

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

by Yan Huang

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Qiang Wu

University of Technology Sydney
Faculty of Engineering and Information Technology

05/2021

Certificate of Original Authorship Template

CERTIFICATE OF ORIGINAL AUTHORSHIP

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

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

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

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

Signature: Production Note:
Signature removed prior to publication.

Date: 21/05/2021

ABSTRACT

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

by

YAN HUANG

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

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

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

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

Dissertation directed by Associate Professor Qiang Wu

School of Electrical and Data Engineering

Acknowledgements

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

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

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

cutting-edge techniques of computer vision.

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

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

Yan Huang
Sydney, Australia, 2021.

List of Publications

Journal Papers

- J-1. **Y. Huang**, Q. Wu, J. Xu, Y. Zhong and Z Zhang. “Unsupervised Domain Adaptation with Background Shift Mitigating for Person Re-Identification,” *Springer International Journal of Computer Vision (IJCV)*, 2021.
- J-2. **Y. Huang**, Q. Wu, J. Xu, Y. Zhong, P. Zhang and Z. Zhang. “Alleviating Modality Bias Training for Infrared-Visible Person Re-Identification,” *IEEE Transactions on Multimedia (TMM)*, accepted, 2021.
- J-3. **Y. Huang**, J. Xu, Q. Wu, Y. Zhong, P. Zhang and Z. Zhang, “Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 30, no. 10, pp. 3459-3471, 2020.
- J-4. **Y. Huang**, J. Xu, Q. Wu, Z. Zheng, Z. Zhang and J. Zhang, “Multi-Pseudo Regularized Label for Generated Data in Person Re-Identification,” *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 3, pp. 1391-1403, 2019.
- J-5. **Y. Huang**, Y. Zhong, Q. Wu, E. Dutkiewicz and T. Jiang, “Cost-effective foliage penetration human detection under severe weather conditions based on auto-encoder/decoder neural network,” *IEEE Internet of Things Journal (IoTJ)*, vol. 6, no. 4, pp. 6190-6200, 2019.
- J-6. P. Zhang, J. Xu, Q. Wu, **Y. Huang** and X. Ben, “Learning Spatial-temporal Representations over Walking Tracklet for Long-term Person Re-Identification in The Wild,” *IEEE Transactions on Multimedia (TMM)*, 2020.
- J-7. L. Huang, Q. Yang, J. Wu, **Y. Huang**, Q. Wu and J. Xu, “Generated Data With Sparse Regularized Multi-Pseudo Label for Person Re-Identification,”

- IEEE Signal Processing Letters (SPL)*, vol. 27, pp. 391-395, 2020.
- J-8. P. Zhang, J. Xu, Q. Wu, **Y. Huang** and J. Zhang, “Top-Push Constrained Modality-Adaptive Dictionary Learning for Cross-Modality Person Re-Identification,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2020.
- J-9. Y. Zhong, Y. Yang, X. Zhu, **Y. Huang***, E. Dutkiewicz, Z. Zhou and T. Jiang, “Impact of Seasonal Variations on Foliage Penetration Experiment: A WSN-Based Device-Free Sensing Approach,” *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 56, no. 9, pp. 5035-5045, 2018.
- J-10. D. Liu, **Y. Huang**, Q. Wu, R. Ma and P. An. “Multi-Angular Epipolar Geometry based Light Field Angular Reconstruction Network,” *IEEE Transactions on Computational Image (TCI)*, 2020.
- J-11. Y. Zhong, S. Wu, J. Wang, T. Jiang, **Y. Huang*** and Q. Wu. “Multi-Location Human Activity Recognition via MIMO-OFDM Based Wireless Networks: An IoT-Inspired Device-Free Sensing Approach,” *IEEE Internet of Things Journal (IoTJ)*, 2020.
- J-12. J. Wu, **Y. Huang***, Q. Wu, Z. Gao, J. Zhao and L. Huang. “Dual-Stream Guided-Learning Via Priori Optimization for Person Re-identification,” *Springer Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, accepted, 2020.

Conference Papers

- C-1. **Y. Huang**, Q. Wu, J. Xu and Y. Zhong, “SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification,” *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 9526-9535, Oct, 2019.
- C-2. **Y. Huang**, Q. Wu, J. Xu and Y. Zhong, “Celebrities-ReID: A Benchmark for Clothes Variation in Long-Term Person Re-Identification,” *Proc. IEEE International Joint Conference on Neural Network (IJCNN)*, pp. 1-8, July, 2019.

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

Submitted Papers

- J-1. **Y. Huang**, Q. Wu, J. Xu, Y. Zhong, P. Zhang and Z Zhang. “Learning Directly from Style Invariant Component for Infrared-Visible Person Re-Identification,” *IEEE Transactions on Cybernetics*, 2020 (Under review).
- J-2. X. Ding, T. Jiang, J. Yang, Y. Zhong, **Y. Huang** and S. Wu. “WiLISensing: Device-free Location-independent Human Activity Recognition via Wi-Fi,” *IEEE Internet of Things Journal*, 2020 (Under review).

Contents

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	vii
List of Figures	xv
List of Tables	xxiii
Abbreviation	xxvii
1 Introduction	1
1.1 Background	1
1.1.1 Challenges on Variations of Illumination, Viewpoint, And Pose in ST-reID	3
1.1.2 Challenges on Footages Taken by Different Environments in ST-reID	6
1.1.3 Challenges on Clothing Change When A Person Being Identified in LT-reID	8
1.1.4 Challenges on Footages Taken by Different Styles of Cameras in LT-reID	10
1.2 Research Problems	11
1.3 Thesis Contribution	13
1.4 Thesis Organization	16

2	Literature Review	18
2.1	Review of Single-Domain ST-reID	18
2.1.1	Non-Deep Learning Approaches	18
2.1.2	Deep Learning Approaches	19
2.1.3	Boosting Tricks	20
2.2	Review of Cross-Domain ST-reID	21
2.2.1	General Inter-Domain Style Transfer for Cross-Domain ST-reID	21
2.2.2	Clustering-Based Cross-Domain ST-reID	22
2.3	Review of Clothing Change LT-reID	23
2.3.1	Biometrics Based Approaches	23
2.3.2	Depth Information Compensation Approaches	24
2.4	Review of Infrared-Visible LT-reID	25
2.4.1	Feature Alignment	25
2.4.2	Feature Alignment + Data Alignment	26
2.5	Datasets for Person Re-ID In The Wild	27
2.5.1	Existing Datasets for ST-reID	27
2.5.2	Existing Datasets for Clothing Change LT-reID	29
2.5.3	Existing Datasets for Infrared-Visible LT-reID	29
2.6	Summary	30
3	Labelling Generated Data for Single-Domain ST-ReID	31
3.1	Motivation	31
3.2	Labeling Generated Data Using Multi-Pseudo Regularized Label . . .	34
3.2.1	LSRO for Person Re-ID	35
3.2.2	Multi-Pseudo Regularized Label	36

3.2.3	Training Strategy	38
3.3	Benefits of Multi-pseudo Regularized Label	40
3.3.1	One-Hot Binary Label vs. Multi-Association-Value Label	42
3.3.2	Single Unified Label vs. Multiple Different Virtual Labels	42
3.3.3	Different Weighing on Virtual Labels	43
3.4	Experiments	44
3.4.1	Person Re-ID Datasets	45
3.4.2	Experimental Setup	46
3.4.3	Baseline Performance	51
3.4.4	Performance Improved on The Identif Network by Using Generated Data	51
3.4.5	Performance Evaluation under Different Implementations of MpRL	53
3.4.6	Comparison with Existing Virtual Labeling Approaches	53
3.4.7	Performance Evaluation by Using Different GAN Models	57
3.4.8	Performance Comparison with The State-of-The-Art Methods	58
3.5	Conclusion	61
4	Background Shift Issue in Cross-Domain ST-ReID	62
4.1	Motivation	62
4.2	Overview of The Proposed Approach	67
4.3	SBSGAN for Generating Soft-Mask Image	68
4.3.1	Objective Functions in SBSGAN	70
4.3.2	Indicators for Data Generation	72
4.4	DA-2S	73

4.4.1	Initial DA-2S Training	73
4.4.2	DA-2S Update	75
4.5	Experiments	79
4.5.1	Datasets for Evaluations	80
4.5.2	Evaluation Criteria	80
4.5.3	Implementation Details	80
4.5.4	Qualitative Evaluation	82
4.5.5	Quantitative Evaluation	87
4.6	Conclusion	94
5	New Benchmarks And Solutions for Clothing Change	
	LT-reID	96
5.1	Motivation	96
5.2	Celeb-reID: New Benchmarks	101
5.3	Vector-Neuron Capsules for Clothing Change LT-reID	104
5.3.1	Visual Feature Extraction Module	105
5.3.2	ID and Dressing Perception Module	105
5.3.3	Auxiliary Module	107
5.4	Experiments	108
5.4.1	Datasets for Evaluations	109
5.4.2	Experimental Setup	110
5.4.3	Ablation Study of ReIDCaps	111
5.4.4	Scalar Neuron <i>vs.</i> Vector Neuron with Quantitative and Qualitative Analyses	115
5.4.5	Comparison with State-of-The-Art Methods	117

5.5 Conclusion	124
6 Modality Bias Training Issue for Infrared-Visible LT-reID	125
6.1 Motivation	125
6.2 Alleviating MBT Issue via Dual-Level Learning Strategy	131
6.2.1 The First-Level Learning Strategy in DLS	131
6.2.2 The Second-Level Learning Strategy in DLS	134
6.3 Experiments	138
6.3.1 Infrared-visible LT-reID Datasets and Evaluation Metrics . . .	138
6.3.2 Implementation Details	140
6.3.3 Effectiveness of The First-Level Learning Strategy	141
6.3.4 Effectiveness of The Second-Level Learning Strategy	146
6.3.5 Using Other ImageNet-Trained Model for Evaluation	150
6.3.6 Comparison with State-of-the-Art Methods	151
6.4 Conclusion	154
7 Conclusions and Future Work	156
7.1 Conclusion	156
7.2 Future Work	158

List of Figures

1.1	An example of a person re-ID application in the wild. Two disjoint public places are simultaneously monitored by camera A and B, respectively.	1
1.2	Three persons under different camera views. Large variations such as illumination, viewpoint, and pose can be observed.	3
1.3	Schematic diagram of cross-domain ST-reID. It can be seen that compared with the single-domain ST-reID issue in which the training and testing datasets are the same, the performance is dropped a lot when testing is conducted on another dataset.	5
1.4	Comparison between samples in ST-reID dataset (<i>i.e.</i> , Market-1501 (Zheng et al., 2015)) and clothing change LT-reID dataset. For each dataset, only one ID is shown in the figure.	8
1.5	Examples of infrared-visible LT-reID. Even human performance is hard to distinguish the ID when using RGB images as query and IR images as gallery.	10
1.6	Two main scenarios in person re-ID. Categorizing according to the time gaps of a person reappears under different cameras. For each category, two important challenges need to be addressed for person re-ID in the wild.	13
1.7	Thesis structure.	16

- 3.1 Label distribution of predefined training classes (c) for generated images (a) and (b). Only the maximum predicted probability of predefined training classes is activated along with the training process (see (c)). Distinguishable label distributions can be observed between (a) and (b). 33
- 3.2 The label distributions of real and generated data. The ground-truth label is assigned to the real data (a). For a generated image, all-in-one (b) assigns a new label to it. One-hot pseudo (c) uses only one predefined training class with maximum predicted probability. LSRO (d) uses a single unified label distribution, while the proposed MpRL (e) considers different weights for different predefined training classes. 41
- 3.3 Examples of generated data and their corresponding representations in the real data domain. The left side shows four generated data with distinct visual differences (in red, yellow, white and green clothes). For each generated data, the right side gives ten nearest representations which represent each predefined training class in the real data domain. 43
- 3.4 Examples of generated (by DCGAN (Radford et al., 2015)) and real person images. (a)-(d) show the generated person images (first two rows) and real person images (the third row) on Market-1501, DukeMTMC-reID, CUHK03, VIPeR, and CUHK01, respectively. Note that all fake images do not belong to any of IDs in real data. . . 47
- 3.5 (a) is the Identif network presented in (Zheng et al., 2016a,b), (b) is the Two-stream network introduced in (Zheng et al., 2018). Both networks use resnet-50 as a basic component of CNN. 49

3.6	Examples of generated and real person images. (a)-(c) show the generated person images (first two rows) and real person images (the third row) on Market-1501, DukeMTMC-reID, and VIPeR respectively. Images in the first and second rows are respectively generated by the WGAN-GP (Gulrajani et al., 2017) and the DCGAN (Radford et al., 2015).	57
4.1	Comparison between different input images for cross-domain ST-reID. Images from Market-1501 and DukeMTMC-reID show distinct background shift. Images generated by SPGAN (Deng et al., 2018) and PTGAN (Wei et al., 2018a) do not suppress the background noise and have the background shift problem. The hard-mask solution, <i>i.e.</i> , JPPNet (Liang et al., 2018) damages the foreground. The proposed SBSGAN considers all the impact.	64
4.2	Pipeline of the proposed approaches.	67
4.3	Overview of SBSGAN. Three domains are used as an example. (a) shows the training process of the generator G . Given an input image from Domain1, G can generate the corresponding soft-mask image and transfer the input image to different domain styles (<i>e.g.</i> , Domain1 to Domain2) according to the indicators. The foreground mask is obtained by JPPNet. (b) All real and fake images are used to minimize the adversarial loss and the domain classification loss in D	69
4.4	Overview of DA-2S. In (a), ISDC, GAP, FC, and CE respectively represent Inter-Stream Densely Connection, Global Average Pooling, Fully-Connected layer, and Cross-Entropy loss. \oplus represents element-wise summation.	73

4.5	The effectiveness of DCCV. Some samples become outliers after clustering (left figure). After conducting DCCV, unreliable samples that are far from their density centers are marked as outliers (right figure).	78
4.6	Comparison between hard-mask and soft-mask images. Images are selected from three different person re-ID datasets. The original images are listed in the first row. The second and the third rows respectively show hard-mask images by Mask-RCNN (Abdulla, 2017; He et al., 2017) and JPPNet (Liang et al., 2018). The last row shows soft-mask images generated by the proposed SBSGAN.	83
4.7	The effectiveness of different loss functions. Best viewed in color.	84
4.8	Data visualization. 5000 images are randomly selected from Market-1501 and DukeMTMC-reID to learn data distributions via the Barnes-Hut t-SNE (Van Der Maaten, 2014a), respectively. Another 200 images of each domain are used for visualization. The red circle and blue triangle respectively represent images belonging to Market-1501 and DukeMTMC-reID. The center points (<i>i.e.</i> , ‘C’) are shown using their corresponding domain color. Domain distance (<i>i.e.</i> , L_1 distance) is given between center points.	85
4.9	The changes of estimated ID number (a) and rank-1 accuracy (b) that come with the increases of N_{iter} (from 1 to 30). This experiment is conducted with DukeMTMC-reID dataset as the target domain.	86

4.10 Feature distributions of training and testing data. 5,000 training data from source (Market-1501) and target (DukeMTMC-reID) domains are selected to extract $f^{s_{xy}}$ and $f^{t_{xy}}$ via pretrained DA-2S and updated DA-2S respectively. Another 5,000 testing data are selected from target domain to extract $f^{s_{xy}}$ and $f^{t_{xy}}$ using pretrained DA-2S and updated DA-2S respectively. Barnes-Hut t-SNE (Van Der Maaten, 2014a) is used to learn the distribution of features extracted from two different DA-2S models. The center point of each distribution is denoted by ‘C-’. $D(\cdot, \cdot)$ represents the distance (*i.e.*, L_1 distance) between center points. 88

5.1 SN *vs.* VN capsule in person re-ID. The (a)-(d) and (A)-(D) are images belonging to two different IDs in the Celeb-reID dataset. The (d) and (A) include two persons with similar dark clothes. The VN capsules use the length of the vector to represent different IDs, while its orientations are used to perceive different types of clothes. With two-dimensional perception capability, different IDs can be distinguished easier by using the length of capsules. Instead, typical SN cannot make a decision between confused appearance (*e.g.*, some images in (d) are regraded as the ID in the green bounding box). Best viewed in color. 99

5.2 The pipeline of data acquisition. Four main steps are included. . . . 101

5.3 Three rows represent three different IDs in the Celeb-reID dataset. For each ID, a great number of clothing changes can be found. . . . 102

5.4 Statistic information of the proposed Celeb-reID dataset. (a), (b), and (c) respectively show the distributions of age, gender, and nationality. 103

5.5	Architecture of the proposed ReIDCaps network. Given an input image, an ImageNet-trained CNN backbone network (<i>i.e.</i> , DenseNet-121 (Huang et al., 2017a)) is used to extract low-level visual features. The output of the backbone network is fed to three branches, including capsule modules (ID and dressing perception), FSR and SEA (two auxiliary modules).	104
5.6	Sensitivity to parameter γ in Eq. 5.1. The x-axis and y-axis respectively represents the γ and mAP. Experiment is conducted on Celeb-reID.	111
5.7	Sensitivity to parameter γ in Eq. 5.1. The x-axis and y-axis respectively represents the γ and rank-1 accuracy. Experiment is conducted on Celeb-reID.	111
5.8	The SN-based and VN capsule networks. The upper network uses VN capsules. The lower one uses the tradition CNN layers. Both networks use the same input image size and backbone network (<i>i.e.</i> , the DenseNet-121). FC, BN, ReLU, and L_{CE} respectively represent the fully-connected layer, batch normalization, ReLu activation function, and the Cross-Entropy loss.	114
5.9	Intra-class variation visualization using C-Caps. Four types of clothing (9 images) belonging to the same person (76 images in total) are selected in the training set of Celeb-reID. ‘a’ to ‘d’ represent different clothes and ‘1, 2, ...’ represents the index of sample images. The cosine similarity is used to calculate the similarity between two images using the VN capsules in C-Caps where the ID is presented. An activation map is used to represent the similarity between any two images. The red and green colors respectively represent the most and the least similar pairs. Elements in the diagonal are self-similar.	116

5.10	Body parts partitions. The whole image is denoted as G . P_{11} , P_{12} , and P_{13} (also P_{21} and P_{22}) are parts equally divided from G	118
5.11	Pose estimation results (the bottom row). The head images (the top row) are extracted according to the location of keypoint of the neck.	123
6.1	Simplified network of existing infrared-visible LT-reID approaches. A and B represent different modalities. If $A=RGB$, then $B=IR$ and vice versa. The ID label is the same (different) between an image sample from modality A and the ID-tied (ID-exclusive) image sample. The GAP, IDE, and FC are short for the Global Average Pooling, ID-discriminative Embedding, and Fully Connected layers, respectively.	126
6.2	The MBT issue presented in infrared-visible LT-reID. All images belong to five people IDs (<i>i.e.</i> , ID1 to ID5). The triangles, circles, and lozenges respectively represent features learned by the modality of RGB, IR, and both of them. Shared features learned from the two modalities are inclined to bias towards the modality of RGB. Best viewed in color.	129
6.3	(a) and (b) are DCGAN models used in conventional re-ID and the proposed solution, respectively. 100-dimensional random vectors are fed into the generator for fake image generation. (c) shows the second-level learning strategy in DLS by jointly training the third modality images (the generated image by (b)) and real images using a CNN network (<i>e.g.</i> , classic IDE or IDE-T network).	134
6.4	(a): LSRO, (b): the proposed dIDEs. In (b), the IDE inputs are used as an example and assume that $K = 2$. Best viewed in color.	137
6.5	Image samples of one people ID from the SYSU-MM01 dataset (upper) and two people IDs from the RegDB dataset (lower).	139

6.6	The (a)-(d) show distributions of IDE features by Barnes-Hut t-SNE (Van Der Maaten, 2014b). The SHA is short for shared features extracted from a model trained on two modalities. RGB and IR respectively represent the RGB features and IR features extracted from the model trained on a single modality. ‘C-’ represents the center point of data distribution. $D(x,y)$ is the L1 distance between C-x and C-y. ‘RD(,)’ is short for the modality Relative Distance (RD), which is used to measure the distance of (C-IR, C-SHA) and (C-RGB, C-SHA)	142
6.7	The performance of changing γ in IDE-T on SYSU-MM01.	145
6.8	The performance of changing the weight β in IDE-T on SYSU-MM01.	146
6.9	Generated third modality images by DCGAN using the SYSU-MM01 dataset. Images in orange and gray bounding boxes are real RGB and IR images, respectively. Best viewed in color. . . .	147
6.10	Visualization of data distribution by using Barnes-Hut t-SNE (Van Der Maaten, 2014b). For the visualization, 500 RGB, IR, and third modality images are respectively used. ‘Img’ is short for ‘Image’. ‘TM’ is short for the ‘Third Modality’ generated images. ‘C-’ represents the center point for different types of data. $D(x,y)$ represents the L1 distance between C-x and C-y.	148
6.11	The performance of changing the weight α in IDE on SYSU-MM01. The baseline is the performance when $\alpha = 0$ (no third modality image is used in training).	150

List of Tables

2.1	Basic information of some exiting person re-ID datasets.	28
3.1	Comparison between virtual labels.	44
3.2	Performance of the Identif and Two-stream networks. Only the real images are used. Rank-1 accuracy and mAP are listed.	50
3.3	Comparison between LSRO and dMpRL-II on five datasets. Identif network is used by adding 24,000, 1,200, and 4,000 generated images on the three large re-ID datasets, VIPeR, and CUHK01, respectively. The improvements is shown in the <i>italic</i> and bold font by using LSRO and the proposed MpRL, respectively.	52
3.4	Comparison of all-in-one, pseudo, LSRO, and MpRLs under different numbers of generated data on Market-1501 by using the Identif network. The best improvement of different methods is highlighted in bold . Rank-1 accuracy and mAP are shown.	54
3.5	Comparison of LSRO and the proposed dMpRL-II under different numbers of generated data on VIPeR with the Identif network. The best improvement of different methods is highlighted in bold . Rank-1 accuracy and mAP are listed.	56
3.6	Comparison between using generated data by DCGAN and WGAN-GP. Two approaches are used, including LSRO and the proposed dMpRL-II. Experiments conducted on three datasets: Market-1501, DukeMTMC-reID, and VIPeR. Rank-1 accuracy and mAP are listed.	58

3.7	Performance comparison with the published state-of-the-art methods. The best and the second-best results are shown in bold and <u>underline</u> , respectively. Rank-1 accuracy and mAP are listed. The ReK means re-ranking.	59
4.1	Baseline performance of cross-domain ST-reID. Market-1501 is for training and DukeMTMC-reID is for testing.	89
4.2	Ablation study of DA-2S. Market-1501 is used for training, and DukeMTMC-reID is used for testing. The baseline does not use SEBlocks and any ISDC modules. This experiment also tries to add SEBlocks to every ISDC module to re-weight the output of ISDC in the middle layers (denoted as ISDC-SE). The DA-2S [†] (DA-2S [‡]) means only using the style-transferred images (soft-mask images) as the inputs of the 2-stream network.	90
4.3	Ablation study of DA-2S update. Market-1501 (DukeMTMC-reID) is used for training (testing). The 1 st line represents baseline performance (pretrained DA-2S). The 2 nd and 3 rd lines are performance when only one stream is used (without using ISDC and SEBlock). The 4 th to 6 th lines are using different combinations of L_t in Eq. 4.13. The last two lines shows the effectiveness of DCCV. S1 and S2 respectively represent Stream1 and Stream2.	91
4.4	Comparison with SOTA methods. X→Y means training is conducted on X and testing is conducted on Y. The performance of pretrained DA-2S is denoted as SBSGAN+DA-2S. ‘-I’ represents the updated DA-2S network. ‘-II’ is the result when re-ranking post-processing trick (Zhong et al., 2017) is adopted on SBSGAN+DA-2S-I. ‘-’ means the performance is not released.	93

5.1	Data split of the Celeb-Reid dataset. In the testing set, around 30% of images of the 420 IDs belong to the query set, the other 70% of images belong to the gallery set.	102
5.2	Ablation study of the proposed ReIDCaps network. mAP and rank-N (N=1, 5, and 10) are listed. The best performance is highlighted in Blod	112
5.3	Ablation study of different training strategies on the proposed ReIDCaps model over two different person re-ID scenarios.	113
5.4	Performance comparison between the SN-based IDE+ model (the lower network in Fig. 5.8) and the proposed Caps _{iter=4} model (the upper network in Fig. 5.8).	115
5.5	Comparison with state-of-the-art methods. The best result is shown in bold . Rank-N accuracy and mAP are listed.	119
5.6	Performance by using different weights on different body parts. The part partition can refer to Fig. 5.10. The result of Celeb-reID is mainly evaluated. The weight assigned to Celeb-reID-light is similar to Celeb-reID since both belong to the LT-reID scenario. Another group of weights on Market1501 are used by considering the contribution of different body parts.	120
5.7	Robustness score evaluation between LT-reID (Celeb-reID (C) or Celeb-reID-light (C-1)) and ST-reID (Market1501 (M) or DukeMTMT-reID (D)) re-ID scenarios using the RS (see Eq. 5.8). The mAP and rank-1 (r1) accuracy (from Tab. 5.5) are used as evaluation indexes to the RS.	122
5.8	Comparison between re-ID performance when only head images are used.	123

6.1	Building blocks of existing infrared-visible LT-reID approaches. I , I_p , and I_n respectively represent input samples, ID-tied image samples, and ID-exclusive image samples in a training minibatch.	127
6.2	Difference between D^m , D^t , and D^e settings for training.	133
6.3	The performance of IDE (with D^m , D^t , and D^e) and IDE-T (with $\alpha=0$, $\beta=1$, and $\gamma=0.1$ in Eq. 6.6) on SYSU-MM01. ResNet50 and DenseNet121 are pretrained on ImageNet. The IDE* means training from scratch. ‘R-1’ means Rank-1.	144
6.4	Comparison with state-of-the-art virtual labels for infrared-visible LT-reID.	149
6.5	Evaluate the proposed DLS using AGW on SYSU-MM01. AGT means replacing the WRT loss used in (Ye et al., 2020) with the classic triplet loss.	151
6.6	Comparison with state-of-the-art infrared-visible LT-reID approaches on the SYSU-MM01 and RegDB datasets. For IDE and IDE-T based model, ImageNet-trained DenseNet121 is adopted in comparison. The AGT is modified from AGW (Ye et al., 2020).	152

Abbreviation

Person re-ID - Person re-IDentification

ST-reID - Short-Term person re-IDentification

LT-reID - Long-Term person re-IDentification

GAN - Generative Adversarial Network

MBT - Modality Bias Training

DLS - Dual-level Learning Strategy

CNN - Convolutional Neural Network

SBSGAN - Suppression of Background Shift Generative Adversarial Network

MpRL - Multi-pseudo Regularized Label

DA-2S - Densely Associated 2-Stream

SN - Scalar Neuron

VN - Vector Neuron

IDE - ID-discriminative Embedding

HOF - Histograms of Optical Flow

MBH - Motion Boundary Histogram

LSRO - Label Smooth Regularization for Outliers

LSR - Label Smoothing Regularization

sMpRL - static Multi-pseudo Regularized Label

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

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

DPM - Deformable Part Model

ISDC - Inter-Stream Densely Connection

DCCV - Dynamic Clustering Confidence Value

GAP - Global Average Pooling
FC - Fully-Connected layer
CE - Cross-Entropy loss
BN - Batch Normalization
mAP - mean Average Precision
SEA - Soft Embedding Attention
FSR - Feature Sparse Representation
ReIDCaps - Re-Identification Capsule network
P-Caps - Primary Capsules
C-Caps - Classification Capsules
R-by-A - Routing-by-Agreement
SE - Squeeze-and-Excitation
TS1 - Training Strategy 1
TS2 - Training Strategy 2
RS - Robustness Score
dIDeS - dynamic ID-exclusive Smooth
CMC - Cumulated Matching Characteristics
SGD - Stochastic Gradient Descent
WRT - Weighted Regularization Triplet