UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

LONG-TERM PERSON RE-IDENTIFICATION IN THE WILD

by

Peng Zhang

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

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Authorship/Originality

I, Peng Zhang, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference 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: 16/06/2020

Acknowledgements

It is a cherished memory to pursue PhD in UTS. I would like to express my sincere gratitude to all those who help me complete my doctoral study.

I would like to express my deepest appreciation, first and foremost, to my supervisor, A/Prof. Qiang Wu, for his professional guidance and warm encouragement. I am deeply impressed by his insight understanding on computer vision and his rich skills on research writing and presentation. He can always inspire me with fancy ideas and rigorous logic. His enthusiasm, attitude and devotion towards academy have deeply influenced me, which provides instructions to my future career. It is the luckiest thing that has him as my supervisor.

I also want to express my sincere attitude to my co-supervisor Dr. Jingsong Xu and A/Prof. Jian Zhang. Jingsong gives me a lot of constructive suggestions for my research and helps me polish papers. Besides, he always encourages me to focus on the advanced techniques that motivate the research going deeper. Jian organized many interesting seminars that provide us opportunities to share and communicate our research progress.

Then, I wish to give thanks to A/Prof. Xianye Ben, who was my supervisor during the master study in Shandong University. She guided me into the field of computer vision and provided me endless support to pursue PhD study. Without her recommendation, I might not have opportunity to study in UTS and work with A/Prof. Qiang Wu.

And, I appreciate the help, support and friendship from my dear colleagues and friends during my doctoral study. Thanks to Yifan Zuo, Zongjian Zhang, Qian Li, Yan Huang, Xunxiang Yao, Lingxiang Yao, Muming Zhao, Xiaoshui Huang, Junjie Zhang, Lina Li, Lu Zhang, Yongshun Gong, Zhibin Li and all other labmates for their

collaboration and discussion. Without their help, I cannot collect my experimental data. I am also grateful to all my friends in Sydney: Lei Sang, Lin Zhu, Tao Shen, Mengyao Li, Xiaolin Zhang, Wentao Li, Mingjie Li, Zhuo Tang, Xin Ba and Shuo Yang for their encouragement and companion. Thank you guys that bring me to try delicious food and visit beautiful landmarks in Sydney.

Finally, I would like to express my sincere thanks to my parents and girlfriend Weiyu for their endless support, trust, encouragement and love throughout my studies these years.

Peng Zhang

February 2020 @ UTS.

List of Publications

Journal Papers

J-1. **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*, Early Access, 2019.

J-2. X. Ben, **P. Zhang**, Z. Lai, R. Yan, X. Zhai and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognition*, vol. 90, pp. 87-98, 2019.

J-3. X. Yao, Q. Wu, **P. Zhang** and F. Bao, "Adaptive rational fractal interpolation function for image super-resolution via local fractal analysis," *Image and Vision Computing*, vol. 82, pp. 39-49, 2019.

J-4. X. Ben, C. Gong, **P. Zhang**, X. Jia, Q. Wu and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3142-3157, 2019.

J-5. X. Ben, C. Gong, **P. Zhang**, R. Yan, Q. Wu and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 734-747, 2020.

J-6. 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*, Early Access, 2019.

J-7. X. Yao, Q. Wu, **P. Zhang** and F. Bao, "Weighted Adaptive Image Super-Resolution Scheme based on Local Fractal Feature and Image Roughness", *IEEE Transactions on Multimedia*, Accepted, 2020.

Conference Papers

C-1. **P. Zhang**, Q. Wu and J. Xu, "VT-GAN: View Transformation GAN for Gait RecognitionAcross Views," *The International Joint Conference on Neural Network (IJCNN)*, Budapest, 14-19 July, 2019.

C-2. **P. Zhang**, Q. Wu and J. Xu, "VN-GAN: Identity-preserved Variation Normalizing GAN for Gait Recognition," *The International Joint Conference on Neural Network (IJCNN)*, Budapest, 14-19 July, 2019.

C-3. **P. Zhang**, Q. Wu, J. Xu and J. Zhang, "Long-term person re-identification using true motion from videos," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 12-15 March, 2018, pp. 494-502.

C-4. H. Song, H. Dong, and **P. Zhang**, "A virtual instrument for diagnosis to substation groundinggrids in harsh electromagnetic environment," *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Turin, 22-25 May, 2017.

Submitted Papers

J-1. **P. Zhang**, Q. Wu, J. Xu and Y. Huang, "Learning Hybrid Representations over Walking Tracklet for Long-term Person Re-Identification in The Wild," *IEEE Transactions on Multimedia*, Under review, 2020.

J-2. Y. Huang, Q. Wu, J. Xu, Y. Zhong, **P. Zhang** and Z. Zhang, "Learning from Decoupled Semantic Cue for Infrared-Visible Person Re-identification", *IEEE Transactions on Information Forensics and Security*, Under review, 2020.

J-3. 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*, Under review, 2020.

Contents

List of Figures

List of Tables

ABSTRACT

LONG-TERM PERSON RE-IDENTIFICATION IN THE WILD

by

Peng Zhang

Person re-identification (re-ID) has been attracting extensive research interest because of its non-fungible position in applications such as surveillance security, criminal investigation and forensic reasoning. Existing works assume that pedestrians keep their clothes unchanged while passing across disjoint cameras in a short period. It narrows person re-ID to a short-term problem and incurs solutions using appearance-based similarity measurement. However, this assumption is not always true in practice. For example, pedestrians are high likely to re-appear after a long-time period, such as several days. This emerging problem is termed as long-term person re-ID (LT-reID).

Regarding different types of sensors deployed, LT-reID is divided into two subtasks: person re-ID after a long-time gap (LTG-reID) and cross-camera-modality person re-ID (CCM-reID). LTG-reID utilizes only RGB cameras, while CCM-reID employs different types of sensors. Besides challenges in classical person re-ID, CCM-reID faces additional data distribution discrepancy caused by modality difference, and LTG-reID suffers severe within-person appearance inconsistency caused by clothing changes. These variations seriously degrade the performance of existing re-ID methods.

To address the aforementioned problems, this thesis investigates LT-reID from four aspects: motion pattern mining, view bias mitigation, cross-modality matching and hybrid representation learning. Motion pattern mining aims to address LTG-reID by crafting true motion information. To this point, a fine motion encoding method is proposed, which extracts motion patterns hierarchically by encod-

ing trajectory-aligned descriptors with Fisher vectors in a spatial-aligned pyramid. View bias mitigation targets on narrowing discrepancy caused by viewpoint difference. This thesis proposes two solutions: VN-GAN normalizes gaits from various views into a unified one, and VT-GAN achieves view transformation between gaits from any two views. Cross-modality matching aims to learn modality-invariant representations. To this end, this thesis proposes to asymmetrically project heterogeneous features across modalities onto a modality-agnostic space and simultaneously reconstruct the projected data using a shared dictionary on the space. Hybrid representation learning explores both subtle identity properties and motion patterns. Regarding that, a two-stream network is proposed: the space-time stream performs on image sequences to learn identity-related patterns, e.g., body geometric structure and movement, and skeleton motion stream operates on normalized 3D skeleton sequences to learn motion patterns.

Moreover, two datasets particular for LTG-reID are presented: Motion-reID is collected by two real-world surveillance cameras, and CVID-reID involves tracklets clipped from street-shot videos of celebrities on the Internet. Both datasets include abundant within-person cloth variations, highly dynamic background and diverse camera viewpoints, which promote the development of LT-reID research.

Abbreviation

CCM-reID - cross camera modality person re-identification

CLT-reID -contemporary long-term person re-identification

CMC - cumulative matching characteristic

CNN -convolutional neural network

CST-reID - conventional short-term person re-identification

CVGLT-reID - cross-view gait-based long-term person re-ID

DT -dense trajectory

FITD - fine motion encoding

GAN - generative adversarial network

GCN - graph convolutional network

GEI - gait energy image

GMM - Gaussian mixture model

LT-reID - log-term person re-ID

LTG-reID -person re-ID after long-time gap

mAP - mean average precision

PCA - principle component analysis

re-ID - re-identification

SILTP - Scale Invariant Ternary Pattern

SOTA - state-of-the-art

TCMDL - top-push constrained modality-adaptive dictionary learning

TSI - Target subject of interest

VN-GAN - variational normalizing generative adversarial network

VT-GAN - view transformation generative adversarial network

Chapter 1

Introduction

1.1 Background

With growing concern on public security and social order, large surveillance networks are deployed in public places, e.g. shopping malls, teaching buildings, hospitals, airports, subway station and parks. These surveillance cameras span most public areas with disjoint views to provide seamless coverage. Such surveillance networks generate large volumes of video stream data, which is either monitored by the security department to keep public order or utilized to provide digital evidence for forensic purposes. With the explosively increasing of video data, tedious video analysis by human force is time-consuming and labour expensive. Thus, automated analysis of these large volumes of video data is necessary. Such analysis is vital for timely alerts by detecting suspicious behaviours and predicting undesirable events.

Regarding this, comprehension of a surveillance scenario by automated machine understanding is required to detect and track people across multiple cameras. A need for tracking people across multiple non-overlapping cameras is particularly urgent, and person re-ID is an avoidable part of multi-camera tracking by bridging the tracked target across scenes. Person re-ID refers to identifying a target pedestrian across a network of disjoint surveillance cameras at distinct locations and time. According to the time interval, it can be divided into two categories, i.e., short-term person re-ID and long-term person re-ID. Most existing works focus on the short-term person re-ID problem while the long-term person re-ID is merely researched.

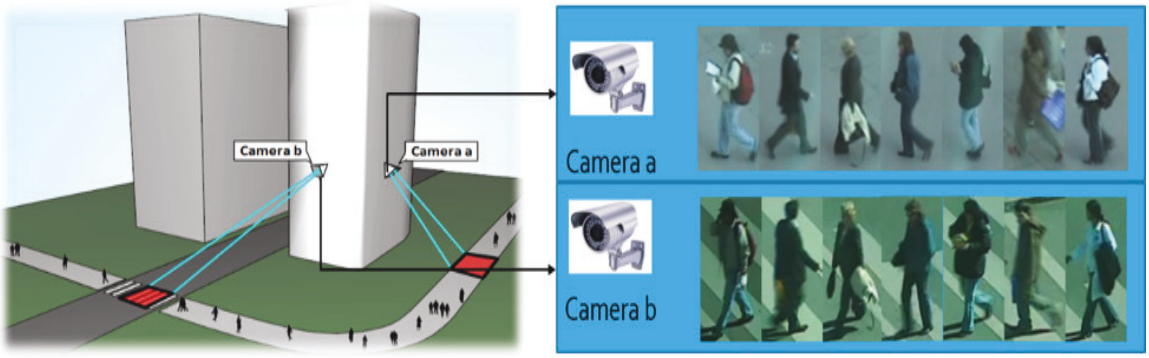


Figure 1.1 : An example of the CST-reID setting

1.1.1 Conventional Short-term Person re-ID

Figure 1.1 shows an example of conventional short-term person re-ID (CST-reID) setting. Two cameras are deployed to monitor the only route with non-overlapping surveillance views. As a person walks from monitoring region of Camera #a to Camera #b, person re-ID bridges the discontinuous tracking between cameras. From the perspective, person re-ID is also critical in many other related fields, such as human-computer interaction and automated multi-camera person retrieval, in addition to applications on surveillance security.

As shown in Figure 1.2, given a probe image (video sequence) captured by Camera #a, the aim of person re-ID is to search where and when the same person will appear in the Camera #b. Figure 1.2 lists the first ten matches in the gallery set with a specific query sample, from left to right, from top to bottom. Figure 1.3 gives pipeline of a complete person re-ID system that includes four steps: people detection, person tracking, feature extraction and classification. However, human detection and person tracking are usually regarded as two independent fields in the community that can be achieved by the SOTA methods (Felzenszwalb et al. 2009; Redmon et al. 2016; Girshick 2015; Liu et al. 2016; Dehghan et al. 2015; Son et al. 2017; Zhu et al. 2018a). Thus, the narrowly defined person re-ID involves

Figure 1.2 : An example of person Re-ID as the retrieval application

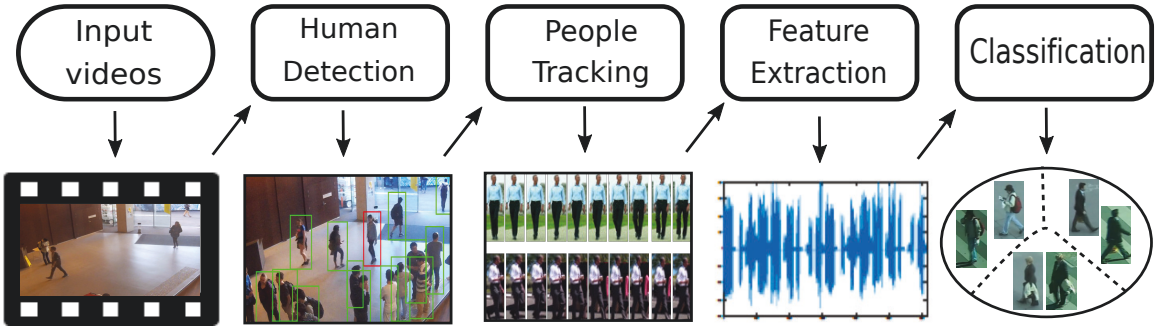


Figure 1.3 : Pipeline of a complete person Re-ID system

only the last two phases: feature extraction and classification. The former refers to design or learn representations from training samples, and the latter classifies the obtained representation of Target Subject of Interest (TSI) to its corresponding identity. Most previous works focus on these two stages, which incurs approaches in terms of hand-crafted feature representation, metric learning and end-to-end deep learning.

Obviously, it takes only several minutes to walk from location Camera #a to Camera #b in the classical setting in Figure 1.1 since the two neighbouring cameras

are disjointly deployed along the only route. In this scene, TSIs maintain their appearance unchanged, which re-ID can be achieved by similarity matching based on appearance properties. This kind of works is termed as CST-reID, and most previous works focus on the classical task.

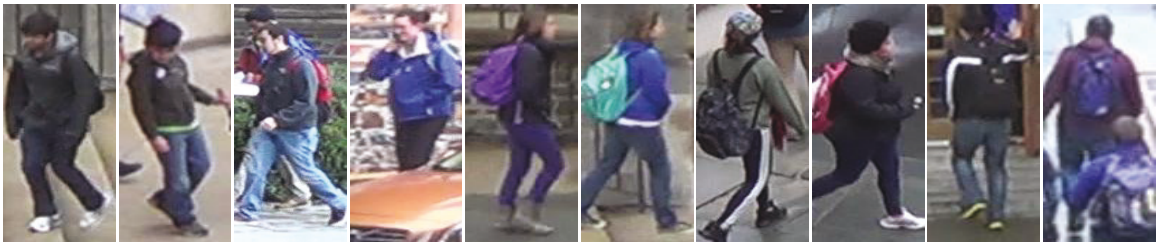
In the past several years, CST-reID has achieved significant progress and great improvement in terms of accuracy. However, it is still a challenging problem due to the large intra-class discrepancy and inter-class ambiguity caused by perturbations, such as variations of lighting condition, camera viewpoint, human pose, background clutters, occlusion, etc. Figure 1.4 lists these challenges in four benchmark datasets, i.e., VIPeR (Gray and Tao 2008), CUHK03 (Li et al. 2014), MARS (Zheng et al. 2016a) and DukeMTMC-reID (Zheng et al. 2017b). These challenges interact mutually which seriously constrains performance of the re-ID system.

Existing works on person re-ID mainly focus on erasing the effect of these variations. There are many ways to categorize these works based on, for example, (a) types of cameras (b) the attribute of input samples and (c) the pipeline of person re-ID system (Zheng et al. 2016b). Regarding the type of sensors, common RGB camera, near-infrared (NIR) camera and RGB-D sensors are utilized as shown in 1.5. Most existing works (Gray et al. 2007; Li et al. 2014; Zheng et al. 2015; Liao et al. 2015; Zheng et al. 2016a; Wang et al. 2016; Zheng et al. 2017a; Chen et al. 2019) address variations in re-ID using RGB cameras since they are cheap and deployed widely in surveillance networks. Recently, NIR cameras are adopted because of their robustness to poor lighting condition, e.g., night (Wu et al. 2017b; Ye et al. 2018c,a; Dai et al. 2018; Wang et al. 2019b). However, it raises cross-modality problem. Some solutions (Barbosa et al. 2012; Haque et al. 2016) leverage consumer RGB-D sensors, Kinects, because they can capture depth information that is useful for 3D human body reconstruction. This benefits the clothing impaired scenarios that usually happen in the long-term re-ID task. Though RGB-D cameras can be

(a) VIPeR

(b) CUHK03

(c) MARS



(d) DukeMTMC-reID

Figure 1.4 : A comparison of variations in four benchmark datasets: (a) VIPeR, (b) CUHK03, (c) MARS, (d) DukeMTMC-reID

.

(a) RGB camera (b) NIR camera (c) RGB-D sensor

Figure 1.5 : An example of three different types of sensors.

used for the short-term re-ID, they achieve inferior performance compared to visual RGB cameras. Moreover, RGB-D sensors are expensive that are not practical to be widely deployed in the real-world now.

According to the inputs, it can be categorized as image-based person re-ID (Hirzer et al. 2011; Koestinger et al. 2012; Liao et al. 2015; Matsukawa et al. 2016; Zhao et al. 2017; Li et al. 2018b; Sun et al. 2019) and video-based person re-ID (Liu et al. 2015; You et al. 2016; Zheng et al. 2016a; McLaughlin et al. 2016; Wang et al. 2016; Chen et al. 2018a; Zhu et al. 2018b). Image-based re-ID takes one or multiple frames as input while video-based person re-ID takes clipped videos with multiple successive frames as input. In particular, image-based methods depend highly on appearance-based attributes such as colour and texture of clothes, because these attributes preserve constant in the short-term re-ID scenarios. In addition to appearance properties, video-based methods also explore motion information that is proved helpful to re-ID performance. This is also verified by experiments in Chapter 3, which the appearance independent motion patterns achieve remarkable performance in the long-term person re-ID task.

According to different stages in the pipeline of person re-ID system, previous works focus on three aspects: feature extraction/mining, metric learning and end-

to-end deep learning. Early works mainly focus on designing or crafting representations from images or videos, for example, histograms in HSV and LAB colour spaces (Hirzer et al. 2012; Xiong et al. 2014; Zheng et al. 2016c), local binary patterns (Hirzer et al. 2012; Zheng et al. 2016c), ensemble of local features (Gray and Tao 2008), local maximal occurrence (Liao et al. 2015), STFV3D (Liu et al. 2015), HOG3D (You et al. 2016), high-level features learned by CNN (Cheng et al. 2016; Zhao et al. 2017; Xiao et al. 2016) and some fusion representations of the above descriptors. To further explore the discriminability of these representations, metric learning is utilized such as KISSME (Koestinger et al. 2012), Local Fisher Discriminant Analysis (Pedagadi et al. 2013), Large Margin Nearest Neighbour (Weinberger and Saul 2009), Top-push Distance Learning (You et al. 2016), Cross-view Quadratic Discriminant Analysis (Liao et al. 2015) and Null Foley-Sammon Transfer (Zhang et al. 2016). With the increasing scale of databases, data-driven approaches become popular and achieve state-of-the-art performance. These methods combine feature learning and classification in one model and perform in an end-end manner. However, all these approaches are developed for the CST-reID that assumes no intra-person clothing changes.

Moreover, there are many datasets collected for person re-ID research. These datasets are collected in either images or videos using security RGB cameras and annotated either manually or automatically. In addition to consider the variations mentioned above, they also involve different scales and a different number of cameras. Especially, some large-scale datasets (Zheng et al. 2016a; Li et al. 2014; Zheng et al. 2017b) with at least eight cameras are proposed, which are significantly helpful to the development of person re-ID. The representative datasets which are extensively used in the existing works include, *e.g.*, VIPeR (Gray and Tao 2008), PRID2011 (Hirzer et al. 2011), CUHK01-03 (Li et al. 2012; Li and Wang 2013; Li et al. 2014), iLIDS-VID (Wang et al. 2016), Market1501 (Zheng et al. 2015), MARS (Zheng et al.

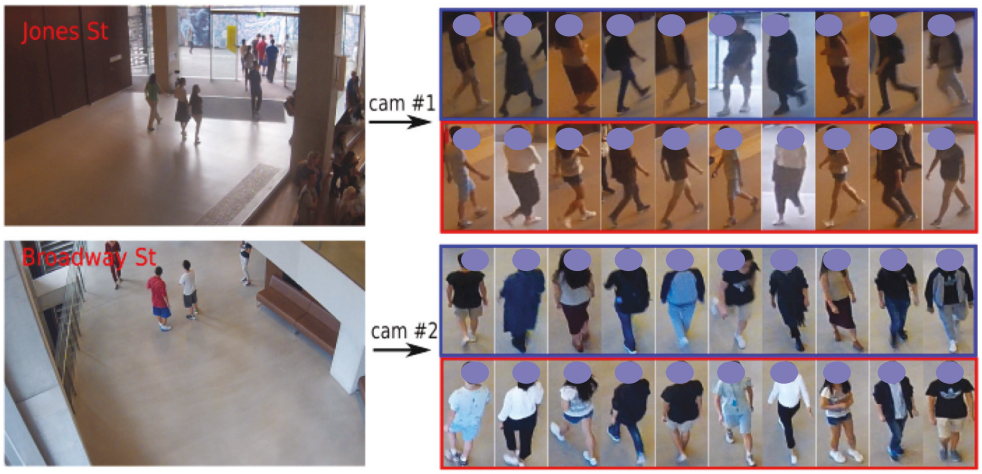


Figure 1.6 : An example of long-term person re-ID setting which includes two cameras deployed at distinct gates in one building. Pedestrians collected in the same day are put in boxes with the same colour.

2016a), DukeMTMC-reID (Zheng et al. 2017b), etc. However, none of these datasets considers the long-term re-ID scenarios, which means they include no drastic appearance changes, especially the clothes. Though these datasets meet the requirements of CST-reID research, they cannot be used for training models for long-term person re-ID.

1.1.2 Challenging Long-term Person Re-ID

Long-term person re-ID is an emerging task that is merely researched before. It considers the scenarios where TSIs re-appear in the same camera network wearing different clothes a few days later or even a few months later. It is also a practical problem that has many applications, for example, criminal investigation, forensic reasoning and customer behaviour analysis.

In the application of long-term person re-ID, different types of sensors are usually used in real-world scenarios, i.e., RGB camera, depth sensor and NIR camera. Thus,

Figure 1.7 : Illustration of challenges in long-term person re-ID. Images in the same row is taken from the same TSI

it involves two sub-problems: LTG-reID and CCM-reID. LTG-reID address the person identification problem using a single kind of sensor, i.e., RGB camera, between different days. For LTG-reID, intra-person appearance suffers larger variations than CAT-reID because the TSI is very likely to change their clothes. For example, there are two cameras deployed in different gates of the same building as in Figure 1.6. It is easy to recall the practical fact that the TSI would pass through either gate in different days span on the surveillance network. Both cameras monitor pedestrians passing through the gate, and capture images of these pedestrians in different days. In this case, different TSIs in similar clothes may appear in the same camera which imposes the intra-person similarity, and the same TSI is likely to wear significantly distinct clothes between two shots in different days. This makes the intra-person appearance more dispersive and meanwhile inter-person appearance prone to more ambiguity.

Figure 1.7 gives an example showing challenges in LTG-reID. It includes images from two TSIs, where images in the same row belong to the same person. As shown in the figure, LTG-reID is affected by common factors as CST-reID, i.e.,

	Short-term re-ID	Long-term re-ID
Time Interval	Short time period	Long time period
Common Challenges	Illumination, viewpoint, pose, background clutter, *etc.*	
Unique Challenges	No	Clothing changes (except NIR camera)
Sensors	RGB	RGB, NIR camera, RGB-D sensor

Table 1.1 : A comparison between short-term person and long-term person re-ID.

lighting condition, viewpoint, etc. In addition to these variations, LTG-reID also suffers extra intra-person impaired appearance problem. For example, the TSI in the blue box of Figure 1.7 wears two different clothes in term of colour and texture. This significantly enlarges the intra-person differences. In contrast, the two TSIs in red box dress very similar clothes which are dark T-shirts and light-coloured pants. It is even hard to distinguish them by the naked eye. This seriously imposes the positive samples, which would result in failure prediction. Table 1.1 compares between the short-term and long-term person re-ID and summarizes their commons and differences.

For CCM-reID, different types of sensors are utilized. For example, NIR cameras are usually utilized to deal with the case that matching TSI captured between day and night. It is practical because RGB camera cannot obtain sufficient information in the night. Moreover, RGB-D cameras are widely used to deal with the case that TSIs' clothes are changed in long-term person re-ID scenarios. Since the fact that RGB-D sensors can provide depth information, these methods usually reconstruct 3D body or estimate skeleton that is irrelevant to the clothing changes. As shown in Figure 1.8, these different sensors take their advantages to sense images in both the day and night. However, CCM-reID raises another problem of how we can identify the persons with such images taken by different types of sensors. In other

Figure 1.8 : An comparison of CST-reID and CCM-reID. Modality of query sample and gallery samples are same for CST-reID, but it is different for CCM-reID.

words, CCM-reID suffers extra modality difference in addition to challenges in CST-reID. Such extra challenge enlarges the intra-person separability and makes existing methods fail while applying to CCM-reID without adaptation.

1.2 Research Problems

As discussed in Sec. 1.1, long-term person re-ID includes two sub-problems: LTG-reID and CCM-reID. Both suffer larger intra-person discrepancy compared to CST-reID. For LTG-reID, it is caused by appearance difference. However, there are still some properties that can be explored, such as movement patterns, subtle identity information. To address it, we investigate three sub-problems: (1) Pure motion estimation in Sec. 1.2.1, (2) View difference mitigation in Sec. 1.2.2, and (3) Hybrid representation learning in Sec. 1.2.4. For CCM-reID, The discrepancy is caused by camera modality difference. We thus study the modality bias mitigation problem in Sec. 1.2.3. Since LTG-reID is a new topic, there is still no public dataset available for such research. This thesis also studies the LTG-reID dataset collection problem in Sec. 1.2.5. We visualize the relationship of our research problems in

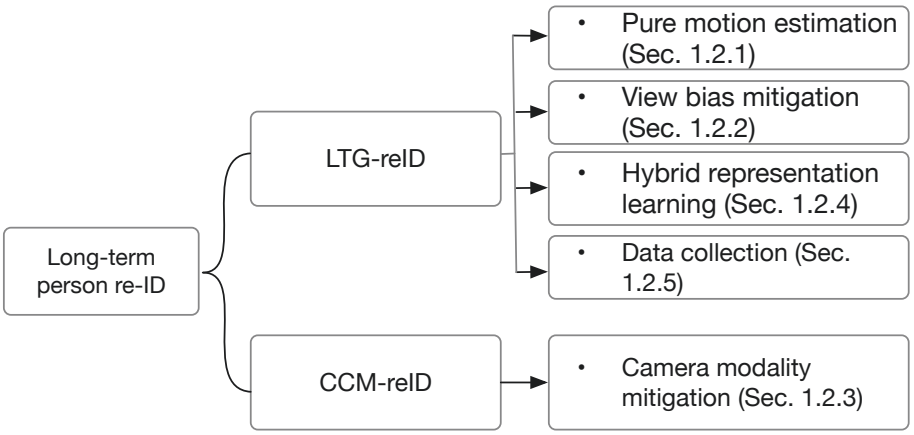


Figure 1.9 : Relationship structure of our research problems

Figure 1.9.

1.2.1 Pure Motion Estimation from Dense Trajectories

As LTG-reID involves great perturbation from appearance variation, it is unreliable to perform identification using appearance-based representations that are widely adopted in CST-reID any more. Rather than appearance properties, previous video-based re-ID methods demonstrate that motion patterns of human do help increase the performance of re-ID systems. In another aspect, motion pattern has been extensively utilized in gait recognition because of its benefits such as uniqueness, un-contacting, hard to disguise, etc. These studies also demonstrate the effectiveness of motion pattern for recognizing people's identity.

As for scenarios of LTG-reID, pedestrian's motion attribute is stable because it is slightly influenced by clothing changes. These motivate us to tackle the long-term re-ID problem using motion characteristics. Gou et al. (2016) has tried to address it using the same idea. In the meantime, they fail to mine the true motion patterns because they extract dense trajectories from normalized areas of human rather than raw images. However, action recognition has provided many choices to explore real motion attribute from videos. Especially, dense trajectory reflects how pedestrian

walks and characters dynamic motion patterns of the human body from raw footage without any scalability normalization. It supposes to be beneficial if true motion attribute from dense trajectories could be applied to the long-term person re-ID. In this thesis, a fine motion encoding method that explores true motion attribute from dense trajectories is proposed in Chapter 3.

1.2.2 View Bias Mitigation via GAN

Viewpoint difference is extensively studied in person re-ID. Prior approaches for this problem rely on aligning features by either body segmentation or pose guided image generation. However, these methods are developed on the basis of appearance-based feature representation. For long-term person re-ID, it also experiences the viewpoint variation problem. This drastically degrades the matching performance, especially under the case of performing person re-identification using gaits such as GEI (Han and Bhanu 2005). This problem is also studied in gait recognition, which adopts either view transform model (Kusakunniran et al. 2009, 2014) or coupled subspace learning (Zhang et al. 2015; Ben et al. 2019b,a). Recently, generative adversarial network (GAN) (Goodfellow et al. 2014; Yu et al. 2017b) is used to address the viewpoint problem since GAN can not only generate visually realistic results but also makes deep models interpretable. This thesis studies the view bias and proposes to mitigate it through two GAN variants in Chapter 4.

1.2.3 Camera Modality Bias Mitigation

In addition to RGB cameras, NIR and RGB-D sensors are also applied to the long-term person re-ID. As above discussed, NIR cameras are insensitive to poor light condition, and RGB-D sensors are robust to clothing changes. These cameras are usually utilized in specific scenarios. This raises the cross-modality problem, *i.e.*, matching across images/videos of TSI shoot by different types of cameras. For example, RGB cameras are used in daylight while NIR cameras are used in the

night (Wu et al. 2017b; Ye et al. 2018a,c). Due to the distinct imaging theory of these sensors, images or videos from these sensors obey significantly different data distribution. The large data bias across modalities certainly causes the mismatching problem. This thesis thus discusses this problem and proposes to address it by combining the theory of asymmetric mapping (Zhang et al. 2015; Ben et al. 2019a,b) and discriminative dictionary learning (Cheng et al. 2017; Lu et al. 2017). Detailed solution is introduced in Chapter 5.

1.2.4 Learning Hybrid Representation via Neural Networks

CNN achieves competitive performance and becomes an increasingly predominant choice for many computer vision tasks because of its strong expressing ability. This is also true for CST-reID. However, most existing works leverage it to explore appearance cues that are not affected by the case of long-term person re-ID. In another aspect, Zheng et al. (2019) demonstrate that there are some subtle identities which are invariant to clothing changes, e.g., body structure. Moreover, implicit motion attributes of walking human do help recognize one's identity as aforementioned. Thus, automatically learning and fusing all these useful properties should be beneficial to the long-term re-ID problem. Details of the solution are presented in Chapter 6.

1.2.5 Dataset Collection

LTG-reID is an emerging topic. Most prior works try to address it using the NIR and RGBD cameras while solutions using RGB cameras are merely addressed. However, RGB cameras are more practical that have been widely deployed in existing surveillance networks. Gou et al. (2016) make the first attempt to tackle long-term person re-ID using RGB cameras. They collect a toy dataset and demonstrate the effectiveness of this idea. Unfortunately, their dataset is not opened for public use. Thus, it is essential to collect and annotate datasets particular for training and

evaluating proposed methods for long-term person re-ID. The related descriptions of dataset collection are embodied in Chapter 3 and 6.

1.3 Thesis Contribution

The contributions of the work are concluded as following:

- Developed a novel fine motion encoding model that characterizes motion patterns pyramidally from both global and local body action units, which achieves significant performance for the LTG-reID task and aids for the CST-reID problem. (Chapter 3)

- Proposed two GAN variants, i.e., VN-GAN and VT-GAN, to address the view difference problem in LTG-reID using gaits. VN-GAN aims to normalize gaits of different views into a unified one. It adopts a coarse-to-fine design that achieves view normalization and identity injecting simultaneously. VT-GAN aims to achieve view transformation across any two views using a single model. And, it also proposes an identity preserver that keeps the identity information while performing view transformation. Both methods achieve competitive performance for LTG-reID using gaits. Moreover, they provide good visual quality of generated gaits that makes cross-view LTG-reID using gaits visually interpretable. (Chapter 4)

- Proposed to join asymmetric feature mapping and discriminative dictionary learning in a unified scheme for CCM-reID. It alleviates data biases across modalities in the projected subspace, and thus heterogeneous data can be represented by a shared discriminative dictionary. Moreover, a top-push ranking constraint is reformulated and integrated into the unified model, which makes the dictionary learning more effective to person re-ID. (Chapter 5)

- Proposed to learn hybrid representation by a dual-stream network for LTG-reID. The framework includes two tracks, i.e., SSIN and SMIN, which learns both subtle identity attributes and motion patterns. In SSIN track, a mask-guided cross-attention module is proposed to capture relative location changes of the local body. In SMIN track, a novel method is presented to mine motion attributes from normalized 3D skeletons with GCN. (Chapter 6)

- Proposed two datasets that especially collected for long-term re-ID, Motion-reID and CVID-reID. Motion-reID is the first dataset that is collected in real-world scenarios for long-term person re-ID. CVID-reID involves videos of celebrities clipped from video sharing platform. It includes diverse variations in practical scenarios such as clothes, background and viewpoint, etc. This is the largest long-term person re-ID dataset up to now. (Chapter 3, 6)

1.4 Thesis Structure

The thesis studies the long-term person re-ID task that includes five research problems in term of challenges from clothing changes, viewpoint variation, camera modality difference, etc. The rest of the thesis is organized as followed:

- *Chapter 2*: This chapter introduces basic knowledge of and reviews literature in fields that are coherently related to the thesis, for example, person re-identification, motion analysis, person re-ID using gaits, cross-modality person re-ID and hybrid feature learning using CNNs.

- *Chapter 3*: This chapter studies the sub-problem, pure motion estimation, and presents a fine motion encoding model. In addition, a dataset called Motion-reID is proposed.

- *Chapter 4*: This chapter studies the view difference problem in long-term

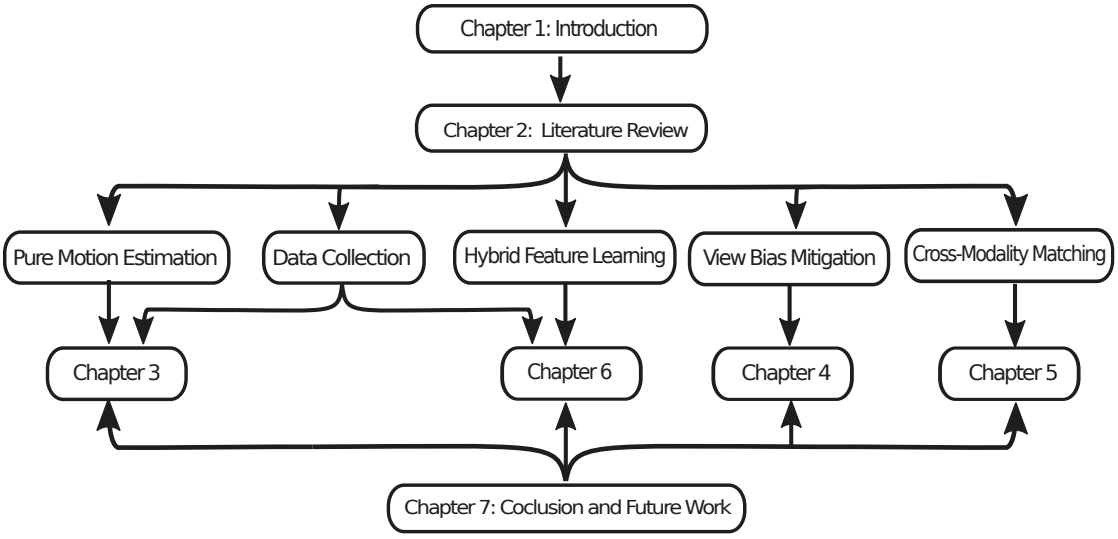


Figure 1.10 : Illustration of thesis structure

person re-ID using gaits and proposes two GAN variants, i.e., VN-GAN and VT-GAN, to address it.

- *Chapter 5*: This chapter studies the CCM-reID problem and proposes a top-push constrained modality-adaptive dictionary learning model to mitigate the data discrepancy.

- *Chapter 6*: This chapter analyses the subtle identity properties and implicit motion attributes of walking TSIs for LTG-reID. In particular, a hybrid representation learning method is developed to mine these characteristics.

- *Chapter 7*: This chapter concludes the thesis and gives insights/trends of future works.

The structure of this thesis is summarized in Figure 1.10.

Chapter 2

Literature Review and Related Theories

This chapter comprehensively reviews SOTA works and benchmark datasets for person re-ID, including CST-reID, CCM-reID and LTG-reID. These works provide baselines and inspirations for tackling the long-term person re-ID in the wild. In addition, we also provides a brief review of related theories that are coherently correlated to proposed approaches regarding to research problems in Section 1.2, such as motion attribute analysis (in Section 2.2), heterogeneous person matching including cross-modality matching (in Section 2.3.1), and cross-view person re-ID using gaits (in Section 2.3.2), etc.

2.1 Review on Current Works of Person Re-ID

This section provides a comprehensive review of works for person re-ID. Considering time variation between two shots, from short to long, we divide person re-ID into two divisions: conventional short-term person re-ID (CST-reID) and contemporary long-term person re-ID (CLT-reID). These divisions involve different perturbations, such as changes of illumination, distinct imagery properties of different sensors, and varying cloth, which has been completely discussed in Section 1.1. It thus incurs significantly different solutions and datasets for addressing these different kinds of person re-ID tasks.

2.1.1 Conventional Short-term Person Re-ID

Most existing approaches for person re-ID are developed based on scenarios of CST-reID. Early works mainly focus on either crafting feature representations em-

pirically by characterizing colour, texture, gradient, attribute and spatial-temporal information, etc., or learning discriminative metrics to measure the similarity between image pairs collected from distinct cameras. These works promote the development of person re-ID. Recently, end-to-end data-driven methods become a prevalent choice for many vision tasks and also achieve predominant performance on person re-ID. In this section, we thus make an in-depth review of these prior works in terms of feature representation, metric learning and data-driven methods.

Feature Representation

Significant efforts have been made to develop or learn better features that are at least partially robust to illumination changing, viewpoint variation, pose indeterminacy and occlusions. These representations are designed from either images or videos, which can be classified into two categories, e.g., image-based feature and video-based representation.

Image-based features which are usually generated from one or multiple discontinuous frames, for example, Gray and Tao (2008) presented a viewpoint invariant representation learned from local features using AdaBoost. Similarly, Farenzena et al. (2010) also proposed to extract local appearance-based features from different body parts. However, they considered the symmetry and asymmetry perceptual properties that are robust to viewpoint difference. Moreover, Zhao et al. (2013) proposed to perform patch matching using patch salience information in order to overcome body misalignment problem. In another aspect, Xiong et al. (2014) utilized fusion features of RGB, YUV, HSV colour-based histograms and LBP to evaluate the effect of different kernels embedded in metric learning algorithms for Re-ID. To overcome illumination and background variations, Ma et al. (2012) proposed a feature representation that characterized texture information using covariance descriptors. Liao et al. (2015) proposed a descriptor LOMO that is insensitive to illumination dif-

ference benefiting of Retinex transformation and SILTP texture histograms. And, LOMO is also robust to viewpoint changes by maximizing horizontal occurrence of local features. In addition, Matsukawa et al. (2016) proposed to model region/patch using hierarchical Gaussian distributions, which achieved SOTA performance.

On the other hand, video-based features (Wang et al. 2016; You et al. 2016; Liu et al. 2015) are usually extracted from consecutive walking sequences, e.g., both Wang et al. (2016) and You et al. (2016) applied HOG3D to extract spatial-temporal information from walking pedestrians. After combining them with colour or texture histograms, they improved the performance by a large margin. For the space-time alignment problem, Liu et al. (2015) proposed a spatial-temporal appearance representation named STFV3D which encodes local descriptors by Fisher Vector with respect to a body-action unit model. The STFV3D exactly solves body misalignment caused by viewpoint change to some content. However, it is worth pointing out that both above-mentioned features in current works are closely relevant to ones appearance, which will be unreliable once his appearance is drastically changed such as matching the same individual with different clothing or carrying conditions.

Metric Learning for Person Re-ID

Metric learning is to learn a similarity metric in the projected discriminative subspace in which distance of positive pairs is closer than corresponding negative pairs. In general, The metric is defined in the form of Mahalanobis-like distance, such as

$$dis_W(x, y) = (y - x)^\top W(y - x) \tag{2.1}$$

where x, y are two samples, W is the learned discriminative space over training samples. Most existing metric learning models applied to person re-ID are developed based on it, for instance, KISSME (Koestinger et al. 2012), LFDA (Pedagadi et al. 2013), LMNN (Weinberger and Saul 2009), TDL (You et al. 2016), XQDA (Liao

et al. 2015) and NFST (Zhang et al. 2016). We briefly review these works below.

Weinberger and Saul (2009) proposed LMNN that aims at performing KNN classification for data from arbitrary cameras. For each instance, negative neighbours are heavily penalized, and data from different classes are isolated by a large margin. Similar to LMNN, TDL proposed by You et al. (2016) not only penalize negative neighbours but also consider a minimum inter-class distance. Both the methods are based on Mahalanobis distance which requires a degree of supervision. It is infeasible for large-scale scenarios since fully labels for large-scale data are hard to obtain. Towards this problem, Koestinger et al. (2012) introduced KISSME to process large scale data, which avoids expensive iteration from the statistical perspective. To reduce the effect of data piling for high dimensional data, Pedagadi et al. (2013) presented LDFA that projects PCA-reduced (principal component analysis) features of two samples into an embedding space. In this space, negative instances are more separable. Considering preprocessing by PCA reduction might destroy the original data structure, Liao et al. (2015) proposed to simultaneously learn a discriminant low dimensional subspace and a QDA metric on this learned subspace. The model works well together with LOMO features. Differently, NFST (Zhang et al. 2016) overcomes small sample size problem caused by high data dimensions and limited training samples by learning a discriminative null space. In the null space, samples from the same class are collapsed into a point. And thus, intra-class scatter is minimized to extreme and margins between classes are maximized. No matter what the intuitions of the metric learning models are, the performance of these models highly depends on the quality of data. For person Re-ID, it is feature representations introduced in Section 2.1.1. In fact, these models collaborating with above hand-crafted features have achieved promising performance for CST-reID.

Data-driven Methods for Person Re-ID

With the increasing scale of data, data-driven approaches using convolutional neural networks (CNNs) have been predominant choices in many computer vision tasks. This is particularly true for person re-ID. Thanks to the strong representation of CNN, CST-reID approaches approximately saturate performance. Typically, these methods address re-ID problem prone to either enlarging inter-person differences or minimizing intra-person discrepancy, as discussed below.

To maximize inter-person variations, a family of previous methods dedicate to optimize CNNs using metric learning losses, such as contrastive loss (Hadsell et al. 2006), triplet loss (Cheng et al. 2016), quadruplet loss (Chen et al. 2017), etc. These losses empower the feature extractor (CNNs) discriminative ability, which tremendously improves the re-ID performance. Contrastive loss is adopted to learn discriminative embeddings in early data-driven re-ID works. It aims to distinguish negative pairs with a predefined margin. For example, Varior et al. (2016a,b) weave the contrastive loss into a Siamese network, which separates negative pairs by enforcing a large margin in the embedding space. Different from contrastive loss, triplet loss takes a triplet set of anchor sample, positive sample and negative sample as input, which simultaneously performs positive aggregation and negative separation. It is widely adopted in the field of person re-ID. For example, Ding et al. (2015) compared the relative distance of image pairs and constrain/optimize model (CNNs) using triplet loss. Cheng et al. (2016) proposed to learn representations using part-based CNN and jointly optimize them with gradients from an improved triplet loss. Hermans et al. (2017) derived a variant of triplet loss, *i.e.*, batch-hard mining, which can calculate triplet loss within a batch. This enables the data-driven deep models to train in an end-to-end way. Song et al. (2019) introduced a logit-triplet loss that benefits generalizable re-ID with strong supervision. In addition to the above works, triplet loss prototype is also weaved into many other re-ID frame-

works (Su et al. 2016; Zhao et al. 2017; Liu et al. 2018; Suh et al. 2018), which improves discriminability of these frameworks and gains tremendous progress of re-ID accuracy. Quadruplet loss is proposed by Chen et al. (2017), which regularizes the model to output representations with a larger inter-person difference and smaller intra-person discrepancy than triplet loss. It demonstrates competitive performance on benchmark datasets.

In contrast to enlarge inter-person variation, minimizing intra-person discrepancy is another solution to boost re-ID performance. This is usually achieved by explicitly or implicitly adressing the viewpoint difference, which can be categorized into three facets: (1) local body matching via pedestrian segmentation (Wei et al. 2017; Su et al. 2017; Li et al. 2017b,a; Sun et al. 2018; Suh et al. 2018); (2) body alignment (Ma et al. 2017; Qian et al. 2018; Ge et al. 2018; Zheng et al. 2019); and (3) attention-aware matching (Liu et al. 2017; Xu et al. 2018; Li et al. 2018b; Yang et al. 2019; Chen et al. 2019; Fang et al. 2019).

For the part-based matching, prior works tend to learn discriminative representations in patch level. These approaches split the human body into patches in a hard or soft way based on body structure prior. For example, Cheng et al. (2016) divided heatmaps of global convolution layer into four equal parts spatially and learned features from both local and global parts. In the same way, Li et al. (2017b) also proposed to manually divide heatmaps of the middle layer into several stripes and learn both local and global representations by jointly optimizing multi-loss. Rather than pre-defined regions, Wei et al. (2017) segmented body into three regions according to four key points (i.e., upper-head, neck and two hips) estimated by Deeper Cut (Insafutdinov et al. 2016). Representations are then learned both locally and globally using CNNs. Similarly, Li et al. (2017a) also proposed to localize local parts in an automatic way. In practice, they adapted STN (Jaderberg et al. 2015) as a part locator and optimized simultaneously with feature extractor. Different from the

above, Sun et al. (2018) proposed to divide the output of base CNN uniformly and penalize logits from each patch. These part-based approaches are resistant to body deformation, which thus narrow intra-person discrepancy and improve performance.

In another aspect, some works dedicated to addressing pose variation in a generative way. These methods usually align pedestrian by synthesizing images constrained by pose prior. Benefiting from the success of generative models in image-to-image translation, such as GAN (Goodfellow et al. 2014), VAE (Diederik et al. 2014), etc., it is possible to boost performance by augmenting more samples of synthesized images (Zheng et al. 2017b; Huang et al. 2019). However, these models only work for image generation, which cannot operate on human bodies. Recently, conditional GAN (Mirza and Osindero 2014; Isola et al. 2017) provides possibility to achieve this. In person re-ID, it enables to correct viewpoint discrepancy under prior pose knowledge. For example, Ma et al. (2017) proposed a two-stage framework to synthesize person images guided by input skeleton prior. Qian et al. (2018) proposed to synthesize human of eight canonical poses in order to address pose variations. However, these two works synthesize images controlled by auxiliary pre-defined pose heatmaps. Different from them, Zheng et al. (2019) considered to decouple appearance and structure and take them as the condition in order to realize pose and appearance exchange. However, these works perform pose correction and feature extraction separately. Unlike them, Ge et al. (2018) firstly proposed to perform pose normalization and feature distilling simultaneously. These approaches achieve promising re-identification results. Moreover, they provide good perceptional effects.

In addition, attention-aware matching is also a popular way to extract pose-invariant features. For example, Liu et al. (2017) presented comparative attention networks that compare the local appearance of image pairs and locate the relevant parts. Xu et al. (2018) proposed a dual-stream pipeline that pose-guided part attention discriminates the rigid and non-rigid parts in order to filter undesirable features

and attention-aware feature composition produces the re-weighted features controlling by part visibility scores. Li et al. (2018b) proposed to solve the misalignment problem by performing soft pixel attention and hard region attention jointly, which achieves SOTA performance. Recently, Chen et al. (2019) presented a self-critical attention learning scheme that provides supervision for feature learning. Fang et al. (2019) proposed a bilinear attention block to interact between the pairwise parts and investigated high-order statistics using attention in attention scheme. These works provide an automatic body alignment mechanism that enables to distinguish different people using the discriminant pairwise local parts.

Because of the strong representation ability, data-driven methods are naturally resistant to other variations such as illumination and background. These properties empower data-driven approaches to dominate research on person re-ID. Moreover, these methods are usually trained in an end-to-end manner with classification loss, ranking loss or both. It increases the robustness of these models to both intra- and inter- person variations, and thus improves identification performance.

2.1.2 Contemporary Long-term Person Re-ID

CLT-reID is an emerging topic that attracts research interest recently. In practice, different types of cameras are used as discussed in Sec. 1.1.2. Thus, two sub-problems are reviewed, i.e., CCM-reID and LTG-reID.

Cross Camera Modality Person Re-ID

CCM-reID is raised recently, which large modality/domain discrepancy seriously deteriorates the performance of conventional re-ID methods. One typical example is the RGB-NIR re-ID problem that matches different types of person images collected in the day by visual-light/RGB cameras and in the night by NIR sensors.

The first work on the problem was published in 2017 by Wu et al. (2017b). They

first discussed the CCM-reID using RGB and NIR images, which learned domain-specific and domain-shared feature via a one-stream network by padding zeros to images. Though the work achieves promising results, data gaps still exist due to the hard threshold used to determine whether a node is domain-specific. In 2018, Ye et al. (2018b) proposed to map modality-specific features from two modalities into a consistent space and learn modality-shared features using a two-stream network (TONE), which improves performances than only using TONE. Furthermore, Ye et al. (2018d) proposed a novel bi-directional dual-constrained top-ranking loss to optimize the two-stream network, which further improves the performance. Rather than designing new loss, Dai et al. (2018) proposed to narrow the domain discrepancy using generative adversarial training, which significantly improves the performance. In 2019, Wang et al. (2019c) proposed to mitigate co-existed modality discrepancy and appearance discrepancy by simultaneously training a GAN model and a feature learning extractor. It achieves the SOTA performance.

In general, CCM-reID is regarded as a long-term re-identification problem, especially targeting the drastic lighting changes in one day. This thesis also considers the problem, which will be further discussed in Chapter 5.

2.1.3 Person Re-ID after Long-time Gap

LTG-reID is an emerging topic that is studied recently. These works can be divided into two categories considering the used sensors, i.e., methods using depth/RGB-D sensors and RGB cameras.

In early works, RGB-D sensors are widely adopted because the reconstructed 3D human is invariant to clothes. For example, Barbosa et al. (2012) proposed to explore 3D soft-biometric cues estimated by 3D body reconstruction and achieve identification by body signature measurement. Munaro et al. (2014a) proposed to perform re-identification by matching wrapped 3D point clouds of body that are

computed from depth information and skeletons by RGB-D sensors. Haque et al. (2016) proposed to extract latent spatio-temporal signatures from 4D point cloud sequences by recurrent attention models. Though these approaches achieve compelling performance, they overly depend on expensive RGB-D sensors that have not been widely deployed in practical scenarios. Recently, methods using common surveillance cameras come into attention. A typical idea under these tentative works is to distil identity information from motion patterns. For example, Gou et al. (2016) proposed to learn dynamic-based feature from dense short trajectories. This work attempts to manually craft or design representations using the popular dense trajectory (Wang and Schmid 2013). However, this requires strong experience/motivation and heavily depends on accurate dense trajectory estimation.

LTG-reID is a more challenging task than CST-reID, in particular the scenarios that only RGB cameras are available. This thesis mainly focuses on the case, which will be further investigated in Chapter 3 and Chapter 6.

2.1.4 Benchmark Datasets for Person Re-ID

Precious works usually restrict re-ID as a cross camera tracking problem. Thus, most existing datasets are constructed under the assumption that pedestrians successively pass across multiple sensors without clothing changes. For these datasets, RGB cameras are usually adopted, because they are cheap and widely deployed. And, they include either images, e.g., VIPeR (Gray et al. 2007), CUHK01 (Li et al. 2012), Market1501 (Zheng et al. 2015), DukeMTMC-reID (Zheng et al. 2017b), etc., or videos/tracklets/image sequences, e.g., PRID2011 (Hirzer et al. 2011), MARS (Zheng et al. 2016a), etc. Accordingly, these datasets stimulate methods using one-shot, multiple-shot images and video tracklets, on which promising performance has been achieved. In addition to differences in sample format, these datasets also distinct from each other by attributes such as scale (e.g., number of identities, cam-

eras and images/videos), label method (e.g., manual or automatic), bounding box (e.g., varying or fixed size), etc. A detailed comparison can be found in Table 2.1. In another aspect, these datasets are collected under CST-reID scenarios where challenges come from disturbances incurred by changes of illumination, viewpoints, poses, background, etc. These datasets are valuable to early re-ID studies, but they cannot fulfil the requirement of re-ID in long-term scenarios. This is because CLT-reID suffers severe clothing or modality difference while none of these datasets considers it.

Several datasets have been proposed to investigate disturbances in CLT-reID as summarized in Table 2.1. These datasets are usually collected in different days where TSIs will change their clothes with a high probability. One special is SYSU-MM01 (Wu et al. 2017b) that includes images on the day time using RGB cameras and night time using NIR sensors on the same day. It provides an alternative for CLT-reID between daylight and night in one way. For the rest datasets, we divide them into two categories according to the sensors used, e.g., RGB-D datasets (Barbosa et al. 2012; Munaro et al. 2014a; Haque et al. 2016) and RGB datasets (Gou et al. 2016; Zhang et al. 2018a). The former are usually collected using RGB-D sensors such as the Kinect or Xtion Pro in the controlled laboratory environment. Since RGB-D sensors provide depth information and skeleton data, identification can be achieved by 3D body measurement (Barbosa et al. 2012), 3D point cloud matching (Munaro et al. 2014a) or latent spatio-temporal patterns from 3D point clouds (Haque et al. 2016). However, RGB-D based approaches are not used widely in common security surveillance cases. Besides, the scale of these datasets is usually small, which are not suitable for CNN model training. TSD (Gou et al. 2016) is collected by common RGB cameras on different days. It is built using only one camera, which imitates LTG-reID scenario by changing their clothes. Unfortunately, the scale of the dataset is too small and not open for public use.

Table 2.1 : A comparison of existing datasets for person re-ID in both short-term and long-term scenarios.

#Format	#Dataset	#IDs	#Sensors	#Samples	#Clothing Changes	# Setting	#Interval
Datasets for Short-term Scenarios							
Images	VIPeR (Gray et al. 2007)	632	RGB, 2	1264	×	Outdoor	Daytime in the same day, usually less than 1 hour.
	CUHK01 (Li et al. 2012)	971	RGB, 2	3884	×	Outdoor	
	Market1501 (Zheng et al. 2015)	1501	RGB, 6	32217	×	Outdoor	
	DukeMTMC-reID (Zheng et al. 2017b)	1812	RGB, 8	36441	×	Outdoor	
Videos	PRID2011 (Hirzer et al. 2011)	200	RGB, 2	400	×	Outdoor	
	MARS (Zheng et al. 2016a)	1261	RGB, 6	20000	×	Outdoor	
Datasets for Long-term Scenarios							
Images	SYSU-MM01 Wu et al. (2017b)		RGB, 4; NIR, 2		×	Outdoor & Indoor	Same day (day & night)
Videos	PAVIS Barbosa et al. (2012)	79	RGB-D (OpenNI), 1	316	✓	Lab	Different days.
	BIWI (Munaro et al. 2014a)	50	RGB-D (Kinect SDK), 1	50	✓	Lab	
	IAS-Lab (Munaro et al. 2014a)	11	RGB-D (OpenNI & NITE), 1	33	✓	Lab	
	DPI-T (Haque et al. 2016)	12	RGB-D, 1	300	✓	Lab	
	TSD (Gou et al. 2016)	9	RGB, 1	81	✓	Indoor	
Ours							
Videos	Motion-ReID (Zhang et al. 2018a)	30	RGB, 2	240	✓	Indoor	Different days
	CVID-reID	90	RGB, vary	2980	✓	Outdoor	

In all, there is no available dataset for the LTG-reID research up to now. Thus, this thesis also investigates the LTG-reID dataset collection problem. In specific, we collected two datasets: Motion-reID and CVID-reID. The two datasets involve significantly different scenarios. Motion-ReID (Zhang et al. 2018a) is a real long-term re-ID dataset that is collected on different days. Most of TSIs in the dataset have changed their clothes. Similar to TSD, it is gathered at indoor scenes, i.e., airport and university building. Differently, CVID-reID is collected under practical outdoor scenarios. It contains more samples makes it suitable for large-scale training. These two datasets will be further introduced in Sec. 3.4.1 and Sec. 6.3, respectively.

2.2 Motion Attribute for Person Re-ID

This section provides an investigation on theories of motion attribute analysis and their applications to person re-ID. In detail, we briefly introduce related theories of three facets: dense trajectory, fisher vector and gait energy image (GEI). The former two are used to tackle the true motion estimation in 1.2.1, which extracts and encodes motion patterns of a walking person as illustrated in Chapter 3. The last one is a classical spatial-temporal template that will be adopted as feature representation for the problem of person re-ID using gaits in Chapter 4.

2.2.1 Dense Trajectory

Dense trajectory (DT) (Wang et al. 2011, 2013; Wang and Schmid 2013) is firstly proposed to address action recognition problem. It samples space-time interest points densely and tracks these points using optical flow field. Dense trajectories depict displacement information of these interest points, which have demonstrated momentous success on activity description combining with encoding descriptors such as HOG (Dalal and Triggs 2005), HOF (Laptev et al. 2008) and MBH (Wang et al. 2013). In person re-ID, Gou et al. (2016) extracted soft biometrics from motion

by encoding dense short trajectories from normalized bounding areas with Hankelet descriptors, which has shown advantage against appearance impaired case.

Formally, given a video, DT firstly samples space-time interest points by a step of W pixels in S scales. Since there is no motion in homogeneous image areas, points in these areas will be filtered out by eigenvalues of the auto-correlation matrix. It means points corresponding to small eigenvalues will be removed. For any frame I_t, the threshold to screen small eigenvalues is set to

$$T = 0.001 \times \max_{i \in I_t} \min(\lambda_i^1, \lambda_i^2) \tag{2.2}$$

where $(\lambda_i^1, \lambda_i^2)$ are eigenvalues of the i-th sampled interest point in the t-th frame I_t. For a point $P_{s,t} = (x_{s,t}, y_{s,t})$ in scale s of t-th frame, the corresponding point $P_{s,t+1}$ in the $t+1$ frame is tracked as

$$P_{s,t+1} = (x_{s,t+1}, y_{s,t+1}) = (x_{s,t}, y_{s,t}) + (M * \omega_{s,t})|_{(x_{s,t}, y_{s,t})} \tag{2.3}$$

where $\omega_{s,t} = (u_{s,t}, v_{s,t})$ is the dense optical flow field, M is kernel of a median filter, which is usually set to 3×3. In practice, the length of DT is limited to 15 frames in order to avoid significant drift, and shorter DTs will be discarded. Moreover, a new point will be tracked if the tracked point is missed in its $W \times W$ neighbours of the frame. To avoid tracking error, sudden drastic displacement will be discarded since it merely happens in practice.

Raw dense trajectories describe the displacement of moving points, which is usually encoded by statistical descriptors such as HOG (Dalal and Triggs 2005), HOF (Laptev et al. 2008), MBH (Dalal et al. 2006), etc. As the number of dense trajectories is varying, it is necessary to pool them into a unified dimension using approaches such as Bag-of-words (Csurka et al. 2004) or Fisher Vector (Sánchez et al. 2013), which will be introduced in the next section.

2.2.2 Fisher Vector

The Fisher Vector (Sánchez et al. 2013) was first proposed to describe images including a varying number of local descriptors. It has gained great success in applications such as large scale image classification (Sánchez et al. 2013), activity recognition (Wang and Schmid 2013), image retrieval (Perronnin et al. 2010) and even person re-ID (Ma and Su 2012; Gou et al. 2016).

Given a sample of T local descriptors $X = \{x_t \in \mathbb{R}^D\}_{t=1}^T$, where D is the dimension of these descriptors, Fisher Vector uses a K-component Gaussian Mixture Model (GMM) (Everitt 2014) to approximate these descriptors, where the GMM is parametrized by $\lambda = \{\mu_k, \Sigma_k, w_k\}_{k=1}^K$, μ_k and Σ_k are mean and covariance of the k-th Gaussian distribution and w_k is its corresponding mixture weights. Thus, the generation process is denoted as

$$\Psi_\lambda(x_t) = \sum_{k=1}^{K} w_k \psi_k(x_t), t = 1, 2, \cdots, T, \tag{2.4}$$

$$\text{s.t.} \;\; w_k \geq 0, \;\; \sum_{k=1}^{K} w_k = 1 \tag{2.5}$$

where ψ_k is the k-th Gaussian distribution, as following,

$$\psi_k(x_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(x_t - \mu_k)^\top \Sigma_k^{-1}(x_t - \mu_k)\} \tag{2.6}$$

To obtain parameters of the GMM, FV optimizes the maximum likelihood probability of local descriptors in the dataset, which derives the normalized gradients of k-th Gaussian,

$$\mathcal{G}_{\alpha_k}^X = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} (\gamma_t(k) - w_k), \tag{2.7}$$

$$\mathcal{G}_{\mu_k}^X = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) (\frac{x_t - \mu_k}{\sigma_k}), \tag{2.8}$$

$$\mathcal{G}_{\sigma_k}^X = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \frac{1}{\sqrt{2}} [\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1]. \tag{2.9}$$

$$\tag{2.10}$$

where σ_k is the standard variance of Gaussian k, whose square is diagonal element of Σ_k, α_k is the normalized weights, which is rectified by a softmax function, *i.e.*,$w_k = \frac{\exp(\alpha_k)}{\sum_{j=1}^{K} \exp(\alpha_j)}$. $\gamma_t(k)$ is the posterior probability that determines whether descriptor x_t is generated by the k-th Gaussian, formulated as

$$\gamma_t(k) = \frac{w_k u_k(x_t)}{\sum_{j=1}^{K} w_j \mu_j(x_t)} \tag{2.11}$$

Thus, the final FV representation is obtained as $f(X) = [\mathcal{G}_{\alpha_1}^X; \mathcal{G}_{\mu_1}^X; \mathcal{G}_{\sigma_k}^X; \cdots; \mathcal{G}_{\alpha_K}^X; \mathcal{G}_{\mu_K}^X; \mathcal{G}_{\sigma_K}^X]$, which is a concatenation of K normalized gradients of α_k, μ_k and σ_k. Since gradients $\mathcal{G}_{\alpha_k}^X$ is a scalar, $\mathcal{G}_{\mu_k}^X$ and $\mathcal{G}_{\sigma_k}^X$ are D-dimensional vectors, therefore dimension of the final FV $f(X)$ is $(2D + 1)K$, which is positively correlated to the number of Gaussian distributions.

2.2.3 Gait Energy Image

Gait energy image (GEI) (Han and Bhanu 2005) is a synthetic spatio-temporal template that describes motion characteristics of human. It is extensively adopted in previous works (Yu et al. 2006; Lu et al. 2008; Martín-Félez and Xiang 2012; Iwama et al. 2012; Hu et al. 2013; Kusakunniran et al. 2013; Yu et al. 2017b; Ben et al. 2019a,b), which significantly promotes development of person identification using gaits.

As gait is a cyclic activity that human walks in a similar manner but with different moving range, it is possible to express gait properties by accumulating gaits in a cycle without considering the moving order. GEI inherits the idea, which is formulated as

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y) \tag{2.12}$$

where $B_t(x, y)$ is a binary silhouette that stands for the pixel in position (x, y) of t-th frame, N is the number of frames in a gait cycle.

2.3 Key Techniques of Heterogeneous Person Re-ID

This section briefly reviews literatures that are most related to research problems in Sec. 1.2.2 and Sec. 1.2.3. In particular, Sec. 2.3.1 introduces works using dictionary learning to address the modality discrepancy problem in Sec. 1.2.3 and Sec. 2.3.2 reviews classical methods to address view bias problem in Sec. 1.2.2 in person re-ID using gaits. The former motivates us to tackle cross-modality person re-ID problem using dictionary as in Chapter 5, and the latter inspires us to solve view difference problem person re-ID using gaits as in Chapter 4.

2.3.1 Cross-modality Dictionary Learning

Benefiting of great expressive ability, dictionary learning and its variations have been applied to many fields during the last decades. Among them, several works attempt to achieve domain adaptation for the cross-modality matching task. For example, Shekhar et al. (2013) proposed to jointly map heterogeneous data into a common space and represented data using a shared dictionary in a common space for object detection. To preserve discriminative ability, they regularized the dictionary rather than encoding coefficients, which are significantly different from us. More recently, Liu et al. (2014) proposed semi-supervised coupled dictionary learning

for re-ID that learns two separated dictionaries to encode images from different domains to address alignment problem. Inspired by (Liu et al. 2014), Li et al. (2015) proposed a cross-view projective dictionary learning method for re-ID, which also learned two distinct dictionaries for each camera view. However, their model should be supervised by using paired samples across views, and paired dictionaries should be learned which highly rely on the expressive of the paired dictionaries. Peng et al. (2016) proposed to conduct transfer learning to achieve cross-dataset person re-identification. In 2017, Zhou et al. (2017a) proposed a joint model for person re-ID that performs dictionary and metric learning simultaneously. However, they targeted the traditional person re-ID and treated images from different cameras equally.

Motivated by these works, this thesis also considers reconstructing data across modalities by building a dictionary in order to address cross-modality person re-ID problem. It will be further discussed in Chapter 5.

2.3.2 Cross-view Person Re-ID Using Gaits

The development of approaches for cross-view person re-ID using gaits includes three stages, i.e., model-based methods, mapping-based methods and deep learning-based methods.

In the early stage, approaches are focused on constructing and analysing human body model, i.e., body parameters measurement (Bobick and Johnson 2001) and 3D reconstruction (Zhao et al. 2006). The former is simple but suffers poor performances. And in contrast, the 3D-based model achieved promising performance, but it is expensive and complicated since multiple calibrated cameras are needed.

After that, mapping-based approaches using 2D images are prevailing to address the cross-view gait recognition problem. These methods either learn to map gaits from one view to another view, i.e., VTM-based (view transformation model) models

(Kusakunniran et al. 2009, 2010; Zheng et al. 2011b; Muramatsu et al. 2016), or asymmetrically project gaits from different views to a shared space, i.e., CCA-like (canonical correlation analysis) models (Mansur et al. 2014; Ben et al. 2019b; Xing et al. 2016). Though this kind of approaches achieves promising performance, they heavily rely on view estimation. Moreover, these approaches only perform gait recognition across a couple of views which would yield abundant of models with an increasing number of views.

Recently, approaches using neural network achieves significant performances. For instance, Wu et al. (2017c) proposed to learn differences between gaits from arbitrary views with CNN. Shiraga et al. (2016) proposed GEINet based on CNN framework, which demonstrates effectiveness when view changes are small. However, these models suffer the interpretation problem due to the black-box characteristics of CNN. More recently, GAN is applied to gait recognition which tries to explicitly visualize the synthesized GEIs while keeping competitive performances, e.g. gaitGAN (Yu et al. 2017b) and MGANs (He et al. 2019). These models normalized gaits from different views into a reference one which behaves as a regressor but only uses one single model.

2.4 Summary

This chapter firstly revisits previous works on various person re-ID variants, including approaches and datasets for CST-reID, CCM-reID and LTG-reID. After that, we introduce some background knowledge and principle theories that coherently related to proposed methods in the thesis, such as techniques for human motion analysis, theories for the heterogeneous person matching and so on. In the following chapters, detailed solutions will be elaborated for addressing CLT-reID problems raised in Sec. 1.2.

Chapter 3

Long-term Person Re-identification using True Motion from Videos

This chapter introduces a novel person re-ID method based on true motion information called FIne moTion encoDing (FITD) to tackle the LTG-reID* problem. Unlike methods for CST-reID, FITD is proposed particularly for the long-term scenarios where TSIs are likely to re-appear and change their cloth after long-time interval in practice. It argues that motion characteristics are more reliable than static appearance feature to describe a walking person in LTG-reID applications because people's motion patterns are more stable than clothing with time lapse. FITD is designed to extract motion patterns hierarchically by encoding trajectory-aligned descriptors with Fisher vectors in a spatial-aligned pyramid, which shows promising performance. Moreover, this chapter introduces a new dataset called Motion-reID, which is the most challenging indoor LTG-reID dataset.

3.1 Introduction

3.1.1 Problem Formulation

As discussed in Section 2.1, most SOTA re-ID works concentrate on CST-reID, which yields methods building on the fact that the TSI keeps similar appearance between query and gallery set, e.g. clothing colour and texture do not change. This is true because people rarely change their clothes within a short interval. Regarding

*In this chapter, long-term person re-ID (CLT-reID) refers to the case of person re-identification after a long-time gap (i.e., LTG-reID).

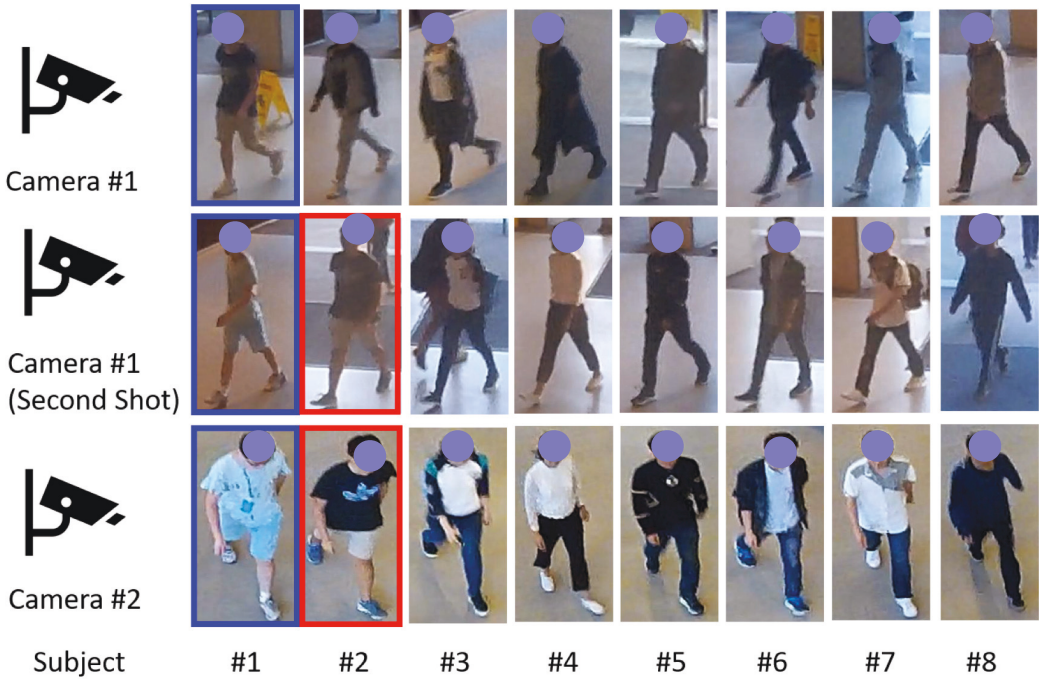


Figure 3.1 : Illustration of LTG-reID challenges. Each line of images is collected from same camera, and each column of images belongs to the same subject. Images from last two rows are captured by a long-time interval with the top row.

the fact, different appearance-based features have been exploited.

For LTG-reID, however, the TSIs are very likely to change their clothes. It causes larger intra-class appearances variations. That is, positive pairs are easier to be imposed by other subjects due to the dressing change. Intra-class cluster based on appearance becomes more dispersive while inter-class cluster based on appearance is more ambiguity than CST-reID. Features based on visual appearances, such as colour and texture histograms, are at a disadvantage when matching the same subject with distinct clothes from either the same camera or different cameras (see Figure 3.1 query subject bounded by blue line at top row is in a dark T-shirt and light short pant, but the positive matching in the gallery set at the rest rows are in blue clothing which is easily imposed by subject #2 in the gallery causing a

mismatch). It implies that different feature descriptors using fewer appearance cues are essential for the re-ID issue in LTG-reID.

3.1.2 Motivation

Inspired by the success of soft biometrics in gait and activity recognition across views (Kusakunniran et al. 2014; Wang and Schmid 2013), this chapter formulates a FITD model based on dynamic cues. The proposed FITD is true motion information extracted from dense trajectories, which characters dynamic motion patterns of the human body from raw footage without any scalability normalization. Especially, we adopt a patch-wise strategy (Cheng et al. 2016; Li et al. 2017b; Liu et al. 2015; Ma et al. 2016; Prosser et al. 2010) which divides human body into several fundamental body-action primitives. Fisher vectors are then utilized to respectively summarize the trajectory-aligned descriptors, e.g. Histograms of Optical Flow (HOF) (Laptev et al. 2008) and Motion Boundary Histogram(MBH) (Dalal and Triggs 2005), within each body-action unit (comprised by the fundamental body-action primitives) in the predefined body-action pyramid model. By this, both local and global motion statistics are computed. And, the final unified motion representation FITD is obtained by concatenating the bag of visual descriptors from all the body-action units.

In contrast to the method of (Gou et al. 2016), our FITD leverages trajectory-based true motion patterns from raw video volumes, and trajectory-aligned descriptors are embedded before Fisher encoding to get more robust motion information. Besides, a body-action pyramid model is considered to obtain both global and local motion information to boost feature discriminability for re-ID tasks.

3.2 Fine Motion Encoding

This section presents a novel spatio-temporal motion representation for person re-ID specific in long-term scenarios. As depicted in Figure 3.2, the proposed frame-

work includes two phases at which *model-training* learns a feature codebook consisting of discriminative motion primitives and *feature-extraction* encodes motion vocabularies to generate unified feature vectors. Particularly, each stage is performed on the basis of trajectory-aligned motion statistics with respect to body-action units corresponding to various levels of motion primitives in the predefined body-action pyramid (see section 3.2.1).

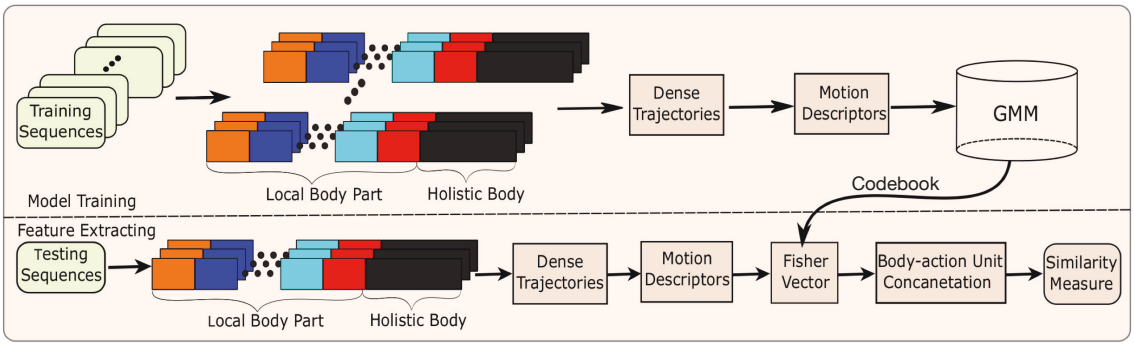


Figure 3.2 : Framework of the proposed FITD model. It consists of two phases: model training and feature extracting.

3.2.1 Body-action Pyramid Model

The human body is a non-rigid object which generates complex movement traces concerning its flexible joints while walking. This causes movement of the human body different from part to part as in Figure 3. It will lose local information if we only consider the motions of human body from a global view. Due to this, our body-action pyramid model (BPM) takes motion patterns of three-level body-action units into account.

Inspired by successes of some body-part based models (Liu et al. 2015; Ma et al. 2016; Li et al. 2017b), we define a BPM which divides human body from coarse to fine sub-regions with respect to some prior knowledge of geometry structure and kinematical characteristics of the human body. In specific, we depict the entire

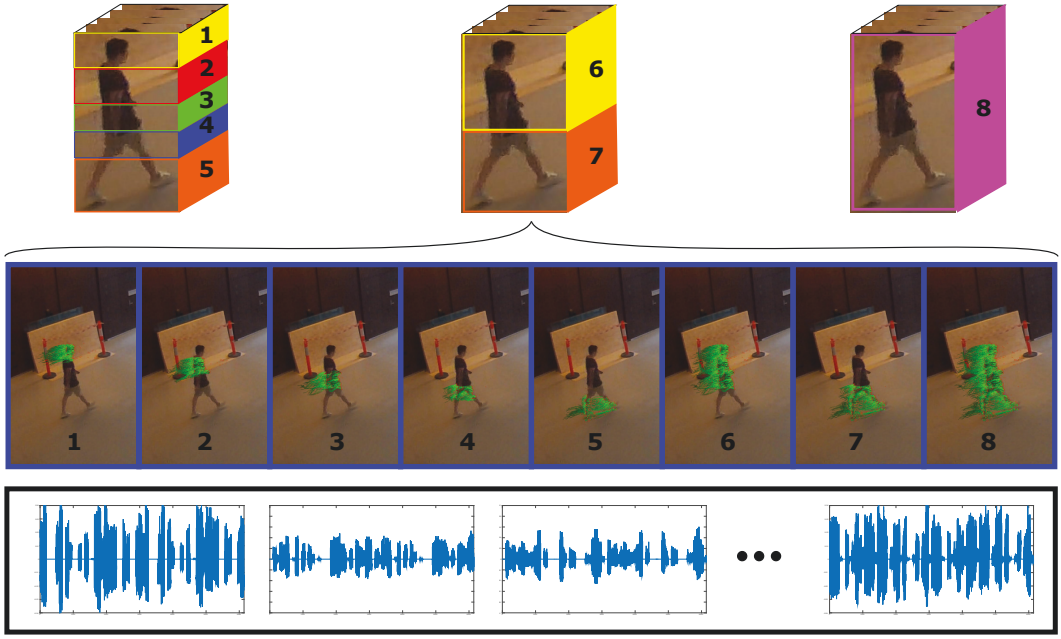


Figure 3.3 : Top: our proposed body-action pyramid model consists of eight body-action units, which is labelled from 1 to 8. Middle: dense trajectories in each body-action unit. Bottom: Fisher vectors correspond to the eight units.

human body in three levels, each of which includes a unique number of patches corresponding to various combinations of the neighbored action primitives, e.g. head (20%), upper torso (20%), lower torso (15%), upper leg (15%) and lower leg (30%). The action primitive template is empirically derived on the basis of the spatial structure of walking pedestrians from multiple benchmarks and fine-tuned in terms of motion characteristics. As shown in Figure 3.3, the top level (label number 8) describes the entire body, which is divided into two horizontal strips locating upper body (55%, label number 6) and leg (45%, label number 7) respectively due to the motion characteristics. Further, the upper body part is subdivided into three sections corresponding to the first three action primitives (label number 1, 2and 3)) whilst the leg section is subdivided into two parts corresponding to the last two action primitives (label number 4 and 5), so as to characterize the motion patterns in a finer way. The total of eight parts from three levels compromise our BPM.

According to the segmentation of the input video sequence above, the patches corresponding to pyramids of body parts are subsequently divided into eight body-action units, as shown in Figure 3.3,

$$\begin{aligned} P_m &= \{(x_t, y_t) | (x_t, y_t) \in P_{m,t}\}, \\ m &= 1, 2, \cdots, 8; t = 1, 2, \cdots, T \end{aligned} \tag{3.1}$$

where $P_{m,t}$ denotes the m-th patch of the t-th frame, (x_t, y_t) is the absolute position in the input video sequence.

In practice, the obtained body-action units are used to restrict action regions and identify whether an untracked feature point in the region should be appended to the tracking process. Detailed usage of the BPM will be introduced in the next section.

3.2.2 Motion Trajectories for Re-ID

To capture motion patterns of a walking pedestrian, we extract dense trajectories in each body-action unit, respectively. For ease of discussion, we take one single unit as an example. Inspired by the framework of dense trajectories (Wang and Schmid 2013), we tracked the sampled feature points $(x_t, y_t) \in P_{m,t}$ to the next frame $t + 1$ in a dense optical flow field $\omega = (\mu_t, \nu_t)$.

$$\begin{aligned} (x_{t+1}, y_{t+1}) &= (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)} \\ s.t. &(x_{t+1}, y_{t+1}) \in padding(P_{m,t+1}) \end{aligned} \tag{3.2}$$

where $padding(P_{m,t+1})$ denotes extending the patch $P_{m,t+1}$ by padding pixels from the neighbour areas. In our case, padding is carried out along vertical direction. That is, the area of size of 10% of body height up-ward and down-ward is added into the original patch. M is the filter kernel, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t).

It is worth noting that our feature point detection and tracking process are limited inside each body action unit. By this, feature points inside each body-action

unit are not only tracked but also avoided drifting to other units. Since the tracking process produces trajectories based on optical flow, very short trajectories related to homogeneous image areas (e.g. background) are pruned. Moreover, we consider bounding box of TSI in our framework, which largely suppresses the effects of other moving objects and generates pure trajectories of the pedestrian. As shown in Figure 3.3, dense trajectories are produced inside each pre-defined body-action units. These motion trajectories contain sufficient soft-biometric characteristics which are able to distinguish distinct human motion patterns such as different walking speed and stride.

3.2.3 Trajectory-aligned Motion Statistics

Local descriptors have been proved efficient by embedding motion characteristics in dense trajectories for many applications of activity recognition. To utilize motion information in our trajectories, we consider both HOF and MBH, which are popular to represent action characteristics. When formulating the above descriptors, we follow the setting in (Wang and Schmid 2013) to describe descriptors around the trajectories in a space-time volume.

In practice, HOF is applied to estimate local motion information with 9 bins covering all the orientations. It well describes the latent motion cues of walking styles because HOF is invariant to motion direction and strength. While HOF describes absolute motion, MBH encodes the relative motion between feature points. In particular, MBH descriptor treats horizontal (MBHx) and vertical (MBHy) components of optical flow separately, which yields motion information in both directions. It can be observed that, for walking person, motion information is stronger in the horizontal direction but subtle in the vertical direction. In this paper, we subdivide the trajectory volume into $2 \times 2 \times 3$ spatial-temporal grid and compute a descriptor within each grid. We further quantize each component, i.e., MBHx and

MBHy, equally along with orientation into 8 bins, and totally two 96-dimensional descriptors ($2 \times 2 \times 3 \times 8$) are obtained for each cell. MBH expresses human walking effectively, which yields excellent re-ID performances.

Other than the above two pure motion-based encoding descriptors, HOG is also considered due to its powerful gradient representation. However, HOG is commonly considered as a type of appearance-based feature. Different from the previous methods which apply HOG directly to an image, we implement it to a space-time video volume around dense trajectories embedding to our model. For HOG, we also perform on $2 \times 2 \times 3$ grid and quantize orientations into 8 bins. This is implicitly related to human's motion patterns.

3.2.4 Fisher Vector Encoding of Motions

As discussed in Section 2.2.2, the Fisher Vector (Sánchez et al. 2013) was first proposed to describe an image for large-scale visual classification and has gained remarkable success in many applications, e.g. activity recognition, image retrieval and even person Re-ID. Given a body-action unit P_m in the proposed body-action pyramid model, we describe the unit of a sample with N aforementioned descriptors, denoting as $X = \{x_n | x_n \in R^D, n = 1, \ldots, N\}$. For example, we extract N HOG descriptors to characterize one sample that each descriptor is from a spatial-temporal cell with $D = 96$ dimensions. To make the descriptors compact, we model them with K probabilistic visual vocabularies (PVVs) (Sánchez et al. 2013) which make the body-action unit complying a distinct distribution $P(X|\Theta)$, where $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$ is the parameters for the K PVVs. In this paper, the K PVVs are learned by a GMM Ψ with parameters $\theta_k = \{\mu_k, \sigma_k^2, \rho_k\}$, where μ_k, σ_k, ρ_k are respectively the mean vector, standard deviation and mixture weight,

$$\Psi = \sum_{k=1}^{K} \rho_k \psi_k, s.t. \rho_k \geq 0, \sum_{k=1}^{K} \rho_k = 1 \tag{3.3}$$

$$\psi_k(x;\mu_k,\sigma_k) = \frac{1}{(2\pi)^{D/2}|\sigma_k|} exp\{-\frac{1}{2}\|\sigma_k^{-1}(x-\mu_k)\|_2^2\} \tag{3.4}$$

where ψ_k is the k-th Gaussian component and x is the a feature descriptor mentioned above such as HOF and MBH. Once the GMMs are obtained, the Fisher vector of the sample in the body-action unit P_m is a concatenation of the deviations α_k^X, μ_k^X and ε_k^X such as $f(X) = [\alpha_1^X;\mu_1^X;\varepsilon_1^X;\ldots;\alpha_K^X;\mu_K^X;\varepsilon_K^X]$,

$$\alpha_k^X = \frac{1}{N\sqrt{\rho_k}} \sum_{n=1}^{N} (\gamma_{nk} - \rho_k) \tag{3.5}$$

$$\mu_k^X = \frac{1}{N\sqrt{\rho_k}} \sum_{n=1}^{N} \gamma_{nk}\rho_k^{-1}(x_n - \mu_k) \tag{3.6}$$

$$\varepsilon_k^X = \frac{1}{N\sqrt{2\rho_k}} \sum_{n=1}^{N} \gamma_{nk}\{\sigma_k^{-2}(x_n - \mu_k)^2 - e\} \tag{3.7}$$

where γ_{nk} is the posterior probability which determines whether descriptor x_n is generated by the k-th component or not, e is a D dimensional vector whose elements are all 1. By concatenating Fisher vectors along with all body-action units, we obtained our final high-level feature which depicts human's motion characteristics in a fine-grained way.

3.2.5 Feature Fusion

Feature fusion plays an important role when multiple features are available. In our case, different kinds of descriptors are considered such as DT, HOG, HOF and MBH. A straightforward strategy is concatenating descriptors on the feature level, which utilizes the mixed feature to train GMM model and generate Fisher vectors.

Another fusion strategy is aggregating the similarity metrics in the score-level, which refers to sum up the weighted similarity scores from different descriptors,

$$s_j = \sum_k \omega_k s_j^{(k)}, \forall_k : \omega_k \geq 0, \sum_k \omega_k = 1 \tag{3.8}$$

where $s_j^{(k)}$ is the similarity score between query sample and the j-th gallery sample with the k-th descriptor, ω_k weights the contribution of the k-th descriptor. For the sake of ease, we leverage Euclidean distance as our similarity score. Before fusion, the distances between samples using certain descriptors are first normalized to $[0, 1]$ by a min-max manner as in (Gou et al. 2016).

3.3 Dataset

As mentioned in Chapter 1, person re-ID are firstly proposed to track people among multiple non-overlap cameras, this implicitly restricts re-ID to the short-term scenarios and yields many corresponding benchmarks, e.g. VIPeR (Gray and Tao 2008), CUHK01-03 (Li et al. 2014), PRID2011 (Hirzer et al. 2011), iLIDS-VID (Wang et al. 2016), MARS (Zheng et al. 2016a). However, these benchmarks are insufficient to cover our case re-identifying a subject with long-term intervals, which explicitly increases the difficulties, i.e. more drastic illumination variation and clothing changes. This invokes us to construct a new dataset specific to the CLT-reID problem. This section first briefly revisits the existing benchmarks, especially benchmark video-based PRID2011, then introduces one new CLT-reID dataset together with its evaluation protocols.

3.3.1 Benchmark Datasets

Person re-ID research is further split into several different sub-topics each of which has different focuses, such as single-shot re-ID (Liao et al. 2015; Hirzer et al. 2012; Ma and Su 2012; Prosser et al. 2010), multi-shot re-ID (Cheng et al. 2016; Zheng et al. 2017a) and video-based re-ID (You et al. 2016; Liu et al. 2015; Wang et al. 2016; Gou et al. 2016). According to the specific research focus, various datasets are created. We refer readers to Section 2.1.4 for a comprehensive review.

Among the existing benchmarks, we take PRID2011 to evaluate our proposed

FITD for CST-reID. This is because PRID2011 is the only one with raw video and annotation information, which is necessary for our proposed FITD to extract true motion trajectories. PRID2011 dataset is captured under two disjoint cameras, where 385 and 749 identities are recorded by each camera, respectively. Among them, only the first 200 subjects appear in both cameras. Since the dataset is collected under the outdoor environment, multiple factors are included, e.g. viewpoint variance, lighting condition and background difference.

In this paper, we follow the protocol adopted in previous works (Gou et al. 2016; Wang et al. 2016; You et al. 2016). Only 178 of first 200 subjects are used, which have more than 25 frames in the footage of each subject. This is required in order to extract dense trajectories.

3.3.2 New Motion-ReID Dataset

Since the proposed FITD is specific to solve re-ID problem in long-term scenarios, we collect and annotate a new dataset named Motion-ReID. Some samples are shown in Figure 3.4. It includes video sequences extracted from two disjoint static surveillance cameras deployed in an office building, which covers the field of two distinct entrance gates respectively. We have collected total of 120 video clips from 30 persons, which half of them are captured by camera #1 and the rests are captured by camera #2. In particular, each subject is recorded twice under the same camera with a long-time interval which is at least one week. To make it clear, we list the recording timeline of one subject as shown in Figure 3.5. Opposite walking directions are separately recorded for one recording when person enters or exits a door (We use *front* and *back* to represent the distinct directions in the following sections). Each video sequence includes approximately 20 to 204 frames with an average 102 frames which cover at least one walking cycle. We have provided hand-crafted bounding boxes with varying size,

Figure 3.4 : Illustration of samples in the proposed Motion-ReID dataset.

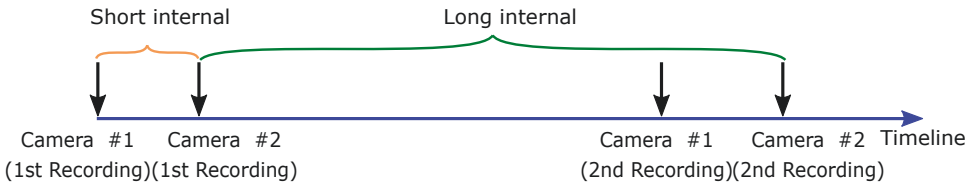


Figure 3.5 : Video recording timeline of a subject.

which makes the dataset is easy to evaluate re-ID algorithms. Considering the specific task, we developed eight challenging validation sets in terms of camera $C = \{camera\#1, camera\#2\}$, walking direction $D = \{back, front\}$ and recording time $R = \{1st, 2nd\}$. For all the sets, gallery and probe sets have significantly different recording time. As shown in Figure 3.6, the gallery and probe sets in the first four validation sets are recorded by same camera in same walking direction, e.g. $S_{1-4} = \{(C_1, D_1, R_1; C_1, D_1, R_2), (C_1, D_2, R_1; C_1, D_2, R_2), (C_2, D_1, R_1; C_2, D_1, R_2), (C_2, D_2, R_1; C_2, D_2, R_2)\}$, while the gallery and probe set in the rest sets are recorded by different cameras, e.g. $S_{5-8} = \{(C_1, D_1, R_1; C_2, D_1, R_2), (C_1, D_2, R_1; C_2, D_2, R_2), (C_1, D_1, R_2; C_2, D_1, R_1), (C_1, D_2, R_2; C_2, D_2, R_1)\}$. The validation setting covers all the long-term Re-ID situations. Here we do not consider walking directions since entering and exiting a door are exactly opposite direction that has little effect on re-ID algorithms.

Compared to current benchmarks, the dataset is more challenging because (1) The dataset is collected by real surveillance cameras rather than self-deployed ones, which makes the scenario more challenge; (2) Our dataset is specially collected for CLT-reID task, which brings out new challenges, e.g. wearing (clothing style and colour) and carrying condition changes.

3.4 Experiments

In this section, we evaluate the proposed FITD on both the long-term Motion-ReID dataset and benchmark PRID2011 dataset.

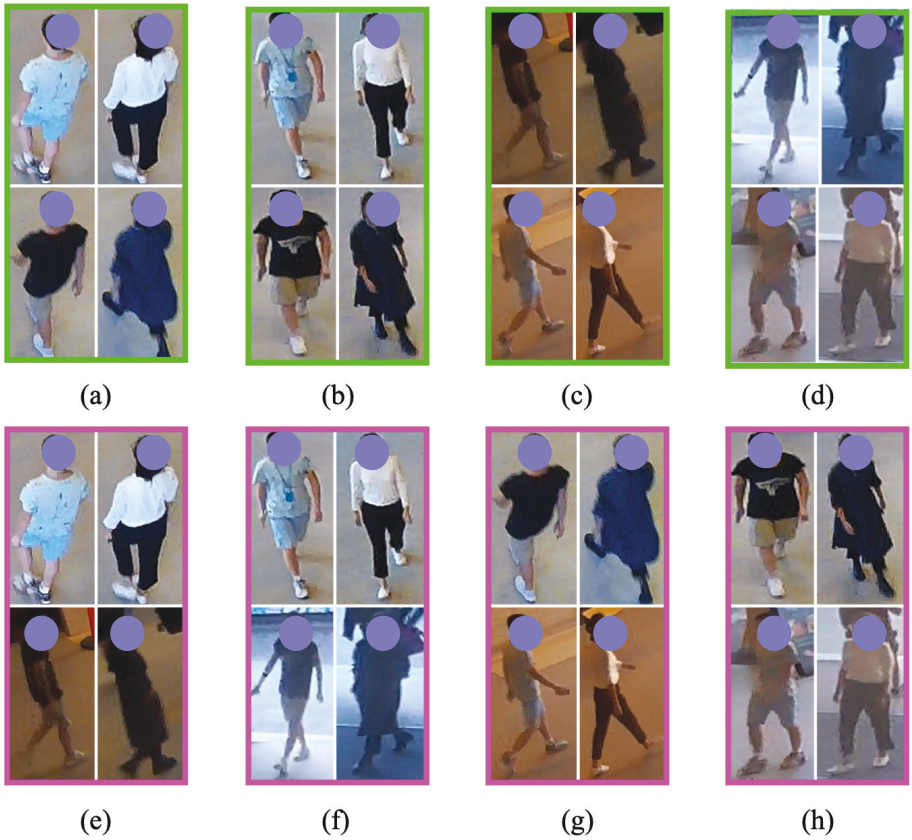


Figure 3.6 : An example of the eight evaluation subsets in Motion-ReID.

3.4.1 Experiment Setting

To highlight the importance of true motion cues for re-ID, we conduct all experiments without using any supervised metric learning method as in (Gou et al. 2016). For PRID2011, we randomly split the dataset into two equal subsets and compute the ranking scores for 10 trials (Gou et al. 2016; Wang et al. 2016; You et al. 2016). Average Cumulative Matching Characteristics (CMC) are reported for comparison. We also repeat the evaluation for extremely wearing similar case named PRID BK as in (Gou et al. 2016), which picks 35 samples with dark clothing forming testing set and 89 from the rest samples forming the training set. For Motion-ReID, we evaluate our method on all the 8 validation sets. Due to the small size of the dataset, we leverage 5-fold cross-validation method and repeat our experiments for

10 times. Average rank-1 matching scores are reported for all the validation sets.

Considering the periodicity of walking and tracking drift problem, we set trajectory length $L = 14$ and $L = 12$ respectively for PRID2011 and Motion-ReID, which roughly equal to half of a walking cycle. We find that the number of GMM components has little impact on the performance as in (Liu et al. 2015). Thus, we simply set to 32 in all our experiments. In practice, only μ_k^X and ε_k^X are reserved to construct Fisher vectors and thus the length of Fisher vectors for DT, HOG, HOF, MBHx and MBHy descriptors in one body-action unit are respectively 2048, 6144, 6912, 6144 and 6144 dimensions. The final FITD is the concatenation of all the body-action units in a fixed order. During testing, nearest neighbourhood classifier based on Euclidean distance is utilized to calculate the matching scores for all the methods.

3.4.2 Experiment Analysis

Effectiveness of FITD for CST-reID

In this section, we evaluated the proposed FITD on benchmark PRID2011 dataset and the cropped PRID BK dataset. To achieve stable performance, we use our Body-action Pyramid Model and concatenate Fisher vectors of all the units. A comprehensive comparison of different types of features for CST-reID case is conducted, and the results are reported in Table 3.1.

In Table 3.1, the features are roughly divided into three categories. Rows 1-3: a single appearance-based component, i.e. *colour or texture*; rows 4-8: ensemble appearance-based feature; rows 9-12: spatial-temporal feature. Noting that we use different trajectory-aligned descriptors for PRID 2011 and PRID BK, e.g. FITD with HOG encoding descriptor for PRID 2011 and FITD with HOGMBHx fused in score level for the PRID BK dataset, this is determined by different properties of the two datasets.

Dataset	PRID2011				PRID BK			
Rank	R-1	R-5	R-10	R-20	R-1	R-5	R-10	R-20
colour (Hirzer et al. 2012)	**9.33**	**29.78**	**39.21**	**60.00**	**12.86**	**31.43**	**41.43**	**70.00**
colourHist (Mignon and Jurie 2012)	2.36	10.45	19.21	35.62	2.86	22.86	32.86	65.71
LBP (Mignon and Jurie 2012)	3.03	14.49	21.91	35.62	7.14	25.71	32.86	68.57
colour & LBP (Hirzer et al. 2012)	10.22	27.19	38.65	60.79	7.14	25.71	41.43	65.71
colourHist & LBP (Xiong et al. 2014)	13.26	29.55	40.79	55.62	11.43	30.00	48.57	67.14
ELF (Gray and Tao 2008)	2.36	11.01	21.80	33.93	1.43	14.29	32.86	62.86
LOMO (Liao et al. 2015)	**22.81**	**61.46**	**77.19**	**88.31**	**40.00**	**67.14**	**82.86**	**94.29**
LDFV (Ma and Su 2012)	14.27	34.16	46.97	60.45	14.29	32.86	47.14	65.71
HOG3D (Wang et al. 2016)	22.92	46.52	59.78	73.15	22.86	50.00	64.29	85.71
STFV3D (Liu et al. 2015)	42.10	71.90	84.40	91.60	40.00	68.57	82.86	91.43
DynFV (Gou et al. 2016)	17.63	47.54	65.00	83.85	40.57	79.57	90.57	99.86
FITD **(Ours)**	**58.65**	**81.91**	**89.33**	**95.17**	**54.29**	**82.86**	**97.14**	**100**

Table 3.1 : A comparison of proposed FITD with other popular features on PRID2011 dataset and PRID BK dataset .

Single Appearance-based Components: Among the three single appearance-based feature component, using colour (Row 1) achieves better performance than that using histogram of colour (Row 2) and LBP (Row 3) on both the PRID 2011 and PRID BK datasets.

Ensemble Appearance-based Features: Rows 4-8 are SOTA appearance-based features used in CST-reID. Among them, LOMO achieves the best performance. The conclusion can be interpreted in two-fold. First, LOMO leverages Retinex images, which weakens illumination and colour gaps across cameras. This makes LOMO can extract refined colour features than extracting from raw images; and second, LOMO also extracts texture features using Scale Invariant Local Ternary Pattern (SILTP) which are more robust to noises than LBP. As expected, when taking the smaller testing set into account, the performances of appearance-based features which utilize colour cues drop notably, e.g. colourHist, colour & LBP, colourHist & LBP and

ELF.

Spatial-temporal Features: All of the four spatial-temporal features are developed to extracting information from videos. Notably, performances of all the spatial-temporal features do not decline when applying to the impaired PRID BK dataset. Both STFV3D and DynFV utilize Fisher vectors to describe features of a human, however, STFV3D represents more appearance-based feature, e.g. pixel value and gradient, while DynFV focuses more on motion patterns. Thus, DynFV is less discriminative than STFV3D on PRID2011 whist more powerful on PRID BK. Compared to STFV3D, our FITD with HOG extracts texture in space-time volume around dense trajectory, thus leading to higher performances on PRID2011. Compared to DynFV, our FITD with HOGMBHx extracts texture and motion from true video volume rather than normalized image sequences, and we use trajectory-aligned descriptors instead of raw trajectories. This explains why our FITD outperforms the DynFV by a large margin.

Effectiveness of FITD for CLT-reID

In this section, we evaluated our FITD on the proposed Motion-ReID dataset. Table 3.2-3.4 report our results on all the eight subsets. To better prove the benefits of our FITD, we evaluate it from three perspectives: trajectory-alignment encoding methods, fusion strategies and feature representations.

Trajectory-aligned Encoding Methods: Table 3.2 compares different encoding methods on all the eight subsets. Out of these encoding methods, FITD with motion descriptors achieves better performances than HOG in the first four subsets while FITD with HOG performs best in the last four subsets. This is not surprising because 1) Motion patterns are more discriminative in the first four subsets since video sequences from both gallery and probe in the first four subsets are captured from the same camera and clothing variation is the leading influential factor. 2) For

S_i	# 1	# 2	# 3	# 4	#5	#6	#7	#8
DT	55.5	60.0	40.3	42.0	20.7	19.3	19.0	22.3
HOG	56.7	55.0	57.7	48.3	27.3	**27.3**	**24.3**	18.3
HOF	52.0	58.7	54.7	49.0	**28.0**	20.7	19.0	22.7
MBHx	62.7	**65.0**	**59.7**	**55.7**	18.0	18.0	22.7	20.7
MBHy	**65.0**	60.7	58.3	50.7	16.7	14.7	17.7	**24.0**

Table 3.2 : A comparison of proposed FITD with different encoding methods.

the last four subsets, huge view difference between cameras affects motion seriously and consequently causes motion-based features less discriminative.

Fusion Strategies: Table 3.3 shows results of different types of fusion methods, Typically, Row 1-5 are fusion at the feature level, and Row 6-10 are fusion at the score level. As see the table, fusions in the score level outperform fusions in the feature level in most cases. Compared to the performance of FITD using single encoding descriptor, the fusion methods are more stable and improve the overall performance to some extent. Considering differences between first four subsets and last four subsets, we extract features by fusion representations HOGHOFMBH and HOGHOF in score level respectively for the two scenarios.

Feature Representations: Table 3.4 compares our FITD model with some SOTA feature representations as in the last section.

Single Appearance-based Components: Different from results on PRID2011, LBP achieves the best performance when gallery and probe samples are from the same camera, while colour is more discriminative when gallery and probe samples are from different cameras with huge viewpoint difference. However, performances of the appearance-based feature using no matter colour or texture degrade significantly

	Subset S_i	#1	# 2	#3	# 4	#5	# 6	#7	# 8
Feature Fusion	HOGHOF	55.00	**65.00**	55.67	53.00	**22.67**	20.33	23.33	19.67
	MBH	**64.67**	64.67	**60.67**	**56.67**	9.67	22.33	20.33	**26.00**
	HOFMBH	56.33	63.33	59.00	53.67	22.00	19.33	21.33	25.67
	HOGMBHx	59.00	60.67	58.00	54.67	21.33	**24.33**	24.33	23.33
	HOGHOFMBH	59.00	**65.00**	59.33	55.00	18.67	18.67	**27.33**	24.67
Score Fusion	HOGHOF	60.33	66.00	55.33	53.00	**30.33**	**26.33**	21.67	24.00
	MBH	64.33	64.67	**60.67**	54.00	18.00	16.00	20.33	22.67
	HOFMBH	61.33	66.00	60.33	55.67	21.67	15.00	18.33	**24.67**
	HOGMBHx	62.00	**66.67**	59.00	**56.67**	22.00	22.33	**23.33**	21.67
	HOGHOFMBH	**65.67**	**66.67**	**60.67**	55.33	24.00	18.00	21.33	24.00

Table 3.3 : A comparison of proposed FITD with different fusion methods.

Subset S_i	#1	# 2	#3	# 4	#5	# 6	#7	# 8
colour (Hirzer et al. 2012)	33.00	35.67	33.33	29.33	**26.00**	**25.00**	**26.33**	**23.33**
colourHist (Mignon and Jurie 2012)	33.33	37.00	28.67	**33.33**	23.33	17.33	17.33	14.33
LBP (Mignon and Jurie 2012)	**51.67**	**39.33**	**34.33**	27.67	19.00	19.67	18.00	20.33
colour & LBP (Hirzer et al. 2012)	38.67	38.67	**35.67**	30.67	18.67	**24.67**	**26.67**	**23.00**
colourHist & LBP (Xiong et al. 2014)	39.67	40.67	29.67	34.67	**24.67**	24.00	19.00	18.00
ELF (Gray and Tao 2008)	31.67	36.00	33.00	32.67	22.67	22.00	20.33	16.67
LOMO (Liao et al. 2015)	27.67	35.00	23.22	27.33	20.00	18.00	14.33	17.33
LDFV (Ma and Su 2012)	**49.33**	**41.67**	34.00	**35.33**	18.67	19.33	15.33	16.33
HOG3D (Wang et al. 2016)	39.67	30.33	37.33	34.67	13.67	14.33	17.67	20.67
STFV3D (Liu et al. 2015)	39.00	44.33	31.67	39.33	15.67	26.00	17.00	20.33
DynFV (Gou et al. 2016)	48.33	45.00	46.67	37.67	22.00	18.00	21.00	20.00
FITD **(Ours)**	**65.67**	**66.67**	**60.67**	**55.33**	**30.33**	**26.33**	**21.67**	**24.00**

Table 3.4 : A comparison of proposed FITD with other popular features on Motion-ReID dataset.

with camera changing and view enlarging. This is because camera changing and viewpoint variation impact texture and colour differently.

Ensemble Appearance-based Feature: Similar to single appearance-based components, performances of the commonly used ensemble appearance-based features also decline drastically. Out of these features, LDFV achieves the best performance in the first four subsets as in (Gou et al. 2016) where the gallery and query samples are obtained from the same camera. However, the performance of LDFV drops more sharply than other appearance features, which is the least discriminative among these features.

Spatial-temporal Features: Among the four spatial-temporal feature representation, motion-based features, e.g. DynFV and our FITD, outperform appearance-based features by a large margin in the first four subsets. The results prove the effectiveness of motion-based feature for long-term Re-ID. Since our FITD model extracts dense trajectory from raw/true video sequence rather than the normalized bounding area and encodes the trajectories with trajectory-aligned descriptors, our FITD model achieves better performance than DynFV. However, motion patterns are more easily affected by camera view changing, which causes performance of motion-based features declining sharply. This is also one of our future research points, which aims to solve view difference problem when using motion-based features. Noting that appearance-based features also perform regularly in the subsets, it is because several targets wore the same clothes and some partially changed their clothes in the dataset.

3.5 Summary

This chapter focuses on the LTG-reID problem and proposes to address it by exploring motion patterns from true/raw video sequences. The proposed model characterizes motion patterns by the trajectory-aligned descriptors in a three-level

body-action pyramid and Fisher vector encoding. Comprehensive experiments show that our method with appropriate trajectory-aligned descriptors benefits for the person re-ID, especially the LTG-reID problem. This tentative work fills the research blank in the field of CLT-reID. However, motion-based features suffer some new challenges, e.g. large walking view and camera viewpoint differences, partial occlusion and background movement. These problems will be discussed in the following chapters.

Chapter 4

GANs for CVGLT-reID

This chapter involves our exploration of view changing problem in long-term person re-ID using gaits. We term this case as cross-view gait-based long-term person re-ID (CVGLT-reID). In specific, it includes two kinds of solutions: view normalization and view transformation using GANs. The former introduces an identity-preserved Variation Normalizing Generative Adversarial Network (VN-GAN) that adopts a coarse-to-fine manner to normalize gaits from various views to an identical one while preserving identity information. The latter presents a View Transformation Generative Adversarial Networks (VT-GAN) that achieves view transformation on gaits across two arbitrary views using only one uniform model. Both frameworks generate visually promising/interpretable gaits and achieve competitive performance for the task of CVGLT-reID.

4.1 Problem Formulation

In the field of surveillance, gait is regarded as one of the most potential biometric features for person identification (Kusakunniran et al. 2009, 2014; Makihara et al. 2006; He et al. 2019; Yu et al. 2017b; Ben et al. 2019a). Unfortunately, person re-ID using gaits suffers serious view difference problem (Bobick and Johnson 2001; Kusakunniran et al. 2014; Yu et al. 2017b). It is because TSIs walk along with different directions against deployed cameras. The viewpoint difference causes gait pattern discrepancy between different views, which degrades the performance of traditional methods (Han and Bhanu 2005; Tao et al. 2007). Recently, some works try to address the problem, such as VTM-based models (Kusakunniran et al. 2009;

Makihara et al. 2006) and GAN-based models Yu et al. (2017b, 2019). However, the former requires to train multiple transformation models each of which correlates two views, while the latter cannot guarantee correct transformation because there is no constraint to force the normalized gaits to keep their identity information. This chapter focuses on the view different problem that aims to achieve view normalization/transformation using only one model and meanwhile preserve identity information while performing gait normalization/transformation.

4.2 Identity-preserved Variation Normalizing GAN

This section introduces the identity-preserved VN-GAN to learn purely identity-related representations. It adopts a coarse-to-fine manner which firstly generates initial coarse images by normalizing view to an identical one and then refines the coarse images by injecting identity-related information. In specific, Siamese structure with view classifier for camera view angles and identity preserver for human identities is utilized to achieve view normalization and identity preservation in two stages, respectively. In addition, reconstruction loss and identity-preserving loss are integrated, which forces the generated images to be the same in view and to be discriminative in identity. This ensures to generate identity-related images under the same view with good visual effect for person re-ID using gaits. Extensive experiments demonstrate that VN-GAN can generate visually interpretable results and achieve promising performance.

4.2.1 Motivations

To mitigate the effect of view difference, previous works concentrate on learning the latent relationship between gaits across different views. Two representative frameworks are view transformation model (VTM) (Kusakunniran et al. 2009, 2010; Makihara et al. 2006; Muramatsu et al. 2016) and coupled subspace learning (CSL)

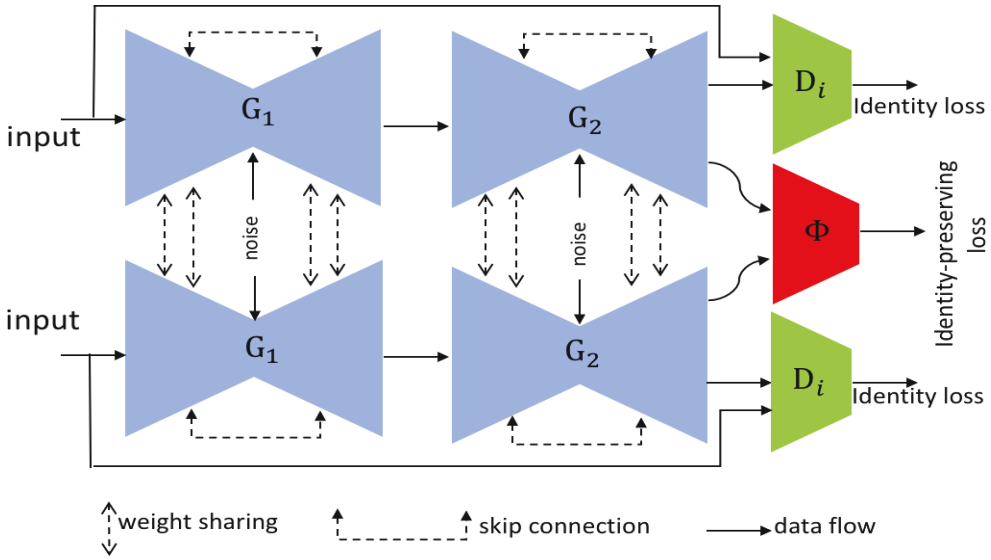


Figure 4.1 : Overview of Siamese structure of the proposed VN-GAN. It includes two-stream networks where each stream consists of the proposed coarse-to-fine design in Figure 4.2. Noting that the variation discriminator $\mathbf{D}_v$ and variation classifier $\mathbf{P}$ in Stage-I are not shown in the figure, because the figure only aims to show the Siamese design. Details of each branch is illustrated in Figure 4.2.

(Ben et al. 2016; Xing et al. 2016; Ben et al. 2019a). VTM aims to transform gaits from one view to another view by learning a regressor while CSL tries to map gaits from different views into a latent space in which view-specific feature is suppressed. Learning-based methods achieve great success for CVGLT-reID, which significantly improves the identification accuracy. However, these methods heavily depend on accurate view estimation and require to learn mutually independent models for each pair of views. Such solutions will incur abundant models and cause great computation costs.

Recently, some works attempt to extract view-invariant features using a single model such as ViDP (Hu et al. 2013) and CNN (Wu et al. 2017c) which achieve promising performances. However, the learned features are hard to explain because

of black-box characteristics of neural network. With the rise of GAN, it is available to normalize gaits to a unified view such as gaitGAN (Yu et al. 2017b) and MGANs (He et al. 2019). Moreover, these GAN-based methods not only achieve competitive performances but also generate visually realistic results that make the model more interpretable.

Considering benefits of GAN, this section also presents a two-stage GAN-based framework that normalizes gait templates, i.e., gait energy images (GEIs), from arbitrary views to a unified one such as 90° in a coarse-to-fine manner. Figure 4.1 shows the overall framework of the proposed VN-GAN. It includes a two-branch Siamese network (Taigman et al. 2014) that each branch consists of a coarse-to-fine generation structure (Ma et al. 2017) as in Figure 4.2. Our motivation is three-fold. Firstly, GAN shows successful in image generation Goodfellow et al. (2014) and image-to-image translation (Zhu et al. 2017). Moreover, it also achieves promising results in person generation guided by conditions such as pose (Ma et al. 2017; Ge et al. 2018; Zheng et al. 2019) and view Yu et al. (2017b, 2019). It inspires us to treat view normalization of gaits as an image translation task and translates gaits from different views into a unified one. Secondly, it is difficult to achieve view normalization and identity preservation simultaneously using an end-to-end network. Inspired by PG^2 (Ma et al. 2017), we propose to achieve the purpose in a divide-and-conquer strategy, which synthesizes the view-normalized gaits in Stage-I and refine the coarse results meanwhile inject identity information in Stage-II. Thus, we propose to embed a view classifier to Stage-I and an identity preserver to Stage-II. Finally, Siamese network (Bromley et al. 1994) demonstrates great success in many identification tasks, such as face verification (Taigman et al. 2014), person re-ID (Chung et al. 2017), etc. To explore identity discriminability, we adopt it as our basic structure and regularize the training process using prevailing contrastive loss (Sun et al. 2014).

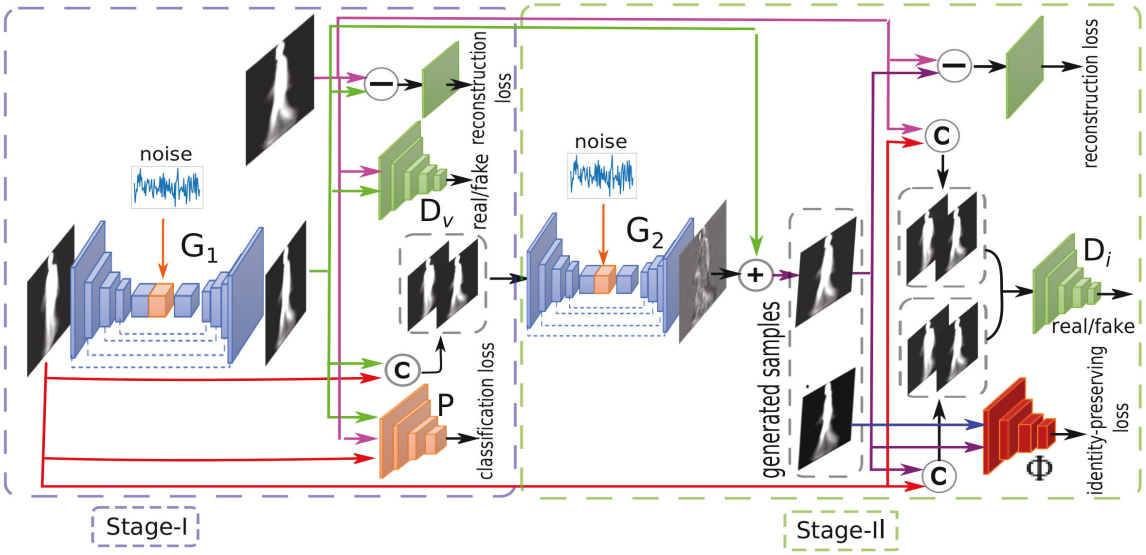


Figure 4.2 : Overview of the coarse-to-fine structure. It is sub-branch of the proposed VN-GAN which consists of two stages: coarse gait image generation and refinement. In the framework, ⓒ denotes channel concatenation and ⊖ denotes image difference, ⊕ denotes image summation.

4.2.2 Proposed Method

Our task is to reduce the data discrepancy caused by view variation in person re-ID using gaits. It is achieved by normalizing gait images of different views to a typical view, i.e. side view 90°. Figure 4.1 shows the overall framework of the proposed VN-GAN. It adopts the prevailing two-branch Siamese network as a basic structure that each branch includes a coarse-to-fine generation process as shown in 4.2. Such a divide-and-conquer design makes it easier to achieve both view normalization and identity preservation simultaneously.

In Stage-I, we adopt a GAN to achieve the coarse generation. It includes a generator $\mathbf{G}_1$ and a discriminator $\mathbf{D}_v$ that beats against each other. The generator $\mathbf{G}_1$ takes gaits x_{src} from any view and noise z as input and tries to produce gaits $x_{dst}^{(1)}$ in the identical view. The discriminator $\mathbf{D}_v$ takes the generated gaits $x_{dst}^{(1)}$ and

real identical-view gaits x_{dst} as input to distinguish whether the input is generated. To ensure the produced gaits are from the identical view, we propose to use a view classifier $\mathbf{P}$ to regularize the generation process.

In Stage-II, we adopt another GAN to refine the coarse results from Stage-I and meanwhile implant the identity information. Inspired by refinement method in (Ma et al. 2017), the generator $\mathbf{G}_2$ takes the concatenation of original input x_{src} and produced coarse result $x_{dst}^{(1)}$ as input, and produces the refinement/different map between final output $x_{dst}^{(2)}$ and coarse result $x_{dst}^{(1)}$. Such design helps supplement more details from x_{src} to refine the coarse results $x_{dst}^{(1)}$. Since generator $\mathbf{G}_2$ takes x_{src} as input, it may mislead $\mathbf{G}_2$ directly produce x_{src} rather than refine $x_{dst}^{(1)}$. To tackle the issue, discriminator $\mathbf{D}_i$ takes real pair (x_{dst}, x_{src}) and fake pair $(x_{dst}^{(2)}, x_{src})$ as input. Such design encourages $\mathbf{D}_i$ to discriminate between $x_{dst}^{(2)}$ and x_{dst} instead of learn difference between real and generated gaits. To make the view-normalized gaits keep their identity information, we propose to enforce a identity preserver $\mathbf{\Phi}$ and regularize Stage-II with a contrastive loss. The following gives details of the coarse-to-fine design.

Stage-I: View Normalization

At Stage-I, we aim to normalize gaits from arbitrary views to a specific view. As in Figure 4.2, this stage includes a generator $\mathbf{G_1}$, a variation discriminator $\mathbf{D}_v$ and a view preserver $\mathbf{P}$.

Generator $\mathbf{G_1}$. we utilize a U-Net-like architecture (Ma et al. 2017) to generate coarse gaits of the specific view. Specially, we first use several stacked convolutional layers to extract feature embeddings in bottleneck from source images x_{src} in arbitrary views. To simulate variations among intra-class samples, we integrate a random noise z in the bottleneck layer. After that, the feature embedding is decoded using a set of stacked deconvolutional layers which is symmetric to the encoder.

For the sake of information loss, skip connections between encoder and decoder are utilized in the U-Net to propagate gait information directly from source to target. The generated coarse gait in Stage-I is denoted as $x_{dst}^{(1)} = \mathbf{G}_1(x_{src}, z)$.

View Discriminator $\mathbf{D_v}$. In GANs, discriminator competes against generator to distinguish whether the input is generated or not. In Stage-I, we adopt a view discriminator $\mathbf{D}_1$ that forces the generated gaits $x_{dst}^{(1)}$ to follow the distribution of target samples $\mathbf{x}_{dst}$ of the same subject in the reference/target view. That is, $\mathbf{D}_v$ takes the synthesize image $x_{dst}^{(1)}$ and real image $\mathbf{x}_{dst}$ in reference view as input, and best distinguish generated image against the real image.

View Classifier P. To ensure view consistency between generated image $x_{dst}^{(1)}$ and target image x_{dst}, we regularize the generation process by a view classifier $\mathbf{P}$. It distinguishes the view angle of the generated image against the view of real image. In practice, the view classifier $\mathbf{P}$ can be pre-trained beforehand. In this work, we integrate it into our proposed framework and train it alternatively with $\mathbf{G}_1$ and $\mathbf{D}_v$ by following an end-to-end manner. When training the generator $\mathbf{G}_1$, gradients from $\mathbf{P}$ are back-propagated to jointly optimize $\mathbf{G}_1$ with adversarial loss, and regularize generated gaits on the reference view, i.e., the normalized standard view.

To force the synthesised gaits near the ground truth, we impose l_1 distance between generated image $x_{dst}^{(1)}$ and target images x_{dst} in the generation process, which is defined as

$$\ell_{id_1}(x_{src}, x_{dst}) = \|x_{dst}^{(1)} - x_{dst}\|_1 \tag{4.1}$$

Therefore, objective functions for generator $\mathbf{G}_1$, discriminator $\mathbf{D}_v$ and view classifier

$\mathbf{P}$ in Stage-I are defined respectively,

$$\begin{aligned}
\ell_{G_1} &= \ell_{bce}(\mathbf{D}_v(\mathbf{G}_1(x_{src}, z), 1)) + \alpha\ell_{id_1}(\mathbf{G}_1(x_{src}, z), x_{dst}) \\
&\quad + \beta\ell_{bce}(\mathbf{P}(\mathbf{G}_1(x_{src}, z)), 1), \\
\ell_{D_v} &= \ell_{bce}(\mathbf{D}_v(\mathbf{G}_1(x_{src}, z)), 0) + \ell_{bce}(\mathbf{D}_v(x_{dst}), 1), \\
\ell_P &= \ell_{bce}(\mathbf{P}(x_{src}), 0) + \ell_{bce}(\mathbf{P}(x_{dst}), 1),
\end{aligned} \tag{4.2}$$

where ℓ_{bce} denotes binary cross-entropy loss, α and β are trade-off parameters of ℓ_{id_1} loss and view classifier loss ℓ_P, respectively. α controls the level of similarity between generated artifacts and real targets in terms of appearance. Large α incurs blurry results since ℓ_1 loss encourages to average all possible cases, while small α causes artifacts due to domination of adversarial loss at training phase (Ma et al. 2017). β determines the contribution of view classifier, which also averages all possible cases and blurs results specified by the target view if it is too large.

In Stage-I, the output of $\mathbf{G}_1$ captures the global structure information specified by the target view, as shown in Figure 4.2, which normalizes the source gaits from arbitrary views into the specific one. Though Stage-I encourages the appearance between produced gait and the real target looks similar, it cannot guarantee the identity discriminability. Thus, Stage-II refines the produced coarse gait using another GAN and meanwhile enforce identity discriminability.

Stage-II: Identity Embedding

Since Stage-I has synthesised coarse gait image with a similar appearance to real ground truth in both global structure and view angle, Stage-II will target on making the generated gaits more identity discriminative, which pulls the generated gaits of the same identity closer while pushes that of different identities further.

Generator $\mathbf{G}_2$. Similar to (Ma et al. 2017), we also adopt a variant of DCGAN (Radford et al. 2015) conditioned on the output of Stage-I as our base model.

Considering that the initial output of Stage-I is globally similar to the target, we propose to simultaneously narrow their gaps and incorporate more details by generating appearance difference map. In particular, we use a similar U-Net architecture Ma et al. (2017) as $\mathbf{G}_1$ to achieve generator $\mathbf{G}_2$. But, $\mathbf{G}_2$ works differently from $\mathbf{G}_1$ which takes concatenation of source image x_{src} and coarse result $x_{dst}^{(1)}$ from Stage-I as input and produces difference map between $x_{dst}^{(1)}$ and ground truth x_{dst}. Such design benefits to supplement more details and speed up the training process.

Identity Discriminator $\mathbf{D}_i$. Different from Stage-I, we do not directly produce a single target gait image but generate difference map. The final synthesised gait $x_{dst}^{(2)}$ is the summation of initial results $x_{dst}^{(1)}$ and the difference map. As $\mathbf{G}_2$ takes x_{src} as input, it may mislead $\mathbf{G}_2$ directly synthesize x_{src} instead of refining $x_{dst}^{(1)}$. To tackle the issue, discriminator $\mathbf{D}_i$ takes image pairs $(x_{dst}^{(2)}, x_{src})$ and (x_{dst}, x_{src}) as input. This makes $\mathbf{D}_i$ not only recognize the difference between synthesised and natural images but also the identity distinction between $x_{dst}^{(2)}$ and x_{dst}. In another words, the pairwise input encourages $\mathbf{G}_2$ to synthesize identity-preserving result with target. This is beneficial to the identification task.

Identity Preserver $\mathbf{\Phi}$. To keep identity discriminability of synthesised gaits, we integrate an identity preserver $\mathbf{\Phi}$ in Stage-II. Benefits of the Siamese structure, we impose an identity-preserving loss to the end of $\mathbf{\Phi}$, which pulls the synthesised gaits of the same subject together and push the synthesised gaits of different subjects away. In this work, we adopt the contrastive loss which is defined as

$$\ell_{\Phi}(x_1, x_2, y) = \frac{1}{2}\{yd^2 + (1-y)[max(0, \rho - d)]^2\}, \tag{4.3}$$

where d denotes the Euclidean distance between normalized input embeddings of two generated gaits x_1 and x_2 from two-branch of the Siamese structure, which is defined as $d(x_1, x_2) = \|\Phi(x_1) - \Phi(x_2)\|_2$. And, y denotes the binary label of the input pair which indicates x_1 and x_2 are positive pair if $y = 1$ and 0 otherwise. $\rho \geq 0$ is

the margin that determines the separability of generated samples in the embedding space defined by Φ. If $\rho = 0$, only gradient of positive pairs will be back-propagated and thus it dominates the training process. While $\rho > 0$, both negative and positive pairs are taken into account.

Reconstruction loss. As stated in (Isola et al. 2017) and (Ge et al. 2018), it is helpful to reduce blurriness and generate human-perceivable images if regularising generation process by ℓ_1 distance between fake and real images. Therefore, we introduce a reconstruction loss by minimizing the ℓ_1 distance between the synthesised gaits $x_{dst}^{(2)}$ in Stage-II and their corresponding real targets x_{dst},

$$\ell_{id-2}(x_{src}, x_{dst}) = \mathbb{E}_{x_{src} \sim p_{data}(x_{src})} \| x_{dst}^{(2)} - x_{dst} \|_1 \tag{4.4}$$

where $x_{dst}^{(2)}$ represents output of Stage-II which is the sum of initial result in Stage-I and the generated difference map in Stage-II, denoting as $x_{dst}^{(2)} = x_{dst}^{(1)} + \mathbf{G}_i([x_{src}; x_{dst}^{(1)}], z)$, and $[\cdot;\cdot]$ represents tensor concatenation.

As mentioned above, losses of the generator, identity discriminator and identity preserver work collaboratively to generate identity-preserved gaits in Stage-II. Thus, losses for $\mathbf{G}_2$ and $\mathbf{D}_i$ are defined respectively as

$$\begin{aligned} \ell_{G_2} &= \ell_{bce}(\mathbf{D}_i([x_{src}; x_{dst}^{(2)}]), 1) + \alpha \ell_{id-2} + \gamma \ell_{\Phi} \\ \ell_{D_i} &= \ell_{bce}(\mathbf{D}_i([x_{src}; x_{dst}]), 1) + \ell_{bce}(\mathbf{D}_i([x_{src}; x_{dst}^{(2)}]), 0) \end{aligned} \tag{4.5}$$

where α and γ are trade-off parameters which balance the contribution of reconstruction loss and identity-preserving loss. In practice, we alternatively optimize generator $\mathbf{G}_2$, identity discriminator $\mathbf{D}_i$ and identity preserver $\mathbf{\Phi}$ by minimizing losses ℓ_{G_2}, ℓ_{D_i} and ℓ_{Φ}, respectively.

4.2.3 Network Architecture Implementation

In this section, we summarize the architecture of the proposed VN-GAN. In both stages, we adopt the U-net structure (Ronneberger et al. 2015) with skipped connec-

tions to construct our generators. It means that each generator includes an encoder of N Convolution-InstanceNorm-LeakyReLu blocks to down-sample the input images into bottleneck and a decoder of N Deconvolution-InstanceNorm-ReLu blocks to up-sample to images according to reference distribution. In our experiment, we set $N = 4$. It is worth note that random noise is imposed to concatenate with features from encoder in the bottleneck and an additional Convolution-InstanceNorm is performed before the concatenated feature transmitting to decoder. In addition, skip connections between encoders and decoders are applied which can be seen in Figure 4.1 and Figure 4.2. For discriminators in both stages, we utilize the structure of PatchGAN (Isola et al. 2017), which models high-frequencies of image. This is critical to prevent much more smoothness caused by reconstruction loss. At Stage-I, variation classifier $\mathbf{P}$ consists of one Convolution-ReLu layer, three Convolution-BatchNorm-LeakyReLU layers and one convolution layer. At Stage-II, identity preserver $\mathbf{\Phi}$ shares the same structure with $\mathbf{P}$ except an additional fully connected layer. In the framework, all convolution/deconvolution layers contain 4×4 filters and the number of filters is doubled/halved with each block except last convolution layers which are 1 for both variation discriminator and variation classifier. In addition, strides are set to 2 for all convolution layers except last layer of discriminator, variation classifier and identity preserver, which are set 1, 4 and 4, respectively.

4.2.4 Experiments

In this section, we evaluate the proposed VN-GAN on widely-used gait database, i.e., CASIA(B) dataset (Yu et al. 2006), against various influencing factors such as view differences, changes of carrying conditions and clothing variation.

Experiment Setup

Dataset. CASIA(B) gait dataset includes about 13640 sequences of 124 subjects from 11 views, i.e., $0°, 18°, \ldots, 180°$ with $18°$ interval. For each view, there are 10 sequences of which six are collected under normal walking (*NM01-NM06*), two are captured under walking with a bag (*BG01-BG02*) and the rest two are taken under walking with a coat (*CL01-CL02*). Noting that the popular spatio-temporal template for gait, i.e., gait energy image (GEI) (Han and Bhanu 2005), formed by aligned silhouettes in a walking cycle has been provided with the original videos. It has been proved noise robustness and computation efficiency by previous works (Han and Bhanu 2005; Ben et al. 2016; Wu et al. 2017c; Kusakunniran et al. 2014). Figure 4.3 lists an example of GEIs from 11 views of different people, carry conditions and cloth. Therefore, we directly use the provided GEIs as our input to perform person re-ID using gaits.

Settings. In all experiments, we follow the setting in (Yu et al. 2017b) and (Yu et al. 2017c) which splits the first 62 people to form the training set and the rest 62 people to form testing set. When performing training, all samples of the first 62 people are used to normalize the GEIs from arbitrary views into a reference one, i.e., $90°$. In the testing phase, four of normal walking sequences (Gallery: *NM01 – NM04*) are taken as gallery set. And, three challenging probe sets against different influencing factors are built, i.e., the rest two normal walking sequences (i.e., ProbeNM: *NM05, NM06*), two walking with a bag (i.e., ProbeBG: BG01, BG02), two walking with a coat (i.e., ProbeCL: *CL01, CL02*). In our experiments, all GEIs are resized into 64×64 pixels.

Training Details. We use Tensorflow (Abadi et al. 2016) to train the proposed VN-GAN on the training set. In particular, we utilize Adam optimizer (Kingma and Ba 2014) with parameters $\beta_1 = 0.5, \beta_2 = 0.999$ and set initial learning rate

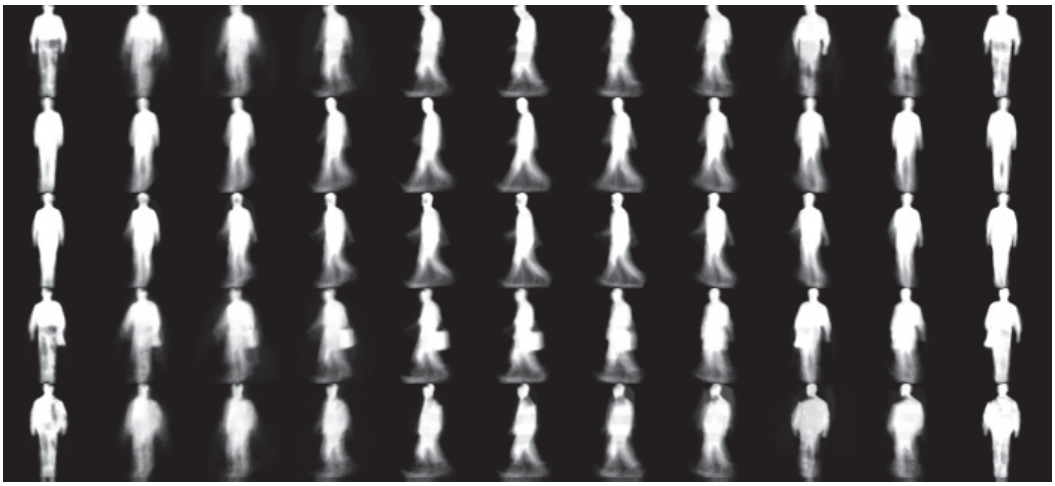


Figure 4.3 : An example of GEIs from 11 views, i.e., $0°, 18°, \cdots, 180°$ from left to right, and images at each column stand for GEIs from the same view angle. The first three rows compares GEIs from three different people in normal walking pattern. The last two rows includes the same subject with the first row but different walking conditions, i.e., the subject carries a bag in the forth row and wear a coat in the last row.

to 0.0002. We respectively train each stage with a mini-batch of size 100 for $5k$ iterations. Empirically, the balance parameters are set $\alpha = 10, \beta = 5, \gamma = 10$. For the contrastive loss, we empirically set margin $\rho = 0.8$. In the training process of Stage-I, we alternatively optimize Generator $\mathbf{G}_1$, discriminator $\mathbf{D}_v$ and variation classifier $\mathbf{P}$. However, the variation classifier $\mathbf{P}$ can also be pretrained ahead of our model training. When the training of Stage-I is finished, we fixed the variables in Stage-I and alternatively optimized generator $\mathbf{G}_2$, identity discriminator $\mathbf{D}_i$ and identity preserver $\mathbf{\Phi}$. As in Figure 4.2, Stage-II takes the output of Stage-I as input and generates difference maps, which refines the results of Stage-I.

Evaluation Metrics. Nearest Neighbour Classifier (NNC) based on Euclidean distance is adopted to evaluate performances of the proposed VN-GAN. Before performing distance measurement, linear discriminant analysis (Mika et al. 1999) is

Table 4.1 : Identification accuracy on ProbeNM subset.

Acc. (%)		Probe View										
		0	18	36	54	72	90	108	126	144	162	180
Gallery View	0	99.2	86.3	66.1	58.9	40.3	41.1	46.0	54.8	52.4	72.6	81.5
	18	92.7	100.0	95.2	80.6	66.1	62.9	68.5	62.9	67.7	78.2	66.1
	36	73.4	92.7	96.8	96.0	83.9	75.8	78.2	79.0	78.2	69.4	56.5
	54	49.2	76.6	92.7	97.6	92.7	86.3	89.5	81.5	82.3	64.5	46.8
	72	46.8	63.7	84.7	91.1	97.6	96.0	92.7	87.1	68.5	58.9	41.9
	90	41.9	56.5	80.6	87.1	96.0	96.0	95.2	88.7	78.2	56.5	37.9
	108	38.7	57.3	71.8	84.7	93.5	95.2	96.8	96.8	87.9	62.1	39.5
	126	44.4	59.7	75.0	83.9	87.9	87.1	93.5	96.8	94.4	76.6	50.0
	144	49.2	54.8	71.0	80.6	79.8	75.8	89.5	96.0	99.2	86.3	51.6
	162	68.5	75.0	68.5	62.1	51.6	54.8	62.9	73.4	85.5	98.4	84.7
	180	83.9	63.7	53.2	42.7	37.1	37.1	40.3	51.6	53.2	83.9	98.4

firstly utilized to pre-process the generated GEIs. Identification accuracies at rank-1 place are reported on all the probes.

Evaluation

In this section, we evaluate performances of the proposed VN-GAN on three challenging probe sets, i.e., ProbeNM, ProbeBG and ProbeCL, which covers variations against view, clothing and carrying conditions.

Effects of Influencing Factors. We measure identification accuracies between each view pair in this section. The results are reported in Table 4.1 – Table 4.3 in terms of three influencing factors in CASIA(B) gait dataset. There are total of 121 pairs of view combination in the table, and we measured identification accuracy across views of each pair. Table 4.1 lists the results on ProbeNM, i.e., for each view pair, we take *NM01-04* in one view as gallery and *NM05-06* in the other view as

Table 4.2 : Identification accuracy on ProbeBG subset.

Acc. (%)		Probe View										
		0	18	36	54	72	90	108	126	144	162	180
Gallery View	0	74.2	61.3	50.0	32.3	32.3	21.0	22.6	28.2	37.1	48.4	60.5
	18	63.7	81.5	72.6	53.2	46.0	32.3	33.9	35.5	42.7	51.6	41.9
	36	48.4	73.4	83.9	69.4	54.0	47.6	37.9	47.6	50.8	47.6	29.0
	54	36.3	52.4	71.8	78.2	68.5	53.2	56.5	68.5	49.2	41.9	25.0
	72	32.3	48.4	62.1	67.7	79.8	66.1	62.1	69.4	50.8	39.5	21.8
	90	26.6	38.7	58.9	61.3	75.0	71.0	68.5	66.9	52.4	37.1	24.2
	108	28.2	45.2	59.7	62.9	75.0	71.8	75.0	71.8	53.2	40.3	21.0
	126	32.3	46.0	51.6	59.7	58.9	54.0	64.5	79.0	65.3	50.8	20.2
	144	33.1	47.6	50.0	49.2	53.2	43.5	55.6	71.0	74.2	54.0	32.3
	162	46.0	54.0	45.2	37.9	32.3	31.5	31.5	43.5	49.2	69.4	57.3
	180	58.1	52.4	42.7	25.8	28.2	21.0	19.4	31.5	45.2	52.4	74.2

Table 4.3 : Identification accuracy on ProbeCL subset.

Acc. (%)		Probe View										
		0	18	36	54	72	90	108	126	144	162	180
Gallery View	0	31.5	25.8	22.6	15.3	11.3	12.9	16.1	16.1	22.6	17.7	21.8
	18	28.2	32.3	32.3	24.2	22.6	21.8	20.2	24.2	18.5	25.0	15.3
	36	23.4	33.9	39.5	39.5	37.1	29.8	29.0	30.6	28.2	26.6	13.7
	54	20.2	27.4	32.3	39.5	36.3	35.5	32.3	29.0	30.6	23.4	12.9
	72	17.7	26.6	31.5	32.3	50.8	40.3	36.3	32.3	29.8	26.6	12.9
	90	19.4	27.4	32.3	41.9	50.8	52.4	41.9	41.1	25.8	23.4	17.7
	108	22.6	25.0	29.8	37.1	45.2	41.1	46.8	37.1	30.6	29.8	17.7
	126	25.0	26.6	31.5	33.1	31.5	31.5	33.9	33.9	34.7	27.4	14.5
	144	28.2	25.8	25.0	22.6	25.0	25.0	35.5	36.3	42.7	36.3	18.5
	162	25.0	23.4	22.6	16.1	21.0	16.9	20.2	22.6	31.5	33.9	25.0
	180	21.8	20.2	20.2	12.9	8.9	11.3	16.1	15.3	14.5	19.4	22.6

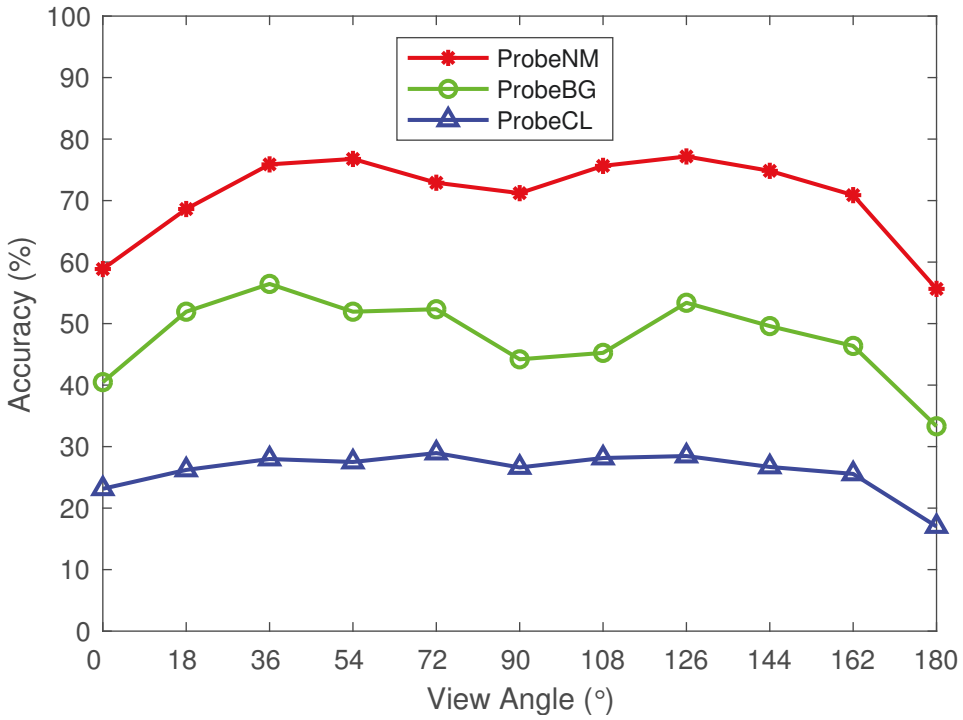


Figure 4.4 : Average accuracies except the identical view on three probe sets.

query. It responds to normal walking conditions. Table 4.2 and Table 4.3 report the results on ProbeBG and ProbeCL, respectively, which use the same gallery set but different probe set compared to Table 4.1. From the three tables, it is easy to observe that identification accuracy is severely influenced by view changes. In specific, larger view difference incurs poorer performance. However, the performance increases at the symmetrical view such as 54° *vs*. 126°. Thus, there two identification accuracy peaks except for the case that probe view is 90° because 90° is in the middle. Figure 4.4 also verifies the conclusion.

In another aspect, we can observe that carry condition and wearing a coat also degrade the identification accuracies drastically. And, wearing a coat causes the worst performance among all the mentioned variations. This is reasonable because 1) wearing a coat changes the shape of human silhouettes, and 2) wearing a coat also occludes human body which causes incomplete motion patterns. This conclusion is

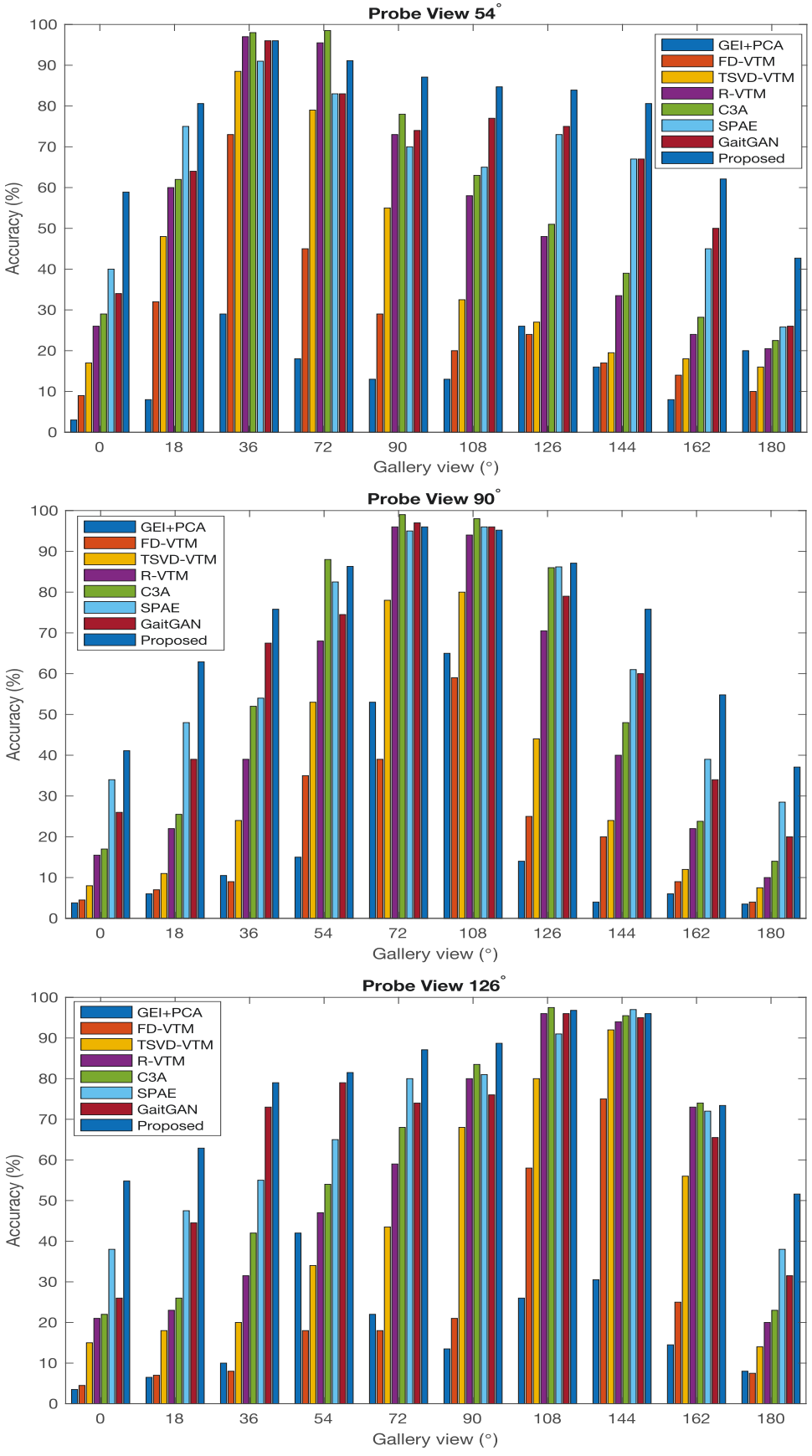


Figure 4.5 : Comparison with the SOTA approaches on ProbeNM subset at three distinct probe views $54°, 90°$ and $126°$.

Table 4.4 : Comparison of average accuracies of 10 gallery views (The corresponding view is excluded.) on ProbeNM with probe view $54°, 90°$ and $126°$.

Method	Probe View			
	54°	90°	126°	Average
Baseline (Yu et al. 2006)	0.16	0.23	0.17	0.19
ViDP (Hu et al. 2013)	0.64	0.60	0.65	0.63
C3A (Xing et al. 2016)	0.57	0.55	0.58	0.57
SPAE (Yu et al. 2017c)	0.63	0.62	0.66	0.64
GaitGAN (Yu et al. 2017b)	0.65	0.58	0.66	0.63
MGANs (He et al. 2019)	0.77	0.67	0.79	0.74
Proposed	0.77	0.71	0.77	0.75

also supported by Figure 4.4.

Comparison with SOTA methods. In this part, we compare performances with some SOTA methods for CVGLT-reID, e.g., GEI+PCA (Han and Bhanu 2005), FD-VTM (Makihara et al. 2006), TSVD-VTM (Kusakunniran et al. 2009), R-VTM (Kusakunniran et al. 2012), C3A (Xing et al. 2016), SPAE (Yu et al. 2017c), GaitGAN (Yu et al. 2017b), ViDP (Hu et al. 2013) and MGAN (He et al. 2019). Typically, we select three probe views, i.e., 54°, 90° and 126°. Figure 4.5 compares the performances of each probe angle against the rest view angles. From the figure, we can observe that view changes significantly affect identification accuracy and larger view difference causes worse performance. This is consistent with the result in Table 4.1-4.3. In another aspect, it is easy to see that the proposed VN-GAN significantly outperforms the SOTA methods in most cases.

Table 4.4 compares average accuracies of the three probe view against other

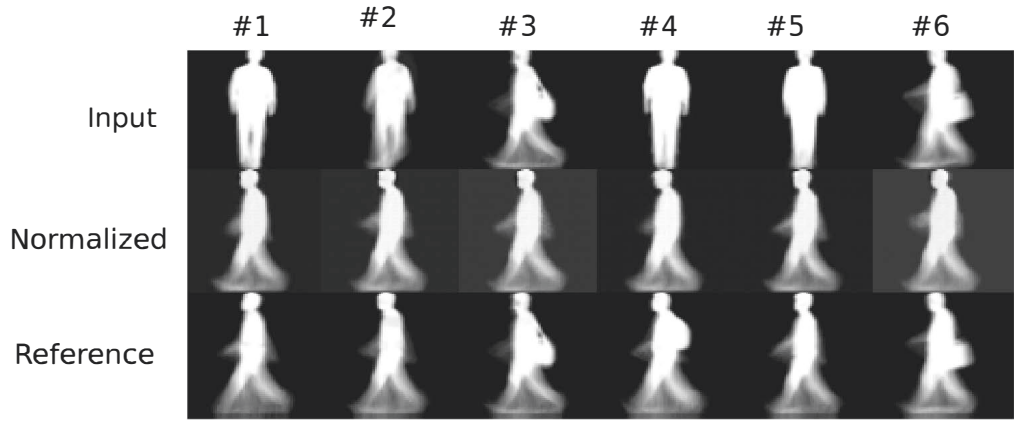


Figure 4.6 : An example of normalized gaits of six subjects. The three rows stand for input gaits, generated gaits and reference gaits, respectively. Each column represents one subject.

views in recent literature. It is easy to see that the proposed VN-GAN achieve promising performances. In specific, it improves 56% (from 19% to 75%) compared with baseline, and 12% (from 63% to 75%) compared with GaitGAN in terms of average accuracy of the three probe view. In addition, the proposed VT-GAN synthesize GEIs in reference view, which can provide visual interpretation to the result. As in Figure 4.6, the generated GEIs are in good visual quality and looks similar to the corresponding target in appearance.

Ablation Study

This section studies how the two-stage design affects the identification accuracies. In the proposed VN-GAN, Stage-I attempts to normalize GEIs from arbitrary views to the reference one, and Stage-II aims to implant identity information. Hence, the study is equivalent to evaluate the effect of variation normalization and identity preserving. To explore the effect of each stage, we train Stage-I alone (denoting as VT-GAN(s1)) and the whole VT-GAN framework, respectively. We report the results in Table 4.5. From the table, we can observe that each stage achieves promising

performance compared to SOTAs and contributes to the ultimate results together.

Table 4.5 : Average accuracies of 10 gallery views (The corresponding view is excluded.) on ProbeNM with 5 distinct probe views. 's1' demotes training with Stage-I of the proposed VN-GAN.

Method	Probe View					
	18°	54°	90°	126°	162°	Average
GaitGAN (Yu et al. 2017b)	0.54	0.65	0.58	0.66	0.54	0.59
MGANs (He et al. 2019)	0.66	0.77	0.67	0.79	0.72	0.72
VN-GAN(s1)	0.65	0.75	0.66	0.74	0.67	0.69
VN-GAN	0.69	0.77	0.71	0.77	0.71	0.73

4.2.5 Conclusions

This section presents a novel framework, i.e., VN-GAN, to address view angle diversity problem in person re-ID using gaits. Typically, the proposed VN-GAN is inspired by GAN theory and tries to normalize gaits under various conditions in a coarse-to-fine manner. In Stage-I, A GAN variant that includes a variation discriminator $\mathbf{D}_v$ and a variation classifier $\mathbf{P}$ is used to achieves view normalization, and in Stage-II, another GAN variant that consists of an identity discriminator $\mathbf{D}_i$ and an identity preserver $\mathbf{\Phi}$ is utilized to implant identity information. Moreover, we integrate the two-stage design to the Siamese structure and preserve identity using identity-preserving loss. It is worth to note that the proposed VN-GAN is also effective to other variations such as dressing changes and carrying variation. Comprehensive experiments on CASIA(B) dataset exhibit that the proposed VN-GAN benefits to identification performance. Moreover, VN-GAN generates visually

promising gaits in the reference view, which provides a visual interpretation of the experimental results.

4.3 View Transformation GAN

This section investigates VT-GAN to achieve view transformation of gaits across two arbitrary views using only one uniform model. In specific, we generated gaits in target view conditioned on input images from any views and the corresponding target view indicator. In addition to the classical discriminator in GAN, which makes the generated images look realistic, a view classifier is imposed. This controls the consistency of generated images and conditioned target view indicator and ensures to generate gaits in the specified target view. On the other hand, retaining identity information while performing view transformation is another challenge. To solve the issue, an identity distilling module with triplet loss is integrated, which constrains the generated images inheriting identity information from inputs and yields discriminative feature embeddings. The proposed VT-GAN generates visually promising gaits and achieves promising performances for CVGLT-reID, which exhibits great effectiveness of VT-GAN.

4.3.1 Motivations

To address CVGLT-reID, previous research experiences three stages which are model-based methods, mapping-based methods and deep learning-based methods. In the early stage, researchers mainly focus on body measurement based on human body model, e.g., extracting view-invariant features by measuring geometric parameters of human body (Bobick and Johnson 2001), or reconstructing 3D body using depth sensors and then re-projecting to 2D in arbitrary view (Zhao et al. 2006). However, the former is not accurate, and the latter is too complicated. In addition, the depth sensors are not practical in real scenarios due to their higher expenses. In

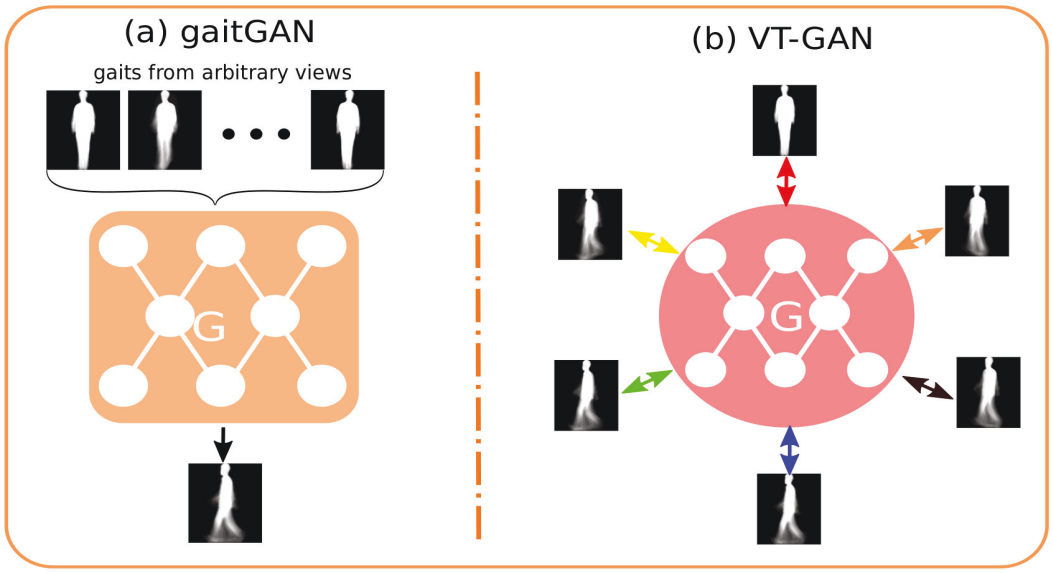


Figure 4.7 : Comparison between gaitGAN and the proposed VT-GAN. (a) The former normalizes GEIs from arbitrary views to a reference one, and (b) the latter directly transform gait images between any pair of views.

contrast, gait-based person re-ID using visual RGB sensors is more practical since the fact that surveillance cameras are deployed everywhere. Thus, approaches using 2D images are arising in the middle stage, which correlate gait templates across views by linear mapping such as regressing gait templates from one view to another view (Kusakunniran et al. 2009, 2010; Mansur et al. 2014) or projecting gait templates across views into a common feature space via asymmetric mapping (Ben et al. 2019b; Mansur et al. 2014; Xing et al. 2016; Ben et al. 2019a). These approaches are popular in past years because of their effectiveness and computational efficiency. However, this kind of approaches can only process gaits across views in pair, which will yield multiple models with the increase of views. In addition, these methods rely on view estimation, though it is not a problem nowadays. With the development of deep neural networks, it is a fashion to distil view-invariant features using CNN (Wu et al. 2017c; Shiraga et al. 2016). This kind of methods significantly improves

the accuracy of CVGLT-reID. However, CNN is a black-box, which is hard to interpret visually. In comparison, view transformation using encoder-decoder model (Yu et al. 2017c) or GAN (Yu et al. 2017b; He et al. 2019) as regressor exactly solves the interpretable problem, which transforms gait templates from various views to a canonical one only using a single model. However, it can only transform gaits to a specific one, which is hard to visualize view relationship between pairs from any two views. If we want to transform gaits across any two views, a lot of models are needed, i.e., the number of views. In contrast, this paper tries to visualize the view transformation relationship between any two views using only one model and thus achieve CVGLT-reID explicitly. The difference between current GAN-based view normalization and the proposed VT-GAN is shown in Figure 4.7.

Recently, image translation (Isola et al. 2017; Choi et al. 2018; Ma et al. 2017) using conditional GAN becomes one of the hottest topics in computer vision, which transforms images from the source domain to target domain conditioned on specific requirements such as style transfer, image inpainting, image editor, *etc.* For instance, some recent works try to generate persons in a specific pose, synthesize faces in conditioned expression, hairstyle or even gender. In specific, the popular starGAN (Choi et al. 2018) can simultaneously process several attributes using only a single model, which demonstrates great success in face editing. Inspired by the idea of starGAN, the paper proposes to perform view transformation between gaits from any two views name view transformation GAN (VT-GAN). As in Figure 4.8, the proposed VT-GAN architecture consists of a generator, a discriminator and a similarity preserver. The generator **G** takes the gait templates from source view and the corresponding reference view indicator as input, and generates images in the reference view. And, the discriminator **D** learns to not only distinguish if the input images are real but also classify the synthesised images to its corresponding view class. To preserve the identity information and make the synthesised gaits

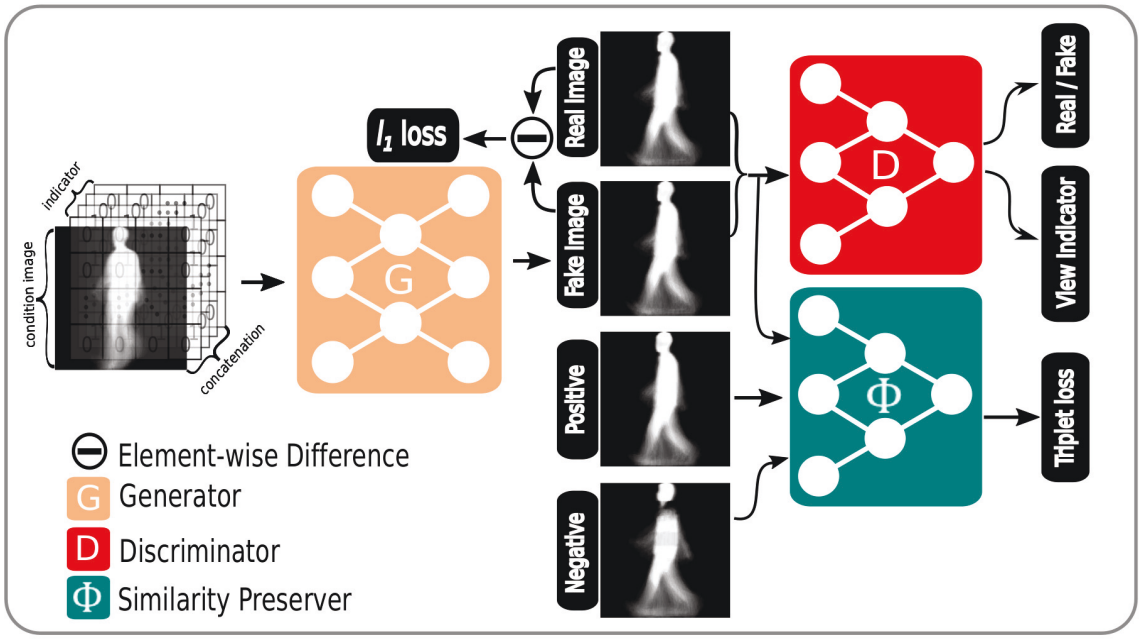


Figure 4.8 : The architecture of the proposed VT-GAN. It consists of three modules, i.e., generator **G**, discriminator **D** and similarity preserver **Φ**. Generator **G** inputs the concatenation of source/condition gait image from arbitrary view and target view indicator, and synthesizes gait image from target view. Discriminator **D** learns to distinguish between synthesised gait image and real gait image and classify real gait image to its corresponding view. Similarity preserver **Φ** learns to pull positive gait pairs together and push negative gait pairs away.

discriminative, we impose two regularization terms, i.e., identity preserving term and identity discriminative term. The former forces the generated gaits close to the target reference one and the latter pulls feature embedding of generated images extracted by a similarity preserver **Φ** close to their corresponding positive samples in target view and push the negatives far away with a margin. Through such a supervised training strategy, the synthesised gait image (on the target view) created by generator can be distinguished against the same/different subject under the same view (i.e. target view). In this way, we achieved identity-preserved view

transformation by a single model and depictedrelationships across any two views.

4.3.2 Proposed Framework

The proposed VT-GAN aims to transform gaits from one specific view to arbitrary views using a single model. The overall framework of the proposed VT-GAN is shown in Figure 4.8 which consists of a generator $\mathbf{G}$, a discriminator $\mathbf{D}$ and an identity preserver $\mathbf{\Phi}$. The generator $\mathbf{G}$ takes the gaits x_{src} from source view θ and the target view indicator c as input, and generate gaits in the target view ϑ, i.e., $\mathbf{G}(x_{src}, c) \mapsto x_{dst}$. In the proposed VT-GAN, we use a one-hot vector to denote the view indicator where the target view is assigned 1 and other views are assigned 0. Thus, x_{src} from the source view θ can be translated to x_{dst} from arbitrary view ϑ specified by indicator c. And, the discriminator $\mathbf{D}$ learns to not only distinguish if the input images are real but also classify the synthesised images to its corresponding domain. That is, the discriminator produces probability distributions for both real/fake discrimination and view classification (Choi et al. 2018), $\mathbf{D} : x \mapsto \{\mathbf{D}_{dis}(x), \mathbf{D}_{cls}(x)\}$, where x is the input to discriminator. To preserve the identity information and make the synthesised gaits discriminative, we impose two regularization terms, i.e., identity preserving term and identity discriminative term. The former forces the generated gaits close to the target reference one and the latter pulls feature embedding of generated images extracted by a similarity preserver $\mathbf{\Phi}$ (Deng et al. 2018) close to their corresponding positive samples in target view and push the negatives far away with a margin.

Adversarial Loss. We apply adversarial learning in our training process to constrain the output of $\mathbf{G}$ visually similar to the reference gait x_{dst} from view ϑ. The adversarial loss is defined as

$$\begin{aligned}\mathcal{L}_{adv}(\mathbf{G}, \mathbf{D}_{dis}, p_\theta, p_\vartheta, p_c) = \mathbb{E}_{x_{dst}\sim p_\vartheta(x_{dst})}[\log \mathbf{D}_{dis}(x_{dst})] \\ +\mathbb{E}_{x_{src}\sim p_\theta(x_{src}), c\sim p_c(c)}[\log(1 - \mathbf{D}_{dis}(\mathbf{G}(x_{src}, c)))],\end{aligned} \tag{4.6}$$

where p_θ and p_ϑ are sample distributions in source view θ and target view ϑ, respectively. p_c denotes the distribution of view indicator. $\mathbf{G}$ tries to generate gaits $\mathbf{G}(x_{src}, c)$ in target view ϑ conditioned on gaits x_{src} from source view θ and the corresponding view indicator c, while $\mathbf{D}_{dis}$ aims to distinguish between generated gait image $\mathbf{G}(x_{src}, c)$ and real gait image x_{dst}. $\mathbf{G}$ tries to minimize the objective while $\mathbf{D}_{dis}$ competes against it which maximizes the objective, i.e., $\min_{\mathbf{G}} \max_{\mathbf{D}_{dis}} \mathcal{L}_{adv}(\mathbf{G}, \mathbf{D}_{dis}, p_\theta, p_\vartheta, p_c)$.

It is worth noting that the generated gait image cannot be mapped back to the source one. This is because the person may carry a bag or wear a coat whose style is unpredictable in source view and our aim is to synthesize gait image in target view as well as normal walking condition.

View Classification Loss. For a given gait image x_{src}, one of our aims is to translate it to another view specified by view indicator c. To fulfil the condition, an auxiliary view classifier is added on the top of $\mathbf{D}$ as in (Choi et al. 2018) and incurs a view classification loss when iteratively optimizing $\mathbf{G}$ and $\mathbf{D}$. In specific, $\mathbf{D}$ learns to classify the real GEI x_{dst} to its corresponding view while $\mathbf{G}$ tries to synthesize gait image $\mathbf{G}(x_{src}, c)$ that can be classified to the view specified by indicator c. To optimize $\mathbf{D}$, we minimize the following loss function,

$$\mathcal{L}^D_{cls}(\mathbf{D}_{cls}, p_\vartheta, p_c) = \mathbb{E}_{x_{dst}\sim p_\vartheta(x_{dst}), c\sim p_c(c)}[-\log \mathbf{D}_{cls}(c|x_{dst})] \tag{4.7}$$

where $\mathbf{D}_{cls}(c|x_{dst})$ is the probability distribution of input gait image over view indicator computed by $\mathbf{D}$. It is worth noting that $\mathbf{D_{cls}}$ is learned from real gaits in the target view. When optimizing $\mathbf{G}$, generated gait image and its corresponding view indicator are used. The loss function is defined as

$$\mathcal{L}^G_{cls}(\mathbf{G}, p_\theta, p_c) = \mathbb{E}_{x_{src}\sim p_\theta(x_{src}), c\sim p_c(c)}[-\log \mathbf{D}_{cls}(c|\mathbf{G}(x_{src}, c))]. \tag{4.8}$$

By minimizing the loss function, $\mathbf{G}$ attempts to synthesize gait image that can be classified to the view specified by c.

Reconstruction Loss. To make the generated gait image similar to the real target at low frequencies, we additionally add a reconstruction loss, i.e., we use an ℓ_1 loss to minimize the reconstruction error between the generated gait image and real target gait image which can be denoted as

$$\mathcal{L}_{id}(\mathbf{G}, p_\theta, p_c) = \mathbb{E}_{x_{src}\sim p_\theta(x_{src}), c\sim p_c(c)}[\|x_{dst} - \mathbf{G}(x_{src}, c)\|_1]. \tag{4.9}$$

It has been proved by previous works that ℓ_1 distance (Isola et al. 2017; Ma et al. 2017; Deng et al. 2018) helps regularise the adversarial training process and preserve appearance consistency.

Identity-preserving Loss. Since our ultimate goal is to achieve CVGLT-reID, identity preserving becomes essential while translating gait image in source view to target views. As analysed before, we not only bridge view gaps between gaits in source and target view but also make them identity-distinguishable while performing gait view translation. Thus, exploring the potential and underlying identity-related information rather than only image style is significant. To achieve this, we integrate a three-stream network $\mathbf{\Phi}$ (also be called identity preserver) to regularize the generation process as shown in Figure 4.8. That is, Φ learns to distinguish real gait images in target view by identity while $\mathbf{G}$ tries to generate gait images from source view that satisfy the classification criterion. Thus, we decompose the loss function into two parts: a similarity preserving loss of real gait images for $\mathbf{\Phi}$ optimization and another similarity preserving loss of fake gait images for $\mathbf{G}$ optimization. We use triplet loss (Schroff et al. 2015) to train the identity preserver module, i.e.,

$$\mathcal{L}_{tri}(x_{anc}, x_{pos}, x_{neg}) = \max\{d(x_{anc}, x_{pos}) - d(x_{anc}, x_{neg}) + \rho, 0\} \tag{4.10}$$

where x_{anc}, x_{pos} and x_{neg} are three normalized embeddings which denotes anchor, positive and negative sample, respectively. In specific, the embeddings are extracted by $\mathbf{\Phi}$. $d(\cdot, \cdot)$ denotes the Euclidean distance between two input elements, and $\rho \geq 0$ represents the margin between classes in the embedding space. We set $\rho = 0.8$

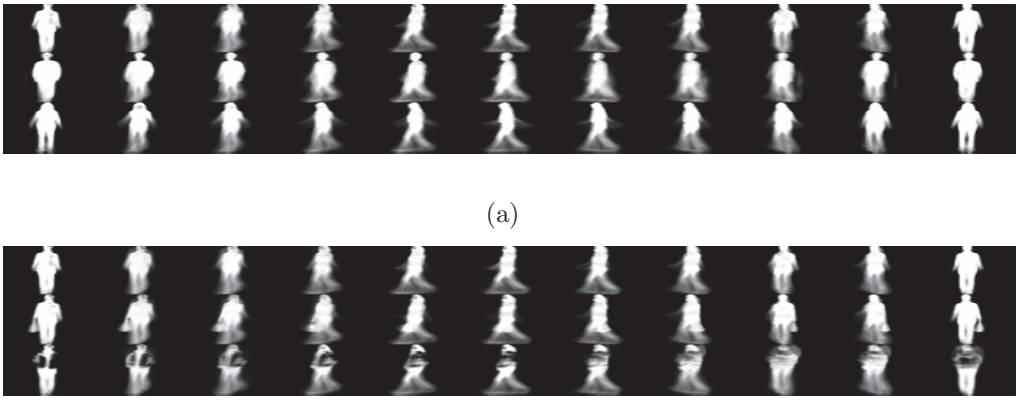
empirically in our experiment. By minimizing the loss, $d(x_{anc}, x_{pos})$ tends to be 0 and $d(x_{anc}, x_{neg})$ tends to be greater than $d(x_{anc}, d_{pos})$ with a margin ρ. When the condition is fulfilled, the loss becomes 0 and no gradient would be back-propagated. When optimizing $\mathbf{\Phi}$, embeddings of real GEIs in target view are utilized, the loss function is denoted as $\mathcal{L}_{tri}^{\Phi}$. On the other hand, the anchor x_{anc} is replaced to embedding of translated GEI $\mathbf{G}(x_{src}, c)$ when optimizing $\mathbf{G}$. The loss function loss function is thus denoted as $\mathcal{L}_{tri}^{G}$. The main difference between $\mathcal{L}_{tri}^{\Phi}$ and $\mathcal{L}_{tri}^{G}$ are where the anchor embedding is from, i.e., real gait images or translated gait images.

Overall Objective. As analysed, our aim is to transform gait image from one specific view to arbitrary views and meanwhile preserve the identity information. To achieve the aim, the above losses work cooperatively, and the overall objective function is

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cls}^{D} + \lambda_2 \mathcal{L}_{id} + \lambda_3 \mathcal{L}_{tri}^{G} \tag{4.11}$$

where $\lambda_t, t \in \{1, 2, 3\}$ are trade-off hyper-parameters which balance contribution of four losses. In practice, we set $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = 10$ in our experiments.

Training Strategies. As in Figure 4.8, the proposed VT-GAN includes three modules, i.e., identity preserver $\mathbf{\Phi}$, discriminator $\mathbf{D}$ and generator $\mathbf{G}$. In the training phase, we alternatively optimize each component. This is, we update parameters of one module while keeping parameters of the other two modules fixed. The discriminator learns to distinguish whether the input gait image is generated and classify it to its corresponding view. The similarity preserver learns to separate the input gait image by its identity. And, the generator tries to generate gaits to satisfy the requirements of both discriminator and identity preserver. We do not stop the training procedure until the objective is converged or maximum iterations are reached.



(a)

(b)

Figure 4.9 : An example of GEIs from 11 views. Each row includes GEIs of the same subject from 0° to 180° with 18° interval. (a) shows GEIs of three different subjects in terms of three rows. (b) exhibits GEIs in three different conditions of the same subject, i.e., NN, BG and CL.

4.3.3 Experiments

In this section, we describe details about the experimental setup, implementation settings and evaluate the proposed VT-GAN on CVGLT-reID task.

Experiment Settings

Dataset. The CASIA(B) gait dataset (Yu et al. 2006) is adopted for evaluation of view transformation and CVGLT-reID in our experiment. Figure 4.9 lists some examples of GEIs from 11 views in which (a) shows GEIs from different subjects in NM condition and (b) shows GEIs in different conditions from the same subject. Following Sec. 4.2.4, this section utilizes the same dataset setup and evaluation method for experiment analysis.

Implementation Details.

Network Architecture. There are three modules in the proposed VT-GAN, generator **G**, discriminator **D** and identity preserver **Φ**. Adapted from (Choi et al. 2018), **G** includes two Conv-InstanceNorm-ReLU blocks for down-sampling, three residual blocks (He et al. 2016), and two TransposeConv-InstanceNorm-ReLu blocks for up-sampling. **D** follows the theory of PtachGAN (Isola et al. 2017) to discriminate real/fake on local patches which is benefit to sharp the synthesised image. It consists of four Conv-LeakyReLu blocks and another two convolution layers with the stride of in parallel, where one is for real/fake discrimination and the other is for view classification. **Φ** is adapted from (Yu et al. 2017b) which includes three convolution layers together with max-pooling with 2×2 kernels and a stride of 2. In addition, 4×4 filters with stride 2 are involved in all convolution and transposed convolution layers above except residual blocks.

Training Details. The model is trained using Adam (Kingma and Ba 2014) with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and an initial learning rate of 0.0001. We train the model for 20k iterations which the learning rate keeps fixed during the first 10k iterations and then linearly decay to 0. Considering the GPU resources, the batch size is set to 100. In our experiments, we empirically set trade-off parameters in objective Eq. (4.11) as $\lambda_1 = 1$, $\lambda_2 = 10$ and $\lambda_3 = 10$. When training the model, we alternatively optimise generator **G**, discriminator **D** and identity preserver **Φ**. In specific, we perform a single optimisation of generator **G** after every five optimisations of discriminator **D** and **Φ**.

Quantitative Analysis

In this section, we evaluate the proposed VT-GAN on CVGLT-reID and analyse the effect of identity preserving loss to identification accuracy.

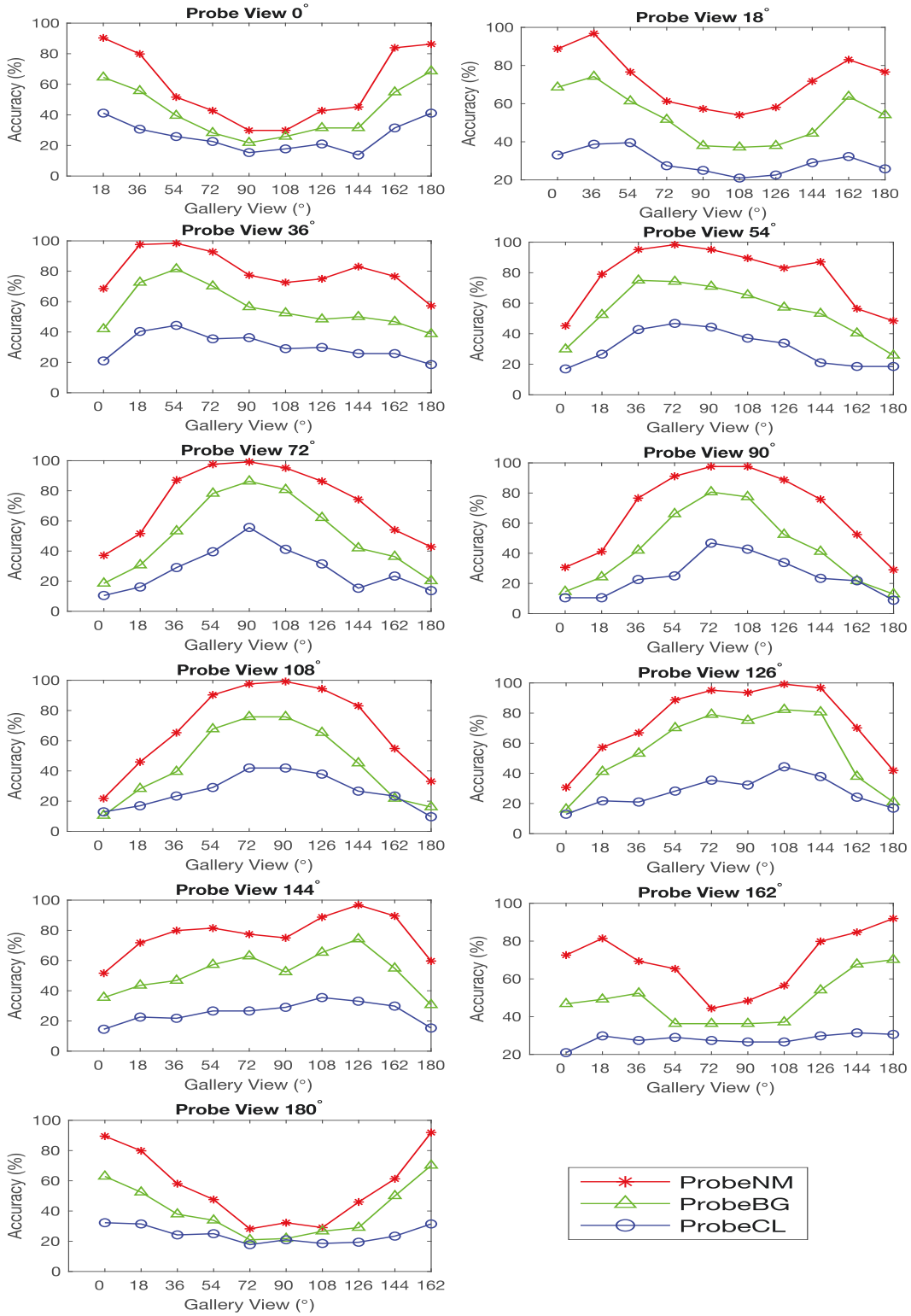


Figure 4.10 : Comparison of accuracies on the three probe subsets at the 11 probe views. For each probe view, accuracies of the rest views are excluded.

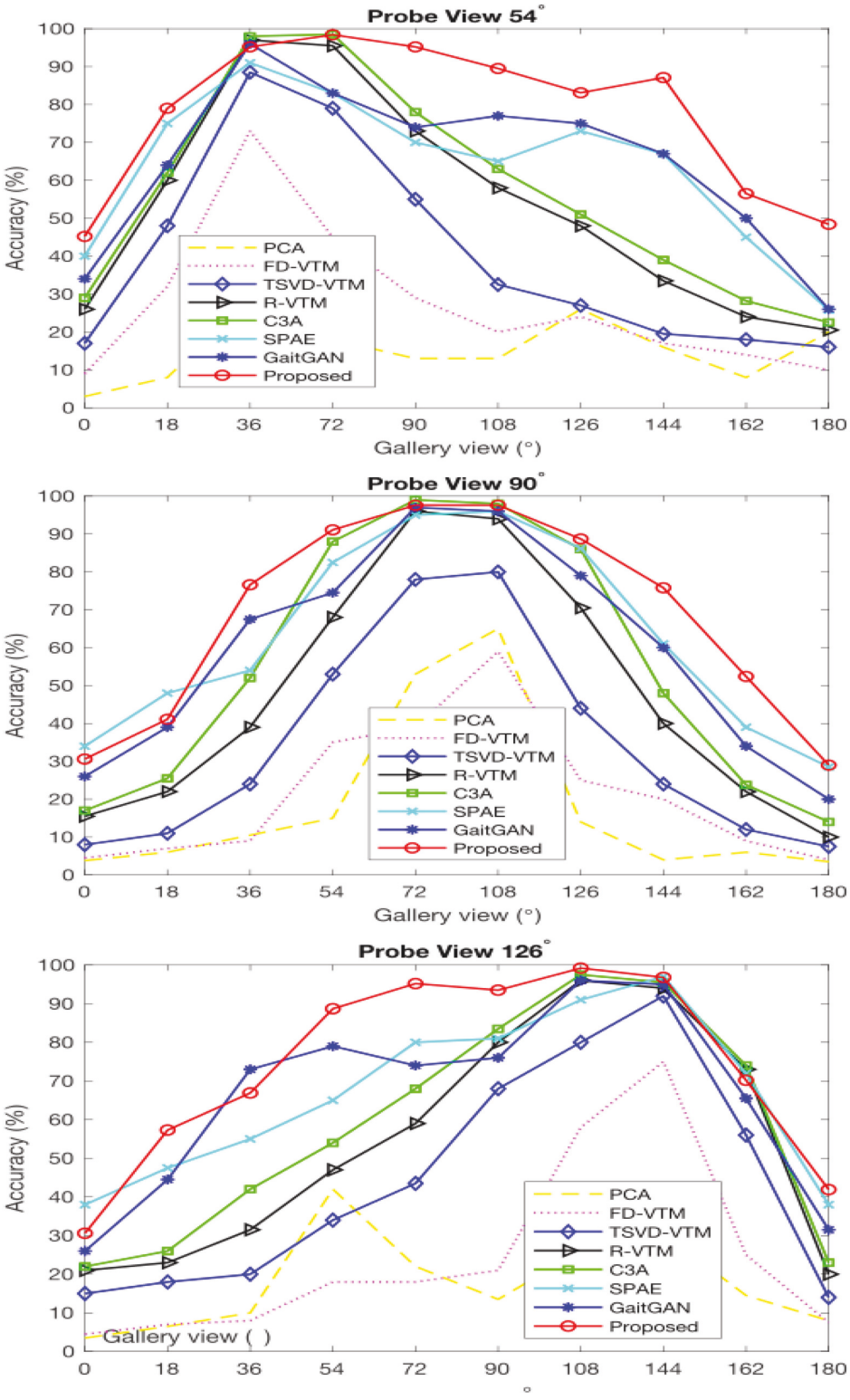


Figure 4.11 : Comparison with previous methods at three representative probe views on ProbeNM.

Effect of View Changes. In order to analyse how the view variation affects the performance of person re-ID using gaits, we evaluate the proposed VT-GAN on three challenging probe sets, i.e., ProbeNM, ProbeBG and ProbeCL, and report the result between any two views. Fig. 4.10 shows results on each probe view versus all views except the identical one. From the figure, it is easy to observe that view variation significantly affects the performance of person re-ID using gaits. In specific, 1) view difference degrades the identification accuracy, and larger one causes worse performance, 2) performance would improve near the symmetric view in terms of centre view 90° when the probe view is fixed, 3) carrying a bag or wearing a coat also lower the identification accuracy and wearing a coat results in worst performances.

It is reasonable for the above conclusions. This is because 1) view difference incurs heterogeneous data distribution and unaligned information structure, 2) gaits from symmetric views can be mutually mirror-reflected, and 3) carrying a bag or wearing a coat severely change silhouette shape of human body and occlude the walking traits. Though the proposed VT-GAN weakens the effect of these factors, they still degrade the identification accuracy.

Performance Comparison. As in (Yu et al. 2017b), we select three probe view 54°, 90° and 126° that are identical for CVGLT-reID to compare performances with SOTA methods. In this section, we compared some representative approaches such as PCA (Han and Bhanu 2005), FD-VTM (Makihara et al. 2006), TSVD-VTM (Kusakunniran et al. 2009), R-VTM (Kusakunniran et al. 2012), C3A (Xing et al. 2016), SPAE (Yu et al. 2017c), ViDP (Hu et al. 2013), GaitGAN (Yu et al. 2017b), gaitGANv2 (Yu et al. 2019) and MGAN (He et al. 2019). Figure 4.11 compares aCVGLT-reID performances between each probe view and other rest views. From the figure, it is clear that the results support conclusions in *Effect of View Changes*. Moreover, we can observe the proposed VT-GAN improves the matching performance to some extent. Table 4.6 reports the average identification accuracies of the

Table 4.6 : Comparison with recent works on CVGLT-reID on ProbeNM. Average identification accuracies except the corresponding view are reported.

Method	Probe View			
	54°	90°	126°	Average
ViDP Hu et al. (2013)	0.64	0.60	0.65	0.63
SPAE Yu et al. (2017c)	0.63	0.62	0.66	0.64
GaitGAN Yu et al. (2017b)	0.65	0.58	0.66	0.63
MGAN He et al. (2019)	**0.77**	0.67	**0.79**	**0.74**
GaitGANv2 Yu et al. (2019)	0.72	0.65	0.73	0.70
Proposed	**0.77**	**0.68**	0.76	**0.74**

three probe views against rest views. From the table, it is easy to see that the proposed VT-GAN outperforms other recent GAN-based works, i.e., we improve 12%, 10% and 10% in terms of the three probe views compared with GaitGAN. Along with Figure 4.11, the proposed VT-GAN achieves promising performances.

Effect of Identity-preserving Loss. In this part, we trained the vanilla VT-GAN and pruned VT-GAN, respectively. Compared vanilla VT-GAN, the pruned VT-GAN does not include the identity preserver $\mathbf{\Phi}$ which is denoted as VT-GAN (*w/o* $\mathbf{\Phi}$). Table 4.7 reports results of the two models. It is easy to observe that the identity preserver does benefit to CVGLT-reID. In addition, the VT-GAN can achieve equivalent performances even without this module, because the reconstruction loss keeps appearance similarity while performing view translation.

Qualitative Analysis

As analysed, one benefit of the proposed VT-GAN is that it provides visual interpretation for CVGLT-reID. As in Figure 4.12, we visualise synthesised gaits

Table 4.7 : Evaluation of effectiveness of the identity preserver in VT-GAN.

Method	Probe View			
	54°	90°	126°	Average
GaitGAN Yu et al. (2017b)	0.65	0.58	0.66	0.63
VT-GAN (*w/o* $\mathbf{\Phi}$)	0.75	0.64	0.73	0.71
VT-GAN	0.77	0.68	0.76	0.74

between any two views. From the figure, we can see that 1) the VT-GAN synthesises visually promising images, which is hard to distinguish from real images, 2) The VT-GAN successfully achieves view transformation between any two views, 2) The generated images are appearance similar to the target gaits in gallery set. Figure 4.13 shows the synthesised gaits in 90° from various views of three different subjects in normal walking condition. It is easy to observe that smaller view difference results in more similar artifact, i.e., it is easy to generate appearance similar results when the view of probe gaits is near 90° in the figure. Figure 4.14 shows the synthesised gaits 90° from various views of the same subject different walking conditions. It is obvious that carrying a bag or wearing a coat cause poorer synthesised gaits. In Figure 4.15, we also list some bad samples caused by poor segmentation or limited silhouettes and their corresponding synthesised artifacts by VT-GAN. These samples undoubtedly degrade identification accuracy. However, the proposed VT-GAN also tries to translate them and makes the generated gaits look normally. This hints that the VT-GAN can correct minor GEI faults caused by poor segmentation and incomplete walking cycle. We suggest that this is consistent with the cases such as carrying a bag or wearing a coat.

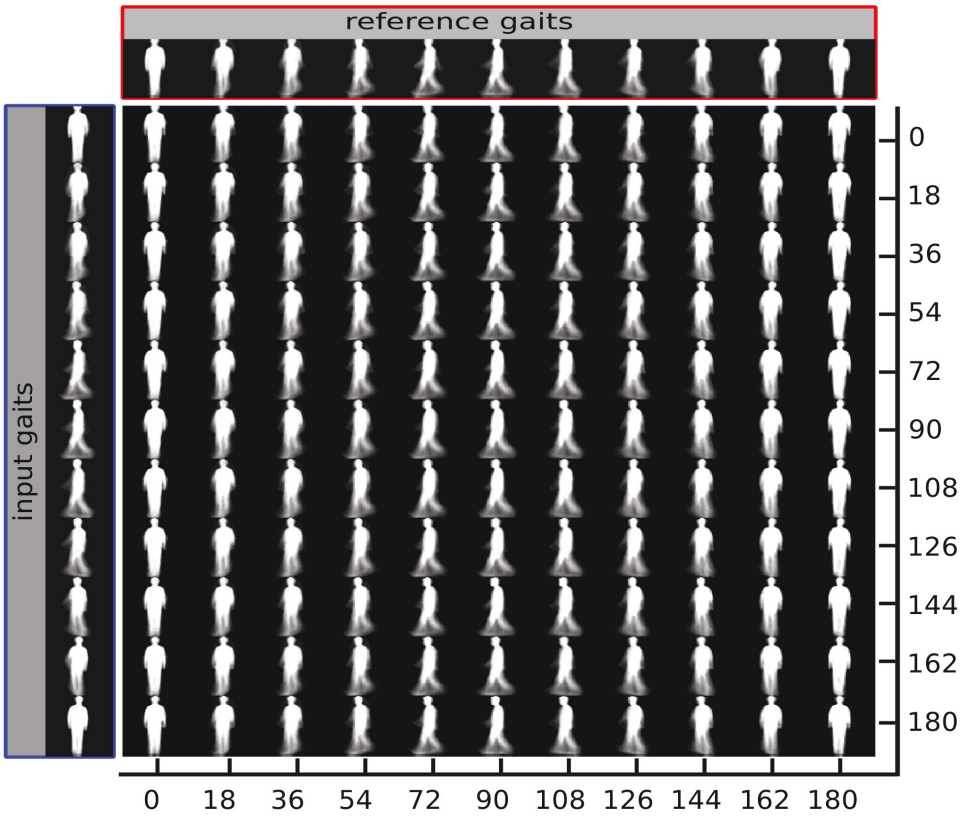


Figure 4.12 : Visualization of synthesised gaits between any two views. Images in blue box are input gaits, and the ones in red box are reference gaits. The rest 11×11 images are synthesised gaits conditioned on the corresponding input image and view indicator.

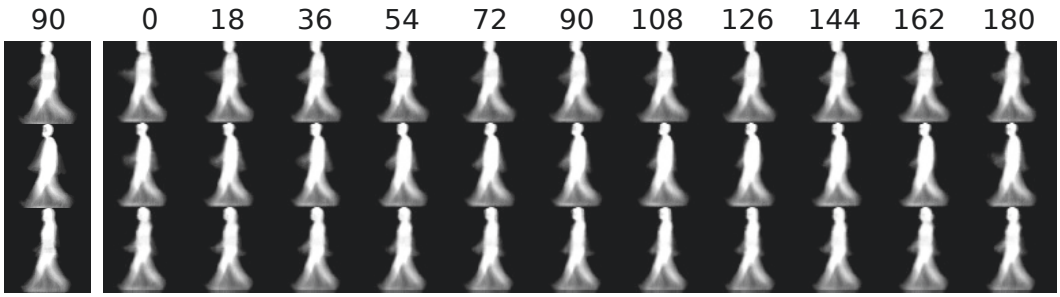


Figure 4.13 : Visualization of synthesised gaits of three different subject. Each row includes gaits from the same subject. And, the first column is reference gaits in 90°, the rest rows are synthesised gaits from input gaits in 0° to 180°.

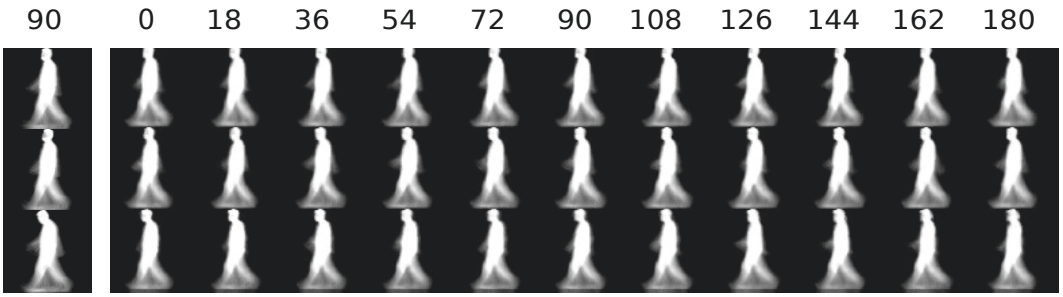


Figure 4.14 : Visualization of synthesised gaits of the same subject in three distinct conditions, i.e., normal walking, carrying a bag and wearing a coat corresponding to three rows, respectively. The first column includes reference gaits of the same subject in 90° and the rest columns consist of synthesised gaits from various views.

Figure 4.15 : An example of some bad samples and their corresponding translated gaits. For each pair, the left lists input gaits and the right shows synthesised gaits.

Further Analysis

Relation to GaitGAN. As in Figure 4.7, GaitGAN aims to normalize gaits in various views to a reference one, i.e., 90°. Thus, only gaits from 90° are used to train the discriminator. If we want to map gaits in an identical view to arbitrary views, multiple models should be built. Different from GaitGAN, the proposed VT-GAN could translate gaits to arbitrary view using only one single model benefiting from the view indicator input. This means that view indicator constrains the generation process and force the generator to synthesize gaits in view corresponding to the specific indicator. Considering this, GaitGAN is a special case of the proposed VT-GAN, i.e., we set the view indicator to 90°. In another aspect, the proposed VT-GAN adopts distinct framework with GaitGAN because of their different goal.

GaitGAN is adapted from classical GAN model while the proposed VT-GAN is derived from conditional GAN, which takes view indicator as the condition. This brings out different discriminator design, i.e., GaitGAN adopts traditional real/fake discriminator while the proposed VT-GAN adapts the discriminator and makes it not only distinguish the artifacts but also discriminate view of input gaits.

We also studied the influences of reference view, i.e., we changed the view indicator and translated gaits of arbitrary views to each reference view. Figure 4.11 compares average accuracies of all the 121 view pairs with respect to different reference views. From the figure, we can observe that side view 90° adopted in GaitGAN does not achieve the highest performance. In contrast, two peak identification accuracies appear at 54° and 126°. The result verifies the view symmetry theory in Sec. 4.3.3 because 54° and 126° are the centrers of two half sphere separated by 90°.

Potential Benefits of VT-GAN. In addition to the above benefits, the proposed VT-GAN brings other applications such as data augmentation and unseen view prediction. As we know, lack of data restricts the performance of CNN models. In person re-ID using gaits, most of the existing datasets only contain limited samples with minor view variation. It is insufficient to train large-scale models for complicated scenarios. The proposed VT-GAN provides a solution to the problem because it can synthesize gaits of arbitrary views from one specific view. Moreover, it can synthesize gaits of arbitrary views from other existing datasets which only contain minor view variation. This significantly enlarges the volume of data. In another aspect, the proposed VT-GAN can synthesize gaits of one specific subject in unseen views. It is beneficial to predict the subject's identity appearing in the view for the first time.

4.3.4 Conclusions

This section studies view transformation problem in long-term person re-ID using gaits and proposes to achieve view-to-view transformation via a single model, i.e., VT-GAN. Instead of normalizing to one unified view as previous GAN-based works, VT-GAN achieves to translate gaits between any two views only using a single model. In specific, VT-GAN includes three modules, a generator, a discriminator and a similarity preserver. The generator attempts to generate gaits in reference view from source view conditioned on a view indicator and fool the real/fake discriminator while the discriminator tries to distinguish whether its input is artificial and meanwhile constrain it to its corresponding view. The similarity preserver supervises the generation process, which forces to synthesize gaits sharing the same identity with its positive sample in the reference view. Both qualitative and quantitative analysis are conducted on CVGLT-reID task, the experiment results show promising performances of the proposed VT-GAN. Furthermore, the proposed VT-GAN brings other usages such as data augmentation (Huang et al. 2019) and unseen view prediction, which will be researched in the future study.

4.4 Summary

This chapter investigates the influence of view changes for long-term person identification using gaits. In particular, two kinds of models using GAN are raised, which one aims to normalize gaits of various views into an identical one and the other one tends to achieve view transformation across any two arbitrary views using a unified model. Both models demonstrate promising identification performance on benchmark CASIA(B) dataset and show the potential of such interpretable models. Moreover, this chapter predicts future applications of such methods in order to further boost the performance of person identification, for instance, data augmentation and unseen view prediction.

Chapter 5

Top-push Constrained Modality-adaptive Dictionary Learning for Cross-modality Person Re-ID

This chapter studies the CCM-reID problem that TSIs are captured by different types of cameras in long-term person re-ID. In contemporary surveillance, cameras of different modalities such as NIR cameras and depth sensors are adopted in broad scenarios in order to ease the effect of light variation between day and night. However, re-identifying the persons across such cameras of different modalities brings severe modality/domain gap. This chapter focuses on the challenging problem and proposes a top-push constrained modality-adaptive dictionary learning (TCMDL) model. The proposed TCMDL asymmetrically projects the heterogeneous features from dissimilar modalities onto a common space. In this way, the modality-specific bias is mitigated. And, the heterogeneous data thus can be reconstructed by a shared dictionary simultaneously in the canonical space. In addition, a top-push ranking graph regularisation is proposed to constrain the model, which improves the discriminability and boosts the matching accuracy. Extensive experiments demonstrate the superior performance of the proposed TCMDL on CCM-reID problem.

5.1 Introduction

5.1.1 Problem Formulation

Most existing approaches for person re-ID depend on the fundamental assumption that person images are collected in the daytime by RGB cameras since they are cheap and informative (Xiong et al. 2014; Liao et al. 2015; You et al. 2016;

Cheng et al. 2017). However, people are hard to be captured by RGB cameras in poor lighting scenarios, e.g., night or cloudy. In these cases, from the perspective of applications, alternative sensors whose image-forming principle is invariant to visible light are necessary such as near-infrared (NIR) camera (Wu et al. 2017b) and RGB-D camera (Barbosa et al. 2012; Wu et al. 2017a). These different sensors take their advantages to sense images in both the day time (good lighting) and the night time (poor lighting). However, this case raises another problem of how we can identify the persons with such images taken by different types of sensors, i.e., CCM-reID problem. Due to the differences of image-forming principle between different types of sensors, people's appearance between RGB and NIR/RGB-D cameras are heterogeneous. This violates the prior assumption that images from different cameras follow the same distribution and causes serious data bias between image across camera modalities as shown in Figure 5.1. The data bias enlarges intra-person discrepancy and causes person mismatch. This creates another layer of difficulty that requires to simultaneously mitigate data bias and explore identity discriminability.

5.1.2 Motivation

To address the CCM-reID problem, this chapter tries to alleviate the data biases between modalities and improves the representative power of feature via asymmetric feature learning and discriminative dictionary learning. Our motivations are three-fold. Firstly, the coupled metric learning (Li et al. 2009; Ben et al. 2019a; Yu et al. 2017a) is able to bridge gaps between heterogeneous data. It motivates us to learn a pair of mapping matrices to project original feature representation from different modalities into a shared subspace. The learned two mapping matrices are asymmetric, which contains different values and projects features of distinct modalities into a modality-agnostic subspace, respectively. Since samples from each modality are transformed to the shared subspace by an independent matrix, i.e., asymmetrically,

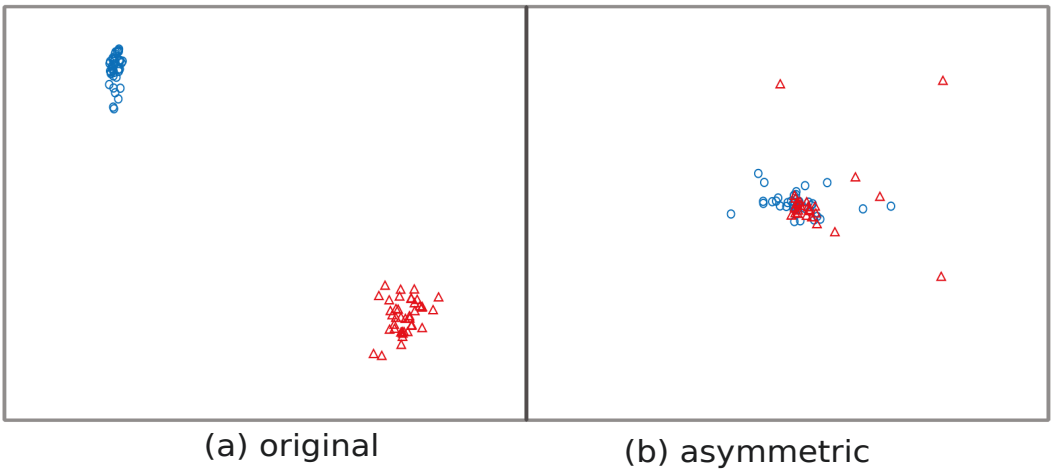


Figure 5.1 : Illustration of how our asymmetric mapping briataaadges the data gaps. We performed PCA on samples from BIWI RGBD-ID dataset Munaro et al. (2014b) for visualization. Each shape (circle or triangle) represents samples from one modality. (a) original data distribution, (b) distribution in the shared space learned by asymmetric mapping.

data heterogeneity across modalities is thus mitigated in the subspace as shown in Figure 5.1. This is essential for distance measurement across modalities. Secondly, the success of discriminative dictionary learning (Cheng et al. 2017) inspires us to impose Laplacian-like graph regularisation to perform retrieval task with a ranking formulation. Finally, triplet constraint (Wang et al. 2016) is usually utilised in the classical ranking scheme. However, inter-person feature differences are more ambiguous in the cross-modality setting. Thus more stringent regularisation is necessary. Inspired by top-push ranking (You et al. 2016) which forces intra-class difference to be smaller than the minimum inter-class difference, we reformulate the top-push ranking to a Laplacian-like graph and integrate it to a unified objective. Benefiting of the above three aspects, the unified objective can simultaneously mitigate the data biases across modalities and keep the powerful feature representation ability of dictionary as well as the discriminative ability of top-push constraint.

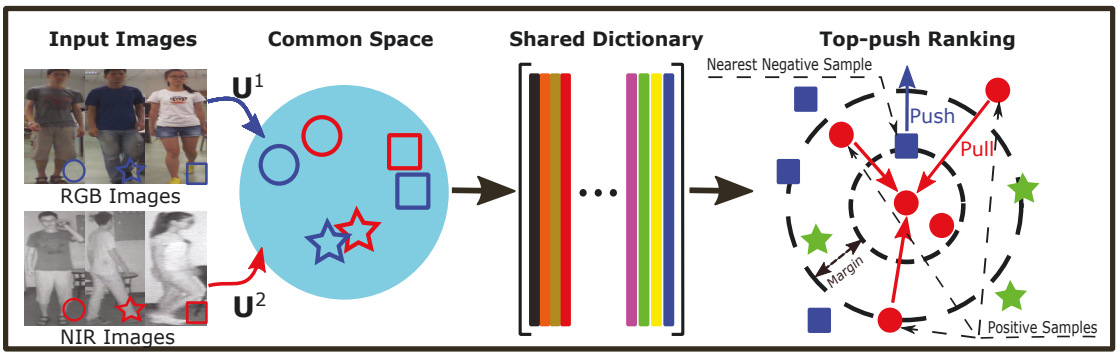


Figure 5.2 : Illustration of the proposed TCMDL model. Objects with the same shape (hallow/solid) represent the same person. Input images from cross modalities are mapped into a subspace in which a shared dictionary is learned. Meanwhile, the encoding coefficients are regularized by a top-push ranking constraint embedded Laplacian-like graph.

Based on the above motivations, this chapter proposes a Top-push Constrained Modality-adaptive Dictionary Learning (TCMDL) model for CCM-reID as illustrated in Figure 5.2. This model simultaneously learns the latent subspace and discriminative dictionary for cross-modality retrieval problem. In detail, our model consists of four parts. One is asymmetric feature and dictionary learning, which jointly map the heterogeneous data into a common subspace and learn a shared dictionary for the heterogeneous data. Since data biases are alleviated by asymmetric feature learning, the projected features can be represented by a shared dictionary. Different from cross-view dictionary learning that learns two distinct dictionaries, we represent the same person across modalities with a shared dictionary in the learned common subspace. We also add three regularisation terms that aim to avoid information loss while performing feature mapping, keep consistent information of the same person across modalities and preserve discriminative ability, respectively. Especially, we reformulate top-push distance learning model into a Laplacian-like graph and impose it to the coding coefficients through dictionary learning. It is

critical for CCM-reID to differentiate minor variations. As far as we know, this is the first work to integrate asymmetric feature mapping and discriminative dictionary learning into a uniform framework and achieve these two purposes simultaneously to solve the CCM-reID problem.

5.2 Top-push Constrained Modality-Adaptative Dictionary Learning

This section formulates the proposed TCMDL model for CCM-reID, which involves four parts: joint asymmetric mapping and dictionary learning, energy-preserving regularisation, cross-view consistency regularisation and top-push constrained Laplacian graph regularisation. It learns a shared discriminative dictionary in a common subspace by joint asymmetric feature mapping and top-push constraint regularized dictionary learning.

5.2.1 Framework Overview

The aim of CCM-reID is to retrieve the person of interest from volumes of gallery images captured by a series of disjoint cameras in the surveillance network. Currently, most algorithms are developed for the two-camera single-modality setting, in which both cameras are based on visible light. In contrast, we consider the CCM-reID and develop the TCMDL model in a general multi-modality way not only for the cross-modality but also for multi-modality scenarios.

Figure 5.2 gives the pipeline of the proposed TCMDL for CCM-reID. In the figure, two modalities, images captured by RGB cameras and NIR sensors are taken as an example. Here, different modalities refer to different styles of person images taken by different sensors such as RGB cameras, NIR sensors and depth sensors. In Figure 5.2, features extracted from RGB images and NIR images are simultaneously inputted to the model and mapped into a common subspace by a pair of asymmet-

ric mapping matrices in which a shared dictionary is learned to reconstruct features from both modalities. Due to data gaps are mitigated when performing asymmetric mapping as shown in Figure 5.1, the features from two domains (modalities) can be represented using the shared dictionary in the common subspace. To keep discriminability, we impose a top-ranking regularisation on the encoding coefficients with respect to features from each domain. This regularisation term ensures that the distance between positive samples is smaller than any pair of their corresponding negative sample.

5.2.2 Objectives

Without loss of generality, we denote training sets collected from P modalities across disjoint cameras as $\mathbf{X}^p = [\mathbf{x}_1^p, \mathbf{x}_2^p, \ldots, \mathbf{x}_{N_p}^p] \in \mathbf{R}^{n_p \times N_p} (p = 1, \ldots, P)$ respectively, where each $\{\mathbf{x}_i^p; l_i^p\}$ $(i = 1, \ldots, N_p)$ corresponds to a n_p-dimensional feature of the i-th image from the p-th camera and l_i^p is its class label.

Joint Asymmetric Mapping and Dictionary Learning (AsyDic)

Due to large data gaps between P cameras, we wish to learn a set of mapping matrices $\mathbf{U}^p \in \mathbf{R}^{n \times n_p}, n \leq min(n_1, n_2, \ldots, n_p)$ to project the heterogeneous features in terms of each sensor modality into a common low-dimensional subspace in which dictionary learning can be performed by learning a shared K-atom dictionary $\mathbf{D} \in \mathbf{R}^{n \times K}$. Thus, the dictionary learning in the learned subspace is

$$\begin{aligned} \mathcal{J}_1(\mathbf{U}^p, \mathbf{D}, \mathbf{A}^p) &= \sum_{p=1}^{P} (\|\mathbf{U}^p \phi(\mathbf{X}^p) - \mathbf{D}\mathbf{A}^p\|_F^2 + \alpha \|\mathbf{A}^p\|_F^2) \\ \text{s.t.} \quad & \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \ldots, K \end{aligned} \tag{5.1}$$

where $\phi(\cdot)$ denotes the feature representation function, which can be hand-crafted feature or data-driven feature learned by deep neural networks, $\mathbf{A}^p \in \mathbf{R}^{K \times N_p}$ is the encoding coefficients spanned over the shared dictionary $\mathbf{D}$, $\mathbf{d}_i$ is the i-th column of $\mathbf{D}$, $\|\cdot\|_F$ denotes Frobenious norm of a matrix and α is the trade-off parameter.

It is worth noting that we regularize the coefficient $\mathbf{a}_i^p \in \mathbf{R}^K$ corresponding to each sample $\mathbf{x}_i^p$ with l_2-norm rather than l_1-norm. This is because less sparsity helps identification and improves computation efficiency as described in (Peng et al. 2017).

Energy-preserving regularisation

To avoid information loss of original signals, it is fashionable to impose an energy-preserving regularisation (Shekhar et al. 2013; Lu et al. 2017), defined as

$$\mathcal{J}_2(\mathbf{U}^p) = \sum_{p=1}^{P} \|\mathbf{U}^{p\top}\mathbf{U}^p\phi(\mathbf{X}^p) - \phi(\mathbf{X}^p)\|_F^2 \tag{5.2}$$

where superscript $\top$ denotes matrix transpose operation.

Cross-view Consistency regularisation

Intuitively, the learned mapping matrices $\mathbf{U}^p(p = 1, \ldots P)$ are arbitrarily inconsistent due to distinct data pattern, which focuses on alleviating data gap between cameras but sacrifices discriminativeness. This is inconsistent with our expectation since images of the same person from different cameras are inherently correlated. As in (Yu et al. 2017a), we add another regularisation term to keep the cross-view consistency, given as

$$\mathcal{J}_3(\mathbf{U}^p) = \sum_{i \neq j} \|\mathbf{U}^i - \mathbf{U}^j\|_F^2 \tag{5.3}$$

It is worth to point out that the term is specifically designed to the case when data from two modalities can be represented using the same feature descriptor, i.e., the person images across two domains do not change too much. This term requires dimensional consistency between mapping matrices. However, the term can be omitted while matching between two significant feature patterns, e.g., text-image retrieval and cross-biometric recognition.

Top-push Constrained Laplacian Graph regularisation (Top-push)

To make the learned dictionary discriminative, Laplacian graph regularisation is usually imposed to minimize the difference between coefficient vectors of samples but from the same class and maximize the difference between coefficient vectors of visually similar samples from different classes, *e.g.*, (Lu et al. 2017) and (Zheng et al. 2011a). In person re-ID, we also require the learned dictionary to be discriminative that the distance between samples from different people should be larger than that of the same person by a margin of ρ. Different from previous works, we consider a top-push ranking metric embedded graph regularisation inspired by the success of ranking matching in Re-ID. Since the coefficient vectors of samples from different modalities are learned in a common subspace spanned on a shared dictionary, $\mathbf{A}^p (p = 1 \dots P)$ can be treated equally. Thus, we define $\mathbf{A} = [\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^P]$ and its class labels l_i corresponding to the i-th sample $\mathbf{a}_i$ in $\mathbf{A}$. Following the principle, ranking methods minimizing the hinge loss of triplets achieve significant success (Wang et al. 2016). Compared with triplet loss, the top-push constraint enhances top-rank matching, which only considers the relationship between the distance of the positive pair and minimum distance of its related negative pairs, given as

$$\min \sum_{l_i=l_j} \max\{\mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) - \min_{l_i \neq l_k} \mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) + \rho, 0\} \tag{5.4}$$

where $\mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) = (\mathbf{a}_i - \mathbf{a}_j)^\top \mathbf{W}^\top \mathbf{W}(\mathbf{a}_i - \mathbf{a}_j)$ is the squared Mahalanobis distance, which indicates distance of a positive pair is closer than its corresponding negative pairs.

It is reasonable to reformulate the top-push constraint into a Laplacian graph representation by ignoring some constant terms, scaling coefficients and imposing a regularisation term on $\mathbf{W}$ inspired by (Cheng et al. 2017), denoting as

$$\mathcal{J}_4(\mathbf{A}, \mathbf{W}) = trace(\mathbf{WALA}^\top \mathbf{W}^\top) + \gamma \|\mathbf{W}\|_F^2 \tag{5.5}$$

where $\gamma(\gamma \geq 0)$ is the trade-off parameter, $trace(\cdot)$ represents trace of a matrix, $\mathbf{L}$ is the Laplacian matrix, defined as $\mathbf{L} = \mathbf{G} - (\mathbf{S} + \mathbf{S}^\top)/2$, $\mathbf{G}$ is a diagonal matrix whose i-th diagonal element is $g_{ii} = \sum_{j=1, j\neq i} \frac{s_{ij}+s_{ji}}{2}$, and s_{ij} is the entry of the weight matrix $\mathbf{S}$ of graph edges which denotes the similarity between the adjacent pairwise vertices $(\mathbf{a}_i, \mathbf{a}_j)$, defining as

$$s_{ij} = \begin{cases} \varepsilon \left[\mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) - \min\limits_{\substack{k\in[1,N],\\ l_i \neq l_k}} \mathcal{D}_{\mathbf{W}}(\mathbf{a}_j, \mathbf{a}_k) + \rho \right]_{l_i = l_j}, & i \neq j, \\ -\varepsilon \left[\max\limits_{\substack{k\in[1,N],\\ l_i = l_k}} \mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) - \mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) + \rho \right]_{l_i \neq l_j}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{5.6}$$

where $\varepsilon[\cdot]$ is an indicator function whose value is zero for negative input, and one otherwise, $N = \sum_{p=1}^{P} N_p$.

To take the benefits of Eq. 5.1, 5.2, 5.3 and 5.5 that simultaneously complete asymmetric mapping and discriminative dictionary learning, our overall optimization objective is

$$\begin{aligned} &\min_{\mathbf{U}^p, \mathbf{D}, \mathbf{A}^p, \mathbf{W}} \mathcal{J}_1 + \lambda_1 \mathcal{J}_2 + \lambda_2 \mathcal{J}_3 + \lambda_3 \mathcal{J}_4 \\ &\text{s.t.} \quad \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \ldots, K; \lambda_1, \lambda_2, \lambda_3, \gamma \geq 0 \end{aligned} \tag{5.7}$$

where λ_1, λ_2, and λ_3 are three trade-off parameters which balance the contributions of different terms. The objective jointly learns mapping matrices with respect to each modality to transform samples from heterogeneous domains into a shared subspace and a discriminative encoding dictionary to reconstruct features in the shared space. Similarity scores between samples then can be computed by Mahalanobis distance of encoding coefficients.

5.2.3 Optimization

To optimize the objective function Eq. 5.7, we first make some simplification and rewrite it to a compact form. For convenience, we define some auxiliary matrices

$$\mathbf{U} = [\mathbf{U}^1, \mathbf{U}^2, \cdots, \mathbf{U}^P], \mathbf{A} = [\mathbf{A}^1, \cdots, \mathbf{A}^P]$$
$$and\ \phi(\mathbf{X}) = \begin{bmatrix} \phi(\mathbf{X}^1) & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \phi(\mathbf{X}^2) & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \phi(\mathbf{X}^P) \end{bmatrix}, \tag{5.8}$$

where $\mathbf{0}$ is a zero matrix whose entries are all 0. Therefore, the first term $\mathcal{J}_1$ can be simplified as

$$\mathcal{J}_1(\mathbf{U}, \mathbf{D}, \mathbf{A}) = \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{D}\mathbf{A}\|_F^2 + \alpha\|\mathbf{A}\|_F^2 \tag{5.9}$$

By ignoring some constant terms, according to Shekhar et al. (2013), the second term $\mathcal{J}_2$ can be rewritten as

$$\mathcal{J}_2(\mathbf{U}) = -trace(\mathbf{U}\phi(\mathbf{X})\phi(\mathbf{X})^\top\mathbf{U}^\top) \tag{5.10}$$

And the third term can be rewritten as

$$\mathcal{J}_3(\mathbf{U}) = trace(\mathbf{U}\mathbf{Z}\mathbf{U}^\top) \tag{5.11}$$

where

$$\mathbf{Z} = \begin{bmatrix} (P-1)\mathbf{I} & -\mathbf{I} & \cdots & -\mathbf{I} \\ -\mathbf{I} & (P-1)\mathbf{I} & \cdots & -\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & -\mathbf{I} & \cdots & (P-1)\mathbf{I} \end{bmatrix}$$

and $\mathbf{I}$ denotes the identity matrix.

By substituting Eq. [5.9, 5.10, 5.11] and Eq. 5.5 into Eq. 5.7, the optimization

problem can be finally simplified as

$$\begin{aligned}\min_{\mathbf{U},\mathbf{D},\mathbf{A},\mathbf{W}} & \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{D}\mathbf{A}\|_F^2 - trace\{\mathbf{U}(\lambda_2\mathbf{Z} - \lambda_1\phi(\mathbf{X})\phi(\mathbf{X})^\top)\mathbf{U}^\top\} \\ & + \lambda_3 trace(\mathbf{W}\mathbf{A}\mathbf{L}\mathbf{A}^\top\mathbf{W}^\top) + \alpha\|\mathbf{A}\|_F^2 + \gamma\|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \ldots, K; \lambda_1, \lambda_2, \lambda_3, \gamma \geq 0\end{aligned} \tag{5.12}$$

It is clear that the objective function in Eq.5.12 is not jointly convex to variables $\mathbf{U}$, $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{W}$. The formulation cannot be directly solved by convex optimization. Following by previous works (Li et al. 2015; Karanam et al. 2015; Lu et al. 2017; Peng et al. 2017; Cheng et al. 2017), we adopt an iteration optimization procedure which alternatively optimizes one variable by fixing others, as follows:

(1) *Initialization.* Considering efficiency of the optimization, some initializations on variables are made based on empirical experience: (a) Mapping matrices, i.e., $\mathbf{U}^p (p = 1, \ldots, P)$ and $\mathbf{W}$, are all initialized as identity matrix; (b) The shared dictionary $\mathbf{D}$ and the corresponding coefficients $\mathbf{A}$ are initialized by solving the standard dictionary learning problem in which input feature matrix $\mathbf{X}$ is defined as in Eq. 5.9.

(2) *Given* $\mathbf{D}, \mathbf{A}$ *and* $\mathbf{W}$*, update* $\mathbf{U}$. By ignoring the irrelevant terms regarding variable $\mathbf{U}$, we can rewrite the optimization problem Eq. 5.12 as

$$\begin{aligned}\min_{\mathbf{U}} \mathcal{L}(\mathbf{U}) = \min_{\mathbf{U}} \ & \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{D}\mathbf{A}\|_F^2 \\ & + trace\{\mathbf{U}(\lambda_2\mathbf{Z} - \lambda_1\phi(\mathbf{X})\phi(\mathbf{X})^\top)\mathbf{U}^\top\}\end{aligned} \tag{5.13}$$

By setting $\frac{\partial\mathcal{L}(\mathbf{U})}{\partial\mathbf{U}} = 0$, we get the analytical solution of $\mathbf{U}$: $\mathbf{U} = \mathbf{D}\mathbf{A}\phi(\mathbf{X})^\top\mathbf{\Omega}^{-1}$, where $\mathbf{\Omega} = [(1 - \lambda_1)\phi(\mathbf{X})\phi(\mathbf{X})^\top + \lambda_2\mathbf{Z}]$. When $\mathbf{U}$ is obtained, $\mathbf{U}^p$ can be computed by splitting $\mathbf{U}$ into slices according to Eq. 5.8.

(3) *Given* $\mathbf{U}, \mathbf{A}$ *and* $\mathbf{W}$*, update* $\mathbf{D}$. The optimization problem is reduced to

$$\begin{aligned}\min_{\mathbf{D}} \ & \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{D}\mathbf{A}\|_F^2, \\ \text{s.t.} \quad & \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \ldots, K\end{aligned} \tag{5.14}$$

Define $\tilde{\mathbf{X}} \triangleq \mathbf{U}\phi(\mathbf{X})$, the quadratic constrained least square problem can be solved using the Lagrange dual approach. As in Lee et al. (2007), the optimal solution of Eq. 5.14 is $\mathbf{D}^* = \tilde{\mathbf{X}}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \Lambda)^+$, where superscript + denotes pseudo inverse operation and Λ is a diagonal matrix whose diagonal entries are dual variables.

(4) *Given* $\mathbf{U}, \mathbf{D}$ *and* $\mathbf{W}$*, update* $\mathbf{A}$. With $\mathbf{U}, \mathbf{D}$ and $\mathbf{W}$ fixed, we obtain the following loss function

$$\mathcal{F}(\mathbf{A}) = \|\tilde{\mathbf{X}} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda_3 trace(\mathbf{W}\mathbf{A}\mathbf{L}\mathbf{A}^\top\mathbf{W}^\top) + \alpha\|\mathbf{A}\|_F^2 \tag{5.15}$$

It is noticeable that the Laplacian matrix $\mathbf{L}$ in term $trace(\mathbf{W}\mathbf{A}\mathbf{L}\mathbf{A}^\top\mathbf{W}^\top)$ explicitly depends on variable $\mathbf{A}$ during the iterations, which causes the objective Eq. 5.15 intractable. Inspired by Cheng et al. (2017), we pre-calculate $\mathbf{L}$ using the prior $\mathbf{A}$ which ensures the objective to be convergent, and then update $\mathbf{L}$ with the new $\mathbf{A}$. Since $\mathbf{L}$ is non-positive semi-definite, we alternatively optimize objective Eq. 5.15 with gradient descent method, given as

$$\mathbf{A}^{(t)} := \mathbf{A}^{(t-1)} - \eta \bigtriangledown \mathcal{F}(\mathbf{A}^{(t-1)}), \quad t \geq 1 \tag{5.16}$$

where $\mathbf{A}^{(t)}$ denotes the t-th step to update variable $\mathbf{A}$, $\eta(\eta \geq 0)$ is the learning rate, and the gradient of Eq. 5.15 with respect to $\mathbf{A}$ is calculated by

$$\bigtriangledown\mathcal{F}(\mathbf{A}) = 2\mathbf{\Theta}\mathbf{A} + \lambda_3\mathbf{W}^\top\mathbf{W}\mathbf{A}(\mathbf{L}^\top + \mathbf{L}) - 2\mathbf{D}^\top\tilde{\mathbf{X}} \tag{5.17}$$

where $\mathbf{\Theta} \triangleq \mathbf{D}^\top\mathbf{D} + \alpha\mathbf{I}$. When updating $\mathbf{A}$ in the t-th iteration, $\mathbf{L}$ is firstly pre-computed with fixed $\mathbf{A}^{(t-1)}$. After obtaining $\mathbf{A}^{(t)}$ by Eq. 5.17, $\mathbf{L}$ is subsequently updated.

(5) *Given* $\mathbf{U}, \mathbf{D}$ *and* $\mathbf{A}$*, Update* $\mathbf{W}$. When $\mathbf{U}, \mathbf{D}$ and $\mathbf{A}$ are fixed, the objective function in Eq.5.12 can be rewritten as

$$\mathcal{H}(\mathbf{W}) = \lambda_3 trace(\mathbf{W}\mathbf{A}\mathbf{L}\mathbf{A}^\top\mathbf{W}^\top) + \gamma\|\mathbf{W}\|_F^2 \tag{5.18}$$

Algorithm 1: Top-push Constrained Modality-Adaptive Dictionary Learning

Input: Features of training images from P modalities: $\phi(\mathbf{X}^p), p = 1, \ldots, P$, parameters $\alpha, \gamma, \rho, \eta$, and $\lambda_i, i \in \{1, 2, 3\}$, iteration number T.

Output: The feature learning matrices $\mathbf{U}^p, p = 1, \ldots, P$, the learned shared dictionary $\mathbf{D}$ and the learned projection matrix $\mathbf{W}$.

Initialize $\mathbf{U} = [\mathbf{U}^1, \mathbf{U}^2, \ldots, \mathbf{U}^P], \mathbf{D}, \mathbf{W}$ and $\mathbf{A} = [\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^P]$ as described in subsection 5.2.3;

for $\kappa = 1 \longrightarrow T$ **do**

  Update $\mathbf{U}^p$ by Eq. 5.13;

  Update $\mathbf{D}$ according to Eq. 5.14;

  Calculate Laplacian matrix $\mathbf{L}$ using current $\mathbf{W}$ and $\mathbf{A}$ by Eq. 5.5 and 5.6;

  while *Non-convergence* **do**

    Update $\mathbf{A}$ by Eq.5.15;

  end

  Calculate Laplacian matrix $\mathbf{L}$ using current $\mathbf{W}$ and $\mathbf{A}$ by Eq. 5.5 and 5.6;

  while *Non-convergence* **do**

    Update $\mathbf{W}$ by Eq. 5.18;

  end

end

As in step (3), the Laplacian matrix $\mathbf{L}$ is also relevant to $\mathbf{W}$, so $\mathbf{L}$ should be kept fixed when updating $\mathbf{W}$. Gradient descent method is also utilised to optimize $\mathbf{W}$, and the corresponding gradient is deduced as

$$\bigtriangledown \mathcal{H}(\mathbf{W}) = \mathbf{W}[\lambda_3 \mathbf{A}(\mathbf{L}^\top + \mathbf{L})\mathbf{A}^\top + 2\gamma \mathbf{I}] \tag{5.19}$$

After obtaining $\mathbf{W}$, we update $\mathbf{L}$ subsequently according to Eq. 5.4 and Eq. 5.5.

5.2.4 Complexity Analysis

The complete algorithm is summarized in **Algorithm 1**. In practice, the objective Eq. 5.12 can converge to the local optimum after $T = 30$ iterations. According to the procedure, computational costs are mainly caused by inverse operations in Eq. 5.13 and learning dictionaries in Eq. 5.14, which is $\mathcal{O}((\sum_{p=1}^{P} n_p)^3)$ and $\mathcal{O}(K^3)$ respectively in each iteration. Thus, the computational complexity in T iterations is $\mathcal{O}(T[(\sum_{p=1}^{P} n_p)^3 + K^3])$.

5.2.5 Matching for Cross-modality Person Re-ID

Given a query person feature vector $\phi(x^p)$ from the p-th modality and gallery person feature vectors from the g-th modality $\phi(\mathbf{x}_i^g), i = 1, \ldots, N_g$, the encoding coefficients $\mathbf{a}^p$ and $\mathbf{a}_i^g$ can be computed by

$$\begin{aligned} \mathbf{a}^p &= \arg\min_{\mathbf{a}} \|\mathbf{U}^p\phi(\mathbf{x}^p) - \mathbf{D}\mathbf{a}\|_2^2 + \alpha\|\mathbf{a}\|_2^2 \\ \mathbf{a}_i^g &= \arg\min_{\mathbf{a}} \|\mathbf{U}^g\phi(\mathbf{x}_i^g) - \mathbf{D}\mathbf{a}\|_2^2 + \alpha\|\mathbf{a}\|_2^2 \end{aligned} \tag{5.20}$$

with respect to the shared dictionary $\mathbf{D}$. Then, the similarity scores between the query person and gallery persons can be calculated by

$$Score(i) = -\|\mathbf{W}(\mathbf{a}^p - \mathbf{a}_i^g)\|_2, \forall i, i = 1, \ldots, N_g \tag{5.21}$$

Thus, we assign the query sample $\mathbf{x}^p$ to the category corresponding to the largest score in the gallery set.

5.3 Experiments

In this section, we evaluate our method TCMDL (AsyDic +Top-push) on two benchmark datasets: NIR versus VIS Re-ID dataset SYSU-MM01 (Wu et al. 2017b) and the classical RGB-D person Re-ID dataset BIWI RGBD-ID (Munaro et al. 2014b,a). Moreover, two variants of the proposed method, *i.e.*, Dic+Top-push by

removing asymmetric mapping and AsyDic+Triplet by changing top-push constraint to classical triplet ranking constraint, are compared to demonstrate the benefits of combining AsyDic and top-push constraint. In the paper, two popular evaluation metrics are adopted: Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) (Zheng et al. 2017a). We reported the rank-k accuracy on both datasets, which is the cumulative identification rate of the true matches in the top k ranks. Moreover, mAP is reported on both datasets, which are mean of average precision scores for each query. We randomly repeat our evaluation for 10 times and report the average performances. All of the experiments are performed using Matlab on a desktop with a configuration of 64-bit OS, Intel(R) Core(TM) i5-6300U CPU @ 2.4 GHz and 8GB RAM.

Baselines. In the paper, four types of approaches are compared, which are state-of-the-art metric learning approaches for person Re-ID, cross-modality retrieval models, dictionary learning approaches and popular deep learning models. The representative metric learning approaches include KISSME (Koestinger et al. 2012) and XQDA (Liao et al. 2015). The cross-modality retrieval methods include CCA (Rasiwasia et al. 2010), GMA (Sharma et al. 2012), SCM (Zhang and Li 2014) and CRAFT (Chen et al. 2018b). We also compare with supervised dictionary learning method DicRW (Cheng et al. 2017) and unsupervised cross-dataset transfer learning UMDL (Peng et al. 2016). For deep learning methods, we compare four state-of-the-art models on thermal-visible* person Re-ID, DeepZero (Wu et al. 2017b), TONE (Ye et al. 2018b), BCTR (Ye et al. 2018d) and BDTR (Ye et al. 2018d). For KISSME, XQDA and the dictionary-based methods, we compute matching scores using the Mahalanobis distance. For all other methods, matching scores are directly calculated by Euclidean distance. Since the feature dimensions of depth images and RGB images in BIWI RGBD-ID dataset are different, PCA is

*In this paper, we use thermal-visible and infrared-RGB exchange.

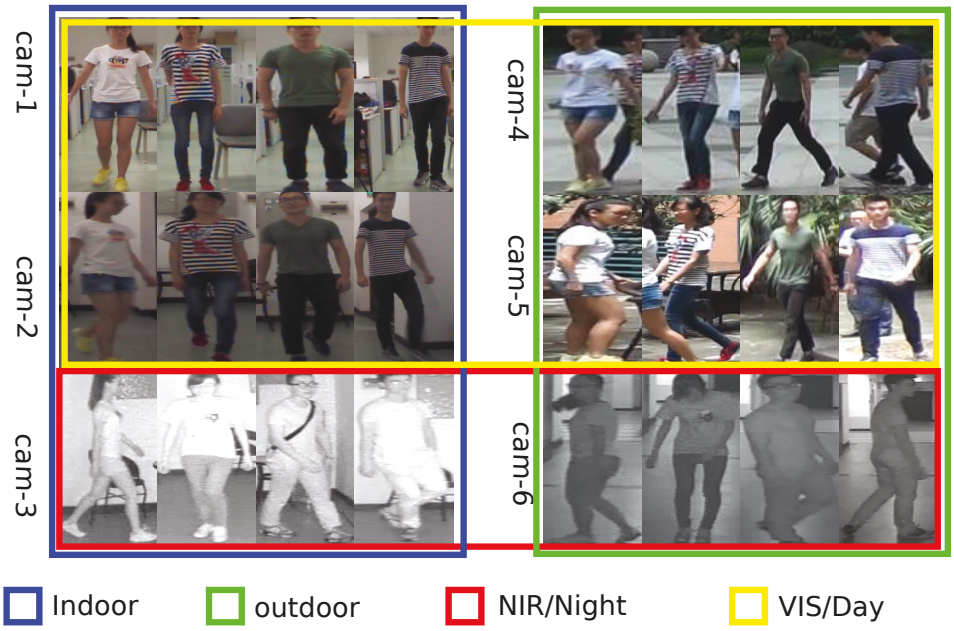


Figure 5.3 : Examples of samples in SYSU-MM01 dataset. Images from cameras 1-3 in the blue box are captured on indoor scenes while images from camera 4-6 in the green box are captured on outdoor scenes. Cameras 1, 2, 4, 5 are visual light sensors and cameras 3, 6 are near-infared sensors. Every column represents images from the same person.

firstly applied to get a fixed dimensional (i.e., 80) feature.

5.3.1 Experiments on SYSU-MM01

SYSU-MM01 is the largest dataset for CCM-reID. The dataset includes 287,628 RGB images from 4 VIS cameras in bright environments and 15,792 NIR images from 2 NIR cameras in dark environments of 491 valid identities. Some examples are shown in Figure 5.3. It is clear that images captured by NIR sensors are different from those captured by VIS sensors from perceptual experience. In NIR images, colour and texture information which are critical for traditional person Re-ID using VIS images are seriously degraded. The large data biases cannot be generalized by previous approaches for person Re-ID and thus incurs the cross-modality Re-ID

problem. Although only two modalities are in our experiment, i.e., NIR images and VIS images, the proposed method is also suitable for the scenarios where multiple modalities are available.

Setting

Feature Representation. We use two kinds of feature representations $\phi(\cdot)$ to evaluate our approach, i.e., LOMO (Liao et al. 2015) and Deep Zero-Padding (DZP) (Wu et al. 2017b). LOMO is a state-of-the-art hand-crafted feature representation for classical single-modality person Re-ID which characterizes person using colour and texture information. DZP is learned by a one-stream network which extracts features of heterogeneous data by learning domain-specific nodes.

Evaluation Protocol. We follow the evaluation protocol of (Wu et al. 2017b) which 296 fixed identities are for training and another 96 identities for testing. Differently, we leverage one image per identity for training in the training set, which is identical to single-shot person Re-ID. During testing, we follow the two validation modes of (Wu et al. 2017b), *all-search* mode and *indoor-search* mode. In both modes, all images of NIR images from two NIR cameras form the probe set. Particularly, images from all VIS cameras form the gallery set for *all-search* mode while images from VIS camera #1 and #2 deployed indoor form the gallery set for *indoor-search* mode. In both modes, we follow the single-shot setting in (Wu et al. 2017b) that only chooses one image for each identity in the gallery set and all images in the probe set (3803 query images).

Parameter Setting. In our experiments, we empirically set balance parameters λ_1 and λ_2 as 0.002 and 0.001, parameters for regularisation terms α and γ as 0.05. λ_3 is set to $\beta/N(\rho)$ where $N(\rho)$ is the number of triplet sets, β is set to 800 and 50 empirically for LOMO and DZP respectively. As in metric learning approaches (You et al. 2016; Wang et al. 2016), the margin ρ is simply set to 1. More detailed

parameter analysis is in Subsection 5.3.3.

Evaluation

LOMO Feature Representation. We extract LOMO using the code provided by (Liao et al. 2015) with default parameters. All images from SYSU-MM01 dataset are resized into 160×60 due to varying bounding box size and thus generate 35722-dimensional features. To overcome dimension curse, we perform PCA on LOMO vectors from each modality respectively and reduce dimension to 300.

Table 5.1 lists the CMC and mAP results using LOMO for *all-search* mode and *indoor-search* mode, respectively. We can observe cross-modality retrieval methods such as GMA, SCM and ours achieve better performances than classical metric learning for person Re-ID. Compared to using baseline Euclidean metric, these methods achieve 0.37 (from 1.6% to 1.97%), 0.40% (from 1.60% to 2.00%) and 1.01% (from 1.6% to 2.61%) improvements of rank-1 accuracy in *all-search* mode and 0.34% (from 2.56% to 2.90%), 0.54% (from 2.56% to 3.10%), 2.09% (from 2.56% to 4.65%) improvements of rank-1 accuracy in *indoor-search* mode respectively. This is because cross-modality methods can mitigate data biases between heterogeneous data while classical metric learning methods cannot generalize large differences. However, CCA and CRAFT achieve poor performance due to severe noises, i.e., colour information. Among the cross-modality methods, our model achieves better performances in all modes, especially in larger ranks, e.g., more than rank-10. Compared to the single-modality dictionary learning DicRW (Cheng et al. 2017), our model improves the performance with a large margin in both modes (from 1.84% to 2.61% for *all-search* and from 3.10% to 4.65% for *indoor-search*). This validates the effectiveness of our model to mitigate data biases for CCM-reID. In particular, our method TCMDL improves rank-1 accuracy from 1.99% to 2.61% and mAP from 4.30% to 5.07% for *all-search*, improves rank-1 accuracy from 3.45% to 4.65% and mAP from 9.94% to

Table 5.1 : Results using LOMO (%). '-' means result is not reported.

	Method	*all-search*					*indoor-search*				
		mAP	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20
Metric Learning	Euclidean	3.63	1.60	7.30	14.07	26.62	8.52	2.56	12.22	22.90	41.43
	KISSME (Koestinger et al. 2012)	4.43	1.76	9.12	17.55	32.46	9.78	3.20	14.18	26.27	48.56
	XQDA (Liao et al. 2015)	4.22	1.88	9.06	17.17	31.15	9.27	3.16	13.65	25.18	45.82
Cross-Modality Retrieval	CCA (Rasiwasia et al. 2010)	3.57	1.37	6.84	13.06	25.54	8.09	2.06	11.40	22.43	41.36
	GMA (Sharma et al. 2012)	4.22	1.97	8.64	16.18	29.68	9.36	2.90	13.51	25.48	46.25
	SCM (Zhang and Li 2014)	4.08	2.00	8.87	16.42	30.00	9.31	3.10	13.61	25.36	46.27
	CRAFT (Chen et al. 2018b)	3.55	1.53	7.17	13.69	25.93	8.09	2.31	11.40	21.42	40.26
Dictionary Learning	DicRW(Dic+Triplet) (Cheng et al. 2017)	4.06	1.84	8.86	15.65	29.03	8.93	3.10	13.30	24.43	43.29
	UMDL (Peng et al. 2016)	4.61	2.46	10.22	18.24	32.19	9.35	3.09	13.25	24.86	45.18
Ours	**Dic+Top-push**	4.30	1.99	8.97	16.49	30.04	9.94	3.45	14.71	27.04	48.21
	AsyDic+Triplet	4.97	2.48	10.86	19.60	35.08	11.72	4.52	**18.34**	**31.87**	**54.56**
	TCMDL	**5.07**	**2.61**	**11.21**	**20.17**	**35.65**	**11.79**	**4.65**	18.33	31.65	54.24

11.79% for *indoor-search* compared to Dic+Top-push. The results show the effectiveness of asymmetric mapping. In another aspect, our method TCMDL improves rank-1 accuracy from 2.48% to 2.61% and mAP from 4.97% to 5.07% for *all-search*, improves rank-1 accuracy from 4.52% to 4.65% and mAP from 11.72% to 11.79% for *indoor-search* compared to AsyDic+Triplet. The results demonstrate the effectiveness of top-push constrained regularisation.

However, performances of the proposed methods are still limited, which the rank-1 and mAP are 2.61% and 5.07% for *all-search* mode and 4.65% and 11.79% for *indoor-search* respectively. This is because LOMO is characterized by colour and texture information which are degraded seriously in NIR images. The imbalanced feature information from different cameras restricts the identification accuracy. Moreover, performances in *indoor-search* mode are much better than that in *all-search* mode because illumination and background interferences can be better controlled under indoor scenarios.

Deep Zero-Padding Feature Representation. In this paper, 256-dimension DZP extracted by (Wu et al. 2017b) are utilised. Both *all-search* mode and *indoor-search* mode are adopted with respect to the protocol above. The experiments are also conducted 10 times with random sample selection. Table 5.2 gives the results when using the DZP. Compared to LOMO, DZP is specially designed for CCM-reID, which results in better performances than that in Table 5.1 when using the same method. As in Table 5.2, our model achieves the best performances no matter CMC ranks or mAP. Though larger data biases have been already migrated when learning DZP, our model obtains about 3.4% (from 15.95% to 19.36%) and 6.4 % (from 26.92% to 33.35%) mAP improvements compared to the baseline using Euclidean metric. It shows strong feature representation augmentation power of our model. In particular, our model outperforms other cross-modality models with a large margin, e.g., more than 13% and 19% rank-1 accuracy improvement than

Table 5.2 : Results using DZP (%). '-' means result is not reported.

Method		*all-search*					*indoor-search*				
		mAP	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20
Metric Learning	Euclidean Wu et al. (2017b)	15.95	14.80	-	54.12	71.33	26.92	20.58	-	68.38	85.79
	KISSME (Koestinger et al. 2012)	14.81	12.24	34.69	50.47	69.30	27.36	16.44	45.99	64.08	82.46
	XQDA (Liao et al. 2015)	18.42	15.87	40.95	57.57	75.72	31.16	20.04	51.08	68.70	86.15
Cross-Modality Retrieval	CCA (Rasiwasia et al. 2010)	19.10	16.71	42.43	58.46	76.12	32.11	21.46	51.68	68.91	85.91
	GMA (Sharma et al. 2012)	12.02	10.34	29.69	43.25	59.66	21.93	13.40	35.91	51.78	71.18
	SCM (Zhang and Li 2014)	4.70	2.60	10.78	19.44	34.17	10.07	3.56	15.19	27.60	48.86
	CRFAT (Chen et al. 2018b)	4.79	3.08	11.14	19.00	32.07	10.10	3.96	15.87	27.24	46.97
Dictionary Learning	DicRW(Dic+Triplet) (Cheng et al. 2017)	17.60	14.41	38.92	54.48	72.55	30.60	20.47	50.23	68.25	85.46
	UMDL (Peng et al. 2016)	17.45	15.35	39.68	55.04	72.31	28.67	18.82	46.33	63.05	80.60
Ours	**Dic+Top-push**	17.66	15.85	40.75	56.42	73.36	30.98	21.51	49.05	64.72	81.52
	AsyDic+Triplet	**19.32**	16.57	42.62	58.62	**77.23**	32.19	21.49	52.02	69.37	86.20
	TCMDL	19.30	**16.91**	**42.74**	**58.83**	76.64	**32.27**	**21.60**	**54.26**	**71.38**	**87.91**

Table 5.3 : Comparison with the state-of-the-art methods for thermal-visible Re-ID on SYSU-MM01 dataset.

Methods	mAP	rank-1	rank-10	rank-20
DeepZero (Wu et al. 2017b)	15.95	14.80	54.12	71.33
TONE (Ye et al. 2018b)	14.42	12.52	50.72	68.60
BCTR (Ye et al. 2018d)	19.15	16.12	54.90	71.47
BDTR (Ye et al. 2018d)	**19.66**	**17.01**	55.43	71.96
Ours	19.30	16.91	**58.83**	**76.64**

CRFAT for *all-search* mode and *indoor-search* mode, respectively. In addition, our model achieves more than 2% mAP improvement than DicRW, which shows strong feature representation ability when data biases exist. Similar to the conclusion using LOMO, the proposed model outperforms Dic+Top-push and AsyDic+Triplet in most cases. This validates the benefits of combining asymmetric mapping and top-pushed constrained dictionary learning together.

Table 5.3 compares the performances of *all-search* mode with the state-of-the-art thermal-visible person Re-ID methods on SYSU-MM01 dataset. These methods either use one-stream or two-stream neural networks to mitigate data biases across RGB visual images and infrared images. From the table, it is easy to observe that our method achieves equivalent performances with these methods. Especially, it outperforms DeepZero and TONE with a large margin since we use deep zero-padding features as input and further mitigate modality biases by jointly performing asymmetric mapping and discriminative dictionary learning.

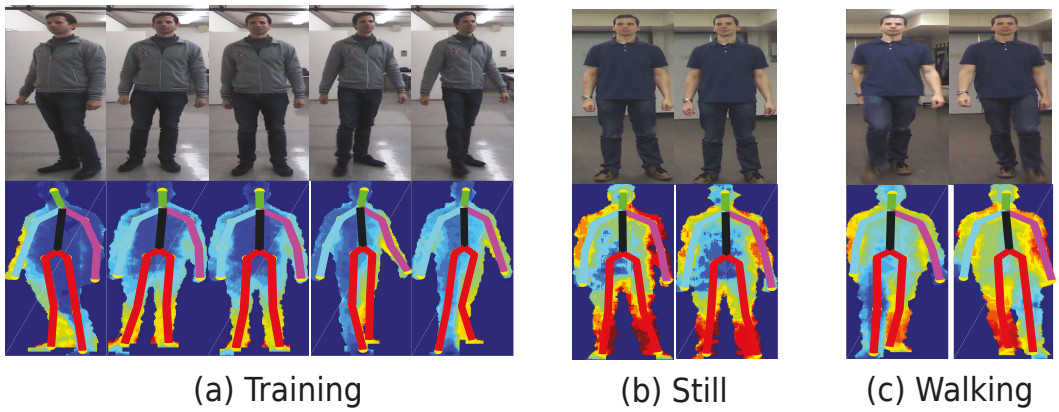


Figure 5.4 : Examples of images in the BIWI RGBD-ID dataset. Images in the top row are RGB images and in the bottom row are depth images (shown by pseudo-colour) as well as skeletons.

5.3.2 Experiments on BIWI RGBD-ID dataset

The BIWI RGBD-ID dataset (Munaro et al. 2014b,a) is originally built for long-term person Re-ID that uses depth information instead of RGB images since depth image is robust to cloth changing and illumination. Different from (Munaro et al. 2014b,a) and (Wu et al. 2017a), we evaluate CCM-reID on the dataset, which considers depth images as query set and corresponding RGB images as gallery set. The dataset includes video sequences captured from 50 different people by a Kinect at about 10 frames per second. For each sequence of a subject, there are about 300 frames of RGB images and their corresponding depth images as well as skeletons. In the dataset of BIWI RGBD-ID, these sequences are originally provided in two groups corresponding to two folders "*Training*" and "*Testing*" which are denoted as *TR* and *TE* in the following. *TE* is sub-divided into two groups named as "*Still*" and "*Walking*". Only 28 out of 50 persons appeared in both *TR* and *TE* (*i.e., Still and Walking*), and they are collected on a different day which thus most subject dressed differently. In *TR*, the subject performs actions such as walking and rotation. In *Still*, people stand in front of the sensor and move slightly. In *Walking*,

Figure 5.5 : Examples of point clouds. Images in top row is the RGB images in BIWI RGBD-ID dataset and images in bottom row are their corresponding visualization of point clouds.

people walk frontally and diagonally against the Kinect. Some examples are shown in Figure 5.4.

Setting

Feature Representation. Due to the large gaps between RGB images and depth images, we extract features with different methods for them. For depth images, we first convert them into point clouds as in (Munaro et al. 2014b) and describe body shape and skeleton of people using 510-dimension Eigen-depth feature (Wu et al. 2017a) and 13-dimension skeleton-based feature (Munaro et al. 2014a). Some examples of RGB images of persons in the dataset and their corresponding point clouds are shown in Figure 5.5. For RGB images, we also hope the visual fea-

tures could describe body shape rather than colour information because the colour is varying between *TR* and *TE* groups. In this paper, we utilised LBP (Ojala et al. 2002) and HOG (Dalal and Triggs 2005) to describe texture and body silhouette, respectively. When extracting LBP and HOG features, we firstly resize all RGB images to 128×48 and convert them to grayscale images. 8×8 cells are used for LBP and HOG feature extraction, which results in 5664 and 2700 dimensional features.

Evaluation Protocol. To evaluate the proposed method on BIWI RGBD-ID dataset, we follow the data protocol in (Wu et al. 2017a) which images of 28 people who appeared in both *TR* and *TE (Walking and Still)* are used for testing, and the remaining 22 subjects who only appeared in *TR* are used for training the model. Different from (Wu et al. 2017a), we take RGB images to construct the gallery set and take depth images as the probe. Thus, three groups of testing sets are built, i.e., images of the 28 persons who appeared in all of *TR*, *Walking* and *Still*. We termed the three groups of testing sets as subset #1, #2 and subset #3. Since depth information is not complete for all frames in one sequence, we selected samples as advised in (Munaro et al. 2014b) and 5 frames are used for each sequence.

Parameter Setting. For BIWI RGBD-ID dataset, we also empirically set trade-off parameters λ_1 as 0.002, parameters for regularisation terms α and γ as 0.005. λ_3 is set to $\beta/N(\rho)$ where $N(\rho)$ is the number of triplet sets, β is set to 0.8. As in metric learning approaches (You et al. 2016; Wang et al. 2016), the margin ρ is simply set to 1. Since the large gaps between different feature representation from RGB images and depth images, we omit the cross-modality consistency term in Eq. 5.10 and set the dimension of asymmetric mapping matrices to be 80. Considering the number of people in the dataset, we set the dictionary size to 50.

Table 5.4 : Results on BIWI RGBD-ID subset #1 (%).

Method	mAP	rank-1	rank-5	rank-10	rank-20
Euclidean	9.82	2.14	10.71	24.29	54.29
KISSME (Koestinger et al. 2012)	15.97	3.57	22.86	35.71	74.29
XQDA (Liao et al. 2015)	19.70	6.43	26.43	47.86	**90.00**
CCA (Rasiwasia et al. 2010)	18.62	7.14	26.43	37.86	70.71
GMA (Sharma et al. 2012)	14.42	3.57	17.86	39.29	73.57
CRAFT (Chen et al. 2018b)	12.03	0.71	18.57	40.00	72.86
Ours	**19.96**	**10.71**	**29.29**	**48.57**	84.29

Table 5.5 : Results on BIWI RGBD-ID subset #2 (%).

Method	mAP	rank-1	rank-5	rank-10	rank-20
Euclidean	10.76	0.00	12.14	40.00	75.00
KISSME (Koestinger et al. 2012)	13.82	3.57	18.57	42.14	75.71
XQDA (Liao et al. 2015)	17.31	6.43	17.86	**47.86**	74.29
CCA (Rasiwasia et al. 2010)	15.51	4.29	20.00	35.00	**77.86**
GMA (Sharma et al. 2012)	14.13	3.57	17.86	36.43	72.14
CRAFT (Chen et al. 2018b)	16.50	4.29	**24.29**	42.14	70.71
Ours	**18.88**	**7.14**	20.00	41.43	77.14

Evaluation

Table 5.4-5.6 list results on the three testing subsets of BIWI RGBD-ID dataset, i.e., *TR*, *Walking* and *Still*. It is easy to observe that our proposed method achieves significant results on all the three subsets. In detail, our proposed method achieves

Table 5.6 : Results on BIWI RGBD-ID subset #3 (%).

Method	mAP	rank-1	rank-5	rank-10	rank-20
Euclidean	10.01	0.71	13.57	25.71	64.29
KISSME (Koestinger et al. 2012)	13.26	5.00	16.43	30.71	67.86
XQDA (Liao et al. 2015)	16.32	5.00	21.43	**43.57**	72.86
CCA (Rasiwasia et al. 2010)	13.40	5.00	14.29	27.86	58.57
GMA (Sharma et al. 2012)	14.23	3.57	21.43	39.29	68.57
CRAFT (Chen et al. 2018b)	**17.73**	6.43	**22.86**	37.86	71.43
Ours	17.53	**7.14**	20.00	42.14	**76.43**

10.71%, 7.14% and 7.14% accuracy on the three subsets at rank-1, which outperforms baseline Euclidean with a large margin. For other ranks, our method also achieves top and stable performances in most cases. This shows the effectiveness of our proposed method.

In another aspect, results on subset #1 and subset #3 are better than that on subset #2. It is reasonable because data distribution of subset #1 is more similar than other subsets and subset #3 suffers less motion variation than other subsets. As in the tables, cross-modality methods achieve relevantly good results than single-modality methods. It is because asymmetric mapping bridges the data gaps to some content. It is interesting that XQDA achieves remarkable performances in some cases in benefits of powerful discriminative ability of metric learning. However, it still cannot reach the performance of the proposed method due to the large gaps between modalities.

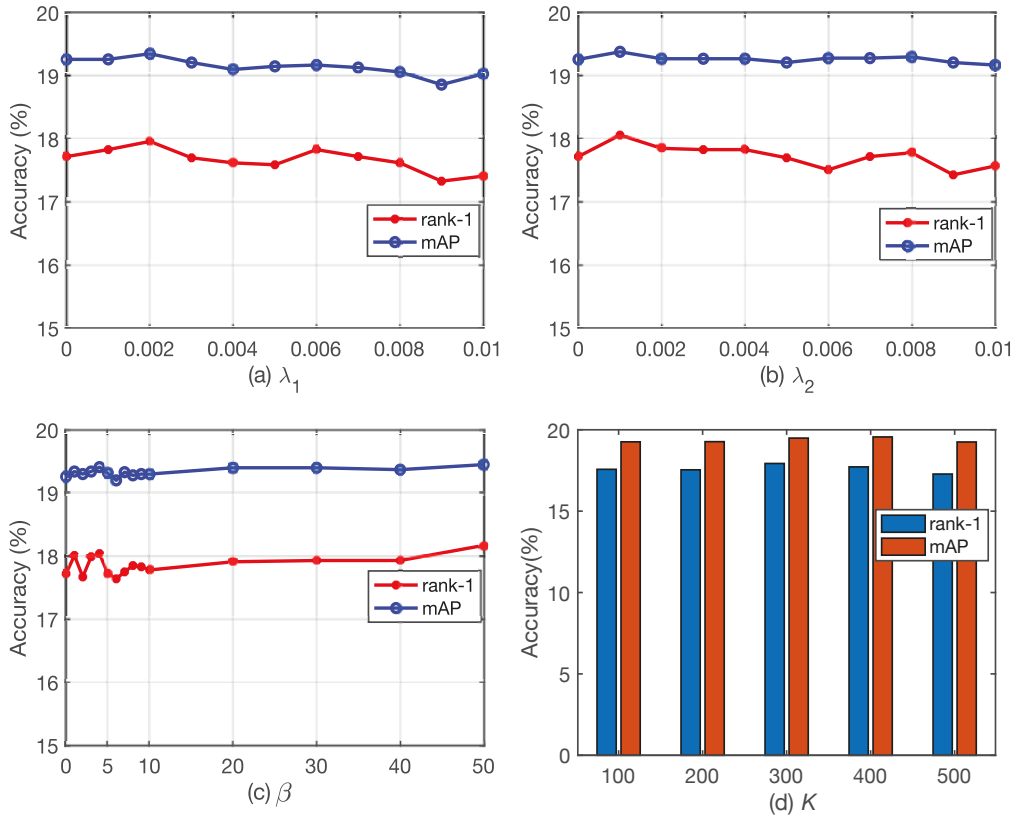


Figure 5.6 : Parameter analysis using DZP for all-search mode. Rank-1 and mAP accuracy with different parameters (a) λ_1, (b) λ_2, (c) β and dictionary size (d) K are reported.

5.3.3 Parameter Analysis

Analysis on SYSU-MM01

As in the objective function Eq. 5.12, our model includes four parts, and corresponding three parameters λ_1, λ_2 and $\lambda_3 = \beta/N(\rho)$ to balance contributions of each part. Figure 5.6(a-c) shows the performance changes under two evaluation metrics (*i.e.*, rank-1 and mAP) in terms of the three different trade-off parameters on the SYSU-MM01 dataset when using DZP, respectively. From the figure, it can be observed that our model is less sensitive to λ_1 and λ_2 than λ_3. However, we can still find that both rank-1 and mAP performance rise with increasing of λ_1 and achieve

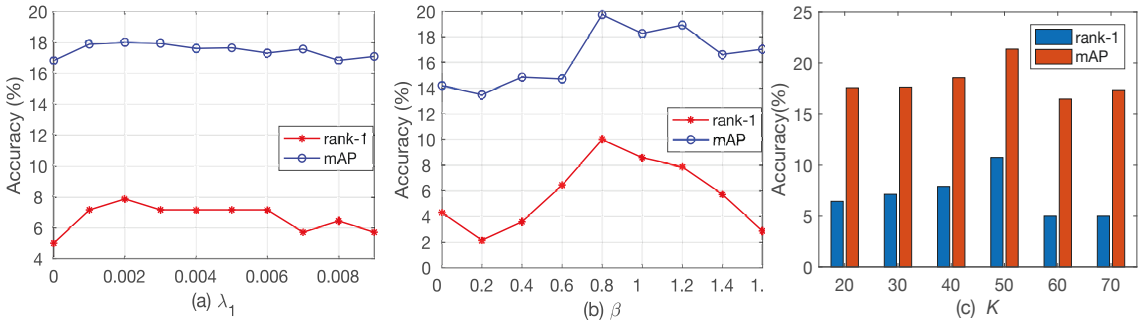


Figure 5.7 : Parameter analysis on subset #1 of BIWI RGBD-ID dataset. Rank-1 and mAP accuracy in terms of different parameters (a) λ_1, (b) β, and dictionary size (c) K are reported.

peak values at 0.002. After that, they show a downward trend with small fluctuation. And, λ_2 varies in a similar way with λ_1 except achieves peak performance at 0.001. Though the performance is relatively stable when the choices of λ_1 and λ_2 in suitable ranges, it is easy to get better performance when setting λ_1 and λ_2 to 0.002 and 0.001 from the above analysis. From Fig. 5.6(c), we can see that the performances fluctuate drastically when β is less than 20. However, it can achieve stable performance when β is larger than 20. To balance each part in the objective, we empirically set parameters $\lambda_1 = 0.002, \lambda_2 = 0.001$ and $\beta = 50$ when using DZP. As for LOMO, λ_1 and λ_2 comply the same variation rule and thus can be set to same value. But for β, we empirically find the optimum value is 800.

Dictionary size K of $\mathbf{D}$ is another important parameter. Fig. 5.6(d) shows rank-1 and mAP accuracies for K in $[100, 500]$. We can observe that our model is not sensitive to dictionary size in suitable ranges. Considering the running time, we set $K = 300$ in all our experiments.

Analysis on BIWI RGBD-ID

Since different feature representations are used to describe RGB images and depth images on BIWI RGBD-ID dataset, we do not use cross-view consistency regularisation as described in Sec. 5.2.2. Thus, only two parameters λ_1 and $\lambda_3 = \beta/N(\rho)$ are left to control the contribution of two regularisation terms. Figure 5.7(a) illustrates two evaluation metrics rank-1 and mAP against λ_1. It can be observed that both rank-1 and mAP performance rise with increasing of λ_1 and achieve the highest performance at 0.002. After that, the graph shows a downward trend with small fluctuation. Fig. 5.7 (b) displays the relationship between accuracies measured by rank-1 and mAP and β. From the figure, both rank-1 and mAP performance show an upward trend and achieve peak performance at 0.8. After that, rank-1 performance drops drastically while mAP performance shows a downward trend with small fluctuations. From the above analysis, it is supposed to set $\lambda_1 = 0.002$ and $\beta = 0.8$ on BIWI RGBD-ID dataset. Fig. 5.7(c) illustrates the rank-1 and mAP accuracies for K in range $[20, 70]$. We can observe that both rank-1 and mAP performance achieve the highest performance when $K = 50$. This is much smaller than K for SYSU-MM01 dataset. We suppose the reason is that the number of people in BIWI RGBD-ID dataset is relatively small, which only has 22 for training and 28 for testing.

5.3.4 Effects of Energy Terms

As in the proposed objective Eq. 5.7, there are four components which control dictionary learning in shared subspace, prevent information loss, keep mapping matrices consistent and force dictionary discriminative, respectively. Basically, we want to reconstruct data from different modalities using a shared dictionary in a common latent subspace. Considering the propose, reconstruction error is minimized to ensure the learned dictionary to be representative. Three penalty terms are adopted

to regularize the dictionary learning. We evaluate the effects of these regularisation terms on both SYSU MM01 and BIWI RGBD-ID dataset and report results in Figure 5.6 and Figure 5.7. The energy-preserving regularisation controlled by λ_1 tries to preserve as much information while performing the asymmetric mapping. As in Figure 5.6(a) and Figure 5.7(a), the performance will drop if λ_1 is too small due to too much information loss. However, there will be redundant if this term is too large. The cross-view consistency term aims to present the learned mapping matrices for different modalities varying too much. However, it incurs the mapping matrices to be the same if the term dominates the whole objective and thus cause poor performance. This is verified by results in Figure 5.6(b). The last penalty term attempts to make the learned dictionary discriminative. As in Figure 5.6(c) and Figure 5.7(b), both small and large β will cause poor performance. This is because small β causes the term contributing less to the whole objective and thus makes the learned dictionary less discriminative while large β incurs over-fitting. Thus, it is important to choose appropriate trade-off parameters to balance the contribution of these different penalty terms.

5.4 Summary

This chapter presents the study on CCM-reID problem. In particular, a top-push constrained modality-adaptive dictionary learning model is proposed, which simultaneously maps the features from different modalities into a common subspace and learns a shared dictionary for the projected features of all modalities in the subspace. Moreover, we impose a top-push constraint to the coding coefficients, which improves the discriminative ability of the learned dictionary. Experiments on two benchmark datasets demonstrate the effectiveness of the proposed TCMDL.

Chapter 6

Learning Hybrid Representations over Tracklets for Person Re-ID in the Wild

This chapter investigates to learn hybrid representations over tracklets for LTG-reID in a data-driven manner. In particular, this chapter introduces a dual-stream network named SpTSkM, including a spatial-temporal stream and a skeleton motion stream. The former performs directly on image sequences, which tends to learn identity-related spatial-temporal patterns such as body geometric structure and body movement. The latter operates on normalized 3D skeletons by adapting graph convolutional network, which tends to learn pure motion patterns from skeleton sequences. Both streams extract fine-grained level time-gap stable information that is robust to appearance changes in LTG-reID and meanwhile maintains sufficient discriminability to differentiate different people. The final matching metric is obtained by mixing information of the two streams in a score-level fusion strategy. In addition, this chapter introduces a new video-based dataset particular for LTG-reID that is the largest one till now. Extensive experiments demonstrate difficulty of the new dataset and also validate effectiveness of the proposed SpTSkM, showing the best performance.

6.1 Introduction

6.1.1 Problem Formulation

Appearance inconsistency seriously limits performance of existing CST-reID methods while applying to CLT-reID*. In tentative works, hand-crafting representations from video tracklets are adopted (Gou et al. 2016; Zhang et al. 2018a). These approaches tend to explore dynamic or motion characteristics of human body from dense trajectories, which are invariant to drastic changes in appearance space. This kind of idea marginally increases the identification performance. It shows effectiveness of motion characteristic for CLT-reID. However, such hand-craft features are designed empirically that cannot completely mine discriminative properties.

In addition to motion attribute, it is demonstrated that there are some other subtle identity attributes that are time-gap stable (Zheng et al. 2019), e.g., geometric structure of human body, hair color and style, etc. These properties also bring potential benefits to the long-term re-ID task. We differentiate these time-gap stable patterns from time-gap sensitive patterns as in Table 6.1. It is important for the long-term person re-ID to explore all these identity-related and time-stable properties in order to improve identification performance. Unfortunately, none of the previous methods considers all these attributes up to now.

Thus, this chapter introduces the hybrid representation learning problem that tries to learn representations simultaneously from both motion pattern analysis and subtle identity attributes.

6.1.2 Motivation

Considering the strong expression ability, this chapter tries to learn the hybrid representation using a dual-stream network. The dual-stream network includes two

*In this section, CLT-reID refers to the case of person re-ID after long-time gap (i.e., LTG-reID).

Table 6.1 : Summarization of potential properties for person re-ID. It is categorized by attribute's time-gap stability.

Time-gap Sensitive Patterns	Time-gap Stable Patterns
clothing/shoes color, texture and style, background, carrying, pose, etc.	Motion/gait characteristics, geometric structure of human body, hair color and style, etc.

subnets: Set-based Subtle Identity Net (SSIN) and Skeleton-based Motion Identity Net (SMIN), which learns subtle identity properties from video tracklets and motion attributes from 3D skeleton, respectively. Our motivation is three-fold. Firstly, subtle identity information lies in images (Zheng et al. 2019) and spatial-temporal patterns (Chao et al. 2019) hides in video sequences. It motivates us to produce subtle identity features related to individual in frame level and learn relative location changes of local body parts in video level. Secondly, self-attention mechanism (Zhang et al. 2019a) characterizes location responses of all spatial regions. It motivates us to adapt it to model mutual relation (or non-local responses) of local body parts (Wang et al. 2018) and capture geometric characteristics of human body. However, background noise may incur useless response that corrupts the attended feature map. It motivates us to suppress the noise by masking the background. To achieve both, it inspires us to fuse self-attention mechanism (Zhang et al. 2019a) and mask attention together and propose a mask-guided cross-attention module. Finally, considering the strong representative power of graph on the topology structure of body joints, graph convolutional network (GCN) (Yan et al. 2018) has shown great success in action recognition. It motivates us to adapt the popular GCN to our re-ID task and force it to extract identity-related motions by using an identity loss.

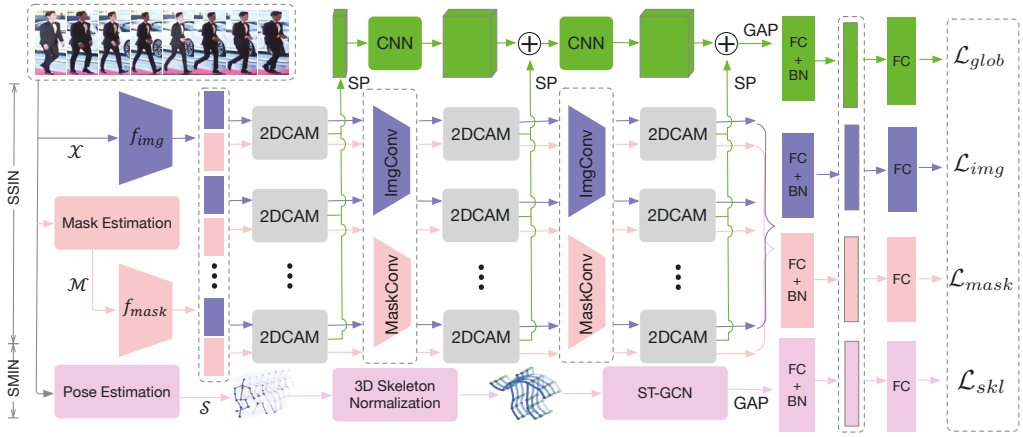


Figure 6.1 : An overview of the proposed SpTSkM framework. "$\oplus$" denotes element-wise addition, "SP" denotes set pooling, "GAP" denotes global average pooling. Different color represents different processing flow.

6.2 The Proposed Method

6.2.1 Structure Overview

We formulate the proposed SpTSkM from two perspectives: Learning subtle identity properties directly from image sequences (SSIN) and exploring motion patterns from 3D skeletons (SMIN), as shown in Figure 6.1. Given a dataset of N sequences $\mathcal{X} = \{\mathcal{X}_i; l_i\}_{i=1}^{N}$, where l_i is the identity label, $\mathcal{X}_i = \{x_i^j\}_{j=1}^{N_i}$ is a sequence of N_i images from the same person. To reduce the effect of background and highlight human geometric structure information, we first estimate the body masks on each frame using pre-trained LIP (Liang et al. 2018). The SSIN stream learns representations from both image sequences and their corresponding masks. It is implemented by cascading several mask-guided cross-attention blocks on the head of CNN backbones f_{img} and f_{mask}. Moreover, we use set pooling to aggregate the frame-level features and then use two CNN blocks further get final representation. Details of SSIN are described in Sec. 6.2.2.

For the SMIN stream, we aim to explore motion patterns from 3D skeleton

sequences. To achieve the purpose, we first extract 3D skeletons using the pre-trained weakly-supervised 3D pose estimation method in (Zhou et al. 2017b). As persons walk along with various directions against camera along with different poses, the estimated skeletons inherently suffer variations from the viewpoint and pose. Thus, it is hard to extract motion patterns from the estimated skeletons for person identification without view and pose normalization. In light of the problem, we propose a pose normalization module that aligns and normalizes these skeletons in a pre-defined direction. While the skeletons are normalized, we then extract motion patterns using a spatial-temporal GCN network (Yan et al. 2018) to mine motion-related representation. The stream will be further introduced in Sec. 6.2.3.

6.2.2 Set-based Subtle Identity Net

The goal of set-based subtle identity net (SSIN) is to explore discriminative properties from image sequences. As shown in Figure 6.1, the SSIN takes a sequence of images $\mathcal{X}_i = \{x_i^j\}_{j=1}^{N_i}$ as input, and firstly extracts frame-level features using a CNN backbone, such as ResNet50 (He et al. 2016), formulated as $\{\hat{x}_i^j\}_{j=1}^{N_i} = f_{img}(\{x_i^j\}_{j=1}^{N_i})$. Meanwhile, SSIN estimates mask regarding to each frame using the pre-trained LIP (Liang et al. 2018), formulated as $\{m_i^j\}_{j=1}^{N_i} = f_{lip}(\{x_i^j\}_{j=1}^{N_i})$, and extracts their frame-level mid-features using another CNN network, formulated as $\{\hat{m}_i^j\}_{j=1}^{N_i} = f_{mask}(\{m_i^j\}_{j=1}^{N_i})$. To learn mutual relation (or non-local responses (Wang et al. 2018)) of different local parts, we propose to cascade several no-local blocks which capture part dependencies between any two positions of human body. It enlarges receptive field of each pixel and builds correspondence between any two pixels. We achieve this in both spatial-level (i.e., 2DCAM) and spatial-temporal level (i.e., 3DCAM). Moreover, we integrate mask attention into the no-local block in order to filter out noise background and highlight human geometric structure, named as *Cross-Attention Module*.

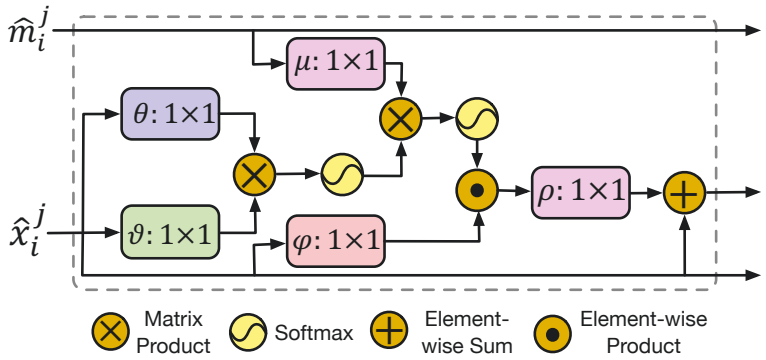


Figure 6.2 : The proposed 2D cross-attention module (2DCAM) for an image. The boxes denote 2D convolution operations with 1×1 kernels.

Cross-attention Module. Figure 6.2 shows the structure of a basic 2D cross-attention module (2DCAM). As in Figure 6.2, we denote features of a certain frame from prior hidden layer $\hat{x}_i^j \in \mathbb{R}^{C \times S}$ and features of its corresponding mask $\hat{m}_i^j \in \mathbb{R}^{C \times S}$. Here C and S denote number of channels and feature locations in current layer, respectively. We first calculate the self-attention map of the image, formulated as

$$\xi_i^j(v,k) = softmax\{\theta(\hat{x}_i^j(k))^\top \vartheta(\hat{x}_i^j(v))\} \tag{6.1}$$

where $\xi_i^j(v,k)$ indicates the responses of the v-th location with respect to all possible k-th locations. $\theta(\hat{x}_i^j) = W_\theta \hat{x}_i^j$ and $\vartheta(\hat{x}_i^j) = W_\vartheta \hat{x}_i^j$. Here $W_\theta \in \mathbb{R}^{\bar{C} \times C}$ and $W_\vartheta \in \mathbb{R}^{\bar{C} \times C}$ are two weight matrices, $\bar{C}$ is the number of reduced channels. In practice, they are implemented by 1×1 convolutions. To filter responses from background area, a mask embedding is computed, i.e., $\mu(\hat{m}_i^j) = W_\mu \hat{m}_i^j$, and multiplied to the self-attention map, formulated as

$$\zeta_i^j(v) = \sum_{i=1}^{N} \xi_i^j(v,k) \mu(\hat{m}_i^j(k)) \tag{6.2}$$

where $\zeta_i^j \in \mathbb{R}^{\bar{C} \times S}$ indicates attention map after filtering. To obtain non-local feature maps, another linear embedding is utilized, i.e., $\varphi(\hat{x}_i^j) = W_\varphi \hat{x}_i^j$. It multiplies mask-guided map $softmax(\zeta_i^j)$ to filter noise background, formulated as

$$\nu_i^j = softmax(\zeta_i^j) \odot \varphi_i^j \tag{6.3}$$

where $\odot$ denotes element-wise product. The hidden map is then re-mapped to the original space, which produces the non-local feature maps, i.e., $\rho(\nu_i^j) = W_\rho \nu_i^j$, where $W_\rho \in \mathbb{R}^{C \times \bar{C}}$. We adopt a residual-like architecture which produces the final feature by adaptively summing the non-local feature maps and the original input, formulated as

$$o_i^j = \lambda \rho(\nu_i^j) + \hat{x}_i^j \tag{6.4}$$

where λ is a learning parameter. The CAM achieves to learn location dependencies over spatial structure. The temporal relation is then explored by set aggregation as below description, which produces the global features. It is worth noting that CAM does not alter feature maps $\hat{x}_i^j$ and $\hat{m}_i^j$. They carry frame-level information from images and masks, which are directly transmitted to the next layer to produce high-level semantics. In Figure 6.1, flows of $\hat{x}_i^j$ and $\hat{m}_i^j$ are represented by blue and orange arrowline, respectively. Moreover, we also extend CAM to 3D cross-attention (3DCAM) by replacing 2D convolution with 3D convolution of $1 \times 1 \times 1$ kernels. However, the 3DCAM performs on a sequence $\mathcal{X}_i$ and its corresponding mask sequence $\mathcal{M}_i$ as inputs. It indicates that 3DCAM can capture space-time non-local relations, and response in position (k, v, t) can be represented by all other positions over the whole sequence, where k, v, t denote the spatial position (k, v) in the t-th frame. Similar to 2DCAM, we also leverage a set aggregation strategy to calculate global features along with the temporal direction.

Set Aggregation. To obtain the global features over the whole sequence, we borrow the idea of set pooling (SP) (Chao et al. 2019), which is formulated as

$$\begin{aligned} \pi_i = max\{O_i + conv(cat[max(O_i), \\ mean(O_i), median(O_i)]_{\times N_i}) \odot O_i\} \end{aligned} \tag{6.5}$$

where $O_i = [o_i^1; o_i^2, \cdots; o_i^{N_i}]$ is the tensor of N_i feature maps in the sequence, $max(\cdot), mean(\cdot), median(\cdot)$ denote three different pooling operations along set, $cat[\cdot,\cdot]_{\times N_i}$ denotes duplicating N_i times across set dimension after channel concatenation, and *conv* denotes 1×1 convolution. In practice, we observed that directly performing max pooling along set achieves competitive results as illustrated in Sec. 6.4. For saving memory, we directly use max pooling during our experiments. In another aspect, we perform set pooling in different levels. It helps aggregate multi-level information, such as low-level spatial and high-level semantic attributes.

Classification loss. Since SSIN outputs hybrid representations of images, masks and global sequence, it is supposed that each sub-representation can be classified to the correct identity. Thus, we use three cross-entropy losses to converge the training process, for instance,

$$\begin{aligned} \mathcal{L}_{img} &= -\frac{1}{N_b}\sum_{i=1}^{N_b}\log p(l_i|\mathcal{X}_i) \\ \mathcal{L}_{mask} &= -\frac{1}{N_b}\sum_{i=1}^{N_b}\log p(l_i|\mathcal{M}_i) \\ \mathcal{L}_{glob} &= -\frac{1}{N_b}\sum_{i=1}^{N_b}\log p(l_i|\mathcal{X}_i, \mathcal{M}_i) \end{aligned} \tag{6.6}$$

where N_b is the number of sequences in a mini-batch. $p(l_i|\mathcal{X}_i)$ and $p(l_i|\mathcal{M}_i)$ are the predicted probabilities that $\mathcal{X}_i$ and $\mathcal{M}_i$ belong to their ground-truth identity l_i. Here, we pool the sequences before predicting the probability so that each sequence produces one prediction. $p(l_i|\mathcal{X}_i, \mathcal{M}_i)$ is the probability of global output of image sequence $\mathcal{X}_i$ and $\mathcal{M}_i$, which classifies inputs to the correct classes. We balance the three parts and formulate the classification loss as

$$\mathcal{L}_{cls}^{SSIN} = \alpha_1 \mathcal{L}_{img} + \alpha_2 \mathcal{L}_{mask} + \alpha_3 \mathcal{L}_{glob} \tag{6.7}$$

where $\{\alpha_i\}_{i=1}^{3}$ are trade-off parameters, and $\sum_{i=1}^{3} \alpha_i = 1$. Influence of these hyper-parameters are discussed in Sec. 6.4.7.

Discriminative Loss As re-ID is an identification task, we further enforce a discriminative loss on the final representations of the SSIN stream, such as batch-hard triplet loss (Hermans et al. 2017). Suppose a training batch includes image sequences from P persons and each person is sampled K sequences, the discriminative loss can be formulated as

$$\begin{aligned} \mathcal{L}_{dis}^{SSIN} = \sum_{i=1}^{P} \sum_{a=1}^{K} \max\{\epsilon + \max_{p \in [1,K]} dis(y_a^i, y_p^i) \\ - \min_{\substack{j \in [1,P] \\ n \in [1,K] \\ j \neq i}} dis(y_a^i, y_n^j), 0\} \end{aligned} \tag{6.8}$$

where y is $L2$ normalized embeddings produced from the SSIN stream. $dis(\cdot, \cdot)$ calculates the Euclidean distance of two input embedding vectors, and ϵ is the margin that is set to 0.5 empirically. Thus, the overall loss of SSIN is

$$\mathcal{L}_{SSIN} = (1 - \gamma)\mathcal{L}_{cls}^{SSIN} + \gamma \mathcal{L}_{dis}^{SSIN} \tag{6.9}$$

where γ is a hyper-parameter that trade off the contribution of classification loss and discriminative loss. In our experiment, we empirically set it to 0.4.

6.2.3 Skeleton-based Motion Identity Net

The goal of skeleton-based motion identity net (SMIN) is to learn identity-related movement patterns from 3D skeleton sequences. In our work, we estimate 3D human skeletons of each image sequence $\mathcal{X}_i = \{x_i^j\}_{j=1}^{N_i}$ using a pre-trained weakly-supervised approach in (Zhou et al. 2017b), denoting as $\mathcal{S}_i = \{s_i^j\}_{j=1}^{N_i}$. It outputs skeletons of 16 key-points as in Figure 6.3, which treats middle hip as the origin of the coordinate.

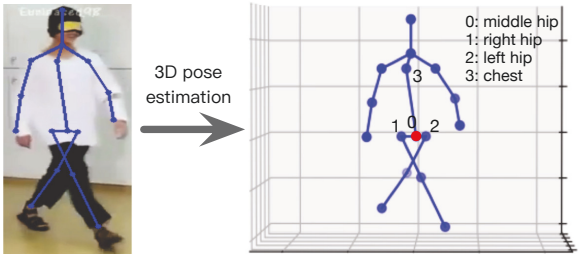


Figure 6.3 : An illustration of 3D pose estimation. Joints are highlighted by circle. Origin of coordinate is located on the middle hip of the estimated skeleton.

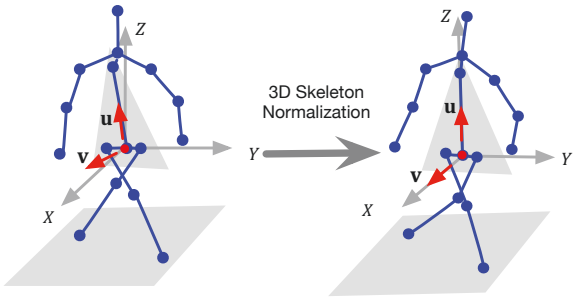


Figure 6.4 : An illustration of 3D skeleton normalization. Origin of coordinate is denoted by the red circle.

As shown in Figure 6.1, the proposed SMIN includes two cascaded stages: 3D skeleton normalization and dynamic feature learning. The former aims to transform the original 3D skeletons to a unified view, and the latter tries to model/learn discriminative attributes from these moving skeletons by spatial-temporal graph convolutional network (ST-GCN) (Yan et al. 2018).

3D Skeleton Normalization. Since people walk against cameras with different views, the estimated skeletons suffer serious view difference. This results in misaligned motion patterns that are ineffective for person re-identification. Thus, we propose to normalize the raw 3D skeletons to a unified view before dynamic feature learning. In specific, as illustrated in Figure 6.4, for a 3D skeleton s_i^j, we first determine a reference plane Θ according to positions of the chest and two hips.

As the origin of coordinate locates on the middle hip of the raw skeleton produced by (Zhou et al. 2017b), it cannot ensure the origin of coordinate roots in the plane Θ. Thus, we re-determine the centre of left and right hips as root point and translate it to the origin of coordinate using 3D translation R_t with an offset, such as $R_t : (t_x^k, t_y^k, t_z^k) \Rightarrow (t_x^k + \delta_x, t_y^k + \delta_y, t_z^k + \delta_z)$, where (t_x^k, t_y^k, t_z^k) is location of the k-th joint of s_i^j, and $(\delta_x, \delta_y, \delta_z)$ is the translation offset. After that, we correct the skeleton in a unified direction by geometric transformation in 3D space, for example, rotating around axes to make unit vector from root point to chest $\mathbf{u}$ parallel to Z-axis and unit normal vector $\mathbf{v}$ of plane Θ parallel to X-axis, which is formulated as

$$\hat{s}_i^j(k) = R s_i^j(k) \tag{6.10}$$

where $s_i^j(k) \in \mathbb{R}^3$ is a column vector formed by coordinate values of the k-th joint, R is the transformation matrix, which can decompose to rotating operations around three axes, such as $R = R_x R_y R_z$, where R_x, R_y, R_z are the corresponding transformation matrices, defined as

$$R_X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_X & -\sin\phi_X \\ 0 & \sin\phi_X & \cos\phi_X \end{bmatrix}, R_Y = \begin{bmatrix} \cos\phi_Y & 0 & \sin\phi_Y \\ 0 & 1 & 0 \\ -\sin\phi_Y & 0 & \cos\phi_Y \end{bmatrix}, R_Z = \begin{bmatrix} \cos\phi_Z & -\sin\phi_Z & 0 \\ \sin\phi_Z & \cos\phi_Z & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.11}$$

where ϕ_X, ϕ_Y and ϕ_Z are the corresponding rotating angles, which are easily calculated by vectors $\mathbf{u}$ and $\mathbf{v}$. Benefiting of 3D skeleton normalization, the perturbation from view difference is tackled, which thus enables to learn discriminative dynamic patterns from skeleton sequence for person re-ID.

Dynamic Feature Learning. As skeleton is a natural graph, recent works (Yan et al. 2018; Shi et al. 2019) have tried to encode and model motion dynamics of skeletons using GCN. It achieves great success in the field of action recognition. On the other hand, as we know, identity attribute also implicitly exists in actions

which has been testified in gait recognition (Ben et al. 2019b). Therefore, we borrow the idea from action recognition and adapt ST-GCN (Yan et al. 2018) to solving person re-ID problem.

Figure 6.5(a) presents our constructed spatio-temporal graph of a skeleton sequence, where joints are naturally connected in space and adjacently connected in time. Here, the spatial connection follows the one in previous works (Yan et al. 2018; Shi et al. 2019), which ignore the chest and center hip since they change little during walking. Figure 6.5(b) shows the used spatial graph partition, which determined according to distance from the node to the gravity center in a neghbor set. For a root node $\mathcal{V}_{ki}$ (i.e., the i-th joint of the k-th skeleton frame, $\hat{s}^k(i)$)[†], its neighbor set is defined as $\mathcal{B}(\mathcal{V}_{ki}) = \{\mathcal{V}_{kj} | disG(\mathcal{V}_{kj}, \mathcal{V}_{ki}) \leq D\}$, where $disG(\mathcal{V}_{kj}, \mathcal{V}_{ki})$ denotes the shortest graph distance from node $\mathcal{V}_{kj}$ to $\mathcal{V}_{ki}$, D is pre-defined distance that is set 1 in our work. Formally, the neighbor set is divided into three subsets $\{\mathcal{B}_i\}_{i=1}^{3}$,

$$p_{ki}(\mathcal{V}_{kj}) = \begin{cases} 0 & if\ r_j = r_i \\ 1 & if\ r_j < r_i \\ 2 & otherwise \end{cases} \tag{6.12}$$

where $p_{ki}(\mathcal{V}_{kj})$ is a mapping function for neighbour set of root node $\mathcal{V}_{ki}$ that categorizes vertex $\mathcal{V}_{kj}$ into the corresponding subset, i.e., 0 for root node, 1 for centripetal subset, 2 for centrifugal group. r_i is the average Euclidean distance between joint i and gravity centre of all the skeletons in the dataset.

Given the defined skeleton graph and partition strategy, we encode them with several spatio-temporal graph convolution layers and predict probability logits over identities using a global average pooling (GAP) layer and a softmax layer. Here introduces the graph convolution (GC) operation (Yan et al. 2018) on vertex $\mathcal{V}_{ki}$ in

[†]We use node $\mathcal{V}$ instead of joint s in graph representation, i.e. both $\mathcal{V}_{ki}$ and $s^k(i)$ represent i-th joint of k-th frame in a skeleton sequence.

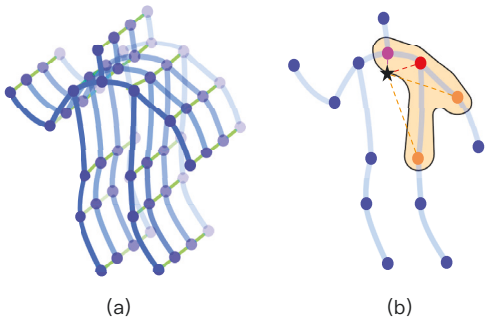


Figure 6.5 : (a) An illustration of constructed spatio-temporal graph; (b) An illustration of partition strategy. $\star$ denotes gravity center. For a neghbor set, it is categorized into three subsets: root joint (in red), centripetal joints (in magenta) and centrifugal joints (in orange).

space,

$$\hat{f}_{out}(\mathcal{V}_{ki}) = \sum_{\mathcal{V}_{kj} \in \mathcal{B}(\mathcal{V}_{ki})} \frac{1}{N_{ki}(\mathcal{V}_{kj})} \hat{f}_{in}(\mathcal{V}_{kj}) \cdot w(p_{ki}(\mathcal{V}_{kj})) \tag{6.13}$$

where $\hat{f}$ denotes feature map, w denotes weight function which provides a weight vector to compute inner product with given input, and $N_{ki}(\mathcal{V}_{kj})$ is the normalization term, i.e., $N_{ki}(\mathcal{V}_{kj}) = |\{\mathcal{V}_{kn} | p_{ki}(\mathcal{V}_{kn}) = p_{ki}(\mathcal{V}_{kj})\}|$, $|\cdot|$ denotes set cardinality. As joints are consecutively connected in time, it is easy to extend the GC to spatio-temporal domain by considering temporally connected joints for neighbor set $\mathcal{B}$ and partition function p, which is discussed in (Yan et al. 2018).

In practice, GC operation in space is not implemented straightforwardly. For a skeleton, Equation (6.13) is firstly rewritten to

$$\hat{\mathbf{f}}_{out} = \sum_{n}^{3} \mathbf{W}_n (\hat{\mathbf{f}}_{in} \mathbf{A}_n) \odot \mathbf{M}_n \tag{6.14}$$

where input feature map $\hat{\mathbf{f}}_{in}$ is a $C \times N_s$ tensor, C is channel number, N_s is number of joints. $\mathbf{W}_n \in \mathbb{R}^{C_{in} \times C_{out} \times 1 \times 1}$ is channel stacked weight matrix of 1×1 convolution. $\mathbf{A}_n = \mathbf{\Lambda}_n^{-\frac{1}{2}} \overline{\mathbf{A}}_n \mathbf{\Lambda}_n^{-\frac{1}{2}}$ is normalized adjacent matrix, $\overline{\mathbf{A}}_n \in \mathbb{R}^{N_s \times N_s}$ is the adjacent

matrix whose entry $\overline{\mathbf{A}}_n(i,j)$ indicates whether vertex $\mathcal{V}_{ki}$ belongs to partition subset $\mathcal{B}_n(\mathcal{V}_{kj})$ of vertex $\mathcal{V}_{kj}$. $\mathbf{\Lambda}$ is the normalizing diagonal matrix whose entry is defined as $\mathbf{\Lambda}_n(i,i) = \sum_j \overline{\mathbf{A}}_n(i,j) + \delta$, where $\delta = 0.001$ is used to prevent empty rows in $\mathbf{A}_n$. $\mathbf{M}_n \in \mathbb{R}^{N_s \times N_s}$ is a learnable attention matrix that weights the edge importance.

Concretely, input feature map $\hat{\mathbf{f}}_{in} \in \mathbb{R}^{C \times T \times N_s}$ is a 3D tensor with fixed T consecutive skeletons in practice. For the spatio-temporal case, we implement GC by performing a conventional 2D convolution of kernel size $\Gamma \times 1$ on the above output feature map. Here, Γ is the temporal kernel size.

Loss Function As SMIN aims to classify people using their walking skeleton sequence, we define the classification loss as

$$\mathcal{L}_{skl} = \frac{1}{N_b} \sum_{i=1}^{N_b} \log p(l_i|\mathcal{S}_i) \tag{6.15}$$

where $p(l_i|\mathcal{S}_i)$ is the predicted probability of skeleton sequence $\mathcal{S}_i$ over class l_i from SMIN stream. Analogous to SSIN, we also impose a discriminative loss on SMIN to boost ranking result, denoting as $\mathcal{L}_{dis}^{SMIN}$. Thus, the overall loss of SMIN stream is

$$\mathcal{L}_{SMIN} = (1-\gamma)\mathcal{L}_{skl} + \gamma\mathcal{L}_{dis}^{SMIN} \tag{6.16}$$

6.2.4 Two-stream Networks

As discussed in Sec. 6.1.1, both subtle identity characteristics and motion patterns benefit to person identification. Though SSIN explores all these properties, it is prone to investigating spatial patterns, because such kind of temporal set pooling is not excelled in motion extraction. Fortunately, the SMIN stream that is dedicated to action description just remedies the defect. The overall framework is illustrated in Figure 6.1. Given a walking tracklet, we extract representations by feeding it to the SSIN stream and SMIN stream, respectively. The matching scores between the query and gallery samples are thus calculated for both streams based on Euclidean

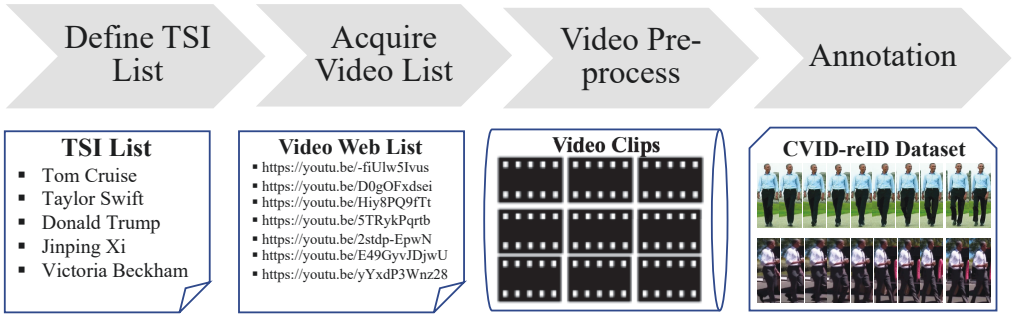


Figure 6.6 : The pipeline of data acquisition process for the proposed CVID-reID.

distance. We fuse to obtain final matching metric as illustrated in Eq. **??**, and predict the query's identity, such as

$$l^* = \arg\min_{l_i} s(\mathcal{X}_{query}, \mathcal{X}_i) \tag{6.17}$$

where $\mathcal{X}_{query}$ is the query sample and $\mathcal{X}_i$ denotes the i-th sample in gallery set. Thus, the predicted identity of query is same to the sample who is closest to the query in gallery set.

6.3 CVID-reID Dataset

In this section, we introduce the proposed CVID-reID dataset in terms of data acquisition and its uniqueness.

Data Acquisition

We follows the manner of ActivityNet (Caba Heilbron et al. 2015) to collect videos. As shown in Figure 6.6, our pipeline includes four stages:

- **Define TSI List.** Since TSIs change their clothes in long-term re-ID scenarios, it is difficult and time-consuming to annotate the same TSI from a large volume of videos taken by surveillance cameras. Thus, we propose to collect data of celebrities from the Internet because their street-shots in different websites are usually taken in entirely different dates and their clothes are usually

distinct to each other. To facilitate acquisition, we first define a TSI list that consists of 150 celebrities considering attributes such as nationality, age and gender, etc.

- **Acquire Video List.** Given the TSI list, we collect videos of these celebrities from publicly available sources on Internet. We directly use their name as keywords to obtain a video list of one specific TSI and then keep videos that contain the walking scene. Finally, we select 1500 videos.

- **Video Pre-process.** In this stage, we edit the obtained videos because only fragments of the walking scene are useful. We crop one sequence per video in order to guarantee intra-person clothing variation.

- **Annotation.** The last stage is to crop and label TSI from raw video sequences. We leverage CVAT[‡] to manually annotate the human body area in key frames and automatically generate bounding boxes of the whole video clip with interpolating mode. In addition, we cut these videos into tracklets with a frequency of 30 successive bounding boxes, which ensures each tracklet involves at least one walking cycle. Finally, total of 77935 bounding boxes from 2980 tracklets are obtained.

Data Uniqueness

The dataset is proposed particularly for long-term person re-ID, which is different from existing datasets as follows:

- **Clothing Variation.** Intra-person clothing difference is one of the most remarkable characteristics in long-term person re-ID. In our dataset, we define at least 10 videos per celebrity from different *URLs*, which incurs at least 5

[‡]The tool is available on `https://github.com/opencv/cvat`

Figure 6.7 : An example of six celebrities in the proposed CVID-reID dataset. It can be seen abundant of intra-person clothing changes exist in the dataset.

different clothes per celebrity, as shown in Figure 6.7. Up to now, this is the one involving the largest number of intra-person clothing changes.

- **Severe Disturbance.** In addition to clothing difference, the dataset also involves more severe challenges in terms of dynamic backgrounds, diverse viewpoints, varying illumination and indeterminate poses than prior datasets. This is because our dataset is collected from Internet whose video footages are taken by different cameras in distinct locations and time. This data set is even more challenging that many footages were taken by mobile devices such as mobile phone, action camera, aerial drone, etc.

- **Large Scale.** The proposed dataset is consist of total 77935 bounding boxes from 2980 tracklets of 90 identities, which is the largest dataset for long-term person re-ID till now.

6.4 Experiments

To evaluate the proposed method, we conducted extensive experiments on two different datasets: the proposed CVID-reID for long-term person re-ID and MARS for the traditional short-term person re-ID. They demonstrated the difficulty of the proposed CVID-reID dataset and also validated the effectiveness of the proposed SpTSkM for person re-ID in the wild, particularly for the long-term case.

6.4.1 Network Structure and Implementation Details

We implement the proposed method on two 16G Quadro P5000 GPUs. For SSIN, we utilize ResNet50 with parameters pre-trained on ImageNet (He et al. 2016) as backbone to extract representations from the last residual layer for images (f_{img}). And, we adopt ResNet structure with only the first residual block of each residual layer as the backbone to output heatmaps before the GAP layer (f_{mask}). After that, we cascade two 2DCAM-Conv blocks that consist of 2DCAMs processing each image and its mask pair of a sequence, ImgConv for images and MaskConv for masks. ImgConv includes four Conv-BN-ReLu blocks and MaskConv contains two Conv-ReLU blocks. We aggregate cross-attention maps with set pooling and process the global representation using CNN. We implement CNN using a ImgConv. For SMIN, we borrow the structure of ST-GCN (Yan et al. 2018) which adapts it for 3D skeleton with 14 joints. We add an FC-BN-Dropout layer at the head of each leaf branch to get final representations and map them to their corresponding class with an additional FC layer.

During training, we train the two streams separately with their optimization objective, for example, optimizing SSIN with $\mathcal{L}_{SSIN}$ and optimizing SMIN with $\mathcal{L}_{SMIN}$. For Celeb-ReID, we use SGD optimizer to train both streams, whose initial rate is set to 0.008 for pre-trained weights, 0.08 for others while optimizing SSIN and 0.05 for parameters of SMIN. We train both streams for 200 epochs. For SSIN, we decay the learning rate every 40 epochs with a decay rate of 0.1. For SMIN, the learning rate is decayed at 50 and 100 epochs with a decay rate 0.1. Due to the limitation of GPU resource, we train SSIN with a batch size of 16 tracklets whose length is 6. For the SMIN counterpart, we optimize it with a batch of 128 tracklets, and each includes 25 frames of skeletons. For both streams, we use a Dropout layer before the classification (FC) layer whose probability is set to 0.6. For MARS, we use Adam optimizer to speed up convergence, whose initial learning rate is set to

0.0003. We train 1000 epochs and decay the learning rate by 0.1 every 200 epochs.

6.4.2 Datasets and Experiment Setting

CVID-reID: The details of CVID-reID are elaborated in Sec. 6.3. We partition long-term CVID-reID dataset following the division method on classical benchmarks, which sequences of the first 50 identities are for training and the rest for testing. We further divided testing set into gallery and probe, for example, gallery set involves of first 4 sequences of each identity and probe set includes the rest. During training, we resized each image into a fixed shape 256×128 , and augmented the training set by random expansion-crop and randomly horizontal flip.

MARS: MARS (Zheng et al. 2016a) is the largest video-based benchmark for CST-reID till now. It is collected by 6 RGB cameras, which totally includes around 20 thousand tracklets from 1261 identities. In our experiment, we use tracklets that contains no less than 30 frames. For fair comparison, we follow the widely adopted split in previous works (Zheng et al. 2016a; Gao and Nevatia 2018), which evenly divides the dataset into training and testing set.

Following popular evaluation metrics for person re-ID, we report results in terms of cumulative matching characteristics (CMC) and mean Average Precision (mAP) for both long-term person re-ID on CVID-reID and conventional video-based person re-ID on MARS.

6.4.3 Challenges of Long-term Person Re-ID

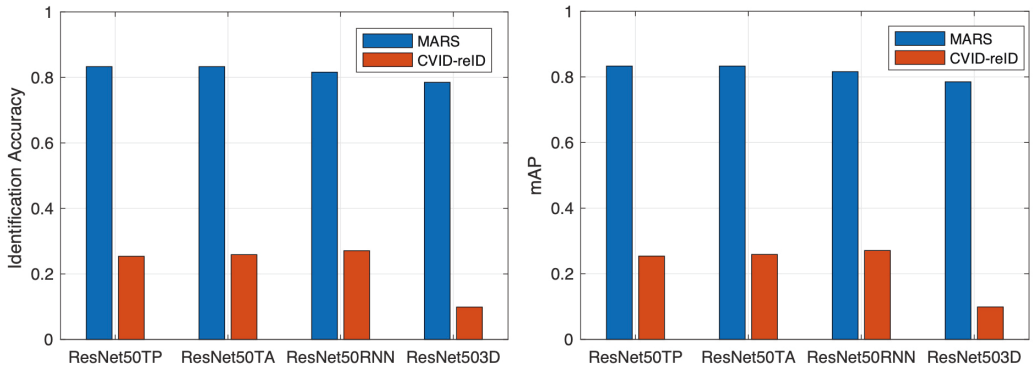
In this section, we elaborate challenges of long-term person re-ID. To achieve the purpose, we evaluate SOTA methods on the proposed long-term CVID-reID datasets and benchmark short-term MARS dataset. In particular, we chose (Gao and Nevatia 2018) as baseline because they evaluate video-based approaches with different prevailing temporal aggregation methods, such as temporal pooling (ResNet50TP),

temporal attention (ResNet50TA), RNN (ResNet50RNN) and 3D Convnet (ResNet503D). These ideas are widely inherited by recent works (McLaughlin et al. 2016; Rao et al. 2018; Li et al. 2019; Zhou et al. 2017c; Li et al. 2018a). Figure 6.8 illustrates performance of the baseline methods on two distinct type of datasets.

From Figure 6.8, we can observe that the performance of all methods drops drastically when applying them on the long-term CVID-reID dataset. In specific, the Rank-1 accuracy drops 57.9% (from 83.3% to 25.4%) for ResNet50TP, 57.4% (from 83.3% to 25.9%) for ResNet50TA, 54.5% (from 81.6% to 27.1%) for ResNet50RNN and 68.6% (from 78.5% to 9.9%), respectively. Correspondingly, the mAP performance drops 56.2% (from 76.5% to 20.3%) for ResNet50TP, 57.5% (from 76.7% to 19.2%) for ResNet50TA, 52.9% (73.9% to 21.0%) for ResNet50RNN and 62.8% (from 70.5% to 7.7%) for ResNet503D, respectively. These results verify the difficulty of long-term person re-ID and great challenges of the proposed long-term CVID-reID dataset. The difficulties have been elaborated in Sec. 6.3. It is worth highlighting again that CVID-reID suffers larger intra-person variations due to cloth changes along time lapse than MARS, in addition to other common variations in person re-ID, as illustrated in Figure 6.9.

6.4.4 Effectiveness of SpTSkM for Long-term Person Re-ID

This section evaluates the effectiveness of the proposed SpTSkM for long-term person re-ID. The results are listed in Table 6.2. We compare our solution with SOTA methods on proposed CVID-reID dataset. For SOTA methods, we select representative video-based methods that implementation codes are available. For example, We choose (Gao and Nevatia 2018) as a baseline. This paper revisited video-based methods with different temporal aggregation modelling methods, such as temporal pooling (ResNet50TP), temporal attention (ResNet50TA), RNN (ResNet50RNN) and 3D Convnet (ResNet503D). In addition, we also compared



(a) Rank-1 (b) mAP

Figure 6.8 : Performance comparison of four baseline methods on MARS and CVID-reID in terms of (a) identification accuracy (Rank-1) and (b) mAP.

Figure 6.9 : A comparison of intra-person variations between CVID-reID (top row) and MARS (bottom row).

some recent methods, e.g., RCNN (McLaughlin et al. 2016), M3D (Li et al. 2019), STA (Rao et al. 2018) and ST-reID (Wang et al. 2019a). It is worth noting that we omit using timestamp for ST-reID since it is unavailable and meaningless for long-term person re-ID. We re-train these methods on our long-term CVID-reID dataset and report their CMC and mAP in Table 6.2.

It can be observed that the proposed SpTSkM outperforms SOTA methods with a large margin. In specific, it improves ResNet50TP by 8.2% (from 25.4% to 33.6%) at rank #1, 11.1% (from 43.0% to 54.1%) at rank #5, 9.0% (from 55.1% to 64.1%)

Table 6.2 : Performance comparison between SOTA video-based person re-ID methods and proposed SpTSkM on CVID-reID for long-term person re-ID (%).

Methods	Source	R-1	R-5	R-10	R-20	mAP
State-of-the-art Methods						
RCNN (McLaughlin et al. 2016)	CVPR'16	5.5	18.2	29.3	43.9	6.1
ResNet50TP (Gao and Nevatia 2018)	ArXiv'18	25.4	43.0	55.1	67.2	20.3
ResNet50TA (Gao and Nevatia 2018)	ArXiv'18	25.9	43.7	54.5	62.9	19.2
ResNet50RNN (Gao and Nevatia 2018)	ArXiv'18	27.1	42.1	52.2	64.4	21.0
ResNet503D (Gao and Nevatia 2018)	ArXiv'18	9.9	20.4	28.2	41.2	7.7
M3D (Li et al. 2019)	AAAI'19	16.3	37.0	49.5	64.6	10.8
STA (Rao et al. 2018)	AAAI'19	4.9	13.5	21.5	39.8	4.7
ST-reID (Wang et al. 2019a)	AAAI'19	14.8	21.6	26.8	31.2	14.4
SpTSkM	**Ours**	<u>33.6</u>	**54.1**	**64.1**	**74.7**	**24.6**

at rank #10, 7.5% (from 67.2% to 74.7%) at rank #20. It increases ResNet50TP by 4.3% (from 20.3% to 24.6%) in terms of mAP. ResNet50RNN achieves best performance in SOTA methods, which reached 27.1% at rank #1 and 21.0% in terms of mAP. It exploits temporal dependences of a walking sequence using RNN. However, it is not tailor-made for long-term person re-ID application, which cannot completely explore subtle identity properties. In comparison, the proposed SSIN overcomes the drawback, which surpasses ResNet50RNN by 7.5% (from 27.1% to 34.6%) at rank #1 and 3.4% (from 21.0% to 24.4%) in terms of mAP.

In another aspect, it is harmful if the model pays much attention to model temporal cues and neglects these subtle identity cues in long-term person re-ID. For example, methods using 3D convolution, i.e., ResNet503D and M3D, achieves much inferior performance. This is because 1) quality of temporal information is seriously affected by viewpoint and pose variation, which causes temporal patterns unstable and 2) subtle identity properties in appearance is easily hidden if it is not well exploited. In comparison, the proposed SpTSkM is particularly designed for long-

term person re-ID, which involves a dual stream architecture: SSIN models set-based subtle identities, and SMIN learns temporal motion from skeleton sequences. Especially, SSIN learns representations from three sources, i.e., image frames, masks and videos, which correspond to image branch, mask branch and global branch, respectively. It utilizes cross-attention module to enhance the subtle identity attributes, which significantly boosts the matching performance. This is further discussed in Sec. 6.4.6. The SMIN focuses on motion encoding, which supplements the additional pure motion characteristics.

6.4.5 Robustness for CST-reID

This section evaluates the robustness of the proposed SpTSkM for CST-reID. We compare our SpTSkM with SOTA video-based approaches published in prevailing conferences and journals recently on video-based benchmark MARS. It is worth to point out that the selected SOTA methods almost cover popular ideas of temporal aggregation for video-based person re-ID, such as temporal pooling (Gao and Nevatia 2018; Subramaniam et al. 2019), temporal attention (Xu et al. 2017; Chen et al. 2018a; Gao and Nevatia 2018), RNN (Gao and Nevatia 2018), 3D convolution (Gao and Nevatia 2018; Li et al. 2019) and their combinations (Zhou et al. 2017c; Wu et al. 2018). In recent years, hand-crafted features are still adopted in different works (You et al. 2016; Zhang et al. 2018b). One of the representative works, (Zhang et al. 2018b), is also chosen in our comparison. For the proposed SpTSkM, we extracted representations by SSIN subnet, in order to better exploit appearance patterns in CST-reID. The results are listed in 6.3.

From the table, it can be observed that our SpTSkM also achieves promising performance on CST-reID. It improves baseline ResNet50TP by 2.4% (from 83.3% to 85.7%) in terms of Rank #1 and 4.7% (from 76.5% to 81.2%) in terms of mAP, respectively. Moreover, it also outperforms recent COSAM in terms of Rank #1 by

Table 6.3 : Performance comparison between SOTA video-based person re-ID methods and proposed SpTSkM on MARS for short-term person re-ID (%). Best and second-best results are highlighted by bold and underline, respectively.

Methods	Source	R-1	R-5	R-10	R-20	mAP
State-of-the-art Methods						
ASTPN (Xu et al. 2017)	ICCV'17	44.0	70.0	74.0	81.0	-
Zhou et al. (Zhou et al. 2017c)	CVPR'17	70.6	90.0	-	97.6	50.7
Zhang et al. (Zhang et al. 2018b)	TCSVT'18	55.5	70.2	-	80.2	-
Wu et al. (Wu et al. 2018)	TMM'18	69.7	83.4	88.3	93.6	-
Li et al. (Li et al. 2018a)	CVPR'18	82.3	-	-	-	65.8
ResNet50TP (Gao and Nevatia 2018)	ArXiv'18	83.3	93.0	95.3	96.8	76.5
ResNet50TA (Gao and Nevatia 2018)	ArXiv'18	83.3	93.8	<u>96.0</u>	97.4	76.7
ResNet50RNN (Gao and Nevatia 2018)	ArXiv'18	81.6	92.8	94.7	96.3	73.9
ResNet503D (Gao and Nevatia 2018)	ArXiv'18	78.5	90.9	93.9	95.9	70.5
Zhang et al. (Zhang et al. 2019b)	TNNLS'19	83.1	91.3	-	-	69.9
Liu et al. (Liu et al. 2019)	AAAI'19	84.4	93.2	-	96.3	72.7
M3D (Li et al. 2019)	AAAI'19	81.0	-	-	-	70.0
COSAM (Subramaniam et al. 2019)	ICCV'19	<u>84.9</u>	**95.5**	-	**97.9**	<u>79.9</u>
Proposed Methods						
SpTSkM	**Ours**	**85.7**	<u>95.2</u>	**97.0**	**97.9**	**81.2**

0.8% (from 84.9% to 85.7%) and mAP by 1.3% (from 79.9% to 81.2%). In another aspect, it surpasses hand-crafted feature, (Zhang et al. 2018b), by a large margin. It increases (Zhang et al. 2018b) from 55.5% to 85.7% in terms of Rank #1. It demonstrates the effectiveness of learning-based methods and proposed SpTSkM. In all, the results demonstrate that the proposed SpTSkM is also robust for CST-reID, in addition to the challenging CLT-reID.

6.4.6 Ablation Study

In this section, we perform ablation studies to verify the effectiveness of the proposed modules. In specific, we conduct analysis of each stream to show the influence of different components.

Ablation Study of SSIN

This section conducts ablation study of SSIN stream. In general, SSIN includes three branches, i.e., image branch, mask branch and global branch. The image branch takes image frames as input and produces representation by pooling frame-level embeddings. The mask branch counterpart takes corresponding masks as input and generates representation in the same aggregation way. The global branch refers to the attentive video-level stream, which considers video feature maps as input. In addition, the global branch is related to the cross-attention module, e.g., 2DCAM and 3DCAM, whose differences have been discussed in Sec. 6.2.2. Table 6.4 compares results in terms of these components.

In the table, the first three groups, i.e., (a) Img, (b) Mask, (c) Glob, report results of the three corresponding branches, respectively. It can be observed that global branch achieves the highest accuracy, i.e., 32.4% at rank #1, which surpasses image branch by 4.1% and mask branch by 14.3%. Accordingly, it also achieves the highest mAP, i.e., 22.5%, which exceeds image branch by 1.6% and mask branch by 6.5%. It is reasonable because the global branch gathers mask-guided attentive features in a video. It can better exploit subtle identity properties in terms of geometric body structure and shape and suppress background noise. As there is mask corruption, it loses too much useful information and causes inferior performance than image branch. However, the three branches may include some unique attributes that are beneficial to person re-ID. Thus, the combination of these three branches further boosts identification performance. For example, it improves image branch by 6.3% (from 28.3% to 34.6%), mask branch by 16.5% (from 18.1% to 34.6%) and global branch by 2.2% (from 32.4% to 34.6%) at rank #1. Accordingly, it increases mAP by 3.5% (from 20.9% to 24.4%), 8.4% (from 16.0% to 24.4%) and 1.9% (from 22.5% to 24.4%) in terms of image branch, mask branch and global branch, respectively. In addition, the last two groups compare different kinds of attention mechanism, i.e,

Table 6.4 : Ablation study of SSIN stream on CVID-reID dataset (%).

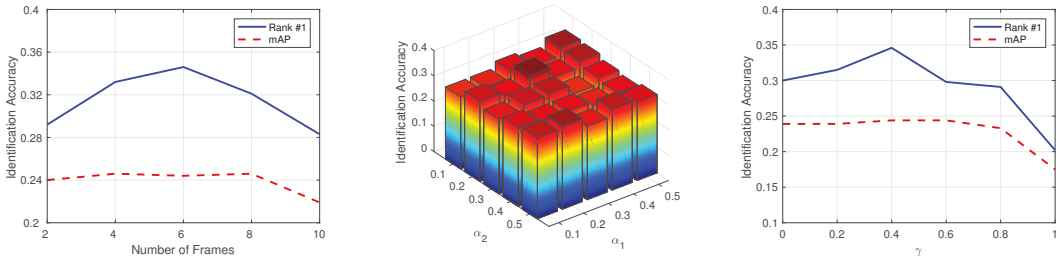
Group	SSIN					Performance		
	Img	Mask	Glob	2DCAM	3DCAM	R-1	R-20	mAP
a	✓					28.3	70.4	20.9
b		✓				18.1	71.7	16.0
c			✓	✓		32.4	71.3	22.5
d	✓	✓	✓	✓		34.6	74.3	24.4
e	✓	✓	✓		✓	32.9	73.8	25.2

Table 6.5 : Ablation Study of Set Aggregation Method (%).

Group	Set Aggregation			Performance			
	Max	Mean	Median	R-1	R-5	R-20	mAP
a	✓			34.6	53.1	74.3	24.4
b		✓		32.4	52.2	73.9	25.5
c			✓	33.8	50.7	71.6	25.2
d	✓	✓	✓	29.2	47.3	71.8	23.0

2DCAM and 3DCAM. It can be observed that 2DCAM gains slightly better scores in rank #1 and rank #20 than 3DCAM but lower mAP. However, the performance gap is trivial. It proved that 2DCAM is capable to capture the relationship of different body parts. Though 3DCAM can capture pixel dependencies from both intra- and inter- frames, it improves the overall mAP marginally from 24.4% to 25.2%. Thus, we use 2DCAM in our experiment in order to save memory.

Moreover, we studied the impact of different set aggregation methods in SSIN. Particularly, we evaluated three pooling approaches and the fusion of them as illustrated in Eq. 6.5. The corresponding results are listed in Table 6.5. We can observe that max pooling achieves promising performance in most cases. As claimed in Sec. 6.2.2, we use max pooling as default unless otherwise specified.



(a) Number of Frames (b) α_i (c) γ

Figure 6.10 : Analysis of identification performance versus hyper-parameters (a) Number of Frames in a training tracklet, (b) α_i in Eq. 6.7, and (c) γ in Eq. 6.9.

Table 6.6 : Ablation Study of SMIN on CVID-reID (%).

Method	R-1	R-5	R-10	R-20	mAP
SMIN (*w/o* 3D-KNM)	7.1	21.4	32.2	46.5	6.6
SMIN	10.6	26.7	39.2	53.2	7.5

Ablation Study of SMIN

This section investigates the effectiveness of the proposed 3D skeleton normalization module (3D-KNM). We conducted experiments on CVID-reID by considering the cases without and with 3D-KNM in SMIN. The results are listed in Table 6.6. It is easy to observe that the performance of SMIN significantly increases when 3D-KNM is utilized before motion modelling. In specific, it improves rank-1 from 7.1% to 10.6% and increases mAP from 6.6% to 7.5%. This verifies our assumption that view difference degrades matching performance. On the other hand, it also demonstrates the effectiveness of the proposed 3D-KNM.

Ablation Study of Dual Stream Network

This section discusses the effectiveness of dual stream network for long-term person re-ID. As elaborated before, the proposed SpTSkM aims to capture time-gap

Table 6.7 : Ablation Study of Dual Stream Networks on CVID-ReID (%).

Methods	Source	R-1	R-5	R-10	R-20	mAP
SSIN (2D)	Ours	**34.6**	53.1	63.4	74.3	24.4
SSIN (3D)	Ours	32.6	51.2	61.9	73.0	23.8
SMIN	Ours	10.6	26.7	39.2	53.2	7.5
SpTSkM	Ours	33.6	**54.1**	**64.1**	**74.7**	**24.6**

stable patterns, which is fulfilled by a dual stream framework, e.g., SSIN and SMIN. We perform ablation study on the long-term CVID-reID dataset. The results are reported in Table 6.7.

It can be observed that both SSIN and SMIN are beneficial to person re-ID. In detail, SSIN is more capable of learning time-gap stable spatial-temporal patterns related to subtle identity properties while SMIN tries to explore pure motion pattern. However, SSIN achieves much better performance than SMIN. For example, SSIN achieves 34.6% at rank #1 and 24.4% mAP while SMIN only obtains 10.6% at rank #1 and 7.5% mAP on CVID-reID. We argue that this is caused by skeleton estimation error and representation limitation. As we know, it is hard to estimate motion from occluded videos. However, it replenishes missed action attributes in SSIN. As we can see, the fusion SpTSkM increases mAP from 24.4% to 24.6% comparing to SSIN on CVID-reID. And, it also improves matching accuracy in most ranks, such as rank #5, rank #10 and rank #20 on CVID-reID.

6.4.7 Parameter Analysis

This section investigates the influence of different parameters, such as the number of frames in a training sequence, trade-off parameters α_i and γ.

Influence of Frame Number

Figure 6.10(a) illustrates the rank #1 accuracy in terms of the number of frames in a training video tracklet. It can be observed that the performance is affected by the number of video frames. Particularly, the mAP curve keeps stable between 4 frames to 8 frames and suffers performance drop otherwise. In contrast, Rank #1 curve fluctuates significantly, which peak accuracy when we train our model with 6 frames in a tracklet. According to the analysis, we set the training tracklet length to 6 by default.

Influence of α_i

In this section, we evaluate the relationship between performance in terms of trade-off parameters $\alpha_i, i = 1, \cdots, 3$. From Table 6.4, it seems that the global branch is prone to providing more identity information, the image branch comes second, and the mask branch contains less. Figure 6.10(b) shows the performance histogram with changes of hyper-parameters α_1 and α_2. Since $\sum_{i=1}^{3} \alpha_i = 1$, the histogram reflects relationship between performance and the three branches. From the figure, we derive that prior performance can be achieved if α_3 approaches 0.5. Accordingly, we balance the contribution of image branch and mask branch. It shows stable performance if α_1 is approximately equal to α_2. In our experiment, we empirically set $\alpha_1 = 0.3$ and $\alpha_2 = 0.2$.

Influence of γ

Figure 6.10(c) describes the performance curves varying in terms of the trade-off parameter γ. From the figure, we can observe that mAP improves gradually along with the increase of γ until reaching saturation at 0.4. After that, th e mAP keeps stable until γ is larger than 0.6, which the corresponding mAP drops drastically. In another aspect, the Rank #1 accuracy presents an ascending trend with the increase

of γ. It reaches peak value while γ equals to 0.4. After than, the Rank #1 accuracy decreases drastically. According to the two evaluation metrics, we set γ to 0.4 in our experiment.

6.5 Summary

This chapter proposes a dual-stream framework, SpTSkM, to tackle the challenging CST-reID problem. Particularly, it includes two main branches: SSIN and SMIN. The SSIN explores subtle identity properties involving geometric structure, spatial-temporal information, etc. To realize the purpose, we propose to learn representation from both image-level by weaving non-local cross-attention module into SSIN in order to highlight discriminative attributes and video-level by temporal set pooling in order to learn information temporally. The SMIN is tailored for motion description which remedies the defect of SSIN on temporal motion extraction. To achieve this, we propose to use ST-GCN to model normalized 3D skeleton sequences. By combining them, the approach demonstrates its effectiveness on re-ID tasks, especially the long-term case. Moreover, this work collects a dataset, i.e., CVID-reID, particularly for the long-term person re-ID problem. As far as we know, it is the largest LTG-reID dataset up to now. Extensive experiments on our CVID-reID and benchmark MARS demonstrate the effectiveness of the proposed SpTSkM for CLT-reID and also robustness for the CST-reID.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Person re-ID has been a hot research topic during the past decade. Previous works mainly restrict it as a short-term event while the long-term case is merely discussed. In practice, the person of interest is highly likely to re-appear in the surveillance network after a long period. This thesis focuses on the CLT-reID task. According to practical application scenarios, CLT-reID includes two typical cases: LTG-reID and CCM-reID. In addition to challenges in CST-reID, such as illumination variation, viewpoint difference, cluttered background and pose indeterminacy, the former suffers additional appearance changes, and the latter involves extra camera modality bias. Correspondingly, this thesis proposes to tackle them from five aspects, e.g., pure motion estimation, view bias mitigation, cross modality-matching, hybrid feature learning and data collection.

Chapter 3 proposes to address the LTG-reID problem by true motion estimation from raw video sequences. The proposed model characters motion patterns by the trajectory-aligned descriptors in a three-level body-action pyramid and encodes these descriptors by Fish Vector. It fills the research blank in the filed of CLT-reID. In addition, this chapter contributes a new dataset particular for LTG-reID, which includes videos captured by practical security cameras under the long-term re-ID setting. Comprehensive experiments demonstrate effectiveness of proposed motion-based representation on the LTG-reID task.

Chapter 4 focuses on the viewpoint problem in CVGLT-reID, a particular case

of LTG-reID. Two GAN-based methods, i.e., VN-GAN and VT-GAN, are proposed to mitigate view bias and improve identification accuracy. VN-GAN aims to normalize gaits from different views into a unified one using only one model. It is achieved by a coarse-to-fine strategy, which generates gaits from various views by introducing a view classifier to GAN in Stage-I and injects identity information by imposing an identity preserver into another GAN in Stage-II. VT-GAN aims to alleviate view difference by performing view to view transformation using only one model. By introducing viewpoint indicator to input and a discriminator as identity preserver, the proposed VT-GAN successfully achieved view to view gait transformation. Experimental results show that both methods achieved promising identification performance and good visual effect.

Chapter 5 tackles the CCM-reID problem and presents a top-push constrained modality-adaptive dictionary learning framework. Such a model jointly learns a common subspace by asymmetric mapping and a shared discriminative dictionary by top-push constrained dictionary learning. In the subspace, modality discrepancy is mitigated and discriminative ability is improved. Experiments on two CCM-reID datasets show the effectiveness of the proposed method.

In Chapter 6, a data-driven hybrid representation learning framework is proposed. It includes two main branches: SSIN and SMIN. The SSIN explores subtle identity properties involving geometric structure, spatial-temporal information, etc. It is achieved by weaving non-local cross-attention module and temporal set pooling. The SMIN is tailored for motion description which remedies the defect of SSIN on temporal motion extraction. It is achieved by 3D skeleton normalization and spatial-temporal GCN, which explores human motion attribute from normalized 3D skeleton sequences. By combining both streams, the proposed representation demonstrates promising performance on person re-ID after long-time gap problem. In addition, we collected a large scale video-based dataset from Internet. It is built

particularly for long-term person re-ID, which is the largest and most challenging dataset up to now. Sufficient experiments not only demonstrate the challenge of the collected dataset but also show the effectiveness of the proposed framework.

In the thesis, chapter 3, 4 and 5 are supported by papers published on prevailing conferences and journals. Chapter 6 is developed by a submitted journal paper. All of these papers are listed on **List of Publications**.

7.2 Future Work

Recently, adversarial learning has been widely adopted to address the distribution discrepancy problem (Ganin and Lempitsky 2015; Tzeng et al. 2017; Shrivastava et al. 2017; Sankaranarayanan et al. 2018; Zhang et al. 2019c). In CLT-reID, it has been used to mitigate the effect of modality bias (Dai et al. 2018). However, it is a feature-level method that tackles the modality discrepancy and explores feature discriminability simultaneously by enforcing constraints on the feature-level patterns. Recently, dual-level methods (Wang et al. 2019c,b) attract extensive interest, which tries to alleviate modality difference in the image space and exploit discriminability in the feature space. However, these kinds of dual-level methods depend on image generation, which may introduce noise. For future work, the benefits of both techniques are expected to be considered. That is, a dual-level strategy is employed, which learns domain/modality-agnostic semantic space and explores feature discriminability jointly. However, it is expected to narrow modality discrepancy by adversarial learning in the semantic space rather than image space and further exploit discriminability in the latent space.

Bibliography

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

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

Ben, X., Gong, C., Zhang, P., Jia, X., Wu, Q. & Meng, W., 2019a, 'Coupled patch alignment for matching cross-view gaits', *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3142–3157.

Ben, X., Zhang, P., Lai, Z., Yan, R., Zhai, X. & Meng, W., 2019b, 'A general tensor representation framework for cross-view gait recognition', *Pattern Recognition*, vol. 90, pp. 87–98.

Ben, X., Zhang, P., Meng, W., Yan, R., Yang, M., Liu, W. & Zhang, H., 2016, 'On the distance metric learning between cross-domain gaits', *Neurocomputing*, vol. 208, pp. 153–164.

Bobick, A. F. & Johnson, A. Y., 2001, 'Gait recognition using static, activity-specific parameters', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , vol. Ipp. 423–430.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R., 1994, 'Signature

verification using a" siamese" time delay neural network', *Advances in Neural Information Processing Systems (NIPS)*, pp. 737–744.

Caba Heilbron, F., Escorcia, V., Ghanem, B. & Carlos Niebles, J., 2015, 'Activitynet: A large-scale video benchmark for human activity understanding', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970.

Chao, H., He, Y., Zhang, J. & Feng, J., 2019, 'Gaitset: Regarding gait as a set for cross-view gait recognition', *AAAI Conference on Artificial Intelligence (AAAI)*, , vol. 33pp. 8126–8133.

Chen, D., Li, H., Xiao, T., Yi, S. & Wang, X., 2018a, 'Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1169–1178.

Chen, G., Lin, C., Ren, L., Lu, J. & Zhou, J., 2019, 'Self-critical attention learning for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 9637–9646.

Chen, W., Chen, X., Zhang, J. & Huang, K., 2017, 'Beyond triplet loss: a deep quadruplet network for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 403–412.

Chen, Y., Zhu, X., Zheng, W. & Lai, J., 2018b, 'Person re-identification by camera correlation aware feature augmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 392–408.

Cheng, D., Chang, X., Liu, L., Hauptmann, A. G., Gong, Y. & Zheng, N., 2017, 'Discriminative dictionary learning with ranking metric embedded for person re-identification', *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 964–970.

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, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. & Choo, J., 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.

Chung, D., Tahboub, K. & Delp, E. J., 2017, 'A two stream siamese convolutional neural network for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1983–1991.

Csurka, G., Dance, C., Fan, L., Willamowski, J. & Bray, C., 2004, 'Visual categorization with bags of keypoints', *ECCV Workshop on Statistical Learning in Computer Vision*, pp. 1–2.

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

Dalal, N. & Triggs, B., 2005, 'Histograms of oriented gradients for human detection', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , vol. 1pp. 886–893.

Dalal, N., Triggs, B. & Schmid, C., 2006, 'Human detection using oriented histograms of flow and appearance', *European Conference on Computer Vision (ECCV)*, pp. 428–441.

Dehghan, A., Modiri Assari, S. & Shah, M., 2015, 'Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking',

The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4091–4099.

Deng, W., Zheng, L., Kang, G., Yang, Y., Ye, Q. & 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.

Diederik, P. K., Welling, M. et al., 2014, 'Auto-encoding variational bayes', *The International Conference on Learning Representations (ICLR)*, .

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.

Everitt, B. S., 2014, 'Finite mixture distributions', *Wiley StatsRef: Statistics Reference Online*.

Fang, P., Zhou, J., Roy, S. K., Petersson, L. & Harandi, M., 2019, 'Bilinear attention networks for person retrieval', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 8030–8039.

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)*, IEEE, pp. 2360–2367.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D., 2009, 'Object detection with discriminatively trained part-based models', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645.

Ganin, Y. & Lempitsky, V., 2015, 'Unsupervised domain adaptation by backpropagation', *International Conference on Machine Learning (ICML)*, pp. 1180–1189.

Gao, J. & Nevatia, R., 2018, 'Revisiting temporal modeling for video-based person reid', <https://arxiv.org/abs/1805.02104>.

Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X. et al., 2018, 'Fd-gan: Pose-guided feature distilling gan for robust person re-identification', *Advances in Neural Information Processing Systems (NIPS)*, pp. 1222–1233.

Girshick, R., 2015, 'Fast r-cnn', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448.

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*, pp. 2672–2680.

Gou, M., Zhang, X., Rates-Borras, A., Asghari-Esfeden, S., Sznaier, M. & Camps, O., 2016, 'Person re-identification in appearance impaired scenarios', *British Machine Vision Conference (BMVC)*, pp. 48.1–48.14.

Gray, D., Brennan, S. & Tao, H., 2007, 'Evaluating appearance models for recognition, reacquisition, and tracking', *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, Citeseer, pp. 1–7.

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.

Hadsell, R., Chopra, S. & LeCun, Y., 2006, 'Dimensionality reduction by learning an invariant mapping', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , vol. 2IEEE, pp. 1735–1742.

Han, J. & Bhanu, B., 2005, 'Individual recognition using gait energy image', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322.

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., 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.

He, Y., Zhang, J., Shan, H. & Wang, L., 2019, 'Multi-task gans for view-specific feature learning in gait recognition', *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 102–113.

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

Hirzer, M., Beleznai, C., Roth, P. M. & Bischof, H., 2011, 'Person re-identification by descriptive and discriminative classification', *Scandinavian Conference on Image Analysis (SCIA)*, Springer, pp. 91–102.

Hirzer, M., Roth, P. M., Köstinger, M. & Bischof, H., 2012, 'Relaxed pairwise learned metric for person re-identification', *European Conference on Computer Vision (ECCV)*, Springer, pp. 780–793.

Hu, M., Wang, Y., Zhang, Z., Little, J. J. & Huang, D., 2013, 'View-invariant discriminative projection for multi-view gait-based human identification', *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2034–2045.

Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z. & Zhang, J., 2019, 'Multi-pseudo regularized label for generated data in person re-identification', *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1391–1403.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. & Schiele, B., 2016,

'Deepercut: A deeper, stronger, and faster multi-person pose estimation model', *European Conference on Computer Vision (ECCV)*, Springer, pp. 34–50.

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.

Iwama, H., Okumura, M., Makihara, Y. & Yagi, Y., 2012, 'The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition', *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521.

Jaderberg, M., Simonyan, K., Zisserman, A. et al., 2015, 'Spatial transformer networks', *Advances in Neural Information Processing Systems (NIPS)*, pp. 2017–2025.

Karanam, S., Li, Y. & Radke, R. J., 2015, 'Person re-identification with discriminatively trained viewpoint invariant dictionaries', *The IEEE International Conference on Computer Vision (CVPR)*, pp. 4516–4524.

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

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M. & Bischof, H., 2012, 'Large scale metric learning from equivalence constraints', *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 2288–2295.

Kusakunniran, W., Wu, Q., Li, H. & Zhang, J., 2009, 'Multiple views gait recognition using view transformation model based on optimized gait energy image', *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1058–1064.

Kusakunniran, W., Wu, Q., Zhang, J. & Li, H., 2010, 'Support vector regression for multi-view gait recognition based on local motion feature selection', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 974–981.

Kusakunniran, W., Wu, Q., Zhang, J. & Li, H., 2012, 'Gait recognition under various viewing angles based on correlated motion regression', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 6, pp. 966–980.

Kusakunniran, W., Wu, Q., Zhang, J., Li, H. & Wang, L., 2013, 'Recognizing gaits across views through correlated motion co-clustering', *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 696–709.

Kusakunniran, W., Wu, Q., Zhang, J., Li, H. & Wang, L., 2014, 'Recognizing gaits across views through correlated motion co-clustering', *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 696–709.

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

Lee, H., Battle, A., Raina, R. & Ng, A. Y., 2007, 'Efficient sparse coding algorithms', *Advances in Neural Information Processing Systems (NIPS)*, pp. 801–808.

Li, B., Chang, H., Shan, S. & Chen, X., 2009, 'Coupled metric learning for face recognition with degraded images', *Advances in Machine Learning*, pp. 220–233.

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, J., Zhang, S. & Huang, T., 2019, 'Multi-scale 3d convolution network for

video based person re-identification', *AAAI Conference on Artificial Intelligence (AAAI)*, , vol. 33pp. 8618–8625.

Li, S., Bak, S., Carr, P. & Wang, X., 2018a, 'Diversity regularized spatiotemporal attention for video-based person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 369–378.

Li, S., Shao, M. & Fu, Y., 2015, 'Cross-view projective dictionary learning for person re-identification.', *The International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2155–2161.

Li, W. & Wang, X., 2013, 'Locally aligned feature transforms across views', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3594–3601.

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.

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.

Liu, H., Feng, J., Qi, M., Jiang, J. & Yan, S., 2017, 'End-to-end comparative attention networks for person re-identification', *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506.

Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S. & Hu, J., 2018, 'Pose transferrable person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4099–4108.

Liu, K., Ma, B., Zhang, W. & Huang, R., 2015, 'A spatio-temporal appearance representation for viceo-based pedestrian re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3810–3818.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. & Berg, A. C., 2016, 'Ssd: Single shot multibox detector', *European Conference on Computer Vision (ECCV)*, Springer, pp. 21–37.

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C. & Bu, J., 2014, 'Semi-supervised coupled dictionary learning for person re-identification', *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3550–3557.

Liu, Y., Yuan, Z., Zhou, W. & Li, H., 2019, 'Spatial and temporal mutual promotion for video-based person re-identification', *AAAI Conference on Artificial Intelligence (AAAI)*, , vol. 33pp. 8786–8793.

Lu, H., Plataniotis, K. N. & Venetsanopoulos, A. N., 2008, 'Mpca: Multilinear principal component analysis of tensor objects', *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 18–39.

Lu, J., Wang, G. & Zhou, J., 2017, 'Simultaneous feature and dictionary learning for image set based face recognition', *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4042–4054.

Ma, B. & Su, Y., 2012, 'Local descriptors encoded by fisher vectors for person re-identification', *Europe Conference Computer Vision on Workshops and Demonstrations*, Springer, pp. 413–422.

Ma, B., Su, Y. & Jurie, F., 2012, 'Bicov: a novel image representation for person re-identification and face verification', *British Machive Vision Conference (BMVC)*, pp. 11–pages.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T. & Van Gool, L., 2017, 'Pose guided person image generation', *Advances in Neural Information Processing Systems (NIPS)*, pp. 406–416.

Ma, L., Liu, H., Hu, L., Wang, C. & Sun, Q., 2016, 'Orientation driven bag of appearances for person re-identification', *arXiv preprint arXiv:1605.02464*.

Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T. & Yagi, Y., 2006, 'Gait recognition using a view transformation model in the frequency domain', *European Conference on Computer Vision (ECCV)*, Springer, pp. 151–163.

Mansur, A., Makihara, Y., Muramatsu, D. & Yagi, Y., 2014, 'Cross-view gait recognition using view-dependent discriminative analysis', *IEEE International Joint Conference on Biometrics (IJCB)*, .

Martín-Félez, R. & Xiang, T., 2012, 'Gait recognition by ranking', *European Conference on Computer Vision*, Springer, pp. 328–341.

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.

McLaughlin, N., Martinez del Rincon, J. & Miller, P., 2016, 'Recurrent convolutional network for video-based person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1325–1334.

Mignon, A. & Jurie, F., 2012, 'Pcca: A new approach for distance learning from sparse pairwise constraints', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 2666–2672.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B. & Mullers, K.-R., 1999, 'Fisher discriminant analysis with kernels', *IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, pp. 41–48.

Mirza, M. & Osindero, S., 2014, 'Conditional generative adversarial nets', *arXiv preprint arXiv:1411.1784*.

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

Munaro, M., Fossati, A., Basso, A., Menegatti, E. & Van Gool, L., 2014b, 'One-shot person re-identification with a consumer depth camera', *Person Re-Identification*, Springer, pp. 161–181.

Muramatsu, D., Makihara, Y. & Yagi, Y., 2016, 'View transformation model incorporating quality measures for cross-view gait recognition', *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1602–1615.

Ojala, T., Pietikainen, M. & Maenpaa, T., 2002, 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987.

Pedagadi, S., Orwell, J., Velastin, S. & Boghossian, B., 2013, 'Local fisher discriminant analysis for pedestrian re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3325.

Peng, P., Tian, Y., Xiang, T., Wang, Y., Pontil, M. & Huang, T., 2017, 'Joint semantic and latent attribute modelling for cross-class transfer learning', *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T. & Tian, Y., 2016, 'Unsupervised cross-dataset transfer learning for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1306–1315.

Perronnin, F., Liu, Y., Sánchez, J. & Poirier, H., 2010, 'Large-scale image retrieval with compressed fisher vectors', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 3384–3391.

Prosser, B. J., Zheng, W.-S., Gong, S., Xiang, T. & Mary, Q., 2010, 'Person re-identification by support vector ranking', *British Machine Vision Conference (BMVC)*, pp. 1–10.

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.-G. & Xue, X., 2018, 'Pose-normalized image generation for person re-identification', *European Conference on Computer Vision (ECCV)*, pp. 650–667.

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

Rao, S., Rahman, T., Rochan, M. & Wang, Y., 2018, 'Video-based person re-identification using spatial-temporal attention networks', <`https://arxiv.org/abs/1810.11261`>.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R. & Vasconcelos, N., 2010, 'A new approach to cross-modal multimedia retrieval', *ACM Multimedia Conference (ACM MM)*, ACM, pp. 251–260.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A., 2016, 'You only look once: Unified, real-time object detection', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788.

Ronneberger, O., Fischer, P. & Brox, T., 2015, 'U-net: Convolutional networks for biomedical image segmentation', *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, pp. 234–241.

Sánchez, J., Perronnin, F., Mensink, T. & Verbeek, J., 2013, 'Image classification with the fisher vector: Theory and practice', *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D. & Chellappa, R., 2018, 'Generate to adapt: Aligning domains using generative adversarial networks', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8503–8512.

Schroff, F., Kalenichenko, D. & Philbin, J., 2015, 'Facenet: A unified embedding for face recognition and clustering', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823.

Sharma, A., Kumar, A., Daume, H. & Jacobs, D. W., 2012, 'Generalized multiview analysis: A discriminative latent space', *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 2160–2167.

Shekhar, S., Patel, V. M., Nguyen, H. V. & Chellappa, R., 2013, 'Generalized domain-adaptive dictionaries', *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 361–368.

Shi, L., Zhang, Y., Cheng, J. & Lu, H., 2019, 'Two-stream adaptive graph convolutional networks for skeleton-based action recognition', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12026–12035.

Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T. & Yagi, Y., 2016, 'Geinet: View-invariant gait recognition using a convolutional neural network', *International Conference on Biometrics (IJB)*, .

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W. & Webb, R., 2017, 'Learning from simulated and unsupervised images through adversarial training', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2107–2116.

Son, J., Baek, M., Cho, M. & Han, B., 2017, 'Multi-object tracking with quadruplet convolutional neural networks', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5620–5629.

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.

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. 3960–3969.

Su, C., Zhang, S., Xing, J., Gao, W. & Tian, Q., 2016, 'Deep attributes driven multi-camera person re-identification', *European Conference on Computer Vision (ECCV)*, Springer, pp. 475–491.

Subramaniam, A., Nambiar, A. & Mittal, A., 2019, 'Co-segmentation inspired attention networks for video-based person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 562–572.

Suh, Y., Wang, J., Tang, S., Mei, T. & Mu Lee, K., 2018, 'Part-aligned bilinear representations for person re-identification', *The European Conference on Computer Vision (ECCV)*, pp. 402–419.

Sun, Y., Chen, Y., Wang, X. & Tang, X., 2014, 'Deep learning face representation by joint identification-verification', *Advances in Neural Information Processing Systems (NIPS)*, pp. 1988–1996.

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)*, pp. 480–496.

Taigman, Y., Yang, M., Ranzato, M. & Wolf, L., 2014, 'Deepface: Closing the gap to human-level performance in face verification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1701–1708.

Tao, D., Li, X., Wu, X. & Maybank, S. J., 2007, 'General tensor discriminant analysis and gabor features for gait recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700–1715.

Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T., 2017, 'Adversarial discriminative domain adaptation', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7167–7176.

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

Varior, R. R., Shuai, B., Lu, J., Xu, D. & Wang, G., 2016b, 'A siamese long short-term memory architecture for human re-identification', *European Conference on Computer Vision (ECCV)*, Springer, pp. 135–153.

Wang, G., Lai, J., Huang, P. & Xie, X., 2019a, 'Spatial-temporal person re-identification', *AAAI Conference on Artificial Intelligence (AAAI)*, , vol. 33pp. 8933–8940.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y. & Hou, Z., 2019b, 'Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment', *The IEEE International Conference on Computer Vision (CVPR)*, pp. 3623–3632.

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L., 2011, 'Action recognition by dense trajectories', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 3169–3176.

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, H. & Schmid, C., 2013, 'Action recognition with improved trajectories', *The IEEE international Conference on Computer Vision (ICCV)*, pp. 3551–3558.

Wang, T., Gong, S., Zhu, X. & Wang, S., 2016, 'Person re-identification by discriminative selection in video ranking', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2501–2514.

Wang, X., Girshick, R., Gupta, A. & He, K., 2018, 'Non-local neural networks', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803.

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.

Wei, L., Zhang, S., Yao, H., Gao, W. & Tian, Q., 2017, 'Glad: Global-local-alignment descriptor for pedestrian retrieval', *ACM International Conference on Multimedia (ACM MM)*, ACM, pp. 420–428.

Weinberger, K. Q. & Saul, L. K., 2009, 'Distance metric learning for large margin nearest neighbor classification', *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244.

Wu, A., Zheng, W.-S. & Lai, J.-H., 2017a, 'Robust depth-based person re-identification', *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2588–2603.

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

Wu, L., Wang, Y., Gao, J. & Li, X., 2018, 'Where-and-when to look: Deep siamese attention networks for video-based person re-identification', *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1412–1424.

Wu, Z., Huang, Y., Wang, L., Wang, X. & Tan, T., 2017c, 'A comprehensive study on cross-view gait based human identification with deep cnns', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 209–226.

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.

Xing, X., Wang, K., Yan, T. & Lv, Z., 2016, 'Complete canonical correlation analysis with application to multi-view gait recognition', *Pattern Recognition*, vol. 50, pp. 107–117.

Xiong, F., Gou, M., Camps, O. & Sznaier, M., 2014, 'Person re-identification using kernel-based metric learning methods', *European Conference on Computer Vision (ECCV)*, Springer, pp. 1–16.

Xu, J., Zhao, R., Zhu, F., Wang, H. & Ouyang, W., 2018, 'Attention-aware compositional network for person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2119–2128.

Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S. & Zhou, P., 2017, 'Jointly attentive spatial-temporal pooling networks for video-based person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4733–4742.

Yan, S., Xiong, Y. & Lin, D., 2018, 'Spatial temporal graph convolutional networks for skeleton-based action recognition', *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7444–7452.

Yang, F., Yan, K., Lu, S., Jia, H., Xie, X. & Gao, W., 2019, 'Attention driven person re-identification', *Pattern Recognition*, vol. 86, pp. 143–155.

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

Ye, M., Wang, Z., Lan, X. & Yuen, P. C., 2018c, 'Visible thermal person re-identification via dual-constrained top-ranking.', *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1092–1099.

Ye, M., Wang, Z., Lan, X. & Yuen, P. C., 2018d, 'Visible thermal person re-identification via dual-constrained top-ranking', *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 1092–1099.

You, J., Wu, A., Li, X. & Zheng, W.-S., 2016, 'Top-push video-based person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1345–1353.

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

Yu, S., Chen, H., Reyes, G., Edel, B. & Poh, N., 2017b, 'Gaitgan: invariant gait feature extraction using generative adversarial networks', *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 30–37.

Yu, S., Chen, H., Wang, Q., Shen, L. & Huang, Y., 2017c, 'Invariant feature extraction for gait recognition using only one uniform model', *Neurocomputing*, vol. 239, pp. 81–93.

Yu, S., Liao, R., Weizhi, A., Haifeng, C., García, E. B., Huang, Y. & Poh, N., 2019, 'Gaitganv2: Invariant gait feature extraction using generative adversarial networks', *Pattern Recognition*, vol. 87, pp. 179–189.

Yu, S., Tan, D. & Tan, T., 2006, 'A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition', *International Conference on Pattern Recognition (ICPR)*, , vol. 4IEEE, pp. 441–444.

Zhang, D. & Li, W.-J., 2014, 'Large-scale supervised multimodal hashing with semantic correlation maximization.', *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2177–2183.

Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A., 2019a, 'Self-attention generative adversarial networks', *International Conference on Machine Learning (ICML)*, pp. 7354–7363.

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, P., Ben, X., Jiang, W., Yan, R. & Zhang, Y., 2015, 'Coupled marginal discriminant mappings for low-resolution face recognition', *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 23, pp. 4352–4357.

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

Zhang, W., He, X., Lu, W., Qiao, H. & Li, Y., 2019b, 'Feature aggregation with reinforcement learning for video-based person re-identification', *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3847–3852.

Zhang, W., Hu, S., Liu, K. & Zha, Z., 2018b, 'Learning compact appearance representation for video-based person re-identification', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2442–2452.

Zhang, Y., Tang, H., Jia, K. & Tan, M., 2019c, 'Domain-symmetric networks for adversarial domain adaptation', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5031–5040.

Zhao, G., Liu, G., Li, H. & Pietikainen, M., 2006, '3d gait recognition using multiple cameras', *International Conference on Automatic Face and Gesture Recognition (FGR)*, pp. 529–534.

Zhao, L., Li, X., Zhuang, Y. & Wang, J., 2017, 'Deeply-learned part-aligned representations for person re-identification', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3219–3228.

Zhao, R., Ouyang, W. & Wang, X., 2013, 'Person re-identification by salience matching', *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2528–2535.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S. & Tian, Q., 2016a, 'Mars: A video benchmark for large-scale person re-identification', *European Conference on Computer Vision (ECCV)*, Springer, pp. 868–884.

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., 2016b, 'Person re-identification: Past, present and future', *arXiv preprint arXiv:1610.02984*.

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

Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G. & Cai, D., 2011a, 'Graph regularized sparse coding for image representation', *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1327–1336.

Zheng, S., Zhang, J., Huang, K., He, R. & Tan, T., 2011b, 'Robust view transformation model for gait recognition', *IEEE International Conference on Image Processing (ICIP)*, pp. 2073–2076.

Zheng, W.-S., Gong, S. & Xiang, T., 2016c, 'Towards open-world person re-identification by one-shot group-based verification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 591–606.

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.

Zhou, Q., Zheng, S., Ling, H., Su, H. & Wu, S., 2017a, 'Joint dictionary and metric learning for person re-identification', *Pattern Recognition*, vol. 72, pp. 196–206.

Zhou, X., Huang, Q., Sun, X., Xue, X. & Wei, Y., 2017b, 'Towards 3d human pose estimation in the wild: a weakly-supervised approach', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 398–407.

Zhou, Z., Huang, Y., Wang, W., Wang, L. & Tan, T., 2017c, 'See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification', *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4747–4756.

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W. & Yang, M.-H., 2018a, 'Online multi-object tracking with dual matching attention networks', *European Conference on Computer Vision (ECCV)*, pp. 366–382.

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.

Zhu, X., Jing, X.-Y., You, X., Zhang, X. & Zhang, T., 2018b, 'Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics', *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5683–5695.