.3 Review of Clothing Change LT-reID

Clothing change can be regarded as a commonly seen case for LT-reID in the wild. Owing to the dramatic clothing change for each individual, such an issue cannot be well sorted out using genetic solutions. In order to deal with the clothing change issue, existing approaches are categorized into two types: 1) biometrics based approaches, 2) depth information compensation approaches.

2.3.1 Biometrics Based Approaches

Biometric traits have been adopted for clothing change LT-reID, including motion (Zhang et al., 2018b), body contour/shape (Yang et al., 2019a; Qian et al., 2020; Li et al., 2020a), and face (Yu et al., 2020; Wan et al., 2020b). **1) Motion.** (Zhang et al., 2018b) extract motion features (Histograms of Optical Flow (HOF) (Laptev

et al., 2008) and Motion Boundary Histogram (MBH) (Wang et al., 2013)) to associate different persons for LT-reID. However, in order to extract robust motion features, a complete motion cycle is normally required, making it difficult to be applied to image-based LT-reID scenario. **2) Body Contour/Shape.** (Yang et al., 2019a) use body contour information for LT-reID. To achieve this, a person only can change her/his clothing moderately (*i.e.*, wears clothes of a similar thickness), which is confined to a limited LT-reID application scenario. (Qian et al., 2020) employ a pose detector to detect and localize body joints which are used for learning body shape features based on the spatial relationship between joints. However, the body shape is sensitive towards camera shooting angles, which may not be effective in the real world. **3) Face.** (Wan et al., 2020b) and (Yu et al., 2020) learn face features for LT-reID. However, the face only takes up a small part of the body region, and it is not always available when the image quality or camera view (*e.g.*, back view) is poor. In order to deal with clothing change LT-reID in the wild, theoretically, biometric features should be robust. However, it heavily relies on high-quality footage taken by different cameras. Due to the low image resolution, poor shooting viewpoint, and occlusion issue, extracting robust biometric features is hard to be guaranteed.

2.3.2 Depth Information Compensation Approaches

RGB-D images taken by depth camera (*e.g.*, Kinect) are used in existing clothing change LT-reID works. (Barbosa et al., 2012a) propose extracting 3D soft-biometric features for LT-reID using the depth information captured by Kinect. (Munaro et al., 2014) transform persons' point clouds to a standard pose via depth information. The transformed point clouds are used for composing 3D models for LT-reID to eliminate impacts caused by clothing change. (Haque et al., 2016) leverage raw depth video data as training inputs, and propose a recurrent attention model that re-identifies persons by focusing on small, discriminative body regions to tackle clothing change.

However, the depth sensor is hard to be widely deployed for real-world LT-reID due to the complexity of camera setup and the limitation on sensing distance.

In order to deal with clothing change LT-reID, existing approaches usually try to extract biometric information or utilize RGB-D camera to achieve depth information. However, there are two main concerns: 1) biometric information heavily relies on the quality of footage, which may not be robust, 2) using RGB-D camera needs to introduce extra complexity to the camera setup, which is still far from real practice. Therefore, in order to tackle the clothing change issue for person re-ID in the wild, more solid and practical solutions should be proposed.

2.4 Review of Infrared-Visible LT-reID

In order to recognize the same person who appears under an RGB camera and an infrared camera, infrared-visible LT-reID is an immediate problem that should be addressed. Existing approaches can be categorized into two types: 1) feature alignment approaches, 2) feature alignment + data alignment.

2.4.1 Feature Alignment

The main concern of the feature alignment approach is to learn modality shared features from both RGB and IR images by adding certain constraint on feature level in a network training stage. (Wu et al., 2017) concatenate zero paddings with input images to learn cross-modality shared features through a one-stream network. (Ye et al., 2018a) propose a two-stream network that joint ID loss and contrastive loss to reduce the modality gap on the feature level. (Dai et al., 2018) cast the infrared-visible LT-reID problem as a generative adversarial training task in order to learn discriminative modality shared features across RGB and IR images. By considering cross-modality and intra-modality variations simultaneously, (Ye et al., 2018b, 2019) introduce a dual constrained top-ranking loss to enhance the discriminability of

shared features learned from two different modalities. (Feng et al., 2020) learn modality shared features from two independent modality-specific networks via cross-modality Euclidean constraint. This approach can tackle the cross-modality issue for Infrared-visible LT-reID somehow. However, it does consider the large modality discrepancy on the data level. Directly mapping images of two different modalities into feature space may not be effective (Wang et al., 2019c).

2.4.2 Feature Alignment + Data Alignment

In addition to the feature space alignment, data level alignment is also important to alleviate the large discrepancy across different modalities for infrared-visible LT-reID. (Wang et al., 2019c) use a generator to generate cross-modality images from a latent vector z, which is encoded by an image encoder. A real image and its corresponding cross-modality generated counterpart are concatenated along the channel dimension to construct a multi-spectral image used to learn modality shared features. That is, it mixes the original image on a modality and the corresponding make-up image generated by the generator together for the following re-ID process. (Wang et al., 2019a) propose generating IR images based on given real RGB images. That is, it converts a given RGB image to an IR image. Thus, the cross-modality image matching issue becomes a common image matching issue between the generated IR images and real IR images. (Yang et al., 2020) propose a random walk solution for mining reliable relationships between images by traversing heterogeneous manifolds in feature space of IR and RGB modalities. The main focus of (Yang et al., 2020) is to alleviate noisy similarities between modalities.

Existing approaches generally follow the IDE or IDE-T network architecture. These approaches normally adopt ImageNet-trained ResNet50 (He et al., 2016) to learn shared features across two modalities. Consequently, the learned shared feature cannot acquire the true characteristics that are not biased to either modality.

Instead, it will contain more information from RGB modality due to the ImageNet-trained ResNet50 is pretrained using RGB images only. Therefore, how to deal with this modality bias training issue is still an open question for infrared-visible LT-reID.

2.5 Datasets for Person Re-ID In The Wild

In past years, several person re-ID datasets are proposed. Most of them are the short-term scenario (used for the single-domain ST-reID task and cross-domain ST-reID task). Some of them are used for the clothing change LT-reID. Two datasets are presented for the infrared-visible LT-reID study. This section will introduce some typical existing person re-ID datasets.

2.5.1 Existing Datasets for ST-reID

There are several ST-reID datasets that have been proposed along with the development of researches in the community. In general, most of them belong to the ST-reID scenario. These datasets do not have the change of clothing for a person or involve different types of cameras (*e.g.*, RGB-D cameras or infrared cameras). Amongst them, VIPeR (Gray and Tao, 2008), CUHK01 (Li et al., 2012), CUHK03 (Li et al., 2014), Market1501 (Zheng et al., 2015), and DukeMTMC-reID (Zheng et al., 2017b), are most widely used ST-reID datasets that contain very diverse situations in the wild (*e.g.*, dramatic variations caused by the changes of illuminations, viewpoints, poses, and background). Tab. 2.1 lists some basic information of the five widely used ST-reID datasets. These datasets can be used for either single-domain ST-reID studies (both training and testing are conduced using the same dataset) or cross-domain ST-reID studies (training on one dataset and testing on another dataset) since there is no clothing change or cross-modality issue.

Table 2.1 : Basic information of some exiting person re-ID datasets.

Dataset	#IDs	#Images or #sequences	Environments	Type of cameras
Traditional ST-reID Datasets (images)				
VIPeR (Gray and Tao, 2008)	632	1,264	indoor+outdoor	RGB cameras
CUHK01 (Li et al., 2012)	971	3884	campus	RGB cameras
CUHK03 (Li et al., 2014)	1467	13,164	campus	RGB cameras
Market1501 (Zheng et al., 2015)	1501	32,217	campus	RGB cameras
DukeMTMC-reID (Zheng et al., 2017b)	1404	36,441	campus	RGB cameras
Clothing Change LT-reID Datasets (video sequences)				
PAVIS (Barbosa et al., 2012b)	79	316	under controlled	Kinect
BIWI (Munaro et al., 2014)	50	50	under controlled	Kinect
IAS-Lab (Munaro et al., 2014)	11	33	under controlled	Kinect
DPI-T (Haque et al., 2016)	12	300	under controlled	Kinect
RGB Video-Based Dataset (Zhang et al., 2018b)	30	240	under controlled	RGB cameras
Infrared-Visible LT-reID Datasets (images)				
SYSU-MM01 (Wu et al., 2017)	491	22,258 RGB images + 11,909 IR images	campus	RGB + near-infrared cameras
RegDB (Nguyen et al., 2017)	412	8240	outdoor	RGB + far-infrared cameras

2.5.2 Existing Datasets for Clothing Change LT-reID

Current approaches that are mainly designed for the ST-reID scenarios may not be suitable for the LT-reID scenario. To achieve a sophisticated re-ID system, the clothing change problem should be explicitly considered for person re-ID in the wild. Unfortunately, existing ST-reID datasets do not involve such characteristics. Therefore, this thesis will introduce new clothing change LT-reID datasets to facilitate future researches in the community.

To the best of our knowledge, five small-scale datasets consider the clothing change problem for person re-ID (refer to Tab. 2.1). They can be categorized into two types: RGB-D and normal RGB video-based datasets. To handle the challenge of clothing changes, the depth information has been leveraged to extract additional 3D soft-biometric beyond the RGB color cue. Accordingly, RGB-D datasets such as PAVIS (Barbosa et al., 2012b), BIWI (Munaro et al., 2014), IAS-Lab (Munaro et al., 2014), and DPI-T (Haque et al., 2016) have been proposed using the Kinect camera under controlled environments. These datasets are helpful on proof of concept by exploiting 3D information using depth information. There is also a video-based clothing change LT-reID dataset (Zhang et al., 2018b). However, the scale of this dataset is also too small, and the environment is controlled under an indoor camera that is not suitable for person re-ID in the wild. Therefore, a new clothing change LT-reID dataset that can satisfy the requirement of practical usage collected under a highly diverse environment with multiple camera views is needed.

2.5.3 Existing Datasets for Infrared-Visible LT-reID

Two publicly available infrared-visible datasets are widely used for evaluation, including SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017). For each person, SYSU-MM01 (RegDB) uses RGB and near-infrared (far-infrared) cameras to take footages. Both datasets are collected under uncontrolled situations in the

real-world environment, which is suitable for the evaluation of person re-ID in the wild. Tab. 2.1 shows some necessary information of the two datasets.

2.6 Summary

This chapter revisits the existing person re-ID approaches and datasets. According to different person re-ID tasks, this chapter respectively introduces the corresponding background knowledge, principal theories, and mainstream technologies. This thesis will present detailed solutions for different person re-ID tasks in the wild from the next chapter.

Chapter 3

Labelling Generated Data for Single-Domain ST-ReID

3.1 Motivation

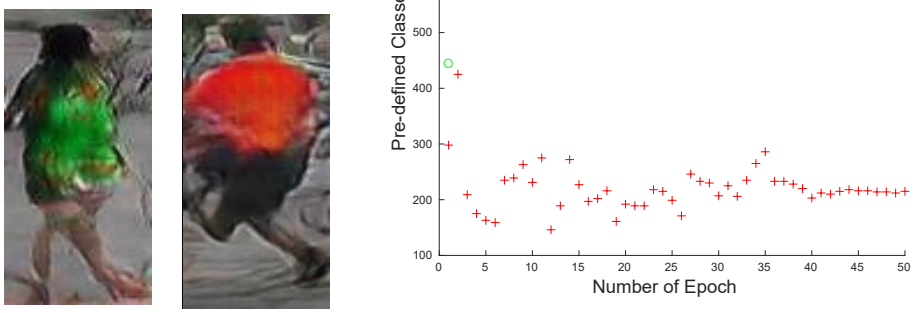
Person re-ID is a challenging task of recognizing a person amongst different camera views. It is a typical computer vision problem that requires sufficient training data to learn a discriminative model. In the past few years, deep learning has demonstrated its performance in person re-ID by producing several state-of-the-art methods (Huang et al., 2017b; Qian et al., 2017; Lin et al., 2017; Zheng et al., 2016a,b, 2018). To this end, sufficient labeled training data is essential to train deep models in a supervised learning fashion in order to tackle challenges caused by variations of illumination, viewpoint, and pose. Although some large datasets, *e.g.*, Market-1501 (Zheng et al., 2015), DukeMTMC-reID (Zhedong et al., 2017), and CUHK03 (Li et al., 2014) have been proposed. However, due to the expensive cost of data acquisition that needs to manually find corresponding labels of pedestrians who appear under different camera views, the diverse of data is still limited. In 2014s, Generative Adversarial Network (GAN) was proposed to generate data (images) with perceptual quality (Goodfellow et al., 2014). Since then, several improved approaches (Radford et al., 2015; Arjovsky et al., 2017; Gulrajani et al., 2017) were presented to further improve the quality of generated data. However, how to use the data is still an open question.

Using generated data to solve the problem of data diversity is a promising solution. Therefore, this thesis attempts to use unlabeled data generated by GANs

to improve the person re-ID performance further. In all existing methods by using GAN, there are two main challenging points in order to assure the better performance: 1) high quality data generated by GAN (Radford et al., 2015; Arjovsky et al., 2017; Gulrajani et al., 2017), 2) a better strategy to use the generated data into the training model (Zhedong et al., 2017). Many works focus on the first point. This thesis particularly focuses on the second point. This thesis follows the same pipeline in (Zhedong et al., 2017) that incorporates generated data with real data to train deep models in a semi-supervised learning fashion. Compared with previous attempts (Salimans et al., 2016; Radford et al., 2015) that perform semi-supervised learning in the discriminator of GANs, sufficient unlabeled generated data will directly participate in training as the supplementary of limited labeled real data.

In 2017s, a related work was first proposed in (Zhedong et al., 2017) that introduced a method called Label Smooth Regularization for Outliers (LSRO). This method assigns virtual labels to generated data with a single unified label distribution for different predefined training classes. The single unified distribution considers weights of all predefined training classes equally. More specifically, if the number of predefined training class is K, the weight of each class is equally divided into $1/K$. By doing so, LSRO shows two undesirable characteristics: 1) On the real data domain, the weights for different predefined training classes are identical. 2) On the generated data domain, all data share the same virtual label.

For the first fact, since every individual predefined training class of real data has the same weight, the data generated by GAN should be able to embed equal properties of all predefined training classes. However, during the actual GAN training process, only a random minibatch of real data samples are used in each iteration. That is, only certain real data from some classes (not all predefined training classes) are used in GAN training in each iteration to generate artificial data following a continuous noise distribution (Goodfellow et al., 2014; Radford et al., 2015). Con-



(a) (b) (c) Fitting generated images (a) and (b) into predefined training classes using markers 'o' and '+', respectively.

Figure 3.1 : Label distribution of predefined training classes (c) for generated images (a) and (b). Only the maximum predicted probability of predefined training classes is activated along with the training process (see (c)). Distinguishable label distributions can be observed between (a) and (b).

sequently, the data distribution between the generated and real data is biased by equally utilizing the weights from all predefined training classes in the real data domain. **A label needs to be assigned to generated data, which can reflect the different weights of different predefined training classes in GAN training.** For the second fact, it may not be correct to assign the same label to certain different generated data if the generate data has the distinct visual differences. Otherwise, ambiguous predictions may happen in training. Fig. 3.1(a) and 3.1(b) show two generated images with red and green clothes respectively. When these two images are fitted into predefined training classes (only using the maximum prediction probability) through 50 training epochs, distinguishable label distribution can be observed in Fig. 3.1(c). Therefore, using the same virtual label for all

generated data is improper. **Dynamically assigning different virtual labels to each generated data is required.**

Although LSRO has demonstrated its effectiveness in (Zhedong et al., 2017), the above problems still limit its effectiveness. To solve this problem, a Multi-pseudo Regularized Label (MpRL) is proposed as a virtual label assigned to generated data. Unlike LSRO, main differences of the proposed MpRL can be summarized in two-fold:

- Compared with LSRO using single unified label distribution, the proposed MpRL assigns each generated data a corresponding label which shows the likelihood of the affiliation of the generated data to all predefined training classes. Thus, the relationship between the generated data and predefined training classes can be substantially built, which makes generated data more informative when they incorporate with the real data in training.

- By differentiating different generated data, MpRL mitigates ambiguous prediction in training. Intuitively, different generated data present distinct visual differences and should have different impacts to the training. The proposed method is to embed such characteristics into the training model.

3.2 Labeling Generated Data Using Multi-Pseudo Regularized Label

In this section, the state-of-the-art virtual label LSRO (Zhedong et al., 2017) for person re-ID is revisited first. Then MpRL is introduced. Finally, three training strategies are proposed for MpRL.

3.2.1 LSRO for Person Re-ID

LSRO assumes that the generated data does not belong to any predefined training class and uses the single unified label distribution on each of them to address over-fitting (Zhedong et al., 2017). LSRO is inspired by Label Smoothing Regularization (LSR) (Szegedy et al., 2016) which assigns less confidence to the ground-truth label and assigns small weights to other classes. Formally, given a generated image g, its label distribution $q^g_{LSRO}(k)$ is defined as follows:

$$q^g_{LSRO}(k) = \frac{1}{K}, \tag{3.1}$$

where K is the number of predefined training classes in the real data domain, $k \in [1, ..., K]$ represents the k-th predefined training class. In training, the loss of LSRO to a generated image is defined as follows:

$$l_{LSRO} = -\frac{1}{K}\sum_{k=1}^{K} log(p(X_k)), \tag{3.2}$$

where X_k represents the output of k-th predefined training class, $p(X_k) \in (0, 1)$ is the softmax predicted probability of X_k belonging to the predefined training class k, defined as follows:

$$p(X_k) = \frac{e^{X_k}}{\sum_{j=1}^{K} e^{X_j}}. \tag{3.3}$$

- In Eq. 3.2, the forward loss is as follows:

$$\begin{aligned} l_{LSRO} &= -\frac{1}{K}\sum_{k=1}^{K} log(\frac{e^{X_k}}{\sum_{j=1}^{K} e^{X_j}}) \\ &= -\frac{1}{K}\sum_{k=1}^{K}(X_k) + log(\sum_{j=1}^{K} e^{X_j}). \end{aligned} \tag{3.4}$$

- While, the backward gradient is as follows:

$$\frac{\partial l_{LSRO}}{\partial X_k} = -\frac{1}{K} + \frac{e^{X_k}}{\sum_{j=1}^{K} e^{X_j}}. \tag{3.5}$$

3.2.2 Multi-Pseudo Regularized Label

This thesis uses the proposed MpRL to assign virtual labels to generated data when they are fed into the network during training. However, unlike LSRO, MpRL does not set the virtual label as a single unified distribution for different predefined training classes (*i.e.*, $1/K$). The weights for different predefined training classes are different in the proposed MpRL. In this way, a dictionary α is built to record the weights. Compared with LSRO (see Eq. 3.1), for a generated image g, its label is defined as follows:

$$q^g_{MpRL}(k) = \frac{\alpha_k}{K}, \tag{3.6}$$

where α_k represents the weight of k-th predefined training class in the dictionary α. The reason why different weights are considered in the proposed MpRL will be discussed in Sec. 3.3.3. The proposed MpRL does not belong to a specifically predefined training class but is constituted by different weights for each of them. To obtain α_k, this thesis first formulates the set of predicted probabilities $p(X)$ of a generated image over K predefined training classes as:

$$p(X) = \{p(X_k)|k \in [1, ..., K]\}\,. \tag{3.7}$$

Then, all elements in $p(X)$ are sorted from the minimum to maximum and saved to $p_s(X)$:

$$p_s(X) = \{p_s(X_n)|n \in [1, ..., K]\}\,, \tag{3.8}$$

where $p_s(X_1) == min(p(X))$ and $p_s(X_K) == max(p(X))$. α_k is obtained by taking the corresponding index of $p(X_k)$ in the set of $p_s(X)$:

$$\alpha_k = \phi_{p_s(X)}(p(X_k)), \tag{3.9}$$

where $\phi_{p_s(X)}(\cdot)$ returns the index of $p(X_k)$ in $p_s(X)$. By doing so, the corresponding relationship between real data and a generated image is built by utilizing different weights obtained through the predicted probabilities for different predefined training

classes. Combining Eq. 3.6 with Eq. 3.9, the proposed MpRL can assign different virtual labels to a generated image g when it is fed into the network in training:

$$q^{g}_{MpRL} = \left\{ \frac{\alpha_k}{K} | k \in [1, ..., K] \right\}, \tag{3.10}$$

'Multi-pseudo' is used to name the proposed virtual label because when compared with the one-hot pseudo label that only the maximum predicted probability is activated, all predicted probabilities are used in MpRL. To address over-fitting (*e.g.*, after several training iterations some weights from predefined training classes will become larger, while others may decrease to a pretty small value), Eq. 3.10 regularizes the gap between two contiguous weights to $1/K$. In this way, the proposed MpRL retains the weights for all predefined training classes, even though some of them may not or just producing a tiny contribution to the generated data.

By combining the generated data with real data in training, the cross-entropy loss is used to build the proposed MpRL as follows:

$$l_{MpRL} = -(1-y)log(p(X_c)) - y \cdot \lambda \cdot \sigma \sum_{k=1}^{K} (\frac{\alpha_k}{K} \cdot log(p(X_k))), \tag{3.11}$$

where c represents the label of a real image, $\frac{\alpha_k}{K}$ is defined in Eq. 3.6. λ is the parameter for the trade-off between generated and real data. If not specified, λ is set to 1. σ is a normalization factor. In Eq. 3.11, if the weights over K per-defined training classes are summed up (*i.e.*, $\sum_{k=1}^{K} \frac{\alpha_k}{K}$), the total weight equals to $\frac{(1+K)\cdot K}{2}$. Therefore, to normalize weights over K predefined training classes, σ is set to $\frac{2}{1+K}$.

For a real image $y = 0$, Eq. 3.11 is equivalent to softmax loss. For a generated image $y = 1$, only the MpRL is used. Overall, the network has two types of losses: one for real data and the other for generated data.

- In Eq. 3.11, the forward loss is as follows:

For a real image, $y = 0$:

$$\begin{aligned} l_{MpRL} &= -log(\frac{e^{X_c}}{\sum_{j=1}^{K} e^{X_j}}) \\ &= -X_c + log(\sum_{j=1}^{K} e^{X_j}). \end{aligned} \tag{3.12}$$

For a generated image, $y = 1$:

$$\begin{aligned} l_{MpRL} &= -\lambda \cdot \sigma \sum_{k=1}^{K} (\frac{\alpha_k}{K} \cdot log(\frac{e^{X_k}}{\sum_{j=1}^{K} e^{X_j}})) \\ &= -\lambda \cdot \sigma \sum_{k=1}^{K} (\frac{\alpha_k}{K} X_k - \frac{\alpha_k}{K} log(\sum_{j=1}^{K} (e^{X_j}))). \end{aligned} \tag{3.13}$$

- While, the backward gradient is as follows:

For a real image, $y = 0$:

$$\frac{\partial l_{MpRL}}{\partial X_c} = -1 + \frac{e^{X_c}}{\sum_{j=1}^{K} e^{X_j}}. \tag{3.14}$$

For a generated image, $y = 1$:

$$\frac{\partial l_{MpRL}}{\partial X_k} = -\lambda \cdot \sigma \cdot \frac{\alpha_k}{K} (1 - \frac{e^{X_k}}{\sum_{j=1}^{K} (e^{X_j})}). \tag{3.15}$$

3.2.3 Training Strategy

To further investigate the effectiveness of the proposed MpRL, three different training strategies, including one static (constant virtual labels) and two dynamic (iteratively updated) approaches are introduced. Descriptions are as follows:

- **Static MpRL (sMpRL).** The sMpRL is assigned to each generated data before training the network. A pretrained Identif network (refer to Sec. 3.4.2) is used to assign sMpRL. Specifically, 1) the Identif network is pretrained by using all real training data; 2) Eq. 3.3 is utilized to calculate the predicted probability over K predefined training classes for each generated data; 3) Eq. 3.10

Algorithm 1: The training strategy of the dMpRL-II: dynamically update MpRL from the intermediate point to change the likelihood of the affiliation of the generated data to all predefined training classes iteratively.

Input: Real data set: R;

Generated data set: G;

Merged data set: $D = R \cup G$;

Loss for the real data set: l_1;

Loss for the generated data set: l_2.

```
for number of training epochs do
    Shuffle D ;
    for number of training iterations in each epoch do
        Set l_1 = 0, l_2 = 0;
        Sample minibatch from D → D';
        Select real data R' from D';
        Set y = 0 in Eq.3.11;
        Calculate loss l_1 for R';
        if number of epochs ≥ 20 then
            Select generated data G' from D';
            Assign MpRL to G' using Eq.3.10;
            Set y = 1 in Eq.3.11;
            Calculate loss l_2 for G';
        Calculate the final loss = l_1 + l_2 × 0.1 ;
        Backward propagation;
        Update parameters;
final;
```

is used to assign a sMpRL to each generated data, and the assigned sMpRL remains unchanged during the whole training process. This implementation is similar to LSRO except that the proposed MpRL considers different weights for different predefined training classes instead of regarding them equally.

- **Dynamic MpRL-I (dMpRL-I): Dynamically Update MpRL from scratch.** During training, dMpRL-Is are dynamically assigned to each generated data using Eq. 3.10, and they will be updated iteratively to change the likelihood of the affiliation of the generated data to all predefined training classes. Therefore, the same generated data may receive a different dMpRL-I each time when it is fed into the network. This dynamic progress starts from the first minibatch fed into the network until the training is completed. Notably, generated data will assign random dMpRL-Is if they are involved in the first training iteration.

- **Dynamic MpRL-II (dMpRL-II): Dynamically Update MpRL from the intermediate point.** This strategy tries to assign dMpRL-II to generated data after 20 epochs when the CNN model becomes relatively stable, and also they will be updated iteratively. That is, in Eq. 3.11, $y = 0$ until 20-th epochs, it is set to 1. Also, the loss is set to 0.1 and 1 for the generated and real data respectively. Therefore, under this training strategy, λ is set to 0.1 in Eq. 3.11. The detailed training strategy is shown in Algorithm 1.

3.3 Benefits of Multi-pseudo Regularized Label

The all-in-one (Odena, 2016; Salimans et al., 2016), one-hot pseudo (Lee, 2013), and LSRO (Zhedong et al., 2017) are used as the comparison experiments. Fig. 3.2(b), (c) and (d) respectively illustrate the label distributions. Given a generated image, a new label that does not belong to any predefined training class is assigned to it

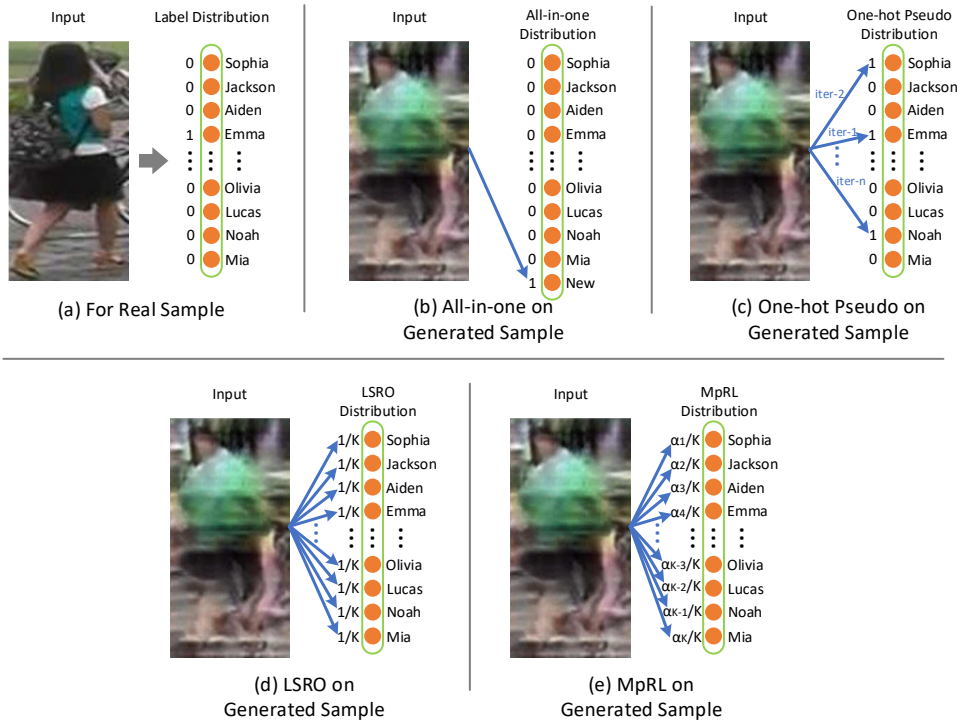


Figure 3.2 : The label distributions of real and generated data. The ground-truth label is assigned to the real data (a). For a generated image, all-in-one (b) assigns a new label to it. One-hot pseudo (c) uses only one predefined training class with maximum predicted probability. LSRO (d) uses a single unified label distribution, while the proposed MpRL (e) considers different weights for different predefined training classes.

by using the all-in-one (see Fig. 3.2(b)). Using the one-hot pseudo, only the maximum predicted probability of predefined training classes is used as a virtual label (see Fig. 3.2(c)). A single unified label distribution $1/K$ is utilized by LSRO (see Fig. 3.2(d)). Our MpRL is illustrated in Fig. 3.2(e). The $\alpha = \{\alpha_k | k \in [1, ..., K]\}$ (defined by Eq. 3.6 to Eq. 3.9) is used to record different weights for different predefined training classes. In this section, the differences between MpRL and the other three virtual labels will be discussed in three aspects: 1) one-hot binary vs. multi-association-value label, 2) single unified label vs. multiple different virtual labels,

and 3) different weighing on virtual labels. Quantitative experiments with detailed performance analysis are provided to support the claim on the benefits of MpRL.

3.3.1 One-Hot Binary Label vs. Multi-Association-Value Label

The all-in-one and pseudo are two standard one-hot binary labels. All-in-one uses a new class not the same as any predefined class as the virtual label assigned to generated data. Pseudo label assigns different virtual label to each generated data by taking the maximum value of the probability prediction on all predefined class in training data. Compared with multi-association-value label that retains information from all predefined training classes, the one-hot binary label may produce inadequate regularization power in training which is critical to prevent over-fitting. In the one-hot binary label, the network may incline to learn a discriminative feature based on certain class of training data which may not have sufficient training samples. While using multi-association-value label, the network better prevent over-fitting (Szegedy et al., 2016; Zhedong et al., 2017). MpRL is is a kind of multiple-association value label. In Sec. 3.4.6, corresponding experiments demonstrate the benefits of multi-association-value label.

3.3.2 Single Unified Label vs. Multiple Different Virtual Labels

Two strategies can be used to assign virtual labels to generated data: 1) using the same virtual label for all generated data, 2) assigning different virtual labels to different generated data. Both all-in-one and LSRO follow the first strategy, while one-hot pseudo and MpRL go with the second one. Compared with the second strategy, assigning each generated data with the same label potentially leads to ambiguous predictions in training. In Fig. 3.3, four different generated images with distinct visual differences (in red, yellow, white and green clothes) are given to find their top ten nearest representations which represent different predefined training classes in the real data domain. The four groups visually show clear differences.

Figure 3.3 : Examples of generated data and their corresponding representations in the real data domain. The left side shows four generated data with distinct visual differences (in red, yellow, white and green clothes). For each generated data, the right side gives ten nearest representations which represent each predefined training class in the real data domain.

If the same virtual label is assigned to these four generated images, it will cause confusion while training network. The proposed MpRL follows the second strategy that assigns each generated data with a weight-based virtual label according to different predicted probabilities. Corresponding experiments can be found in Sec. 3.4.6 to show that by assigning different virtual labels to generated data, the proposed MpRL can achieve a better performance.

3.3.3 Different Weighing on Virtual Labels

LSRO assumes that the weight from each predefined training class is identical. Thus a generated image is assumed to have the capability to simulate all predefined training classes equally. This is impractical when considering the actual GAN training process for two reasons (details can be found in (Goodfellow et al., 2014; Radford et al., 2015)). First, in each training iteration, a minibatch of random noise

Table 3.1 : Comparison between virtual labels.

Method	Label type	Label Assigning	Weights on Pre-defined Classes
All-in-one	One-hot binary	Single unified	–
Pseudo	One-hot binary	Multiple different	–
LSRO	Multi-association-value	Single unified	Same
Our MpRL	Multi-association-value	Multiple different	Different

is fed into a generator to simulate another minibatch of real data. This indicates that the generation capability of the inputs is limited in a small scope, specifically, within a minibatch of real data. Second, the scale and the quality of data in different classes of training data are different so their contributions in training generator are different. Thus, such differences need to be recognised in the generated data. To address the problem of LSRO, In this thesis, such differences are embedded in the virtual labels of corresponding generated data by adopting proper weighing. That is, the proposed MpRL uses different weights from predefined training classes (see Sec. 3.2.2). In experiments, this thesis observes that the proposed MpRL can outperform the state-of-the-art LSRO method on three large and two small-scale person re-ID datasets (see Sec. 3.4.6).

Through the above discussion, Tab. 3.1 summaries the properties between the proposed MpRL and other labels. The proposed MpRL takes the advantages of all properties and achieves a better performance than others. The numerical evidences which show the superiority of MpRL are presented in Sec. 3.4.

3.4 Experiments

In this section five person re-ID datasets are used to verify the effectiveness of the proposed MpRL, including three large-scale datasets (Market-1501 (Zheng et al.,

2015), DukeMTMC-reID (Zhedong et al., 2017), and CUHK03 (Li et al., 2014)) and two small-scale datasets (VIPeR (Gray and Tao, 2008) and CUHK01 (Li et al., 2012)). The proposed MpRL is mainly evaluated using Market-1501 and VIPeR since they have different scales of data volumes.

3.4.1 Person Re-ID Datasets

Market-1501 is collected from six cameras in Tsinghua University. It contains 12,936 training images and 19,732 testing images. The number of identities is 751 and 750 in the training and testing sets respectively. There is an average of 17.2 images per training identity. All pedestrians are detected by the Deformable Part Model (DPM) (Felzenszwalb et al., 2010). Both single and multiple query settings are used.

DukeMTMC-reID is collected from eight cameras. The original dataset is used for cross-camera multi-target pedestrian tracking (Ristani et al., 2016). The re-ID version benchmark (Zhedong et al., 2017) is used for evaluation. It contains 1,404 identities in which 702 identities for training and the remaining 702 identities for testing. The total training images are 16,522. In the testing set, one query image for each identity is picked up in each camera and put the remaining images in the gallery. There are 2,228 query images and 17,661 gallery images for the 702 testing identities.

CUHK03 is captured by six cameras on the CUHK campus. It contains 14,097 images of 1,467 identities, and each identity is observed by two disjoint camera views. There is an average of 9.6 training identity images in this set. CUHK03 contains two image settings: one is annotated by hand-drawn bounding boxes, and the other is produced by the DPM (Felzenszwalb et al., 2010). The detected bounding boxes and the single query setting are used in experiments.

VIPeR is a small-scale dataset that only contains 632 identities. Each identity

has two images which are observed by two different camera views. There are 1,264 images in which half identities are for training and the remaining is for testing.

CUHK01 has 971 identities, each with four images captured from two disjoint camera views. There are totally 3884 images. Two different settings can be found on this dataset: 1) 871 identities for training, and 2) 485 identities for training. The latter one is chosen to verify the effectiveness of the proposed approach since the scale of training data is much more limited than the former one. The multiple query setting is used in testing.

3.4.2 Experimental Setup

GAN Models for Generating Data

GAN simultaneously trains two models: a generator that simulates the distribution of real data, and a discriminator that estimates the probability that a image comes from the real data set rather than the generator (Goodfellow et al., 2014). The DCGAN model (Radford et al., 2015) is used that follows the same settings in (Zhedong et al., 2017) for fair experimental comparisons. For the generator, 100-dim random noise is fed into a linear function to produce a tensor with size of $4 \times 4 \times 16$. Then, five deconvolutional functions with a kernel size of 5×5 and a stride of 2 are used to enlarge the tensor. A rectified linear unit and batch normalization are used after each deconvolution. Also, one deconvolutional layer with a kernel size of 5×5 and a stride of 1 are added to fine-tune the result followed by a tanh activation function. Finally, $128 \times 128 \times 3$ sized images can be generated. The input of the discriminator includes generated and real data. Five convolutional layers are used to classify whether the generated image is fake with a kernel size of 5×5 and a stride of 2. In the end, a fully-connected layer is added to perform a binary classification.

The Tensorflow (Abadi et al., 2016) and DCGAN packages are used to train

(a) Market-1501. (b) DukeMTMC. (c) CUHK03. (d) VIPeR. (e) CUHK01.

Figure 3.4 : Examples of generated (by DCGAN (Radford et al., 2015)) and real person images. (a)-(d) show the generated person images (first two rows) and real person images (the third row) on Market-1501, DukeMTMC-reID, CUHK03, VIPeR, and CUHK01, respectively. Note that all fake images do not belong to any of IDs in real data.

the GAN model. All images are resized to 128×128 and randomly flipped before training. The adam stochastic optimization (Kingma and Ba, 2014) is used with parameters $\beta 1 = 0.5, \beta 2 = 0.99$. The training stops after 30 and 60 epochs on large and small-scale re-ID datasets respectively. During testing, a 100-dim random vector ranged in [-1, 1] with Gaussian distribution is fed into the GAN to generate a person image. Finally, all generated data are resized to 256×256 and will be used to train CNN models with the proposed MpRL.

Fig. 3.4 illustrates the generated and real data on the five different re-ID datasets. For each dataset, it can observe that all generated images have similar visual style comparing with real data. Although the generated data can be easily recognized as fake by human, they remain effective in improving the performance by adding the proposed MpRL as virtual labels in experiments. Note that these generated images are employed to improve the performance of CNN models by its regularization power instead of providing more actual subjects beyond the scope of the raw dataset only

with real images in training. The DCGAN we used for the data generation task can only ensure the overall distribution of fake data being close to the real data. Therefore, each generated image does not belong to any of IDs in the real data. The proposed MpRL virtual label will be assigned to these generated data during training (refer to Sec. 3.2).

Performance Evaluation on Person Re-ID

Two CNNs are adopted to evaluate the proposed MpRL. These two networks have been used to evaluate the performance of the all-in-one, one-hot pseudo, and LSRO labels in (Zhedong et al., 2017). The first is an Identif network (Zheng et al., 2016a,b) that takes person re-ID as a multi-classification task according to the number of predefined training classes in the real data domain. The Identif network is used as a baseline when only the real data is used. Furthermore, to compare the performance of different virtual labels, generated images are incorporated into real images as inputs. The second one is a Two-stream network (Zheng et al., 2018) that combines the Identif network with a verification function to train the network. Given two input images, the verification function will classify them into two classes (belong to the same or different identities). This Two-stream network is used to achieve better results by adding generated data in training. In the experiment, both Identif and Two-stream networks use the pretrained resnet-50 (He et al., 2016) as the backbone. The last fully-connected layer is changed to have K neurons to predict K classes, where K is the number of predefined training classes. Since there is no need to add any extra class for generated data by using the proposed MpRL, the last fully-connected layer remains K neurons in training.

Fig. 3.5(a) and Fig. 3.5(b) respectively show the Identif and Two-stream networks. MpRLs are assigned to generated data when they are fed into the network. In the Two-stream network, squared Euclidean distance is used as a similarity mea-

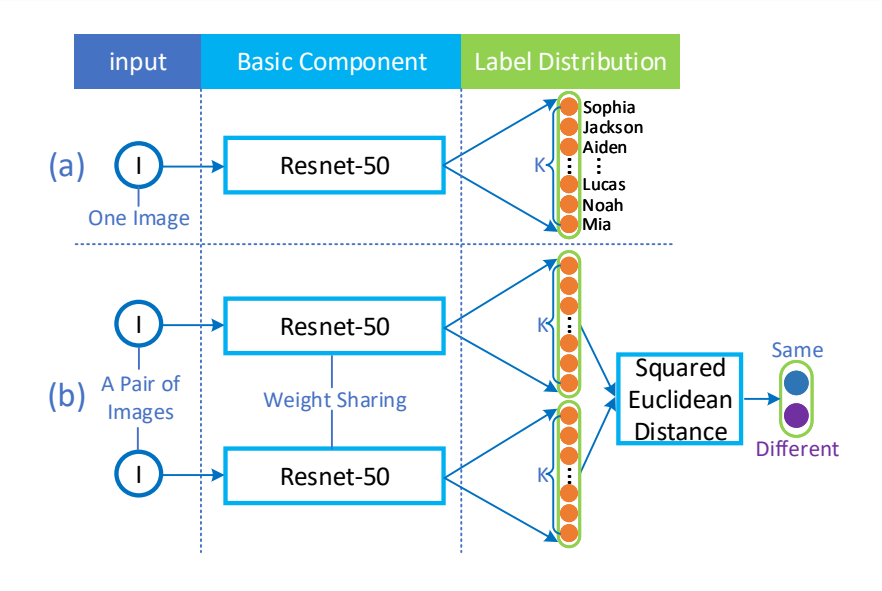


Figure 3.5 : (a) is the Identif network presented in (Zheng et al., 2016a,b), (b) is the Two-stream network introduced in (Zheng et al., 2018). Both networks use resnet-50 as a basic component of CNN.

sure between two outputs of the K neurons, and parameters are shared between the two resnet-50 components. Since generated images are unlabeled data that do not belong to any classes, only real images participate in the verification function.

The Matconvnet (Vedaldi and Lenc, 2015) package is used to implement the Identif network and the Two-stream network. All images are resized to 256×256 before being randomly cropped into 224×224 with random horizontal flipping. A dropout layer is inserted before the final convolutional layer of the resnet-50. The dropout rate is set to 0.75 for Market-1501 and DukeMTMC-reID, and 0.5 for CUHK03, VIPeR, and CUHK01. The fully-connected layer of resnet-50 is modified to have 751, 702, 1,367, 316 and 485 neurons for Market-1501, DukeMTMC-reID, CUHK03, VIPeR, and CUHK01 respectively. For the verification function in the Two-stream network, a dropout layer with a rate of 0.9 is adopted after the similarity measure. Stochastic gradient descent is used on both networks with momentum 0.9. The learning rate is set to 0.1 and decay to 0.01 after 40 epochs, and the

Table 3.2 : Performance of the Identif and Two-stream networks. Only the real images are used. Rank-1 accuracy and mAP are listed.

Dataset	CNN	mAP	rank-1
Market-1501	Identif (Zheng et al., 2016a,b)	52.68%	74.08%
	Two-stream (Zheng et al., 2018)	64.09%	81.83%
DukeMTMC-reID	Identif (Zheng et al., 2016a,b)	42.20%	61.94%
	Two-stream (Zheng et al., 2018)	51.04%	72.62%
CUHK03	Identif (Zheng et al., 2016a,b)	68.36%	63.10%
	Two-stream (Zheng et al., 2018)	85.20%	81.88%
VIPeR	Identif (Zheng et al., 2016a,b)	46.38%	40.76%
	Two-stream (Zheng et al., 2018)	59.38%	51.84%
CUHK01	Identif (Zheng et al., 2016a,b)	63.60%	65.33%
	Two-stream (Zheng et al., 2018)	76.38%	77.78%

training stops after the 50-th and 60-th epochs on the Identif network and Two-stream network, respectively. For the Identif network, the batchsize is set to 64. For the Two-stream network, the batchsize is set to 32 and 48 on large and small-scale re-ID datasets respectively. During testing, for both networks, a 2,048-dim CNN embedding in the last convolutional layer of the resnet-50 is extracted. The similarity between two images is calculated by a squared Euclidean distance before ranking. Naturally, the small-scale dataset cannot train a network from the scratch. In order to build certain initial network parameters, the three large scale re-ID datasets are used to pretrain two evaluation CNN models (*i.e.*, the Identif network and the Two-stream network). Then, small datasets VIPeR and CUHK01 along with the generated data are used to fine-tune the network.

3.4.3 Baseline Performance

Using the experimental setup in Sec. 3.4.2, the Identif and Two-stream networks are trained on Market-1501, DukeMTMC-reID, CUHK03, VIPeR and CUHK01, respectively. Tab. 3.2 shows the experimental results using the real data only. With the Identif (Two-stream) network, the rank-1 accuracy achieves 74.08% (81.83%), 61.94% (72.62%), 63.10% (81.88%), 40.76% (51.84%), and 65.33% (77.78%) on Market-1501, DukeMTMC-reID, CUHK03, VIPeR, and CUHK01, respectively. The result shown in Tab. 3.2 is a baseline, and the goal is to improve the performance of the two networks by using the proposed MpRL with generated data in training.

3.4.4 Performance Improved on The Identif Network by Using Generated Data

The result of the Identif network is given to evaluate the proposed MpRL. Tab. 3.3 shows that when 24,000 GAN generated images are added to train the Identif network on three large-scale datasets, the proposed dMpRL-II significantly improves the re-ID performance on the strong baseline of Market-1501. The improvements are +5.91% (from 52.68% to 58.59%) and +6.29% (from 74.08% to 80.37%) in mAP and rank-1 accuracy, respectively. For DukeMTMC-reID, +6.38% (from 42.20% to 48.58%) and +6.30% (from 61.94% to 68.24%) improvements are obtained in mAP and rank-1 accuracy, respectively. For CUHK03, the improvements are +5.12% (from 68.36% to 73.48%) and +5.58% (from 63.10% to 68.68%) in mAP and rank-1 accuracy, respectively. The effectiveness of the proposed method is also testified on two small-scale datasets, including VIPeR and CUHK01. +5.87% (mAP) and +5.84% (rank-1) improvements can be observed on VIPeR by adding 1,200 generated images in training. Meanwhile, +2.77% (mAP) and +3.48% (rank-1) improvements can be observed on CUHK01 by adding 4,000 generated images in training. The above results indicate the proposed MpRL can effectively yield

Table 3.3 : Comparison between LSRO and dMpRL-II on five datasets. Identif network is used by adding 24,000, 1,200, and 4,000 generated images on the three large re-ID datasets, VIPeR, and CUHK01, respectively. The improvements is shown in the *italic* and **bold** font by using LSRO and the proposed MpRL, respectively.

Dataset	Method	mAP	rank-1
Market-1501	baseline	52.68%	74.08%
	LSRO (Zhedong et al., 2017)	56.33%	78.21%
	Improvement	*+3.65%*	*+4.14%*
	dMpRL-II	58.59%	80.37%
	Improvement	**+5.91%**	**+6.29%**
DukeMTMC-reID	baseline	42.20%	61.94%
	LSRO (Zhedong et al., 2017)	46.66%	66.92%
	Improvement	*+4.46%*	*+4.98%*
	dMpRL-II	48.58%	68.24%
	Improvement	**+6.38%**	**+6.30%**
CUHK03	baseline	68.36%	63.10%
	LSRO (Zhedong et al., 2017)	71.60%	66.30%
	Improvement	*+3.24%*	*+3.20%*
	dMpRL-II	73.48%	68.68%
	Improvement	**+5.12%**	**+5.58%**
VIPeR	baseline	46.38%	40.76%
	LSRO (Zhedong et al., 2017)	49.94%	43.57%
	Improvement	*+3.56%*	*+2.81%*
	dMpRL-II	52.25%	46.60%
	Improvement	**+5.87%**	**+5.84%**
CUHK01	baseline	63.60%	65.33%
	LSRO (Zhedong et al., 2017)	64.47%	66.98%
	Improvement	*+0.87%*	*+1.65%*
	dMpRL-II	66.37%	68.81%
	Improvement	**+2.77%**	**+3.48%**

improvements over the baseline on both large and small-scale re-ID datasets.

3.4.5 Performance Evaluation under Different Implementations of MpRL

Three different implementations are used in experiments to demonstrate the effectiveness of the proposed MpRL (see Sec. 3.2.2). This experiment is conducted using the Identif network. Tab. 3.4 gives performance the comparison on Market-1501. It is observed that by dynamically updating the likelihood of the affiliation of the generated data to predefined training classes in training, dMpRL-I (+4.74% and +4.87% improvements in mAP and rank-1 accuracy respectively) and dMpRL-II (+5.91% and +6.29% improvements in mAP and rank-1 accuracy respectively) achieve better improvements compared with the sMpRL (+3.08% and +4.77% improvements in mAP and rank-1 accuracy respectively). This is because when discriminative ability of the network getting better and better during training, MpRL assigned to each generated data will be more and more robust. Also, compared with dMpRL-I, dMpRL-II achieves better improvement when the network becomes relatively stable after 20 training epochs.

3.4.6 Comparison with Existing Virtual Labeling Approaches

The proposed MpRL is compared with other three existing competitive virtual labels: all-in-one, one-hot pseudo, and LSRO. Amongst them, LSRO (Zhedong et al., 2017) is the state-of-the-art method using generated data for person re-ID. Tab. 3.4 provides the comparison results. Different number of generated data is used in training to show the different performance. By adding 30,000 and 18,000 generated images, the all-in-one achieves the best improvements in mAP (+3.51%) and rank-1 accuracy (+3.32%), respectively. The one-hot pseudo achieves +4.22% (mAP) and +3.87% (rank-1) improvements when 24,000 and 30,000 generated images are respectively added. Compared with them, LSRO obtains a better rank-1 accuracy improvement (+4.13%) when adding 24,000 generated images. However, the im-

Table 3.4 : Comparison of all-in-one, pseudo, LSRO, and MpRLs under different numbers of generated data on Market-1501 by using the Identif network. The best improvement of different methods is highlighted in **bold**. Rank-1 accuracy and mAP are shown.

#GAN Img	All-in-one		Pseudo		LSRO	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
0 (base)	52.68%	74.08%	52.68%	74.08%	52.68%	74.08%
12000	55.68%	76.96%	55.69%	76.52%	55.22%	77.17%
18000	55.59%	**77.40%**	56.04%	77.95%	55.28%	76.96%
24000	56.07%	77.21%	**56.90%**	77.62%	**56.33%**	**78.21%**
30000	**56.19%**	77.17%	56.54%	**77.95%**	55.40%	77.46%
36000	55.24%	75.92%	56.38%	77.42%	55.82%	77.91%
48000	53.98%	75.16%	55.86%	76.72%	54.87%	76.90%
Improvement	+3.51%	+3.32%	+4.22%	+3.87%	+3.65%	+4.13%
#GAN Img	sMpRL		dMpRL-I		dMpRL-II	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
0 (base)	52.68%	74.08%	52.68%	74.08%	52.68%	74.08%
12000	55.27%	77.73%	55.84%	77.88%	58.14%	79.22%
18000	55.05%	77.73%	56.21%	78.36%	58.31%	79.81%
24000	55.59%	**78.85%**	56.10%	77.79%	**58.59%**	**80.37%**
30000	**55.76%**	77.82%	57.15%	78.65%	57.69%	79.16%
36000	55.45%	78.32%	**57.42%**	**78.95%**	57.61%	79.90%
48000	55.02%	77.45%	56.01%	77.57%	57.03%	78.73%
Improvement	+3.08%	+4.77%	+4.74%	+4.87%	**+5.91%**	**+6.29%**

provement of mAP (+3.65%) is slightly less than the one-hot pseudo. In this experiment, the same generated data are used for different methods; the improvements are similar to the result reported in (Zhedong et al., 2017). Although the improvement of mAP (+3.08%) is less than other virtual labels by using sMpRL, better rank-1 accuracy improvements are obtained under all implementations of the proposed MpRL (+4.77%, +4.87%, and +6.29%, respectively). dMpRL-I and dMpRL-II also outperform other methods in mAP by +4.74% and +5.91% respectively. By adding 24,000 generated images, dMpRL-II improves the mAP and rank-1 accuracy of the Identif network from 52.68% and 74.08% to 58.59% and 80.37%, respectively. The proposed method outperforms the previous state-of-the-art method LSRO to a certain degree (mAP: +3.65%→ +5.91%, rank-1 accuracy: +4.13%→ +6.29%). It can be observed that when 12,000 generated images are used, there is limited regularization capability to improve the re-ID performance for different virtual labels. Meanwhile, if too many generated images are added in training, *e.g.*, 48,000, the performance is dropped since the network tends to converge towards the generated data instead of real data. To balance the number of generated data in training, it is empirically set to 24,000 over the three large-scale datasets.

In Tab. 3.4, it is clear to see that the multi-association-value label (LSRO and MpRL) can always outperform the one-hot binary label (all-in-one and pseudo) in the rank-1 accuracy. The reason can be found in Sec. 3.3.1. Besides, it is also observed that compared with the way using the single unified label, assigning multiple different labels to generated data can achieve better results (*i.e.*, MpRL *vs.* LSRO and pseudo *vs.* all-in-one). The reason can be found in Sec. 3.3.2.

To further evaluate the performance of the proposed MpRL, it is also evaluated on two small-scale re-ID datasets. Tab. 3.5 lists the result on VIPeR. The proposed dMpRL-II improves the mAP and rank-1 accuracy on this dataset by +5.87% and +5.84% respectively when adding 1,200 generated images in training, and outper-

Table 3.5 : Comparison of LSRO and the proposed dMpRL-II under different numbers of generated data on VIPeR with the Identif network. The best improvement of different methods is highlighted in **bold**. Rank-1 accuracy and mAP are listed.

#GAN Img	LSRO (Zhedong et al., 2017)		dMpRL-II	
	mAP	rank-1	mAP	rank-1
0 (base)	46.38%	40.76%	46.38%	40.76%
600	48.98%	42.80%	48.59%	42.61%
1200	**49.94%**	**43.57%**	**52.25%**	**46.60%**
1800	49.41%	43.39%	50.51%	44.24%
2400	45.95%	40.65%	49.36%	43.77%
12000	43.34%	37.12%	44.25%	37.66%
Improvement	+3.56%	+2.81%	**+5.87%**	**+5.84%**

forms the LSRO method. Since VIPeR is a small dataset (only 632 images for training), adding too many generated images, *e.g.*, 12,000 leads to inferior results. Therefore, the number of generated data is set to approximate double than that of the number of real data on small datasets. Specifically, 1,200 and 4,000 generated images are used for VIPeR and CUHK01, respectively. The result on VIPeR is mainly reported by changing the number of generated data. The results of CUHK01 can be found in Tab. 3.3 and 3.7.

Using the Identif network, Tab. 3.3 shows comparison results between the proposed dMpRL-II and LSRO on three large-scale datasets by adding 24,000 generated images. Also, two small-scale datasets are used to evaluate the proposed method by adding 1,200 and 4,000 images respectively. By using different weights from predefined training classes, dMpRL-II can always outperform previous state-of-the-art virtual label LSRO over the five datasets. The reason can be found in Sec. 3.3.3.

(a) Market-1501. (b) DukeMTMC-reID. (c) VIPeR.

Figure 3.6 : Examples of generated and real person images. (a)-(c) show the generated person images (first two rows) and real person images (the third row) on Market-1501, DukeMTMC-reID, and VIPeR respectively. Images in the first and second rows are respectively generated by the WGAN-GP (Gulrajani et al., 2017) and the DCGAN (Radford et al., 2015).

3.4.7 Performance Evaluation by Using Different GAN Models

In addition to the DCGAN, other GAN models such as WGAN-GP (Gulrajani et al., 2017) has demonstrated its superior in generating high quality person images. Thus, data generated by WGAN-GP are also used in experiments. Then, the relationship between the quality of generated images and the proposed MpRL can be testified by using different GAN models. In this experiment, two large and one small-scale datasets are used individually to generate images. Fig. 3.6 shows the generated data by using different GAN models. It can be observed that the WGAN-GP exhibits better capability of generating person images on these datasets. In order to verify the impacts of image quality created by different GAN approaches, the proposed MpRL is evaluated on the two different generated data sets. Tab. 3.6 lists the comparison results. It is observed that by using generated data with different quality through different GAN approaches, the re-ID performance is not significantly affected. This is because these generated data are employed to improve the performance of CNN models by its regularization power instead of providing more actual

Table 3.6 : Comparison between using generated data by DCGAN and WGAN-GP. Two approaches are used, including LSRO and the proposed dMpRL-II. Experiments conducted on three datasets: Market-1501, DukeMTMC-reID, and VIPeR. Rank-1 accuracy and mAP are listed.

Method	Market-1501			
	DCGAN		WGAN-GP	
	mAP	rank-1	mAP	rank-1
LSRO (Zhedong et al., 2017)	56.33%	78.21%	55.53%	78.32%
dMpRL-II	58.59%	80.37%	59.04%	79.75%
	DukeMTMC-reID			
LSRO (Zhedong et al., 2017)	46.66%	66.92%	46.79%	66.97%
dMpRL-II	48.58%	68.24%	49.30%	68.76%
	VIPeR			
LSRO (Zhedong et al., 2017)	49.41%	43.39%	48.47%	43.14%
dMpRL-II	52.25%	46.60%	52.16%	46.39%

subjects beyond the scope of the raw dataset in training. Therefore, better generated data can bring superior perceptual quality but cannot dramatically boost the effectiveness of regularizer although some marginal improvements can be observed.

3.4.8 Performance Comparison with The State-of-The-Art Methods

Although the main contribution in this thesis focuses on using the generated data to improve the performance of CNNs, but not on producing a state-of-the-art result, a comparison is still conducted with several state-of-the-art methods. Tab. 3.7 lists the comparison results. It is clear to see that although the performance of the original Two-stream network is competitive, it still be inferior to many methods such as Resnet+OIM (Xiao et al., 2017), SSM (Bai et al., 2017), JLML (Li et al., 2017b), SVDNet (Sun et al., 2017), and PDC (Su et al., 2017). However, by incorporating

Table 3.7 : Performance comparison with the published state-of-the-art methods. The best and the second-best results are shown in **bold** and underline, respectively. Rank-1 accuracy and mAP are listed. The ReK means re-ranking.

Method		Large-Scale Datasets								Small-Scale Datasets			
		Market-1501				DukeMTMC-reID		CUHK03		VIPeR		CUHK01	
		Single Query		Multiple Query		Single Query		Single Query (detected)		Single Query		Multiple Query	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	rank-1		rank-1	
Gate-reID (Varior et al., 2016)	ECCV16	39.55%	65.88%	48.45%	76.04%	–	–	58.84%	68.10%	37.80%		–	
SI-CI (Wang et al., 2016a)	CVPR16	–	–	–	–	–	–	–	52.17%	35.76%		–	
GOG+XQDA (Matsukawa et al., 2016)	CVPR16	–	–	–	–	–	–	–	65.50%	49.70%		57.80%	
SCSP (Chen et al., 2016)	CVPR16	26.35%	51.90%	–	–	–	–	–	–	53.54%		–	
DNS (Zhang et al., 2016)	CVPR16	35.68%	61.02%	46.03%	71.56%	–	–	–	54.70%	51.17%		69.09%	
Resnet+OIM (Xiao et al., 2017)	CVPR17	–	82.10%	–	–	–	68.10%	–	–	–		–	
Latent Parts (Li et al., 2017a)	CVPR17	57.53%	80.31%	66.70%	86.79%	–	–	–	67.99%	38.08%		–	
P2S (Zhou et al., 2017)	CVPR17	44.27%	70.72%	55.73%	85.78%	–	–	–	–	–		–	
ReRank (Zhong et al., 2017)	CVPR17	63.63%	77.11%	–	–	–	–	–	–	–		–	
CADL (Lin et al., 2017)	CVPR17	55.60%	80.90%	–	–	–	–	–	–	–		–	
SpindleNet (Zhao et al., 2017a)	CVPR17	–	76.90%	–	–	–	–	–	–	53.80%		**79.90%**	
SSM (Bai et al., 2017)	CVPR17	68.80%	82.21%	76.18%	88.18%	–	–	–	72.70%	53.73%		–	
JLML (Li et al., 2017b)	IJCAI17	65.50%	85.10%	74.50%	89.70%	–	–	–	80.60%	50.20%		76.70%	
SVDNet (Sun et al., 2017)	ICCV17	62.10%	82.30%	–	–	56.80%	76.70%	84.80%	81.80%	–		–	
PDC (Su et al., 2017)	ICCV17	63.41%	84.14%	–	–	–	–	–	78.29%	51.27%		–	
Part Aligned (Zhao et al., 2017b)	ICCV17	63.40%	81.00%	–	–	–	–	–	81.60%	48.70%		75.00%	
LSRO (Zhedong et al., 2017)	ICCV17	66.07%	83.97%	76.10%	88.42%	47.13%	67.68%	87.40%	84.60%	–		–	
Identif (Zheng et al., 2016a,b)		52.68%	74.08%	64.95%	82.06%	42.20%	61.94%	68.36%	63.10%	40.76%		65.33%	
Identif+dMpRL-II		58.59%	80.37%	70.22%	86.47%	48.58%	68.24%	73.48%	68.68%	46.60%		68.81%	
Two-stream (Zheng et al., 2018)		64.09%	81.83%	73.65%	86.82%	51.40%	72.62%	85.20%	81.88%	51.84%		77.78%	
Two-stream+dMpRL-II		67.53%	85.75%	77.85%	89.88%	58.56%	76.81%	87.53%	85.42%	**54.65%**		78.83%	
Two-stream+dMpRL-II+ReK		**81.18%**	**87.96%**	**86.53%**	**90.97%**	**74.54%**	**81.28%**	**90.16%**	**88.00%**	53.22%		78.08%	

with the proposed dMpRL-II, the Two-stream network achieves the state of the art compared with other methods on Market-1501, DukeMTMC-reID, CUHK03, and VIPeR. To achieve a better performance, after obtaining the rank list by sorting the similarity of gallery images to a query, a re-ranking method (Zhong et al., 2017) is adopted to further boost the performance of the proposed method. The combination of the dMpRL-II and re-ranking on the Two-stream network achieves the best results on the three large-scale datasets. However, the re-ranking approach cannot further improve the performance of the two small-scale datasets with limited number of testing person identities. It is observed that the rank-1 accuracy of the DPFL method (Chen et al., 2017b) proposed in the ICCV17 workshop is slightly higher than result of the proposed method on Market-1501 (88.90% in single query and 92.30% in multiple query). However, DPFL uses an ensemble deep model with multiple granularity inputs for each image. The two-stream network used for evaluation just utilizes a single model and outperforms the DPFL on CUHK03 by a large margin in mAP even without re-ranking (mAP: 87.53% *vs.* 78.10% (DPFL), rank-1: 85.42% *vs.* 82.00% (DPFL)). Also, the performance of the Spindle (Zhao et al., 2017a) approach is slightly higher than the proposed method on CUHK01 (79.90% *vs.* 78.83%). Since VIPeR and CUHK01 are two small-scale datasets, nine different person re-ID datasets are used to pretrain the SpindleNet model and then fine-tuning on the two small datasets respectively. The fine-tuning strategy is also used on these two datasets, but only three datasets are involved in the pretraining stage (see 3.4.2). Except for the CUHK01 dataset, the performance of the proposed method outperforms the SpindleNet on the other small-scale dataset VIPeR and the three large-scale re-ID datasets simultaneously.

3.5 Conclusion

In this chapter, a new virtual label MpRL is proposed for the generated data by GAN. To train a CNN, MpRL is used as virtual label assigned to generated data. These data are used for semi-supervised learning. Two CNNs are adopted to show the effectiveness of the proposed MpRL. Experiments demonstrate that generated data can effectively improve the performance of the two CNNs trained with the proposed MpRL. Compared with the previous state-of-the-art method LSRO (Zhedong et al., 2017), MpRL can always achieve better improvements. In the future, considering the capability of GAN, the proposed virtual label will attempt to be applied on unlabelled generated data for semi-supervised learning to other research fields.

Chapter 4

Background Shift Issue in Cross-Domain ST-ReID

4.1 Motivation

Cross-domain ST-reID is an active research problem since the annotation of all images in the target domain is expensive and infeasible (Yang et al., 2019b; Wu et al., 2019a; Zhong et al., 2019; Song et al., 2019; Qi et al., 2019; Liu et al., 2019; Li et al., 2019a; Chen et al., 2019c). Compared with conventional fully supervised training re-ID task (Zheng et al., 2019, 2017b; Huang et al., 2018a, 2017b; Sun et al., 2019; Yu et al., 2019b; Guo et al., 2019; Dai et al., 2019; Luo et al., 2019; Chen et al., 2019a,b; Zhou et al., 2019), cross-domain ST-reID is more challenging due to the large shift between training (*i.e.*, source domain) and testing domains (*i.e.*, target domain) (Huang et al., 2019b). For instance, person images captured from two different campuses have distinct illumination conditions and backgrounds (*e.g.*, Market-1501 (Zheng et al., 2015) and DukeMTMC-reID (Ristani et al., 2016; Zheng et al., 2017b) datasets). Therefore, the shift between two domains becomes large. Directly training a classifier from one dataset often produces a degraded performance when testing is conducted on another dataset. Thus, it is important to investigate solutions for such a cross-domain ST-reID issue.

Recent cross-domain ST-reID methods adopt (or resort to) *Generative Adversarial Network* (GAN) to learn the domain variations through data transformation (Bak et al., 2018; Deng et al., 2018; Wei et al., 2018a; Zhong et al., 2018b; Qi et al., 2019; Liu et al., 2019; Li et al., 2019a; Chen et al., 2019c). These methods may perform well in certain cases (*i.e.*, domain/camera style change) by transferring person im-

ages from the source domain and keeping their identities while presenting similar styles (*e.g.*, backgrounds, lightings, *etc.*) with person images in the target domain. However, these methods do not consider another factor which is a significant contributor causing differences between domains. That is *background* information. For instance, when a network is trained based on limited background information presented in a source domain, such a network may not well distinguish essential pedestrian features against noise caused by background variations in a target domain. Unfortunately, backgrounds in the target domain are normally very different from the source domain. This chapter formulates this issue as a background shift issue that may significantly degrade the overall performance of cross-domain ST-reID.

One possible solution to sort out the background shift issue is to remove backgrounds using foreground masks in a hard manner directly (*i.e.*, applying the binary masks on original images) (Farenzena et al., 2010; Huang et al., 2016; Song et al., 2018; Tian et al., 2018). However, it is observed that methods, such as JPP-Net (Liang et al., 2018) and Mask-RCNN (Abdulla, 2017; He et al., 2017), specifically being designed for removing the background, may damage the foreground information too. By simply removing backgrounds, this hard manner solution does improve the performance of cross-domain ST-reID to a certain extent. At the same time, it can be seen that this is still an open problem. "*Is there a way to better suppress background shifts to improve cross-domain ST-reID performance*?" This chapter makes the first effort to generate images where backgrounds are mitigated moderately instead of being completed removed in a hard manner.

To address the problem above, a *Suppression of Background Shift Generative Adversarial Network* (SBSGAN) is proposed. Compared with hard-mask solutions, images generated by the proposed SBSGAN can be regarded as foreground images, where the background is suppressed moderately. The generated images by SBSGAN can also be called *soft-mask images*. In addition, previous works (Deng et al.,

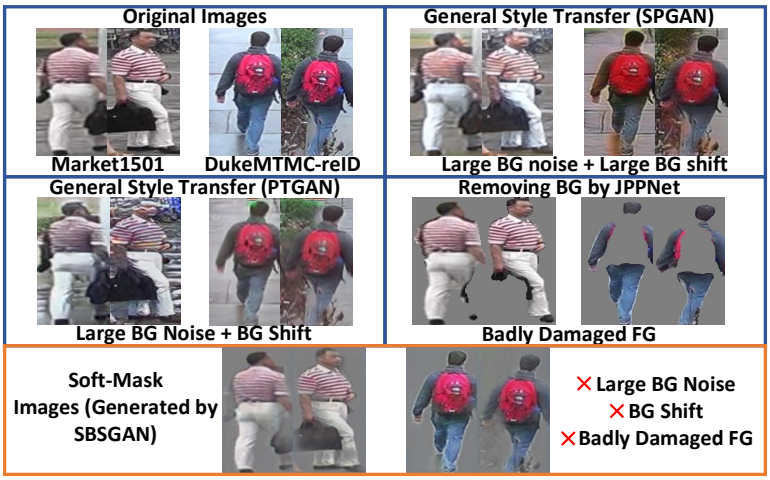


Figure 4.1 : Comparison between different input images for cross-domain ST-reID. Images from Market-1501 and DukeMTMC-reID show distinct background shift. Images generated by SPGAN (Deng et al., 2018) and PTGAN (Wei et al., 2018a) do not suppress the background noise and have the background shift problem. The hard-mask solution, *i.e.*, JPPNet (Liang et al., 2018) damages the foreground. The proposed SBSGAN considers all the impact.

2018; Wei et al., 2018a) show that keeping certain style consistency between data of different domains (*i.e.*, domain style transfer) can improve the performance of cross-domain ST-reID. Such an idea is integrated into SBSGAN to reduce the domain shift further. Fig. 4.1 shows images selected from two different person re-ID datasets. The backgrounds are very different. A model trained on one dataset may easily be biased on another one due to the background shift problem mentioned above. Images generated by recent cross-domain ST-reID approaches, such as SPGAN (Deng et al., 2018) and PTGAN (Wei et al., 2018a), still present some undesirable results. If foreground masks obtained by JPPNet (Liang et al., 2018) is directly used to zero out backgrounds, the foreground can be badly damaged by the noisy masks. On the contrary, every pixel in the generated images is preserved in a soft manner. Fig. 4.1

shows that SBSGAN can generate better images and further reduce the domain differences caused by background shifts.

In order to well utilize the generated foreground information and integrate useful ID-related cues from context into the network, a *Densely Associated 2-Stream* (DA-2S) network is proposed. This work argues that certain context information, *e.g.*, companions and vehicles in the background, can provide ID-related cues. Both images with suppressed backgrounds (generated by the proposed SBSGAN) and the original images with full backgrounds are respectively fed into two individual streams of DA-2S. Unlike previous 2-stream methods (*e.g.*, (Ahmed et al., 2015; Chen et al., 2018; Zheng et al., 2018)), this chapter proposes *Inter-Stream Densely Connection* (ISDC) modules as new components between two streams of DA-2S. With ISDCs, the relationship between two streams can be enhanced by signals coming from two streams in training.

Although SBSGAN is able to mitigate the background shift and thus to reduce the shift between data in source and target domains, so far, the proposed DA-2S is not able to fully explore the data in a target domain for training because this target domain does not has label information for each image in the training set. To learn more discriminative ID features, this chapter makes use of data in a target domain as training data with assigned virtual labels to further update DA-2S. Inspired by recently published cross-domain ST-reID works (Fan et al., 2018; Song et al., 2020; Fu et al., 2019; Zhang et al., 2019b), unsupervised clustering methods (*i.e.*, DBSCAN (Ester et al., 1996)) are used to assign virtual labels to the unlabelled target domain training images. Specifically, the proposed DA-2S network is first pretrained with labelled images from a source domain. However, unlike (Fan et al., 2018; Song et al., 2020; Fu et al., 2019; Zhang et al., 2019b) which use raw data from the source domain to pretrain a re-ID model, the proposed method uses background suppressed images generated by the proposed SBSGAN, which have less

domain shift. Then, such a pretrained DA-2S network is used to extract features of each training image in the target domain. DBSCAN is used to classify these unlabelled images into different clusters. Each unlabelled training image in the target domain is assigned a virtual ID label according to the corresponding cluster. It is clear that the effectiveness of virtual label estimation highly depends on the quality of the clustering result. Therefore, a *Dynamic Clustering Confidence Value* (DCCV) is proposed for each image when it is selected to update the DA-2S network. Specifically, given a cluster j, the mean value of features of every candidate image in j is used as the density center of j. Then, all candidate images in j are used to calculate the distance to the density center. The average distance is employed as DCCV to measure the reliability of images in the corresponding cluster. All images with DCCV$<$0 (close to the density center) are selected for DA-2S update; the others are discarded. Once images in the target domain training set have proper label information, they can be utilized to further update DA-2S previously trained by source domain data only. During the update, the natural characteristics of target domain data are explicitly considered into re-ID model (*i.e.*, DA-2S) training. Thus, performance can be further improved.

The contributions of this chapter are:

1) Background shift is comprehensively investigated to analyze its impact on cross-domain ST-reID. An SBSGAN is proposed to make the first effort by generating soft-mask images in order to reduce the domain shift. In addition, compared with previous methods, backgrounds are mitigated rather than completely removed in the generated images.

2) A DA-2S CNN network with the proposed ISDC components is presented to facilitate complementary information between the generated data and the ID-related cues from the background.

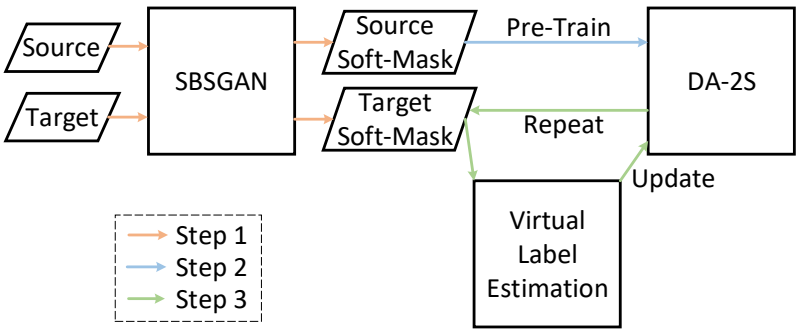


Figure 4.2 : Pipeline of the proposed approaches.

3) Target domain training images can further update the pretrained DA-2S network with estimated virtual labels produced by the DBSCAN clustering method. DCCV is proposed to improve virtual label estimation quality.

4.2 Overview of The Proposed Approach

The proposed method can be divided into three steps. The first step is soft-mask image generation to mitigate the domain shift by the proposed SBSGAN. Then, in the second step, the generated source domain images with ground-truth ID labels are used to pretrained the DA-2S re-ID model. In the third step, the pretrained (or updated) DA-2S is utilized to extract features of each image in target domain training set. The DBSCAN clustering method is adopted to classify these features into different clusters. Virtual labels are assigned to images according to the corresponding cluster. Finally, the images with virtual labels are used to update DA-2S. The step three repeats several times until DA-2S can converge to an approximate optimal solution. Fig. 4.2 illustrates the pipeline of the proposed approach.

The details of each step are introduced from Sec. 4.3 to Sec. 4.4, respectively. Specifically, the SBSGAN for soft-mask image generation is introduced in Sec. 4.3 (**step 1**). The details of the DA-2S network and the pretraining DA-2S model using labelled source domain data are given in Sec. 4.4.1 (**step 2**). Virtual label estimation

for the target domain training data and DA-2S update are presented in Sec. 4.4.2 (**step 3**).

4.3 SBSGAN for Generating Soft-Mask Image

The generator (G) of SBSGAN is adopted based on (Zhu et al., 2017). Given an input image, two down-sampling convolutional layers are used followed by six residual blocks (He et al., 2016) in G. Unlike (Zhu et al., 2017), two branches (without parameters sharing) are respectively used to generate soft-mask and auxiliary style-transferred images followed by the output of the last residual block. Each branch contains two up-sampling transposed convolutional layers with a stride of 2. For the discriminator (D), the PatchGAN (Isola et al., 2017; Zhu et al., 2017) structure is used without any change.

There are two tasks in G. The main task is to generate soft-mask images with suppressed backgrounds. The auxiliary task is to generate inter-domain style-transferred images (retain background) to normalize the style of soft-mask images across domains. D is used to distinguish the real against fake images and classify these images to their corresponding domains. Fig. 4.3 shows the proposed SBSGAN.

Specifically, given an input image (*e.g.*, $I_{\mathbb{D}_s}$) from a source domain $\mathbb{D}_s$, G can generate its corresponding soft-mask image $I_{\bar{\mathbb{D}}}$ by $G(I_{\mathbb{D}_s}, \bar{\mathbb{D}}) \rightarrow I_{\bar{\mathbb{D}}}$. G takes both image (*e.g.*, $I_{\mathbb{D}_s}$) and indicator (*e.g.*, $\bar{\mathbb{D}}$, refer to Sec. 4.3.2) as inputs. In addition, G can also transfer the style of $I_{\mathbb{D}_s}$ to a target domain $\mathbb{D}_k$ ($k \neq s$) via $G(I_{\mathbb{D}_s}, \mathbb{D}_k) \rightarrow I_{\mathbb{D}_k}$. The proposed SBSGAN is able to support multi-domain data as inputs. If there are K domains in training, then all $I_{\mathbb{D}_k} (k \in [1, K] \cap k \neq s)$ and the input image $I_{\mathbb{D}_s}$ are used to normalize the style of $I_{\bar{\mathbb{D}}}$, ensuring the style of $I_{\bar{\mathbb{D}}}$ is consistent across all the K domains.

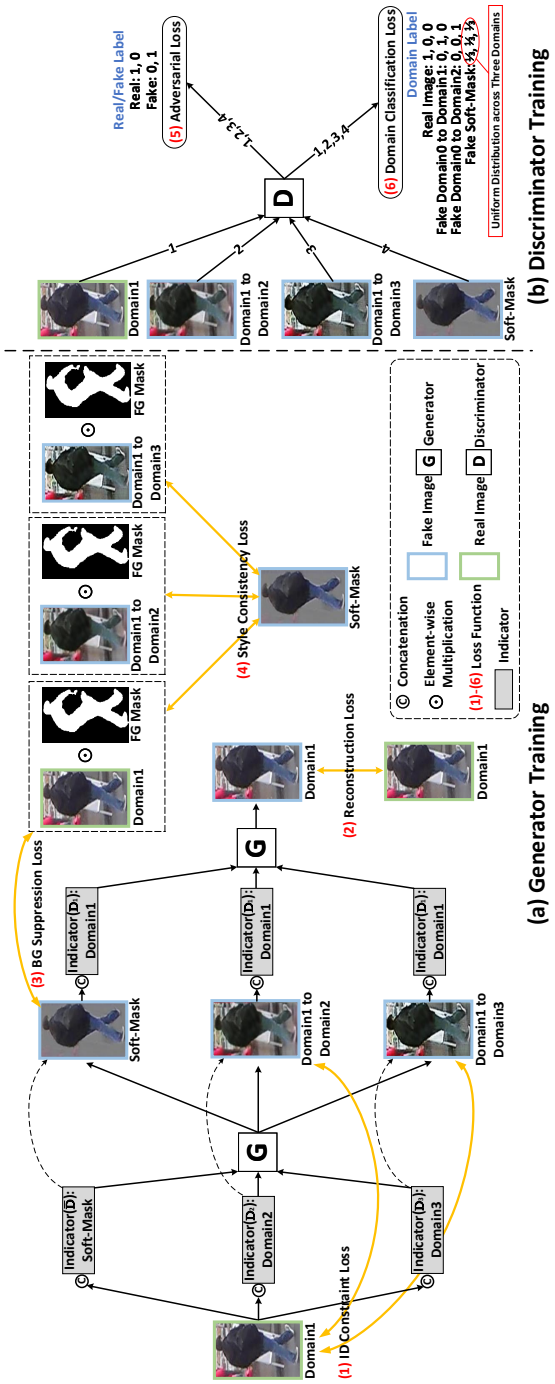


Figure 4.3 : Overview of SBSGAN. Three domains are used as an example. (a) shows the training process of the generator G. Given an input image from Domain1, G can generate the corresponding soft-mask image and transfer the input image to different domain styles (*e.g.*, Domain1 to Domain2) according to the indicators. The foreground mask is obtained by JPPNet. (b) All real and fake images are used to minimize the adversarial loss and the domain classification loss in D.

4.3.1 Objective Functions in SBSGAN

In order to achieve the function above, several loss functions are adopted to train SBSGAN. There are four important losses in terms of the data generation, which control four different things: **color**, **content**, **background**, and **style consistency**.

(1) ID Constraint Loss (color). Without any constraint, G may change the color of generated style-transferred images when the style-transferred images are directly applied to normalize the style of soft-mask images across multiple domains (see Eq. 4.4). Therefore, in order to well preserve the color information of all generated style-transferred images, the *ID Constraint* (IDC) loss (Taigman et al., 2017) is adopted to preserve the underlying color information of image for the auxiliary style-transferred image generation. The IDC loss is defined as follows:

$$\mathcal{L}_{idc} = \mathbb{E}_{I_{\mathbb{D}_s}, \mathbb{D}_k} \left[\| G(I_{\mathbb{D}_s}, \mathbb{D}_k) - I_{\mathbb{D}_s} \|_1 \right]. \tag{4.1}$$

IDC enforces the similarity between the generated image $G(I_{\mathbb{D}_s}, \mathbb{D}_k)$ and the source domain image $I_{\mathbb{D}_s}$ via L_1 norm constraint.

(2) Reconstruction Loss (content). G can generate an input image to different data according to the indicator it receives. However, there is no strong pixel-level supervision to ensure this generation process sufficiently reliable. Therefore, a *RE-Construction* (REC) loss is required to ensure the content consistency between an input image and the corresponding generated image. REC is a conventional objective function for the domain-to-domain image style transfer when strong pixel-level supervision is unavailable (Choi et al., 2018; Deng et al., 2018; Wei et al., 2018a; Zhu et al., 2017). The REC loss is given as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{I_{\mathbb{D}_s}, \mathbb{D}_k \vee \bar{\mathbb{D}}} \left[\left\| G(G(I_{\mathbb{D}_s}, \mathbb{D}_k \vee \bar{\mathbb{D}}), \mathbb{D}_s) - I_{\mathbb{D}_s} \right\|_1 \right], \tag{4.2}$$

where $\vee$ is 'or' operator. By adopting REC, when soft-mask image (or style-transferred image) is generated, it enforces the image content of foreground (or

foreground+background) to be consistent with the corresponding real image. Only the domain-related information in images is expected to be changed by the G.

(3) Background Suppression Loss (background). In order to generate soft-mask images, a *BackGround Suppression* (BGS) loss is proposed to suppress background in data generation. The BGS loss is formulated as follows:

$$\mathcal{L}_{bgs} = \mathbb{E}_{I_{\mathbb{D}_s},\bar{\mathbb{D}}}\left[\left\|I_{\mathbb{D}_s} \odot M(I_{\mathbb{D}_s}) - G(I_{\mathbb{D}_s},\bar{\mathbb{D}})\right\|_2\right]. \tag{4.3}$$

An auxiliary foreground mask $M(I_{\mathbb{D}_s})$ produced by JPPNet (Liang et al., 2018) is used to suppress the background of the input image $I_{\mathbb{D}_s}$. L_2 distance is applied to minimize the loss. The mask produced by JPPNet contains certain segmentation errors. The generated soft-mask image by the proposed SBSGAN is able to tolerate (or refine) such segmentation errors during the data generation process.

(4) Style Consistency Loss (style consistency). With suppressed background, the domain shift can be reduced in the generated soft-mask images. In order to further reduce the domain shift between target and source domain data, a *Style Consistency* (SC) Loss is proposed to encourage the style of soft-mask images to be consistent across all the input domains. The SC loss is given as follows:

$$\begin{aligned}\mathcal{L}_{sc} = \mathbb{E}_{I_{\mathbb{D}_s},\bar{\mathbb{D}},\mathbb{D}_k}[&\left\|G(I_{\mathbb{D}_s},\bar{\mathbb{D}}) - I_{\mathbb{D}_s} \odot M(I_{\mathbb{D}_s})\right\|_1 + \\ &\sum_{k=1,k\neq s}^{K}\left\|G(I_{\mathbb{D}_s},\bar{\mathbb{D}}) - G(I_{\mathbb{D}_s},\mathbb{D}_k) \odot M(I_{\mathbb{D}_s})\right\|_1].\end{aligned} \tag{4.4}$$

As shown in Fig. 4.3 (a), all style-transferred images (*i.e.*, Domain1 to Domain2 and 3) and the original input image are used to encourage the style of $G(I_{\mathbb{D}_s},\bar{\mathbb{D}})$ to be consistent across all three domains. Since the information of soft-mask images concentrates on the foreground, a foreground mask is imposed on $I_{\mathbb{D}_s}$, and all style-transferred images $G(I_{\mathbb{D}_s},\mathbb{D}_k)$ in Eq. 4.4 are used to normalize the foreground information of soft-mask image across all input domains.

(5) Other Common Loss Functions. Besides the above-mentioned loss functions, the conventional adversarial loss ($\mathcal{L}_{adv}$) (Goodfellow et al., 2014) is added to distinguish real and fake images in training. In addition, domain classification loss $\mathcal{L}_{cls}^{r}$ (Choi et al., 2018) is employed to classify the source domains of real images for optimizing D, and $\mathcal{L}_{cls}^{f}$ (Choi et al., 2018) is used to classify the target domains of fake images for optimizing G. Since the style of $I_{\bar{\mathbb{D}}}$ is normalized across all the K domains, a uniform distribution (*i.e.*, $\frac{1}{K}$) over the K domains is used as domain label of $I_{\bar{\mathbb{D}}}$ (refer to Fig. 4.3 (b)).

The overall objective functions of G and D are given as follows:

$$\mathcal{L}_D = \mathcal{L}_{adv} + \mathcal{L}_{cls}^{r}, \tag{4.5}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \mathcal{L}_{cls}^{f} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{idc}\mathcal{L}_{idc} + \lambda_{bgs}\mathcal{L}_{bgs} + \lambda_{sc}\mathcal{L}_{sc}, \tag{4.6}$$

where λ is hyper-parameter to control the contribution weights of different loss functions. This chapter empirically sets $\lambda_{rec} = 10$ and $\lambda_{idc} = \lambda_{bgs} = \lambda_{sc} = 5$ in experiments.

4.3.2 Indicators for Data Generation

The proposed SBSGAN supports multi-domain images as inputs. In experiments, images from three domains (datasets) are used in training. When images are fed into G, an indicator is concatenated after each image on the dimension of channel to let G knows what types of image should be generated. A 3D tensor $\mathbb{D}$ is used as the indicator. The height and width of $\mathbb{D}$ equal to the input image. There are K channels in $\mathbb{D}$. For the auxiliary style-transferred image generation, $\mathbb{D}$ is denoted as $\mathbb{D}_k$; all values in the k-th channel of $\mathbb{D}_k$ are set to 1, and other values in the remaining $K-1$ channels are set to 0. For the soft-mask image generation, $\mathbb{D}$ is denoted as $\bar{\mathbb{D}}$; all values of $\bar{\mathbb{D}}$ are set to $\frac{1}{K}$.

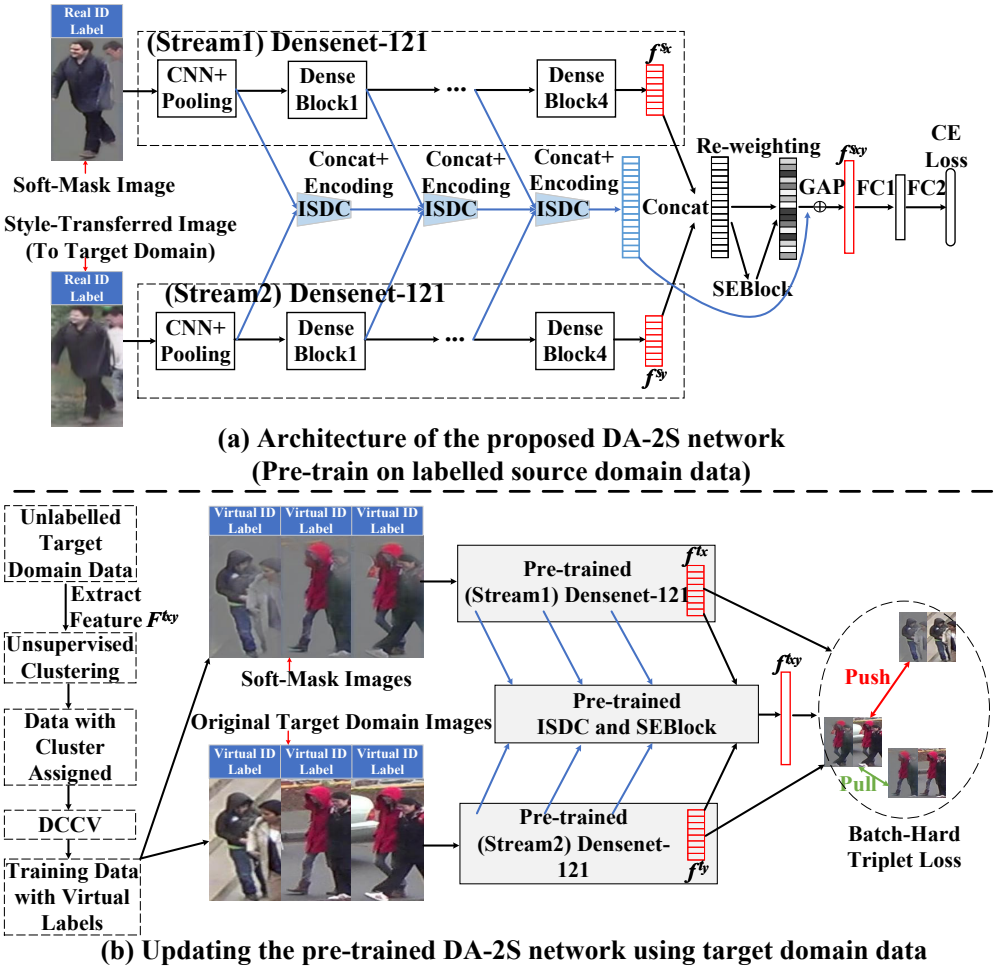


Figure 4.4 : Overview of DA-2S. In (a), ISDC, GAP, FC, and CE respectively represent Inter-Stream Densely Connection, Global Average Pooling, Fully-Connected layer, and Cross-Entropy loss. $\oplus$ represents element-wise summation.

4.4 DA-2S

4.4.1 Initial DA-2S Training

One of the main contributions of this chapter is to deal with the cross-domain ST-reID task based on the proposed inter-domain background shift suppression. Moreover, in order to explore helpful ID-related background cues (*i.e.*, context information), a DA-2S network is proposed. Besides foreground, the context informa-

tion in the background, *e.g.*, companions and vehicles, is also useful in cross-domain ST-reID. The proposed DA-2S network is to enrich person representations by learning features from both foreground and background. Fig. 4.4 (a) shows the DA-2S network. A pair of input images from the source domain (a soft-mask image and its style-transferred image to the target domain) is fed into two ImageNet-trained Densenet-121 (Huang et al., 2017a) networks (without parameters sharing). It can be observed that the companion in white clothes is suppressed as background in the soft-mask image. To use the companion as an ID-related cue presented in background, a style-transferred images (without background suppression) is fed into the second stream. In order to learn the complementarity between two inputs, ISDC is proposed and applied on the first pooling layer and also every Dense Block after. Specifically, the input information of each ISDC module is accumulated from the outputs of both streams as well as previous ISDC module. The output of each ISDC module can be defined as:

$$O_n^{ISDC} = \delta(\mathcal{F}(y \cdot O_{n-1}^{ISDC} \oplus [O_n^{S1}, O_n^{S2}], \{\mathbf{W}_n\})), \tag{4.7}$$

where $S1$ and $S2$ respectively represent two streams, O_n^{S1} and O_n^{S2} are the outputs of two streams after the first pooling layer or after each Dense Block, $n \in [1, 4]$ represents the index of ISDC modules, $\mathcal{F}$ is a CNN encoder parameterized by $\mathbf{W}_n$, $\oplus$ is element-wise summation, $[\cdot]$ refers to concatenation along channel dimension, y indicates whether this is the first (*i.e.*, $n = 1$) ISDC module between two streams. If $n = 1$, $y = 0$, it refers to the first ISDC module. If $n \in [2, ..., 4]$, $y = 1$, element-wise summation is used to transfer the knowledge from previous ISDC module to the next one. δ denotes ReLU (Nair and Hinton, 2010). Also, Batch Normalization (BN) (Ioffe and Szegedy, 2015) is used in front of each ReLU activation function.

The final output of two Densenet-121 backbone networks (after concatenation by channel) is re-weighted using SEBlock (Hu et al., 2018a) to emphasize informative features. The output of the last ISDC is directly connected to the re-weighted feature

maps by an element-wise summation. Then, a Global Average Pooling (GAP) layer is used followed by a fully-connected layer (FC1), BN, and ReLU. Another fully-connected layer (FC2) is used with N neurons, where N is the number of training identities. At last, a cross-entropy loss is adopted by casting the training process as an ID classification problem.

Notably, instead of using ResNet-50, DenseNet-121 is adopted as the backbone network of two streams in DA-2S because: 1) DenseNet-121 demonstrates similar performance to ResNet-50 in some person re-ID models (*e.g.*, the widely used IDE person re-ID architecture (Zheng et al., 2016a; Deng et al., 2018; Zheng et al., 2017b)), and 2) using DenseNet-121 is able to reduce the chance of over-fitting since parameters of two streams in DA-2S are not shared, and the number of parameters in DenseNet-121 is three times less than ResNet-50.

4.4.2 DA-2S Update

In Sec. 4.4.1 a preliminary DA-2S is trained using labelled source domain data. However, although the data distribution of inputs (*i.e.*, the source domain data) of DA-2S is already close to the target domain, it is still difficult to achieve good cross-domain ST-reID performance without perceiving knowledge from the target domain. That is, when DA-2S is pretrained, only data from the source domain is adopted. In this case, if there are some identities with yellow clothes in the target domain, DA-2S may not learn (or perceive) features from the yellow clothes when there is no identity of yellow clothes in the source domain. Consequently, the discriminability of the pretrained DA-2S is limited. To sort out this problem, this chapter encourages DA-2S to learn more characteristics from the target domain. The details of learned knowledge (or natural characteristic) from the target domain onto the pretrained DA-2S network are introduced in this section.

Virtual Label Assignment for Unlabelled Data on Target Domain

In order to learn knowledge from a target domain, the pretrained DA-2S network is used to extract features of training images from each target domain. The features $f^{t_{xy}}$ (as shown in Fig. 4.4 (b)), which contains knowledge from both foreground and ID-related background information, are used as deep representations of images in the target domain. The extracted target domain features are denoted as follows:

$$F^{t_{xy}} = \left\{ f_1^{t_{xy}}, f_2^{t_{xy}}, ..., f_{N_t}^{t_{xy}} \right\}, \tag{4.8}$$

where N_t is the number of training images in the target domain.

Then, unsupervised clustering is employed to classify these deep representations into different clusters. Because the number of IDs is unknown in the target domain (Ester et al., 1996), the DBSCAN (Ester et al., 1996) (as in (Song et al., 2020; Fu et al., 2019; Zhang et al., 2019b)) is adopted for clustering since DBSCAN does not require predefined number of clusters. After clustering, images which are not classified as outliers (do not belong to any cluster) by DBSCAN are assigned with virtual labels according to the clustering result:

$$X_{\mathbb{D}_t} = \{I_i : (y_i \parallel -1)\}_{i=1}^{N_t}, y_i \in [1, ..., N_c] \tag{4.9}$$

where $X_{\mathbb{D}_t}$ represents the training set of target domain with virtual labels, N_c is the number of clusters. Each image I_i in $X_{\mathbb{D}_t}$ is assigned a virtual label $y_i \in [1, ..., N_c]$ according to the cluster that I_i belongs to. If I_i does not belong to any cluster, it is regarded as an outlier and is assigned a label -1. Images with label -1 do not participate in DA-2S update.

Clustering Refinement

After clustering, a training set $X_{\mathbb{D}_t}$ from the target domain can be obtained, where each image has a virtual label. However, since the effectiveness of virtual

label estimation highly depends on the quality of clustering result, *Dynamic Clustering Confidence Value* (DCCV) is proposed to further refine the clustering result dynamically after DBSCAN. The DCCV is used to select more reliable image samples in $X_{\mathbb{D}_t}$, and discard samples with confidence value below the threshold. To achieve this, first, the density center of each cluster is calculated. The density center can be regarded as the mean value of feature vectors belonging to one specified cluster. For instance, given all images that belong to a cluster j, the density center $\overline{f^{t_{xy}}(j)}$ of cluster j can be formulated as follows:

$$\overline{f^{t_{xy}}(j)} = \frac{\sum_{i=1}^{N_t}[f_i^{t_{xy}} \cdot (y_i == j)]}{N_t^j}, \tag{4.10}$$

where N_t^j is the total number of images that belong to the cluster j. If $y_i == j$, $(y_i == j) = 1$, otherwise $(y_i == j) = 0$.

Finally, given an image I_i of cluster j, DCCV calculates the average Euclidean distance between each image of cluster j and the density center $\overline{f^{t_{xy}}(j)}$. This average distance is then used to compare the distance between I_i and $\overline{f^{t_{xy}}(j)}$:

$$\begin{aligned} DCCV(I_i) = &\left\| \overline{f^{t_{xy}}(j)} - f_i^{t_{xy}} \right\|_2 - \\ &\frac{\sum_{i=1}^{N_t} \left[\left\| \overline{f^{t_{xy}}(j)} - f_i^{t_{xy}} \right\|_2 \cdot (y_i == j) \right]}{N_t^j}. \end{aligned} \tag{4.11}$$

According to the definition of DCCV, if $DCCV(I_i) => 0$, I_i is far away from its density center comparing with other samples of cluster j. Thus, I_i is regarded as an unreliable sample, and its virtual label is then set to -1. That is, this sample is regarded as an outlier. Consequently, this sample I_i does not participate in the DA-2S update. If $DCCV(I_i) < 0$, I_i is close to its density center. Therefore, I_i along with its virtual label are used for DA-2S update. Fig. 4.5 illustrates the function of DCCV. It can be observed that, after DCCV, some unreliable samples in different clusters are marked as outliers. The remaining samples are used for updating the pretrained DA-2S model.

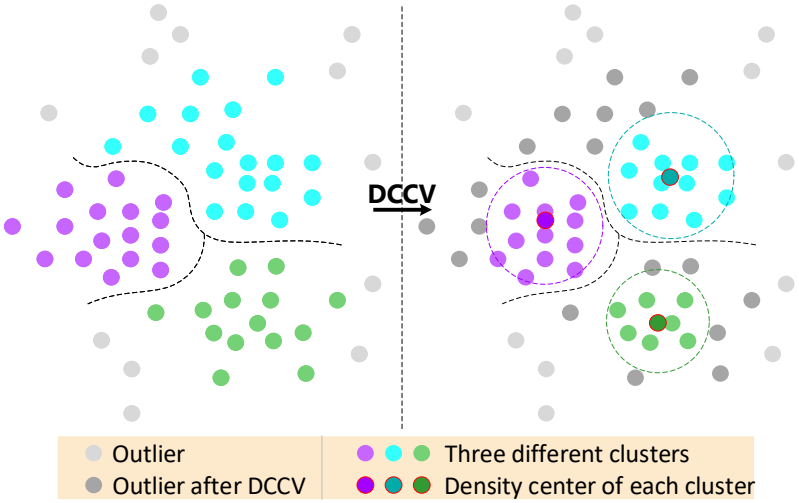


Figure 4.5 : The effectiveness of DCCV. Some samples become outliers after clustering (left figure). After conducting DCCV, unreliable samples that are far from their density centers are marked as outliers (right figure).

Loss Function for DA-2S Update

After DBSCAN and clustering refinement, preliminary N_c clusters (refer to Sec. 4.4.2 and 4.4.2) are obtained. These clustering IDs are assigned to images on the target domain as the virtual labels. However, the actual number of person IDs may not be N_c. Therefore, DA-2S is updated N_{iter} times until it reaches an approximate optimal solution. During every update, N_{epoch} training epochs are used, and N_c is progressively updated when DBSCAN is carried out every time. Since N_c is updated every time, as in (Fan et al., 2018; Song et al., 2020; Fu et al., 2019; Zhang et al., 2019b), the batch-hard triplet loss (Hermans et al., 2017) is used to

update the DA-2S network without a fixed ID number.

$$L_t(f) = \sum_{i=1}^{P}\sum_{a=1}^{Q}[\alpha + \overbrace{\max_{pos=1\ldots Q} \|f_{i,a} - f_{i,pos}\|_2}^{hardest\ positive} - \underbrace{\min_{\substack{neg=1\ldots Q \\ j=1\ldots P \\ j\neq i}} \|f_{i,a} - f_{j,neg}\|_2}_{hardest\ negative}]_+, \tag{4.12}$$

where P and Q represent the number of clusters (IDs) and instances of each cluster in a training minibatch, respectively. Thus, the batchsize is $P \times Q$. f represents feature used to calculate the distance between an anchor sample $f_{i,a}$ and the corresponding hardest positive (negative) sample $f_{i,p}$ ($f_{i,n}$). Finally, for the target domain inputs to DA-2S, features extracted from the soft-mask image (*i.e.*, f^{t_x}), the corresponding original image (*i.e.*, f^{t_y}), and their joint output (*i.e.*, $f^{t_{xy}}$) are used to calculate the batch-hard triplet loss as follows:

$$L_{joint} = L_t(f^{t_x}) + L_t(f^{t_y}) + L_t(f^{t_{xy}}). \tag{4.13}$$

Three different features can refer to Fig. 4.4 (b).

4.5 Experiments

In this section, comprehensive evaluations (qualitative and quantitative) are carried out to verify the effectiveness of the proposed SBSGAN and the DA-2S network for cross-domain ST-reID. The effectiveness of soft-mask images generated by SBSGAN and virtual label assignment to target domain data are verified qualitatively. The performance of DA-2S for cross-domain ST-reID is evaluated quantitatively. The experiments are mainly conducted on Market-1501→ DukeMTMC-reID (using Market-1501 (Zheng et al., 2015) for training and DukeMTMC-reID (Ristani et al., 2016; Zheng et al., 2017b) for testing), since both datasets have fixed training/testing splits. In addition, other results are given on three widely used person

re-ID datasets, including Market-1501 (Zheng et al., 2015), DukeMTMC-reID (Ristani et al., 2016; Zheng et al., 2017b), and CUHK03 (Li et al., 2014).

4.5.1 Datasets for Evaluations

Market-1501 is collected from six cameras in Tsinghua University. It contains 751 IDs with 12,936 images for training. The average training images per ID is 17.2, making it a widely used large person re-ID dataset. The test set contains 750 IDs with 3,368 query images and 19,732 gallery images.

DukeMTMC-reID is collected from eight cameras in Duke University. The original dataset is used for multi-target pedestrian tracking (Ristani et al., 2016). Its re-ID version is used here for evaluation (Zheng et al., 2017b). This dataset contains 702 IDs with 16,522 images for training. Another 702 IDs are used for testing. In the testing set, there are 2,228 query images and 17,661 gallery images.

CUHK03 is captured by six cameras in CUHK. It contains 1,467 IDs with 14,097 images in total. The CUHK03 dataset contains two image settings: one is annotated by hand-drawn bounding boxes, the other one is produced by the DPM detector (Felzenszwalb et al., 2010). Only the detected images are used in the experiment, which is more challenging.

4.5.2 Evaluation Criteria

All query images are used to retrieve corresponding person images in the galley set. Single-query evaluation is adopted. The conventional rank-n accuracy and mean Average Precision (mAP) are used as evaluation criteria (Zheng et al., 2015).

4.5.3 Implementation Details

SBSGAN. All images of three datasets ($K = 3$) are used to train the proposed SBSGAN. Only domain labels are used for training. Input images and their corre-

sponding body masks are resized to 256×128. Adam (Kingma and Ba, 2014) is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batchsize is set to 16. To train G, $\frac{K+1}{16}$ images of each minibatch are randomly selected for soft-mask image generation as well as the auxiliary style-transferred image generation. The remaining images in a minibatch are used for the general style transfer to stabilize the data generation performance in G. The learning rate is initially set to 0.0001 for both G and D, and the model stops training after 5 epochs. One G update is performed after five D updates as in (Gulrajani et al., 2017). In testing, an indicator (*i.e.*, $\bar{\mathbb{D}}$) and an original image (*i.e.*, $I_{\mathbb{D}_s}$) are concatenated for the soft-mask image generation. Notably, there is no need to use any foreground or body mask in testing.

Initial DA-2S Training. Both soft-mask and style-transferred images (to the target domain) are used to pretrain DA-2S (refer to Sec. 4.4). The soft-mask images are generated by the proposed SBSGAN. PTGAN (Wei et al., 2018a) is used to get the general style-transferred images as the input to DA-2S. The batchsize is set to 50. Input images are resized to 256×128 with random horizontal flipping. The SGD is used with momentum 0.9. The initial learning rate is set to 0.1, and decayed to 0.01 after 40 epochs. DA-2S stops training after the 60-th epoch. A reduction rate of 16 is used for SEBlock as in (Hu et al., 2018a). A dropout layer with the rate of 0.5 is inserted after FC1 (see Fig. 4.4 (a)) to reduce the risk of over-fitting. The FC1 layer has 512 neurons. According to the number of training identities, FC2 has 751, 702, and 1,367 neurons when training is conducted on Market-1501, DukeMTMC-reID, and CUHK03, respectively. For each convolutional layer of ISDC, kernel size=3, and padding=1. In addition, stride=2 is used for the first three ISDC modules and stride=1 for the last ISDC module. The number of channels is doubled by each ISDC. Finally, 2,048 channels are obtained after four ISDC modules. The pretrained DA-2S model is trained using labelled source domain data.

DA-2S Update. To update the pretrained DA-2S model, both original target

domain training images and the corresponding soft-mask images are used as inputs (refer to Sec. 4.4.2). The batchsize is set to 64 (16 IDs with 4 instances of each ID). Thus, in Eq. 4.12 $P = 16$ and $Q = 4$. The input images are also resized to 256×128 with random horizontal flipping. During update, FC1, FC2, and the CE loss are removed (see in Fig. 4.4 (a)). Other parts in the pretrained model are directly loaded, including two Densenet-121 streams, ISDC modules, and SEBlock. The $N_{iter} = 30$ and $N_{epoch} = 120$ (refer to Sec. 4.4.2). Thus, DBSCAN is conducted 30 times for virtual label estimation, and for each time, 120 training epochs are executed. In each update process, the initial learning rate is set to 6e-3, and it decays to 6e-4 after 100 training epochs. The α in Eq. 4.12 is set to 0.5. Following the same setting of the pretrained DA-2S, the SGD optimizer is used with momentum 0.9. In testing, 4,096-dim CNN features (f^{t_x}, f^{t_y}, and $f^{t_{xy}}$ as shown in Fig. 4.4 (b)) are extracted for each testing image. The Euclidean distance is used to compute the similarity between query and gallery images.

4.5.4 Qualitative Evaluation

This Experiment is To Confirm That Soft-Mask Images Are Better Than Hard-Mask Images in Suppression of Background Shift. In Fig. 4.6, a comparison is given between the generated soft-mask images and hard-mask images. The hard-mask images are respectively obtained by JPPNet (Liang et al., 2018) and Mask-RCNN (Abdulla, 2017; He et al., 2017). Both methods have shown compelling performance in person parsing or object instance segmentation. However, it can be observed that both methods cannot perform well in body segmentation from the background on existing person re-ID datasets. As shown in Fig. 4.6, when people carry objects (*e.g.*, bags), these objects are regarded as backgrounds and removed by noisy foreground masks with segmentation errors. However, such features are significant to person re-ID, which should be retained rather than removed. On the

Figure 4.6 : Comparison between hard-mask and soft-mask images. Images are selected from three different person re-ID datasets. The original images are listed in the first row. The second and the third rows respectively show hard-mask images by Mask-RCNN (Abdulla, 2017; He et al., 2017) and JPPNet (Liang et al., 2018). The last row shows soft-mask images generated by the proposed SBSGAN.

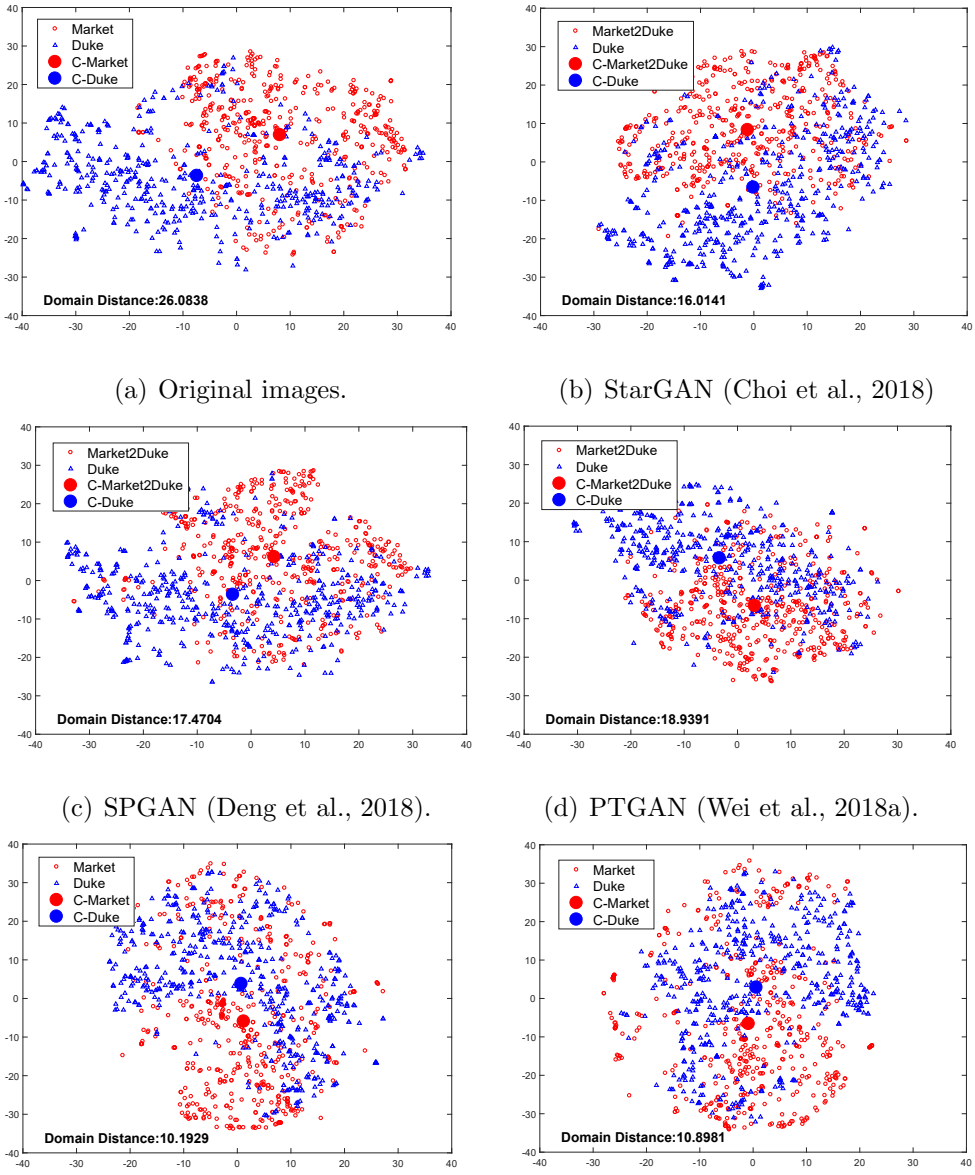
contrary, in soft-mask images, important cues such as bags and body parts can be well generated and retained. This is because the binary body mask is not directly utilized on original images to remove the backgrounds. Although the foreground mask obtained by JPPNet (the third row in Fig. 4.6) is also used to suppress the background (refer to Eq. 4.3), the generated images by the proposed SBSGAN show better results. This phenomenon also shows that the proposed SBSGAN is robust to the noisy masks in the data generation.

This Experiment is To Confirm That The Effectiveness of Loss Functions in SBSGAN. The proposed SBSGAN jointly optimizes over several loss functions (see Eq. 4.5 and Eq. 4.6). Fig. 4.7 shows images generated by SBSGAN using different loss functions. The effectiveness of $\mathcal{L}_{idc}$, $\mathcal{L}_{bgs}$, and $\mathcal{L}_{sc}$ are verified. Others are conventional GAN-based loss functions, and their effectiveness is already evaluated by several previous works (Arjovsky et al., 2017; Choi et al., 2018; Gulrajani et al., 2017; Isola et al., 2017; Taigman et al., 2017; Zhu et al., 2017). It can be observed in Fig. 4.7 that when $\mathcal{L}_{idc}$ and $\mathcal{L}_{bgs}$ are removed, the color information

Figure 4.7 : The effectiveness of different loss functions. Best viewed in color.

of original images cannot be well preserved. In addition, the background cannot be well suppressed. By only removing $\mathcal{L}_{sc}$, SBSGAN can generate soft-mask images which are close to the objective. The $\mathcal{L}_{sc}$ is proposed to encourage the style of generated soft-mask images to be consistent (refer to Eq. 4.4). Apart from the qualitative comparison in Fig. 4.7, a quantitative evaluation can be found in Tab. 4.1 to verify the effectiveness of $\mathcal{L}_{sc}$ further.

This Experiment is To Confirm That Reducing the Background Shift Is Effective to Reduce the Domain Shift. It is Shown By Visualization of Data Distributions Between Two Domains. The distance between different domains is visualized using different types of data, including the popular style-transferred images, hard-mask images, and the generated soft-mask images. Three recently published methods SPGAN (Deng et al., 2018), PTGAN (Wei et al., 2018a), and StarGAN (Choi et al., 2018) are used to transfer the image style from Market-1501 to DukeMTMC-reID, respectively. Fig. 4.8 shows the result. Compared with the general style transferred results, the hard-mask and soft-mask images can reduce the domain shift by a large margin. This phenomenon verifies the effectiveness of reducing the domain shift by considering the background shift problem. The domain distance of hard-mask images is on par with the soft-mask images (10.19 *vs.* 10.90). However, compared with hard-mask images, the generated soft-mask images show a better performance in cross-domain ST-reID (*e.g.*, rank-1: 43.3% *vs.* 38.6%, see



(a) Original images.

(b) StarGAN (Choi et al., 2018)

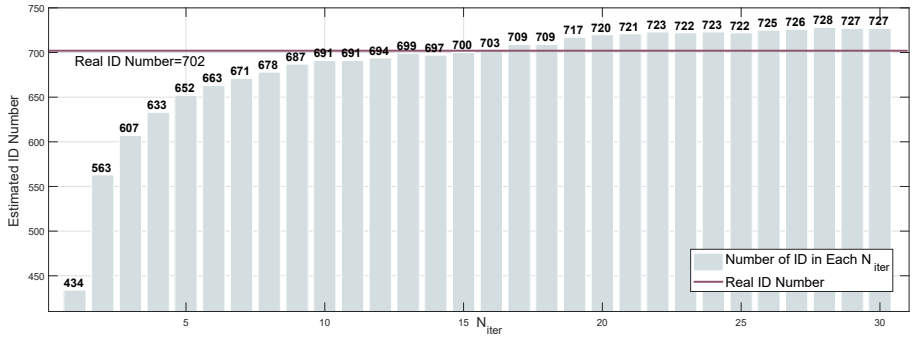
(c) SPGAN (Deng et al., 2018).

(d) PTGAN (Wei et al., 2018a).

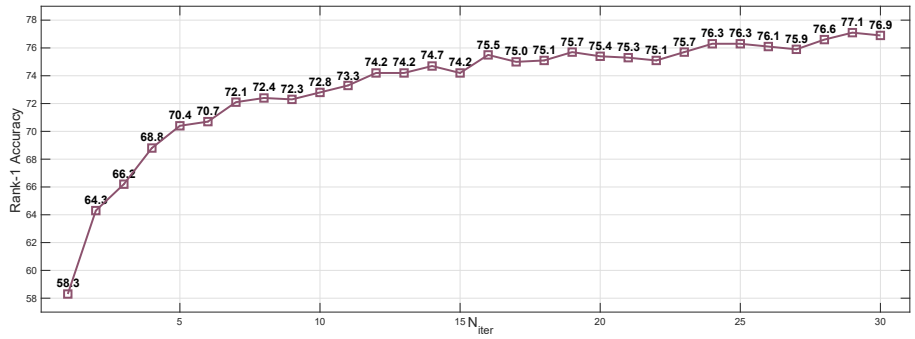
(e) Hard-mask images (JPPNet).

(f) Soft-mask images.

Figure 4.8 : Data visualization. 5000 images are randomly selected from Market-1501 and DukeMTMC-reID to learn data distributions via the Barnes-Hut t-SNE (Van Der Maaten, 2014a), respectively. Another 200 images of each domain are used for visualization. The red circle and blue triangle respectively represent images belonging to Market-1501 and DukeMTMC-reID. The center points (*i.e.*, 'C-') are shown using their corresponding domain color. Domain distance (*i.e.*, L_1 distance) is given between center points.



(a) The number of estimated ID.



(b) Rank-1 accuracy.

Figure 4.9 : The changes of estimated ID number (a) and rank-1 accuracy (b) that come with the increases of N_{iter} (from 1 to 30). This experiment is conducted with DukeMTMC-reID dataset as the target domain.

Tab. 4.1). Naturally, it is unfair to compare the domain distance between soft-mask and hard-mask images directly. This is because many pixel values of hard-mask images are simply zeroed out, making approximate half of the information of hard-mask images already being discarded in the comparison. Alternatively, soft-mask images suppress backgrounds rather than simply removing them.

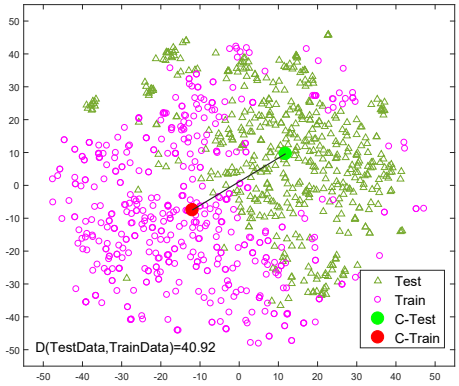
This Experiment is To Confirm That DA-2S Update Can Encourage The Number of Estimated IDs to Be Close to The Actual Number of Real IDs. As shown in Fig. 4.9 (a), the number of estimated IDs are recorded when the DBSCAN clustering approach is conducted every time. The real ID number in the

target domain training set is 702 (*i.e.*, DukeMTMC-reID). In the beginning, for examples in the first five N_{iter}, the estimated ID number increases fast after DA-2S updated each time. It is close to the real ID number when $N_{iter} = 16$. Then, the number of estimated ID gradually reaches a stable range (*i.e.*, [720,728] when $N_{iter} = [20, 30]$). Although the estimated ID number is not exactly the same as the real ID number, this experiment demonstrates that DA-2S update can effectively explore the natural characteristics of the target domain according to the quality of virtual ID label estimation. In Fig. 4.9 (b), it can be observed that the rank-1 accuracy increases dramatically when $N_{iter} = [1, 5]$ (*i.e.*, from 58.3% to 70.4%), which is in line with the increase of estimated ID number. After that, the rank-1 accuracy grows steadily, and it achieves the best result when $N_{iter} = 29$. Finally, N_{iter} is set to 30 according to this experiment.

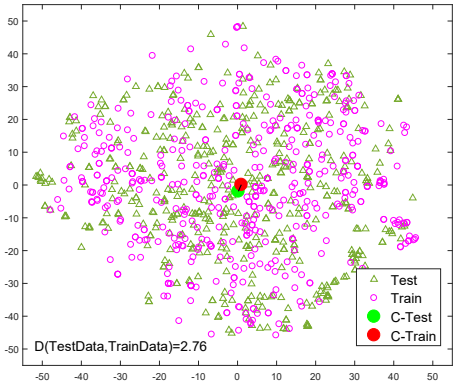
This Experiment is To Confirm That DA-2S Update Can Effectively Reduce the Feature Distribution Shift Between Training and Testing Data. The feature distributions of training and testing data are visualized in Fig. 4.10. The features are extracted by DA-2S. As shown in Fig. 4.10 (a), when only using the pretrained DA-2S network, the feature distribution of source domain training data is far from the feature distribution of target domain testing data (*i.e.*, 40.92). On the contrary, when DA-2S is updated after N_{iter} times, it can clearly see that the feature distributions of target domain training data are close to the feature distribution of target domain testing data (*i.e.*, 2.76). This experiment demonstrates that the DA-2S update can effectively reduce the shift of features between training and testing domains.

4.5.5 Quantitative Evaluation

Soft-Mask Images *vs.* Other Types of Images. The widely used IDE model (Zheng et al., 2016a, 2017b; Deng et al., 2018) with ImageNet-trained DenseNet-



(a) Features extracted using pretrained DA-2S (*i.e.*, Fig. 4.4 (a)).



(b) Features extracted using updated DA-2S (*i.e.*, Fig. 4.4 (b)).

Figure 4.10 : Feature distributions of training and testing data. 5,000 training data from source (Market-1501) and target (DukeMTMC-reID) domains are selected to extract $f^{s_{xy}}$ and $f^{t_{xy}}$ via pretrained DA-2S and updated DA-2S respectively. Another 5,000 testing data are selected from target domain to extract $f^{s_{xy}}$ and $f^{t_{xy}}$ using pretrained DA-2S and updated DA-2S respectively. Barnes-Hut t-SNE (Van Der Maaten, 2014a) is used to learn the distribution of features extracted from two different DA-2S models. The center point of each distribution is denoted by 'C-'. $D(\cdot,\cdot)$ represents the distance (*i.e.*, L_1 distance) between center points.

121 as backbone network is adopted to compare cross-domain ST-reID performance across different types of images, including the generated soft-mask images, the general style-transferred images, and hard-mask images. Tab. 4.1 lists the performance. By directly using the original images, the performance is inferior (mAP: 17.7%, rank-1: 33.5%). A clear performance improvement is achieved by directly removing backgrounds from both training and testing images using foreground masks obtained by JPPNet and Mask-RCNN, respectively. However, the performance of soft-mask images outperforms hard-mask images by +4.7% in rank-1 accuracy (43.3% *vs.* 38.6%). This is because hard-mask images normally contain segmentation errors.

Table 4.1 : Baseline performance of cross-domain ST-reID. Market-1501 is for training and DukeMTMC-reID is for testing.

Training Data	mAP	R-1	R-5	R-10
Original	17.7	33.5	49.3	55.1
Hard-mask Images				
Mask-RCNN (Abdulla, 2017; He et al., 2017)	20.6	37.5	53.4	59.1
JPPNet (Liang et al., 2018)	21.5	38.6	54.3	60.0
Style-transferred Images				
PTGAN (Wei et al., 2018a)	22.7	42.9	58.0	64.2
SPGAN (Deng et al., 2018)	22.8	42.0	57.9	64.1
StarGAN (Choi et al., 2018)	21.6	39.8	53.4	59.9
Soft-mask Images (Ours)				
Soft-mask w/o $\mathcal{L}_{sc}$	21.2	41.7	56.3	62.7
Soft-mask	22.3	43.3	58.2	64.4
Soft-mask$_{2-Domains}$	**23.5**	**44.2**	**59.5**	**65.3**

The general style-transferred results, such as PTGAN and SPGAN achieve competitive performance. However, the soft-mask images obtain the best rank-1 accuracy (43.3%), which shows the effectiveness by considering the background shift problem in cross-domain ST-reID. In addition, without $\mathcal{L}_{sc}$, images generated by SBSGAN can satisfy the visual requirement (see Fig. 4.7), but the performance is dropped by 1.1% in mAP and 1.6% in rank-1 accuracy. This is because $\mathcal{L}_{sc}$ is used to normalize the style of soft-mask images across multiple domains, by which the inter-domain shift can be further reduced. Since SPGAN and PTGAN only support images of two domains as inputs, the proposed SBSGAN is also trained in the same way instead of using images from three domains. Without interference from images of the third domain (*i.e.*, CUHK03), the performance gains by Soft-mask$_{2-Domains}$ (mAP: 23.5%, rank-1: 44.2%). However, multiple domains as inputs are still used in all

Table 4.2 : Ablation study of DA-2S. Market-1501 is used for training, and DukeMTMC-reID is used for testing. The baseline does not use SEBlocks and any ISDC modules. This experiment also tries to add SEBlocks to every ISDC module to re-weight the output of ISDC in the middle layers (denoted as ISDC-SE). The DA-2S† (DA-2S‡) means only using the style-transferred images (soft-mask images) as the inputs of the 2-stream network.

Methods	mAP	R-1
Basel.	28.8	50.2
Basel.+SEBlock	28.9	50.5
Basel.+SEBlock+ISDC-SE	30.4	51.5
Basel.+SEBlock+ISDC (DA-2S)	**30.8**	**53.5**
DA-2S† (2*Style-transfer)	28.4	49.6
DA-2S‡ (2*Soft-mask)	27.0	51.5

experiments to generate soft-mask images. Thus, only one model is needed to be trained instead of training multiple models between any two domains.

Ablation Study of Pretrained DA-2S. An ablation study of the proposed DA-2S network is given in Tab. 4.2. Without SEBlock and ISDC (*i.e.*, baseline), DA-2S achieves 28.8% in mAP and 50.2% in rank-1 accuracy. By using SEBlock, the performance is improved from 50.2% to 50.5% in rank-1 accuracy. To strengthen the inter-stream relationship, Basel.+SEBlock+ISDC produces the best performance (mAP: 30.8%, rank-1: 53.5%), demonstrating the effectiveness of the proposed ISDC modules. If SEBlocks are added to every ISDC module (ISDC-SE), the performance is dropped by 2% in rank-1 accuracy. This is because additional SEBlocks produce more parameters, which can potentially increase the risk of over-fitting. Moreover, the inputs of 2-stream DA-2S is changed to style-transferred images or soft-mask images only (*i.e.*, the network receives two style-transferred images or two soft-

Table 4.3 : Ablation study of DA-2S update. Market-1501 (DukeMTMC-reID) is used for training (testing). The 1^{st} line represents baseline performance (pretrained DA-2S). The 2^{nd} and 3^{rd} lines are performance when only one stream is used (without using ISDC and SEBlock). The 4^{th} to 6^{th} lines are using different combinations of L_t in Eq. 4.13. The last two lines shows the effectiveness of DCCV. S1 and S2 respectively represent Stream1 and Stream2.

Modules								**Market2Duke**	
Update	S1	S2	ISDC&SEBlock	$L_t(f^{t_x})$	$L_t(f^{t_y})$	$L_t(f^{t_{xy}})$	DCCV	mAP	R-1
✗	✓	✓	✓	✗	✗	✗	✗	30.8	53.5
✓	✓	✗	✗	✓	✗	✗	✗	37.8	61.1
✓	✗	✓	✗	✗	✓	✗	✗	48.2	68.7
✓	✓	✓	✓	✗	✗	✓	✗	55.1	73.5
✓	✓	✓	✓	✓	✗	✓	✗	55.6	73.8
✓	✓	✓	✓	✗	✓	✓	✗	55.2	73.6
✓	✓	✓	✓	✓	✓	✓	✗	57.5	74.9
✓	✓	✓	✓	✓	✓	✓	✓	**61.3**	**76.9**

mask images) to show the performance (*i.e.*, DA-2S† and DA-2S‡). The results demonstrate that the combination of two types of images is better than using them independently.

Ablation Study of DA-2S Update. The ablation study of DA-2S update is given in Tab. 4.3. The pretrained DA-2S is used as a baseline (mAP: 30.8%, rank-1: 53.5%). It can be seen that when only using one pretrained stream for updating (without loading the ISDC and SEBlock modules), the performance can be improved from 53.5% to 61.1% and 68.7% in rank-1 accuracy. However, compared with using both streams (*i.e.*, rank-1: 74.9%), the performance of using one stream is still inferior. This phenomenon verifies both foreground and ID-related

information in the background should be jointly considered in DA-2S update. The experiment also tries to remove a part of loss functions in Eq. 4.13 to testify the effectiveness of each loss when the full pretrained DA-2S model is used for updating. It can be observed that both f^{t_x} and f^{t_y} are useful when they participate in DA-2S update. If removing one or both of them, the rank-1 accuracy can be reduced from 74.9% to 73.6% (w/o $L_t(f^{t_x})$), 73.8% (w/o $L_t(f^{t_y})$), and 73.5% (w/o $L_t(f^{t_x})$ & $L_t(f^{t_y})$), respectively. This experiment demonstrates that features extracted from both foreground and ID-related information in the background can improve cross-domain ST-reID performance in DA-2S. At last, the experiment attempts to only remove the DCCV module when the DA-2S update is conducted. The performance is reduced from 76.9% to 74.9% in rank-1 accuracy. This experiment demonstrates the effectiveness of DCCV.

Comparison With State-of-the-Art Methods. The performance of the proposed approach is compared with several recently published *State-Of-The-Art* (SOTA) cross-domain ST-reID methods, including PUL (Fan et al., 2018), PT-GAN (Wei et al., 2018a), SPGAN+LMP (Deng et al., 2018), TJ-AIDL (Wang et al., 2018), HHL (Zhong et al., 2018b), ATNet (Liu et al., 2019), PAUL (Yang et al., 2019b), ECN (Zhong et al., 2019), CR-GAN (Chen et al., 2019c), PDA-Net (Li et al., 2019a), UCDA-CCE (Qi et al., 2019), CASCL (Wu et al., 2019a), PCB-R-PAST (Zhang et al., 2019b), SSG (Fu et al., 2019), and Theory (Song et al., 2020). For all methods, all training images from the source domain have ground-truth ID labels. On the contrary, the ID labels in the target domain are unavailable. Tab. 4.4 lists the comparison results. It is clear to see that the proposed method achieves the best cross-domain ST-reID performance. The proposed approach outperforms the SOTA method SSG++ 3.7% (1.7%) in rank-1 accuracy when testing is conducted on DukeMTMC-reID (Market-1501). Besides, if CUHK03 is used to pretrain DA-2S, the proposed approach outperforms the SOTA approach PCB-R-PAST 2.9%

Table 4.4 : Comparison with SOTA methods. X→Y means training is conducted on X and testing is conducted on Y. The performance of pretrained DA-2S is denoted as SBSGAN+DA-2S. '-I' represents the updated DA-2S network. '-II' is the result when re-ranking post-processing trick (Zhong et al., 2017) is adopted on SBSGAN+DA-2S-I. '-' means the performance is not released.

Methods	Market→Duke		Duke→Market		CUHK03→Duke		CUHK03→Market	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
PUL (Fan et al., 2018) *TOMM2018*	16.4	30.0	20.5	45.5	12.0	23.0	18.0	41.9
PTGAN (Wei et al., 2018a) *CVPR2018*	-	27.4	-	38.6	-	17.6	-	31.5
SPGAN+LMP (Deng et al., 2018) *CVPR2018*	26.4	46.9	26.9	58.1	-	-	-	-
TJ-AIDL (Wang et al., 2018) *CVPR2018*	23.0	44.3	26.5	58.2	-	-	-	-
HHL (Zhong et al., 2018b) *ECCV2018*	27.2	46.9	31.4	62.2	23.4	42.7	29.8	56.8
ATNet (Liu et al., 2019) *CVPR2019*	24.9	45.1	25.6	55.7	-	-	-	-
PAUL (Yang et al., 2019b) *CVPR2019*	35.7	56.1	36.8	66.7	-	-	-	-
ECN (Zhong et al., 2019) *CVPR2019*	40.4	63.3	43.0	75.1	-	-	-	-
CR-GAN (Chen et al., 2019c)+LMP (Deng et al., 2018) *ICCV2019*	33.3	56.0	33.2	64.5	26.9	46.5	30.4	58.5
CR-GAN (Chen et al., 2019c)+TAUDL (Li et al., 2018a) *ICCV2019*	48.6	68.9	54.0	77.7	47.7	67.7	56.0	78.3
PDA-Net (Li et al., 2019a) *ICCV2019*	45.1	63.2	47.6	75.2	-	-	-	-
UCDA-CCE (Qi et al., 2019) *ICCV2019*	36.7	55.4	34.5	64.3	-	-	-	-
CASCL (Wu et al., 2019a) *ICCV2019*	30.5	51.5	35.6	64.7	-	-	-	-
PCB-R-PAST (Zhang et al., 2019b) *ICCV2019*	54.3	72.4	54.6	78.4	51.8	69.9	57.3	79.5
SSG (Fu et al., 2019) *ICCV2019*	53.4	73.0	58.3	80.0	-	-	-	-
SSG+ (Fu et al., 2019) *ICCV2019*	56.7	74.2	62.5	81.4	-	-	-	-
SSG++ (Fu et al., 2019) *ICCV2019*	60.3	76.0	68.7	86.2	-	-	-	-
Theory (Song et al., 2020) *PR2020*	49.0	68.4	53.7	75.8	-	-	-	-
SBSGAN+DA-2S (Our) (Huang et al., 2019b)	30.8	53.5	27.3	58.5	27.8	47.7	28.5	57.6
SBSGAN+DA-2S-I (Our)	61.3	76.9	69.5	86.9	53.2	70.2	67.2	84.5
SBSGAN+DA-2S-II (Our)	**71.5**	**79.7**	**80.0**	**87.9**	**66.5**	**72.8**	**81.0**	**86.2**

(6.7%) in rank-1 accuracy when testing is conducted on DukeMTMC-reID (Market-1501). It can be seen that SSG++ achieves competitive performance comparing with SBSGAN+DA-2S-I (without using re-ranking post-processing trick). This is because SSG++ benefits from 1) body part partition and 2) a joint training strategy via clustering-guided semi-supervised training. The proposed method does not try to divide the input images into different body parts further and conduct clustering on each part. In addition, the clustering-guided semi-supervised training needs to double the loss items in the final objective function (*i.e.*, Eq. 4.13). The proposed approach only focuses on the whole images by considering the importance of foreground and ID-related information in background. Through DA-2S update that benefits from knowledge of target domain data, the proposed approach significantly improves the performance of the preliminary work (Huang et al., 2019b) (*e.g.*, the rank-1 accuracy is improved from 53.5% to 79.7% when testing is conducted on DukeMTMC-reID). In addition, comparing with existing clustering-based cross-domain ST-reID approaches (*i.e.*, PUL, PCB-R-PAST, SSG, and Theory listed in Tab. 4.4), the proposed DA-2S method, which is based on soft-mask images generated by SBSGAN, achieves the best performance.

4.6 Conclusion

In this chapter, the background shift issue is first considered to reduce the domain shift for cross-domain ST-reID. SBSGAN is proposed to generate soft-mask images with the background being suppressed. Compared with hard-mask solutions, soft-mask images are able to suppress the background moderately. Compared with general inter-domain style-transferred approaches, soft-mask images can further reduce the domain shift by considering the background shift problem. A DA-2S model is introduced along with the proposed ISDC module to make use of helpful background cues. To further explore/learn the natural characteristics from unlabelled

target domain training data. An update strategy is given based on the proposed DA-2S network and images generated by SBSGAN. Based on DBSCAN clustering results, the proposed DCCV is used to improve the virtual label estimation quality. Experiment results demonstrate the effectiveness of the proposed approach in both qualitative and quantitative evaluations. SOTA performance is achieved.

Chapter 5

New Benchmarks And Solutions for Clothing Change LT-reID

5.1 Motivation

In past years, several datasets have been proposed for person re-ID and contribute significantly to the community, *e.g.*, VIPeR (Gray and Tao, 2008), CUHK03 (Li et al., 2014), Market1501 (Zheng et al., 2015), DukeMTMC-reID (Zheng et al., 2017b), *etc.* However, most of them are based on the assumption that each person who appears under one camera will re-appear under another one in a short period of time (*e.g.*, less than 30 minutes). Under this assumption, a person is less likely to change clothing. This scenario is defined as ST-reID. On the contrary, if a person appears again after a long-time gap (*e.g.*, more than one day), the chance of changing clothes or carrying different objects will become large. This scenario is regarded as clothing change LT-reID. This chapter will investigate the feasibility of the existing datasets for clothing change LT-reID as defined above. The clothing change LT-reID is a more challenging case commonly seen in large-scale video security surveillance.

According to the existing research outcomes and potential applications in practice, a dataset is suitable for clothing change LT-reID research should satisfy the following requirements: 1) a large number of person IDs, 2) highly diverse environment/backgrounds, 3) multiple camera views, 4) various shooting conditions (*e.g.*, illumination and resolution), and 5) highly dynamic appearance of each person. Existing datasets regardless of being indoor or outdoor, there are mainly two methods for collecting datasets: 1) collecting data in a free setup environment with

non-collaborative people (Gray and Tao, 2008; Li et al., 2014; Zheng et al., 2015, 2017b); 2) collecting data in a constrained environment with collaborative people (*e.g.*, actors) (Barbosa et al., 2012b; Munaro et al., 2014; Haque et al., 2016; Zhang et al., 2018b). Either method has its advantages but also clear problems when the aforementioned requirements are to be satisfied.

The first method can best align with the situations of a real surveillance environment. However, in practice, the data annotation is extremely difficult even just for a few thousand people ID. In order to allocate the same ID to the same person, the annotation staff has to rely on the face information (if that is available) to give a typical ID for the same person when they appear under different cameras. If the face information is not available due to poor image quality or poor camera view (*e.g.*, back view), the samples have to be discarded or marked based on experiences. For the case of ST-reID, such difficulty is still manageable, where the annotation staff may use clothing information to match people across cameras. However, in the case of LT-reID, the assumption above does not exist. Thus, the annotation for clothing change LT-reID purpose is impossible. That is the main reason that current datasets such as VIPeR (Gray and Tao, 2008), Market1501 (Zheng et al., 2015), and CUHK03 (Li et al., 2014) are not suitable for the clothing change LT-reID for the requirement of the highly dynamic appearance of each annotated person, although they are collected in a free control (at least less constrained) environment.

The second method can best simulate many extremely difficult cases, such as very unusual camera views and clothing changes. For example, a person re-ID dataset is proposed in (Zhang et al., 2018b) that demonstrates the scenarios of person re-ID in the case of clothing change. It tries to recognize different people through gait information across view angles. However, the main issue of the datasets conducted by the second method is the scale, including the number of people ID, diversity of environment, and the number of camera views. This is mainly because building-

up such a dataset in a mimic environment relies on a certain number of actors who follow the predefined way to complete the shooting. In practice, it is not manageable to have a few thousand actors.

Based on the experiences of existing datasets, the key challenging problems to build a dataset particularly suitable for clothing change LT-reID are: 1) sufficient large number of people IDs, 2) dynamic shooting with true environments, 3) various clothes on each person. In this chapter, a new dataset called "Celeb-reID" is proposed, which can tackle the aforementioned challenging problem. It is named Celeb-reID because all images are collected using celebrities' street snap-shots. The celebrities are chosen as the target because there are tremendous resources of celebrity images on the Internet. The street snap-shots of celebrities are used to build the dataset because these street snap-shots are more relevant to the real-life scenario. In this way, a large number of people ID can be acquired. Moreover, celebrities are widely seen and taken photos in various scenarios, so it is easier to obtain their images in different environments. The most important thing is that these celebrities normally wear different clothes in different scenes. It well satisfies the requirements in terms of a highly dynamic environment in the scenarios of security surveillance for clothing change LT-reID.

Since appearance reliability (*e.g.*, color and texture) is reduced greatly due to the change of clothing, current re-ID approaches, which mainly rely on appearance, may not be suitable for the LT-reID scenario. Existing person re-ID approaches are mainly designed for the ST-reID scenario (Zheng et al., 2017a; Hermans et al., 2017; Sun et al., 2017; Yu et al., 2017b; Zheng et al., 2018; Chang et al., 2018; Guanshuo et al., 2018; Yuan et al., 2018; Zhang et al., 2019a). Although these approaches have achieved compelling performance, simply concentrating on appearance and overlooking the change of clothing may not be suitable for the LT-reID scenario. Specifically, existing approaches try to distinguish the inter-class clothing changes

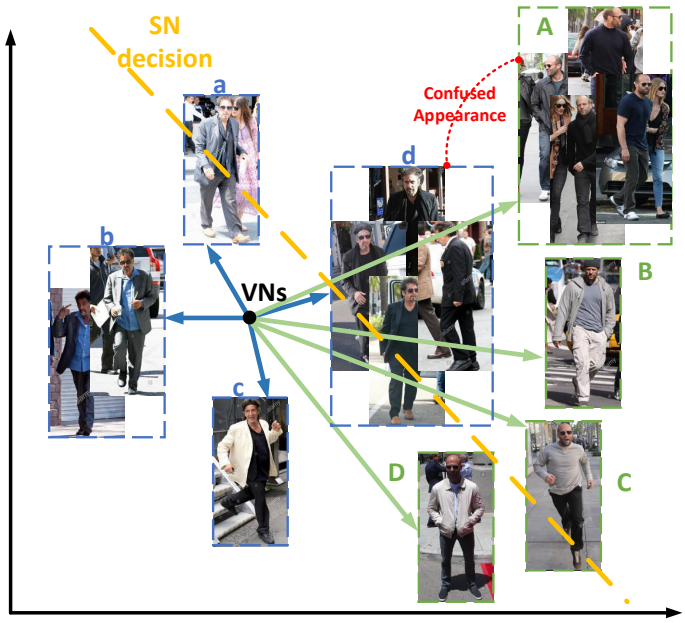


Figure 5.1 : SN *vs.* VN capsule in person re-ID. The (a)-(d) and (A)-(D) are images belonging to two different IDs in the Celeb-reID dataset. The (d) and (A) include two persons with similar dark clothes. The VN capsules use the length of the vector to represent different IDs, while its orientations are used to perceive different types of clothes. With two-dimensional perception capability, different IDs can be distinguished easier by using the length of capsules. Instead, typical SN cannot make a decision between confused appearance (*e.g.*, some images in (d) are regraded as the ID in the green bounding box). Best viewed in color.

but do not pay attention to intra-class clothing changes. Therefore, these methods are not suitable to LT-reID.

Compared with traditional approaches which use the Scalar Neuron (SN), this chapter proposes to use Vector-Neuron (VN) capsules (Sabour et al., 2017) to perceive the clothing change of the same person. In common CNNs, the value of each scalar reflects the likelihood of a neuron belonging to an existing people ID. However, if the clothing is changed, the one-dimensional SN cannot further perceive the clothing change information. Therefore, the two-dimensional VN capsules are inte-

grated into a CNN network. The idea is inspired by (Sabour et al., 2017): the length (the first dimension) of each capsule (denoted as $C_{\mathcal{L}}$) expresses the "existence of the entity"; the orientation (the second dimension) of each capsule (denoted as $C_{\mathcal{O}}$) is forced to represent "the properties of the entity". For clothing change LT-reID case, $C_{\mathcal{L}}$ is expected to reflect the likelihood of an existing people ID while $C_{\mathcal{O}}$ represents different clothes of the same ID. Fig. 5.1 shows the difference between SNs and VNs. With two-dimensional perception capability, VN capsules can discriminate different IDs (through the length of capsules) with confused appearance (perceived by the orientation of capsules), while the SN may make a wrong decision. The proposed model is named "ReIDCaps". Unlike the original application of capsules in (Sabour et al., 2017) (*e.g.*, learning the property of affine transformation for handwritten digits recognition), the proposed ReIDCaps network considers the clothing change of the same person should be further perceived and represented by $C_{\mathcal{O}}$. In the meantime, $C_{\mathcal{L}}$ is used as common SN to distinguish different people ID. Finally, to further enhance the discrimination and generalization power of ReIDCaps, both Soft Embedding Attention (SEA) mechanism and Feature Sparse Representation (FSR) mechanism are integrated into the proposed network.

In general, the main contributions of this chapter can be summarized in two-fold:

- A large-scale LT-reID dataset Celeb-reID is proposed. Approximately more than 70% of the images of each person are in different clothes. There are 1,052 IDs and 34,186 person images in Celeb-reID to make it the largest clothing change LT-reID dataset.

- A ReIDCaps network is introduced to deal with the challenge of clothing change LT-reID. In addition, SEA and FSR mechanisms are integrated to enhance the discrimination and generalization power of the proposed network.

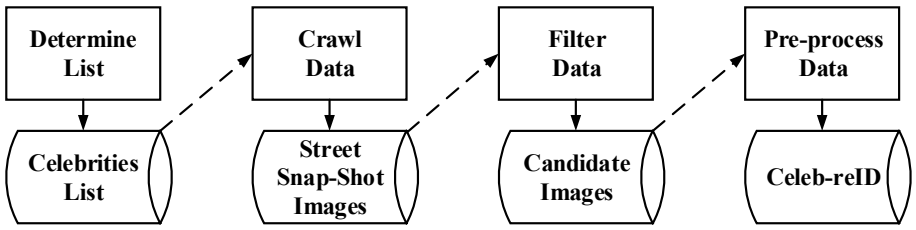


Figure 5.2 : The pipeline of data acquisition. Four main steps are included.

5.2 Celeb-reID: New Benchmarks

In this section, the Celeb-reID dataset is introduced. Fig. 5.2 shows the pipeline of the data acquisition process. It can be summarized into four steps:

i. **Determine Name List.** Since the street snap-shots of celebrities are used to build the dataset, the first step is to determine the name list of celebrities. Initially, 2,600 popular celebrities from all over the world are selected.

ii. **Crawl Data.** Given the name of celebrities, the data of each celebrity are crawled on Google, Bing, and Baidu Images using keywords: name + street snap-shot (*e.g.*, Justin Bieber street snap-shot) with free use license. The top 100 images of each celebrity are crawled. Finally, 2,600×100 candidate street snap-shot images are obtained.

iii. **Filter Data.** In this step, the annotation staff verifies the identity of images for each celebrity. Any wrong returned result is discarded.

iv. **Preprocess Data.** The last step is to crop the bounding box of the person body from the original image. Mask-RCNN (He et al., 2017) is used to detect the bounding box. Pixel paddings from the original image are used to ensure that the ratio of height and width of each image is identical (*i.e.*, 2:1). The final image size is resized to 256×128. Finally, total of 34,186 images of 1052 celebrities are retained.

Figure 5.3 : Three rows represent three different IDs in the Celeb-reID dataset. For each ID, a great number of clothing changes can be found.

Table 5.1 : Data split of the Celeb-Reid dataset. In the testing set, around 30% of images of the 420 IDs belong to the query set, the other 70% of images belong to the gallery set.

split	training	testing		total
subsets	training	query	gallery	total
#ID	632	420	420	1,052
#images	20,208	2,972	11,006	34,186

Similar to traditional setting of ST-reID datasets, Celeb-reID is split into three subsets, including training, gallery, and query (see Tab. 5.1). The query and gallery sets are used for testing. Fig. 5.3 shows three IDs in the proposed dataset. It is clear to see that the clothing change brings about large appearance changes. Notably, a person may wear the same clothing twice. Specifically, more than 70% of the images of each person show different levels of clothing change. Fig. 5.4 gives the statistical information of the proposed dataset, including the distribution of age, gender, and nationality of celebrities.

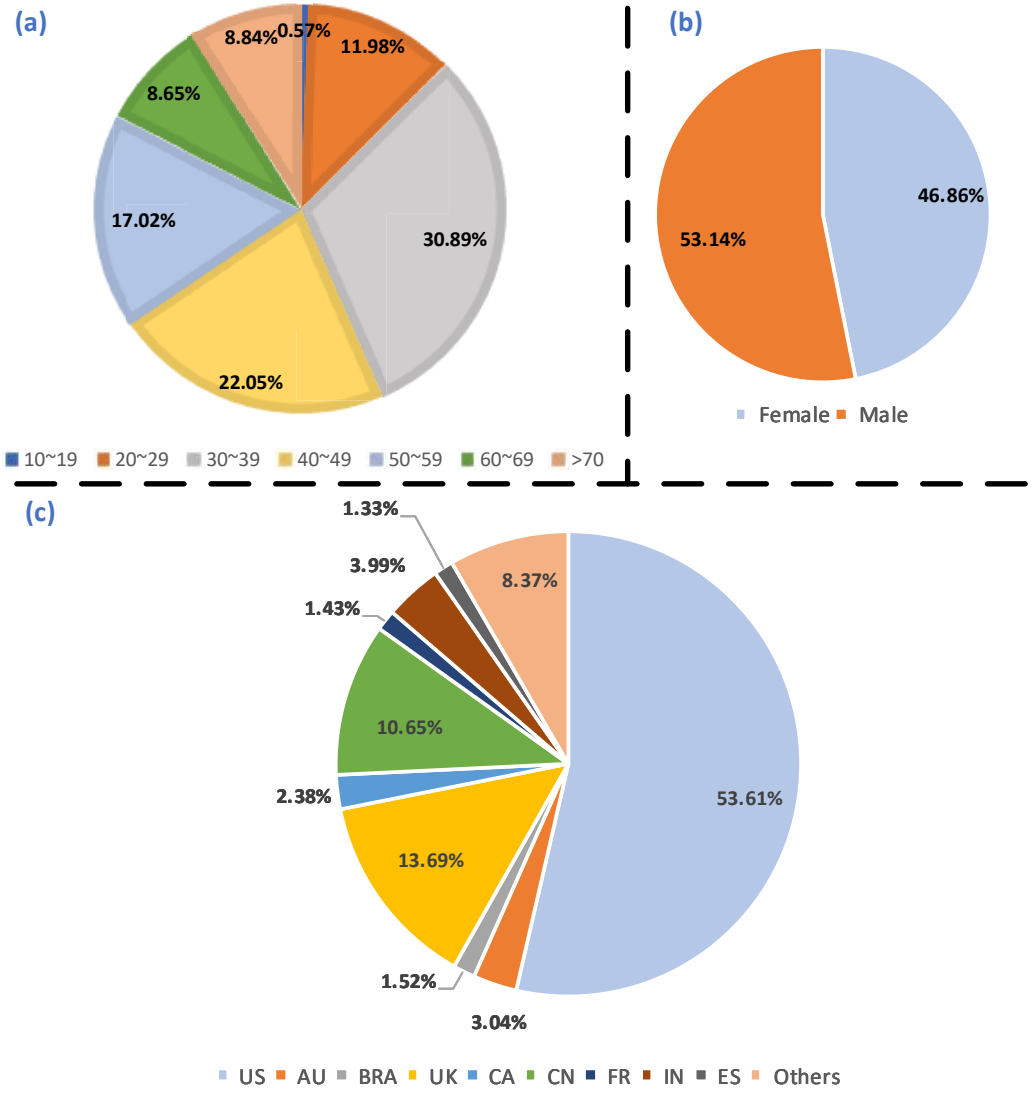


Figure 5.4 : Statistic information of the proposed Celeb-reID dataset. (a), (b), and (c) respectively show the distributions of age, gender, and nationality.

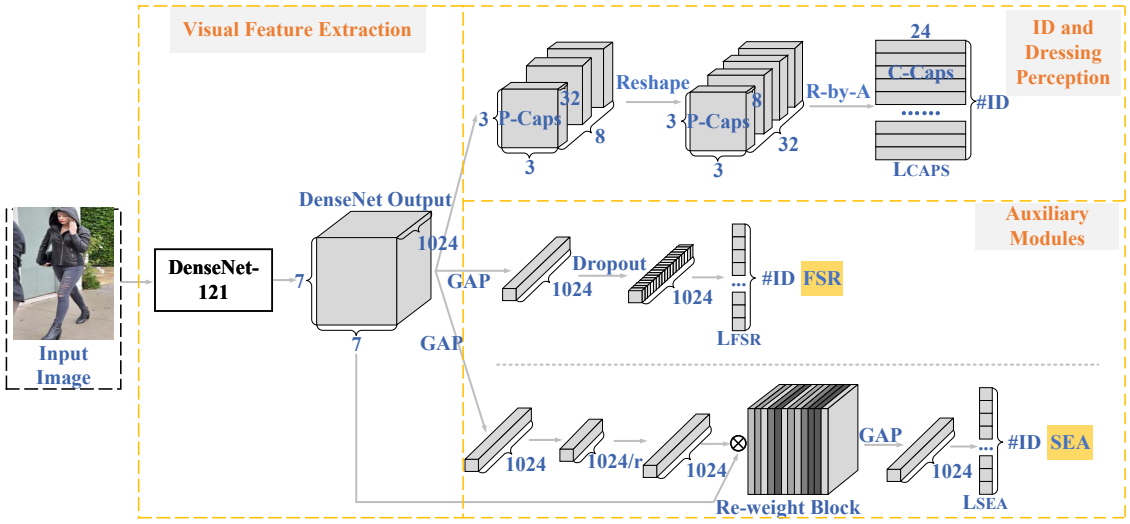


Figure 5.5 : Architecture of the proposed ReIDCaps network. Given an input image, an ImageNet-trained CNN backbone network (*i.e.*, DenseNet-121 (Huang et al., 2017a)) is used to extract low-level visual features. The output of the backbone network is fed to three branches, including capsule modules (ID and dressing perception), FSR and SEA (two auxiliary modules).

5.3 Vector-Neuron Capsules for Clothing Change LT-reID

The proposed ReIDCaps model is divided into three modules: **1) Visual feature extraction module**: the ImageNet-trained model is adopted to extract the low-level visual features of each image. These features are then fed into the capsule layers to perceive the ID and dressing information. This step follows the common practice of Capsule network in (Sabour et al., 2017) which uses several CNN layers to extract visual features of an image before these features forwarding to the capsule layers. **2) ID and dressing perception module**: The capsule layers are adopted to perceive the ID and dressing information of images. The length of each VN mainly perceives the ID information. The dressing information of each person is perceived by the orientation of each VN. **3) Auxiliary module**: The auxiliary modules are used to improve the discriminability of features learned from the proposed ReIDCaps

modules further. The architecture of the proposed ReIDCaps network is shown in Fig. 5.5. This section will introduce each module in details.

5.3.1 Visual Feature Extraction Module

Given an input image I_n^x, where n is the index of an ID, x represents the x-th image of ID n, an ImageNet-trained CNN backbone network is used to extract low-level visual features from the I_n^x. Densenet-121 (Huang et al., 2017a) is selected as the backbone network*. The output of the backbone is denoted as $\mathcal{O}(I_n^x) \in \mathbb{R}^{7\times7\times1024}$. The rest parts of the proposed ReIDCaps are divided into three branches: Capsule layers (main), the FSR, and the SEA. The losses of the three branches are denoted as $\mathcal{L}_{CAPS}$, $\mathcal{L}_{FSR}$, and $\mathcal{L}_{SEA}$ respectively. Amongst the three branches, FSR and SEA are two auxiliary branches. Both FSR and SEA dedicate to enhancing the performance of CNN-based visual feature. Then, these well-learned features can be utilized by the VN-based layers in a more efficient way. The objective function of the proposed ReIDCaps network is given as follows:

$$\mathcal{L} = \mathcal{L}_{CAPS} + \gamma * (\mathcal{L}_{FSR} + \mathcal{L}_{SEA}), \tag{5.1}$$

where γ is used to balance the weight between contributions raised from capsule layers and auxiliary modules. The three different branches are respectively introduced in details:

5.3.2 ID and Dressing Perception Module

The Capsule layers. VN capsules are used to learn the change of clothing. In training, the length of vector capsule $C_{\mathcal{L}}$ is used to reflect the likelihood of an existing people ID while its orientation $C_{\mathcal{O}}$ represents different types of clothes of the same people ID. To achieve this, two different capsule-based layers are adopted after

*Other alternatives also can be used.

$\mathcal{O}(I_n^x)$, including a Primary Capsules (P-Caps) layer and a Classification Capsules (C-Caps) layer (Sabour et al., 2017). Both layers are proposed in (Sabour et al., 2017) for digital recognition. This work changes the parameter setting on both layers to adapt the re-ID task and the ImageNet-trained CNN architecture. Specifically, given $\mathcal{O}(I_n^x)$, eight 32-channel convolutional operations (kernel size 2×2 and stride 2) are used to construct the P-Caps. Then, a reshaped operation is used to concatenate the corresponding channel of each block (8 blocks in total) in P-Caps. After that, 288 ($3 \times 3 \times 32$) VN capsules can be achieved with 8D in P-Caps layer. Finally, a non-linear Squashing Function is used to ensure the length of each VN capsule being normalized:

$$v_k^{8D} = \frac{\left\|v_k^{8D}\right\|^2}{1 + \left\|v_k^{8D}\right\|^2} \cdot \frac{v_k^{8D}}{\left\|v_k^{8D}\right\|}, \tag{5.2}$$

where v_k^{8D} represents k-th 8D VN capsule in P-Caps, $k \in [1, 288]$.

The C-Caps layer is followed by the P-Caps layer. There are N ID capsules in the C-Caps layer; N represents the number of ID in the training set. Each ID capsule in C-Caps is the combination of all VN capsules in the P-Caps layer. Given an 8D VN v_k^{8D} in P-Caps, its dimension is mapped to 24D by:

$$v_k^{24D} = W_k \cdot v_k^{8D}, \tag{5.3}$$

where $W_k \in \mathbb{R}^{24\times8}$ is a weight matrix; v_k^{24D} is a 24D VN capsule after mapping. Then an ID capsule (C_n^{24D}) in the C-Caps layer can be calculated by:

$$C_n^{24D} = \sum_{k=1}^{K} u_k^n \cdot v_k^{24D}, \tag{5.4}$$

where $n \in [1, N]$, u_k^n represents the coupling coefficient which is determined by a Routing-by-Agreement (R-by-A) process between the P-Caps and the C-Caps layers. Notably, all the C_n^{24D} are normalized by the Squashing Function (refer to Eq. 5.2).

The R-by-A process is a key technique to build the relationship between the P-Caps and C-Caps layers. The details of R-by-A can refer to (Sabour et al., 2017).

The R-by-A process is similar to (Sabour et al., 2017). The only difference is that the number of routing iteration is set to four rather than three in (Sabour et al., 2017). Four iterations can achieve better re-ID accuracy in experiments.

Given an input image I_n^x, the Margin Loss is used for ID existence:

$$\begin{aligned} L_{CAPS} = \sum_{n=1}^{N} \{ & y_n \cdot max(0, m^+ - \left\| C_n^{24D} \right\|)^2 \\ & + \lambda \cdot (1 - y_n) \cdot max(0, \left\| C_n^{24D} \right\| - m^-)^2 \}, \end{aligned} \tag{5.5}$$

where y_n represents the existence of ID for the input image I_n^x; $y_n = 1$ if I_n^x belongs to ID n, otherwise $y_n = 0$. $\lambda = 0.5$ is used to balance the weight between the two parts in L_{CAPS}. m^+ and m^- are used to control the length of C_n^{24D}. If the ID is present in I_n^x, the length of C_n^{24D} should be long, otherwise it should be short. The $m^+ = \frac{N-1}{N}$ and $m^- = \frac{1}{N}$.

5.3.3 Auxiliary Module

FSR and SEA mechanisms. Before formally introducing the two mechanisms, the Global Average Pooling (GAP) is adopted to the output of the backbone (*i.e.*, $\mathcal{O}(I_n^x)$) to obtain a CNN representation $\mathcal{F}(I_n^x)$ of the input image I_n^x:

$$\mathcal{F}(I_n^x) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{O}(I_n^x)(i, j), \tag{5.6}$$

where H and W respectively represent the height and width of $\mathcal{O}(I_n^x)$ ($H = 7$ and $W = 7$). After GAP, the input of FSR and SEA is denoted as $\mathcal{F}_{FSR}(I_n^x)$ and $\mathcal{F}_{SEA}(I_n^x)$, respectively.

The FSR mechanism is used to enhance the generalization power of the CNN-based output to affect the learning capability of the whole network. As is well-known, Dropout has demonstrated its effectiveness to prevent the CNN network from over-fitting by randomly dropping out neurons (set the value to zero) in network training (Srivastava et al., 2014). The Dropout is simply adopted to make the $\mathcal{F}_{FSR}(I_n^x)$

to become a sparse representation. The generalization capability of $\mathcal{F}_{FSR}(I_n^x)$ can be transferred to $\mathcal{O}(I_n^x)$ in backward propagation, which potentially improves the robustness of the whole network (particularly for all branches after $\mathcal{O}(I_n^x)$). The dropout rate is set to 0.75 in FSR.

The SEA mechanism aims to boost the discriminative power of the backbone network by applying the Squeeze-and-Excitation (SE) block (Hu et al., 2018b) on $\mathcal{O}(I_n^x)$. By doing so, informative features can be substantially highlighted and the useless information can be suppressed in $\mathcal{O}(I_n^x)$ which is the key link between the CNN-based architecture and the capsule layers. In this way, the prior knowledge learned by Densenet-121 can be used in a more efficient way when the knowledge is fed into the Caps Modules. The GAP after $\mathcal{O}(I_n^x)$ is used to obtain $\mathcal{F}_{SEA}(I_n^x)$. Then, $\mathcal{F}_{SEA}(I_n^x)$ is squeezed by a fully-connected layer to obtain a low-dimensional ($1024/r$-dim) representation, where r is the reduction rate. After that, another 1024-dim fully-connected layer ($\mathcal{F}'_{SEA}(I_n^x)$) is linked after the $1024/r$-dim layer. The final re-weight block is obtained by rescaling $\mathcal{O}(I_n^x)$ with $\mathcal{F}'_{SEA}(I_n^x)$:

$$\mathcal{O}'(I_n^x) = \mathcal{F}'_{SEA}(I_n^x) \otimes \mathcal{O}(I_n^x), \tag{5.7}$$

where $\otimes$ represents channel-wise multiplication between $\mathcal{F}'_{SEA}(I_n^x)$ and $\mathcal{O}(I_n^x)$. In SEA, $r = 16$ (the same as (Hu et al., 2018b)). Relu and Sigmoid activation functions are respectively used after the two fully-connected layers followed by $\mathcal{F}_{SEA}(I_n^x)$.

Finally, Cross-Entropy loss is used on both FSR ($\mathcal{L}_{FSR}$ in Eq. 5.1) and SEA ($\mathcal{L}_{SEA}$ in Eq. 5.1) branches to classify the ID of persons in the training set.

5.4 Experiments

In this section, four different datasets used for evaluation are introduced. Then, an experimental setup is given. An ablation study of ReIDCaps is provided. Moreover, quantitative and qualitative analyses will be given to verify the effectiveness

of VN capsules comparing with the traditional SN. Finally, comprehensive evaluations are carried out by comparing the proposed methods with other state-of-the-art methods. The experiments are mainly conducted on Celeb-reID since the proposed method is particularly designed to tackle the clothing change challenges exposed in the LT-reID scenario.

5.4.1 Datasets for Evaluations

Four different datasets are used to evaluate the proposed method. Two of them are LT-reID datasets. Another two datasets are ST-reID datasets (*i.e.*, Market1501 (Zheng et al., 2015) and DukeMTMC-reID (Zheng et al., 2017b)).

Celeb-reID is introduced in Sec. 5.2. This dataset is used for the clothing-change LT-reID performance evaluation. Details of Celeb-reID can refer to Tab. 5.1. Notably, a person may wear the same clothing (the ratio is less than 30% within each ID) in Celeb-reID.

Celeb-reID-light is a light version of Celeb-reID. Unlike Celeb-reID, a person in Celeb-reID-light appears in different images always with different clothes (Huang et al., 2019a). The images in Celeb-reID-light is picked up from the candidate image set after data filtering (see Fig. 5.2). There are 590 persons. 490 IDs with 9,021 images are used for training, and 100 IDs with 1,821 images are used for testing. In the testing set, 887 images are used as queries, and 934 images are used as galleries. Celeb-reID-light is used to testify the robustness of re-ID methods when a person never wears the same clothing.

Market1501 is one of the widely used ST-reID dataset. There are 751 (12,936) and 750 (19,732) IDs (images) in training and testing sets respectively.

DukeMTMC-reID is another widely used ST-reID dataset. There are 702 IDs in the training and testing sets respectively. The number of training images is

16,522, while the number of testing images is 19,889.

5.4.2 Experimental Setup

The PyTorch package is used to implement the proposed ReIDCaps network. To be consistent, the architecture of the proposed ReIDCaps network is not changed, but use different training strategies for ST-reID and LT-reID scenarios respectively. This is because, for LT-reID, the new model needs to produce more learnable parameters to accommodate additional information about the clothing change. Thus, it may cause certain overfitting issues when it is applied to a relatively simple case of ST-reID where there is no clothing change. For the LT-reID scenario, the initial learning rate is set to 1e-4 in the ImageNet-trained DenseNet-121 and 1e-3 in new layers after DenseNet output (see Fig. 5.5). This training strategy is named 'TS1' (short for Training Strategy 1). In order to mitigate the possible overfitting issue, for the ST-reID scenario, the backbone network (DenseNet-121) is pretrained on common SN (refer to Fig. 5.8). The SN-trained DenseNet-121 backbone network is adopted in ReIDCaps. The initial learning rate is set to 1e-5 in the SN-trained DenseNet-121 and 1e-4 in new layers. This training strategy is named 'TS2' (short for Training Strategy 2).

For both scenarios, all input images are resized to 224×224 and randomly flipped before training. The Adam stochastic optimization (Kingma and Ba, 2014) is used with parameters $\beta 1 = 0.9$, $\beta 2 = 0.999$. The learning rate is decayed by 10 times after 40 training epochs, and the training stops after 50-th epochs. The number of training IDs $N = 632$, 490, 751, and 702 for Celeb-reID, Celeb-reID-light, Market1501, and DukeMTMC-reID respectively.

In testing, the same procedure as previous works (*e.g.*, (Yu et al., 2017b; Zheng et al., 2018; Sun et al., 2018; Guanshuo et al., 2018)) is adopted. Eq. 5.6 is used to extract 1,024D features as the person's description. This feature is used to compare

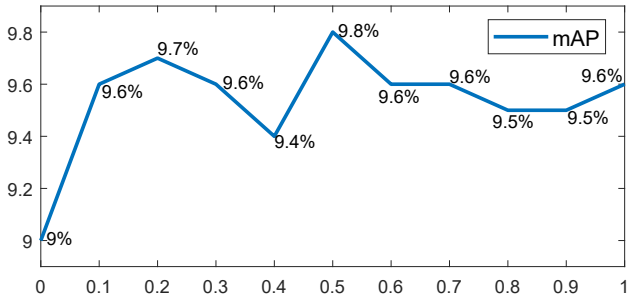


Figure 5.6 : Sensitivity to parameter γ in Eq. 5.1. The x-axis and y-axis respectively represents the γ and mAP. Experiment is conducted on Celeb-reID.

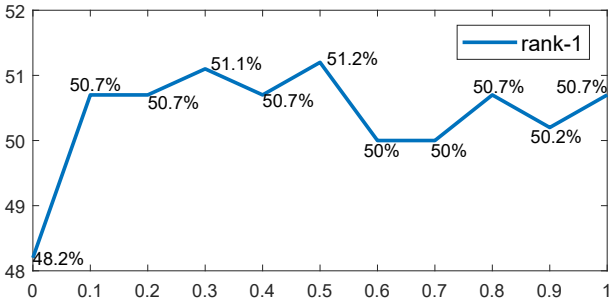


Figure 5.7 : Sensitivity to parameter γ in Eq. 5.1. The x-axis and y-axis respectively represents the γ and rank-1 accuracy. Experiment is conducted on Celeb-reID.

the similarity between any two images in query and gallery sets using the Euclidean distance. The standard rank-N and mAP re-ID evaluation protocols are used in experiments. The single query setting is adopted over four datasets in testing.

5.4.3 Ablation Study of ReIDCaps

An ablation study is given to verify some important settings of ReIDCaps. The proposed Celeb-reID dataset is used.

Hyper-parameter Setting

The parameter γ in Eq. 5.1 is evaluated. The results are shown in Fig. 5.6 and Fig. 5.7, respectively. When $\gamma = 0$ (*i.e.*, without the help of auxiliary modules), the

Table 5.2 : Ablation study of the proposed ReIDCaps network. mAP and rank-N (N=1, 5, and 10) are listed. The best performance is highlighted in **Blod**.

Methods	Celeb-reID			
	mAP	rank-1	rank-5	rank-10
$\text{Caps}_{iter=3,C_n^{16D}}$	7.8%	43.8%	59.2%	67.5%
$\text{Caps}_{iter=3}$	8.7%	46.2%	62.1%	68.9%
$\text{Caps}_{iter=4}$	9.0%	48.2%	62.7%	70.0%
$\text{Caps}_{iter=5}$	8.7%	47.5%	62.4%	69.6%
$\text{Caps}_{iter=4}$+0.5*FSR	9.2%	49.1%	63.1%	70.4%
$\text{Caps}_{iter=4}$+0.5*SEA	9.5%	49.8%	63.8%	70.8%
$\text{Caps}_{iter=4}$+0.5*(FSR+SEA)	**9.8%**	**51.2%**	**65.4%**	**71.9%**

proposed ReIDCaps reduces to the baseline. It can be observed that the proposed method achieves the best performance on both mAP and rank-1 accuracy when the $\gamma = 0.5$. Therefore, $\gamma = 0.5$ in the experiment setting.

Number of Iterations by R-by-A

The Caps modules and the backbone network of ReIDCaps are used to search a proper number of iterations between P-Caps and C-Caps by R-by-A (refer to Sec. 5.3.2). In the original capsule network design (Sabour et al., 2017), the R-by-A algorithm uses 3 iterations between the P-Caps and C-Caps. This setting is also testified in the proposed ReIDCaps network (see $\text{Caps}_{iter=3}$ in Tab. 5.2). When the length of C-Caps capsules is 16 ($\text{Caps}_{iter=3,C_n^{16D}}$, as the same setting in (Sabour et al., 2017)) instead of 24 (see Fig. 5.5), a baseline performance can be achieved (mAP: 7.8%, rank-1: 43.8%). After changing the length of capsules from 16 to 24, the performance can be improved from 43.8% to 46.2% in rank-1 accuracy. Thus, the length of C-Caps capsules is set to 24 in the follow-up experiments. It can be observed that compared with $\text{Caps}_{iter=3}$, better re-ID accuracy can be achieved

Table 5.3 : Ablation study of different training strategies on the proposed ReIDCaps model over two different person re-ID scenarios.

Training Strategies	mAP	rank-1	rank-5	rank-10
Celeb-reID (LT-reID)				
TS1	9.8%	51.2%	65.4%	71.9%
TS2	9.5%	50.3%	64.9%	71.7%
Market1501 (ST-reID)				
TS1	62.5%	84.8%	94.7%	98.1%
TS2	72.7%	89.0%	95.5%	96.9%

when the number of iteration is set to 4 ($\text{Caps}_{iter=4}$, mAP=9.0%, rank-1=48.2%). It can be clear to see that when iter=5 ($\text{Caps}_{iter=5}$), the performance is dropped again. Therefore, the number of iteration is set to 4 in the R-by-A algorithm.

Combination of Different Modules

Different combinations of auxiliary modules (*i.e.*, FSR and SEA) are also used to verify the effectiveness of them in enhancing the overall performance, shown in Tab. 5.2. By combining FSR (SEA) (with the weight is 0.5, refer to Sec. 5.4.3) with $\text{Caps}_{iter=4}$, the re-ID accuracy is improved from 48.2% to 49.1% (49.8%) in rank-1 accuracy. This combination verifies the effectiveness of different auxiliary modules. Moreover, both FSR and SEA are combined with the weight of 0.5. The rank-1 accuracy gains 3% comparing with single $\text{Caps}_{iter=4}$ module. This result verifies the combination of auxiliary modules (*i.e.*, FSR and SEA) can be useful in enhancing the overall performance.

Comparison of Different Training Strategies

In the experimental setup, two different training strategies are used in the LT-reID scenario and the ST-reID scenario, respectively (see Sec. 5.4.2). The two train-

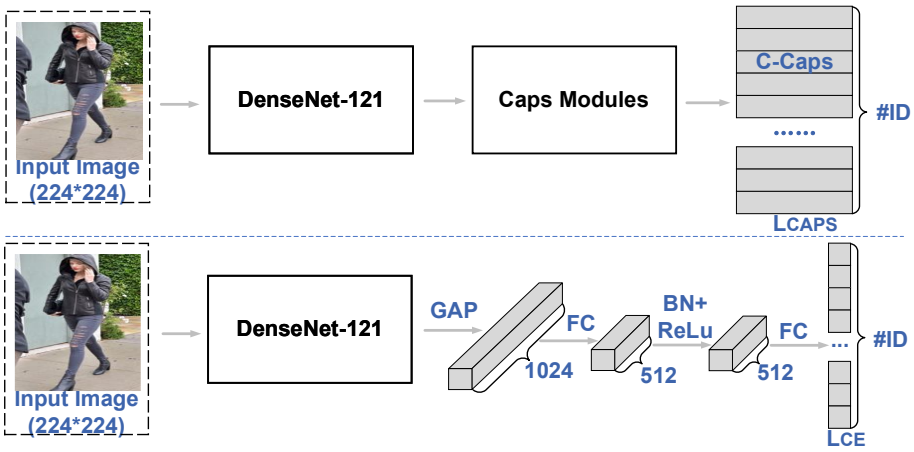


Figure 5.8 : The SN-based and VN capsule networks. The upper network uses VN capsules. The lower one uses the tradition CNN layers. Both networks use the same input image size and backbone network (*i.e.*, the DenseNet-121). FC, BN, ReLU, and L_{CE} respectively represent the fully-connected layer, batch normalization, ReLu activation function, and the Cross-Entropy loss.

ing strategies are used on different re-ID scenarios to verify the function of them. It can be observed in Tab. 5.3 that, compared with TS1, TS2 can improve the performance on Market1501 noticeably. This result verify that using the SN-trained backbone can effectively mitigate the overfitting problem produced by more learnable parameters from VN. On the contrary, when different training strategies are applied for the case of LT-reID, the performance is barely affected. This comparison shows that the proposed method is robust to the LT-reID scenario under different training strategies. Although Celeb-reID also achieves comparable performance by TS2 which is more suitable for the ST-reID scenario, TS1 is used for Celeb-reID since Celeb-reID and Market1501 achieve the best performance under TS1 and TS2, respectively.

5.4.4 Scalar Neuron *vs.* Vector Neuron with Quantitative and Qualitative Analyses

Network Architecture. The VN capsules are used to compare with SN solution. Network architectures are given in Fig. 5.8. Both approaches share the same input and backbone network (*i.e.*, ImageNet-trained DenseNet-121). The VN capsules (Caps modules in Fig. 5.2) and SN-based layers are followed by the output of DenseNet-121 respectively. The SN-based network (also called as IDE+ (Hermans et al., 2017; Sun et al., 2017)) is a widely used benchmark for ST-reID scenario. The result of SN-based IDE+ is achieved by the Deep-Person-reID package[†].

Table 5.4 : Performance comparison between the SN-based IDE+ model (the lower network in Fig. 5.8) and the proposed $\text{Caps}_{iter=4}$ model (the upper network in Fig. 5.8).

Methods	Celeb-reID			
	mAP	rank-1	rank-5	rank-10
IDE+	5.9%	42.9%	56.4%	63.4%
$\text{Caps}_{iter=4}$	**9.0%**	**48.2%**	**62.7%**	**70.0%**

Quantitative Evaluation. Tab. 5.4 shows the quantitative evaluation between the proposed $\text{Caps}_{iter=4}$ and the IDE+ models. By using the VN capsules, the proposed $\text{Caps}_{iter=4}$ outperforms the IDE+ by +3.1% (9.0% *vs.* 5.9%) on mAP and +5.3% (48.2% *vs.* 42.9%) in rank-1 accuracy. It is clear to see that the VN capsules show their superiority in dealing with the clothing change challenge for person re-ID. This is because in addition to distinguish the inter-class variation, VN capsule can potentially perceive the intra-class variation when clothes of the same person are changed.

[†]https://github.com/KaiyangZhou/deep-person-reid

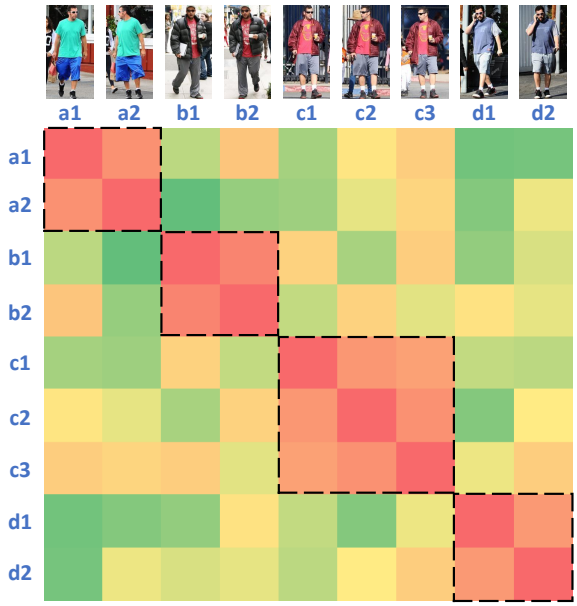


Figure 5.9 : Intra-class variation visualization using C-Caps. Four types of clothing (9 images) belonging to the same person (76 images in total) are selected in the training set of Celeb-reID. 'a' to 'd' represent different clothes and '1, 2, ...' represents the index of sample images. The cosine similarity is used to calculate the similarity between two images using the VN capsules in C-Caps where the ID is presented. An activation map is used to represent the similarity between any two images. The red and green colors respectively represent the most and the least similar pairs. Elements in the diagonal are self-similar.

Qualitative Evaluation. The intra-class variation of the same people ID using the proposed $\text{Caps}_{iter=4}$ is visually illustrated. The cosine similarity is used to calculate the orientation distance between two images. Specifically, given two images (*e.g.*, $I_{n_{id}}^{x_1}$ and $I_{n_{id}}^{x_2}$) belonging to the same ID (e.g., n_{id}), the VN capsules can be achieved for this ID (*i.e.*, $C_{n_{id}}^{24D}(I_{n_{id}}^{x_1})$ and $C_{n_{id}}^{24D}(I_{n_{id}}^{x_2})$) by Eq. 5.4 (after training is completed) to calculate the cosine distance between the two 24D VNs. An activation map of similarity is used to illustrate the effectiveness of the proposed Caps Modules in perceiving the knowledge of intra-class clothing changes. It can be observed in

Fig. 5.9 that the proposed $\text{Caps}_{iter=4}$ can potentially learn the clothing changes of the same person. The same types of clothes show higher similarity than different types to some extent. This result demonstrates that the orientation of VN can carry the information of clothing change of the same person for the LT-reID scenario.

5.4.5 Comparison with State-of-The-Art Methods

The proposed ReIDCaps method is compared with several state-of-the-art re-ID approaches. The proposed method is mainly designed for the LT-reID scenario. It may not be a universal solution to both ST-reID scenario and LT-reID scenario. However, it is also evaluated on the case of ST-reID to verify its robustness under different re-ID scenarios. Tab. 5.5 lists the results. Four different coarse-grained re-ID approaches (*i.e.*, IDE+ (Hermans et al., 2017; Sun et al., 2017), ResNet-Mid (Yu et al., 2017b), Two-Stream (Zheng et al., 2018), MLFN (Chang et al., 2018)), and four different fine-grained re-ID approaches (*i.e.*, HACNN (Li et al., 2018b), Part-Bilinear (Yumin et al., 2018), PCB (Sun et al., 2018), and MGN (Guanshuo et al., 2018)) are picked up to evaluate the challenges exposed in the LT-reID scenario. Coarse-grained methods leverage the whole image as inputs while fine-grained methods take both global information and body parts into consideration. All the methods used in comparison are recently published methods which show promising performance in the traditional ST-reID scenario. In these methods, the IDE+, ResNet-Mid, MLFN, and HACNN are implemented by Deep-Person-Reid package (Zhou and Xiang, 2019) with solid performance. The implementation of rest methods are based on the resources released by the authors of original papers (*i.e.*, Two-Stream, Part-Bilinear, PCB, and MGN).

Comparison with coarse-grained methods. It can be observed in Tab. 5.5, by only using the global body cue, the proposed method outperforms other coarse-grained methods on both Celeb-reID and Celeb-reID-light datasets. Compared with

Figure 5.10 : Body parts partitions. The whole image is denoted as G. P_{11}, P_{12}, and P_{13} (also P_{21} and P_{22}) are parts equally divided from G.

the second best method ResNet-Mid, the proposed ReIDCaps outperforms it on Celeb-reID by a large margin in rank-1 accuracy (51.2% *vs.* 43.3%). The best performance is also achieved on Celeb-reID-light (mAP: 11.2%, rank-1: 20.3%). This result verifies the proposed ReIDCaps can best tackle the extreme case where a person does not wear the same clothing twice. MLFN achieves the best performance on Market1501 (mAP: 74.3%, rank-1: 90.0%). However, it only obtains 6.0% mAP and 41.4% rank-1 accuracy on Celeb-reID. Although the proposed method is not designed for the ST-reID scenario, it also experiments on Market1501 (DukeMTMC-reID) and obtains 89.0% (81.2%) in rank-1 accuracy, which is competitive comparing with other coarse-grained approaches.

Comparison with fine-grained methods. To compare with fine-grained learning approaches, the body parts are utilized in the experiments. Following the same setting in (Guanshuo et al., 2018), a person image is equally divided into three and two parts, respectively. Fig. 5.10 shows the partition result. The proposed ReIDCaps is simply trained and tested on five body parts respectively. Tab. 5.5 shows the re-ID accuracy using different parts on four re-ID datasets. The LT-reID datasets such as Celeb-Reid and its light version involve appearance changes dramatically. Parts such as P_{12}, P_{13}, and P_{22} show much lower re-ID accuracy than P_{11}

Table 5.5 : Comparison with state-of-the-art methods. The best result is shown in **bold**. Rank-N accuracy and mAP are listed.

Methods	LT-reID						ST-reID				
	Celeb-reID (New)			Celeb-reID-light (New)			Market1501			DukeMTMC-reID	
	mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1		mAP	rank-1
Coarse-Grained Learning Approaches (without human body parts partition)											
IDE+ (DenseNet-121)	5.9%	42.9%	56.4%	5.3%	10.5%	24.8%	68.0%	87.8%		57.5%	77.9%
ResNet-Mid (Yu et al., 2017b) ArXiv17	5.8%	43.3%	54.6%	6.0%	10.3%	28.0%	75.6%	89.9%		**64.0%**	**81.6%**
Two-Stream (Zheng et al., 2018) TOMM18	7.8%	36.3%	54.5%	-	-	-	60.9%	80.3%		51.4%	72.6%
MLFN (Chang et al., 2018) CVPR18	6.0%	41.4%	54.7%	6.3%	10.6%	31.0%	**74.3%**	**90.0%**		63.2%	81.1%
ReIDCaps (Ours)	**9.8%**	**51.2%**	**65.4%**	**11.2%**	**20.3%**	**48.2%**	72.7%	89.0%		62.6%	81.2%
Fine-Grained Learning Approaches (with human body parts partition)											
HACNN (Li et al., 2018b) CVPR18	9.5%	47.6%	63.3%	11.5%	16.2%	42.8%	75.7%	91.2%		63.2%	80.1%
Part-Bilinear (Yumin et al., 2018) ECCV18	6.4%	19.4%	40.6%	-	-	-	74.5%	88.8%		64.2%	82.1%
PCB (Sun et al., 2018) ECCV18	8.2%	37.1%	57.0%	-	-	-	77.4%	92.3%		66.1%	81.8%
MGN (Guanshuo et al., 2018) ACMMM18	10.8%	49.0%	64.9%	13.9%	21.5%	47.4%	**86.9%**	**95.7%**		**78.4%**	**88.7%**
ReIDCaps* (Ours)	**15.8%**	**63.0%**	**76.3%**	**19.0%**	**33.5%**	**63.3%**	78.0%	92.8%		67.8%	83.8%
Performance on Different Body Part using ReIDCaps											
ReIDCaps: P_{11}	11.5%	53.6%	69.2%	14.4%	24.5%	54.9%	29.2%	56.2%		38.7%	59.4%
ReIDCaps: P_{12}	3.6%	33.7%	43.2%	3.5%	5.3%	19.2%	40.1%	65.3%		34.5%	56.6%
ReIDCaps: P_{13}	3.7%	32.3%	43.2%	4.0%	7.5%	22.4%	22.9%	41.6%		20.8%	35.0%
ReIDCaps: P_{21}	10.2%	50.9%	66.5%	13.1%	25.3%	51.8%	37.7%	65.4%		45.0%	66.6%
ReIDCaps: P_{22}	4.3%	36.0%	48.0%	4.4%	8.5%	23.8%	45.5%	67.3%		32.9%	52.2%

Table 5.6 : Performance by using different weights on different body parts. The part partition can refer to Fig. 5.10. The result of Celeb-reID is mainly evaluated. The weight assigned to Celeb-reID-light is similar to Celeb-reID since both belong to the LT-reID scenario. Another group of weights on Market1501 are used by considering the contribution of different body parts.

Methods	Celeb-reID		
	mAP	rank-1	rank-5
P_{11}+P_{12}+P_{13}+P_{21}+P_{22}	13.6%	60.7%	74.2%
P_{11}+P_{12}+P_{13}+P_{21}+P_{22}+G	14.2%	62.2%	75.2%
P_{11}+0.9*(P_{12}+P_{13})+P_{21}+0.9(P_{22})+G	14.9%	62.6%	75.4%
P_{11}+0.7*(P_{12}+P_{13})+P_{21}+0.7(P_{22})+G	15.4%	62.9%	76.3%
P_{11}+0.5*(P_{12}+P_{13})+P_{21}+0.5(P_{22})+G	**15.8%**	**63.0%**	**76.3%**
P_{11}+0.3*(P_{12}+P_{13})+P_{21}+0.3(P_{22})+G	15.8%	62.3%	75.7%
P_{11}+0.1*(P_{12}+P_{13})+P_{21}+0.1(P_{22})+G	15.4%	61.4%	75.1%
	Celeb-reID-light		
P_{11}+0.5*(P_{12}+P_{13})+P_{21}+0.5(P_{22})+G	**19.0%**	**33.5%**	**63.3%**
	Market1501		
0.25*(P_{11}+P_{12}+P_{13})+0.5(P_{21}+P_{22})+G	**78.0%**	**92.8%**	**97.6%**

and P_{21} where present more robust appearance cues, *i.e.*, the upper body. However, ST-reID dataset, *i.e.*, Market1501 and DukeMTMC-reID illustrate opposite results. The P_{11} and P_{21} on Market1501 even show lower re-ID accuracy comparing with other body parts. This is because, without substantial appearance changes, cues such as color and texture are more likely to be re-identified as significant factors. When clothing is completely changed (Celeb-reID-light), the appearance cues become even less reliable. However, with benefits of the proposed ReIDCaps network, even the most difficult part (P_{12}) can play a certain role (mAP: 3.5%, rank-1: 5.3%) in Celeb-reID-light.

As in (Guanshuo et al., 2018), the five body parts and the global body cue are integrated to get the final result (denoted as ReIDCaps*). Different weights are assigned to each part according to their contributions. Tab. 5.6 shows the performance. Considering the discriminability of different body parts, the weights of P_{11}, P_{21}, and G are simply set to 1 because the three parts are visually more recognizable than other parts (P_{12}, P_{13}, and P_{22}, refer to Fig. 5.10). Under this setting, the search space (*i.e.*, different combinations) can be greatly reduced. The same weight on parts P_{12}, P_{13}, and P_{22} is used, the weight varies from 0.1 to 1 (gap 0.2). When the weight equals to 0.5, the proposed method achieves the best performance. Since the Celeb-reID and Celeb-reID-light are all LT-reID datasets, the same setting is adopted. The weights on Market1501 and DukeMTMC-reID are different because each person in both datasets does not change clothing. With the help of color information on clothes, it is easier to recognize each person on Market1501 and DukeMTMC-reID with more body part information. Therefore, the weight of the whole body is simply set to 1, 0.5 for large body parts (P_{21} and P_{22}), and 0.25 for small body parts (P_{11}, P_{12}, and P_{13}).

It is clear to see in Tab. 5.5 that the proposed ReIDCaps* outperforms other methods by a large margin on Celeb-reID and Celeb-reID-light. It achieves 15.8% mAP and 63.0% rank-1 accuracy on Celeb-reID. The second best method MGN only achieves 10.8% mAP and 49.0% rank-1 accuracy. Even without using any body part cue, the proposed method (ReIDCaps) can outperform MGN in rank-1 accuracy (51.2% *vs.* 49.0%). In the meantime, the proposed method outperforms MGN on Celeb-reID-light by 12.0% in rank-1 accuracy (33.5% *vs.* 21.5%), which further verifies the effectiveness of the proposed method in dealing with the LT-reID scenario.

Robustness Evaluation. To further verify the robustness of re-ID approaches between the LT-reID scenario and the ST-reID scenario. In this chapter, a Ro-

Table 5.7 : Robustness score evaluation between LT-reID (Celeb-reID (C) or Celeb-reID-light (C-l)) and ST-reID (Market1501 (M) or DukeMTMT-reID (D)) re-ID scenarios using the RS (see Eq. 5.8). The mAP and rank-1 (r1) accuracy (from Tab. 5.5) are used as evaluation indexes to the RS.

Methods	RS(mAP)			
	C & M	C & D	C-l & M	C-l & D
MGN	0.17	0.17	0.20	0.20
ReIDCaps* (Ours)	**0.22**	**0.22**	**0.24**	**0.24**
Methods	RS(rank-1)			
	C & M	C & D	C-l & M	C-l & D
MGN	0.47	0.47	0.26	0.26
ReIDCaps* (Ours)	**0.59**	**0.60**	**0.35**	**0.35**

bustness Score (RS$\in [0, 1]$) is defined. The RS score is used to comprehensively quantify the robustness of a model when it is applied on both ST-reID and LT-reID scenarios. For instance, it is observed that MGN achieves the state-of-the-art re-ID performance on two ST-reID datasets (*i.e.*, Market1501 and DukeMTMC-reID). However, MGN is still hard to tackle the clothing change challenge in LT-reID when people can change their clothing. In order to get a quantitative comparison of the robustness between different re-ID approaches across different scenarios, the RS is defined as follows:

$$RS(\mathbb{P}) = \frac{\sqrt{score_L(\mathbb{P}) \times score_S(\mathbb{P})}}{|score_L(\mathbb{P}) - score_S(\mathbb{P})| + 1}, \tag{5.8}$$

where $score_L$ and $score_S$ represent the LT-reID and ST-reID accuracy respectively; $\mathbb{P} \in \{mAP, rank-N\}$ represents different evaluation indexes. Given a re-ID model, RS returns a large value if the score of both $score_L$ and $score_S$ are high and close to each other. Tab. 5.7 shows the robustness comparison results by using Eq. 5.8. It can observe that the proposed method has better robustness than MGN when the

Figure 5.11 : Pose estimation results (the bottom row). The head images (the top row) are extracted according to the location of keypoint of the neck.

Table 5.8 : Comparison between re-ID performance when only head images are used.

Methods	rank-1	mAP
Celeb-reID (head)		
Two-Stream (Zheng et al., 2018)	27.9%	11.6%
Market1501 (head)		
Two-Stream (Zheng et al., 2018)	29.0%	13.2%

evaluation is conducted on different scenarios.

Reliability of Head Information. Normally, person re-ID (in existing researches) would like to use overall body information rather than head/face information. This experiment clarifies that the proposed dataset does not present high-quality head/face information. Otherwise, any method of person re-ID developed on such dataset may have bias caused by strong head/face information. In order to verify this point, the 2D human pose estimation approach (Cao et al., 2017) is adopted to localize anatomical keypoints on the body of persons. According to the keypoint of the neck, the head part from the body can be extracted. As shown in Fig. 5.11, the head part (including the face) can be obtained from the Celeb-reID dataset and Market1501. It can be clear to see that, in terms of head information,

the proposed dataset shares the same characteristics and presumption (*i.e.*, head information not useful) as other existing datasets for person re-ID research such as Market1501. The Two-Stream (Zheng et al., 2018) model, which contains both identification and verification signals, is used to train and test the re-ID performance when only the head images are adopted. Tab. 5.8 shows the result. It shows that the quality of head information in the proposed dataset is even worse than Market1501. This is because celebrities in the Celeb-reID dataset have a great chance to wear sunglasses or hats, making these faces hard to recognize.

5.5 Conclusion

This chapter introduces a new LT-reID dataset called "Celeb-reID" to the community. This dataset uses the street snap-shots of celebrities as the resource. Compared with previous datasets, the proposed dataset is the largest re-ID dataset with the clothing change of the same people ID. A ReIDCaps model is designed to tackle the clothing change challenge. Compared with the common SN-based CNN, VN capsules are used to perceive the clothing change of the same person. The capsule layers and an ImageNet-trained CNN are integrated together for complex person re-ID data. Comprehensive experiments are given to demonstrate the superiority of the proposed method in the LT-reID scenario.

Chapter 6

Modality Bias Training Issue for Infrared-Visible LT-reID

6.1 Motivation

The task of person re-identification (re-ID) is to associate the IDs of different person images captured by different surveillance cameras. Currently, significant progress has been made in person re-ID under visible lighting conditions (Wan et al., 2020a; Song et al., 2018; Zheng et al., 2019; Ding et al., 2019; Wei et al., 2018b; Wang et al., 2015; Yu et al., 2019a; Luo et al., 2020b,a; Wang et al., 2019b; Zeng et al., 2020; Wang et al., 2019d). However, with the rapid development of person re-ID, a sophisticated system should have the capability to handle various cases such as cross-modality re-ID. As a typical and widely observed case, infrared-visible LT-reID has been brought into focus (Wu et al., 2017; Ye et al., 2018a; Dai et al., 2018; Ye et al., 2018b, 2019; Wang et al., 2019c), which uses different sensing cameras under different lighting conditions (*i.e.*, RGB cameras for good lighting conditions and IR cameras for poor lighting conditions).

To associate person images captured by different types of cameras (RGB and IR), infrared-visible LT-reID is studied to tackle a typical cross-modality issue (Wu et al., 2017). The main challenge of infrared-visible LT-reID is to tolerate the discrepancies between RGB and IR given objects with the same IDs (*i.e.*, ID-tied cross-modality image pairs) (Ye et al., 2018a,b). To achieve this, existing approaches learn shared features across modalities (Wu et al., 2017; Ye et al., 2018a; Dai et al., 2018; Ye et al., 2018b, 2019; Wang et al., 2019c). Amongst the existing methods, the ImageNet-

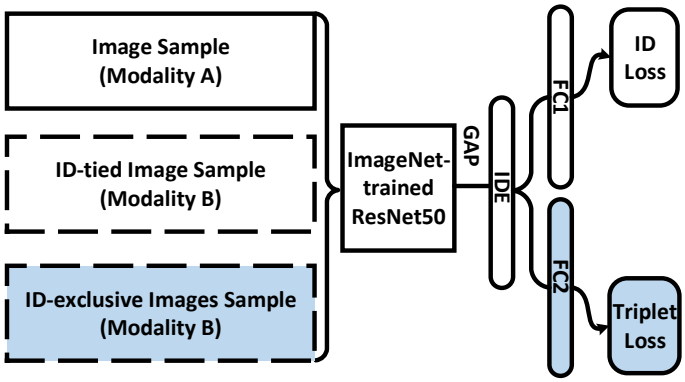


Figure 6.1 : Simplified network of existing infrared-visible LT-reID approaches. A and B represent different modalities. If A=RGB, then B=IR and vice versa. The ID label is the same (different) between an image sample from modality A and the ID-tied (ID-exclusive) image sample. The GAP, IDE, and FC are short for the Global Average Pooling, ID-discriminative Embedding, and Fully Connected layers, respectively.

trained ResNet50 is widely used and has demonstrated a compelling performance (Ye et al., 2018a; Dai et al., 2018; Ye et al., 2018b, 2019; Wang et al., 2019c). In addition, ImageNet-trained ResNet50 is also prevalent in state-of-the-art conventional re-ID approaches (Zheng et al., 2019; Yu et al., 2019a; Luo et al., 2020a,b; Wan et al., 2020a), which demonstrates its vital role in drafting re-ID approaches. However, the performance of ImageNet-trained ResNet50 on the infrared-visible LT-reID task is not promising.

To learn shared features for infrared-visible LT-reID, existing approaches either adopt feature-level constraints (Wu et al., 2017; Ye et al., 2018a; Dai et al., 2018; Ye et al., 2018b, 2019; Hao et al., 2019) or add additional constraints provided by certain image generation processes through Generative Adversarial Network (GAN) (Wang et al., 2019c,a, 2020). These approaches are illustrated in Fig. 6.1. The simplified model is based on the classic ID-discriminative Embedding (IDE) (Zheng et al.,

Table 6.1 : Building blocks of existing infrared-visible LT-reID approaches. I, I_p, and I_n respectively represent input samples, ID-tied image samples, and ID-exclusive image samples in a training minibatch.

Methods	Inputs	Backbone	Extra Functions
HCML (Ye et al., 2018a)	(I,I_p)	IDE	contrastive loss
BDTR (Ye et al., 2018b)	(I,I_p)	IDE	top-ranking loss
eBDTR (Ye et al., 2019)	(I,I_p)	IDE	center-constrained loss
D-HSME (Hao et al., 2019)	(I,I_p)	IDE	KL loss
cmGAN (Dai et al., 2018)	(I,I_p,I_n)	IDE-T	modality classification loss
D^{2}RL (Wang et al., 2019c)	(I,I_p,I_n)	IDE-T	data-level transfer
AlignGAN (Wang et al., 2019a)	(I,I_p,I_n)	IDE-T	data-level transfer
JSIA-ReID (Wang et al., 2020)	(I,I_p,I_n)	IDE-T	data-level transfer

2016a) person re-ID model (ImageNet-trained ResNet50 + identification (ID) loss) (*e.g.*, (Ye et al., 2018a,b, 2019)) or with an additional triplet loss to pull (push) the instance of the same (different) ID closer (farther) (*e.g.*, (Dai et al., 2018; Wang et al., 2019c)). The IDE network with triplet loss is named IDE-T in this work. Tab. 6.1 lists the building blocks of existing cutting-edge infrared-visible LT-reID approaches. These approaches are trained on the basic IDE or IDE-T backbone model shown in Fig. 6.1 with some extra functions.

This work observes that existing infrared-visible LT-reID approaches which normally adopt ImageNet-trained ResNet50 (He et al., 2016) to learn the shared features of ID-tied cross-modality image pairs can lead to an undesired Modality Bias Training (MBT) issue. Consequently, the learned shared features cannot be used to acquire the true characteristics that are not biased to either modality. Instead, the learned shared features will contain more information from the RGB modality due to the significant differences in training data between the RGB modality and

IR modality. For instance, when two images from different modalities need to be classified into the same person ID, ImageNet-trained ResNet50 is likely to learn the discriminative features from its familiar modality (*i.e.*, RGB). Thus, the information from its unfamiliar modality (*i.e.*, IR) can be overwhelmed by the RGB information during training. Regardless of the ID-tied cross-modality image pair or ID-exclusive cross-modality image pair (*i.e.*, two images belong to different IDs and modalities), the ImageNet-trained model is inclined to favor RGB features. That is, an ImageNet-trained model has difficulty learning essential ID features from IR images. Given a pair of images with the same ID, the objective of the model is to learn the shared features between images from two modalities to minimize their differences. However, due to the bias explained above, such an objective is hard to achieve. On the other hand, for ID-exclusive image pairs, the objective is to determine the difference between two images in terms of their ID. That is, the result will be acceptable as long as the model can learn different features from these two images from different modalities. Although such differences may be partially contributed by model bias mentioned above, the objective is easier to achieve from the view of model training. In other words, compared to the first case, the model training in the second case is more efficient and is able to converge more easily. Therefore, compared with directly using ID-tied cross-modality image pairs as the input, a better way to learn shared features is desired for infrared-visible LT-reID.

This chapter investigates the root cause of the problem described above and concludes that the MBT problem is the key cause. Fig. 6.2 provides an intuitive understanding of MBT. It can be seen that there is a clear boundary between the RGB and IR feature spaces if the model is trained using data from a single modality. When an ImageNet-trained model (*e.g.*, ResNet50) is adopted to learn the shared modality features by using data from the combination of these two modalities, the learned feature space is inclined to the RGB feature space.

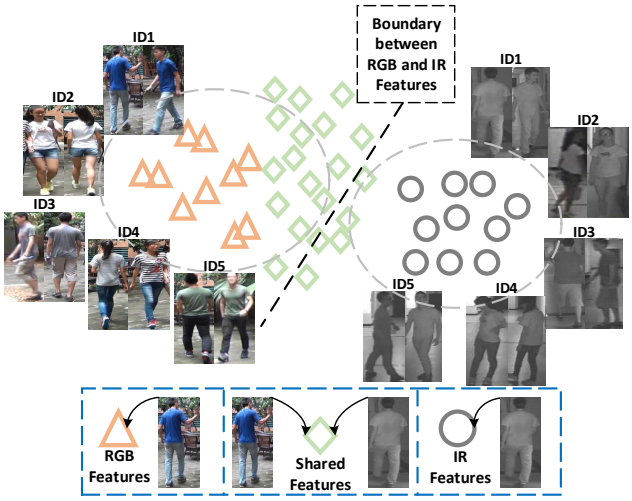


Figure 6.2 : The MBT issue presented in infrared-visible LT-reID. All images belong to five people IDs (*i.e.*, ID1 to ID5). The triangles, circles, and lozenges respectively represent features learned by the modality of RGB, IR, and both of them. Shared features learned from the two modalities are inclined to bias towards the modality of RGB. Best viewed in color.

In this chapter, a Dual-level Learning Strategy (DLS) is presented for infrared-visible LT-reID to mitigate the MBT problem:

First, since ID-tied cross-modality image pairs cause the MBT issue, the proposed DLS should not be heavily tangled with ID-tied information. In each training minibatch, the proposed method forces the network to focus on the ID-exclusive cross-modality image pairs rather than the ID-tied pairs.

Second, third modality data is used to further alleviate the MBT issue in the proposed DLS. The third modality data are generated by GAN. Unlike previous attempts (Zheng et al., 2017b; Huang et al., 2018a) that use GAN to generate RGB images only for person re-ID, the generated images in this work contain the characteristics from both RGB and IR modalities. More specifically, through fooling the discriminator, third modality data can be produced close to RGB and IR images

simultaneously. By doing so, the generated third modality data are able to carry information from the two modalities. Combining real RGB and IR images, the generated third modality data are also used in the training process to further alleviate the MBT issue. The unconditional DCGAN (Radford et al., 2015) is adopted for image generation following (Zheng et al., 2017b; Huang et al., 2018a). Since all the third modality data generated by DCGAN are unlabeled (do not belong to any people ID), a dynamic ID-exclusive Smooth (dIDeS) label is proposed for the third modality data that can then be used in the supervised training. With the addition of the third modality data, the MBT issue can be further alleviated in infrared-visible LT-reID.

To explicitly explore the MBT issue exposed in infrared-visible LT-reID, the classic IDE (and IDE-T) network is adopted directly (refer to Fig. 6.1) with ImageNet-trained ResNet50 as the backbone to conduct experiments. This is because, as mentioned above, the classic IDE model is the pivotal component for designing infrared-visible LT-reID approaches and has been widely adopted in existing infrared-visible LT-reID approaches. In experiments, the MBT issue is clearly exposed in the classic IDE model. Comprehensive evaluations are carried out to demonstrate the effectiveness of the proposed DLS. Note that the main significance of this work is not to attempt to achieve state-of-the-art performance on two infrared-visible LT-reID datasets or to design a fancy network architecture to deal with the task of infrared-visible LT-reID. This chapter is the first work that unveils the MBT issue in existing infrared-visible LT-reID models, which could affect the performance of infrared-visible LT-reID. To tackle (or alleviate) the MBT issue, a solution (*i.e.*, the proposed DLS) is first provided for the community by conducting experiments on the pivotal IDE network (without changing the network architecture) to facilitate future infrared-visible LT-reID works built upon IDE.

The contributions of this chapter are twofold:

- The MBT issue, which can lead to undesired performance degradation, is first unveiled for infrared-visible LT-reID.
- The underlying problem in current infrared-visible LT-reID approaches is investigated deeply. DLS, which enforces the network focus on ID-exclusive cross-modality image pairs and introduces third modality data during training, is proposed to address the undesired MBT issue.

6.2 Alleviating MBT Issue via Dual-Level Learning Strategy

In this section, DLS is presented to alleviate the MBT issue in infrared-visible LT-reID. As mentioned in Sec. 6.1, the first-level learning strategy in DLS is to learn shared features using the classic IDE and IDE-T models that focus on ID-exclusive labels of cross-modality image pairs. The second-level learning strategy in DLS is to leverage third modality data to further balance the distribution of shared features in the training process.

6.2.1 The First-Level Learning Strategy in DLS

To mitigate the MBT issue, ID-exclusive cross-modality image pairs are used as input into a vanilla IDE network (ImageNet-trained ResNet50 + ID classification loss). Then, IDE is evolved into IDE-T, which additionally benefits from triplet loss to enhance the feature discriminability.

IDE Model for infrared-visible LT-reID

First, the classic IDE model (Zheng et al., 2016a) is adopted without triplet loss. The training data of an infrared-visible LT-reID dataset is given as: $\mathcal{D} = \{(I_i^{RGB}, Y_i^{RGB}), (I_j^{IR}, Y_j^{IR})\}_{i=1,j=1}^{N_i,N_j}$, where I_i^{RGB} and I_j^{IR} represent i-th RGB image and j-th IR image, respectively. N_i and N_j respectively represent the total number of images belonging to the modality of RGB and IR. Given an RGB or IR image, its

ID label is denoted as a C-dimensional one-hot vector Y=$[0, ..., 0, 1, 0, ..., 0]^\top$, where 1 is present in the index of ID of the image. In conventional re-ID approaches, a minibatch data (*e.g.*, the batchsize is K) in $\mathcal{D}$ is randomly picked to train an IDE network $\mathbb{M}$. This strategy is widely used in single RGB modality re-ID tasks. For infrared-visible LT-reID, images from two different modalities are simply mixed. Images of a training minibatch are randomly picked from the mixed training set, regardless of the modality information. This randomly mixed setting is denoted as D^m. Compared with existing infrared-visible LT-reID approaches, it is observed that general model D^m achieves better performance when it is applied into a vanilla IDE network without bells and whistles (refer to Sec. 6.3). Investigating why the performance of existing infrared-visible LT-reID approaches is even worse than a simple IDE model D^m is required.

Following the above-mentioned idea, other two settings are introduced. Unlike D^m, paired images (I_i^{RGB}, I_j^{IR}) are fed into $\mathbb{M}$ in each training minibatch. There are two different settings: when the ID label $Y_i^{RGB} == Y_j^{IR}$, I_i^{RGB} and I_j^{IR} belong to the same people ID (ID-tied); Otherwise, I_i^{RGB} and I_j^{IR} belong to different people IDs (ID-exclusive). The two types of settings are respectively denoted as D^t and D^e. If there are K images in each minibatch of D^m, there should be $K/2$ image pairs in D^t or D^e. To prevent the network from MBT, D^e, which uses the cross-modality ID-exclusive label on each image pair, constitutes the first-level learning strategy. Comprehensive evaluations are carried out in experiments to show the effectiveness of D^e in mitigating the MBT issue (refer to Fig. 6.6). Tab. 6.2 shows the three different settings to clearly distinguish the differences between D^m, D^t, and D^e.

IDE-T for infrared-visible LT-reID

Based on D^e, the discriminability of shared features learned by an IDE model can be further enhanced with a cross-modality triplet loss. The cross-modality triplet

Table 6.2 : Difference between D^m, D^t, and D^e settings for training.

Settings	Inputs in each minibatch
D^m	K images are randomly picked from $\mathcal{D}$
D^t	$\frac{K}{2}$ image pairs $\{(I_i^{RGB}, I_j^{IR}) \vert Y_i^{RGB} == Y_j^{IR}\}$ are randomly picked from $\mathcal{D}$
D^e	$\frac{K}{2}$ image pairs $\{(I_i^{RGB}, I_j^{IR}) \vert Y_i^{RGB} \neq Y_j^{IR}\}$ are randomly picked from $\mathcal{D}$

loss can be employed in $\mathbb{M}$ as follows:

$$\mathcal{L}_T = max(\|F_t(I_a^A) - F_t(I_p^B)\|_2^2 - \|F_t(I_a^A) - F_t(I_n^B)\|_2^2 + \xi, 0), \tag{6.1}$$

where A and B represent two different modalities; if A=RGB, B=IR, and *vice versa*. I_a^A is an anchor image sample with the ID label Y_a^A; I_p^B (I_n^B) is a corresponding positive (negative) image sample with the ID label Y_p^B (Y_n^B). The positive (negative) image sample is also called the ID-tied (ID-exclusive) image sample as shown in Fig. 6.1. Thus, $Y_a^A == Y_p^B$, $Y_a^A \neq Y_n^B$. The $F_t(\cdot) \in \mathbb{R}^{1024}$ represents the output of a 1024-d fully connected layer (*i.e.*, FC2 in Fig. 6.1). ξ is a margin parameter.

To use triplet loss, both ID-tied and ID-exclusive image samples are fed into the network. However, as it is mentioned in Sec. 6.1, D^t which uses ID-tied image pairs as inputs is easily biased towards the modality of RGB (refer to Fig. 6.6 (d)). Therefore, the contribution of ID-tied image samples is reduced by introducing a decay factor γ in ID loss. With γ, the ID loss in IDE-T is given as follows:

$$\mathcal{L}_{ID} = -(Y_a^A \odot logP_a^A + \gamma \cdot Y_p^B \odot logP_p^B + Y_n^B \odot logP_n^B), \tag{6.2}$$

where $\odot$ is element-wise multiplication; P_x^X (*e.g.*, $X \in \{A, B\}$, $x \in \{a, p, n\}$) represents the softmax predicted probability of an image I_x^X:

$$P_x^X = P(c|F_{id}(I_x^X)) = \frac{[e^{\rho_1}, e^{\rho_2}, ..., e^{\rho_C}]^\top}{\sum_{c=1}^{C} e^{\rho_c}}, \tag{6.3}$$

where $F_{id}(\cdot) \in \mathbb{R}^C$ is the output of a C-dimensional fully connected layer (*i.e.*, FC1

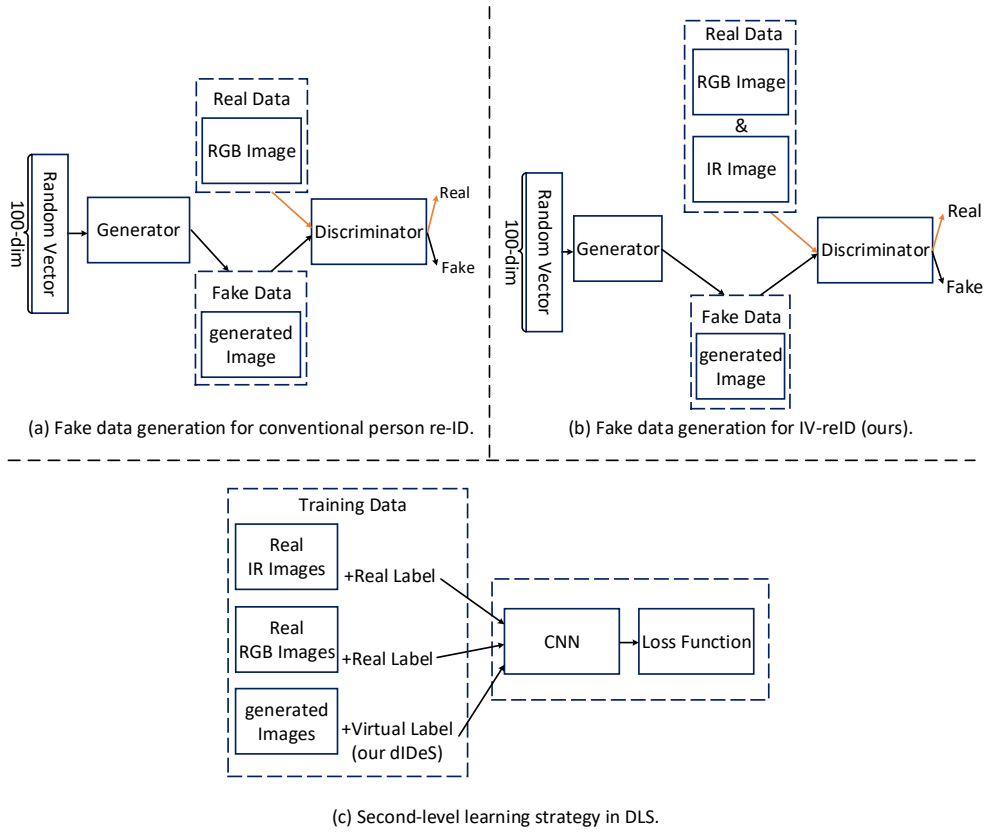


Figure 6.3 : (a) and (b) are DCGAN models used in conventional re-ID and the proposed solution, respectively. 100-dimensional random vectors are fed into the generator for fake image generation. (c) shows the second-level learning strategy in DLS by jointly training the third modality images (the generated image by (b)) and real images using a CNN network (*e.g.*, classic IDE or IDE-T network).

in Fig. 6.1). e^{ρ_c} represents the probability of I_x^X belonging to c-th predefined training class (people ID).

6.2.2 The Second-Level Learning Strategy in DLS

Besides the strategy to move focus to ID-exclusive data, another strategy is to introduce more neutral data that possesses balancing information between RGB and IR to dilute the bias caused by unbalanced training data between two modalities.

Such neutral data (which is also called the third modality data) can be produced by unconditional GAN model (*e.g.* DCGAN (Radford et al., 2015)).

The Data Generation Process

To generate fake images for re-ID model training (*i.e.*, data augmentation), previous conventional person re-ID studies (Zheng et al., 2017b; Huang et al., 2018a) adopt unconditional GANs such as DCGAN (Radford et al., 2015). Thus, this work also follows the same DCGAN setting of (Zheng et al., 2017b; Huang et al., 2018a) to generate fake images. Unlike (Zheng et al., 2017b; Huang et al., 2018a), which only contain images from the RGB modality to train their DCGAN, in the proposed solution, DCGAN is trained with images randomly picked from two modalities. However, this process may make training DCGAN more difficult if the input images are from different modalities (*e.g.*, less stable). To minimize the stability issue, this chapter proposes two ideas to adjust DCGAN training. To best differentiate the ideas, Fig. 6.3 (a) and (b) illustrate changes in DCGAN training to generate the third modality data. First, in each training minibatch in DCGAN, one-half of the images are randomly picked from the RGB modality, and the other half of the images are randomly picked from the IR modality. This process ensures the best information balance between two modalities. Second, in the amended DCGAN described in this work, the discriminator verifies that the distribution of the pattern of the generated data is aligned with the combination of RGB and IR data (rather than single modality data as with conventional DCGAN). Fig. 6.9 shows the generated third modality images. It can be clearly seen that the third modality images have the characteristics of both RGB and IR information.

Since DCGAN belongs to the unconditional GAN model (Radford et al., 2015), all the generated images do not have ID labels (Zheng et al., 2017b; Huang et al., 2018a). Therefore, to use the third modality data in training, dIDeS virtual labels

are proposed and assigned to the third modality data. In training, the real images with real ID labels and generated third modality images with virtual labels are combined together (refer to Fig. 6.3 (c)). Specifically, if the third modality images are added to train IDE with D^e, there are $K/2$ triplet tuples (*i.e.*, $[(I_a^A, I_n^B, I_g^G)_k]_{k=1}^{K/2}$) in each training minibatch, where I_g^G represents a third modality image. To train IDE-T, I_p^B is added in the triplet tuple (*i.e.*, $[(I_a^A, I_p^B, I_n^B, I_g^G)_k]_{k=1}^{K/2}$).

The dIDeS Virtual Label

Since all generated images (the third modality data) by DCGAN are unlabeled, dIDeS virtual labels are proposed for these generated data to ensure a supervised learning process in training. The proposed dIDeS label is inspired by LSRO (Zheng et al., 2017b). LSRO assigns each unlabeled generated image with a smooth virtual label $Y_g^G \in \mathbb{R}^C = [\frac{1}{C}, ..., \frac{1}{C}, ..., \frac{1}{C}]^\top$, where C is the number of people ID in the training set. LSRO is used to prevent the network from over-fitting on a particular people ID (Zheng et al., 2017b). Thus, features extracted from the well-trained network can better generalize on the testing set. **Unlike LSRO**, the proposed dIDeS dynamically zeros out label information corresponding to real images presented in each training minibatch. This is because the third modality data contain information from both RGB and IR modalities. As it is mentioned in Sec. 6.1, the ID-tied cross-modality information may increase the risk of MBT in infrared-visible LT-reID. Formally, given a third modality image I_g^G, its dIDeS label is defined as:

$$Y_g^G = [\frac{\varepsilon_c}{C}]_{c=1}^{C}, \tag{6.4}$$

where c represents the index of people IDs in the training set; $\varepsilon_c \in \{0, 1\}$. If there is no real image (both RGB and IR) in a minibatch belonging to c-th people ID, $\varepsilon_c = 1$, otherwise, $\varepsilon_c = 0$. According to the existence of some real ID labels in each minibatch, different dIDeS labels can be assigned to the same third modality image when the image is picked by different training minibatches. Fig. 6.4 shows

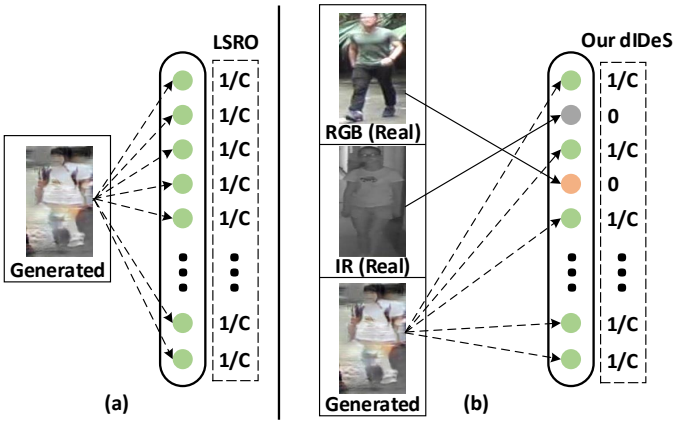


Figure 6.4 : (a): LSRO, (b): the proposed dIDeS. In (b), the IDE inputs are used as an example and assume that $K = 2$. Best viewed in color.

the difference between LSRO and the proposed dIDeS label.

Advantage of The Second-Level Learning Strategy

First, since third modality images are produced by considering information from both RGB and IR modalities, they inherently have a positive impact on learning shared features across the two modalities. **Second**, for real images, more attention is paid to the ID-exclusive image pairs (*i.e.*, D^e). However, ImageNet-trained ResNet50 is more familiar with the RGB modality. Thus, ImageNet-trained ResNet50 is more likely to extract discriminative information from certain person IDs from RGB images rather than from IR images. To mitigate this issue, the proposed dIDeS can enforce the network to not be too confident regarding certain person IDs from RGB images in each training minibatch. **Third**, dIDeS zeros out the label information corresponding to real images in a minibatch. As it is mentioned before, ID-tied information may cause the MBT issue. Since the real images and generated third modality images can be regarded as data from different modalities, the dIDeS (the label for the third modality data) is also regarded as being mutually exclusive to the ground-truth ID labels of real images within a minibatch. That is,

they are ID-exclusive. According to the first-level learning strategy (see Sec. 6.2.1), it is able to mitigate the MBT issue.

For a generated third modality image with dIDes, the ID loss is formulated as:

$$\mathcal{L}_{ID_G} = -\sum_{c=1}^{C} \frac{\varepsilon_c}{C} \cdot log P_g^G(c), \tag{6.5}$$

Finally, the total loss of the proposed DLS is:

$$\mathcal{L} = \frac{1}{3K/2} \cdot \sum_{k=1}^{K/2} \mathcal{L}_{ID}^k + \frac{1}{K/2} \cdot \sum_{k=1}^{K/2} (\alpha \cdot \mathcal{L}_{ID_G}^k + \beta \cdot \mathcal{L}_T^k), \tag{6.6}$$

where α and β are hyper-parameters to control the weight of different objective functions. Since $\mathcal{L}_{ID}^k$ contains anchor, positive, and negative image samples, it needs to be reduced by three times further (*i.e.*, $\frac{1}{3K/2}$).

6.3 Experiments

6.3.1 Infrared-visible LT-reID Datasets and Evaluation Metrics

In experiments, two widely used publicly available infrared-visible LT-reID datasets are adopted (*i.e.*, SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017)) for evaluation. The experiments are mainly conducted on SYSU-MM01. This is because: 1) compared with RegDB (4K training images), SYSU-MM01 (34K training images) is a large-scale infrared-visible LT-reID dataset that is more suitable for evaluations of deep neural network models; 2) the images in SYSU-MM01 are more colorful than those in RegDB, and the bodies of persons are not aligned well with different postures and scales (Wang et al., 2019c), which makes them more difficult to evaluate and closer to reality. Fig. 6.5 shows image samples from the two datasets.

SYSU-MM01 is a large-scale infrared-visible LT-reID dataset collected on a campus with six cameras (three indoors + three outdoors). SYSU-MM01 contains 491 person IDs, 395 for training, and the remaining 96 for testing. There are 22,258

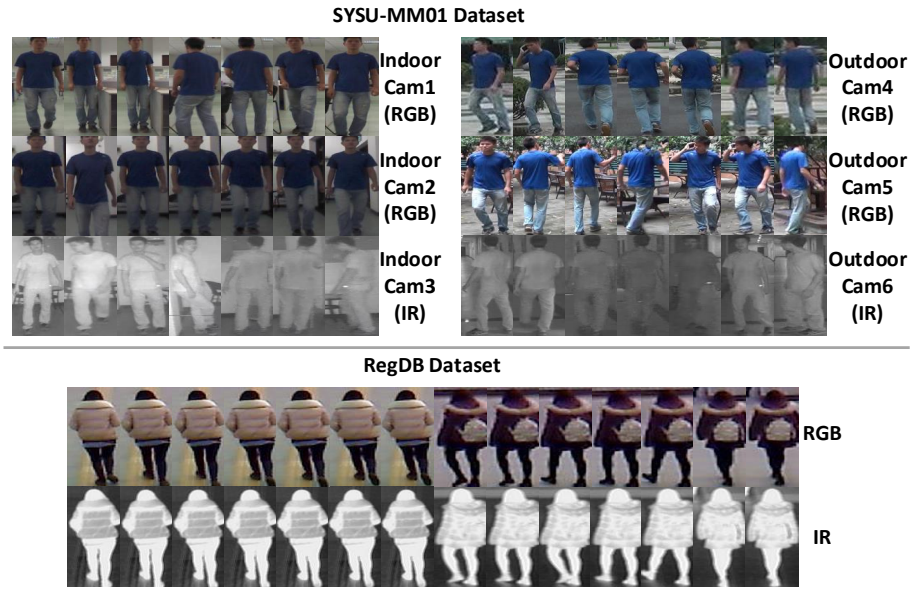


Figure 6.5 : Image samples of one people ID from the SYSU-MM01 dataset (upper) and two people IDs from the RegDB dataset (lower).

RGB images and 11,909 IR images in the training set. The single-shot all-search mode (Wu et al., 2017) is used for testing. Specifically, 3,803 IR images from cameras 3 and 6 of the 96 IDs are used for query, and 301 RGB images are randomly selected from RGB cameras 1, 2, 4, and 5 as the gallery set. The selection of the gallery set is conducted 10 times. In addition, probe images of camera 3 skip the gallery images of camera 2 since both cameras are in the same location. The single-shot all-search setting is chosen on SYSU-MM01 because it is the most challenging and widely used setting on this dataset.

RegDB is a small-scale infrared-visible LT-reID dataset collected by a dual-camera system. RegDB contains 412 people IDs. For each ID, 10 images are captured by an RGB camera, and another 10 images are captured by an IR camera. Following the evaluation protocol in (Ye et al., 2018a), the dataset is randomly split into two halves, one for training and one for testing. In the testing set, all RGB images are used for query, while all IR images are used as the gallery set. The data

splitting procedure is repeated 10 times, and the average performance is reported.

Evaluation Metrics. The standard Cumulated Matching Characteristics (CMC) and mean Average Precision (mAP) are adopted for evaluation.

6.3.2 Implementation Details

The IDE and its evolved IDE-T network architectures (refer to Fig. 6.1) are adopted to evaluate the effectiveness of the proposed DLS in mitigating the MBT issue exposed in infrared-visible LT-reID. In addition to ImageNet-trained ResNet50, ImageNet-trained DenseNet121 is adopted as a CNN backbone to evaluate the proposed method comprehensively. FC1 is set to have C neurons, where C is the number of person IDs in the training set (C=395 for SYSU-MM01, C=206 for RegDB). For IDE-T, FC2 is set to have 1024 neurons to train the cross-modality triplet loss (refer to Eq. 6.1). The margin parameter ξ is set to 1. IR images are selected as anchors, while RGB images are used as positive and negative samples for both datasets. The batchsize K is set to 64. In training, all the images are resized to 288×144 before being randomly cropped to 256×128. Random horizontal flipping is performed on each training image. Stochastic Gradient Descent (SGD) is performed with a momentum of 0.9. For the ImageNet-trained backbone, the learning rate is set to 0.01. For FC1 and FC2, the learning rate is set to 0.1 and 5e-6, respectively. For SYSU-MM01 (RegDB), all the learning rates decay 10 times in the last 10 training epochs, and the training stops after the 20th (50th) epoch. To test the effectiveness of the proposed DLS in alleviating the MBT issue, the IDE features of each image are extracted after GAP (see Fig. 6.1). The squared Euclidean distance calculates the similarity between two images before ranking.

6.3.3 Effectiveness of The First-Level Learning Strategy

The proposed first-level learning strategy is embedded into the IDE and IDE-T architectures, respectively. Both qualitative and quantitative evaluations are given.

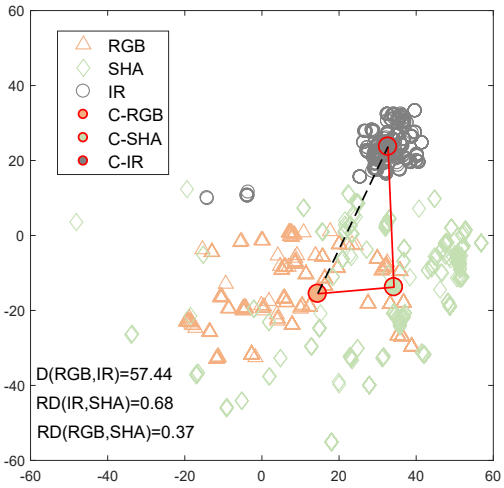
Qualitative Evaluation by Visualizing Distributions of Shared Features v.s. Single Modality Features.

To explicitly demonstrate the MBT issue exposed in infrared-visible LT-reID. The distributions of shared features (*i.e.*, SHA) and features learned from one single modality (*i.e.*, RGB or IR) are visualized. Fig. 6.6 (a)-(d) illustrate the results.

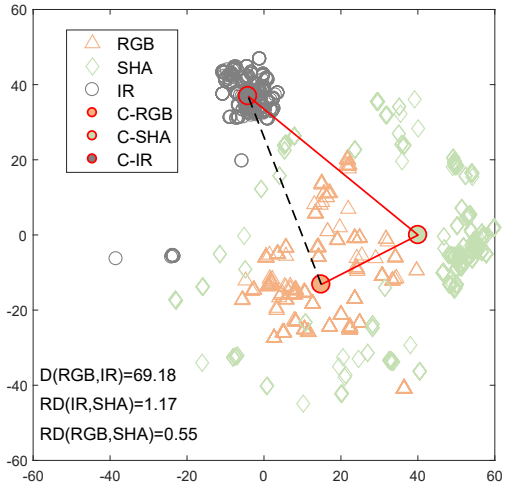
Given the training set of SYSU-MM01, there are several steps used for the visualization: **First**, two IDE models using RGB and IR images are trained, respectively. The two IDE models are used to extract single modality features. **Second**, four IDE models using mixed data (RGB+IR) are trained. The four models use different settings (*i.e.*, D^m, D^t, D^e, and D^e+dIDeS), respectively. The four models are used to extract shared features across modalities. **Finally**, 500 RGB images and 500 IR images are used to visualize the single modality features in Fig. 6.6. Another 500 images (250 RGB + 250 IR) are used to visualize the shared features. In Fig. 6.6 (a)-(d), IDE models with different settings are used to extract shared features (see the captions of each sub-figure).

According to the distribution result shown in the figure, it is clearly seen that compared with D^e (Fig. 6.6 (c)), shared features learned by D^m and D^t show obvious modality bias. In order to indicate how much the proposed DLS resolves the MBT issue for infrared-visible LT-reID, a modality Relative Distance (RD) is defined to measure the degree of modality bias using the center points of RGB, IR, and SHA features shown in Fig. 6.6. The RD is given as follows:

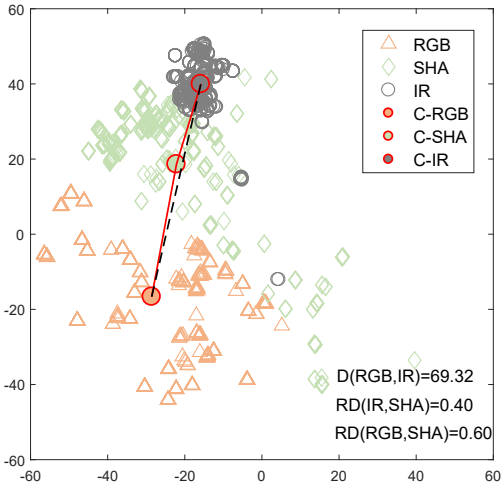
$$RD(M, SHA) = \frac{D(M, SHA)}{D(RGB, IR)}, \tag{6.7}$$



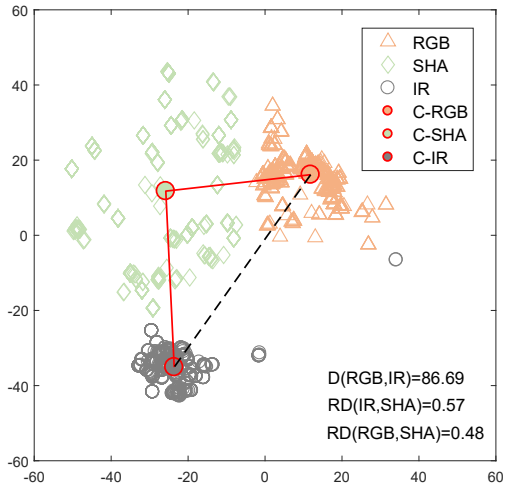
(a) IDE (D^m) (ResNet50).



(b) IDE (D^t) (ResNet50).



(c) IDE (D^e) (ResNet50).



(d) IDE (D^e) (ResNet50)+dIDeS.

Figure 6.6 : The (a)-(d) show distributions of IDE features by Barnes-Hut t-SNE (Van Der Maaten, 2014b). The SHA is short for shared features extracted from a model trained on two modalities. RGB and IR respectively represent the RGB features and IR features extracted from the model trained on a single modality. 'C-' represents the center point of data distribution. D(x,y) is the L1 distance between C-x and C-y. 'RD(,)' is short for the modality Relative Distance (RD), which is used to measure the distance of (C-IR, C-SHA) and (C-RGB, C-SHA)

where $M \in \{\text{RGB or IR}\}$. The relative distance is used because D(RGB,IR) should be the same across (a)-(d) since all RGB and IR features in Fig. 6.6(a)-(d) are extracted from two same IDE models trained on the two modalities respectively. However, when t-SNE is used for visualization, D(RGB,IR) is changed across (a)-(d) due to the impact caused by different SHA.

In Fig. 6.6 (a), RD(IR,SHA)=0.68 and RD(RGB,SHA)=0.37. This result shows that shared features (*i.e.*, SHA) learned by IDE(D^m) are closer to the modality of RGB. The gap between the two RDs is |RD(IR,SHA)-RD(RGB,SHA)|=0.31. That is, the bias is quite obvious. When D^t is adopted for training SHA feature (see Fig. 6.6 (b)), the bias becomes significant (*i.e.*, |RD(IR,SHA)-RD(RGB,SHA)|=0.62). This is due to the input of D^t using ID-tied cross-modality image pairs. The ImageNet-trained ResNet50 is likely to learn features from its familiar modality (*i.e.*, RGB). When D^e is adopted to learn shareable features, it shows much better results (*i.e.*, |RD(IR,SHA)-RD(RGB,SHA)|=0.2). Clearly, the bias is suppressed a lot as indicated in Fig. 6.6 (c). This demonstrates the effectiveness of ID-exclusive information for infrared-visible LT-reID. Fig. 6.6 (d) shows the effectiveness of the second-level learning strategy for alleviating the MBT issue. Corresponding analysis can refer to Sec. 6.3.4.

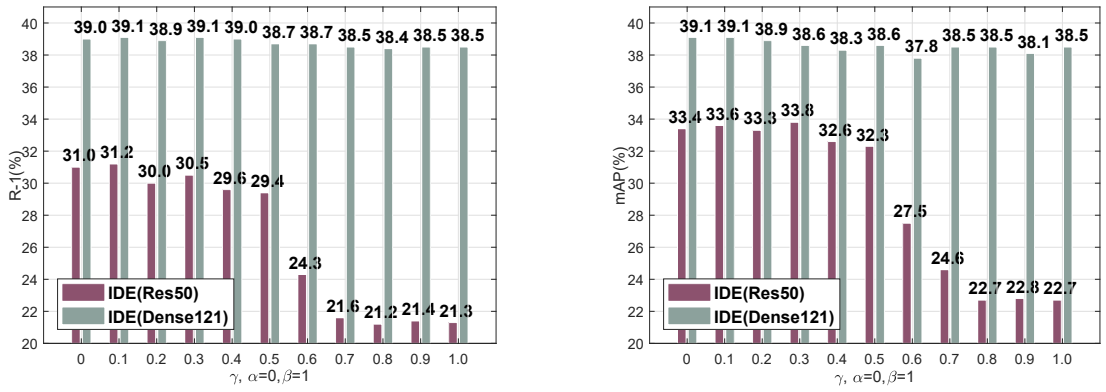
Quantitative Evaluation of The First-Level Learning Strategy.

The performance of D^m, D^t, and D^e is compared by using ImageNet-trained ResNet50 and DenseNet121. Tab. 6.3 shows the results. It can be observed that D^e achieves the best performance when it is adopted by both the ResNet50 and DenseNet121 IDE networks. A significant performance improvement is achieved from D^t to D^e on ResNet50 (*i.e.*, rank-1: 20.2% *vs.* 29.7%). This result shows that state-of-the-art approaches (*e.g.*, (Dai et al., 2018; Ye et al., 2018b, 2019)), which normally adopt ID-tied cross-modality image pairs as inputs into ImageNet-

Table 6.3 : The performance of IDE (with D^m, D^t, and D^e) and IDE-T (with α=0, β=1, and γ=0.1 in Eq. 6.6) on SYSU-MM01. ResNet50 and DenseNet121 are pretrained on ImageNet. The IDE* means training from scratch. 'R-1' means Rank-1.

Methods	ResNet50		DenseNet121	
	mAP	R-1	mAP	R-1
IDE (D^m)	29.5	28.2	33.6	34.0
IDE (D^t)	24.1	20.2	33.8	34.8
IDE (D^e)	*30.3*	*29.7*	*34.7*	*35.7*
IDE* (D^m)	14.6	10.7	16.2	13.5
IDE* (D^e)	14.9	11.1	16.1	13.3
IDE-T *w/o.* $\mathcal{L}_{ID}$	21.2	16.7	21.6	17.2
IDE-T (γ)	**33.6**	**31.2**	**39.1**	**39.1**

trained ResNet50, are easily inclined to a modality bias in training. Compared with ResNet50, DenseNet121 is not adopted by any existing infrared-visible LT-reID work. When DenseNet121 is adopted in the IDE model, MBT issue is less serious. However, Tab. 6.3 still shows that the training using D^e can still boost the performance by $\sim$1% for rank-1 accuracy and mAP. This experiment demonstrates the effectiveness of D^e for alleviating the MBT issue when different backbones are applied. If both models are trained from scratch (*i.e.*, IDE* in Tab. 6.3, without pretraining on ImageNet dataset), then the performance drops significantly, which shows the significance of ImageNet-trained models for infrared-visible LT-reID. Indeed, if IDE* is not pretrained using ImageNet, then it should not have an MBT issue. However, without the pretraining process, parameter initialization in the CNN backbone becomes significantly challenging, which will cause many uncertainties and even jeopardize the overall performance. Thus, this work still adopts the ImageNet pretraining process followed by the proposed solutions to deal with the



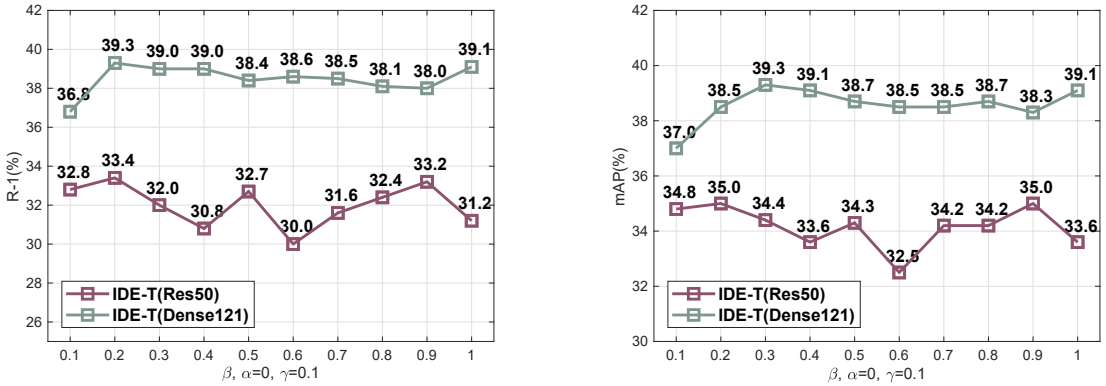
(a) rank-1 accuracy

(b) mAP

Figure 6.7 : The performance of changing γ in IDE-T on SYSU-MM01.

MBT issue. When the IDE-T architecture with the decay factor γ (refer to Eq. 6.2) is used, performance can be further improved (refer to IDE-T(γ) in Tab. 6.3). It is noteworthy that this experiment also tries to remove $\mathcal{L}_{ID}$ for training (*i.e.*, only use triplet loss: IDE-T *w/o.* $\mathcal{L}_{ID}$). Compared with only using $\mathcal{L}_{ID}$ (*i.e.*, IDE), the performance drops significantly on both ResNet50 and DenseNet121. This is because triplet loss is normally used for enhancing the discriminative ability of features, which is regarded as an auxiliary loss term in addition to the ID loss. Thus, the proposed DLS method only focuses on the ID loss.

Parameter Analysis (γ). γ is used to control the weight of ID-tied samples in IDE-T (Eq. 6.2). In Fig. 6.7, it is clear that ImageNet-trained ResNet50 is sensitive to γ for ID-tied samples. This result shows that, due to MBT, ID-tied samples can compromise performance when they are fed into IDE-T. If the ImageNet-trained backbone is changed to DenseNet121, the results become better. But for both ResNet50 and DenseNet121, the best performance is achieved when $\gamma = 0.1$. This experiment shows the ID-tied samples can produce negative effects to infrared-visible LT-reID when triplet loss is adopted for infrared-visible LT-reID. $\gamma = 0.1$ is used in experiments when IDE-T is adopted.



(a) rank-1 accuracy (b) mAP

Figure 6.8 : The performance of changing the weight β in IDE-T on SYSU-MM01.

Parameter Analysis (β). β (refer to Eq. 6.6) is evaluated on both the ImageNet-trained ResNet50 and DenseNet121 backbones. As shown in Fig. 6.8, when β is changed from 0.1 to 1 (stepsize=0.1), DenseNet121 outperforms ResNet50 in terms of both mAP and rank-1 accuracy. However, for both DenseNet121 and ResNet50, the performance fluctuates with the change in β. For the unified management of parameters, β is set to 1 in experiments to maintain the consistency between ResNet50 and DenseNet121.

6.3.4 Effectiveness of The Second-Level Learning Strategy

To demonstrate the effectiveness of the second-level strategy, a visual demonstration of the generated third modality images is used, and then the performance of the proposed dIDeS virtual labels is evaluated.

Visualization of Generated Third Modality Images. Fig. 6.9 shows the generated third modality images. Since the unconditional DCGAN is used for fake data generation (refer to Sec. 6.2.2), it can be observed that each third modality image does not belong to any person ID. The third modality images can be regarded as a type of neutral data lying between the modalities of RGB and IR. Although

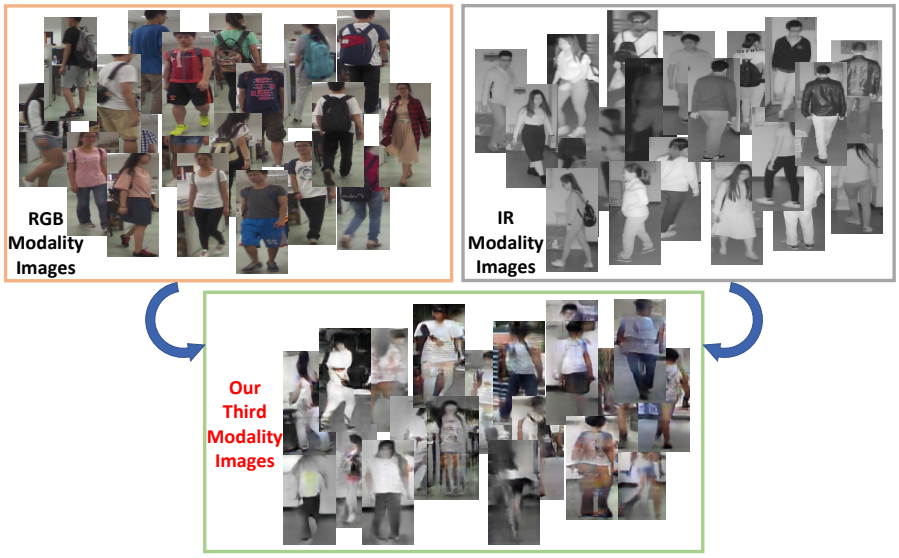


Figure 6.9 : Generated third modality images by DCGAN using the SYSU-MM01 dataset. Images in orange and gray bounding boxes are real RGB and IR images, respectively. Best viewed in color.

the generated third modality images do not look similar to RGB or IR images, they are useful. Early works (Zheng et al., 2017b; Huang et al., 2018a) also support the ideas that generated unlabeled data can improve the performance of person re-ID. Note that, other unconditional GAN models such as WGAN (Arjovsky et al., 2017) and WGAN-GP (Gulrajani et al., 2017) may produce higher quality images than DCGAN used in this work. However, other unconditional GAN models are not used in the comparison. This is because better generators may lead to superior perceptual quality, but may not boost the performance when these generated images are used in re-ID model training (Huang et al., 2018a). Given an infrared-visible LT-reID dataset, N_g third modality images are generated, where $N_g = max(N_i, N_j)$.

Visualization of Data Distribution. To more clearly demonstrate the characteristics of unique third modality (or neutral) data, Fig. 6.10 shows the distribution of the pattern of generated third modality data using t-SNE. The distribution of

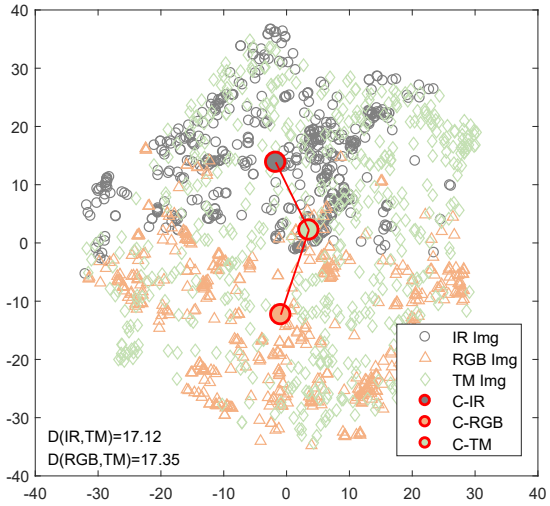


Figure 6.10 : Visualization of data distribution by using Barnes-Hut t-SNE (Van Der Maaten, 2014b). For the visualization, 500 RGB, IR, and third modality images are respectively used. 'Img' is short for 'Image'. 'TM' is short for the 'Third Modality' generated images. 'C-' represents the center point for different types of data. D(x,y) represents the L1 distance between C-x and C-y.

the third modality data lies between the distributions of RGB images and IR images (*e.g.*, |D(IR,TM)-D(RGB,TM)|=|17.12-17.35|=0.23), which indicates that the generated data possess balanced information which can mitigate MBT issues.

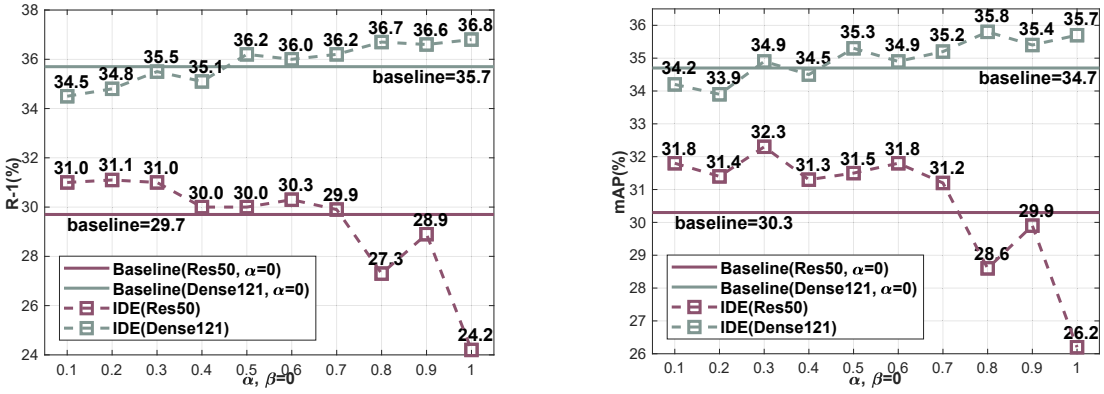
Comparison with Other Virtual Labels. In this experiment, the proposed dIDeS is compared with other state-of-the-art virtual labels for person re-ID, including LSRO (Zheng et al., 2017b), dMpRL-I and dMpRL-II (Huang et al., 2018a). These three methods are used to improve the generalization ability of CNNs on single modality (*i.e.*, RGB) person re-ID tasks. The comparisons are shown in Tab. 6.4. Compared with the proposed dIDeS, **dIDtS** is an inverse operation. Specifically, in Eq. 6.4, if there is no real image in the a minibatch belonging to c-th people ID, $\varepsilon_c = 0$, otherwise, $\varepsilon_c = 1$. The proposed dIDeS achieves the best performance on both ImageNet-trained ResNet50 and DenseNet121 when the IDE

Table 6.4 : Comparison with state-of-the-art virtual labels for infrared-visible LT-reID.

Methods	ResNet50		DenseNet121	
	mAP	R-1	mAP	R-1
IDE (D^e) (baseline)	30.3	29.7	34.7	35.7
IDE (D^e)+LSRO	30.8	30.0	34.8	35.6
IDE (D^e)+dMpRL-I	28.2	27.4	*34.9*	*35.8*
IDE (D^e)+dMpRL-II	*30.6*	*30.1*	34.1	35.1
IDE (D^e)+dIDtS	30.4	28.8	34.4	35.9
IDE (D^e)+Random-Zero	30.6	30.0	34.7	35.8
IDE (D^e)+dIDeS (Our)	**31.8**	**31.0**	**35.7**	**36.8**
IDE-T (γ) (baseline)	33.6	31.2	39.1	39.1
IDE-T (γ)+dIDeS (Our)	**35.0**	**32.6**	**40.6**	**40.5**

network is adopted. The proposed dIDeS outperforms the baseline (without using third modality images in training) by 1.3% and 1.1% in terms of rank-1 accuracy on ResNet50 and DenseNet121 respectively, which shows the effectiveness of the proposed dIDeS virtual label for infrared-visible LT-reID. In addition, when randomly zero out N_z label information for each generated image (N_z equals to the number of zero in its corresponding dIDeS label), the result is also shown in Tab. 6.4 (*i.e.*, IDE (D^e)+Random-Zero). It can be observed that when using a random zero out strategy, the performance is lower than the proposed dIDeS, which demonstrates the effectiveness of the proposed dIDeS label for generated images.

To further illustrate the effectiveness of the proposed dIDeS (with third modality images), a comparison is provided in Fig. 6.6 (d). Compared with the best results obtained by only using the first-level learning strategy (Fig. 6.6 (c)), the MBT issue is further mitigated (*i.e.*, |RD(IR,SHA)-RD(RGB,SHA)|=0.09). Finally, IDE-T is also trained with the proposed third modality images via the proposed dIDeS. The



(a) rank-1 accuracy (b) mAP

Figure 6.11 : The performance of changing the weight α in IDE on SYSU-MM01. The baseline is the performance when $\alpha = 0$ (no third modality image is used in training).

performance can be further improved 1.5% in terms of mAP (*e.g.*, 40.6% *vs.* 39.1%).

Parameter Analysis (α). Fig. 6.11 shows the performance when the weight of third modality images is changed in training. The analysis of α is based on IDE. The proposed dIDeS label is dynamically assigned to each third modality image when it is fed into IDE along with the real images in each minibatch. It can be observed in Fig. 6.11 that the performance shows growth trends along with the increase in α when DenseNet121 is adopted. However, when ResNet50 is adopted, the performance shows the opposite trend. Therefore, α is set to 1 for DenseNet121 and 0.1 for ResNet50.

6.3.5 Using Other ImageNet-Trained Model for Evaluation

To show the effectiveness of the proposed method in broader architectures, another model (*i.e.*, AGW (Ye et al., 2020)) is used to evaluate the proposed MBT mitigation concept, which is also pretrained on ImageNet by using ResNet50. Compared with the other two models, IDE and IDE-T, AGW is regarded as a stronger

Table 6.5 : Evaluate the proposed DLS using AGW on SYSU-MM01. AGT means replacing the WRT loss used in (Ye et al., 2020) with the classic triplet loss.

Methods	mAP	R-1
AGW (reported in (Ye et al., 2020))	47.7	47.5
AGW (run by our)	48.0	47.6
AGW+dIDES (second-level DLS)	48.8	48.4
AGT (modify by our)	46.5	46.2
AGT(γ) (first-level DLS)	48.2	47.9
AGT(γ)+dIDeS (first-level+second-level DLS)	**49.0**	**48.8**

baseline for infrared-visible LT-reID, which performs reasonably well on existing infrared-visible LT-reID datasets (*e.g.*, SYSU-MM01). The proposed DLS is applied to AWG. The result is shown in Tab. 6.5. By directly adding the third modality data introduced in this work into training with the proposed dIDeS virtual label, the performance can be improved from 47.6% to 48.4% in terms of rank-1 accuracy. A minor modification to AGW model is made by changing the Weighted Regularization Triplet (WRT) loss to classic triplet loss. This modified model is named AGT. Then both the proposed first-level and second-level DLS are adopted in AGT. It can be seen in Tab. 6.5 that DLS is able to progressively improve the performance by using first-level and second-level DLSs in AGT.

6.3.6 Comparison with State-of-the-Art Methods

The proposed method is compared with several published available state-of-the-art infrared-visible LT-reID approaches, including Zero-Padding (Wu et al., 2017), HCML (Ye et al., 2018a), BDTR (Ye et al., 2018b), cmGAN (Dai et al., 2018), eBDTR (Ye et al., 2019), D-HSME (Hao et al., 2019), D^{2}RL (Wang et al., 2019c), SDL (Kansal et al., 2020), JSIA-ReID (Wang et al., 2020), and AlignGAN (Wang et al., 2019a). The comparisons are conducted on two datasets.

Table 6.6 : Comparison with state-of-the-art infrared-visible LT-reID approaches on the SYSU-MM01 and RegDB datasets. For IDE and IDE-T based model, ImageNet-trained DenseNet121 is adopted in comparison. The AGT is modified from AGW (Ye et al., 2020).

Methods	SYSU-MM01		RegDB	
	mAP	R-1	mAP	R-1
Zero-Padding (Wu et al., 2017) *2017*	16.0	14.8	18.9	17.8
HCML (Ye et al., 2018a) *2018*	16.1	14.3	20.8	24.4
BDTR (Ye et al., 2018b) *2018*	19.7	17.0	31.8	33.5
cmGAN (Dai et al., 2018) *2018*	27.8	27.0	-	-
eBDTR (Ye et al., 2019) *2019*	28.4	27.8	33.5	34.6
D-HSME (Hao et al., 2019) *2019*	23.1	20.7	47.0	50.9
D^2RL (Wang et al., 2019c) *2019*	29.2	28.9	44.1	43.4
AlignGAN (Wang et al., 2019a) *2019*	*40.7*	*42.4*	*53.5*	*57.1*
SDL (Kansal et al., 2020) *2020*	29.0	28.1	23.2	26.1
JSIA-ReID (Wang et al., 2020) *2020*	36.9	38.1	*49.1*	*48.3*
Ours				
IDE (D^e)	34.7	35.7	32.8	31.2
IDE (D^e)+dIDeS	35.7	36.8	34.2	31.8
IDE-T (γ)	39.1	39.1	48.6	46.1
IDE-T (γ)+dIDeS	40.6	40.5	49.2	47.2
AGT(γ)+dIDeS	**49.0**	**48.8**	**68.1**	**71.1**

SYSU-MM01. It can be observed in Tab. 6.6 that the proposed method outperforms the recently published method JSIA-ReID by +3.7% in terms of mAP and +2.4% in terms of rank-1 accuracy. The mAP and rank-1 accuracy of our method are also higher than those of JSIA-ReID when only the first-level learning strategy is adopted (*e.g.*, IDE (D^e)). Since SYSU-MM01 is a large-scale infrared-visible LT-reID dataset with 34K training images, both ResNet50 and DenseNet121 perform promisingly in the comparison (the result of ResNet50 can refer to Tab. 6.3 and Tab. 6.4). Although the number of parameters of DenseNet121 is less than one-third that of ResNet50, it shows better results.

RegDB. The proposed method is also evaluated on RegDB. Unlike SYSU-MM01, there are only 4K training images in RegDB. To reduce the risk of overfitting on the small-scale dataset, previous re-ID works such as (Huang et al., 2018a; Wang et al., 2019c; Xiao et al., 2016) normally pretrain a CNN model on a large-scale re-ID dataset and use the pretrained model to initialize parameters on the small-scale one. Thus, the ImageNet-trained DenseNet121 is pretrained on SYSU-MM01 using IDE (D^e). The pretrained IDE (D^e) (DenseNet121) is adopted to train the RegDB dataset. The DenseNet121 is used for evaluations on RegDB because it contains fewer parameters than ResNet50. Therefore, it is more suitable to fit the small-scale RegDB dataset.

Tab. 6.6 shows the comparison results. The proposed method (IDE-T (γ)+dIDeS) outperforms all the other methods in terms of mAP. However, both D-HSME and JSIA-ReID show higher performance than the proposed method in terms of rank-1 accuracy. Although D-HSME performs promisingly on RegDB, its performance on SYSU-MM01 is quite low (*e.g.*, rank-1: 20.7%). The performance of JSIA-ReID is also slightly higher than that of the proposed method on RegDB in terms of rank-1 accuracy. However, shared features learned by JSIA-ReID can benefit from cross-modality image transfer. The proposed method does not contain the func-

tion of data-level transfer, and shows significant improvements on two widely used infrared-visible LT-reID datasets simultaneously.

The AlignGAN (Wang et al., 2019a) reports a 40.7% (53.5%) mAP and 42.4% (57.1%) rank-1 accuracy on SYSU-MM01 (RegDB). AlignGAN heavily relies on data-level transfer using GAN for pixel alignment. If only the feature-level alignment between two modalities is considered without a data-level alignment, then it achieves only a 34.1% rank-1 accuracy and only a 36.2% mAP on SYSU-MM01. The performance is much lower than that of the proposed methods. Note that the performance improvement of the proposed solution is only based on the vanilla IDE and IDE-T models with the ImageNet-trained backbone to verify the effectiveness of the proposed solution in mitigating the MBT issue. The proposed solution does not use a fancy CNN architecture. In fact, the proposed solutions can be applied to other state-of-the-art infrared-visible LT-reID architectures to further boost their performance. In this revised version, the proposed method is also applied to a stronger model (*i.e.*, AGT(γ)+dIDeS, refer to Tab. 6.5), and the performance surpasses AlignGAN significantly. This result also further demonstrates the scalability of the proposed method.

6.4 Conclusion

This chapter is the first to investigate and unveil the MBT issue for infrared-visible LT-reID and propose a dual-level learning strategy to tackle this issue by regularizing the ImageNet-trained CNN in training. Rather than designing a new network architecture, the classic IDE or IDE-T is adopted to verify the effectiveness of the proposed approach. The first-level learning strategy forces the network to focus on ID-exclusive cross-modality image pairs to reduce the risk of MBT. The first-level learning strategy demonstrates the effectiveness of the proposed method in both qualitative and quantitative evaluations in experiments. To further alleviate

the MBT issue, the proposed second-level learning strategy combines real images with ground-truth ID labels and the generated third modality images with the proposed dIDeS virtual labels as inputs. Comprehensive evaluations are carried out to verify the effectiveness of the proposed dual-level learning strategy in alleviating the MBT issue for infrared-visible LT-reID. The proposed approach performs favorably against other state-of-the-art approaches without bells and whistles. In addition, since the proposed DLS is based on the vanilla IDE or IDE-T networks (without changing the network architecture), it can facilitate future infrared-visible LT-reID works that are built upon IDE.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

Person re-ID has been widely studied during the past decade. Previous works focusing on a small person re-ID task may not be suitable for real-world security surveillance, which needs to consider more aspects. In order to tackle person re-ID in the wild, this thesis focuses on four challenges, including 1) challenges on variations of illumination, viewpoint, and pose, 2) challenges on footages taken in different environments, 3) challenges on clothing change when a person being identified, and 4) challenges on footages taken by different styles of cameras. Correspondingly, this thesis discusses limitations of existing approaches on four challenges and presents new approaches to tackle issues exposed in these approaches from four aspects, *i.e.*, single-domain ST-reID, cross-domain ST-reID, clothing-change LT-reID, and infrared-visible LT-reID.

Chapter 3 addresses single-domain ST-reID using generated data with a proposed virtual label (*i.e.*, MpRL). By doing so, it can increase the diversity of data in supervised training to sort out the inevitable data limitation issue in the wild. To train a CNN, MpRL is used as a virtual label assigned to generated data. Two CNNs are adopted to show the effectiveness of the proposed MpRL. Experiments demonstrate that generated data can effectively improve the performance of the two CNNs trained with the proposed MpRL.

Chapter 4 considers the BG shift issue in reducing domain gaps for cross-domain ST-reID. SBSGAN is proposed to generate soft-mask images with BG being sup-

pressed. A DA-2S model is introduced with the proposed ISDC module to use helpful BG cues to sort out the re-ID task in the wild. To further explore/learn the natural characteristics from unlabelled target domain training data. An update strategy is given based on the proposed DA-2S network and images generated by SBSGAN. Based on DBSCAN clustering results, the proposed DCCV is used to improve the virtual label estimation quality. Experiment results demonstrate the effectiveness of the proposed approach in both qualitative and quantitative evaluations.

Chapter 5 introduces a new clothing-change LT-reID dataset called Celeb-reID. This dataset uses the street snap-shots of celebrities as the resource. Compared with previous datasets, Celeb-reID is the largest re-ID dataset with clothing change cases for each individual person. In addition, a ReIDCaps model is proposed to tackle the clothing-change challenge exposed in LT-reID. Compared with common scalar-neuron-based CNNs, vector-neuron-based capsules are used to aware of clothing changes of each person. Comprehensive experiments are given to demonstrate the superiority of the proposed method for clothing-change LT-reID in the wild.

Chapter 6 unveils the MBT issue for infrared-visible LT-reID and proposes a dual-level learning strategy to tackle this issue by regularizing the ImageNet-trained CNN during training. Rather than designing a fancy network architecture, IDE and IDE-T are used to verify the effectiveness of the proposed approach. In the first-level learning strategy, the network focuses on ID-exclusive cross-modality image pairs to reduce the risk of MBT. To further alleviate the MBT issue, the proposed second-level learning strategy combines real images with ground-truth ID labels and our generated third modality images with the proposed dIDeS virtual labels as inputs. Comprehensive evaluations are carried out to verify the effectiveness of the proposed dual-level learning strategy in alleviating the MBT issue for infrared-visible LT-reID in the wild.

In the thesis, chapters 3, 4, 5, and 6 are supported by papers published in prevailing conferences and journals. All of these papers are listed on the List of Publications.

7.2 Future Work

Due to the challenges described in this thesis, person re-ID has a long way from becoming an accurate and efficient application. In order to tackle different challenges for person re-ID in the wild, there are more works should be made for both ST-reID scenario and LT-reID scenario. In ST-reID, existing works mainly rely on clothing information to distinguish different pedestrians. However, there are very few works considering the case where many people have similar clothing. For instance, in winter, most people may wear coat with dark colours. In such a case, how to deal with the similar clothing issue is still an open question. Moreover, there is no existing work considering the impact caused by severe weather conditions for both ST-reID and LT-reID. It is conceivable that the quality of person images captured by surveillance cameras in outdoor environments can be badly damaged when there is rain, snow, or fog. Dealing with the impact caused by severe weather conditions is also required and has practical significance for both ST-reID and LT-reID in the wild. As a rarely-considered issue, clothing change is very common in LT-reID, however, it may not necessarily change in some circumstances. Therefore, how to be aware of the true clothing status for each individual person should be concerned in order to better handle clothing-change cases in LT-reID when non-clothing-change cases also exist.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. et al., 2016, 'Tensorflow: A system for large-scale machine learning', *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 265–283.

Abdulla, W., 2017, 'Mask r-cnn for object detection and instance segmentation on keras and tensorflow', .

Ahmed, E., Jones, M. & Marks, T. K., 2015, 'An improved deep learning architecture for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916.

Ali, S., Javed, O., Haering, N. & Kanade, T., 2010, 'Interactive retrieval of targets for wide area surveillance', *ACM Multimedia Conference (ACM MM)*, pp. 895–898.

Arjovsky, M., Chintala, S. & Bottou, L., 2017, 'Wasserstein gan', *arXiv preprint arXiv:1701.07875.*

Arthur, D. & Vassilvitskii, S., 2006, 'k-means++: The advantages of careful seeding', Tech. rep., Stanford.

Bai, S., Bai, X. & Tian, Q., 2017, 'Scalable person re-identification on supervised smoothed manifold', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3356–3365.

Bak, S., Carr, P. & Lalonde, J., 2018, 'Domain adaptation through synthesis for

unsupervised person re-identification', *European conference on computer vision (ECCV)*, Springer, pp. 189–205.

Barbosa, I. B., Cristani, M., Del Bue, A., Bazzani, L. & Murino, V., 2012a, 'Re-identification with rgb-d sensors', *European conference on computer vision (ECCV)*, Springer, pp. 433–442.

Barbosa, I. B., Cristani, M., Del Bue, A., Bazzani, L. & Murino, V., 2012b, 'Re-identification with rgb-d sensors', *European conference on computer vision (ECCV)*, Springer, pp. 433–442.

Campello, R. J., Moulavi, D. & Sander, J., 2013, 'Density-based clustering based on hierarchical density estimates', *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 160–172.

Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y., 2017, 'Realtime multi-person 2d pose estimation using part affinity fields', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299.

Chang, X., Hospedales, T. M. & Xiang, T., 2018, 'Multi-level factorisation net for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2109–2118.

Chen, D., Yuan, Z., Chen, B. & Zheng, N., 2016, 'Similarity learning with spatial constraints for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1268–1277.

Chen, D., Zhang, S., Ouyang, W., Yang, J. & Tai, Y., 2018, 'Person search via a mask-guided two-stream cnn model', *European conference on computer vision (ECCV)*, Springer, pp. 734–750.

Chen, G., Lin, C., Ren, L., Lu, J. & Zhou, J., 2019a, 'Self-critical attention learning

for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 9637–9646.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z. & Wang, Z., 2019b, 'Abd-net: Attentive but diverse person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 8351–8361.

Chen, W., Chen, X., Zhang, J. & Huang, K., 2017a, 'Beyond triplet loss: a deep quadruplet network for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1320–1329.

Chen, Y., Zheng, W. & Lai, J., 2015, 'Mirror representation for modeling view-specific transform in person re-identification', *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3402–3408.

Chen, Y., Zhu, X. & Gong, S., 2017b, 'Person re-identification by deep learning multi-scale representations', *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2590–2600.

Chen, Y., Zhu, X. & Gong, S., 2019c, 'Instance-guided context rendering for cross-domain person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 232–242.

Cheng, D., Gong, Y., Zhou, S., Wang, J. & Zheng, N., 2016, 'Person re-identification by multi-channel parts-based cnn with improved triplet loss function', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1335–1344.

Choi, S., Lee, S., Kim, Y., Kim, T. & Kim, C., 2020, 'Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10257–10266.

Choi, Y., Choi, M. & Kim, M., 2018, 'Stargan: Unified generative adversarial networks for multi-domain image-to-image translation', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8789–8797.

Dai, P., Ji, R., Wang, H., Wu, Q. & Huang, Y., 2018, 'Cross-modality person re-identification with generative adversarial training', *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 677–683.

Dai, Z., Chen, M., Gu, X., Zhu, S. & Tan, P., 2019, 'Batch dropblock network for person re-identification and beyond', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3691–3701.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y. & Jiao, J., 2018, 'Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 994–1003.

Ding, G., Zhang, S., Khan, S., Tang, Z., Zhang, J. & Porikli, F., 2019, 'Feature affinity-based pseudo labeling for semi-supervised person re-identification', *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2891–2902.

Ding, S., Lin, L., Wang, G. & Chao, H., 2015, 'Deep feature learning with relative distance comparison for person re-identification', *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al., 1996, 'A density-based algorithm for discovering clusters in large spatial databases with noise', *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 226–231.

Fan, H., Zheng, L., Yan, C. & Yang, Y., 2018, 'Unsupervised person re-identification: Clustering and fine-tuning', *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, pp. 1–18.

Farenzena, M., Bazzani, L., Perina, A., Murino, V. & Cristani, M., 2010, 'Person re-identification by symmetry-driven accumulation of local features', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2360–2367.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D., 2010, 'Object detection with discriminatively trained part-based models', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645.

Feng, Z., Lai, J. & Xie, X., 2020, 'Learning modality-specific representations for visible-infrared person re-identification', *IEEE Transactions on Image Processing*, vol. 29, pp. 579–590.

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H. & Huang, T. S., 2019, 'Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 6112–6121.

Garcia, J., Martinel, N., Micheloni, C. & Gardel, A., 2015, 'Person re-identification ranking optimisation by discriminant context information analysis', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1305–1313.

Geng, M., Wang, Y., Xiang, T. & Tian, Y., 2016, 'Deep transfer learning for person re-identification', *arXiv preprint arXiv:1611.05244*.

Gong, S., Cristani, M., Yan, S. & Loy, C. C., 2014, *Person re-identification*, Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y., 2014, 'Generative adversarial nets', *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680.

Gray, D. & Tao, H., 2008, 'Viewpoint invariant pedestrian recognition with an ensemble of localized features', *European conference on computer vision (ECCV)*, Springer, pp. 262–275.

Guanshuo, W., Yufeng, Y., Xiong, C., Jiwei, L. & Xi, Z., 2018, 'Learning discriminative features with multiple granularities for person re-identification', *ACM Multimedia Conference (ACM MM)*, pp. 274–282.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C., 2017, 'Improved training of wasserstein gans', *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5767–5777.

Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J.-G. & Han, K., 2019, 'Beyond human parts: Dual part-aligned representations for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3642–3651.

Hao, Y., Wang, N., Li, J. & Gao, X., 2019, 'Hsme: hypersphere manifold embedding for visible thermal person re-identification', *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 8385–8392.

Haque, A., Alahi, A. & Fei-Fei, L., 2016, 'Recurrent attention models for depth-based person identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1229–1238.

He, K., Gkioxari, G., Dollár, P. & Girshick, R., 2017, 'Mask r-cnn', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969.

He, K., Zhang, X., Ren, S. & Sun, J., 2016, 'Deep residual learning for image recognition', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

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

Hu, J., Shen, L. & Sun, G., 2018a, 'Squeeze-and-excitation networks', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141.

Hu, J., Shen, L. & Sun, G., 2018b, 'Squeeze-and-excitation networks', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q., 2017a, 'Densely connected convolutional networks.', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708.

Huang, L., Yang, Q., Wu, J., Huang, Y., Wu, Q. & Xu, J., 2020, 'Generated data with sparse regularized multi-pseudo label for person re-identification', *IEEE Signal Processing Letters*, vol. 27, pp. 391–395.

Huang, Y., Sheng, H., Liu, Y., Zheng, Y. & Xiong, Z., 2015, 'Person re-identification by unsupervised color spatial pyramid matching', *International Conference on Knowledge Science, Engineering and Management (KSEM)*, Springer, pp. 799–810.

Huang, Y., Sheng, H. & Xiong, Z., 2016, 'Person re-identification based on hierarchical bipartite graph matching', *The IEEE International Conference on Image Processing (ICIP)*, pp. 4255–4259.

Huang, Y., Sheng, H., Zheng, Y. & Xiong, Z., 2017b, 'Deepdiff: Learning deep difference features on human body parts for person re-identification', *Neurocomputing*, vol. 241, pp. 191–203.

Huang, Y., Wu, Q., Xu, J. & Zhong, Y., 2019a, 'Celebrities-reid: A benchmark for clothes variation in long-term person re-identification', *The International Joint Conference on Neural Network (IJCNN)*, pp. 1–8.

Huang, Y., Wu, Q., Xu, J. & Zhong, Y., 2019b, 'Sbsgan: Suppression of inter-domain background shift for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 9527–9536.

Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, P. & Zhang, Z., 2021a, 'Alleviating modality bias training for infrared-visible person re-identification', *IEEE Transactions on Multimedia*.

Huang, Y., Wu, Q., Xu, J., Zhong, Y. & Zhang, Z., 2021b, 'Unsupervised domain adaptation with background shift mitigating for person re-identification', *International Journal of Computer Vision*.

Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z. & Zhang, J., 2018a, 'Multi-pseudo regularized label for generated data in person re-identification', *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1391–1403.

Huang, Y., Xu, J., Wu, Q., Zhong, Y., Zhang, P. & Zhang, Z., 2019c, 'Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3459–3471.

Huang, Y., Zhong, Y., Wu, Q., Dutkiewicz, E. & Jiang, T., 2018b, 'Cost-effective foliage penetration human detection under severe weather conditions based on auto-encoder/decoder neural network', *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6190–6200.

Ioffe, S. & Szegedy, C., 2015, 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', *International Conference on Machine Learning (ICML)*, pp. 448—-456.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A., 2017, 'Image-to-image translation with conditional adversarial networks', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134.

Kansal, K., Subramanyam, A., Wang, Z. & Satoh, S., 2020, 'Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification',

IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3422–3432.

Kingma, D. & Ba, J., 2014, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980.*

Laptev, I., Marszalek, M., Schmid, C. & Rozenfeld, B., 2008, 'Learning realistic human actions from movies', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

Lee, D.-H., 2013, 'Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks', *International Conference on Machine Learning Workshops (ICMLW)*, .

Li, D., Chen, X., Zhang, Z. & Huang, K., 2017a, 'Learning deep context-aware features over body and latent parts for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 384–393.

Li, M., Zhu, X. & Gong, S., 2018a, 'Unsupervised person re-identification by deep learning tracklet association', *European conference on computer vision (ECCV)*, Springer, pp. 737–753.

Li, W., Zhao, R. & Wang, X., 2012, 'Human reidentification with transferred metric learning', *Asian Conference on Computer Vision (ACCV)*, Springer, pp. 31–44.

Li, W., Zhao, R., Xiao, T. & Wang, X., 2014, 'Deepreid: Deep filter pairing neural network for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159.

Li, W., Zhu, X. & Gong, S., 2017b, 'Person re-identification by deep joint learning of multi-loss classification', *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2194–2200.

Li, W., Zhu, X. & Gong, S., 2018b, 'Harmonious attention network for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2285–2294.

Li, Y.-J., Lin, C.-S., Lin, Y.-B. & Wang, Y.-C. F., 2019a, 'Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 7919–7929.

Li, Y.-J., Luo, Z., Weng, X. & Kitani, K. M., 2020a, 'Learning shape representations for person re-identification under clothing change', *IEEE Winter Conference on Applications of Computer Vision (WACV)*, .

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L. & Smith, J. R., 2013, 'Learning locally-adaptive decision functions for person verification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3610–3617.

Li, Z., Zhang, J., Gong, Y., Yao, Y. & Wu, Q., 2020b, 'Field-wise learning for multi-field categorical data', *Advances in Neural Information Processing Systems (NeurIPS)*, .

Li, Z., Zhang, J., Wu, Q., Gong, Y., Yi, J. & Kirsch, C., 2019b, 'Sample adaptive multiple kernel learning for failure prediction of railway points', *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2848–2856.

Li, Z., Zhang, J., Wu, Q. & Kirsch, C., 2018c, 'Field-regularised factorization machines for mining the maintenance logs of equipment', *Australasian Joint Conference on Artificial Intelligence*, Springer, pp. 172–183.

Liang, X., Gong, K., Shen, X. & Lin, L., 2018, 'Look into person: Joint body parsing & pose estimation network and a new benchmark', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 871–885.

Liao, S., Hu, Y., Zhu, X. & Li, S. Z., 2015, 'Person re-identification by local maximal occurrence representation and metric learning', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2197–2206.

Liao, S. & Li, S. Z., 2015, 'Efficient psd constrained asymmetric metric learning for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3685–3693.

Lin, J., Ren, L., Lu, J., Feng, J. & Zhou, J., 2017, 'Consistent-aware deep learning for person re-identification in a camera network', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5771–5780.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L., 2014, 'Microsoft coco: Common objects in context', *European conference on computer vision (ECCV)*, Springer, pp. 740–755.

Liu, C., Change Loy, C., Gong, S. & Wang, G., 2013, 'Pop: Person re-identification post-rank optimisation', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 441–448.

Liu, D., Huang, Y., Wu, Q., Ma, R. & An, P., 2020, 'Multi-angular epipolar geometry based light field angular reconstruction network', *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1507–1522.

Liu, J., Zha, Z.-J., Chen, D., Hong, R. & Wang, M., 2019, 'Adaptive transfer network for cross-domain person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7202–7211.

Luo, C., Chen, Y., Wang, N. & Zhang, Z., 2019, 'Spectral feature transformation for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 4976–4985.

Luo, H., Jiang, W., Fan, X. & Zhang, C., 2020a, 'Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification', *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2905–2913.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S. & Gu, J., 2020b, 'A strong baseline and batch normalization neck for deep person re-identification', *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609.

Matsukawa, T., Okabe, T., Suzuki, E. & Sato, Y., 2016, 'Hierarchical gaussian descriptor for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1363–1372.

Munaro, M., Basso, A., Fossati, A., Van Gool, L. & Menegatti, E., 2014, '3d reconstruction of freely moving persons for re-identification with a depth sensor', *The IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4512–4519.

Nair, V. & Hinton, G. E., 2010, 'Rectified linear units improve restricted boltzmann machines', *International Conference on Machine Learning (ICML)*, pp. 807–814.

Nguyen, D., Hong, H., Kim, K. & Park, K., 2017, 'Person recognition system based on a combination of body images from visible light and thermal cameras', *Sensors*, vol. 17, p. 605.

Odena, A., 2016, 'Semi-supervised learning with generative adversarial networks', *arXiv preprint arXiv:1606.01583*.

Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y. & Gao, Y., 2019, 'A novel unsupervised camera-aware domain adaptation framework for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 8080–8089.

Qian, X., Fu, Y., Jiang, Y.-G., Xiang, T. & Xue, X., 2017, 'Multi-scale deep learning

architectures for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 5409–5418.

Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.-G. & Xue, X., 2020, 'Long-term cloth-changing person re-identification', *arXiv preprint arXiv:2005.12633*.

Radford, A., Metz, L. & Chintala, S., 2015, 'Unsupervised representation learning with deep convolutional generative adversarial networks', *arXiv preprint arXiv:1511.06434*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C., 2016, 'Performance measures and a data set for multi-target, multi-camera tracking', *European conference on computer vision (ECCV)*, Springer, pp. 17–35.

Sabour, S., Frosst, N. & Hinton, G. E., 2017, 'Dynamic routing between capsules', *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3856–3866.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. & Chen, X., 2016, 'Improved techniques for training gans', *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2234–2242.

Sheng, H., Huang, Y., Zheng, Y., Chen, J. & Xiong, Z., 2015, 'Person re-identification via learning visual similarity on corresponding patch pairs', *International Conference on Knowledge Science, Engineering and Management (KSEM)*, Springer, pp. 787–798.

Sheng, H., Zhang, B., Huang, Y., Zheng, Y. & Xiong, Z., 2016, 'Discriminative dictionary learning sparse coding for person re-identification', *The IEEE Intelligent Vehicles Symposium (IV)*, pp. 1338–1343.

Song, C., Huang, Y., Ouyang, W. & Wang, L., 2018, 'Mask-guided contrastive

attention model for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1188.

Song, J., Yang, Y., Song, Y.-Z., Xiang, T. & Hospedales, T. M., 2019, 'Generalizable person re-identification by domain-invariant mapping network', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 719–728.

Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C. & Wang, X., 2020, 'Unsupervised domain adaptive re-identification: Theory and practice', *Pattern Recognition*, vol. 102, pp. 107–173.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R., 2014, 'Dropout: a simple way to prevent neural networks from overfitting', *Journal of Machine Learning Research (JMLR)*, vol. 15, no. 1, pp. 1929–1958.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W. & Tian, Q., 2017, 'Pose-driven deep convolutional model for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3980–3989.

Sun, Y., Zheng, L., Deng, W. & Wang, S., 2017, 'Svdnet for pedestrian retrieval', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3820–3828.

Sun, Y., Zheng, L., Li, Y., Yang, Y., Tian, Q. & Wang, S., 2019, 'Learning part-based convolutional features for person re-identification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, Y., Zheng, L., Yang, Y., Tian, Q. & Wang, S., 2018, 'Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)', *European conference on computer vision (ECCV)*, Springer, pp. 501–518.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z., 2016, 'Rethinking the inception architecture for computer vision', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826.

Taigman, Y., Polyak, A. & Wolf, L., 2017, 'Unsupervised cross-domain image generation', *International Conference on Learning Representations (ICLR)*, .

Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J. & Wang, X., 2018, 'Eliminating background-bias for robust person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5794–5803.

Van Der Maaten, L., 2014a, 'Accelerating t-sne using tree-based algorithms', *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245.

Van Der Maaten, L., 2014b, 'Accelerating t-sne using tree-based algorithms', *Journal of Machine Learning Research*, vol. 15, pp. 3221–3245.

Varior, R. R., Haloi, M. & Wang, G., 2016, 'Gated siamese convolutional neural network architecture for human re-identification', *European conference on computer vision (ECCV)*, Springer, pp. 791–808.

Vedaldi, A. & Lenc, K., 2015, 'Matconvnet: Convolutional neural networks for matlab', *ACM Multimedia Conference (ACM MM)*, pp. 689–692.

Wan, C., Wu, Y., Tian, X., Huang, J. & Hua, X.-S., 2020a, 'Concentrated local part discovery with fine-grained part representation for person re-identification', *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1605–1618.

Wan, F., Wu, Y., Qian, X., Chen, Y. & Fu, Y., 2020b, 'When person re-identification meets changing clothes', *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3620–3628.

Wang, F., Zuo, W., Lin, L., Zhang, D. & Zhang, L., 2016a, 'Joint learning of single-image and cross-image representations for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1288–1296.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y. & Hou, Z., 2019a, 'Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3622–3631.

Wang, G.-A., Yang, T. Z., Cheng, J., Chang, J., Liang, X., Hou, Z. et al., 2020, 'Cross-modality paired-images generation for rgb-infrared person re-identification', *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12144–12151.

Wang, H., Gong, S., Zhu, X. & Xiang, T., 2016b, 'Human-in-the-loop person re-identification', *European conference on computer vision (ECCV)*, Springer, pp. 405–422.

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L., 2013, 'Dense trajectories and motion boundary descriptors for action recognition', *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79.

Wang, J., Zhu, X., Gong, S. & Li, W., 2018, 'Transferable joint attribute-identity deep learning for unsupervised person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2275–2284.

Wang, Z., Hu, R., Liang, C., Yu, Y., Jiang, J., Ye, M., Chen, J. & Leng, Q., 2015, 'Zero-shot person re-identification via cross-view consistency', *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272.

Wang, Z., Jiang, J., Wu, Y., Ye, M., Bai, X. & Satoh, S., 2019b, 'Learning sparse and identity-preserved hidden attributes for person re-identification', *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2013–2025.

Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.-Y. & Satoh, S., 2019c, 'Learning to reduce dual-level discrepancy for infrared-visible person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 618–626.

Wang, Z., Wang, Z., Zheng, Y., Wu, Y., Zeng, W. & Satoh, S., 2019d, 'Beyond intra-modality: A survey of heterogeneous person re-identification', *arXiv preprint arXiv:1905.10048*.

Wei, L., Zhang, S., Gao, W. & Tian, Q., 2018a, 'Person transfer gan to bridge domain gap for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 79–88.

Wei, L., Zhang, S., Yao, H., Gao, W. & Tian, Q., 2018b, 'Glad: Global–local-alignment descriptor for scalable person re-identification', *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 986–999.

Wu, A., Zheng, W.-S. & Lai, J.-H., 2019a, 'Unsupervised person re-identification by camera-aware similarity consistency learning', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 6922–6931.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S. & Lai, J., 2017, 'Rgb-infrared cross-modality person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 5380–5389.

Wu, J., Yao, L., Huang, Y., Xu, J., Wu, Q. & Huang, L., 2019b, 'Improving person re-identification performance using body mask via cross-learning strategy', *The IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4.

Xiao, T., Li, H., Ouyang, W. & Wang, X., 2016, 'Learning deep feature representations with domain guided dropout for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1249–1258.

Xiao, T., Li, S., Wang, B., Lin, L. & Wang, X., 2017, 'Joint detection and identification feature learning for person search', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3376–3385.

Yang, F., Wang, Z., Xiao, J. & Satoh, S., 2020, 'Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval', *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12589–12596.

Yang, Q., Wu, A. & Zheng, W.-S., 2019a, 'Person re-identification by contour sketch under moderate clothing change', *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, Q., Yu, H.-X., Wu, A. & Zheng, W.-S., 2019b, 'Patch-based discriminative feature learning for unsupervised person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3633–3642.

Ye, M., Lan, X., Li, J. & Yuen, P. C., 2018a, 'Hierarchical discriminative learning for visible thermal person re-identification', *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7501–7508.

Ye, M., Lan, X., Wang, Z. & Yuen, P. C., 2019, 'Bi-directional center-constrained top-ranking for visible thermal person re-identification', *IEEE Transactions on Information Forensics and Security*, pp. 407–419.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. & Hoi, S. C., 2020, 'Deep learning for person re-identification: A survey and outlook', *arXiv preprint arXiv:2001.04193*.

Ye, M., Wang, Z., Lan, X. & Yuen, P. C., 2018b, 'Visible thermal person re-identification via dual-constrained top-ranking.', *Proc. Int. Joint Conf. Artif. Intell.*, pp. 1092–1099.

Yu, H., Wu, A. & Zheng, W., 2017a, 'Cross-view asymmetric metric learning for unsupervised person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 994–1002.

Yu, H.-X., Zheng, W.-S., Wu, A., Guo, X., Gong, S. & Lai, J.-H., 2019a, 'Unsupervised person re-identification by soft multilabel learning', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2148–2157.

Yu, Q., Chang, X., Song, Y.-Z., Xiang, T. & Hospedales, T. M., 2017b, 'The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching', *arXiv:1711.08106*.

Yu, S., Li, S., Chen, D., Zhao, R., Yan, J. & Qiao, Y., 2020, 'Cocas: A large-scale clothes changing person dataset for re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3400–3409.

Yu, T., Li, D., Yang, Y., Hospedales, T. M. & Xiang, T., 2019b, 'Robust person re-identification by modelling feature uncertainty', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 552–561.

Yuan, Y., Zhang, J. & Wang, Q., 2018, 'Modeling unknown class centers for metric learning on person re-identification', *IEEE Access*, vol. 6, pp. 40602–40610.

Yumin, S., Jingdong, W., Siyu, T., Tao, M. & Kyoung, M. L., 2018, 'Part-aligned bilinear representations for person re-identification', *European conference on computer vision (ECCV)*, Springer, pp. 418–437.

Zeng, Z., Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.-Y. & Satoh, S., 2020, 'Illumination-adaptive person re-identification', *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3064–3074.

Zhang, J., Yuan, Y. & Wang, Q., 2019a, 'Night person re-identification and a benchmark', *IEEE Access*, vol. 7, pp. 95496–95504.

Zhang, L., Xiang, T. & Gong, S., 2016, 'Learning a discriminative null space for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1239–1248.

Zhang, L., Xu, J., Zhang, J. & Gong, Y., 2018a, 'Information enhancement for travelogues via a hybrid clustering model', *Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, pp. 1–8.

Zhang, L., Zhang, J., Li, Z. & Xu, J., 2020a, 'Towards better graph representation: Two-branch collaborative graph neural networks for multimodal marketing intention detection', *IEEE International Conference on Multimedia and Expo*, pp. 1–6.

Zhang, P., Wu, Q., Xu, J. & Zhang, J., 2018b, 'Long-term person re-identification using true motion from videos', *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 494–502.

Zhang, P., Xu, J., Wu, Q., Huang, Y. & Ben, X., 2020b, 'Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild', *IEEE Transactions on Multimedia*.

Zhang, P., Xu, J., Wu, Q., Huang, Y. & Zhang, J., 2020c, 'Top-push constrained modality-adaptive dictionary learning for cross-modality person re-identification', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4554–4566.

Zhang, X., Cao, J., Shen, C. & You, M., 2019b, 'Self-training with progressive augmentation for unsupervised cross-domain person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 8222–8231.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X. & Tang, X., 2017a, 'Spindle net: Person re-identification with human body region guided feature decomposition and fusion', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1077–1085.

Zhao, L., Li, X., Wang, J. & Zhuang, Y., 2017b, 'Deeply-learned part-aligned representations for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3239–3248.

Zhedong, Z., Liang, Z. & Yi, Y., 2017, 'Unlabeled samples generated by GAN improve the person re-identification baseline in vitro', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3774–3782.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. & Tian, Q., 2015, 'Scalable person re-identification: A benchmark', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124.

Zheng, L., Yang, Y. & Hauptmann, A. G., 2016a, 'Person re-identification: Past, present and future', *arXiv preprint arXiv:1610.02984.*

Zheng, L., Zhang, H., Sun, S., Chandraker, M. & Tian, Q., 2016b, 'Person re-identification in the wild', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3346–3355.

Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y. & Tian, Q., 2017a, 'Person re-identification in the wild', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1367–1376.

Zheng, W.-S., Gong, S. & Xiang, T., 2011, 'Person re-identification by probabilistic relative distance comparison', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 649–656.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y. & Kautz, J., 2019, 'Joint discriminative and generative learning for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2138–2147.

Zheng, Z., Zheng, L. & Yang, Y., 2017b, 'Unlabeled samples generated by gan

improve the person re-identification baseline in vitro', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3754–3762.

Zheng, Z., Zheng, L. & Yang, Y., 2018, 'A discriminatively learned cnn embedding for person reidentification', *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, pp. 1–20.

Zhong, Y., Wang, J., Wu, S., Jiang, T., Huang, Y. & Wu, Q., 2020, 'Multi-location human activity recognition via mimo-ofdm based wireless networks: An iot-inspired device-free sensing approach', *IEEE Internet of Things Journal*.

Zhong, Y., Yang, Y., Zhu, X., Huang, Y., Dutkiewicz, E., Zhou, Z. & Jiang, T., 2018a, 'Impact of seasonal variations on foliage penetration experiment: a wsn-based device-free sensing approach', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5035–5045.

Zhong, Z., Zheng, L., Cao, D. & Li, S., 2017, 'Re-ranking person re-identification with k-reciprocal encoding', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3652–3661.

Zhong, Z., Zheng, L., Li, S. & Yang, Y., 2018b, 'Generalizing a person retrieval model hetero-and homogeneously', *European Conference on Computer Vision (ECCV)*, Springer, pp. 172–188.

Zhong, Z., Zheng, L., Luo, Z., Li, S. & Yang, Y., 2019, 'Invariance matters: Exemplar memory for domain adaptive person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 598–607.

Zhou, K. & Xiang, T., 2019, 'Torchreid: A library for deep learning person re-identification in pytorch', *arXiv preprint arXiv:1910.10093*.

Zhou, K., Yang, Y., Cavallaro, A. & Xiang, T., 2019, 'Omni-scale feature learning

for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3702–3712.

Zhou, S., Wang, J., Wang, J., Gong, Y. & Zheng, N., 2017, 'Point to set similarity based deep feature learning for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5028–5037.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A., 2017, 'Unpaired image-to-image translation using cycle-consistent adversarial networks', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232.