Exploring Deep Learning powered Person Re-identification

by

Yunqi Miao

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Warwick Manufacturing Group

Feb 2023

Contents

List of Tables

List of Figures

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor, Prof. Jungong Han for his continuous guidance throughout my study. Thanks to his comprehensive knowledge and valuable experience in the computer vision community, my research ability and writing skills have been greatly improved. I also would like to thank Prof. Kurt Debattista, Prof. Giovanni Montana, Prof. Xiao Ma, and Prof. Jay Bal for their insightful advice at different stages of my study.

Moreover, I would like to thank all fellows and friends in the Data Science group, the WMG accelerator group and the Visualization Group, for always being friendly and supportive. The friendly environment and advanced devices provided by each team encourage me to be fully dedicated to my research and acquire academic achievements.

Furthermore, I would like to thank my families, who give me selfless love and support for each step I take. I would like to thank my best friend, Xiaoyang Ma, for sharing subtle but joyful things all the time, which has incredible power to help me survive the hard time. Also, I would like to thank all the nice people and great things that comfort and accompany me over the past four years, which refresh me and decorate my busy but fulfilling life.

Last but not least, I would like to thank every single thing that happened in the past four years, which makes me who I am and give me the courage to face my future life.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I hereby declare that this thesis has been composed by myself and has not been submitted before for an academic degree. Parts of this thesis have been published in publications / submissions of the author.

Publications

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

- **Miao, Y.**, Han, J., Gao, Y., and Zhang, B.. "ST-CNN: Spatial-Temporal Convolutional Neural Network for crowd counting in videos." Pattern Recognition Letters 125: 113-118, 2019.
- **Miao, Y.**, Lin, Z., Ding, G., and Han, J.. "Shallow feature based dense attention network for crowd counting." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.
- **Miao, Y.**, Lin, Z., Ma, X., Ding, G., and Han, J.. "Learning Transformation-Invariant Local Descriptors With Low-Coupling Binary Codes." IEEE Transactions on Image Processing 30: 7554-7566, 2021.
- **Miao, Y.**, Lattas, A., Deng, J., Han, J., and Zafeiriou, S.. "Physically-Based Face Rendering for NIR-VIS Face Recognition." Advances in Neural Information Processing Systems, 2022.
- Huang, N., Liu, J., **Miao, Y.**, Zhang, Q., and Han, J.. "Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review." Information Fusion, 2022.
- **Miao, Y.**, Huang, N., Ma, X., Zhang, Q., and Han, J.. "On Exploring Pose Estimation as an Auxiliary Learning Task for Visible-Infrared Person Re-identification." arXiv:2201.03859.
- **Miao, Y.**, Deng, J., Ding, G., and Han, J.. "Confidence-guided Centroids for Unsupervised Person Re-Identification." arXiv:2211.11921.

Abstract

With increased security demands, more and more video surveillance systems are installed in public places, such as schools, stations, and shopping malls. Such large-scale monitoring requires 24/7 video analytics, which cannot be achieved purely by manual operations. Thanks to recent advances in artificial intelligence (AI), deep learning algorithms enable automatic video analytics via smart devices, which interpret people/vehicle behaviours in real time to avoid anomalies effectively. Among various video analytical tasks, people search is one of the most critical use cases due to its wide application scenarios, such as searching for missing people, detecting intruders, and tracking suspects. However, current AI-powered people search is generally built upon facial recognition technique, which is effective yet may be privacy-invaded. To address the problem, person re-identification (ReID), which aims to identify person-of-interest without facial information, has become an effective panacea.

Despite considerable achievements in recent years, person ReID still faces some tough challenges, such as 1) the strong reliance on identity labels during feature learning, 2) the tradeoff between searching speed and identification accuracy, and 3) the huge modality discrepancy lying between data from different sources, *e.g.,* RGB image and infrared (IR) image. Therefore, the research interest of this thesis is to focus on the above challenges in person ReID, analyze the advantages and limitations of existing solutions, and propose improved solutions for each challenge.

Specifically, to alleviate the identity label reliance during feature learning, an improved unsupervised person ReID framework is proposed in Chapter 3, which refines not only imperfect cluster results but also the optimisation directions of samples. Based on the unsupervised setting, we further focus on the tradeoff between searching speed and identification accuracy. To this end, an improved unsupervised binary feature learning scheme for person ReID is proposed in Chapter 4, which derives binary identity representations that not only are robust to transformations but also have low bit correlations. Apart from person ReID conducted within a single modality where both query and gallery are RGB images, cross-modality retrieval is more challenging yet more common in real-world scenarios. To handle the problem, a two-stream framework, facilitating person ReID with on-the-fly keypoint-aware features, is proposed in Chapter 5.

Furthermore, the thesis spots several promising research topics in Chapter 6, which are instructive for future works in person ReID.

Abbreviations

CV	Computer vision
NLP	Natural language processing
AI	Artificial intelligence
ReID	Re-identification
CCTV	Closed circuit television
DVR	Digital video recorder
IFSEC	International fire and security exhibition and conference
CNN	Convolutional neural network
FC	Fully connected layer
PCB	Part-based convolutional baseline
GAN	Generative adversarial network
SSL	Self-supervised learning
VAE	Variational auto-encoder
DAE	Denoising auto-encoder
CL	Contrastive learning
CE	Cross entropy
CMC	Cumulative matching characteristic

mAP	Mean average precision
MAE	Mean-average error
MSE	Mean-square error
SOTA	State-of-the-art
UDA	Unsupervised domain adaptation
USL	Purely unsupervised learning
CGC	Confidence-guided centroids
CGL	Confidence-guided pseudo labels
ACM	Adversarial constraint module
VIS	Visible
IR	Infrared
VI-ReID	Visible-infrared person re-identification
HFC	Hierarchical feature constraint

Chapter 1

Introduction

1.1 Research background

As safety demands increase, more and more closed circuit televisions (CCTVs) are located in public places, such as campuses and stations, to continuously monitor population movements. According to IFSEC Globals Video Surveillance Report of 2022 [61], the number of CCTV cameras has exceeded 5.2 million in the UK. In other words, on average, every 13 people can be covered by a surveillance camera. Additionally, among the interviewed surveillance system installers, 81% of them have witnessed the increasing demands of new video surveillance projects in 2022, which demonstrates the enormous and promising value of the research in video surveillance system [61].

A typical video surveillance system is shown in Fig. 1.1, which includes 1) CCTV cameras, capturing source videos, 2) CCTV digital video recorder (DVR), coding, processing, and storing videos, and 3) display devices, displaying the recorded videos on local devices (monitor) via cables or remote devices (PCs, smartphones) via Internet. Generally, according to the requirement of different scenarios, the video surveillance system is installed for multiple purposes. For example, it can be used to facilitate the criminal investigation system via detecting and preventing crimes or used to detect anomalies in public spaces, such as train stations, in real time, or search for missing people in schools and/or hospitals. To achieve the above

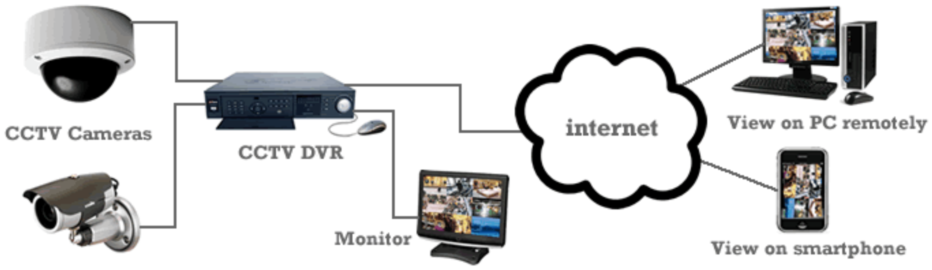
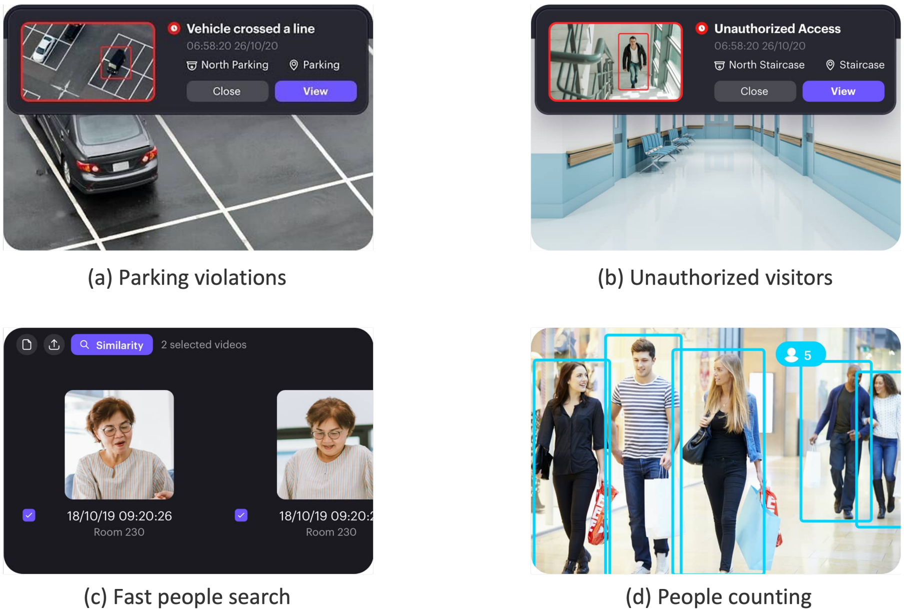


Figure 1.1: A typical video surveillance system [154].



(a) Parking violations

(b) Unauthorized visitors

(c) Fast people search

(d) People counting

Figure 1.2: Features of a typical AI-powered surveillance system - Ava Aware Cloud [5].

purposes, video analytic techniques, such as face or gait recognition, people counting, people search and tracking, are gradually introduced to exploit the potential of surveillance video systems.

Conducting such 24/7 video analytics purely through manual operations is burdensome due to a large number of cameras and the diversity of analytic tasks. Thanks to advances in artificial intelligence (AI) technology, AI-powered surveillance video system has become a panacea [61, 140], which interprets people/vehicle behaviours in videos via deep learning algorithms. According to the aforementioned survey [61], 42% of respondents have either adopted or installed AI cameras in the past year, where 51% of them chose AI-powered surveillance systems, such as IDIS and Ava Aware Cloud, which is built upon deep learning powered video analytics. Jamie Barnfield, Senior Sales Director at IDIS Europe, points out that the latest AI-powered systems can reduce the number of false alarms caused by traditional systems, thereby helping operators to detect suspicious behaviour effectively [140]. To better understand end-to-end solutions provided by AI-powered surveillance systems, some features of a video management system - Ava Aware Cloud [5] are shown in Fig. 1.2. For example, deep learning algorithms can help to identify parking violations in car parks, or identify unauthorized visitors and search for missing people in hospitals, or count the number of people in shopping malls in real time to avoid overcrowding.

Among a large number of use-cases of video analytics, the IFSEC Globals report [61] points out that “people search” is one of the major ones used by respondents, with a proportion of 23%. Another major case - “critical event search”, constituting up to 42%, also involves similar people search related techniques. Such results reveal the importance of research in people behaviour analysis. However, AI-powered “people search” is currently built upon facial recognition, which is an increasingly mature technique in vision-based identity identification with discriminative facial information [157]. Despite the effectiveness of identity recognition, face recognition is criticized for its violation of data privacy [157]. For example, the use of automated facial recognition by South Wales Police has been ruled to not be in accordance with the law by a court of appeal [135]. Additionally, the recognition accuracy is extremely vulnerable to occlusions such as facial masks or poor image quality.

Therefore, person re-identification (ReID), which identifies a person-of-interest with less sensitive yet more credible cues, such as body shape and clothing style instead of facial information, has become a popular alternative. Apart from alleviating data privacy issues, person ReID is capable of identifying a person even when the body is partially occluded by cars or bicycles. Considering its advantages, person ReID has been widely embedded into AI-powered surveillance systems to provide pre-incident warning and post-incident investigations [174]. For example, as shown in Fig. 1.2(b), person ReID can be used to detect unauthorized visitors by searching for the query visitor in the staff management system. Additionally, person ReID enables the real-time suspect identification by simply matching the query suspect with the criminal records or retrieving from resident database. Considering its wide practical value in the industry and the promising research value in academia, the thesis focuses on the topic of person ReID.

1.2 Research framework

This section will first introduce the background of person ReID and the pipeline of the typical person ReID system, and then analyze main challenges of person ReID.

1.2.1 Overview of person ReID

Person ReID typically refers to the identity retrieval problem across multiple non-overlapping cameras [220]. Due to the increasing demand for public safety, more and more surveillance cameras are installed over indoor and outdoor public places, including shopping malls, campuses, stations, *etc.* When encountering situations, such as finding missing children / older people, person ReID can quickly identify the person-of-interest in the surveillance system. Additionally, since person ReID can contribute to effective feature learning, the training scheme and network design can be also adapted to other popular and promising tasks in the computer vision community, such as vehicles ReID [189, 124, 81], face recognition [199, 56, 222], face verification [179, 127]. Given its broad practical applications and promising academic influence, person ReID has drawn increasing attention in recent years [220].

A typical person ReID system is illustrated in Fig. 1.3, which generally includes five steps:

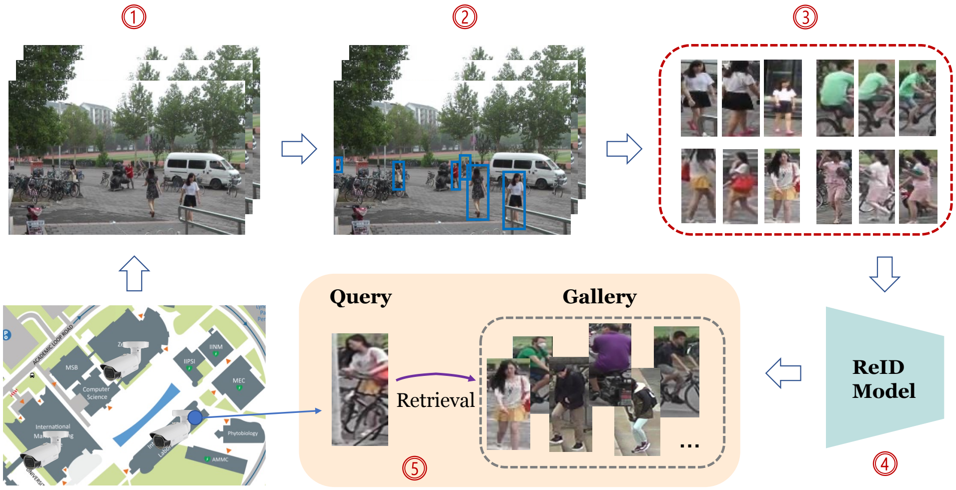


Figure 1.3: A person re-identification (ReID) system, including five steps: 1) data collection, 2) data processing, 3) identity label annotation, which is optional. 4) model training, which can be in a supervised or unsupervised manner, and 5) identity retrieval.

- **S1. Data collection**: Collecting the raw video data from surveillance cameras, which are set at multiple monitoring points within the surveillant area. However, as shown in Fig. 1.3①, the raw data generally severely suffer from cluttered backgrounds as well as humans with varying scales.

- **S2. Data processing**: Extracting the key frames from the videos. Detecting people within the image via object detection algorithms [43, 121] and tracking people across the image sequence via human tracking algorithms [13, 151]. As shown in Fig. 1.3②, detected people are marked by bounding boxes.

- **S3. Identity label annotation (optional)**: Given the detected persons cropped by bounding boxes, identity labels are required to learn effective identity features for ReID. Specifically, as shown in Fig. 1.3③, people captured by multiple cameras are grouped and assigned with identity labels. Unlike image categorisation where the category discrepancy is huge, such as cat and dog, the differences within the same person captured by different cameras or across multiple persons can be very subtle sometimes. In Fig. 1.4, for the query image (left) of each group, the positive image, shown in the middle, belongs to the same person with the query, while the negative one is shown on the right. As can be seen, the differences between positive and negative samples are subtle. Moreover, such identity labelling is required every time when a new scenario is encountered. Therefore, cross-camera labelling generally requires heavy manual work due to the unsatisfactory performances of the current automatic labelling systems. Fortunately, the step has become less necessary recently due to advances in the representability of networks as

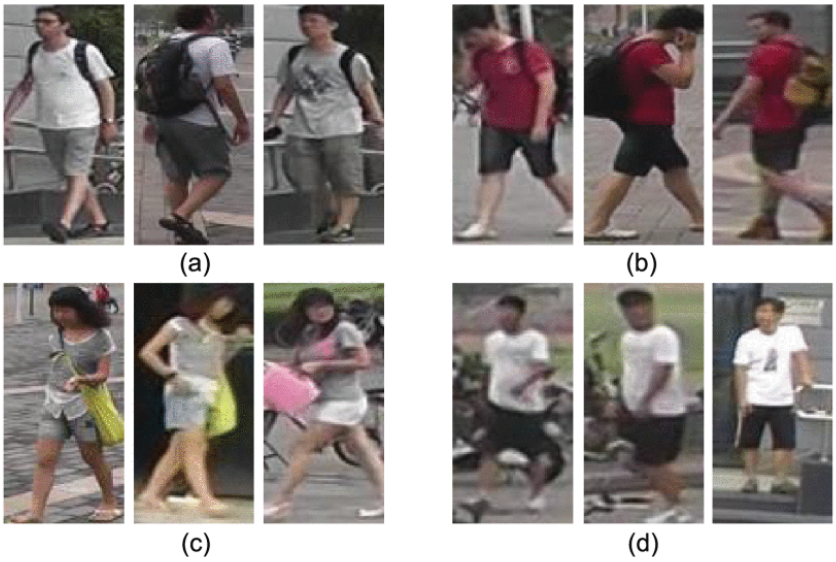


Figure 1.4: The person ReID is very challenging due to the subtle differences between positive samples and negative samples. For each group, images shown from left to right are query, positive sample, negative samples.

well as the effectiveness of training schemes enabling ReID models to learn in an unsupervised manner, which largely alleviates the burdensome annotation work.

- **S4. Model training**: Given the person images and corresponding identity labels, a Re-ID model is trained to learn representative and discriminative identity features. A large amount of related work has been done investigating the network design and training schemes, and extensive solutions have been proposed to handle the challenges of person ReID. This will be discussed elaborately in the following content.

- **S5. Identity retrieval**: Identity retrieval is conducted between a query image and a gallery set, as shown in Fig. 1.3⑤. Specifically, query features and gallery features are first extracted from the ReID model from S4. After performing the distance/similarity calculation with the query, gallery images can be ranked according to their similarities. The more representative identity features are, the more images belonging to the query identity will be ranked at the top of the ranking list. The retrieval performance is evaluated based on the ranking list of a given query. Therefore, to improve the performance, some re-ranking strategies are proposed to refine the initial ranking results.

Among the aforementioned five steps, this thesis mainly focuses on **S4** and **S5**, *i.e.,* the training and testing phase of ReID models. All to-be-discussed works in the thesis are built upon public person ReID datasets, where raw video clips, pedestrian images cropped by bounding boxes, and identity labels are provided. The main focus of this work is to derive representative identity representations from discriminative ReID models, thereby improving the ReID performance under various challenging settings.

1.2.2 Challenges of person ReID

A wide variety of real-world scenarios pose multiple challenges for person ReID. For example, although the rapidly increasing amount of data can benefit discriminative feature learning, it also brings the burdensome and time-consuming identity annotation work for training. To address the problem, label-free training schemes [60, 224, 88, 201, 82, 114, 37, 32] of ReID models are required. Moreover, the large-scale data also requires higher matching efficiency, which cannot be achieved by the distance calculation between real-valued features. Therefore, binary features (descriptors) can serve as a remedy to speed up the distance calculation and comparison [231, 22, 123, 182, 221]. Apart from searching within a single modality, the lack of high resolution cameras or poor lighting conditions will inevitably deteriorate the image quality, which requires the people search to be conducted across images with multiple resolutions [186, 107]. Additionally, the demand for uninterrupted 24/7 security ongoing surveillance systems also requires the people search to be conducted cross RGB images of daytime and infrared images of night time [194, 220, 244, 86].

To sum up, in the thesis, the author mainly focuses on three ReID research challenges:

- **Unsupervised person ReID**, which requires the training of ReID models NOT involving identity labels. Labels, typically, provide the most intuitive cues for models to learn from in the plain person ReID setting. However, without such informative cues, ReID models can only learn from pair-wise, triplet-wise or group-wise image similarity, which means the learning is extremely susceptible to label noise resulting from incorrect clustering. How to reduce the label noise is one of the main challenges of unsupervised person ReID.

- **Fast unsupervised person ReID**, which is more challenging because it not only requires the aforementioned label-free training but also requires identity features being represented by a binary value, *i.e.,* 0/1. Since binary codes carry information far more less than real-valued ones, the retrieval performance drop is unavoidable. Therefore, how to trade off between retrieval accuracy and search efficiency is the dilemma of the task.

- **Cross-modality person ReID**, which requires identity retrieval across different modalities, such as visible-infrared matching, text-image matching, and sketch-visible matching. When compared to the plain person ReID setting, cross-modality person ReID is more challenging since the ReID model needs to handle variations not only within each modality but also across different modalities. The former refers to variations in human poses, camera viewpoints, and various scenes while the latter refers to differences caused by different data sources, such as RGB images and infrared (IR) images. How to reduce the intra-modality discrepancy and the cross-modality discrepancy at the same time is one of the most essential problems in solving the cross-modality person ReID problem.

1.3 Contributions

This thesis aims to handle the aforementioned three challenges of person ReID, including unsupervised person ReID, fast unsupervised person ReID, and cross-modality person ReID. The contributions of this thesis can be summarised as follows:

- An **unsupervised person ReID framework** is proposed, which boosts ReID performances via pseudo label refinement. Specifically, confidence-guided centroids (CGC) are proposed to provide cluster-wise prototypes for feature learning. The reliability of centroids is improved via filtering out low-confidence samples during formation. Additionally, to overcome the problem that the identity information of low-confidence samples is rarely presented in their assigned centroids, confidence-guided pseudo labels (CGL) is used during training. Apart from the originally assigned centroid, instances are also encouraged to approach other centroids where their identity information is potentially embedded.

- A **fast unsupervised person ReID framework** is proposed. Fast unsupervised person ReID is a challenging problem due to the combination of the unsupervised learning manner and the learning of binary representation. Therefore, the author firstly focuses on unsupervised binary descriptors for general visual retrieval tasks, such as image retrieval and patch matching. Specifically, an unsupervised binary descriptor learning framework is proposed, where transformation-invariant and low-coupling binary descriptors are learned. To ensure the transformation invariance of binary local descriptors, contrastive training is applied to enforce original patches to have the same binary codes as their augmented counterparts while having discriminative codes with other patches. Additionally, to address the problem of high correlations between bits within binary descriptors, the author proposes an adversarial constraint module (ACM), where low-coupling binary codes generated externally are employed to guide the learning of binary descriptors. Then, the learning strategy is adapted to person ReID to achieve the fast unsupervised person ReID.

- A **cross-modality person ReID framework** is proposed, where pose estimation acts as an auxiliary learning task to assist the identity feature learning for cross-modality person ReID. Both pose and identity constraints are imposed on both backbone features and partitioned feature stripes to ensure that ID-related cues are embedded from global to local features. Additionally, to further enhance the discriminability consistency of global features and local features, Hierarchical Feature Constraint (HFC) is proposed, which regulates their corresponding predictions by the knowledge distillation strategy.

1.4 Outline

The literature on person ReID is elaborately reviewed in Chapter 2. The author first reviews person ReID methodologies under the plain setting, *i.e.,* both query and gallery images are RGB images and identity labels are involved during the training process. Moreover, existing methods proposed to handle different person ReID challenges are detailedly reviewed, including unsupervised person ReID, fast unsupervised person ReID, and cross-modality person ReID. Subsequently, targeting the limitations of existing methods for each challenge, the author provides the solutions in Chapters 3, 4, and 5, respectively. In each chapter, the existing works and their limitations are firstly reviewed. On top of that, the author provides her own solutions, where technical details and advantages are elaborately introduced. Extensive experiments are conducted on benchmarks to demonstrate the advancement of proposed methods. Additionally, ablation studies are conducted to quantify the contribution of each component of the proposed methods. Moreover, various visualisation results are also provided to intuitively demonstrate the effectiveness of the proposed methods. Finally, in Chapter 6, the summary of the proposed methods is given. Moreover, some potential future research directions are provided.

Chapter 2

Literature review

In this chapter, basic methodologies and evaluation metrics for plain person ReID are introduced in Section 2.1. Then, the main focus of the thesis, *i.e.,* three challenging scenarios for person ReID, is discussed. Specifically, unsupervised person ReID will be discussed in Section 2.2, fast unsupervised person ReID in Section 2.3 and cross-modality person ReID in Section 2.4. Finally, several conclusions will be drawn in Section 2.5.

2.1 Plain Person ReID

For the plain person ReID setting, which only involves RGB images and the model is trained in a supervised manner, existing ReID approaches have two main focuses: 1) how to learn effective identity features for identity retrieval, *i.e.,* feature learning and 2) how to regulate features with advanced constraints, *i.e.,* metric learning. The taxonomy of plain person ReID approaches is illustrated in Fig. 2.1.

2.1.1 Feature learning

A pioneering person ReID work [247] proposes the baseline for end-to-end person ReID, where the learning of identity features is regarded as a multi-class classification task with each identity being a distinct category. On top of the baseline, many works have been investigating how to facilitate the learning of discriminative features for identity retrieval, which generally can be considered in two aspects: 1) network design, *i.e.,* mining network with powerful learning ability, and 2) auxiliary information leveraging, *i.e.,* employing informative explicit cues, such as human masks, during training.

Network design can be further categorized to networks enhanced with local features [166, 254, 58, 118, 143, 239, 164, 248, 240, 165, 131], networks enhanced with the attention mechanism [104, 24, 161], and networks with various backbone architectures [81].

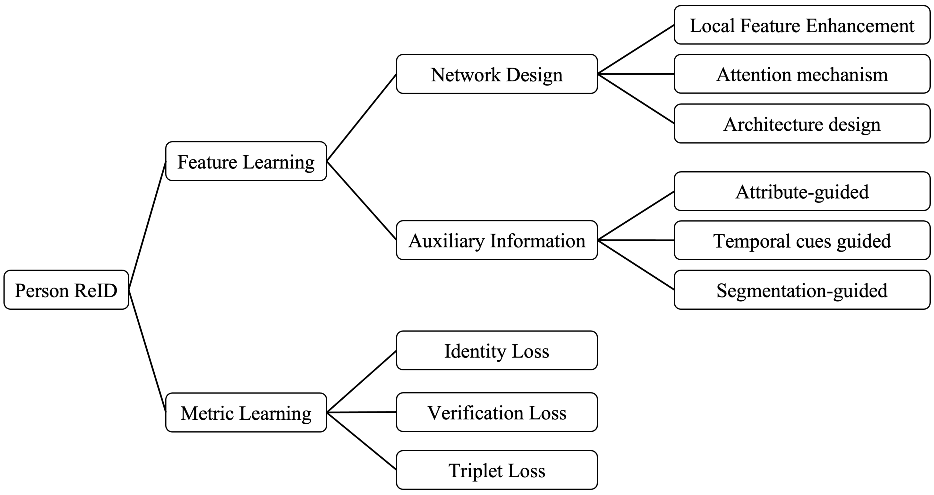


Figure 2.1: Taxonomy of plain person ReID approaches.

- **Local feature enhancement** facilitates the feature learning with fine-grained local features, which are obtained by horizontally striping backbone features [166, 217, 254] or partitioning human body with body joints detected by human pose estimation works [58, 118, 143, 239, 164, 248, 240, 165, 131].

  For methods that based on horizontal striped local features, despite the simpleness, the part-based convolutional baseline (PCB) [166] shows the effectiveness in identity retrieval performance improvement, whose structure is shown in Fig. 2.2. As can be seen, backbone feature maps are firstly horizontally striped and vectorized to get feature vectors, followed by 1×1 convolutional layers for the dimension reduction. Then, each stripe, *i.e.,* local feature, is input into an individual classifier for the identity classification. By enforcing each local feature to embed identity information with identity labels, representative features can be derived from the ReID network. Considering its effectiveness, PCB has been widely adopted by later person ReID works [254, 217].

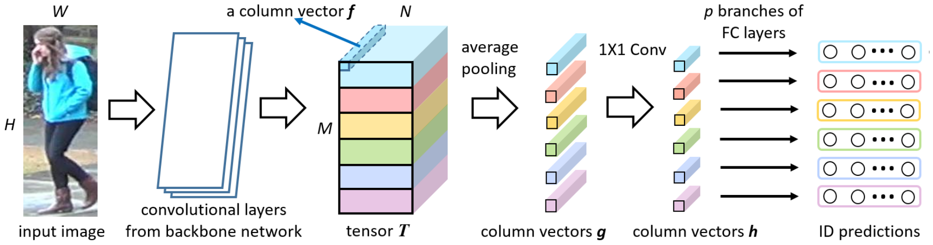


Figure 2.2: Structure of PCB [166]. Backbone features are striped horizontally and vectorised by average pooling to obtain column vectors g. Then, 1×1 convolutional layers are applied on g for dimension reduction and output column vectors h, each of which is then fed into an FC layer for classification. Taken from PCB [166].

Additionally, human pose estimation are generally used to benefit person ReID by providing the region/position of body joints. Such information can help to 1) generate person images with various poses, *i.e.,* data augmentation [58, 118, 143], and 2) align body parts [239, 164, 248, 240, 165, 131].

For data augmentation, with the help of generative adversarial networks (GAN), PN-GAN [143] synthesizes eight new images for an identity following a target canonical pose set. Apart from considering the realisticity of generated images, Pose-Transfer [118] aims to preserve the identity consistency between original image and its variants via a guider module. Additionally, Pose-Transfer applies the label smoothness scheme on the identity classification loss to balance the contribution between real images and generated ones to the training.

In terms of the body part alignment in person ReID, SpindleNet [239] segments the human body into seven regions by the human landmark information and integrates part representations (local) with backbone features (global) for identity retrieval. PDC [164] segments the body into six parts according to body joints, which are then affine-transformed to ensure the authenticity of regrouped part features. The problem of pose variations can be eliminated by re-grouping all person images with a standard pose. PBF [248] re-organizes people to a standard pose by stitching the segmented body regions with affine transformations. However, such unnatural stitch destroys the authenticity of human and requires an elaborately designed fusion method to fuse the local features. Instead of using detected body regions rigidly, Zhao *et al.* [240] and PGFA [131] employ body joint maps to refine image feature maps, where discriminative body parts for person ReID are emphasised. As a departure from directly utilizing the pose estimation results, PABR [165] adopts a two-stream structure, where one branch generates appearance (global) features while the other generates body part (local) features. To fuse global and local features, a bilinear-pooling layer is employed to generate the final descriptor for identity retrieval.

- **Attention mechanism** is leveraged to encourage ReID networks to attend to human regions rather than backgrounds. For example, HA-CNN [104] employs a joint learning scheme, where soft pixel attention and hard regional attention are integrated to locate the most discriminative parts for feature learning. Abd-net [24] is composed of two attention modules, *i.e.,* channel attention module (CAM) and position attention module (PAM). CAM aims to group channels according to semantic contexts such that channels carrying body parts cues or channels indicating backgrounds can be grouped together. PAM is then applied to aggregate human groups at the pixel level in the spatial domain. The integration of channel-wise and spatial-wise attention enables the learning of attentive features from informative regions. MGCAM [161] employs a mask-guided contrastive attention model, which is capable of generating a pair of attention maps that focus on human regions and backgrounds via spatial attention, respectively.

- **Architecture design** is upgraded as the evolution of the network. On the basis of the widely used backbone network, *i.e.,* ResNet50 [77], some work improves ReID models with several modifications that are beneficial to the identity retrieval performance, such as replacing the averaging pooling with adaptive average pooling [166] or generalized mean pooling [37], employing a bottleneck layer with BN after the last pooling layer [217]. Apart from modifying ResNet50, OSNet [257] proposes a lightweight ReID model by exploiting point-wise and depth-wise convolutions, which achieves better ReID performance than the backbone model (ResNet50).

  Despite the advanced performances of CNN-based ReID methods, they inevitably suffer from the information loss caused by the pooling/downsampling operation during the network forward. To address the problem, TransReID [81] aims to use the emerging network, *i.e.,* transformer, to facilitate the person ReID. The framework of TransReID is illustrated in Fig. 2.3, which adopts the structure of ViT [45]. As can be seen, input images are split into patches to obtain corresponding embeddings via linear projection. Before being sent to transform layers, each embedding is concatenated with a position embedding at the front and side information embedding at back. The position embeddings indicate the patch position regarding the whole image while side information embeddings embed cameras or viewpoints information. Inspired by the previous part feature reinforcement works [166, 217], after enhanced by the in-between relationship by multiple transform layers, two types of outputs are derived by two independent transformer layers, namely global branch and jigsaw branch respectively. The former outputs global identity features for retrieval while the latter shuffles and re-groups the patch features via Jigsaw Patch Module (JPM) to construct local features, where long-range dependencies can be captured. By applying multiple constraints on global features and local features at the same time, more representative ReID features can be derived.

Auxiliary information Research has shown that auxiliary information such as body shape and keypoints *etc*., can significantly facilitate the person ReID task [99, 220]. According to the type of information involved, existing person ReID methods that leverage auxiliary information can be categorized as attribute-guided methods [115, 168], temporal-guided methods [181, 150], segmentation-guided methods [91, 161, 259], and pose-guided methods [58, 118, 143, 239, 164, 248, 240, 165].

- **Attribute-guided methods** include complementary details of identities to facilitate person ReID. MTNet [115] combines the identity classification task and the attribute recognition task in a joint learning framework. The former embeds features with body shape or appearance information while the latter embeds features with information such as age and gender. The attribute branch embeds the learned features with information that has explicit meanings, which provides straightforward cues for the identity retrieval. AANet [168] enhances identity features with attribute attention maps, where

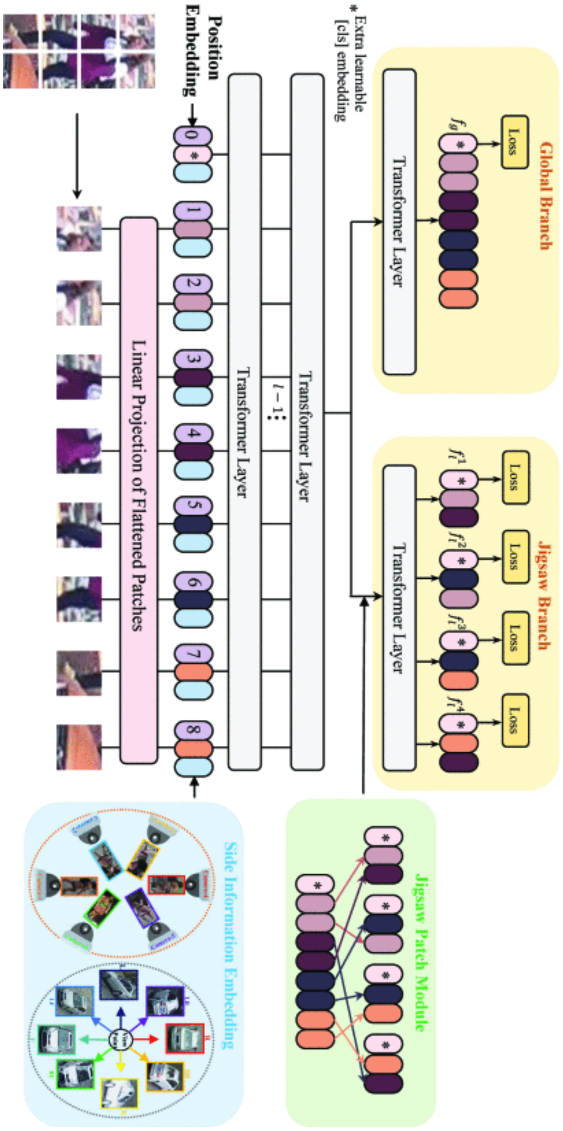


Figure 2.3: Framework of TransReID [81]. Input images are split into patches to obtain corresponding embeddings via linear projection. Before being sent to transform layers, each embedding is concatenated with a position embedding at the front and a side information embedding at back. After being enhanced by the in-between relationship by multiple transform layers, two types of outputs are derived by the global branch and jigsaw branch respectively. The former outputs global identity features for retrieval while the latter shuffles and re-groups the patch features to construct local features. By applying multiple constraints on both global and local features, more representative ReID features can be derived. Taken from TransReID [81].

class-sensitive regions activated by multiple attributes, such as clothing color, hair and gender, are emphasized. Such attention maps encourage the network to focus on human regions instead of backgrounds, which are beneficial to the learning of identity features.

- **Temporal cues guided methods** leverage temporal information, such as the moving direction, to filter out visually similar but temporally irrational samples so that the search space can be narrowed down. For example, st-ReID [181] is a two-stream framework that mines visual features and the spatial-temporal information, respectively. Features from two streams are then integrated to obtain ReID features under the supervision of a joint similarity metric with the logistic smoothing. Adopting the similar multi-stream structure, InSTD [150] decouples spatial patterns and temporal patterns from instance features by constructing their marginal distribution separately. Such disentanglement enables the retrieval of persons that are spatially match but temporally unmatched, and vice versa. On top of that, a joint metric is proposed to adaptively fuse spatial patterns and temporal patterns to better handle outliers.

- **Segmentation-guided methods** are based on the pixel-level body part segmentation, which facilitates the learning of identity features by masking out backgrounds [161] or leveraging fine-grained local features of discriminative body regions [91, 259]. Specifically, a mask-guided contrastive attention model (MGCAM) [161] differentiates human regions (foregrounds) and backgrounds via binary segmentation masks, which provides two valuable cues: 1) indicating background regions, which can be removed in the feature learning stage to avoid noise, and 2) outlining the body contours, which is one of the most important identity features. Similarly, SPReID [91] integrates the human semantic parsing into person ReID. SPReID regards the human parsing as an alternative to bounding boxes and obtains partial body features, which are more fine-grained since fewer backgrounds and more flexible contours are included. Instead of using a binary map where human regions and backgrounds are indicated by 0 and 1 respectively, ISP [259] facilitates person ReID via more fine-grained semantic segmentations. Specifically, ISP locates not only body parts but personal belongings, such as bags and hats, at the pixel level. Such belonging segmentation can be especially useful when dealing with occlusions where body parts are occluded by cars and trees. Following the spirit, P^2-Net [71] adopts a two-stream network for the mining of body part features. Concretely, a human part branch generates human part-aligned embeddings based on human part masks provided by the human parsing. Whilst a latent part branch leverages the self-attention mechanism to group semantically similar part features to obtain non-human part features. By combining the information of human parts and non-human parts, more discriminative identity-related representations can be learned.

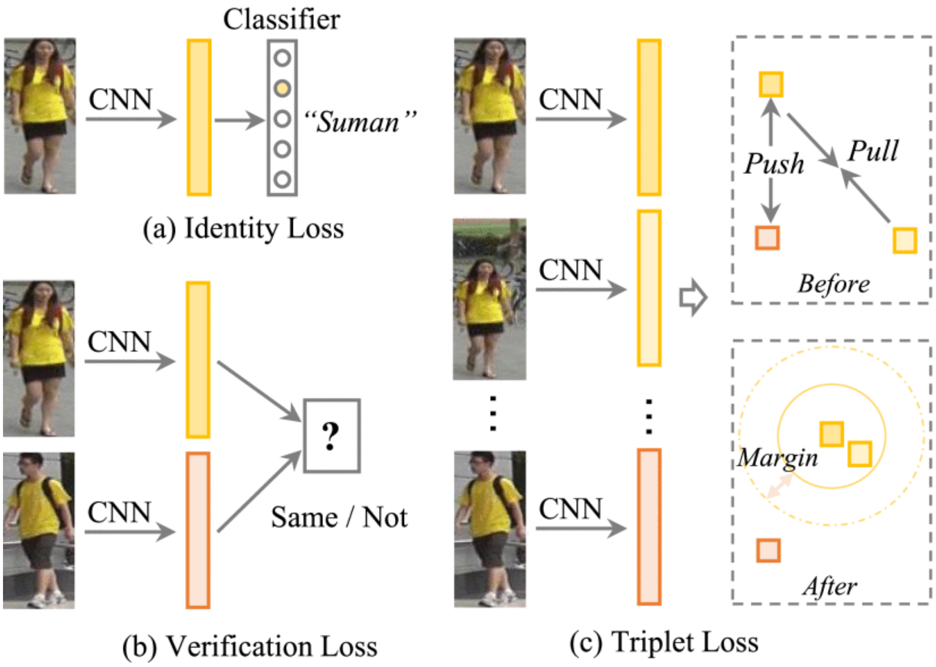


Figure 2.4: Widely used losses for person ReID. (a) Identity loss, which aims to match the image with its identity ID. (b) Verification loss, which aims to verify if two images belong to different persons. (c) Triplet loss, which aims to regulate the distance between an anchor and its positive/negative samples. Taken from AGW [220].

2.1.2 Metric learning

Metric learning, as another main focus of person ReID approaches, has been extensively studied to regulate the network optimization direction so that certain characteristics can be embedded in learned identity features. Three widely used types of loss functions in person ReID are shown in Fig. 2.4, which are identity loss, verification loss, and triplet loss, respectively.

Identity loss aims to match the query image with its identity label. As shown in Fig. 2.4(a), given an image x_i and its identity ID y_i, similar to image classification, ReID features are input to a classifier to obtain the predicted probability of x_i being matched to identity y_i, *i.e.*, $p(y_i|x_i)$. The identity loss is formulated as,

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N} log(p(y_i|x_i)), \tag{2.1.1}$$

where N denotes the number of training samples in a mini-batch. Similar to image classification, some strategies, such as label smoothing [128] and margin-based softmax [68], are employed to avoid the network being over misled by wrong labels. Due to the limited constraint power of identity loss, it is generally integrated with other loss terms to improve the representability of features.

Verification loss aims to verify if two images belong to different persons and is generally achieved by binary verification [103, 58, 115] loss or contrastive loss [161, 20, 21, 28, 245]. Both of them are conducted between sample pairs.

Specifically, given a pair of samples x_i and x_j, the features derived by ReID model can be presented as f_i and f_j. The binary verification loss is imposed on the L2 distance d_{ij} between two features, which is obtained by $d_{ij} = \|f_i - f_j\|^2$. d_{ij} is then input into a classifier to verify if the distance conforms to a positive pair or a negative pair. Formally, the probability is indicated as $p(\delta_{ij}|d_{ij})$, where $\delta_{ij} = 1$ if (x_i, x_j) is a positive pair, otherwise $\delta_{ij} = 0$. The binary verification loss between pair (x_i, x_j) is formulated as follows,

$$L_v(i,j) = -\delta_{ij}\, log(p(\delta_{ij}|d_{ij})). \tag{2.1.2}$$

Recently, inspired by the advances of contrastive learning schemes, contrastive loss is also introduced to minimize the pair-wise distance between positive pairs while maximize that between negative pairs. Specifically, the contrastive loss between pair (x_i, x_j) is formulated as,

$$L_c(i,j) = (1-\delta_{ij})\{max(0, \rho - d_{ij})\}^2 + \delta_{ij} d_{ij}^2, \tag{2.1.3}$$

where ρ denotes the margin.

Triplet loss is similar to contrastive loss, except that it handles the relations among triplets. The basic idea is that, given a triplet consisting of an anchor, its positive sample, and its negative sample, the anchor-positive distance should be smaller than the anchor-negative distance by a pre-defined margin. Formally, given a triplet (x_a, x_p, x_n), the triplet loss is defined as follows:

$$L_{tri}(a,p,n) = max(\rho + d_{ap} - d_{an}, 0), \tag{2.1.4}$$

where d_{ap} and d_{an} denote the anchor-positive distance and the anchor-negative distance, respectively. However, mining all triplets that can be constructed by the mini-batch during training is burdensome. Moreover, the optimization can be dominated by easy triplets, which have already or can be easily refined to satisfy the condition. To solve the problem, some works propose to use hardest triplet loss [83, 116, 223], where only the positive sample that is furthest to the anchor and the negative that is closest to the anchor within the current mini-batch are selected to construct the triplet.

2.1.3 Evaluation metrics

The most common metrics used for the evaluation of person ReID are mean average precision (mAP) [6] and cumulative matching characteristic (CMC) [65] top-1, top-5, top-10 accuracies.

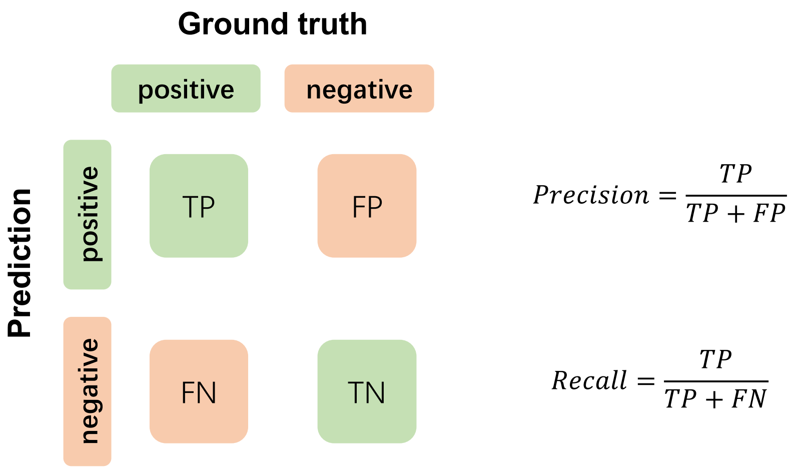


Figure 2.5: Confusion matrix (left). Formulation of precision and recall (right).

mAP is a widely-used evaluation metric for object detection [149], image retrieval [63, 162], face recognition [56], person ReID [246, 190]. Formally, mean Average Precision (mAP) is calculated as follows,

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i, \tag{2.1.5}$$

where n is the number of query samples. AP_i refers to the average precision of i-th query. The formulation of precision, average precision (AP), and mAP for image retrieval will be discussed in this section.

Precision, along with **recall**, are commonly-used metrics for the evaluation of classification models. The calculation is based on the confusion matrix, which is shown in Fig. 2.5. As can be seen, the confusion matrix represents the relationship between ground truth (actual) labels and predictions, which contains 4 attributes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), respectively. Based on the above attributes, precision and recall are formulated as follows,

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}.$$

The former implies the ratio of true positives across all positive predictions while the latter measures how many true positives are correctly classified.

In the context of image retrieval and person re-identification, the definition of precision is slightly different. Specifically, given a query image, precision is defined as the ratio of the matched images over all retrieved images. Matched images are those that have the same ground-truth (identity) label as the query image. Given a query image and top-k retrieved images in the ranking list, *i.e.,* with highest

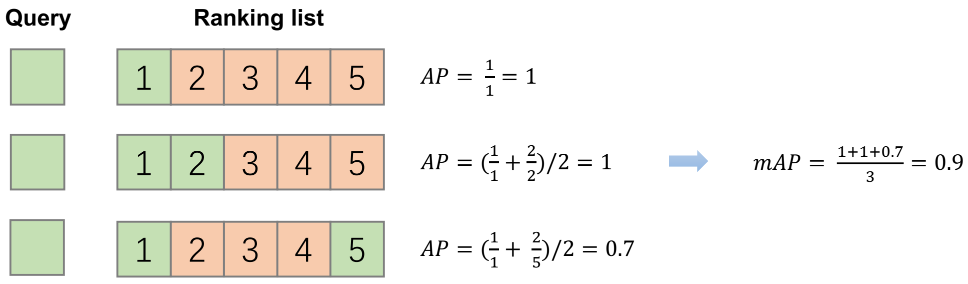


Figure 2.6: Illustration of the calculation of mAP. Matched images are indicated by green while unmatched ones are indicated by orange.

similarity scores, **average precision (AP)** can be obtained as follows,

$$AP = \frac{1}{P}\sum_{i=1}^{k} Precision_i \cdot \mathbb{1}\{g_q = g_i\},$$

where P refers to the total number of matched images among top-k images and $Precision_i$ refers to the precision at i-th index in the ranking list. The formula $\mathbb{1}\{g_q = g_k\}$ equals to 1 if the ground-truth label of i-th image (g_i) is the same as the query image (g_q), *i.e.,* matches the query, otherwise 0. After obtaining AP of all queries, **mAP** can be calculated by Eqn. (2.1.5). To better illustrate the formulation of AP, examples are shown in Fig. 2.6.

CMC curve stands for cumulative matching characteristics (CMC) curve, which is widely used to evaluate person ReID methods [246, 190, 220]. CMC represents the probability that a correct match appears in the top-k retrieved images. Concretely, for a given query and its top-k retrieved images with their similarity being sorted in descending order, the accuracy at i-th index $accuracy_i$ is defined as follows,

$$accuracy_i = \begin{cases} 1, & if\ g_q \in \mathcal{G}_i, \\ 0, & otherwise, \end{cases} \tag{2.1.6}$$

where $\mathcal{G}_i = \{g_j\}_{j=1}^{i}$ presents the identity set at i-th index, which is constructed by the ground-truth ID labels of top-i retrieved images. Generally, the mean top-i accuracy is reported during inference, which refers to the averaged top-i accuracy over all query images in the test set. Empirically, top-1, top-5 and top-10 accuracy are widely used as performance evaluation metrics for person ReID. For simplification, such metrics are denoted as top-1, top-5 and top-10 or Rank-1, Rank-5, and Rank-10 in the following content.

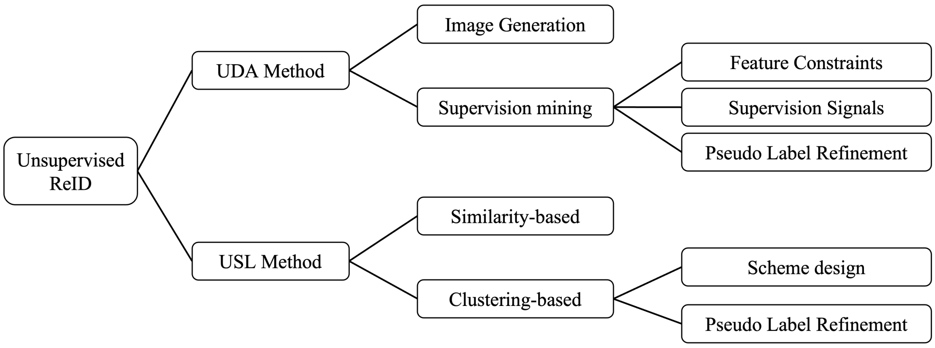


Figure 2.7: Taxonomy of unsupervised person ReID approaches.

2.2 Unsupervised person ReID

Although many works have achieved impressive performances on the basic person ReID, the success is mainly attributed to the massive human-annotated data. However, large-scale identity annotation is not only time-consuming but may involve some privacy concerns. To address the above issue, how to train Re-ID models without identity labels, *i.e.,* unsupervised person ReID, has become a research hotspot.

Unsupervised ReID vs. ReID. Generally, unsupervised ReID and plain ReID differ in the training stage. ReID leverages identity labels as supervision signals for effective feature learning, whereas unsupervised ReID does not exploit "explicit" identity labels of training data. However, some works may leverage the identity-related knowledge from other person datasets (domains) to implicitly facilitate identity feature learning. Unsupervised person ReID follows the inference scheme of ReID, where the identity retrieval is generally conducted between query images and a gallery in a single modality only.

The taxonomy of unsupervised person ReID approaches is illustrated in Fig. 2.7. As can be seen, existing unsupervised person ReID methods can be divided into two categories: 1) Unsupervised Domain Adaptation (UDA) methods, which require both labeled source datasets and unlabelled target datasets [60, 224, 88, 201, 82], and 2) purely Unsupervised Learning (USL) methods, which only require unlabelled datasets [114, 37, 32]. UDA methods will be discussed in Section 2.2.1 and USL methods will be discussed in Section 2.2.2.

2.2.1 UDA methods

To better leverage the knowledge from the labelled source domain (datasets), UDA methods mainly focus on reducing the gap between the source domain and the target domain. Generally, the source datasets can facilitate the unsupervised person ReID task in two ways, 1) image generation [39, 27, 108], and 2) supervision mining [254, 256, 60, 88].

Image generation. Generally, some methods adopt the spirit of image style transfer. For example, a state-of-the-art image-to-image style translation network - CycleGAN [258] is used to transfer the image style of the source dataset to the target dataset by person ReID works [190, 39, 27] to reduce the distribution gap between datasets. Specifically, PTGAN [190] generates high quality person images by transferring persons cross datasets. The transfer network is encouraged to keep identities of the source dataset while presenting the image style, such as lighting, of the target dataset. Such generation reduces the difficulty of unsupervised ReID models training in new scenarios. Following the similar spirit that data translated from the source domain should have similar identity information with source ID, a similarity preserving generative adversarial network (SPGAN) [39] is proposed. Given a generated image, apart from the identity consistency with the source dataset, SPGAN also considers the identity diversity with the target dataset since two domains are supposed to have non-overlapped identities. With the generated people images that have high self-similarity as well as low domain-similarity, supervised learning can be directly conducted in the target domain. Apart from the one-to-one image generation, CR-GAN [27] renders a given identity in the source dataset into multiple target domain contexts to achieve the scenario diversity. Additionally, a dual conditional GAN is employed to avoid the model collapsing into limited styles. Instead of focusing on image styles such as backgrounds and lightings, PDA-Net [108] is proposed to generate visually diverse images with the aid of the pose estimation task. Specifically, PDA-Net decouples pose features and domain content features. The former could help to manipulate images across datasets without identity information while the latter could benefit the pure ReID feature learning.

Despite the ability of the above methods in identity preserving and appearance alignment, they largely ignore the capability of unlabelled data in the target domain, which undermines the performance of ReID models. Moreover, the computational cost and the complexity of generative models are also non-negligible [220].

Supervision mining. Methods in this group aim to seek effective supervisions from source datasets. Most of them follow the training scheme that alternates between the clustering stage and the fine-tuning stage. The scheme is firstly proposed by PUL [51], as shown in Fig. 2.8. As can be seen, the model pre-trained on the labeled source dataset assigns pseudo labels for the unlabelled target dataset via k-means clustering (step ⓪ and step ①). Then, the model is fine-tuned with reliable samples which are close to the cluster centroid (step ②). As the training proceeds, more discriminative features are learned, thereby more samples are involved. Following the scheme, later works start focusing on seeking better feature constraints [234, 224, 88] or robust supervision signals [60, 254, 256, 201], and effective pseudo label refinement [59, 245, 57, 82].

- **Feature constraints.** In terms of effective feature constraints, PAST [234] adopts a self-training strategy which alternates between the conservative stage and the promoting stage during training. Specifically, the former captures the local structure of the target dataset with a ranking-based triplet loss func-

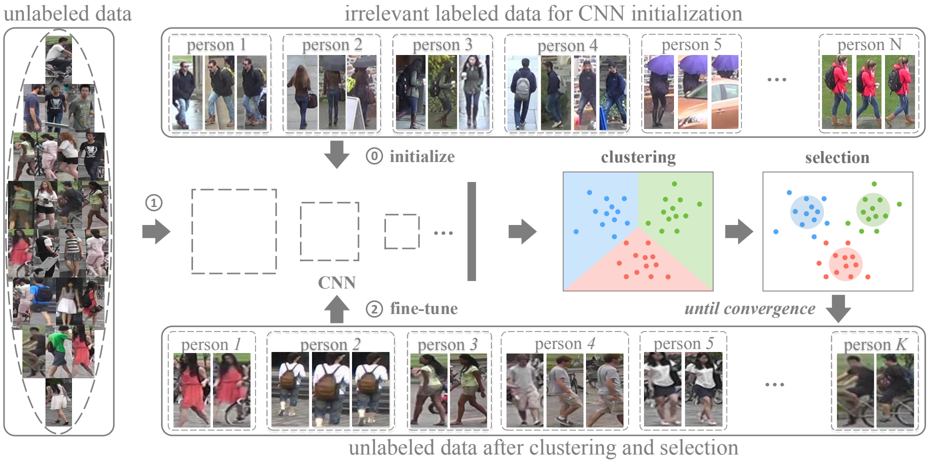


Figure 2.8: Pipeline of PUL [51], which contains three steps: ⓪ model initialization with parameters pre-trained on irrelevant labeled data, ① feature clustering and selection, and ② fine-tuning on the unlabeled target datasets. Taken from PUL [51].

tion that selects triplets based on data similarities. Whilst the latter inserts a classification layer at the end to capture the global data distribution. Instead of applying feature constraints on limited samples in the target domain, Ad-cluster [224] improves the discriminability of ReID models via augmenting confusing (hard) samples. Specifically, the training process follows an adversarial min-max manner, which alternates between the training of an image generator and a feature encoder. The former maximizes the intra-cluster discrepancy in the sample space while the latter minimizes the intra-cluster discrepancy in the feature space. By continuously generating confusing samples in the target domain while improving the discrimination of the network, the method improves the capability of ReID models greatly.

As a departure from PUL [51], labeled images of the source domain not only can provide a better initial model for the training in the target domain, but can benefit the network optimization with a properly designed objective. For example, Mekhazni *et al.* [130] aligns the pair-wise distance distributions within and between the source domain and the target domain at the same time. To achieve the goal, a Dissimilarity-based Maximum Mean Discrepancy (D-MMD) loss is proposed. The framework is shown in Fig 2.9. As can be seen, D-MMD aims to reduce the domain gap between source and target to trigger the potential of labeled source images in network optimization. Similarly, Isobe *et al.* [88] also involves the source data in the joint network optimization. They reduce the source-target domain gap by performing the transition from the source domain to the target domain progressively, with the aid of a weight adjusting function. Additionally, an extra feature constraint, *i.e.,* Fourier augmentation, is proposed, which maximizes the inter-class distance in the Fourier space.

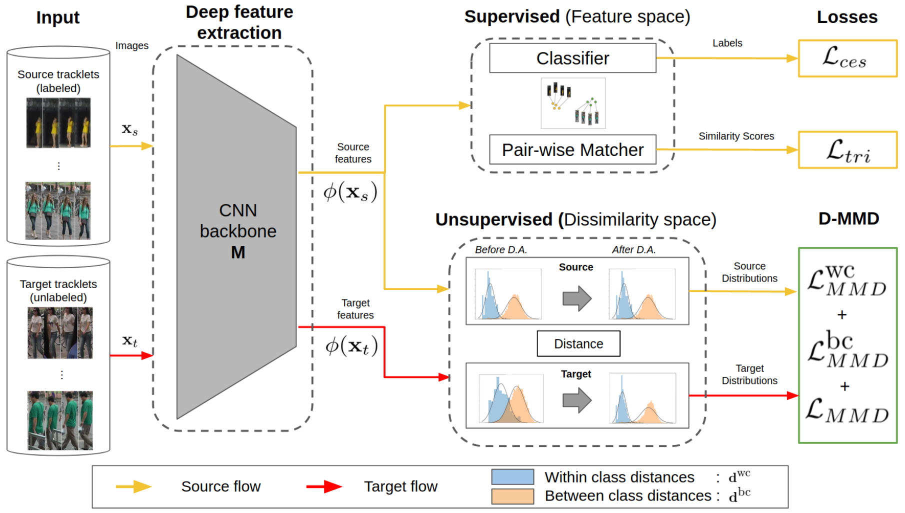


Figure 2.9: Pipeline of D-MMD [130], where labeled source images, along with unlabeled target images, are involved in the network optimization. Note that, s and t refers to the source domain and the target domain, respectively and MMD refers to maximum mean discrepancy (MMD) loss.
Taken from D-MMD [130].

- **Supervision signals.** Moreover, some works argue that simply using the pseudo label provided by the model pre-trained on the source domain is not sufficient for the learning of effective ReID features. Therefore, some strategies are proposed to enhance the supervision signals. For example, SpCL [60] leverages the hybrid memory to generate three types of pseudo labels as supervision for the feature learning, including source-domain class-level labels, target-domain cluster-level labels, and un-clustered instance-level labels. Unlike previous works where source-domain labeled samples and target-domain un-clustered outliers are discarded during training, SpCL distances the given sample from other class (cluster) centroids of its domain, all centroids of the other domain, as well as all outliers. By dynamically updating the hybrid memory, the framework could provide reliable pseudo labels for ReID feature learning. Similarly, ECN [254, 256] investigates the intra-domain variations of the target domain, including exemplar-invariance, camera-invariance, and neighborhood-invariance. By jointly considering the three types of invariance during the network optimization, the network is enforced to push a query image to itself, its nearest neighbors, as well as the generated ones with different camera styles while away from other negative instances. Additionally, an exemplar memory is employed to obtain up-to-date features. Feature instances in the memory are updated in a momentum manner [78] to ensure that the sample similarity can be calculated over the whole training set rather than a single batch. MCRN [201] adopts the contrastive learning scheme and employs the Multi-Centroid Memory (MCM) to collect more reliable positive/negative centroids for each query. For the positive centroid, instead of using the one with the highest similarity, MCM takes the centroid whose similarity ranks in the median of top-K neighbors as the learning target. In terms of negative centroid, MCM leverages the mean vector of all negative centroids of the given query. Apart from MCM, MCRN also proposes a Second-Order Nearest Interpolation (SONI). SONI generates more negative centroids for the given query by interpolating with its nearest negative neighbors in the feature space.

- **Pseudo label refinement.** Apart from mining the underlying relationships within and between identities, some works aim to refine the clustering results to obtain robust pseudo labels, *i.e.,* reducing the pseudo label noise. The noise can be presented as either different identities are mixed in one cluster or the same identity is split into multiple clusters.

  Generally, ReID methods take the one-hot (hard) pseudo labels from clustering as the learning targets. To avoid the model from being over-confident in its predictions or over-dependent on its learning target, label smoothing [167] is employed as a regularization technique in the classification task. Following the spirit, soft pseudo labels are introduced to facilitate the ReID feature learning. MMT [59] employs two collaborative networks for the ReID feature learning where both off-line hard pseudo labels and on-line soft ones jointly serve as supervision. Inspired by the "teacher-student" learning scheme, soft pseudo labels are generated based on the predictions of the other network. To avoid collaborative bias, the temporally average models are adopted to provide predictions. During inference, ONLY the past average model with better performance is employed to generate ReID features. MEB-Net [225] extends the one-to-one "teacher-student" structure to the many-to-one structure. The training process consists of three stages: 1) "Experts" training, *i.e.,* multiple "expert" models with various architectures are firstly pre-trained on the labeled source data, 2) robust feature extraction, *i.e.,* for target unlabeled samples, robust features can be obtained by averaging the features extracted by multiple "experts" with different knowledge, and 3) network optimization, *i.e.,* the network is optimized under the pseudo labels assigned by $k - means$ clustering on robust features. Additionally, a brainstorming-based mutual learning scheme is proposed to collaboratively deliver multi-experts knowledge to the target domain. Specifically, the scheme encourages each expert to learn from other experts, where the predictions serve as the learning target. To ensure the diversity of knowledge and the reliability of the learning target, similar to MMT, the scheme leverages the predictions of its temporally average model. Similar to MMT, Dubourvieux *et al.* [48] also adopts the two-branch framework. But the branches take charge of the source domain and target domain, respectively. The supervision signal for the source domain is the ground-truth identity labels while for the target domain is cluster IDs, *i.e.,* pseudo labels. To better deliver the knowledge from the source domain to the target domain, the low-level features, as well as mid-level ones, are shared by both branches. During training, the two branches are jointly optimized.

Instead of softening pseudo labels, GLT [245] refines pseudo labels in an online manner. Specifically, GLT formulates the problem of the pseudo label refinement as an optimal transport problem, which aims to find the minimal cost of the label assignment. Additionally, GLT assigns each instance with multi-group pseudo labels, *i.e.,* multiple prototypes embedded with different semantic information. Such a grouping strategy can narrow down the search space of a given query instance. The combination of online pseudo label refinement and the multi-group strategy benefits effective ReID feature learning.

Moreover, additional information is leveraged for the pseudo label refinement. For example, JVTC [102] leverages temporal information to refine pseudo labels. Aside from the one-hot pseudo labels assigned at the mini-batch level, the work also considers the temporal consistency at the dataset level. The temporal consistency refers to the high possibility of a person appearing in two cameras at time t and $t + \triangle t$, respectively. With the temporal constraint, some negative samples that are, visually similar but with large time intervals to the query image, can be filtered out. Also, there are some works [57, 82] using the local information lying in body parts to refine global-level pseudo labels of identities as different body parts generally carry discriminative information. For example, SSG [57] encourages the framework to model the similarity relationship within the target dataset by the self-similarity grouping. The motivation of SSG is illustrated in Fig. 2.10, where body parts are involved to boost the representability of the ReID features. Concretely, the clustering is conducted not only on the whole body features (global) but the partial features (local), which are from the upper body and lower body, respectively. By mining the similarity at both global level and local level, fine-grained features can be learned under the supervision of pseudo labels assigned by corresponding clusters. Instead of considering different body parts individually, SECRET [82] improves the pseudo label quality based on the consistency between predictions of global features space and local ones. For each feature space, ONLY instances whose global pseudo labels are in agreement with that in the local feature spaces are retained. Such constraint ensures consistency in the supervision signals of different feature spaces.

Albeit effective, UDA methods typically suffer from the complex source-target adaptation process, and the success is based on an assumption that the gap between the source and target domain is insignificant. Such an assumption is generally intractable for real-world scenarios, which hinders its generalizability. Therefore, USL is the main focus of this work.

2.2.2 USL methods

As discussed above, USL methods are more challenging but are well fit with real-world scenarios. Similar to UDA methods, USL methods face the problem of how to generate high quality pseudo labels for effective feature learning. Typically, for USL methods, pseudo labels are generated either based on the image similar-

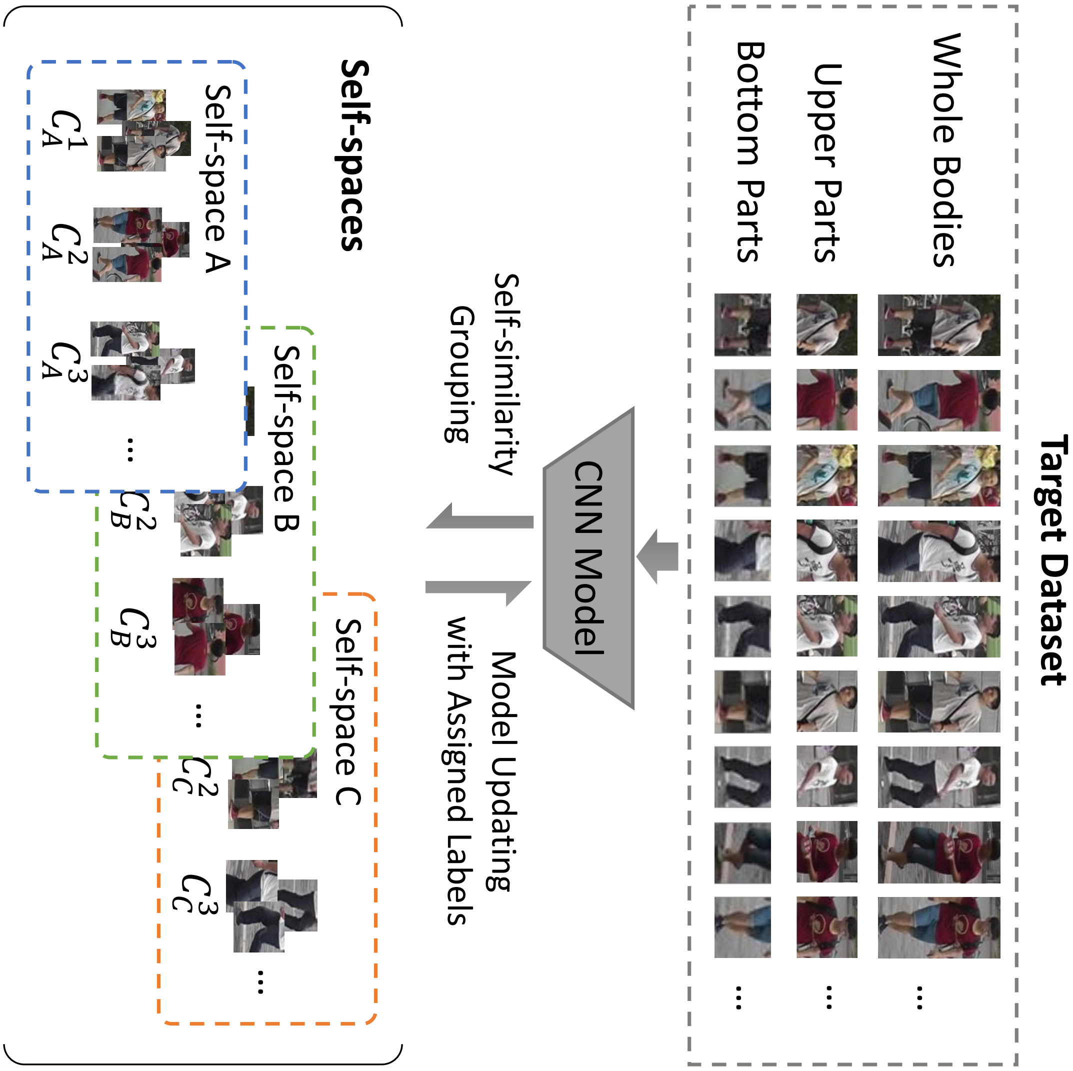


Figure 2.10: Motivation of SSG. Target images are grouped by three cues: whole body, upper parts, and bottom parts, respectively. Each self-space can provide distinctive information for the person representation learning. During the inference, features from multiple self-space are concatenated for identity retrieval. Taken from SSG [57].

ity [114, 178] or clustering results [113, 223, 37, 235]. Details of both schemes will be discussed in the following.

Similarity-based methods. SSL [114] focuses on the problem of the quantization error brought by the hard pseudo label assignment. Instead of using one-hot labels during training, SSL employs image-level similarity as the soften labels. To further handle the camera variance, a cross-camera encouragement term (CCE) is proposed by performing the feature learning under different camera views. MMCL [178] formulates the unsupervised person ReID as a multi-label classification task. The adopted labels consider the visual similarity and the cycle consistency at the same time. In other words, instances from the same class should be close in the feature space as well as sharing similar neighbors. The one-hot labels are transferred to multi-class ones via a memory-based non-parametric classifier. The framework is learned under the supervision of a multi-label classification loss, namely MMCL. Note that, instead of adopting the sigmoid function as an image classification method, MMCL regresses the classification score to $[-1, 1]$. Additionally, to handle the unbalanced number of positive and negative classes, MMCL mines and involves ONLY top-$r\%$ hardest negative classes during training. The stable pseudo labels as well as the proposed constraint for the multi-label classification attributed to the performance boosting.

Clustering-based methods. The focus of existing clustering-based methods can be roughly categorized into two aspects: scheme design [113, 41, 223, 37, 29] and pseudo label refinement [184, 233, 32, 20, 21, 235].

- **Scheme design** enhances the discriminability of networks by adopting proper clustering algorithms to provide better supervision signals for feature learning. According to the clustering algorithms adopted, methods can be further categorized as bottom-up clustering scheme based ones and memory-bank contrastive learning based ones.

  Bottom-up clustering scheme gradually merges the small clusters into bigger ones based on the similarity. Bottom-up clustering based USL methods [113, 41, 223] usually adopt the pipeline illustrated in Fig. 2.11. As can be observed, the training process of the above methods generally iterates among three steps: 1) feature extraction, *i.e.,* extracting features from a backbone network such as CNN, 2) cluster initialization and merging, *i.e.,* starting from each sample being an individual cluster, clusters are then merged under a certain criterion, and 3) cluster ID assignment and embedding learning, *i.e.,* the feature extractor is optimized in a supervised manner with the assigned cluster IDs (pseudo labels).

  Concretely, BUC [113] starts with regarding each sample as a cluster, which imitates the situation where the largest inter-class (identity) diversity is achieved. As the training proceeds, samples (clusters) are gradually grouped based on the minimum distance between two clusters. Two clusters are merged if the

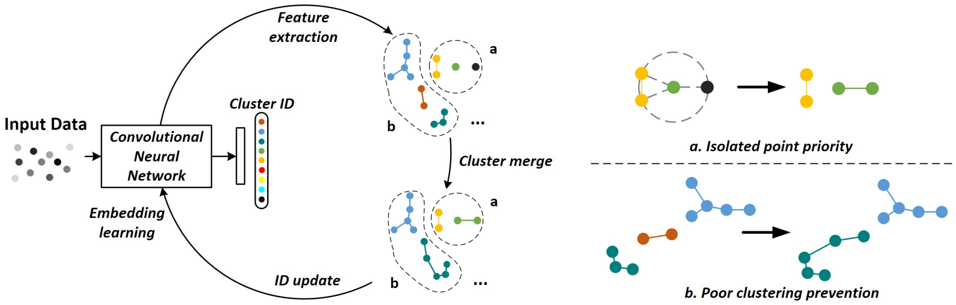


Figure 2.11: Left: pipeline of the bottom-up clustering scheme of USL methods. Right: two advantages of DBC. Taken from DBC [41].

minimum distance is lower than a pre-set threshold. Network optimization can be regarded as a process balancing diversity and similarity. However, BUC encounters the performance drop at the later epochs as only the distance between a single pair of samples from two clusters is considered in the merging criterion. To reduce the incorrect merging, DBC [41] proposes a dispersion based clustering scheme. Following the pipeline shown in Fig. 2.11, DBC merges clusters based on a criterion that considers the intra-class compactness and the inter-class separation at the same time. Such criterion mainly has two advantages, 1) assigning high priority to isolated samples. For previous methods, isolated samples, which are distant from other samples of the corresponding identity, will be further pushed away due to the low similarity. However, as shown in Fig. 2.11 (Right a), DBC merges the given isolated sample (green) to the isolated cluster (black) instead of the grouped ones (yellow), because the query has the same inter-class distance with both clusters yet isolated cluster has a lower (zero) intra-class distance. 2) preventing poor clustering. As discussed in [11], agglomerative clustering is irreversible due to the hierarchical structure. In other words, the incorrect clusters from previous iterations cannot be recovered. However, Fig. 2.11 (Right b), DBC can avoid the undesirable situation by preventing "poor" clusters (blue) with high intra-cluster dissimilarity from being further merged. In common with DBC, HCT [223] also points out that merging based on pair-wise similarity is not favourable. Such pairwise metric cannot effectively separate hard samples, which are closely located in the feature space yet belong to different persons. To solve the problem, HCT replaces the pair-wise criterion of BUC with a cluster-level one, which accounts for all pairwise distances among two clusters. HCT follows the same training scheme shown in Fig. 2.11, instead of the clusters being merged by the cluster-level distance gradually. Additionally, to improve the discriminability of the network to hard samples, a hard-batch triplet loss is leveraged during the optimization. To organize the triplets, HCT builds each training mini-batch via a PK sampling method. Specifically, after PK sampling, each batch contains P identities, each with K images. Both the identities and images are sampled randomly from the training set.

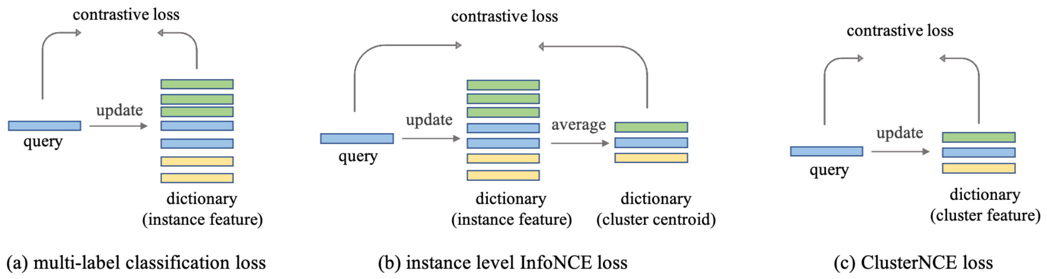


Figure 2.12: Comparisons between different types of memory-based contrastive learning schemes. (a) Both the update of memory and the contrastive loss are at the instance level. (b) The contrastive loss is computed at the cluster level yet the update of memory is instance-level. (c) Both the update of memory and the contrastive loss are at the cluster level. Taken from Cluster-Contrast [37].

Memory-based contrastive learning scheme is widely used for unsupervised representation learning, such as MoCo [78] (see Section 3.2 for more details). Recently, the potential of such a scheme has been gradually revealed in the unsupervised person ReID task.

Although being employed by some UDA methods [60, 254, 256], such scheme generally performs at the instance level. In other words, the memory bank (dictionary) is initialized and updated by instance features, and so does the formulation of the contrastive loss, as shown in Fig 2.12(a) and Fig 2.12(b). One of the drawbacks of instance-level scheme is that, due to the imbalanced distribution of training data, instance features in the memory will be updated at different paces. To handle the inconsistency, Cluster-Contrast [37] presents a cluster-level memory-based contrastive learning scheme for unsupervised person ReID. Cluster-Contrast improves the instance-level scheme from two aspects, 1) adopting a cluster-level memory dictionary, and 2) adopting a cluster-level contrastive loss, *i.e.,* ClusterNCE. The pipeline of Cluster-Contrast is illustrated in Fig. 2.13. Specifically, the memory dictionary is initialized by averaging all instance features within the corresponding cluster. As a result, ONLY a single feature vector is stored in the memory to represent each cluster. As training processes, the cluster representations are gradually updated by the instances in the mini-batch in a momentum manner. Additionally, as illustrated in Fig 2.12(c), ClusterNCE regulates the relationship between the query instance and cluster representations in the memory dictionary. Compared to instance-level schemes, such cluster-level loss is more memory-friendly and time-saving. With the above improvements, Cluster-Contrast outperforms even UDA methods by a large margin.

Previous memory-based USL methods generally discard the outliers given by the cluster results. However, those outliers can be very valuable for ReID feature learning, especially for hard sample mining. HDCRL [29] involves the un-clustered instances during training by embedding them into the contrastive

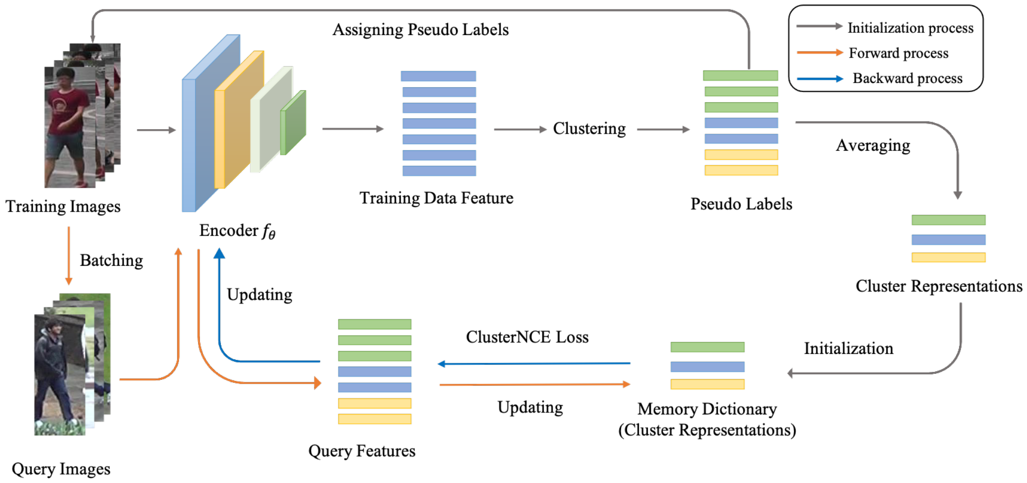


Figure 2.13: Pipeline of Cluster-Contrast, which is taken from Cluster-Contrast [37].

loss, where each serves as an individual negative cluster. Apart from the improved instance-level contrastive loss, a hybrid local-to-global contrastive learning scheme is proposed. Specifically, for each iteration, the hardest positive and negative pairs of the query instance are selected from both memory-bank and the current mini-batch. The former provides informative and valuable examples at the global level. Whilst the latter further stablizes the training in case an incorrect global-level hardest sample is selected. Additionally, HDCRL employs a pseudo probability distillation strategy to improve the network robustness. Specifically, two images with different augmentations are inputted into the teacher and student framework, respectively. The student network is encouraged to imitate the probability distributions output by the teacher network. Discriminative and robust ReID features can be learned in an unsupervised manner under the supervision signals, which is the combination of pseudo labels and the desired probability distillation.

- **Pseudo label refinement** strategies are proposed to facilitate feature learning by reducing the noise within pseudo labels. Due to the blind to ground-truth identity labels and the limited capability of the clustering algorithm, different identities can be inevitably dragged into the same cluster, and meanwhile, samples belonging to the same identity can be scattered into multiple clusters. In other words, pseudo labels are generally contaminated by noise under the unsupervised setting. The noise is especially problematic under the fully unsupervised setting [184, 233, 32, 20]. Therefore, pseudo label noise alleviation has drawn much attention in USL [48, 225, 184, 233, 32, 20, 235]. Inheriting the spirit of pseudo label refinement for UDA methods (Section 2.2.1), USL methods generally refine the pseudo labels by leveraging additional information [184, 233, 32] or alleviate the noise by data augmentation [20, 21, 235].

  Additional information includes camera IDs, clustering results, body parts, *etc.* Specifically, CAP [184] leverages the camera IDs to obtain fine-grained pseudo labels. Specifically, samples with the same cluster ID is further split into several proxies by cameras, *i.e.,* samples captured by the same camera are grouped into one proxy. With the camera-aware pseudo labels, CAP conducts intra-camera and inter-camera contrastive learning for optimization. The former enables discriminative feature learning without the interference of cameras. Whilst the latter enhances the discriminability of features against multiple cameras with the positive proxies and hard negative ones from different cameras. RLCC [233] refines noisy pseudo labels with the temporal clustering consensus, which encourages the consistency between cluster results of two consecutive iterations. Specifically, the consistency is stored in a consensus matrix $C \in \mathcal{R}^{M^{(t-1)} \times M^{t}}$, where $M^{(t-1)}$ and M^{t} denote the pseudo labels of $(t-1)$-th and t-th iterations. Each element $C(i, j)$ refers to the Intersection over Union (IoU) score, *i.e.,* consensus, between i-th pseudo class at $(t-1)$ and j-th pseudo class at t. With the consensus matrix, pseudo labels of the current iteration can be refined in a direction of preserving the consensus. The work suggests that the label consensus is an important cue to mitigate the pseudo label noise. PPLR [32] exploits the complementary relationship between features of the whole body and different body parts to refine pseudo labels at the global level and the local level in a collective manner. Specifically, the work aims to mine reliable local-context information for the global pseudo label refinement, and meanwhile, pseudo labels of each body part are smoothed based on its reliability against global features. To measure the reliability, a metric, *i.e.,* cross agreement score (CAS), is proposed. CAS presents the similarity between the k-nearest neighbors of global features and local features. A higher CAS means higher reliability. Given the one-hot global pseudo labels assigned by clustering and CAS, reliable body parts with rich context information are encouraged to move towards pseudo labels while unreliable ones, such as backgrounds, are encouraged to have random predictions, *i.e.,* uniform vectors. Then the refined local predictions are combined under weights given by CAS, where reliable body parts contribute more and vice versa. By alternating the local refinement and the global refinement, the discriminative ReID features can be learned based on informative labels.

Data/feature augmentation generates extra images/features serving as augmented views for training data and constraints the feature learning by improving the similarity between original images and the corresponding augments. ICE [20] adopts the spirit of instance contrastive learning, which encourages the given instance to be closed to its augmented views while being distanced from other instances. Specifically, the similarity between the original instance and corresponding augmented ones serves as soft pseudo labels to ensure consistency. Additionally, considering that the person ReID task generally involves more than one image for each identity, the one-hot hard pseudo labels assigned by clustering are also employed for hard instance contrast. By enforcing different instances to have the same target, the intra-class discrepancy can be alleviated. GCL [21] employs the generative model (GAN) to provide multiple "views" of a given instance for contrastive learning, which can be regarded as an online data augmentation. As shown in Fig. 2.14(a), GCL incorporates the GAN generation and the contrastive learning scheme in a joint framework. In terms of image generation, instead of leveraging the skeleton information [108], GCL estimates the 3D meshes of unlabelled training images. As can be seen from Fig. 2.14(b), the identity features and structure features, *i.e.,* 3D meshes, are encoded by the identity encoder E_id and the structure encoder E_str, respectively. The generated images, *i.e.,* x'_{ori} and x'_{new}, belong to the same identity yet with different structures. The generator G is encouraged to reconstruct the input image at high quality as well as preserving the identity features. Based on the high quality generated images, a view-invariant contrastive loss is conducted. As shown in Fig. 2.14(c), given a person, the loss maximizes the feature similarity between, 1) the original image and corresponding instance feature stored in the memory bank, 2) the augmented ones and corresponding instance feature stored in the memory bank, and 3) the original image and its augmented counterparts. The proposed mesh-based generation can synthesise person images authentically by involving multiple poses and body shapes, which improves its adaptability to real-world scenarios. Apart from data augmentations at the image level, ISE [235] proposes a feature-level generation method to enhance the discriminability of the network. Specifically, support samples $\tilde{f}$, *i.e.,* samples lying close to the cluster decision boundaries, are generated in the latent space with the formulation $\tilde{f} = f + \lambda \triangle f$, where f represents the actual sample, and λ and $\triangle f$ denote the degree and direction, respectively. Considering that incorrect samples, *i.e.,* noisy samples, have a high possibility of falling in the neighboring clusters, the direction is set from the given actual sample towards its k-nearest neighbors. In terms of the degree, which controls the amount of information absorbed from neighbors, it is updated in a progressive manner. Starting with a small value, the degree increases gradually as the training goes on. Such strategy allows hard samples with richer neighboring contexts to be involved gradually during the optimization. On top of the actual images and generated support samples, a label-preserving loss is conducted. The loss improves the feature discriminability via enforcing the network to correctly classified all samples. Based on a powerful baseline - Cluster-Contrast [37], ISE achieves the state-of-the-art performance under the USL setting.

2.3 Fast unsupervised person ReID

In recent years, person ReID is facing an exponential increase in the data scale. For example, the gallery size of benchmark datasets ranges from 10∼20k [246, 190] to 100∼200k [247, 205]. Apart from the high annotation cost, large scale data also imposes a high requirement on the identity matching speed. Despite the high matching accuracy of the aforementioned person ReID methods, the searching efficiency is quite low due to the slow Euclidean distance computation and complex ranking algorithms, especially when handling the large-scale gallery set. To solve

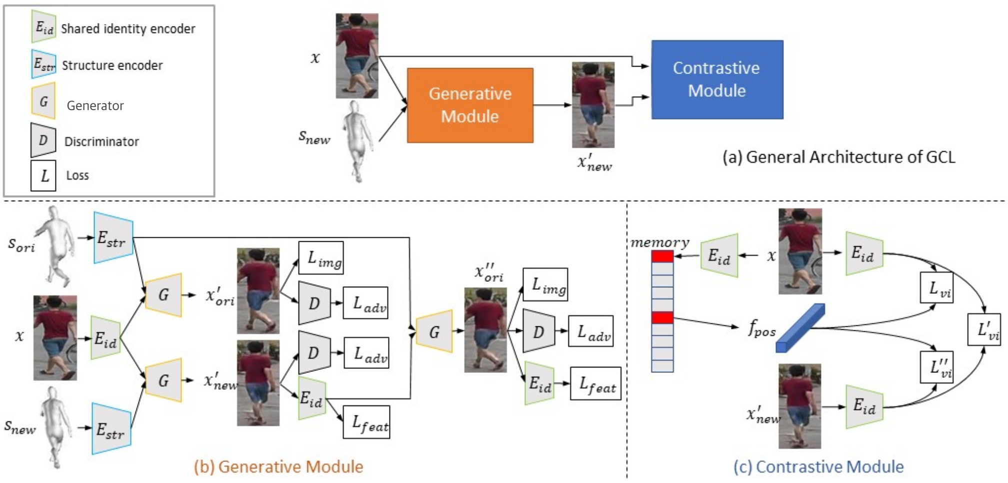


Figure 2.14: (a) Pipeline of GCL. (b) Pipeline of Generative Module. (c) Pipeline of the contrastive module. Taken from GCL [21].

the problem, some works facilitate fast identity retrieval with binary descriptors, where each bit is presented as 0 or 1. Such a binary encoding scheme conducts the Hamming distance computation during retrieval, which is simply implemented by an XOR operation. However, since binary bits carry much less information than real-valued ones, binary encoding will inevitably lead to a severe performance drop. How to adopt binary descriptors while maintaining the ReID performance has become the main focus of fast person ReID.

Fast ReID vs. ReID. Generally, fast ReID differentiates plain ReID in the data type of identity features. ReID adopts the real-valued (floating) features whereas fast ReID adopts binary descriptors to speed up the feature similarity comparison during inference. Except for the feature used, fast person ReID follows the inference scheme of ReID, where the identity retrieval is generally conducted between query images and a gallery in the single modality only.

As a more challenging setting, fast unsupervised person ReID requires the training of fast person ReID models to be processed in an unsupervised manner. In this section, state-of-the-art binary descriptors are reviewed in Section 2.3.1, and state-of-the-art fast person ReID works are reviewed in Section 2.3.2.

2.3.1 Binary descriptors

Hand-crafted based binary descriptors. Early binary descriptors are obtained via a set of pair-wise intensity comparisons with a predefined pattern, such as BRIEF [14], BRISK [100], ORB [153], and FREAK [1]. However, manually predefined sampling modes and intensity comparisons are prone to geometric transformations and distortions, thereby leading to unstable and unsatisfactory performance.

Learning based binary descriptors. Learning based binary descriptors bridge the performance gap by improving the robustness, which are categorized as supervised ones [163, 170, 171] and unsupervised ones [3, 155] according to whether labels are involved during training. For the supervised ones, LDAHash [163] jointly minimizes the intra-class covariance of the descriptors and maximizes the inter-class covariance with Linear Discriminant Analysis (LDA), to produce a binary string from a SIFT descriptor. D-BRIEF [170] projects image patches to a subspace and thresholds the projected patches to obtain discriminative binary descriptors. Binboost [171] aims to learn the illumination and viewpoint invariant binary descriptors with each bit being computed by a boosted binary hash function, which achieves state-of-the-art performance on patch matching tasks. Although learning based binary descriptors outperform the hand-crafted ones, they are built upon the knowledge of labels, which hinders the adaptability of descriptors to other tasks [202].

To avoid the involvement of labels, unsupervised binary descriptors [3, 155, 63, 75] are learned by constraining the feature distance in the latent space with image patch similarities. In other words, similar image patches are enforced to have similar binary descriptors and vice versa. Specifically, locality sensitive hashing (LSH) [3] maps images to low-dimension feature vectors via random projection and then conducts the binarization. Semantic hashing (SH) [155] adopts the multilayer Restricted Boltzmann Machine (RBM) to derive compact binary descriptors for image search. Iterative quantization (ITQ) [63] adopts an iterative optimization strategy to find the optimized projections such that the binarization loss is minimized. K-means Hashing (KMH) [75] obtains binary descriptors based on k-means clustering results, where the Hamming distance between quantized cells and the cluster centroids are minimized.

However, most traditional learning based binary descriptors are based on linear projections, which are incapable of capturing the nonlinear structure of samples, thereby leading to limited representability.

Deep binary descriptors. To address the problem, deep binary descriptors are proposed where deep neural networks are adopted to perform the non-linear projection in an end-to-end manner. Due to the discriminability of neural networks, deep learning based binary descriptors significantly outperform the previous ones. Similarly, according to whether the label information is used to facilitate the training, deep binary descriptors can be categorized into supervised ones [169, 133, 80, 211] and unsupervised ones [112, 47, 46].

Supervised binary descriptors, which rely on the pair-wise/triplet-wise similarity labels of image patches to learn the descriptors, generally achieve better performance. CNNH [203] proposes a two-stage supervised binary descriptor learning scheme. In the first stage, a pairwise similarity matrix of training samples is decomposed to a product of a hash code matrix with each row being the estimated hash code of the corresponding sample. In the second stage, a convolutional neural network is used to extract representative visual features and output hash codes. At the same time, class labels of images are adopted to supervise feature learning. Later, an end-to-end network for binary descriptor learning is proposed by Lai *et al.* [98]. The

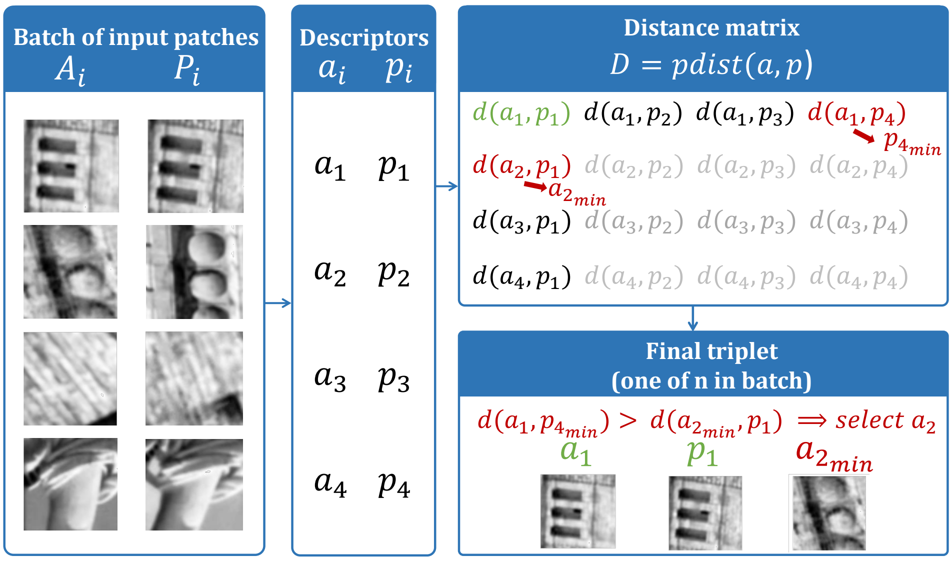

Batch of input patches
A_i P_i
Descriptors
a_i p_i
a_1 p_1
a_2 p_2
a_3 p_3
a_4 p_4
Distance matrix
$D = pdist(a, p)$
$d(a_1,p_1)$ $d(a_1,p_2)$ $d(a_1,p_3)$ $d(a_1,p_4)$
$p_{4_{min}}$
$d(a_2,p_1)$ $d(a_2,p_2)$ $d(a_2,p_3)$ $d(a_2,p_4)$
$a_{2_{min}}$
$d(a_3,p_1)$ $d(a_3,p_2)$ $d(a_3,p_3)$ $d(a_3,p_4)$
$d(a_4,p_1)$ $d(a_4,p_2)$ $d(a_4,p_3)$ $d(a_4,p_4)$
Final triplet
(one of n in batch)
$d(a_1,p_{4_{min}}) > d(a_{2_{min}},p_1) \Rightarrow select\ a_2$
a_1 p_1 $a_{2_{min}}$

Figure 2.15: Sampling strategy proposed by HardNet [133]. Given a list of positive (matched) patch pairs, a distance matrix is calculated based on all descriptors. To form a triplet with the given positive pair (a_1, p_1) (green), two hardest negative pairs (red) are selected regarding to a_1 and p_1, respectively. Among two hardest negative pairs (a_1, p_4) and (a_2, p_1), the hardest one that with the smaller distance $(a_{2_{min}})$ is chosen to form the final triplet. Taken from HardNet [133].

framework consists of a stack of convolutional layers and a divide-and-encode module. Stacked convolutional layers extract effective visual features, which are assigned to multiple branches with each presenting a bit via the divide-and-encode module. A triplet ranking loss serves as the objective of the network optimization, where positive pairs, *i.e.,* belonging to the same class, are enforced to have similar features while negative ones have discriminative ones. HashGAN [15] enhances the learning of compact binary descriptors with both real images and synthesized images, which are generated by a generative model, namely Pair Conditional Wasserstein GAN (PC-WGAN). PC-WGAN consists of a generator and a discriminator. The generator generates a fake image with a vector concatenated by a similarity-related feature of a real image and a random noise. The discriminator is trained to discriminate between real images and synthetic images with adversarial loss. With the synthesized images, the framework is trained to preserve the image similarity as well as reducing the quantization error.

Aside from image retrieval, binary descriptors are also be used to facilitate other visual tasks [169, 133, 211, 7], including patch retrieval, image matching, *etc.* Under different context, the label information used for supervised descriptor learning has different meanings. For instance, for patch retrieval, labels present whether a descriptor matches the given query. For image matching, labels present whether a descriptor can identify two correspondences from two different images. Despite

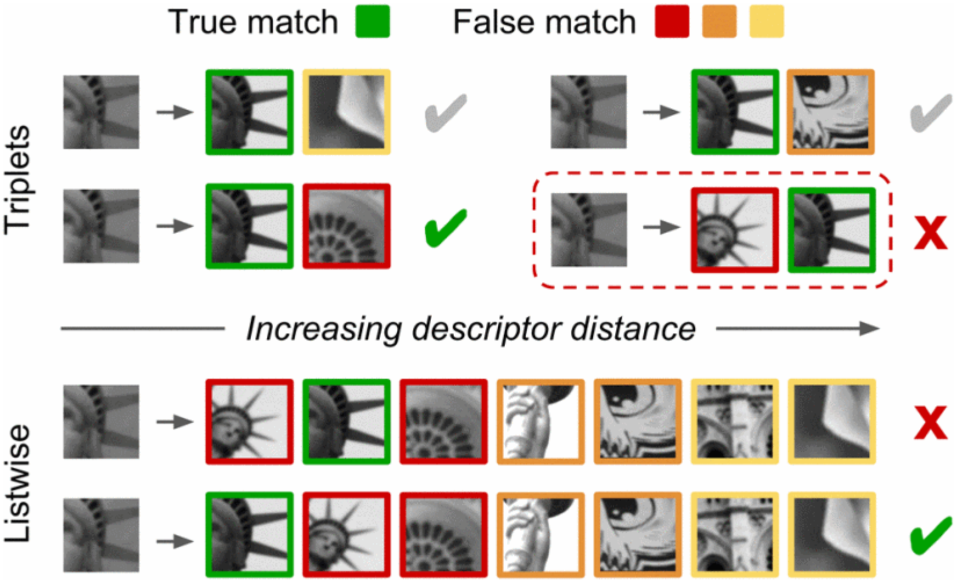


Figure 2.16: Comparison between the triplet relationship mining and list-wise ranking mining. In triplet relationship mining (top), the penalty is exclusively given to the first-ranked negative patch. Whilst in list-wise ranking mining (down), penalties are given to all high-ranked negative samples according to their AP score. List-wise ranking mining enforces correct matches without explicitly building triplets. Taken from DOAP [80].

label differences, the training scheme is quite similar to image hashing. For example, L2-Net [169] is trained to learn binary descriptors by enforcing the batch-wise relative distance among images, *i.e.,* the binary descriptors of neighboring images should also keep the neighboring relationship. Aside from the constraint on the relative distance, L2-Net considers improving the compactness of learned descriptors and reducing the bit correlation to avoid overfitting. Additionally, during training, L2-Net adopts a progressive sampling strategy that enables billions of sample pairs can be accessed within only a few epochs. On top of that, HardNet [133] shows that more powerful binary descriptors can be learned with an improved sampling strategy, which is shown in Fig. 2.15. As can be seen, a distance matrix can be calculated based on binary descriptors of a given list of positive (matched) patch pairs. In order to form a triplet with the positive pair, the two hardest negative pairs are selected regarding to the anchor and positive patch, respectively. Among the two hardest negative pairs, the hardest one with the smaller distance is chosen to form the final triplet. Although adopting the same structure as L2-Net, with the aid of such a sampling strategy, HardNet outperforms L2-Net even without using the other two loss terms. CDbin [211] argue that since image patches are generally simple in semantics, even shallow network can well capture the visual cues. CDbin proposes to learn binary descriptors with a 5-layer CNN, which is trained by an objective with 4 basic terms: 1) triplet loss, ensuring the discriminability, 2)

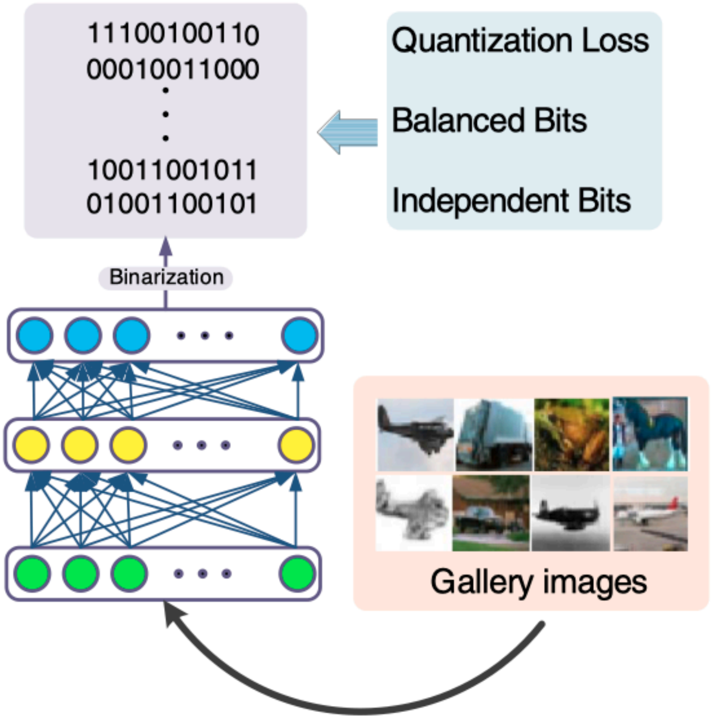


Figure 2.17: Pipeline of DH [49]. Given gallery images, the network is employed to output binary descriptors. Three feature learning constraints used during training are also listed. Taken from DH [49].

quantization loss, reducing information loss during binarization, 3) correlation loss, avoiding highly-correlated bits, and 4) even-distribution loss, enriching the carried information. Instead of exploiting the pair/triplet relationship, DOAP [80] proposes to learn binary descriptors with a list-wise ranking based metric, *i.e.,* average precision. The comparison between the triplet relationship mining and list-wise ranking mining is shown in Fig. 2.16. As can be seen, since triplet loss is calculated between a preset triplet, *i.e.,* (anchor, positive, negative), the loss value is independent of how hard the negative can be discriminated, which is inconsistent with the essence of retrieval/matching tasks. Therefore, a position-sensitive metric, *i.e.,* list-wise ranking, is proposed, where negative samples are given penalties according to the ranking. In other words, high-ranked negative samples will receive a heavy penalty. Optimizing with such ranking-based metrics, correct matches are enforced by DOAP without explicitly building triplets for training.

Despite the state-of-the-arts performance of supervised binary descriptors in visual tasks, they are unfavourable for real-world scenarios considering the cost of large-scale label annotation in terms of data size and the variety of tasks. Therefore, unsupervised binary descriptors, which do NOT require any label information during training, have drawn increasing attention in the computer vision community.

Unsupervised binary descriptors, which do not require any label information during training, have drawn increasing attention recently. Specifically, as one of the pioneering unsupervised binary descriptors learning schemes, evolutionary compact embedding (ECE) [119] proposes a bit-wise learning scheme, where each bit is learned by a weighted binary classifier iteratively trained with genetic programming (GP) strategy. To improve the discriminability of each bit, AdaBoost is employed during training to reweigh the training samples for the learning of the next bit. With the combination of GP and AdaBoost, compact binary descriptors are learned. At the same stage, Deep Hashing (DH) [49] is proposed for large-scale image retrieval. The pipeline of DH is illustrated in Fig. 2.17. As can be seen, a deep neural network is used to derive binary descriptors. Instead of involving label information, DH employs three label-free criteria to supervise the feature learning: 1) quantization loss, *i.e.,* minimizing the distance between real-valued features and binary descriptors, 2) balanced bits, *i.e.,* ensuring each bit to be evenly-distributed globally, and 3) independent bits, *i.e.,* ensuring each bit within the binary code to be independent. The parameters of the network are optimized by the objective via the back propagation algorithm. Following the spirit of DH, DeepBit [112] employs a deep neural network to learn visual features and three constraints are applied on top of derived binary descriptors. However, DeepBit improves DH by considering the robustness of descriptors to transformations. Specifically, a siamese network, taking as inputs image patches and the transformed counterparts, is adopted. An objective term, *i.e.,* transformation invariant, is applied to minimize the distance between the original patch and its corresponding transformed ones in the Hamming space. Meanwhile, the constraints of quantization minimization and even distribution are also adopted to ensure the effectiveness of derived binary descriptors.

On the basis of the framework, GraphBit [46] is proposed to handle "ambiguous bits", which lie near the threshold when adopting the sign function for binarization. Such bits tend to carry ambiguous information for confident binarization, thereby being sensitive to noise. To overcome the problem, GraphBit aims to eliminate ambiguity by exploiting underlying inner relationships between bits. Specifically, the framework is built upon a directed acyclic graph, where each bit presents a node while interactions between bits present edges. By maximizing the mutual information with input images and related confident bits, ambiguous bits can be enhanced by receiving additional guidance. The discriminability of binary descriptors is enhanced by involving more confident bits. Although the drawbacks of sign function based binarization can be alleviated by GraphBit [46], the handcrafted zero threshold is still a sub-optimal choice, which neglects data distributions. To solve the problem, DBD-MQ [47] considers binarization as a multi-quantization problem, where K-AutoEncoders (KAEs) are used to perform a more reasonable binarization. Specifically, training images are first expected to perform the encoding-decoding process with all KAEs. Then each image is associated with the AE with minimum reconstruction loss. Lastly, image features are used to retrain the associated KAEs. The above steps are iteratively executed until convergence is achieved. At the same stage, BinGAN [262], which adopts the structure of GAN network, is also proposed to eliminate the quantization error. The binarized low-dimensional

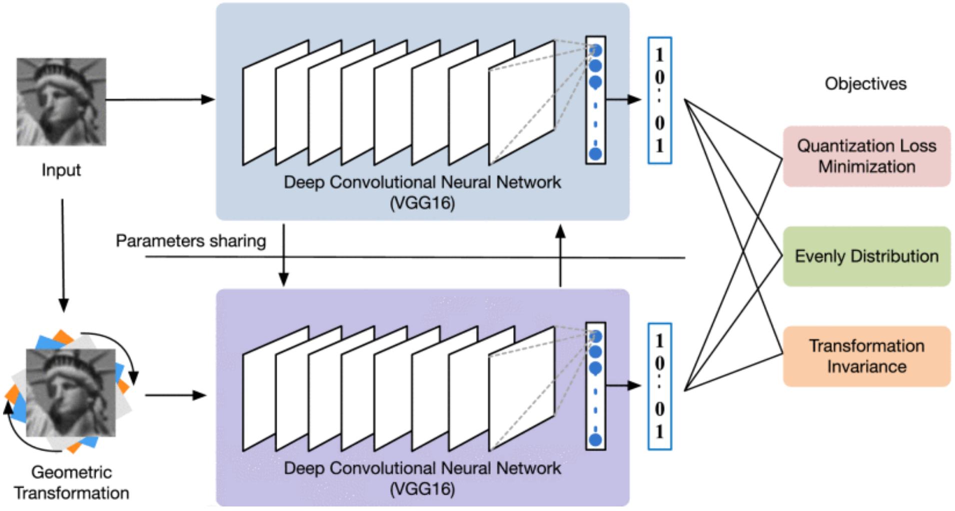
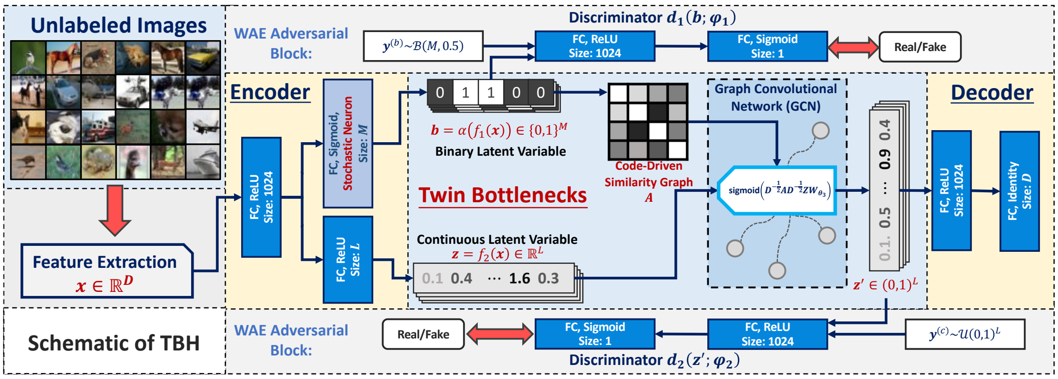

Input
Deep Convolutional Neural Network
(VGG16)
1
0
0
1
Objectives
Quantization Loss Minimization
Evenly Distribution
Transformation Invariance
Parameters sharing
Geometric Transformation
Deep Convolutional Neural Network
(VGG16)
1
0
0
1

Figure 2.18: Pipeline of DeepBit [112]. A siamese network is employed to deal with the original image patch and its transformed ones, respectively. Three objective terms are employed to supervise the learning of binary descriptors. Taken from DeepBit [112].

features output by the penultimate layer of the discriminator are regarded as binary descriptors. To preserve the consistency between high-dimensional floating features of preceding layers and derived binary descriptors, two constraints are applied: 1) reducing the correlation between floating features and binary codes, and 2) propagating the relation between bits from the high-dimensional feature space to low-dimensional ones.

Aside from the backbone framework provided by DH [49], some recent works also investigate the power of auto-encoding schemes in unsupervised binary descriptor learning, where latent variables in the middle serve as the binary codes after binarization. To better adapt auto-encoder(AE) in the binary descriptor learning, Twin-Bottleneck Hashing (TBH) [158] improves AE with twin bottlenecks, which represents binary latent variables and continuous latent variables, respectively. The framework of TBH is illustrated in Fig. 2.19. As can be seen, binary latent variables are used to build a code-driven similarity graph, which is updated dynamically based on Hamming distances during training. Continuous bottleneck, carrying rich semantic information than binary ones, ensures the reconstruction quality. On top of twin bottlenecks, a graph convolutional network (GCN) is adopted to further enhance continuous bottlenecks with similar relationships within the binary latent space. Meanwhile, two discriminators are used as regularizers for twin bottlenecks to avoid high bit correlation and align the distribution of latent spaces. The training of TBH follows the training scheme of adversarial learning, alternating between the auto-encoding step and the discriminating. Although the representability of binary



Unlabeled Images
Feature Extraction
$x \in \mathbb{R}^D$
Schematic of TBH
WAE Adversarial Block:
$y^{(b)} \sim \mathcal{B}(M, 0.5)$
Discriminator $d_1(b; \varphi_1)$
FC, ReLU Size: 1024
FC, Sigmoid Size: 1
Real/Fake
Encoder
FC, ReLU Size: 1024
FC, Sigmoid, Stochastic Neuron Size: M
FC, ReLU Size: L
0 1 1 0 0
$b = \alpha(f_1(x)) \in \{0,1\}^M$
Binary Latent Variable
Twin Bottlenecks
Continuous Latent Variable
$z = f_2(x) \in \mathbb{R}^L$
0.1 0.4 ⋯ 1.6 0.3
Code-Driven Similarity Graph
A
Graph Convolutional Network (GCN)
$\mathrm{sigmoid}\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}ZW_{\theta_3}\right)$
0.1 0.5 ⋯ 0.9 0.4
$z' \in (0,1)^L$
Decoder
FC, ReLU Size: 1024
FC, Identity Size: D
WAE Adversarial Block:
Real/Fake
FC, Sigmoid Size: 1
FC, ReLU Size: 1024
$y^{(c)} \sim \mathcal{U}(0,1)^L$
Discriminator $d_2(z'; \varphi_2)$

Figure 2.19: Framework of TBH [158]. TBH improves the auto-encoder with twin bottlenecks, representing binary and continuous latent variables, respectively. The code-driven similarity graph A is built upon Hamming distances between binary variables, which are used, along with GCN, to refine continuous variables. Two WAE Adversarial blocks are employed to regularize both latent variables. Taken from TBH [158].

descriptors can be improved by enforcing the reconstruction quality, a recent work, *i.e.,* Hashing with Contrastive Information Bottleneck (CIBHash) [144], suggests that such whole-image reconstruction encourages the model to focus on meaningless backgrounds instead of the meaningful similarity preservation. To handle the problem, CIBHash adapts a state-of-the-art contrastive learning scheme [25] into the binary descriptor learning by replacing the projection head with a probabilistic binary representation layer for binarization. On top of this, CIBHash is trained to reduce the mutual information between binary descriptors and input data inspired by information bottleneck (IB) theory, to eliminate the impacts of backgrounds. Similarly, contrastive quantization with code memory (MeCoQ) [183] also adopts the contrastive learning scheme to learn binary descriptors. However, MeCoQ adopts the training scheme of MoCo [78], which leverages the memory bank to provide a large number of negative samples to avoid the requirement of a large batch size. Following MoCo, images are first augmented to obtain different views for training. Images are fed into a feature encoder to derive real-valued embeddings, which are then sent to a trainable quantization module to obtain a "soft" quantization code. Those codes are also used to build and update the memory bank. Then, the contrastive loss is then applied to soft quantization codes and the memory bank to maximize the similarity between positive pairs (images and their augmented ones) while minimizing that between negative ones (images and codes in the memory bank). The combination of contrastive learning and quantization enables the learning of discriminative binary descriptors.

2.3.2 Fast person ReID

Since person ReID follows the same spirit as image retrieval [209], where the retrieval is conducted between query images and the gallery set based on the feature

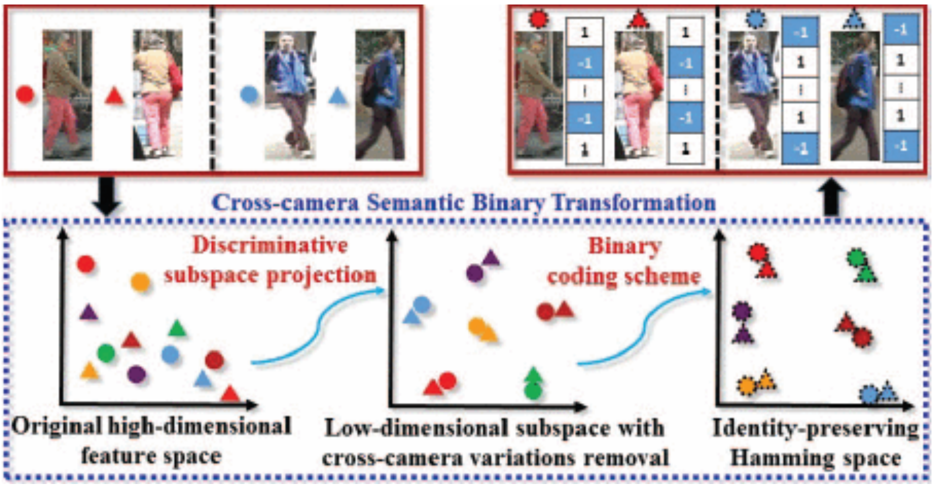

Cross-camera Semantic Binary Transformation
Discriminative subspace projection
Binary coding scheme
Original high-dimensional feature space
Low-dimensional subspace with cross-camera variations removal
Identity-preserving Hamming space

Figure 2.20: Pipeline of CSBT [22]. High-dimensional features are projected to the discriminative subspace where different identities are well separated. Then, a binary coding scheme is used for feature binarization in the subspace. Taken from CSBT [22].

similarity, most fast person ReID methods [231, 22, 123, 182, 221] are inspired by aforementioned methods to obtain effective binary descriptors for identity retrieval.

Specifically, DRSCH [231] generates bit-scalable binary descriptors for fast person ReID via a convolutional neural network (CNN). The framework is trained by a triplet loss, where positive pairs, *i.e.,* belonging to the same identity, are encouraged to have similar binary codes while negative pairs, *i.e.,* belonging to the different identities, are encouraged to be distanced in the Hamming space. Meanwhile, the margin between positive pairs and negative pairs is maximized. Additionally, DRSCH assigns each bit with unequal weights, where insignificant ones can be removed to achieve bit-scalable binary descriptors. To better adapt image hashing to person ReID, CBI [243] considers the impacts of multiple camera views. In other words, although people captured from different views are visually dissimilar, their binary descriptors should be similar. CBI aims to learn different sets of hash functions for different camera views by minimizing the intra-identity distance in the Hamming space while maximizing bit variances and cross-covariance of binary codes. However, CBI can only deal with two views at each time, which lacks flexibility in multi-camera scenarios. To handle multiple cameras, cross-camera semantic binary transformation (CSBT) [22] is proposed, which follows the pipeline shown in Fig. 2.20. CSBT firstly projects high-dimensional features to a subspace, where intra-identity distances are smaller than inter-identity distances. Subsequently, a binary coding strategy is applied in the subspace to learn binary descriptors via preserving the similarity relationship within sample triplets. Additionally, orthogonal constraints are applied to avoid correlated bits within binary descriptors. Instead of learning the projection from high-dimensional features to binary codes, Adversarial Binary Coding (ABC) [123] fits the distribution of real-valued features to

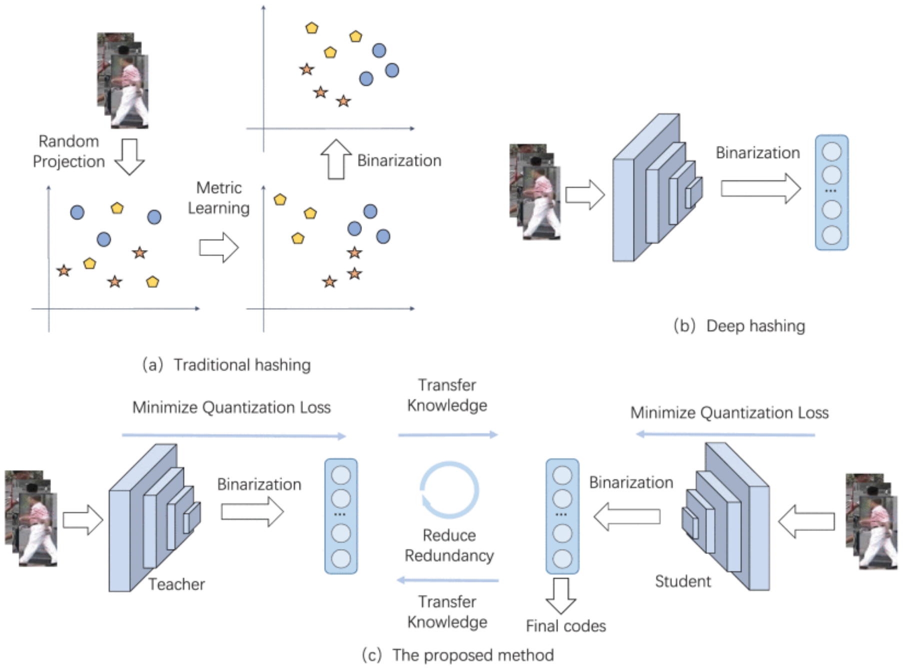


Figure 2.21: Illustration of the learning of binary descriptors for fast person ReID with (a) traditional approaches, (b) deep hashing based approaches, and (c) SIAMH. Taken from SIAMH [238].

that of prior binary codes via adversarial learning. ABC is plugged into a ReID network trained with triplet loss to ensure the semantic discriminability of derived binary codes. The work facilitates person ReID by correlating the compactness and discriminability of binary descriptors. However, DLBC [23] suggests that existing fast-person ReID methods ignore local details, which are discriminative cues to distinguish visually similar persons. However, details presented by local patches are generally inconsistent for the same person due to varying viewpoints. To address the problems, DLBC aims to extract discriminative local features from spatially salient regions, which are then transformed to binary codes via a hash layer. By maximizing the mutual information between salient local features and binary codes, the correlation between binary representations and local regions from different views is strengthened, thereby improving the robustness of learned binary codes.

Despite that much effort has been made to improve the representability of binary descriptors, the above methods generally suffer from unsatisfying performance, especially when short binary codes are applied for identity retrieval, due to the fact that information carried by each binary bit is far less than the real-valued one. To address the problem, SIAMH [238] adopts the mutual learning scheme to achieve effective binary descriptors for fast person ReID. The comparisons between previous fast person ReID works and SIAMH are illustrated in Fig. 2.21. As can be seen, SIAMH (Fig. 2.21(c)) involves a teacher model and a student model, both of which take the structure of deep hashing methods (Fig. 2.21(b)). SIAMH improves deep

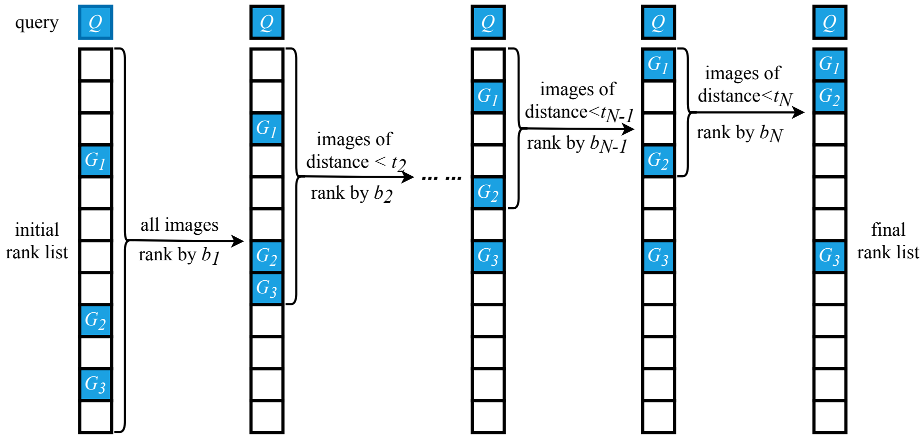


Figure 2.22: Illustration of Coarse-to-fine (CtF) [182] ranking strategy. Q and G denote the query image and the matched ones, respectively. $B = \{b_i\}_1^N$ are binary descriptors of ascending lengths and $T = \{t_i\}_1^N$ are ranking thresholds. At each time, only samples whose distance to the query is lower than the threshold are selected for further comparison. Taken from CtF [182].

hashing models in the redundancy reduction between binary codes. Specifically, codes are encouraged to attend on salience regions, which are updated dynamically during training. Meanwhile, outputs of the teacher model are used to guide the learning of the student model, and in return, classification scores output by the student model serve as dark knowledge to regularize the teacher model. With the aid of mutual learning, SIAMH improves the discrimination of learned binary codes. Instead of employing complex models, CtF [182] improves the ReID performance by taking advantage of both real-valued features and binary descriptors. Specifically, CtF proposes a coarse-to-fine ranking strategy, which is illustrated in Fig. 2.22. Firstly, binary codes of different lengths are generated for each image. During searching, for a given query image, gallery images are ranked based on the Hamming distance of shorter codes to the query. After comparing with the threshold for the current search level, only images whose distances are lower than the threshold are selected to perform the search at the next level. As the level goes deeper, longer binary codes are utilized for the Hamming distance comparison. Meanwhile, to conduct such progressive searching, preserving the information consistency between binary descriptors of different lengths is essential. To this end, the self-distillation learning strategy is used to encourage shorter codes to mimic the distribution of longer codes. Recently, sub-space Consistency Regularization (SCR) [221] proposes to balance the model accuracy and efficiency from the aspects of distance measurement and ranking speed. Specifically, instead of conducting the vector-to-vector distance computation, SCR divides each feature vector into multiple sub-spaces, in which the clustering is then conducted. With the cluster centroids given by each sub-space, the vector-to-vector distances can be transformed to the summation of centroid-to-centroid ones. Since the number of centroids is far less than the feature dimension, the computational cost can be effectively reduced. In terms of the ranking speed, a look-up-table (LUT) is built to record the mapping from cluster indexes to cluster centroids, which can speed up the inference.

However, the aforementioned fast person ReID approaches are trained in a supervised manner, where identity labels are used during training. Inspired by the emerging unsupervised binary descriptors for image retrieval as well as the success of unsupervised person ReID methods, the author aims to propose an unsupervised binary descriptor learning manner for fast person ReID. However, it has not been investigated in any works. To fill in the gap, the author first explores the learning of unsupervised binary descriptors for visual search tasks, and then applies the proposed method on fast ReID.

2.4 Cross-modality person ReID

Despite that advanced performances have been achieved, the aforementioned person ReID approaches can only handle the identity retrieval within a visible modality, where RGB is exclusively involved. However, in some real-world scenarios, the person ReID across multiple modalities [220, 244, 86] is commonly required. For example, criminal investigation commonly requires the search of persons captured by low-resolution cameras among high-resolution databases or searching by verbal descriptions of suspects merely. Considering the above cases, person re-identification across images with different resolutions or images and text can be very helpful. Additionally, 24/7 smart surveillance system requires the identity identification system running from day to night to guarantee security continuously. However, RGB cameras can only capture limited information under low-lighting conditions, near infrared cameras are required to collect effective information via infrared (IR) images during night time. Therefore, the identity retrieval cross visible images and infrared images is also important. Considering the wide applicability, the thesis mainly focuses on Visible-Infrared person ReID (VI-ReID), which handles the identity matching between the daytime visible and night-time infrared images.

VI-ReID vs. ReID. Generally, due to the existence of two modalities, VI-ReID differentiates plain ReID at both the training and inference stages. During training, plain person ReID faces challenges within the modality, such as pose variances, illumination changes, and occlusions. For VI-ReID, apart from the aforementioned intra-modality discrepancy, the existence of two intrinsically different modalities also poses challenges across modalities, such as the differences in the intensity/color information between visible images and infrared images. The noticeable intra- and inter-modality discrepancies make the effective identity feature learning a challenging task. In terms of the inference, as illustrated in Fig. 2.23, the query images and gallery images are from the same modality for ReID, yet different for VI-ReID, which fits the requirement of day-to-night people searching.

As the day-to-night people search demand increases, much effort has been put into the study of VI-ReID in recent years. Existing VI-ReID approaches, as illustrated in Fig. 2.24, mainly improve the ReID performance from four aspects: 1)

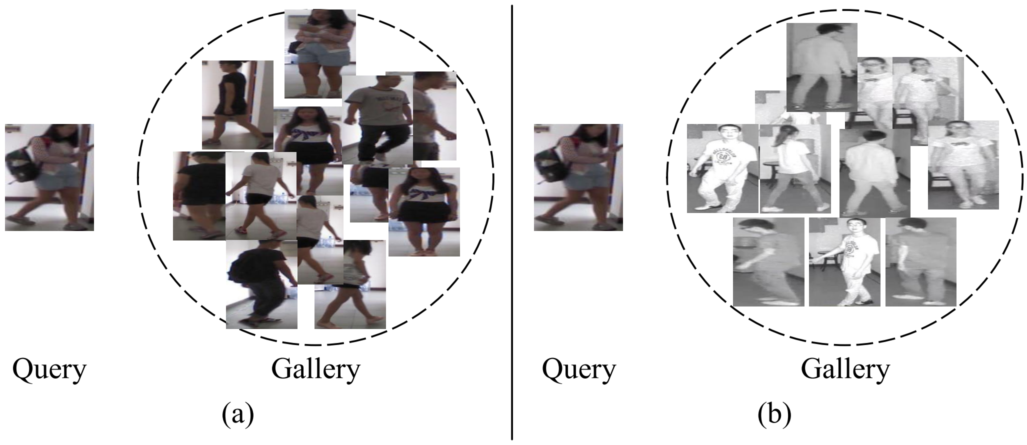


Figure 2.23: Comparison between the inference of (a) person ReID and (b) VI-ReID models. The query and gallery images belong to the same modality for ReID yet from different modalities for VI-ReID.

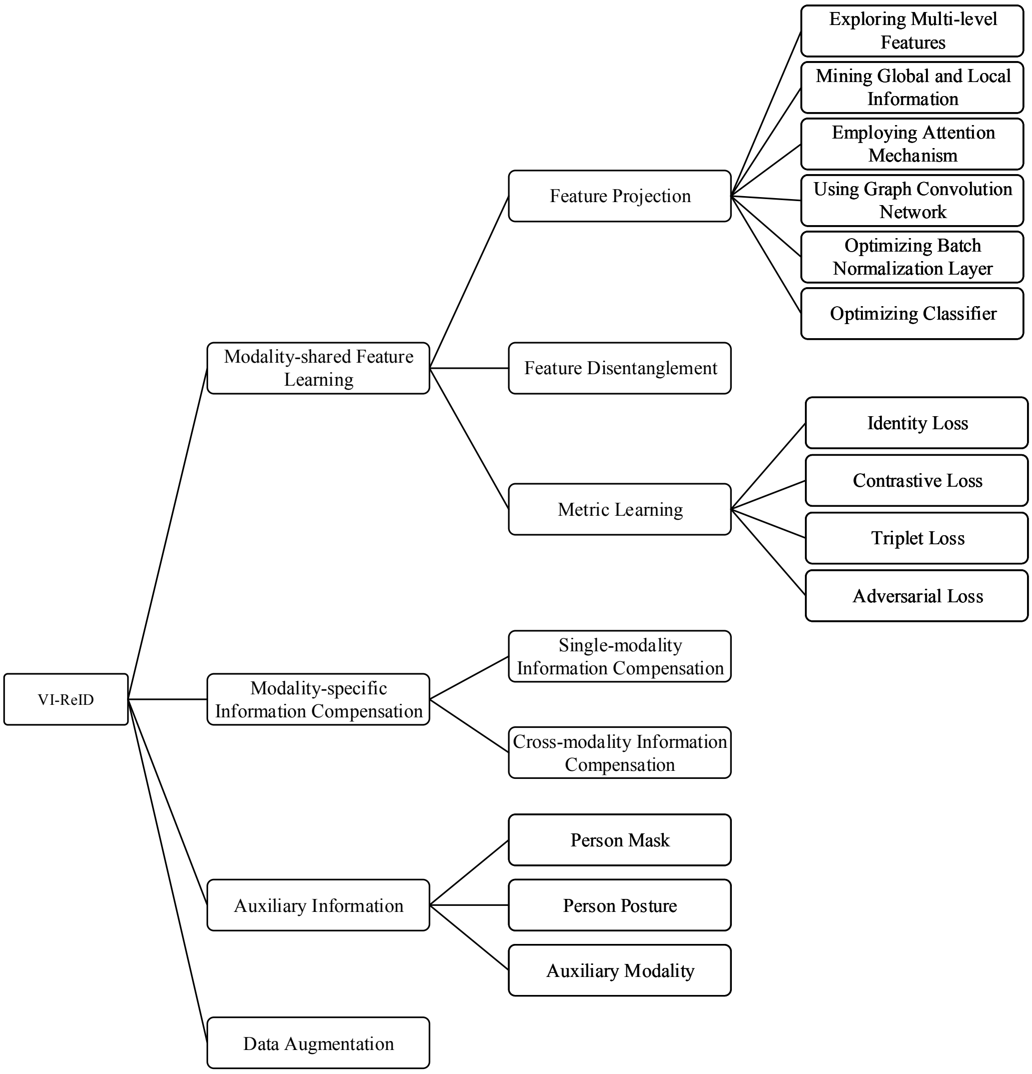
modality-shared feature learning (Section 2.4.1), 2) modality-specific information compensation (Section 2.4.2), 3) auxiliary information (Section 2.4.3), and 4) data augmentation (Section 2.4.4).

2.4.1 Modality-shared feature learning

To perform accurate identity retrieval across the VIS modality and the IR modality, identity-related features that are shared by both modalities, *i.e.,* modality-shared features, are required. In other words, modality-specific features, that will cause confusion during retrieval, are discarded during the feature extraction, thereby leading to a better cross-modality person ReID performance. Many works have been investigated in the learning of modality-shared features in recent years, which mainly include three types: 1) feature projection, 2) feature disentanglement, and 3) metric learning. Multiple pipelines of modality-shared feature learning are shown in Fig. 2.25.

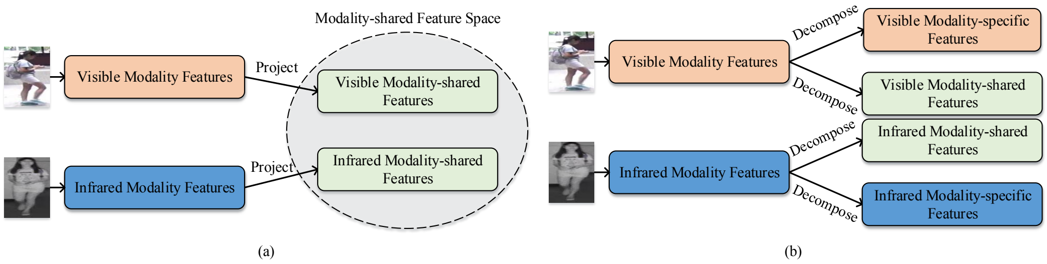
Feature projection based methods project single-modality features of into a modality-shared feature space to reduce the gap between two modalities. Then, discriminative modality-shared features are learned in the modality-shared feature space to handle the discrepancy within the modality. The pipeline of feature projection based methods is shown in Fig. 2.25(a).

As a pioneering VI-ReID works, Zero-pad [194] investigates the structure of three modality-shared feature learning networks, including a one-stream network, a two-stream network and one with an asymmetric FC layer. The illustrations of three structures are shown in Fig. 2.26. As can be seen, **one-stream network** takes the input of both RGB images and IR images and ask both images to share all parameters of the network. **Two-stream network** firstly inputs RGB images and IR images into two different convolutional blocks. Then single-modality features



VI-ReID
Modality-shared Feature Learning
Feature Projection
Exploring Multi-level Features
Mining Global and Local Information
Employing Attention Mechanism
Using Graph Convolution Network
Optimizing Batch Normalization Layer
Optimizing Classifier
Feature Disentanglement
Metric Learning
Identity Loss
Contrastive Loss
Triplet Loss
Adversarial Loss
Modality-specific Information Compensation
Single-modality Information Compensation
Cross-modality Information Compensation
Auxiliary Information
Person Mask
Person Posture
Auxiliary Modality
Data Augmentation

Figure 2.24: Taxonomy of VI-ReID approaches.



Modality-shared Feature Space
Visible Modality Features
Project
Visible Modality-shared Features
Infrared Modality Features
Project
Infrared Modality-shared Features
(a)
Visible Modality Features
Decompose
Visible Modality-specific Features
Decompose
Visible Modality-shared Features
Infrared Modality Features
Decompose
Infrared Modality-shared Features
Decompose
Infrared Modality-specific Features
(b)

Figure 2.25: Pipeline of modality-shared feature learning by (a) feature projection and (b) feature disentanglement.

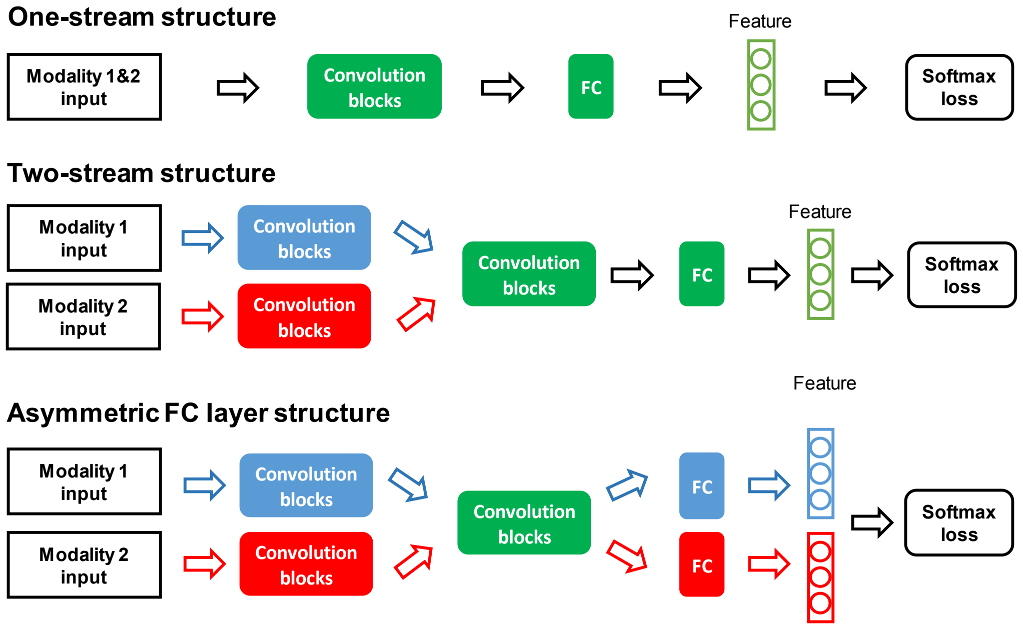


Figure 2.26: Structures of three types of modality-shared feature learning networks, including one-stream network, two-stream network, and one with asymmetric FC layer. Convolutional blocks and FC layers with the same color denote that parameters are shared by both modalities and vice versa.

are merged to extract modality-shared features via subsequent shared-parameter layers. **Network with asymmetric FC Layer** adopts the structure of the shallow layers of a two-stream network. However, backbone features are fed into two FC layers here, with each extracting discriminative modality-shared features within the corresponding modality. Three structures are widely adopted in later feature projection based methods. Especially, the design of the effective parameter-shared part has become the main focus since it directly impacts the modality-shared feature learning. Specifically, according to the solutions proposed by later works for network improvement, current feature projection methods can be categorized into 6 types: 1) exploring multi-level features, 2) mining global and local information, 3) employing attention mechanism, 4) using graph convolution network (GCN), 5) optimizing batch normalization (BN) layer, and 6) optimizing classifier.

- **Exploring multi-level features** aims to take advantage of detail-embedded low-level features and semantic-embedded high-level features and make them provide complementary information mutually. Specifically, MGN [204] is built upon the two-stream structure, where fused single-modality backbone features are fed into a multi-branch network to gather the information of multi-level features and derive modality-shared features. DMCF [30] proposes a multi-granularity dividing method to extract modality-shared features with more details of multiple granularities. The framework is shown in Fig. 2.27. Additionally, a strategy is proposed to apply different divisions to different layers. Since

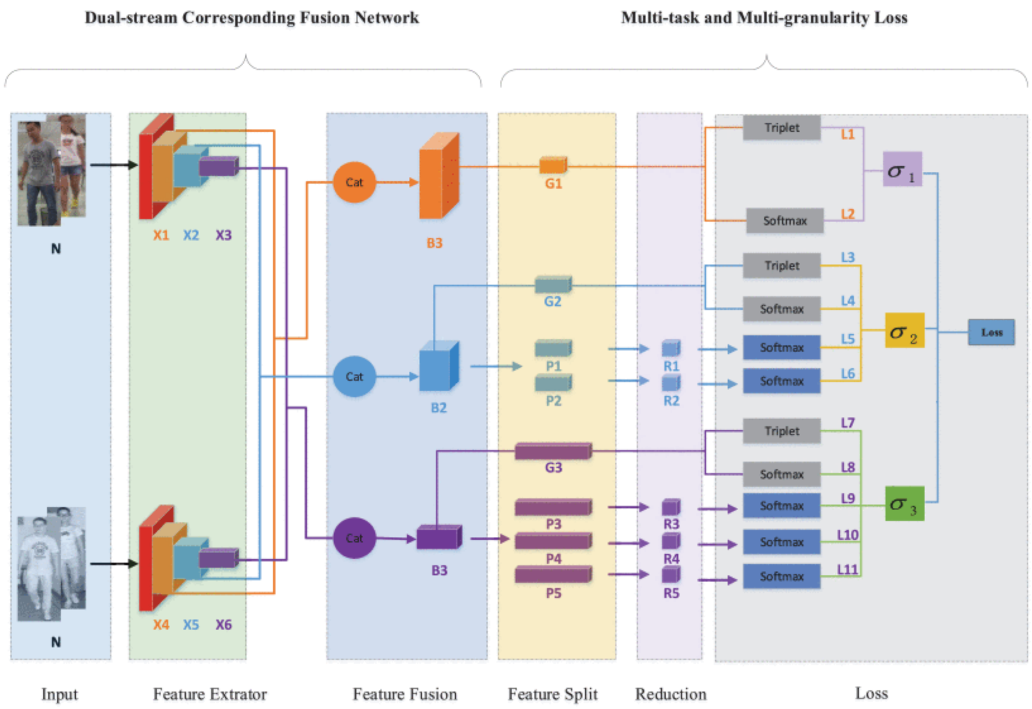


Figure 2.27: Framework of DMCF [30]. Single-modality features are fused according to the granularity and fused features are further divided to mine multi-level information to facilitate modality-shared feature learning. Taken from DMCF [30].

more effective information, as well as less noise present as the layer goes deeper, low-level features output by shallow layers, are not divided while middle-level and high-level features are divided into two and three parts, respectively. In terms of the framework, DMCF is also built upon the two-stream structure with the modification that corresponding layers of two single-modality backbone feature extractors are concatenated together and the parameter-shared blocks are applied on granularity-aware fused single-modality features.

- **Mining global and local information** aims to facilitate robust modality-shared feature learning by encouraging the network to focus on body structure instead of identity-irrelevant information, such as human pose or backgrounds. Specifically, an adaptive body partition (ABP) [192] is proposed to learn distinguishable body part representations. The framework of ABP is illustrated in Fig. 2.28. The work adopts the one-stream structure, where RGB images and IR images are first fed into a feature extractor to get modality-shared features. Then K-means clustering algorithm [129] is applied on backbone features to gather channels carrying information of similar body regions. The resulting M groups serve as M body parts, which are subsequently sent to multiple branches. Constraints are applied to global features and local (part) features separately to ensure the discriminability of features generated by

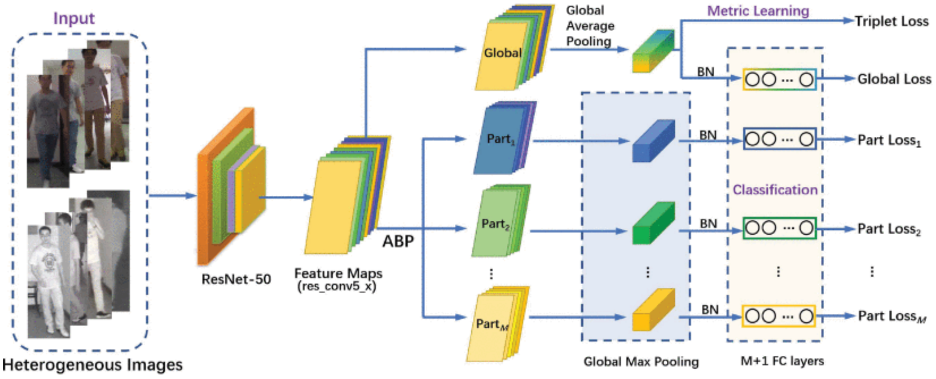


Figure 2.28: Framework of ABP [192]. ResNet-50 serves as the backbone feature extractor to extract modality-shared features, which are divided into M parts via K-means. Constraints are applied to each part to facilitate fine-grained information learning. Taken from ABP [192].

each branch. During inference, features from both global branches and local branches are concatenated for identity retrieval. Later, FBP-AL [193] is proposed to improve ABP with an adaptive weighting strategy, which highlights discriminative body parts during training and an adversarial learning scheme, which encourages the generator to generate common features for the same identity of both modalities and asks the discriminator to distinguish the modality that features comes from. Instead of dividing body parts via clustering, MSPAC [226] directly splits backbone features horizontally to represent body part features. Additionally, local features are gradually aggregated to global features by a network with cascading structure, resulting in enhanced modality-shared features embedded with low-level content information and high-level semantic information. The enhanced features are used for identity retrieval during inference. Despite that effective modality-shared features are extracted by the above methods, LbA [138] argue that such image-level and part-level based features are too coarse to handle the problem of body misalignment. Therefore, LbA aims to exploit pixel-wise features by establishing dense correspondences between RGB images and IR images and enforcing semantically similar local regions to be closely embedded. Such pixel-wise representations facilitate VI-ReID mainly from two aspects: 1) reducing the modality gap since pixel-wise features are modality-irrelevant, and 2) embedding features with more details due to the enforcement of local association.

- **Employing attention mechanism** in the learning of modality-shared features enables the detection of person-related regions or identity-relevant features where discriminative information generally exists. DLIS [200] encourages the network to focus on local patterns apart from the global information by the position attention-guided learning module (PALM). Such a module captures long-range dependencies via the self-attention mechanism [175], where inputs are interacted with each other to mine the relationship in-between so that the most "important" ones can be emphasized. Similarly, a dual-path deep supervision network (DDSN) [31] utilizes the self-attention mechanism to capture potential contextual cues within feature maps. Additionally, several dual-path deep supervision learning modules (DDSL) are inserted at multiple layers to obtain multi-level features, which are fused with contextual cues later to facilitate modality-shared feature learning. MPANet [198] aims to discover nuances in IR images, such as the type of pants or whether carrying bags, via attention mechanism. Specifically, MPANet includes a modality alleviation module and a pattern alignment module. The former alleviates the modality discrepancy via attention-guided instance normalization, where identity-relevant channels are emphasized while identity-irrelevant ones are dislodged. The pattern alignment module utilizes diverse pattern maps generated by the attention mechanism to discover the nuances in identity images. CoAL [191] proposes an end-to-end framework with two types of attention mechanisms: attention lifting and co-attentive learning. The former identifies the identity-relevant features from identity-irrelevant ones, thereby alleviating the intra-modality discrepancy. Whilst co-attentive learning mechanism reduces the gap between visible modality and infrared modality by performing the co-attentive interactions between modality-specific features and modality-shared ones to find out features that contribute more to the cross-modality identity retrieval.

- **Using graph convolutional networks (GCN)** in the modality-shared feature learning mainly has two advantages: 1) mining the relationship between features of the whole body (global) and body parts (local) to embed final representations with fine-grained information, and 2) mining the relationship between different identities across modalities to bridge the noticeable modality gap. Specifically, GLGCN [228] aims to mine the relation among different body parts via GCN to handle occlusions in VI-ReID, where body parts may be hiden by backgrounds. To achieve this, two types of graphs are built, including one built among different body parts within each modality and one built between embeddings of the same identity of two modalities. With both graphs mining local and global relations respectively, both intra- and inter-modality discrepancy are reduced. Similarly, DDAG [217] proposes a dual-attentive aggregation strategy to mine both intra-modality part-level and inter-modality graph-level contextual information for VI-ReID. Concretely, within each modality, the self-attention mechanism is applied to mine the relations between part-level features such that attentive parts can be assigned with higher weights. Additionally, cross-modality graphs are built upon single-modality features to mine the graphical relations between identity features of both modalities. In order to integrate two components adaptively, a parameter-free dynamic dual aggregation learning scheme is developed, enabling the joint optimization of modules with different objectives. Benefitting from contextual cues, CGRNet [53] built graphs at the context level (local and

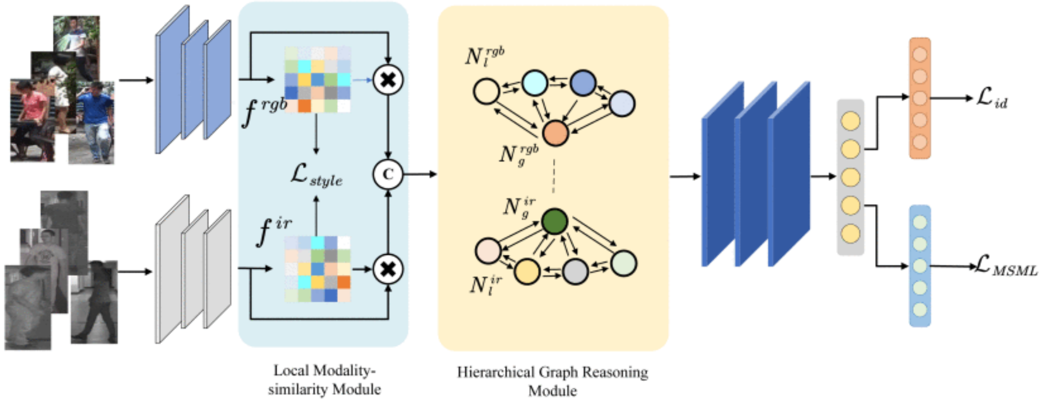


Figure 2.29: Framework of CGRNet [53]. The local modality-similarity module (LMM) minimizes the distance between style similarity grams of two modalities. The hierarchical graph reasoning module (HGR) models the relations across modalities and contexts. Taken from CGRNet [53].

global) and at the modality level (RGB and IR). The framework is illustrated in Fig. 2.29. CGRNet mainly consists of two modules: the local modality-similarity module (LMM) and the hierarchical graph reasoning module (HGR). Firstly, the style similarity grams are built upon modality-specific features and LMM is employed to reduce the distance between two grams such that features of the same identity in two modalities are pushed to each other in the latent space. Then single-modality features are squeezed along the channel and the graph is built upon global vectors and local vectors of both modalities. By mining relations in-between, the identity-consist information is inferred and can be presented as modality-shared features.

- **Optimizing BN layer** aims to solve the gap resulting from BN layers, including inter-mini-batch distribution gap and intra-mini-batch modality distribution gap [105]. The former refers to the gap between each mini-batch within each modality while the latter refers to the cross-modality gap within each mini-batch. To address the problem, a modality batch normalization (MBN) layer is proposed to divide each mini-batch into segments according to modality and the normalization is conducted within each modality such that both gaps can be reduced significantly. Similarly, CM-NAS [55] points out the importance of separating the BN layer in cross-modality person ReID performance boosting. Inspired by the observation, an objective is proposed to investigate the optimal separation scheme, which is illustrated in Fig. 2.30. As can be seen, the scheme provides the network with two options for each BN layer: adopting different parameters for visible modality and infrared modality via trainable separate parameters or adopting the same parameters for both modalities. With such a BN strategy, the network is encouraged to learn the optimal combination of layer-wise BN strategy that contributes to a better

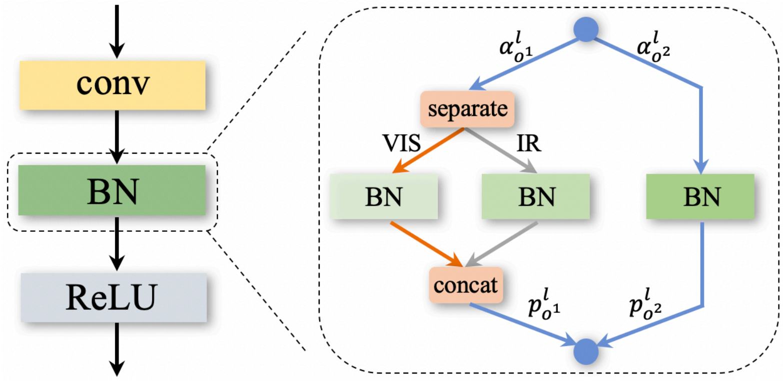


Figure 2.30: Illustration of BN separation in CM-NAS [55]. Two options are provided for BN layer: adopting different parameters for visible modality and infrared modality via trainable separate parameters or adopting the same parameters for both modalities. Taken from CM-NAS [55].

cross-modality identity retrieval performance.

- **Optimizing classifier** aims to improve the modality-shared classifier in existing two-stream structure based frameworks. Specifically, a modality-aware collaborative (MAC) [214] learning scheme is proposed to impose additional constraints apart from that applied to the modality-shared classifier, which is shown in Fig. 2.31. As can be seen, three additional classifiers with different objectives are involved, including two modality-specific classifiers and a modality classifier. The former two aims to capture modality-specific identity-related information at the classifier level while the latter one serves as a discriminator to distinguish which modality that current inputs belong to. Additionally, a collaborative learning strategy is developed to coordinate the joint learning with modality-shared classifiers and modality-specific ones. Based on MAC, they further develop a collaborative ensemble learning strategy in the extension work [216]. Such an ensemble scheme improves the knowledge transfer among different classifiers, which provides informative guidance for each classifier, facilitating the learning of more representative features. Following the idea of knowledge transfer and mutual learning, two modality-specific nearest neighbour classifiers are exploited in CMSP [195]. By conducting the inner product between features and classifier parameters that are assigned by single-modality features, the intra-modality similarity scores and cross-modality ones can be calculated. A modality-aware similarity-preserving loss is applied with similarity scores to enforce the consistency of same-modality retrieval results and cross-modality ones.

Feature disentanglement based methods aim to extract identity-related features by filtering out modality-specific ones from single-modality backbone features. The pipeline of feature disentanglement-based methods is shown in Fig. 2.25(b),

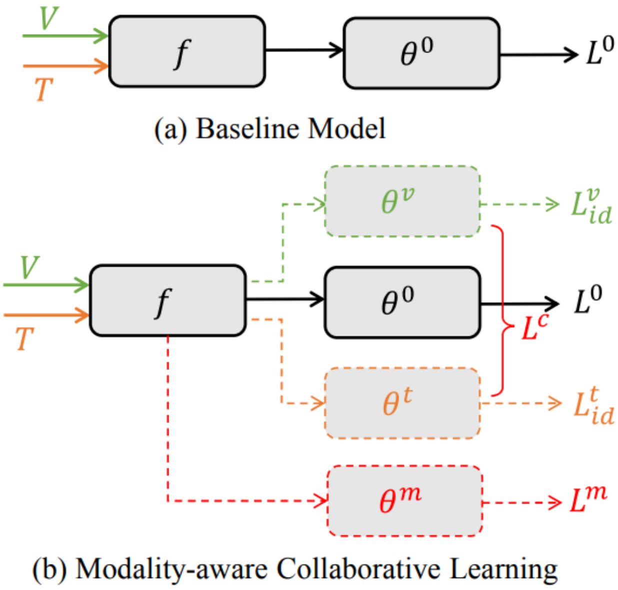


Figure 2.31: Comparison of models with (a) two-stream structure and (b) MAC [214] learning scheme. V and T refers to visible modality and thermal modality, respectively.
Apart from the modality-shared classifier (θ_0), MAC employs two modality-specific classifiers (θ_v and θ_t) and a modality classifier (θ_m). Taken from MAC [214].

where single-modality features are decomposed into two parts: modality-shared features and modality-specific features. The former is what the cross-modality identity retrieval requires. MSR [54] adopts the two-stream structure with two branches attaching to single-modality backbone features to extract modality-specific and modality-shared features separately via the discriminant metric and the identity loss, respectively. Following the structure, SDL [92] attempts to disentangle single-modality features into the identity-related features, *i.e.,* modality-shared ones, and spectrum-disentangled features, *i.e.,* modality-specific ones. SDL names two branches the spectrum dispelling branch and the spectrum distilling branch, respectively. The former aims to extract identity-related features under the supervision of identity loss while the latter aims to distil spectrum-specific cues under the supervision of identity-dispeller loss. Additionally, to further ensure that the two parts are well separated, SDL adopts a disentanglement loss where the loss between single-modality features and the summation of modality-shared features and corresponding spectrum-disentangled features is minimized.

As a departure from previous disentanglement methods, Hi-CMD [33] disentangles features by image reconstruction with generative models. Specifically, single-modality features are disentangled into ID-discriminative factors and ID-excluded factors. Given a pair of RGB-IR features for a person, a decoder is applied to reconstruct a pair of images with their swapped ID-excluded factors, which are composed of illumination attribute code and pose attribute code. By imposing the recon-

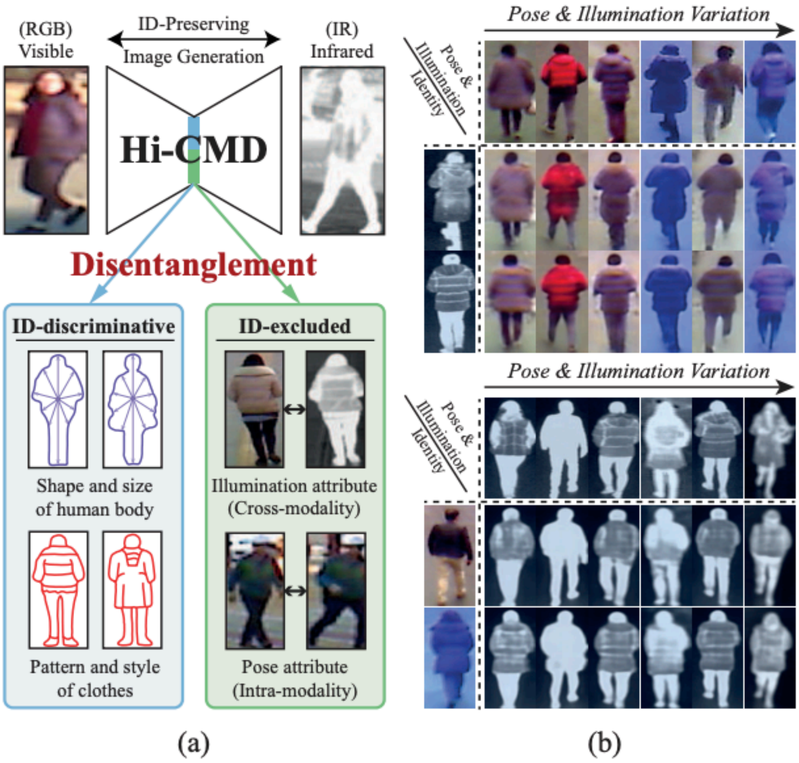


Figure 2.32: (a) Pipeline of Hi-CMD [33]. The method aims to disentangle ID-discriminative features from ID-excluded ones by image reconstruction. (b) Examples of images generated by Hi-CMD in visible modality and infrared modality. As can be seen, Hi-CMD is able to preserve the identity information as well as generating high-quality images. Taken from Hi-CMD [33].

struction loss on the generated pair, the generator learns how to encode and decode ID-excluded factors. Meanwhile, additional reconstruction loss terms are applied to images that are reconstructed within each modality to ensure that ID-discriminative features, including style attribute code and prototype code, are well preserved during the cross-modality reconstruction while ID-excluded ones are well preserved during the same-modality reconstruction. Examples of generated image pairs by Hi-CMD are shown in Fig. 2.32(b). As can be observed, with the elaborate design, Hi-CMD is capable of disentangling ID-discriminative features from ID-excluded ones and generating high-quality images. At the same period, DG-VAE [141] is proposed to disentangle the single-modality features into identity-discriminable information (IDI) and identity-ambiguous information (IAI) via variational auto-encoder (VAE). IDI refers to cues such as the shape and contour of the body, which can be used to distinguish one person from the other. In contrast, IAI refers to cues like spectrum characteristics, which confuses the VI-ReID. To solve the problem that the standard Gaussian distribution cannot well handle the structural discontinuity between disparate classes of IDI, DG-VAE embeds IDI with mixture-of-gaussian (MoG) distribution, which considers intra-class variations and inter-class separability at the same time. Additionally, a triplet swap reconstruction (TSR) strategy is adopted to enhance disentanglement via squeezing IDI and IAI into separate branches.

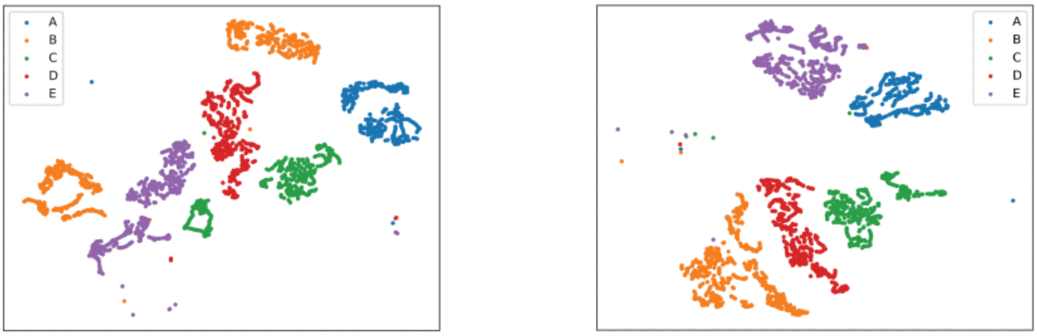
Metric learning based methods improve optimization objectives of identity feature learning, where the network is enforced to extract features with specific properties. As shown in Fig. 2.24, existing loss functions can be categorized as identity loss, contrastive loss, triplet loss, and center loss, which will be detailed as follows.

- **Identity loss** aims to embed identity-related information into the learned features. Generally, it takes the formulation of cross entropy (CE) loss, which is widely used in classification. Differently, HSME [74] argues that existing methods ignore the correlation between the classification loss and the feature embedding loss. Therefore, they propose to use the sphere softmax function, which aims to build a hypersphere embedding space where intra-modality variations and cross-modality variations are constrained.

- **Contrastive loss** aims to bridge the gap between visible modality and infrared modality by enforcing the distance relations between sample pairs. For example, HCML [212] adopts the two stream structure with an additional contrastive loss applying on backbone features, which enforces that the distance between cross-modality negative pairs (belonging to different persons) is larger than that between cross-modality positive pairs (belonging to the same person) by a margin in the latent space.

- **Triplet loss** aims to constrain the distance relationship among triplets built within single-modality or across heterogeneous modalities. For example, a bi-directional top-ranking (BDTR) [213] loss is proposed, which considers both cross-modality constraint and intra-modality constraint. The former bridges the modality gap by enforcing the distance between the anchor and its furthest cross-modality positive should be smaller than that between its nearest cross-modality negative ones by a margin. The latter focuses on the triplet relationship within the modality, *i.e.,* the anchor should be closer to its positive samples than the hardest negative ones. Instead of imposing the triplet loss in the whole feature space, the modality alignment (HMA) [185] loss firstly mines a hard feature subspace with noticeable modality discrepancy and then conducts the triplet loss within the hard subspace to mitigate the imbalance of modality discrepancies. Instead of Euclidean distance, a bi-directional exponential angular triplet (expAT) [210] loss leverages the cosine distance to describe the distance relationship between samples in the latent space, such that angularly separable features can be learned. The feature distribution of models trained with plain triplet loss and expAT loss are shown in Fig. 2.33. As can be seen, more separable features can be derived when the model is trained with expAT loss.

- **Center loss** aims to regulate the identity-wise relationship instead of sample-wise ones, such that the network will not be focusing on meaningless de-



(a) Triplet Loss

(b) expAT Loss

Figure 2.33: Visualization of feature distribution of models trained with (a) triplet loss and (b) expAT loss [210]. Taken from expAT [210].

tails, such as human poses, and accessories. Specifically, the hetero-center (HC) [261] loss is proposed to regulate the distance between feature centers of visible modality and infrared modality, where heterogeneous features of the same identity are pulled closer. Integrating with the CE loss, discriminative modality-shared features are learned with the baseline two stream network. On top of it, PSENet [116] upgrades HC loss into a triplet version, where cross-modality feature centers of different identities are pushed away apart from the distance constraint applied on cross-modality positive center pairs. eBDTR [215] extends the bi-directional dual-constrained top-ranking loss (BDTR) [213] to the center-based ones, where the previous cross-modality constraint and intra-modality constraint are incorporated into a single formula.

2.4.2 Modality-specific information compensation

In addition to extracting modality-shared features, some works [180, 237, 34, 252, 126, 188, 85, 230] facilitate VI-ReID by compensating modality-specific information. According to the modality of generated information, the compensation can be categorized as single-modality information compensation and cross-modality information compensation. The differences between the two types are shown in Fig. 2.34(b) and Fig. 2.34(c). The single-modality information compensation only generates information in a single modality while cross-modality information compensation conducts the generation in both modalities.

Single-modality information compensation is generally conducted by generating fake images from one modality to another so that the matching can be conducted within a single modality, thereby alleviating the cross-modality discrepancy.

Specifically, AlignGAN [180] employs a pixel alignment module to generate fake infrared (IR) images from real RGB images, which adopts the structure of CycleGAN [258]. With the min-max training scheme alternating between the generator

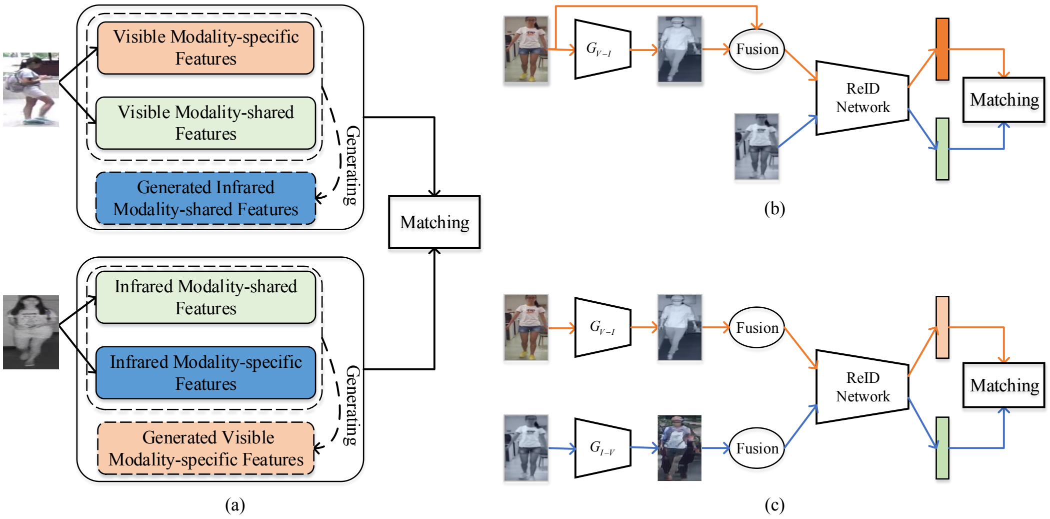

Visible Modality-specific Features
Visible Modality-shared Features
Generated Infrared Modality-shared Features
Generating
Matching
Infrared Modality-shared Features
Infrared Modality-specific Features
Generated Visible Modality-specific Features
Generating
(a)
G_{V-I}
Fusion
ReID Network
Matching
(b)
G_{V-I}
Fusion
G_{I-V}
Fusion
ReID Network
Matching
(c)

Figure 2.34: (a) Framework of FMCNet [230]. Pipelines of modality-specific information compensation methods, including (b) single-modality ones and (c) cross-modality ones.

and the discriminator, RGB images are transferred into IR-like ones. Correspondingly, the cross-modality matching is transferred to an intra-modality one performing between fake and real IR images at the feature level. Later, TS-GAN [237] proposes a teacher-student GAN model, where the teacher model is pretrained on real IR images and then guides the learning of the student model. The student model takes as input both real RGB and fake IR images, which are generated by GAN following the scheme of AlignGAN.

Apart from the visible-infrared generation, some works focus on the generation in an inverse way, *i.e.,* generating fake visible images from infrared ones. For example, CE^{2}L [34] converts the infrared image into visible ones via CycleGAN and conducts the intra-modality matching in visible modality. Instead of using the generative model to obtain visible images, an alternative way is to utilize the colorization. The comparison of pixel correspondences between the two types is shown in Fig. 2.35. As can be seen, due to the lack of single-channel (infrared) to three-channel pixel-wise ground-truths, the generation process is generally unstable and the evaluation of generation quality is ambiguous. Therefore, a colorization network, namely GECNet [252], is proposed to colorize the infrared images into RGB ones. To handle the channel discrepancy between infrared and visible images, point-wise grayscale visible images serve as single-channel ground-truths of the colorization network optimization. Examples of colored images are shown in Fig. 2.36

Cross-modality information compensation generally involves the bidirectional generation between visible modality and infrared modality at the image level or the feature level.

Specifically, D^{2}RL [188] proposes a dual-level discrepancy reduction learn-

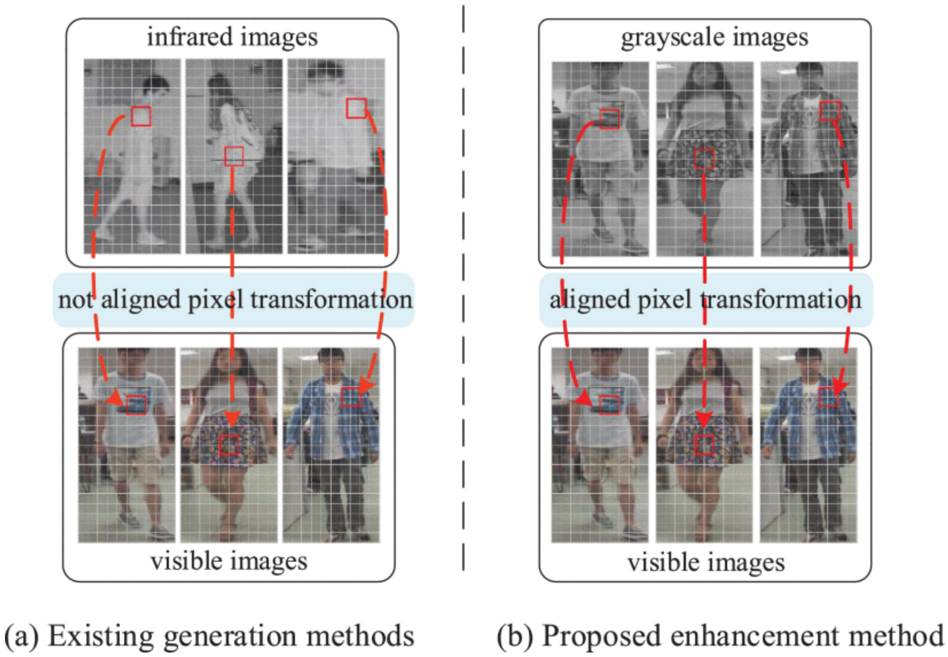

infrared images
not aligned pixel transformation
visible images
grayscale images
aligned pixel transformation
visible images
(a) Existing generation methods
(b) Proposed enhancement method

Figure 2.35: Pixel correspondences of cross-modality images generated by (a) generative models or (b) colorization. Taken from GECNet [252].

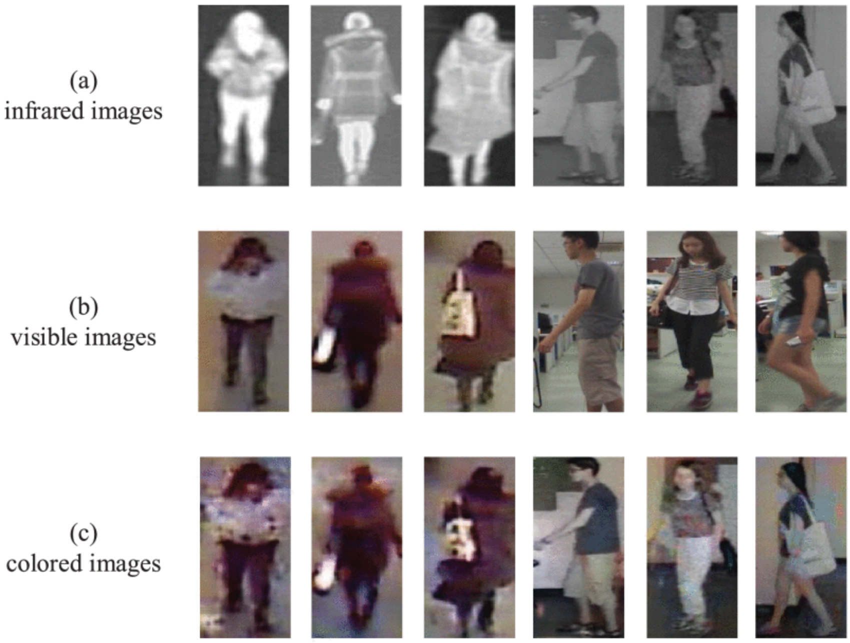


Figure 2.36: Visualization of (a) infrared images, (b) visible images, and (c) colored images. Each column presents the images of the same person. Taken from GECNet [252].

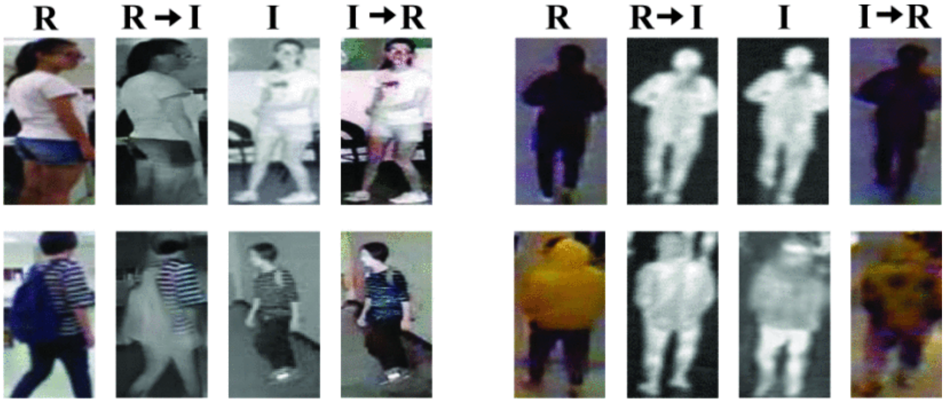


Figure 2.37: Visualization of original images (R, I) and generated images (R→I, I→R) by ADCNet. Taken from ADCNet [85].

ing scheme, where cross-modality discrepancy and intra-modality discrepancy are handled separately. Specifically, to reduce the gap between two modalities, infrared images and visible images are translated mutually via an image-level sub-network. On top of unified multi-spectral images, a feature-level sub-network is trained to handle the intra-modality discrepancy, such as pose variation. ADCNet [85] designs a feature disentanglement network, which consists of an auto-encoder and an adversarial learning module. Paired-images with different modalities are generated by swapping the modality-shared features. By enforcing the visual appearance of images in each modality, modality-shared information can be well disentangled from modality-specific ones. Subsequently, a feature alignment network with multiple second-order correlation blocks is employed to capture long-range relationships underlying modality-shared features, thereby leading to a satisfactory VI-ReID performance. Examples of generated images are shown in Fig. 2.37.

Instead of the image-level compensation, cm-SSFT [126] employs a shared-specific feature transfer to enable the flow between modality-shared and modality-specific features, which serve as the compensation to each other. Specifically, based on the affinity of features in both modalities, the learning of each sample can collect information from its nearest neighbors within both modalities. Such feature-level compensation improves the feature representability by taking advantage of specific information without involving unstable image generation. Following the spirit of feature-level compensation, FMCNet [230] makes up missing modality-specific features of one modality from the modality-shared features of the other, as illustrated in Fig. 2.34(a). Instead of collecting information from neighboring samples, FMCNet performs the feature compensation within the modality-specific features and modality-shared features of the sample itself. Specifically, single-modality features are firstly detangled into the modality-specific part and the modality-shared part. Later, the modality-shared features of one model are used to compensate for the missing modality-specific part of the other model via a feature-level modality compensation module. Embedding with more completed information, more discriminative person-related features can be learned.

2.4.3 Auxiliary information

Empowered by advances in the computer vision community, auxiliary information, including people masks, body keypoints and auxiliary modality, are involved to facilitate identity feature learning.

People mask and pose estimation. To reduce the background noise, some works aim to filter out the human region via people masks, which generally are binary maps with backgrounds being 0 and people being 255. Examples of people masks can be found in Fig. 2.38, indicating by V(IR)-mask images. For example, MDAN [142] utilizes people masks generated by a human phrasing method [106] to filter out human regions. The framework of MDAN is shown in Fig. 2.38. As can be seen, features of original images and human parts are learned by separated branches and combined by inner product, where human parts can be highlighted during training. By applying multiple constraints on modality-specific features and modality-shared features, representative representations can be learned by the framework. Instead of masking out the human part, Zhao *et al.* [242] aims to filter out the clothing region such as tops, pants and shoes. By applying random color jitter augmentations on clothing regions, such as adjusting the illumination or brightness, while keeping non-clothing parts unchanged, people with diverse clothing color can be simulated. Such clothing color transformation operation enables the network to learn color-irrelevant features.

Apart from people masks, body joints, serving as an explicit modality-shared cue, are also introduced in VI-ReID. For example, SPOT [19] exploits the body structure information obtained from key point heatmaps derived by OpenPose [16] and embedded them into feature maps to highlight critical areas.

Auxiliary modality. One of the reasons for the poor performance of VI-ReID is that the significant gap between visible modality and infrared modality makes the learning of modality-shared features very difficult. Therefore, some works start to use auxiliary modality as an intermediary to mitigate the modality gap.

Specifically, homogeneous augmented tri-modal (HAT) [218] introduces an auxiliary grayscale modality as a transition between the matching across visible modality and infrared modality. The grayscale modality not only preserves the structure information of RGB images but adapts the image style of IR images. To achieve the tri-modality optimization, a triplet-mining ranking loss is applied on triplets built among three modalities to constrain the distance relationship. For example, to build a triplet for the anchor from the visible modality, positive samples and negative samples are selected from the infrared modality and grayscale modality, respectively. More combinations can be found in Fig. 2.39. The introduction of grayscale modality enables the network to be optimized with more informative triplets, thereby improving the robustness of learned features. XIV [101] proposes

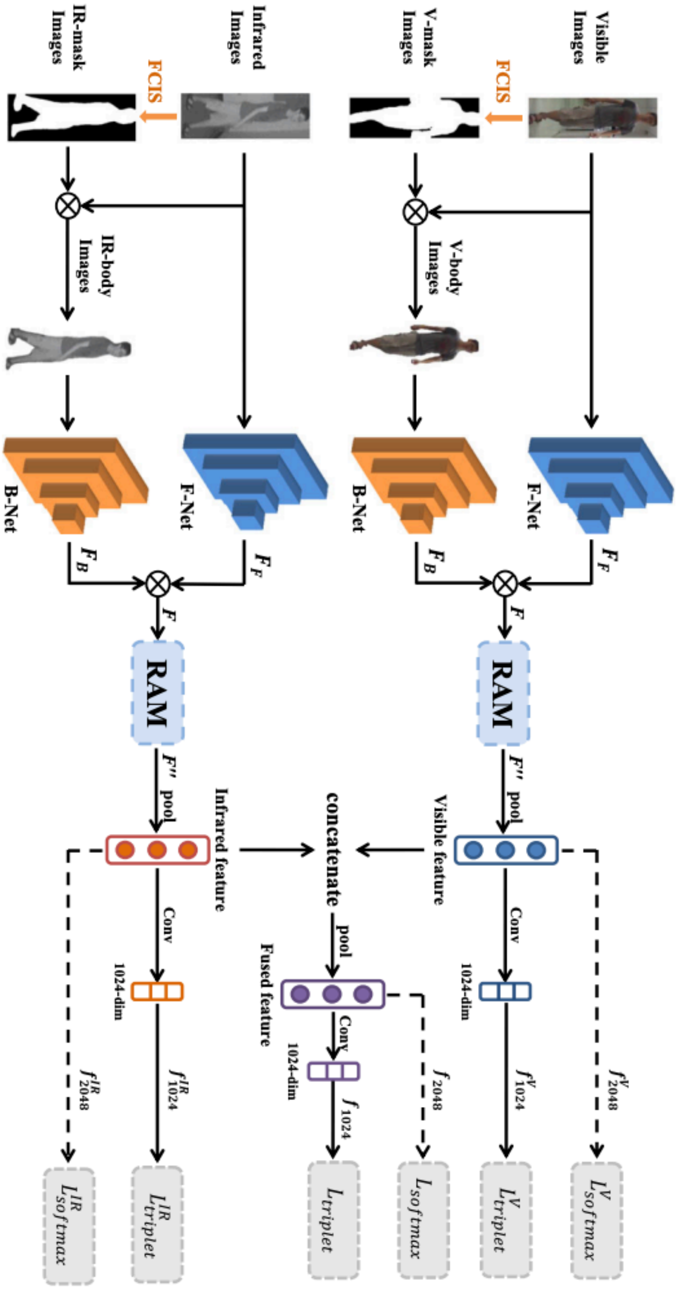


Figure 2.38: Framework of MDAN [142]. The network consists of two branches to extract features from RGB images and IR images respectively. Each branch takes two types of inputs: original images and corresponding people masks generated by FCIS [106]. Correspondingly, F-Net and B-Net are used to extract features from original images and people regions, respectively. Enhanced features F are obtained by combining two groups of features via the inner product. Then, Residual Attention Module (RAM) is applied to enhanced features to improve the discriminability. Multiple constraints are imposed on visible features, infrared features as well as concatenated ones to obtain representative features. Taken from MDAN [142].

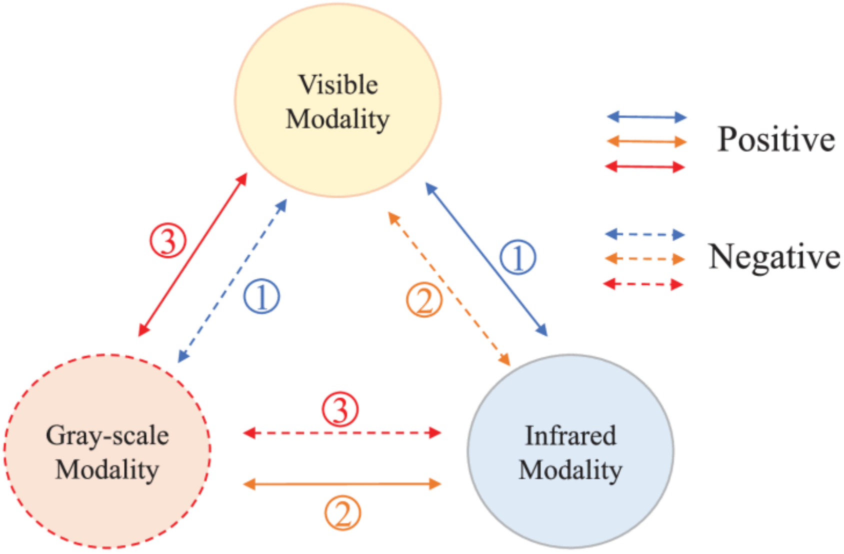


Figure 2.39: Illustration of tri-modality learning, which involves three modalities: visible modality, infrared modality, and gray-scale modality. The triplet-mining ranking loss is applied on triplets, *i.e.,* anchor-positive-negative, built among three modalities, which are ① visible-infrared-grayscale, ② infrared-grayscale-visible, and ③ grayscale-visible-infrared. Taken from HAT [218].

an X-Infrared-Visible ReID cross-modal learning framework, where the X modality is generated by a lightweight network in a self-supervised manner. As shown in Fig. 2.40, X images look like the intermediate status in the transition from visible to infrared images. Compared to visible images, X images seem to carry more red information, *i.e.,* with a longer wavelength, while compared to infrared images, X images seem more colorful, *i.e.,* with a shorter wavelength. By introducing an intermediate modality into the matching problem across visible and infrared modalities, the inherent cross-modality gap can be mitigated.

2.4.4 Data augmentation

Although different kinds of data augmentation are adopted during training, such as random cropping, random erasing, and random flipping, they fail to fully consider imagery properties. However, the key factor differentiating VI-ReID from ReID is the color information resulting from a different spectrum. Some works, therefore, have investigated how color-relevant data augmentation can help cross-modality feature learning.

For example, CDP [52] proposes a multi-spectrum image generation method to generate samples in a different spectrum, including blue, green, red, gray spectrum, which is shown in Fig. 2.41. Such augmented data facilitates the spectrum-irrelevant feature learning. To narrow down the significant modality gap between

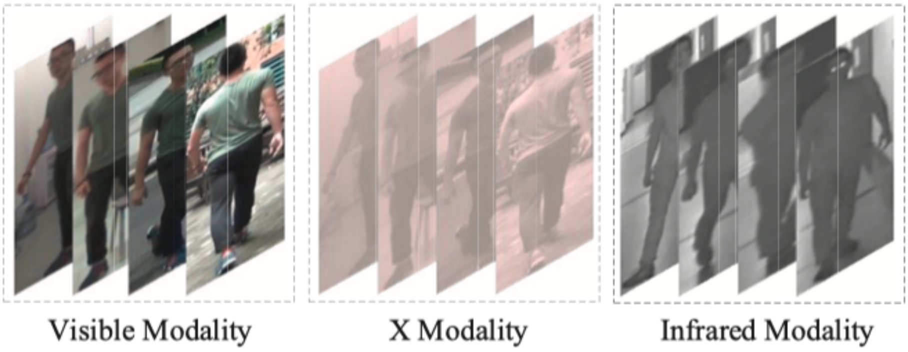


Figure 2.40: Comparison of visible images, X images and infrared images. Taken from XIV [101].

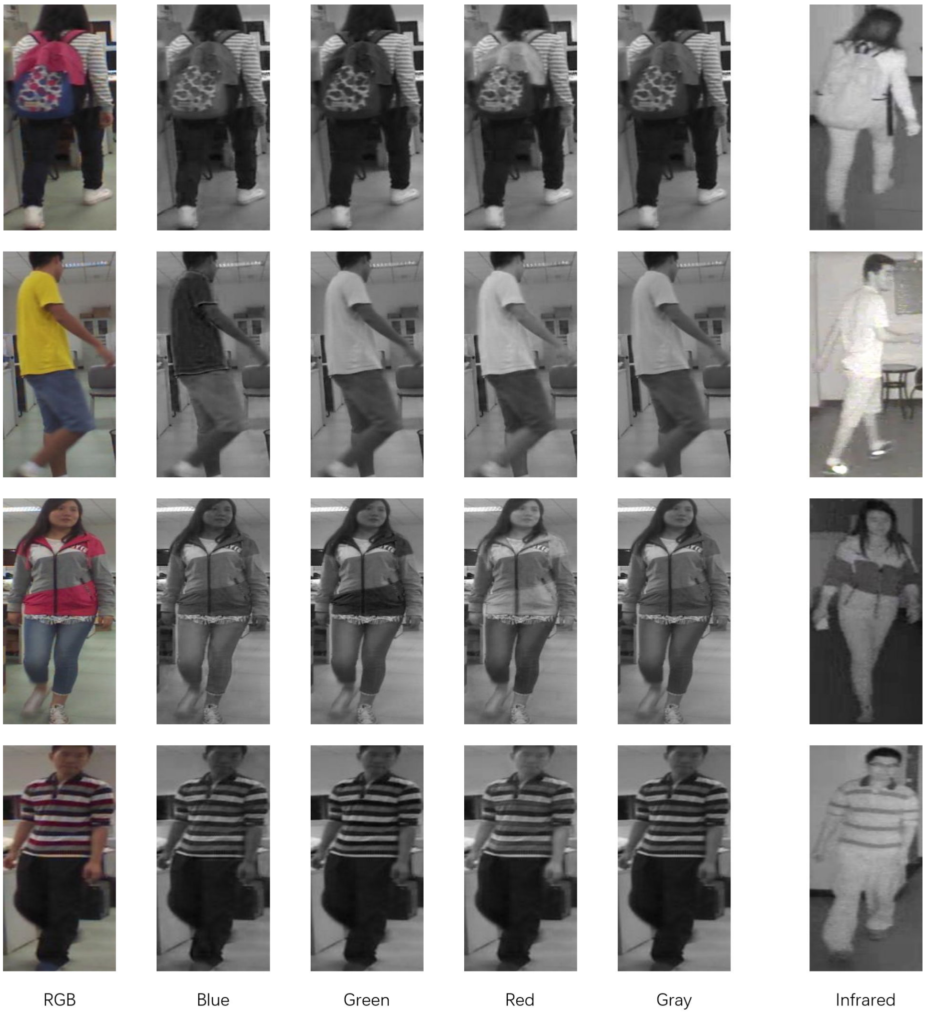


Figure 2.41: Examples of images generated by the cross-spectrum scheme. First column: original images. Second to fifth columns: images generated in blue, green, red, gray spectrum. Six column: corresponding infrared images. Taken from CDP [52].

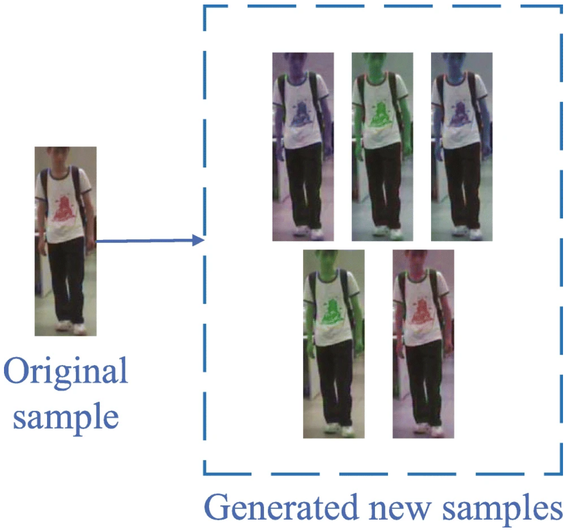


Figure 2.42: Examples of images generated by the channel swapping strategy. Taken from CPN [117].

visible modality and infrared modality, CAJL [219] proposes a channel exchangeable augmentation, where one of the channels (R, G, B) of visible images is randomly selected to replace the other two to form new inputs. The channel reorganization strategy encourages the network to match infrared images with each color channel of visible images, which reduces the learning difficulty. To better adapt the proposed augmentation in VI-ReID, a channel-augmented joint learning scheme is proposed, where channel-augmented images are considered as an auxiliary modality. To this end, the original cross-modality matching is formulated as a tri-modality matching problem, as discussed in Section 2.4.3. Following the spirit of channel-wise data augmentation, instead of composing images with three identical channels, CPN [117] proposes to generate newly visible images by directly randomly swapping R, G, B channels, as shown in Fig. 2.42. Although the generated samples lack authenticity, it is beneficial to color-irrelevant feature learning.

2.5 Conclusion

In this chapter, the basic methodologies of plain person ReID in terms of feature learning and metric learning are firstly reviewed, and the common evaluation metrics for person ReID: mAP and CMC are introduced. Based on plain person ReID settings, three more challenging scenarios are discussed: unsupervised person ReID, fast unsupervised person ReID, and cross-modality person ReID. For each scenario, its differences with plain ReID settings are firstly discussed and existing works are elaborately reviewed. In the following three chapters, the author will introduce her solutions towards these three challenging person ReID problems, the advancement over existing works, and the performance comparisons with state-of-the-art methods.

Chapter 3

Unsupervised person re-identification

3.1 Introduction

Person re-identification (ReID) aims to retrieve a person of interest across multiple cameras [220, 247, 104]. Due to the label-free training manner, unsupervised person ReID has been attracting increasing attention. Existing unsupervised ReID methods can be broadly categorized into two types: unsupervised domain adaptation (UDA) methods [254, 234, 60, 36, 201, 82] and purely unsupervised learning (USL) methods [37, 20, 32, 184, 235]. The former pre-trains a model on person-related datasets, *i.e.,* source domain, and fine-tunes it on ReID-related datasets, *i.e.,* target domain. Apart from requiring additional annotated labels, UDA methods are vulnerable to the large gap between the source domain and the target domain. In contrast, USL methods do not require any labeled data for training, which are more challenging but well fit real-world scenarios. In the work, the author focuses on USL methods.

Existing USL methods generally follow a two-stage training scheme: 1) clustering, *i.e.,* obtaining the pseudo labels via a clustering algorithm such as DBSCAN [50], and 2) network training, *i.e.,* optimizing the network in a "supervised" manner with assigned cluster IDs. Contrastive loss such as InfoNCE [60] or ClusterNCE [37] usually serves as training objectives. Due to the blind trust in imperfect clustering results, the learning is inevitably misled by unreliable pseudo labels, where multiple identities are merged into one cluster or samples of one person are assigned to multiple clusters. Despite that some pseudo label refinement [184, 20, 233, 235, 32] have been proposed, they generally leverage auxiliary information, such as camera IDs [184, 20], body part predictions [32], and generated samples [235]. Given the fact that such auxiliary information is not free in reality, refining pseudo labels by merely exploiting internal characteristics within samples, *i.e.,* the sample-wise clustering confidence, appears to be more valuable.

To measure the sample-wise clustering confidence, *i.e.,* how well a sample fits its cluster, a metric: silhouette score [152] is used in this work. The score

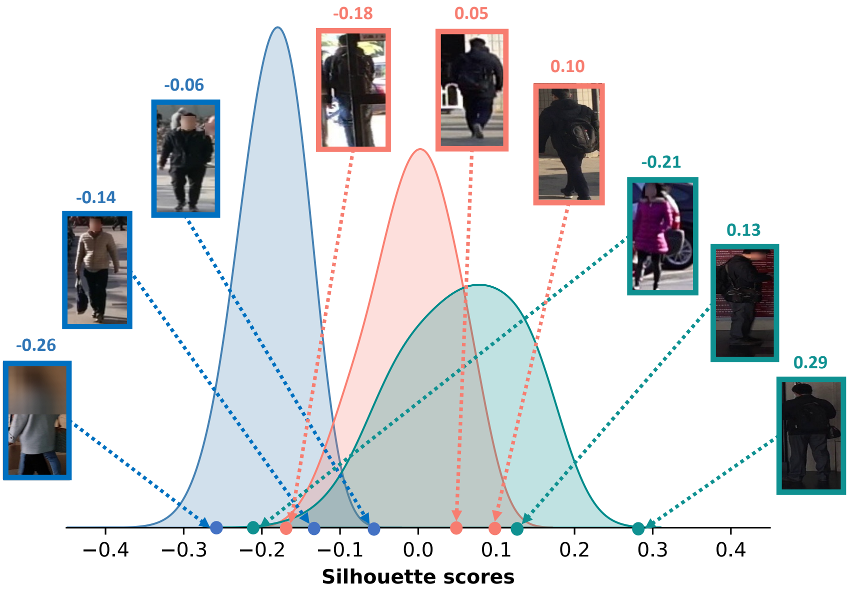


Figure 3.1: Training samples (cluster ID = 1) and corresponding silhouette scores at epoch 0 (blue), epoch 25 (orange), and epoch 50 (green) on MSMT17 [190]. Higher silhouette scores denote samples are clustered at higher confidence. **Best viewed in color**.

presents the ratio between intra-cluster distance and inter-cluster distance, which ranges from -1 to +1 (*higher is better*). To demonstrate the relationship between the clustering confidence and the silhouette score, we visualize silhouette scores of training samples of MSMT17 [190] in Fig. 3.1. Samples are from the same cluster (cluster ID=1) but at different training epochs, *i.e.,* 0, 25, and 50, respectively. As training goes on, the clustering is gradually enhanced by involving more effective features and a more discriminative network. At first, images are grouped by coarse visual features, yet by identity-related information in the end. Meanwhile, sample-wise silhouette scores continuously shift towards higher values during training. Given this consistency, a conclusion can be drawn that, *a higher silhouette score implies the sample better fits its cluster, i.e., being clustered at higher confidence.* Previous learning schemes [37, 235] adopt all-sample based centroids, which are obtained by averaging features of all samples within the cluster, and enforce instances to approach such centroids. However, the observation suggests that low-confidence samples either are poor in quality or belong to other identities. Features of such images will inevitably contaminate centroids regardless of the training stage. In light of this, Confidence-Guided Centroids (CGC) are proposed to provide more reliable cluster-wise prototypes for feature learning.

Although the reliability of cluster centroids has been improved, the conventional one-hot labeling strategy aggravates a problem. Since high-confidence samples exclusively contribute to the formation of cluster centroids, the identity-related information of low-confidence samples can hardly be presented in the assigned centroid. To illustrate the problem, an analysis is conducted on MSMT17 [190], where the author intends to investigate how much identity information of low-confidence samples can be presented in their assigned centroids. The author finds that, with the plain all-sample based cluster centroids, only 5.83% low-confidence samples have their identity information embedded in the assigned centroid at the beginning. Although the ratio gradually climbs to 17.19%, a large proportion of low-confidence samples (over 80%) still are pushed to "wrong" centroids. Unfortunately, the ratio achieves 14.17% at most with confidence-guided centroids. Given the situation, the one-hot labeling strategy, which enforces samples to learn from the assigned centroid solely, is unwise. To address the problem, the author proposes to use confidence-guided pseudo labels (CGL), which encourages instances to approach not only the assigned confidence-guided centroid but also others where their identity information is potentially embedded.

In summary, the contributions of this work are as follows:

- Confidence-Guided Centroids (CGC) are proposed to provide cluster-wise prototypes for feature learning. The reliability of centroids is improved via filtering out low-confidence samples during formation.
- To overcome the problem that the identity information of low-confidence samples is rarely presented in their assigned centroids, the author proposes to use confidence-guided pseudo labels (CGL) during training. Apart from the originally assigned centroid, instances are also encouraged to approach other centroids where their identity information is potentially embedded.
- The proposed method only exploits internal characteristics for unsupervised person re-identification. Extensive experiments on benchmark datasets demonstrate that the proposed method yields better or comparable performances with state-of-the-art ones that largely leverage auxiliary information.

3.2 Previous solutions

3.2.1 Unsupervised learning

As a main learning paradigm of machine learning (ML), unsupervised learning aims to discover data patterns without label annotations [9]. Tasks in unsupervised learning can be categorized into three types:

- **Clustering.** Grouping data based on similarity.
- **Density estimation.** Modelling the distributions of input data.
- **Dimension reduction.** Representing high-dimensional data (features) in low-dimension space.

Since unsupervised person Re-ID is a task which aims to group unlabelled data according to the identity information, in this chapter, we mainly focus on the clustering task.

3.2.2 Clustering

Data clustering is defined as the process of segmenting a group of instances into subgroups of similar ones, *i.e.,* clusters [90]. Generally, clustering algorithms can be categorized, but not limited to, hierarchical clustering, partitioning clustering, density-based clustering, *etc.*

Hierarchical clustering generally forms clusters in two ways: 1) divisive, and 2) agglomerate. The former starts from a broader cluster that includes all points, which is then split into more specialized sub-clusters. On the contrary, the latter starts from the smallest clusters and builds up until the entire set is included in a single cluster. Despite the clear hierarchical relationship among clusters, the incorrect clustering decisions cannot be reversed [11].

Typical hierarchical clustering approaches include BIRCH [232], CURE [69], Chameleon [93], *etc.* BIRCH [232] organizes features with a clustering feature (CF) tree, which dynamically grows when new data points appear. CURE [69] adopts the idea of random sampling, and obtains the final clustering by integrating the partial clustering of partitioned samples. Chameleon [93] partitions data points into sub-clusters with a k-nearest-neighbor graph, where neighboring samples are connected with edges. Then subclusters are merged repeatedly according to the similarity.

Partitioning clustering aims to iteratively relocate data points into k groups based on the characteristics and similarity until an optimal grouping is attained. k is a user-specified number.

Typical partitioning clustering methods include K-means [129], K-Medoid [148], *etc.* K-means [129] firstly initializes k cluster centroids with data points randomly chosen from the dataset. Then the clustering process is iteratively repeated between 1) cluster assignment, where the remaining data points are assigned to the closet clusters with the minimum distance to the corresponding centroid, and 2) centroid update, where cluster centroids are updated by the mean of data points within the corresponding cluster. The clustering stops when the stable cluster assignment is achieved. K-mediods [148] improves K-means by replacing the representation of centroids (intra-cluster mean features) with one of the data points in the cluster, which improves the robustness against outliers.

Density-based clustering detects areas where data points concentrate, *i.e.,* with high density. Data points lying in low-density areas are typically regarded as noise or outliers [96]. Since density-based clustering methods do not require presetting the number of clusters, they naturally conforms to the unsupervised person ReID task where the number of identities is unknown, thereby being widely used in unsupervised ReID.

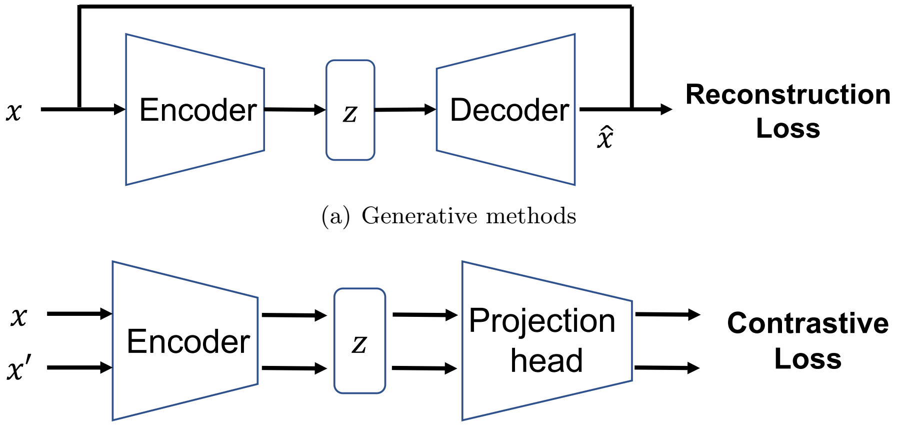
The most popular density-based clustering method is Density Based Spatial Clustering of Applications with Noise (DBSCAN) [50]. DBSCAN requires two hyperparameters: 1) the maximum distance between neighbouring points (*eps*), and 2) the minimal number of points within a region (*min_num*). During clustering, DBSCAN encourages instances to merge their neighbours whose distances are less than *eps* into the cluster. Instances that, have fewer neighbours within the cluster (less than *min_num*) or unreachable, are regarded as outliers. The process stops when all density-connected clusters are completely found.

3.2.3 Self-supervised learning

Recently, an emerging branch of unsupervised learning, *i.e.,* Self-Supervised Learning (SSL), has drawn extensive attention. SSL aims to leverage the relations of input data or the underlying data structure as supervisory signals for discriminative feature learning [89, 122]. Generally, SSL can be roughly categorized as follows [122]:

- **Generative methods** aim to learn the latent code z by reconstructing the inputs x.
- **Contrastive methods** aim to learn the latent code z by maximizing the agreement between original inputs x and the corresponding augmentations x'.

The pipelines of generative methods and contrastive methods are illustrated in Fig. 3.2. As can be seen, the main differences lie in the framework design and objectives.



(a) Generative methods

(b) Contrastive methods

Figure 3.2: Pipelines of generative methods and contrastive methods.

- **Framework design.** Generative methods utilize a decoder to reconstruct the inputs x from latent codes z, while contrastive methods map use a projection head to map z to a low-dimension latent space.
- **Objectives.** Generative methods adopt the reconstruction loss as objectives, whereas contrastive methods adopt contrastive loss.

More details of generative methods and contrastive methods can be found as follows.

Generative SSL. Generally, due to the ability to recover the original data distribution regardless of downstream tasks, generative SSL learning is widely used in the NLP field to conduct text classification [122]. In the computer vision community, the most frequently used generative model for the SSL task is Auto-encoder (AE).

Generally, an AE is formed by an **encoder** $z = E_\phi(x)$ parameterized by ϕ and a **decoder** $x' = D_\theta(z)$ parameterized by θ. The objectives L_{AE} attempts to minimize the Mean-Square Error (MSE) between the inputs x and the reconstructed ones x', which can be formulated as,

$$L_{AE} = \|x - x'\|_2^2. \tag{3.2.1}$$

Among AE models, Variational AE (VAE) [95] and Denoising AE (DAE) [176] are two broadly used in image/video-based SSL tasks.

- **VAE** [95] assumes data are generated from unobserved latent variables $\mathbf{z}$. Given the inputs x, VAE refers to a recognition model $q_\phi(\mathbf{z}|\mathbf{x})$ as the probabilistic **encoder** and refers to $p_\theta(\mathbf{x}|\mathbf{z})$ as the **decoder**. With the spirit of variational inference, VAE aims to maximize the evidence lower bound (ELBO) on the log-likelihood of inputs during training.

  $$L(\phi, \theta; \mathbf{x}^i) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^i)||p_\theta(\mathbf{z})) + E_{q_\phi(z|\mathbf{x}^i)}[\log p_\theta(\mathbf{x}^i|\mathbf{z})], \tag{3.2.2}$$

  where $\mathbf{x}^i$ refers to the i-th data point. Note that the prior $p_\theta(\mathbf{z})$ and the approximate posterior $p_\theta(\mathbf{z}|\mathbf{x})$ are assumed to follow Gaussian distributions in [95]. As can be seen, the objective encourages the posterior to approximate the prior (first term), and meanwhile, encourages the latent variables to reconstruct the original inputs (second term).

  In terms of the applications of VAE in the image/video-based SSL, S3VAE [260] proposes a sequential VAE to learn disentangled representations of sequential data in a self-supervised manner. Swap-VAE [120] employs a regularized VAE to reconstruct the inputs of neural activity with the learned "content" and "style" representations.

- **DAE** improves the robustness of VAE by corrupting the input data with noise. In the NLP domain, Masked language model (MIM), which learns discriminative representations by predicting the input tokens that are masked out before feeding into the network, has achieved a significant success [40, 146, 147].

  Recently, MIM has been introduced to image/video-based SSL by encouraging the network to learn from corrupted (masked) images. Starting from the attempts with CNN [177, 139], MIM with Transformers enlightens the recent SSL works. For example, SimMIM [206] and MAE [79] explore the potential of pixel-wise reconstruction. For both methods, a large ratio of random masking, *e.g.,* above 60%, is applied on the input images. The lightweight decoders are forced to recover the masked counterparts from latent representations. The differences between SimMIM and MAE are 1) the inputs of the encoder. SimMIM takes all image patches, both masked and unmasked, as inputs. MAE takes the unmasked ones ONLY, and 2) the design of the decoder. Compared to MAE, which uses 8 transformer blocks, SimMIM employs a linear layer as the decoder. Instead of manually pre-defining a mask ratio, ADIOS [159] adopts masks generated by an adversarial masking model. Apart from the pixel-wise reconstruction, the token-wise recovery is another mainstream [8, 44]. Both BEiT [8] and PeCo [44] use VQ-VAE [172] as visual tokenizers. However, besides the latent embedding based token classification task conducted by BEiT, PeCo also enforces the perceptual similarity between reconstructed tokens and the original ones.

Despite the significant achievements of generative SSL, the inconsistency between the objectives of pretext tasks and downstream tasks hinders its performance on some CV tasks, such as image classification, image segmentation, *etc.* Therefore, high-level semantic objectives for pretext tasks are still needed.

Contrastive SSL. Contrastive learning (CL) [73] aims to train an encoder for a *dictionary look-up* task. Specifically, given a set of keys in the dictionary $\{k_0, k_1, k_2, ...\}$ and an encoder query q, assume there is a single key k_+ that matches q, the contrastive loss, *i.e.,* InfoNCE, is formulated as,

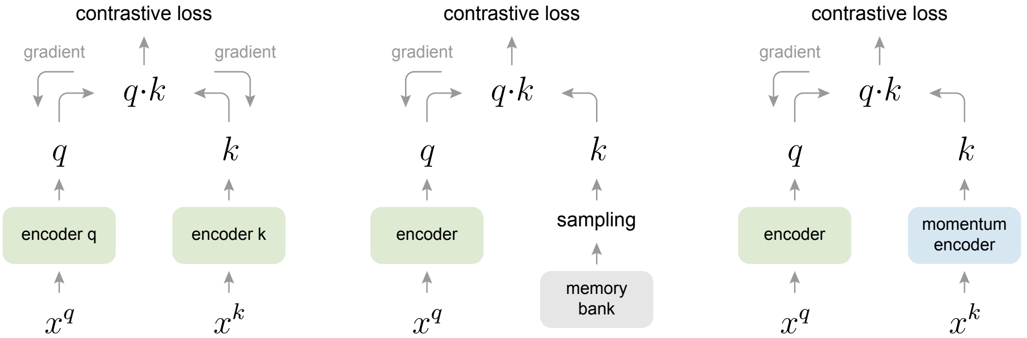
$$L_{InfoNCE} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}, \tag{3.2.3}$$

where τ is a temperature hyper-parameter. K refers to the number of negative samples. InfoNCE encourages the query to be close to its positive samples while distancing itself from its negative ones. Existing contrastive SSL can be roughly categorized as **instance-level** ones and **clustering-level** ones according to the composition of contrast targets.

- **Instance-level CL** [202, 25, 67, 78, 134] studies the relationships between different samples. Generally, the positive/negative samples can be collected from 1) the current mini-batch, 2) the memory bank, and 3) the momentum encoder. The pipelines are illustrated in Fig. 3.3.

  For a given query, early works [25, 67] generally regard the augmented counterpart as the positive sample, while regarding the rest of the samples within the current mini-batch as negative ones. As can be seen from Fig. 3.3(a), the query encoder q and the key encoder k are trained for query feature extraction and key feature extraction, respectively, with the objective Eqn. 3.2.3. However, such a negative collecting scheme requires a large batch size to provide sufficient information, whose performances are highly dependent on the GPU memory [89].

  To get rid of the reliance on the large batch size, the memory bank (see Fig. 3.3(b)), serving as a large look-up dictionary, is proposed to store feature representations of past samples [202, 134]. The keys in the memory bank



(a) Mini-batch (b) Memory bank (c) Momentum enc.

Figure 3.3: Comparisons of typical instance-level CL pipelines. Taken from MoCo [78]. (a) Mini-batch based scheme. (b) Memory bank based scheme. (c) Momentum encoder based scheme.

are updated by the features of corresponding samples from prior epochs in an exponential moving average manner. By replacing the key encoder with a non-backpropagation module, *i.e.,* memory bank, more sufficient negative samples can be collected without increasing the batch size. However, maintaining a large memory bank is computationally expensive and the update of keys is inconsistent as only a limited number of samples are sampled at each iteration [78].

Instead of tracking every sample, MoCo [78] replaces the memory bank with a momentum encoder, which updates the parameters of the encoder in a momentum manner, as can be seen in Fig. 3.3(c). In other words, it updates the dictionary in a queueing manner. Specifically, for each iteration, the current mini-batch is enqueued while the oldest one is dequeued. Formally, the parameters of the momentum encoder are updated as,

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q, \tag{3.2.4}$$

where $m \in [0, 1)$ is the momentum coefficient. The introduction of a momentum encoder eases the requirement of training an additional key encoder as well as reduces the storage cost of the memory bank.

- **Clustering-level CL** [17, 18] extends the contrast target from a single pair of samples to a group of samples with similar attributes, *i.e.,* cluster (prototype). The comparison between pipelines of instance-level CL and clustering-level CL is depicted in Fig. 3.4. DeepCluster [17] is proposed to iteratively use the cluster assignments as pseudo labels to optimize the feature encoder. Recently, SwAV [18] proposes an online learning scheme where image features are projected to a set of trainable vectors, *i.e.,* prototypes. Unlike previous methods regarding prototypes as contrast targets, SwAV encourages the mapping consistency between different views of the given image.

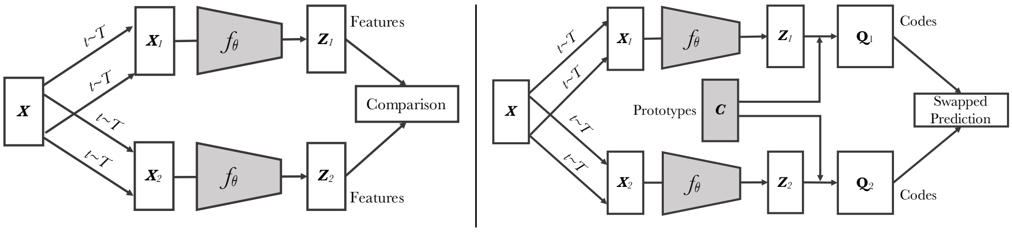


Figure 3.4: Instance-level CL (left) vs. Clustering-level CL (right). Taken from SwAV [18].

Unsupervised person ReID, as a popular branch of unsupervised (self-supervised) representation learning, follows the above CL scheme. However, despite the success of instance-level CL methods in downstream tasks with fine-tuning, they cannot be directly adopted to the unsupervised person ReID task [60]. Specifically, the instance-level CL considers each instance as a class. Such assumption unfits the objective of person ReID, where a query image can match multiple images in the gallery. Therefore, current clustering-based unsupervised person ReID methods generally adopt the scheme of clustering-level CL methods, which alternate between the cluster assignment (pseudo label generation) and feature learning during training.

3.2.4 Pseudo label noise reduction

Recently, how to handle the problem of noisy pseudo labels in clustering-based methods has become a research hotspot. Specifically, SpCL [60] employs a self-paced learning scheme to gradually obtain more reliable clusters for the learning target refinement. MMT [48] and MEBNet [225] adopt the mutual learning strategy, where the predictions of auxiliary teacher models are utilized to refine the pseudo labels. CAP [184] conducts intra-camera and inter-camera contrastive learning based on camera-aware proxies. ICE [20] alleviates the label noise by enhancing the consistency between augmented and original instances. RLCC [233] refines noisy pseudo labels with clustering consensus, which encourages the consistency between cluster results of two consecutive iterations. PPLR [32] employs the complementary relationship between reliable human global and part features for the pseudo label refinement. ISE [235] generates boundary samples from actual samples and their neighbouring clusters to handle the noisy clusters.

Unlike these works which involve the designed learning strategy or additional information, a simple but effective way is proposed to reduce the pseudo label noise. Specifically, samples are encouraged to be close to the high-quality cluster centroids. Additionally, neighboring relationships are used to control to what degree the given sample should be pushed towards the learning targets.

3.3 Problem statement

Let $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}$ denote an unlabeled training dataset, where x_i is the i-th image and $N_\mathcal{D}$ is the number of images. The USL ReID task aims to train a backbone feature extractor E_θ in an unsupervised manner, where ReID features $\mathcal{F} = \{f_i\}_{n=1}^{N_\mathcal{D}}$ are extracted. During the inference, ReID features are used for identity retrieval. Following the training scheme of clustering-based USL methods [60, 37, 184], the training process alternates between two stages:

Stage I: Clustering. At the beginning of each epoch, all training samples are clustered with the DBSCAN [50] clustering algorithm. Cluster IDs are employed as pseudo labels $y_i \in \{1, ..., C\}$ for the network optimization, *i.e.,* samples of the same cluster are regarded as the same person. C is the number of clusters (identity). Note that, outliers, *i.e.,* un-clustered samples, are not included during the training.

Stage II: Network training. The network is optimized in a supervised manner with the pseudo labels obtained in Stage I. Specifically, following the PK-sampling strategy, K identities and P samples per identity are randomly sampled to form a mini-batch. At the same time, a cluster-based memory bank $\mathcal{M} = \{m_i\}_{i=1}^{C}$ is employed to store the cluster centroid as [37]. The memory bank is firstly initialized by averaging features of each cluster as follows,

$$m_i = \frac{1}{|\mathcal{H}_c|} \sum_{f_i \in \mathcal{H}_c} f_i, \tag{3.3.1}$$

where f represents all instances of the cluster $\mathcal{H}_c$, and $|\mathcal{H}_c|$ represents the cluster size, *i.e.,* the number of instances within the cluster. As the training goes on, the memory bank is updated in a momentum manner as follows,

$$m_i \leftarrow \mu \cdot m_i + (1 - \mu) \cdot f, \tag{3.3.2}$$

where μ is the updating factor. f denotes the instance belonging to the i-th cluster in the current mini-batch. The instance can be either the hardest one [37], *i.e.,* one has the largest distance to its cluster centroid, or all samples in the mini-batch [235]. Here this work takes the latter which has been proven to have better performances [235].

The optimization is supervised by the cluster-wise contrastive loss, *i.e.,* ClusterNCE [37], which is formulated as follows,

$$\mathcal{L}_q = -\log \frac{\exp(\mathrm{sim}(f \cdot m_+)/\tau)}{\sum_{j=1}^{C} \exp(\mathrm{sim}(f \cdot m_j)/\tau)}, \tag{3.3.3}$$

where m_+ is the centroid of the cluster that f belongs to, and m_j represents the j-th centroid of the memory bank, and $sim(u \cdot v)$ represents the cosine similarity between vector u and vector v, *i.e.,* $sim(u \cdot v) = u^{\mathrm{T}} v / \, ||u|| \, ||v||$.

Our framework is illustrated in Fig. 3.5, which is based on a state-of-the-art baseline, *i.e.,* Cluster-Contrast [37]. the proposed method differs from previous works mainly in two aspects: 1) cluster centroids. Instead of the all-sample based ones, confidence-guided centroids (CGC) are used to provide reliable cluster-wise prototypes for the feature learning (Section 3.5), and 2) pseudo labels. Apart from the assigned centroid, the proposed confidence-guided pseudo labels (CGL) encourages instances to approach other centroids where their identity information

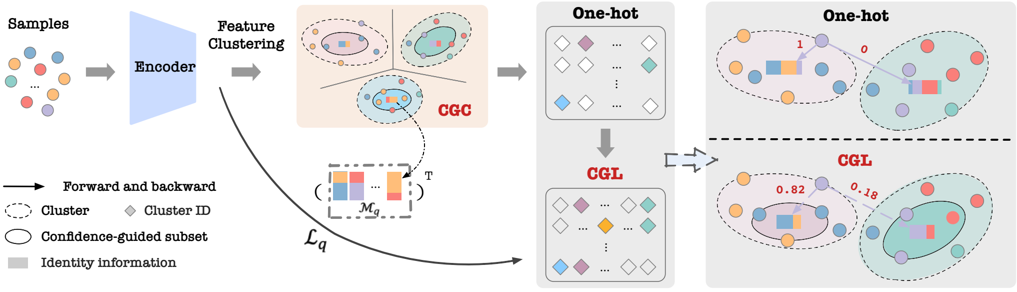


Figure 3.5: Framework of the proposed method. At the beginning of each epoch, training samples are clustered by DBSCAN [50]. Based on the original clustering results, a confidence-guided subset is selected to build the confidence-guided centroids (CGC). During optimization, samples are encouraged to approach not only the assigned centroid but others where their identity information is potentially embedded via the proposed confidence-guided pseudo labels (CGL).

is potentially embedded (Section 3.6).

3.4 Silhouette score

Generally, the quality of clusters is mainly affected by two factors: tightness and separation. The former describes how compact each cluster is, while the latter represents how separate different clusters are. In other words, a good clustering should have smaller intra-class distances (tightness), while having larger inter-class distances (separation). In order to measure the quality of clustering, a metric, *i.e.,* silhouette score [152], is proposed. Formally, for the i-th data point being clustered to cluster C_I, its distance to other data points within the cluster can be calculated as,

$$a_i = \frac{1}{|C_I|} \sum_{j \in C_I} d(i,j), \tag{3.4.1}$$

where $|C_I|$ refers to the number of samples within the cluster C_I and $d(i,j)$ measures the distance between i-th and j-th data points. Basically, a represents the average intra-class distance. Additionally, the inter-class distances are represented by the distance between the i-th data point and all samples belonging to its closest neighboring cluster, which can be denoted as,

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j). \tag{3.4.2}$$

Combining the above metrics, the silhouette score $S = \{s_i\}_{i=1}^{N_{\mathcal{D}}}$ of the i-th data point can be formulated as,

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}. \tag{3.4.3}$$

Note that, since the intra-class distance for clusters with a single data point (outliers) is 0, the silhouette score of outliers is 1. According to Eqn. (3.4.3), the silhouette score ranges from $[-1, 1]$. A higher silhouette score means a smaller intra-class distance as well as a large inter-class distance, *i.e.,* a better clustering result [152].

3.5 Confidence-guided centroids

Based on the observation that images with lower silhouette scores (confidence) are generally containing high uncertainty regarding person identity, previous all-sample based cluster centroids are undoubtedly unwise. To remedy the problem, we build confidence-guided centroids (CGC) with high-confidence images only.

Specifically, the confidence-guided centroid of i-th cluster m_i can be formulated as,

$$m_i = \frac{1}{|\mathcal{C}_q|} \sum_{f_i \in \mathcal{C}_q} f_i, \quad \mathcal{C}_q = \{f_i \in \mathcal{C} | s_i > \delta\}, \tag{3.5.1}$$

where a confidence-guided subset $\mathcal{C}_q$ is selected from the original cluster $\mathcal{C}$ by a silhouette score threshold δ. All confidence-guided centroids are then stored in a confidence-guided memory bank $\mathcal{M}_q = \{m_i\}_{i=1}^{C}$ for network optimization.

According to Fig. 3.1, the proposed confidence-guided centroids can filter out images that are poor in quality or with cluttered backgrounds at early stages. While at later stages, such centroids effectively exclude some low-confidence samples that possibly belong to other identities. In summary, the proposed confidence-guided centroids can provide more reliable cluster-wise prototypes for feature learning.

3.6 Confidence-guided pseudo labels

Another problem of the clustering-based USL methods is that samples, especially low-confidence ones, very likely carry different identity information with their assigned centroids. The proposed confidence-guided centroids also confronts with the problem since only high-confidence samples are included in the formation of centroids, as illustrated in Fig. 3.5. Given the situation, the previous learning scheme, which enforces samples to approach their assigned centroids solely regardless of the identity consistency in-between, is unwise. To alleviate the problem, the author proposes to use confidence-guided pseudo labels (CGL). Such labeling encourages samples to approach not only the assigned centroid but other centroids where their identity information is potentially embedded.

Specifically, a distance matrix $\mathcal{D} \in \mathbf{R}^{N \times C}$ is built, where N and C denote the number of samples and clusters at the current epoch, respectively. In the paper, clusters consisting of one sample are ignored [37]. As normalized identity features and centroids are adopted, $\mathcal{D}(i, j)$ represents the cosine distance between i-th sample and j-th confidence-guided centroid. Since similar samples are more likely to be scattered in neighboring clusters [235], the identity information of boundary samples is probably embedded in neighboring centroids. Therefore, when setting the learning target for samples, neighboring centroids should be assigned with higher confidence while distanced ones should be given lower confidence. To this end, a confidence matrix $\mathcal{P} \in \mathbf{R}^{N\times C}$ is obtained by,

$$\mathcal{P}(i,j) = \frac{p_{i,j}}{\sum_{j=1}^{C} p_{i,j}}, \quad p_{i,j} = \sigma(-\mathcal{D}(i,j)), \tag{3.6.1}$$

where $\mathcal{P}(i,j)$ represents the confidence of j-th centroid given by i-th sample. Note that $\sum_{j=1}^{C} \mathcal{P}(i,j) = 1$. $\sigma(\cdot)$ is the Sigmoid function. By integrating the confidence matrix with the originally assigned one-hot pseudo label y_i, the confidence-guided pseudo label of i-th sample $\tilde{y}_i$ can be formulated as,

$$\tilde{y}_i = \beta \cdot y_i + (1-\beta)\cdot \mathcal{P}(i), \tag{3.6.2}$$

where $\beta \in [0,1]$ is the coefficient for the pseudo label refinement.

Algorithm 1: The pipeline of the proposed method.

```
Input: Unlabeled data with pseudo labels $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$, where
       $y_i \in \{1, \dots, C\}$; Threshold $\delta$ for Eqn. (3.5.1); Coefficient $\beta$ for
       Eqn. (3.6.2).
Output: Backbone encoder $E_\theta$
Initializing the backbone encoder $E_\theta$
for $n$ in $[1, epoch\_num]$ do
    Extracting features $\mathcal{F}$ by $E_\theta$
    Clustering $\mathcal{F}$ into $C$ clusters with DBSCAN
    Building CGC dictionary $\mathcal{M}_q$ by Eqn. (3.5.1)
    for $m$ in $[1, iteration\_num]$ do
        Sampling a mini-batch from $\mathcal{T}$
        Computing CGL with Eqn. (3.6.2)
        Computing loss with Eqn. (3.6.3)
        Updating encoder $E_\theta$
        Updating centroids with Eqn. (3.3.2)
```

Training objective According to a previous work [202], the training objective, *i.e.,* ClusterNCE, can be considered as a non-parametric classifier, where centroids stored in the memory bank serve as the weight matrix of the classification layer. Therefore, the training objective of the proposed method can be rewritten as,

$$\mathcal{L}_q = \frac{1}{N}\sum_{i=1}^{N}\Big[\ell_{ce}\big(\mathcal{M}_q^T f_i, \tilde{y}_i\big)\Big], \tag{3.6.3}$$

where ℓ_{ce} refers to the cross-entropy loss. Compared to Eqn. (3.3.3), the training objective of the proposed method can be obtained by simply applying two modifications: 1) replacing the original $\mathcal{M}$ with the proposed confidence-guided memory

Dataset	Images	Cameras	Persons
Market-1501 [246]	32,668	6	1,501
MSMT17 [190]	126,441	15	4,101

Table 3.1: Details of person ReID benchmark datasets.

bank $\mathcal{M}_q$, and 2) replacing the one-hot pseudo label y_i with the proposed confidence-guided one $\tilde{y}_i$. The training details are presented in Algorithm 1.

3.7 Experiments

3.7.1 Datasets

There are several widely-used benchmark datasets for single-modality person ReID, such as CUHK [103], DukeMTMC [250], Market-1501 [246],MSMT17 [190]. However, CUHK is outdated due to the limited number of identities and images and DukeMTMC has been forbidden due to ethical concerns. Therefore, in this work, the evaluation is conducted on two benchmark person ReID datasets: Market-1501 and MSMT17. Details of two datasets are reported in Table 3.1.

Market-1501 overall includes 32,668 images of 1,501 identities with 6 camera views. Cameras include 5 high-resolution ones and 1 low-resolution one. The dataset randomly divides identities into two parts: training and testing. The training set has 12,936 images of 751 identities. The remaining 19,732 images of 750 identities are used for gallery during testing. 3368 images are randomly selected to build the query set. The dataset is trending for its diversity in scenarios, illumination, and viewpoints and has become a popular benchmark dataset for multiple-person ReID tasks.

MSMT17 contains 126,441 images from 4,101 identities captured by 15 cameras. The training set has 32,621 images of 1,041 identities and the testing set has 93,820 images of 3,060 identities. Among them, 11659 images are used as query and 82161 images are used to build the gallery. MSMT17 is more challenging due to its complex scenarios as well as various distractors such as a longer time span, varying pose variations, and severe body occlusions.

3.7.2 Evaluation metrics and implementation details

Evaluation metrics. The metrics used for evaluation are mean average precision (mAP) [6] and cumulative matching characteristic (CMC) [65] top-1, top-5, top-10 accuracies to evaluate performances. Note that, there are no post-processing operations in testing, *e.g.,* reranking [253].

Implementation details. Following previous works [60, 37, 235], ResNet-50 [77] pre-trained on ImageNet [38] serves as the backbone feature encoder [37]. All layers after layer-4 are replaced by a generalized mean pooling (GeM) [145] layer followed by the batch normalization layer [87]. GeM pools feature maps into vectors via a learnable parameter p. Max pooling and average pooling are special cases of GeM when $p \to +\infty$ and $p = 1$, respectively. During inference, the output 2048-dimensional ReID features are normalized for identity retrieval. The proposed framework is built upon a state-of-the-art USL method [37]. For a fair comparison, the authors follows all experimental settings except for the formation of cluster centroids and the training objectives, as described in Section 3.5 and Section 3.6. The coefficient β in Eqn. (3.6.2) is empirically set as 0.8 to achieve optimal performances.

During training, input images are resized to 256×128 pixel. Random flipping, cropping, and erasing [255] are adopted as data augmentation. Each mini-batch is formed by 16 identities, each with 16 images. Both identity and images are randomly selected from the training set. For the optimization, Adam [94] optimizer with a weight decay of 0.0005 is adopted during training. The learning rate is set to 3.5×10^{-4} initially and is divided by 10 every 30 epochs. The training process takes 70 epochs on Market-1501 [246], and 50 on MSMT17 [190].

3.7.3 Comparison with the state-of-the-art

The performance comparison between the proposed method and state-of-the-art (SOTA) unsupervised person Re-ID methods are reported in Table 3.2 and Table 3.3. Compared with SOTA USL methods, the proposed method outperforms previous ones, except ISE [235], on both benchmarks. Specifically, the proposed method achieves 85.3% mAP and 94.2% top-1 accuracy on Market-1501 and 34.6% mAP and 63.4% top-1 accuracy on MSMT17. As stated in Section 3.2, existing SOTA methods generally leverage auxiliary information to refine pseudo labels. For example, CAP [184] and ICE [20] leverage the camera information, PPLR [32] employs body part predictions, and ISE [235] generates extra support samples in the latent space. As a departure from the above methods, this work does NOT exploit any auxiliary information. The proposed method refines pseudo labels with internal characteristics, *i.e.,* the sample-wise clustering confidence.

Additionally, the performance of some well-known supervised person ReID methods [166, 251] and unsupervised one [20] under the supervised setting are also illustrated in Table 3.2 and Table 3.3. Despite the absence of identity labels, the proposed method even outperforms some supervised person ReID methods. Additionally, by replacing the pseudo labels with the ground-truth identity labels provided by datasets, the proposed method outperforms a USL method (ICE [20]), which proves the potential of the proposed framework.

3.7.4 Ablation study

In the section, we thoroughly analyze the effectiveness of the proposed strategies, *i.e.,* confidence-guided centroids (CGC) and confidence-guided pseudo labels (CGL).

Method	Reference	Market-1501			
		mAP	top-1	top-5	top-10
Purely Unsupervised					
SSL [114]	CVPR'20	37.8	71.7	83.8	87.4
MMCL [178]	CVPR'20	45.5	80.3	89.4	92.3
HCT [223]	CVPR'20	56.4	80.0	91.6	95.2
SpCL [60]	NeurIPS'20	73.1	88.1	95.1	97.0
JNTL-MCSA [208]	CVPR'21	61.7	83.9	92.3	-
GCL [21]	CVPR'21	66.8	87.3	93.5	95.5
IICS [207]	CVPR'21	72.9	89.5	95.2	97.0
JVTC+* [21]	CVPR'21	75.4	90.5	96.2	97.1
OPLG-HCD [249]	ICCV'21	78.1	91.1	96.4	97.7
CAP† [184]	AAAI'21	79.2	91.4	96.3	97.7
ICE[20]	ICCV'21	79.5	92.0	97.0	98.1
ICE† [20]	ICCV'21	82.3	93.8	97.6	98.4
Cluster-Contrast [37]	Arxiv'21	82.6	93.0	97.0	98.1
PPLR [32]	CVPR'22	81.5	92.8	97.1	98.1
PPLR† [32]	CVPR'22	84.4	94.3	97.8	98.6
ISE [235]	CVPR'22	84.7	94.0	97.8	98.8
Cluster-Contrast (Our Baseline)	-	82.4	92.5	96.9	98.0
Baseline+CGC	-	84.1	93.1	97.2	98.2
Baseline+CGL	-	83.4	93.2	97.1	98.2
Ours	-	**85.3**	**94.2**	**97.6**	**98.5**
Fully Supervised					
PCB [166]	ECCV'18	81.6	93.8	97.5	98.5
DG-Net [251]	CVPR'19	86.0	94.8	-	-
ICE (w/ ground-truth) [20]	ICCV'21	86.6	95.1	98.3	98.9
Our (w/ ground-truth)	-	87.4	95.3	98.5	99.0

Table 3.2: Comparison with the state-of-the-art ReID methods on **Market-1501** [246]. The best USL results without camera information are marked with **bold**. † indicates using the additional camera knowledge.

Method	Reference	MSMT17			
		mAP	top-1	top-5	top-10
Purely Unsupervised					
MMCL [178]	CVPR'20	11.2	35.4	44.8	49.8
SpCL [60]	NeurIPS'20	19.1	42.3	55.6	61.2
JNTL-MCSA [208]	CVPR'21	15.5	35.2	48.3	-
GCL [21]	CVPR'21	21.3	45.7	58.6	64.5
IICS [207]	CVPR'21	26.9	56.4	68.8	73.4
JVTC+* [21]	CVPR'21	29.7	54.4	68.2	74.2
OPLG-HCD [249]	ICCV'21	26.9	53.7	65.3	70.2
CAP† [184]	AAAI'21	36.9	67.4	78.0	81.4
ICE[20]	ICCV'21	29.8	59.0	71.7	77.0
ICE† [20]	ICCV'21	38.9	70.2	80.5	84.4
Cluster-Contrast [37]	Arxiv'21	33.3	63.3	73.7	77.8
PPLR [32]	CVPR'22	31.4	61.1	73.4	77.8
PPLR† [32]	CVPR'22	42.2	73.3	83.5	86.5
ISE [235]	CVPR'22	35.0	64.7	75.5	79.4
Cluster-Contrast (Our Baseline)	-	30.1	58.6	69.6	74.4
Baseline+CGC	-	34.1	63.1	75.0	79.0
Baseline+CGL	-	33.7	62.5	73.9	78.4
Ours	-	**34.6**	**63.4**	**74.6**	**79.3**
Fully Supervised					
PCB [166]	ECCV'18	40.4	68.2	-	-
DG-Net [251]	CVPR'19	52.3	77.2	-	-
ICE (w/ ground-truth) [20]	ICCV'21	50.4	76.4	86.6	90.0
Our (w/ ground-truth)	-	51.0	76.6	87.1	90.1

Table 3.3: Comparison with the state-of-the-art ReID methods on **MSMT17** [190]. The best USL results without camera information are marked with **bold**. † indicates using the additional camera knowledge.

Effectiveness of CGC. The performances of models trained with the plain all-sample based cluster centroids ("Baseline"), and with the proposed confidence-guided ones ("Baseline + CGC") are reported in Table 3.2 and Table 3.3, respectively. As can be seen, confidence-guided centroids boost the ReID performance by +1.7% / +0.6% on mAP / top-1 accuracy on Market-1501, and +2.7% / +1.9% on MSMT17. Such improvements reveal the potential of the clustering confidence in the pseudo label refinement.

To better understand how the proposed confidence-guided centroids benefit feature learning, the author analyzes how the sample-wise confidence varies throughout the training process on MSMT17. Specifically, the distribution of silhouette scores at different epochs are visualized in Fig. 3.6. Note that scores of outliers are excluded. Several conclusions can be drawn from the comparison between Fig. 3.6(a) and Fig. 3.6(b).

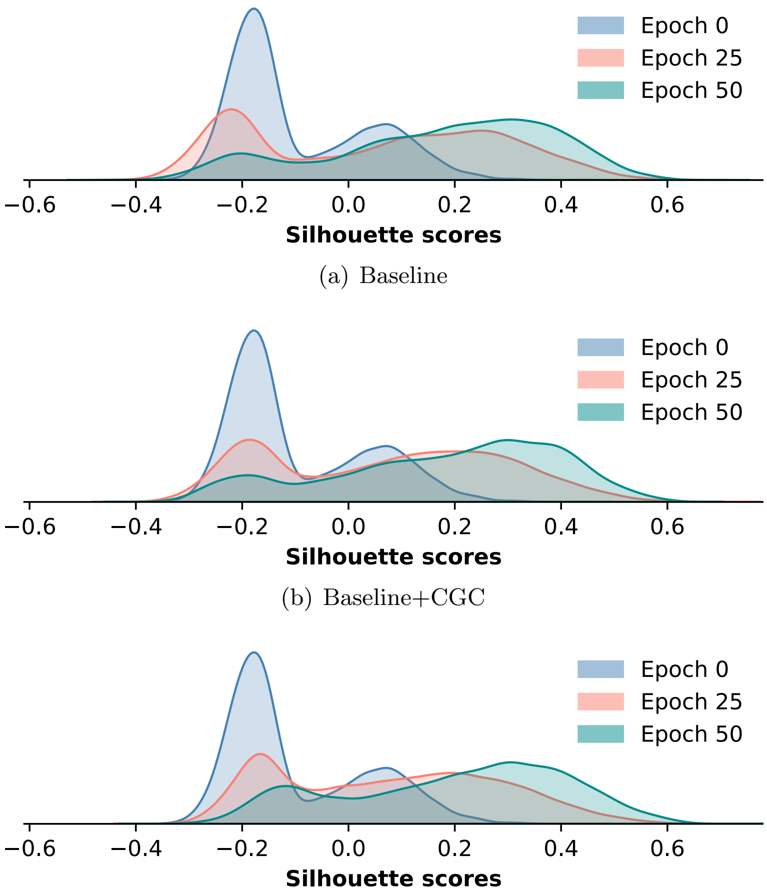
- As training goes on, the number of valid samples gradually increases, representing as larger areas under the curve.
- Starting from the same point (epoch 0), with the proposed confidence-guided centroids, a noticeable shift towards higher scores can be found at epoch 25. The shift implies CGC can effectively reduce the overall number of low-confidence samples while enhancing high-confidence ones.
- The advantage remains until the end of training. At epoch 50, the number of high-confidence samples increases, representing by a higher peak closer to 0.4.

Effectiveness of CGL. The comparison is conducted between the baseline model ("Baseline") and the model trained with confidence-guided pseudo labels ("Baseline + CGL"). The performances are shown in Table 3.2 and Table 3.3, respectively. As can be seen, CGL improves mAP and top-1 accuracy by 1.0% and 0.7% on Market-1501, by 2.3% and 1.3% on MSMT17. When both CGC and CGL are employed during training, improvements are +2.9% and +1.7% on Market-1501, and +3.2% and +2.2% on MSMT17.

In terms of the sample-wise clustering confidence, we visualize the distribution of silhouette scores in Fig. 3.6(c), when CGL is applied during training. Compared to the model trained without CGL (Fig. 3.6(b)), CGL further pushes the score towards a higher value at both epoch 25 and epoch 50. Less low-confidence samples during training imply the proposed CGL contributes to better clustering. In summary, the above qualitative and quantitative results prove the proposed scheme can boost performance by enhancing the sample-wise clustering confidence.

3.7.5 Parameter analysis

Threshold δ in CGC. To obtain the optimal threshold δ in Eqn. (3.5.1) for the proposed confidence-guided centroids (CGC), three types of threshold selection strategies are explored, *i.e.,* linear, dynamic and constant, respectively. For the



(a) Baseline

(b) Baseline+CGC

(c) Baseline+CGC+CGL

Figure 3.6: Silhouette scores of valid samples (MSMT17 [190]) at different epochs. Comparisons are conducted between (a) baseline model, (b) baseline model with confidence-guided centroids (CGC), and (c) baseline model with CGC and confidence-guided pseudo labels (CGL). **Best viewed in color**.

Method	Strategy	δ	Market-1501		MSMT17	
			mAP	top-1	mAP	top-1
Baseline	-	-	82.4	92.5	31.4	61.2
Ours	Linear	-	**85.3**	**94.2**	33.6	63.0
	Dynamic	-	84.9	93.9	33.0	62.8
	Constant	-0.1	83.5	93.4	32.7	62.8
		0	84.9	94.0	**34.6**	**63.4**
		0.1	84.0	93.3	34.0	63.2

Table 3.4: Comparison of threshold selection strategies of confidence-guided centroids (CGC) on benchmark datasets.

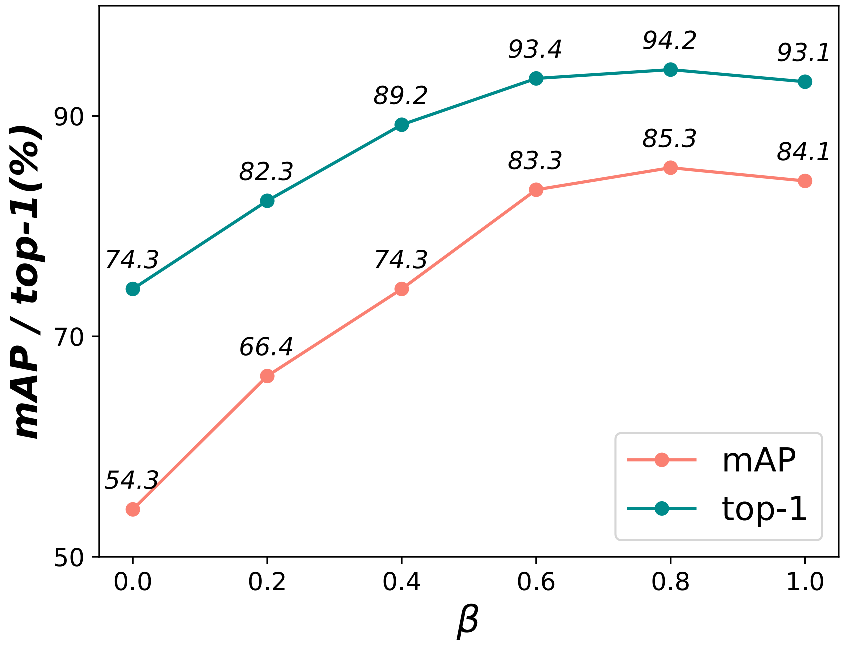
former two strategies, the threshold gradually increases as training goes on. The constant strategy employs a fixed threshold throughout the training process.

Specifically, the linear strategy updates the threshold by $\delta_t = \delta_0 * t/T + \epsilon$, where δ_0 limits the range of threshold and ϵ is the offset. In the work, $\delta_0 = 0.2$ and $\epsilon = -0.1$ are used. t and T denote the current epoch and the overall number of epochs, respectively. In terms of the dynamic strategy, the threshold is updated by $\delta = \delta_0 * tanh(0.1 * (t - T/2))$. δ_0 is set as 0.1 to achieve $\delta \in [-0.1, 0.1]$, which is the same as the linear strategy. The range is set empirically in the consideration of the image quality and the distribution of silhouette scores (see Fig. 3.6). Apart from the varying threshold, the constant strategy is also conducted by fixing the threshold as $\{-0.1, 0, 0.1\}$ respectively. The comparisons between model performances with different strategies are reported in Table 3.4. The best performance is achieved when adopting the linear strategy for Market-1501 and applying a fixed threshold $\delta = 0$ on MSMT17. The optimal settings are employed in all experiments.

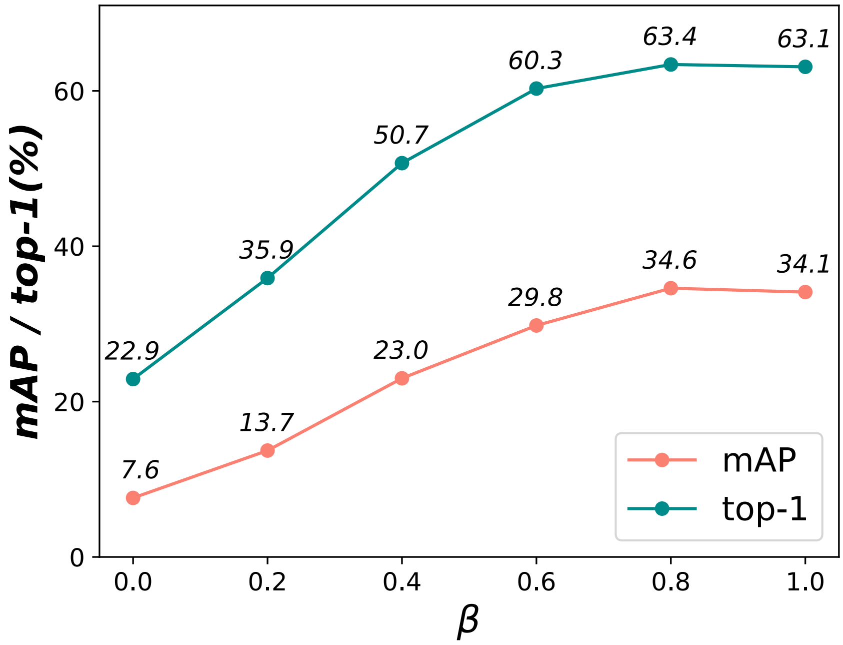
Coefficient β in CGL. To analyze the impact of the coefficient β in the proposed confidence-guided pseudo labels (CGL), the value of parameter β is tuned from 0 to 1 while keeping others fixed. According to Eqn. (3.6.2), when β is set to 0 or 1, the proposed method decomposes down to using the confidence matrix or the one-hot pseudo label exclusively during training. The results on two benchmarks are illustrated in Fig. 3.7. As shown, as β increases from 0 to 0.8, both mAP and top-1 accuracy increase. A slight performance drop can be found when increasing β from 0.8 to 1. To achieve the best performance, $\beta = 0.8$ is used for all experiments.

3.7.6 More discussion

Identity feature distribution. To better understand the advantages of the proposed strategies, the distribution of identity features is visualized via t-SNE [173]. Specifically, 20 identities are randomly selected from Market-1501 and MSMT17, respectively. Features of selected identities are extracted by the baseline model and the proposed model is trained with confidence-guided centroids (CGC) and confidence-guided pseudo labels (CGL). The distribution of identity features is illustrated in Fig. 3.8. As can be seen, due to the vast variety in camera views,



(a) Market-1501



(b) MSMT17

Figure 3.7: Comparison of coefficient β in confidence-guided pseudo labels (CGL) on (a) Market-1501 and (b) MSMT17.

backgrounds, and poses, the feature distribution of MSMT17 is more chaotic than that of Market-1501. Despite such challenges, with the aid of the proposed strategies, features of the same identity are distributed more compactly while features of different identities are further separated.

Identity consistency score. The current learning scheme enforces samples to approach their assigned cluster centroids, where their identity information is embedded. However, the existence of noisy labels will lead samples to "wrong" centroids. It is especially problematic for low-confidence samples, *i.e.,* boundary samples because they can be closer to other centroids than the assigned ones.

To investigate the problem, the author conducts an experiment on MSMT17 to analyze how much the identity information of boundary samples can be presented in the assigned centroids, *i.e.,* the identity consistency in-between. Specifically, clusters whose size is over 100 at each epoch are selected. For each cluster, samples whose silhouette scores rank at the bottom 5% are empirically marked as boundary samples. Formally, let $\mathcal{C} = \{(x_i, g_i)\}_{i=1}^{N_c}$ denote a cluster with N_c samples, where g_i refers to the ground-truth identity label provided by the dataset. An identity set $\mathcal{G} = \{g_k\}_{k=1}^{M}$ is then constructed by overall M identities occurring in the cluster. Following the formation of plain all-sample based cluster centroids (Eqn. (3.3.1)), the identity information embedded in the centroid can be obtained by linearly integrating all identities within the cluster via weights $\mathcal{Q}_c = \{q_k\}_{k=1}^{M}$, where q_k is obtained as,

$$q_k = \frac{1}{|\mathcal{C}|} \sum_{g_i \in \mathcal{C}} \mathbb{1}\{g_i = g_k\}, \tag{3.7.1}$$

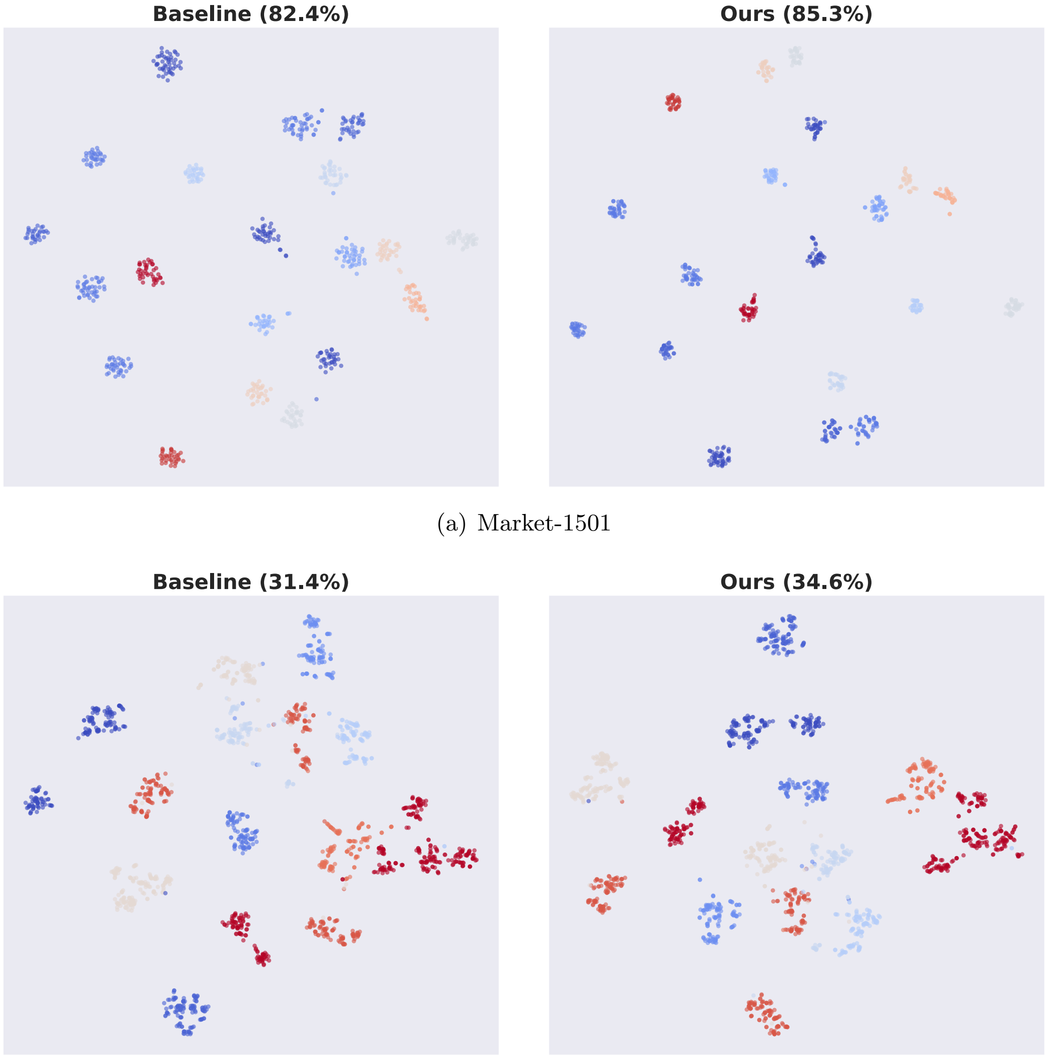
where $|\mathcal{C}|$ denotes the cluster size. $\mathbb{1}\{g_i = g_k\}$ equals to 1 when $g_i = g_k$, otherwise 0. Then, the identity consistency score (ICS) between boundary samples and the corresponding centroid of can be calculated as,

$$ICS = \frac{1}{N_b} \sum_{g_i \in \mathcal{C}} q_k \cdot \mathbb{1}\{g_i = g_k\}, \tag{3.7.2}$$

where N_b denotes the number of boundary samples.

Similar to the plain scheme, ICS of the proposed confidence-guided centroids (CGC) scheme can be computed by simply replacing $\mathcal{C}$ with the confidence-guided subset $\mathcal{C}_q$ during the computation of the weight q_k. Since low-confidence samples are filtered out in the formation of confidence-guided centroids, the identity set $\mathcal{G}$ only includes identities of samples with high confidence scores. The comparison on average ICS throughout the training is conducted between the plain all-sample based cluster centroids and the proposed confidence-guided ones. The ICS at different epochs are shown in Fig. 3.9.

For the plain scheme, only 5.83% boundary samples carry the same identity information with their assigned centroid at the beginning. Although the ratio gradually climbs to 17.19%, a large proportion of boundary samples (over 80%) still are pushed to centroids where their identity information is rarely presented. Unfortu-



(a) Market-1501

(b) MSMT17

Figure 3.8: Visualization of the identity feature distribution via t-SNE [173] on (a) Market-1501 and (b) MSMT17. For each group, features are derived by the baseline model (left) and the model trained with the proposed confidence-guided centroids (CGC) and pseudo labels (CGL) (right), respectively. Model performances (mAP) are also denoted. Different identities are denoted by different colors. **Best viewed in color**.

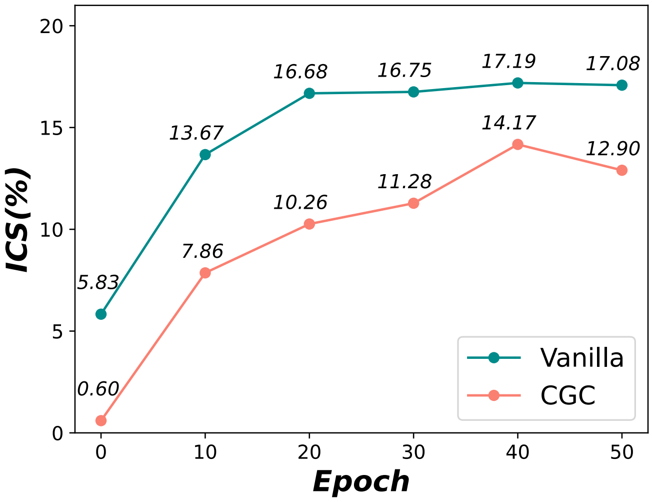


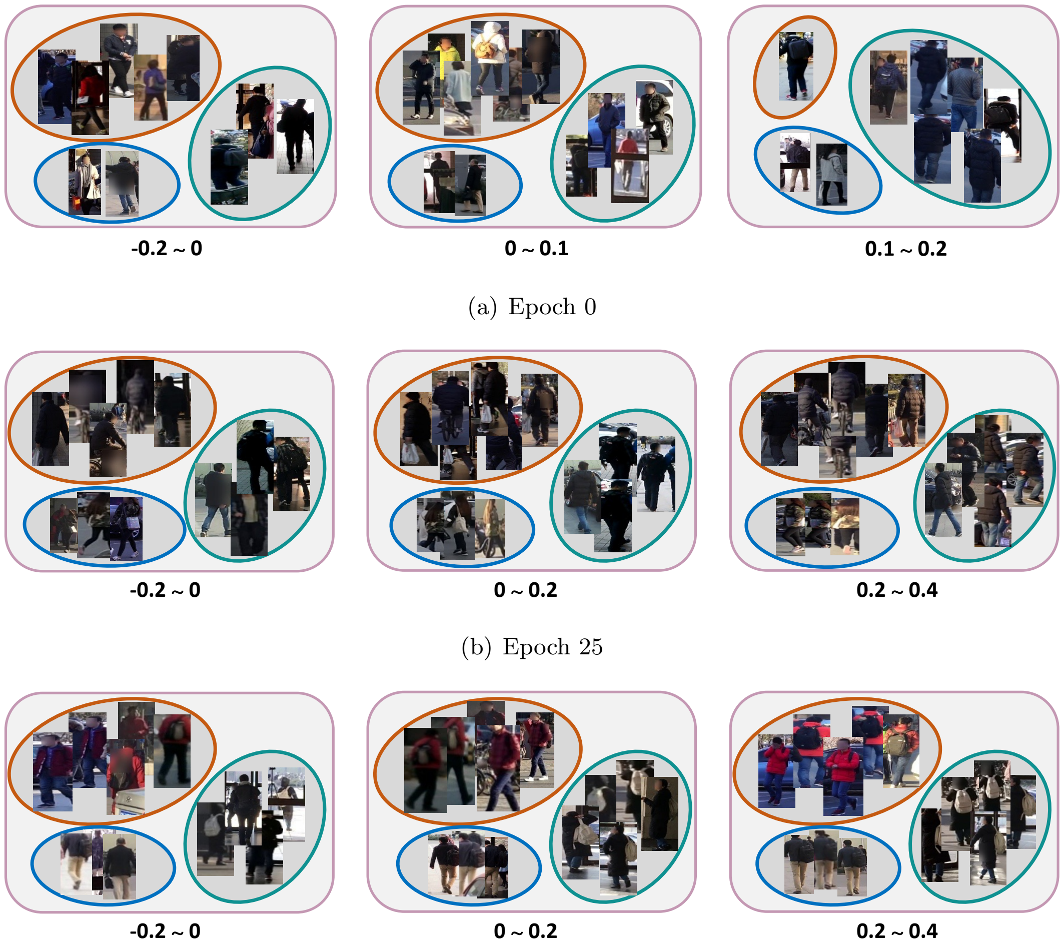
Figure 3.9: Identity consistent score (ICS) of boundary samples at different epochs. Plain and CGC refer to the previous all-sample based cluster centroids and the proposed confidence-guided centroids, respectively.

nately, the problem has been aggravated by confidence-guided centroids, where the ratio achieves 14.17% at most. The low identity consistency scores point out the seriousness of the problem and validate the necessity of the proposed confidence-guided pseudo labels.

More samples with silhouette scores. To better demonstrate the relationship between the sample-wise clustering confidence and silhouette scores, more clustering results of the proposed method are visualized in Fig. 3.10.

As can be seen, samples are coarsely clustered based on basic visual features, such as image quality, at the beginning. As more identity-related information is learned, yet belonging to different identities, samples with similar appearances and poses are gradually grouped together. Additionally, top-ranking images have higher silhouette scores at epoch 25. Finally, at a later stage (epoch 50), better identity information is learned, presenting as images of the same identity are grouped together while those belonging to different identities are scattered into different clusters.

Clustering quality. In the section, the author analyzes the improvements brought about by the proposed strategies in terms of clustering quality. Four evaluation metrics are employed in this work, which are fowlkes_mallows_score, adjusted_rand_score, adjusted_mutual_info_score and v_measure_score, respectively. All the above metrics represent the consistency between clustering results and ground-truth labels (*higher is better*). The author analyzes how the four metrics vary throughout the training process on two benchmark datasets, *i.e.*,Market-1501 and MSMT17. The results are illustrated in Fig. 3.11. To demonstrate the advantages of the proposed method, the comparison on clustering quality is conducted with "Baseline" models and a

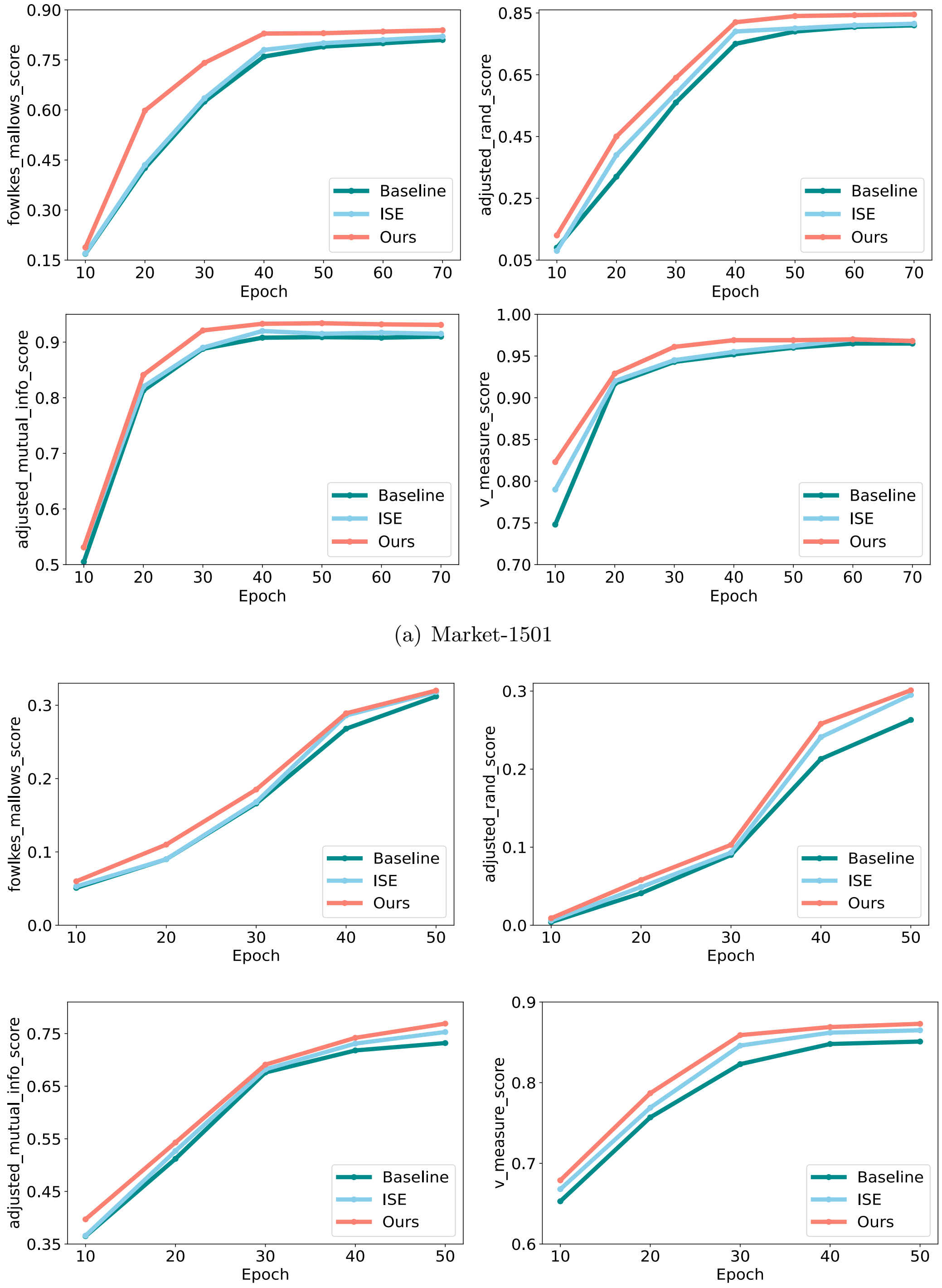


(a) Epoch 0

(b) Epoch 25

(c) Epoch 50

Figure 3.10: Visualization of samples (MSMT17) with silhouette scores at (a) epoch 0, (b) epoch 25, and (c) epoch 50. Samples within different clusters are indicated by different colors. **Best viewed in color**.



(a) Market-1501

(b) MSMT17

Figure 3.11: Comparison of clustering quality during the training between the baseline model (Baseline), ISE [235], and the proposed method (Ours) on (a) Market-1501 and (b) MSMT17. **Best viewed in color**.

state-of-the-art method, ISE [235]. The results are reported in Fig. 3.11. Note that the ISE curve is plotted by the estimated values from their published work.

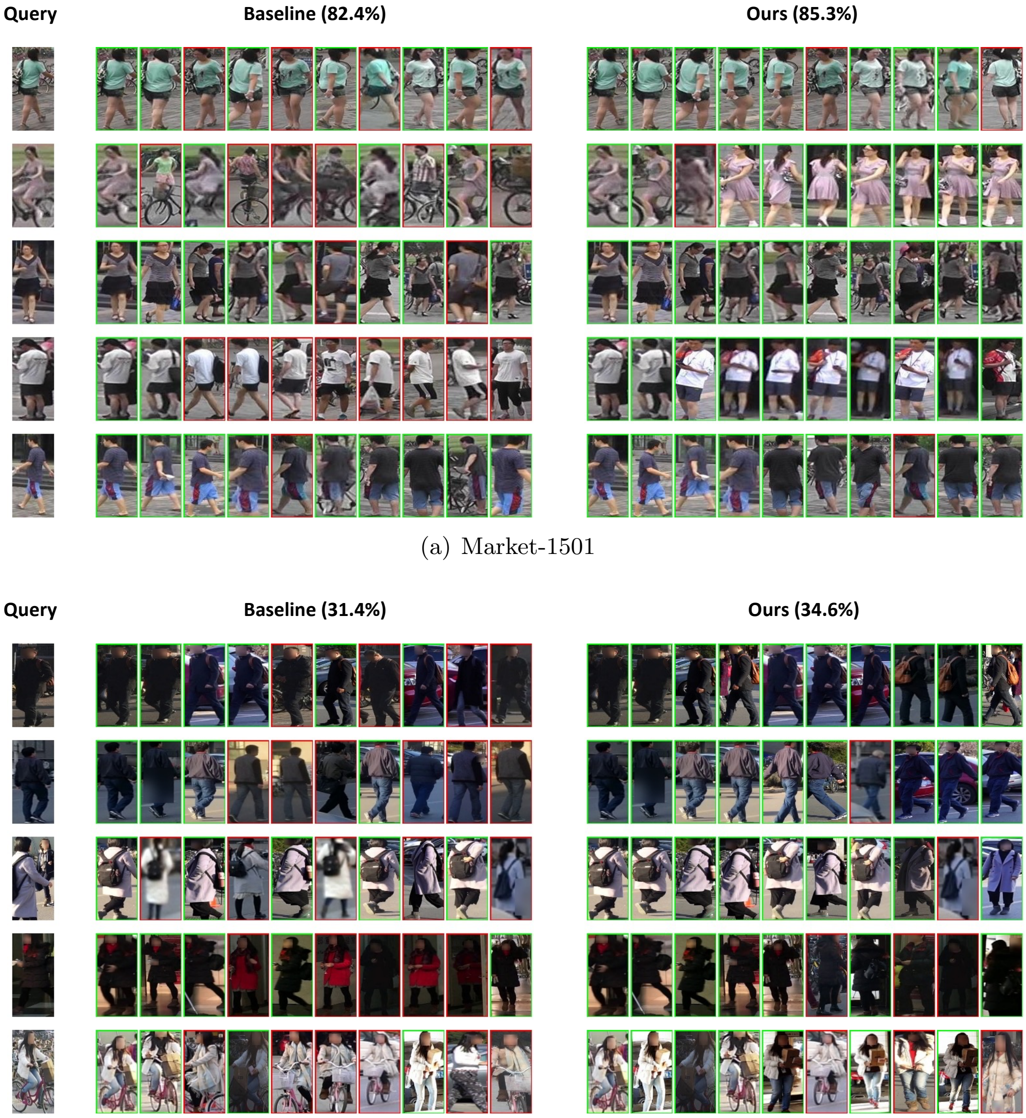
As can be seen, for all training schemes, the clustering quality gradually improves during training due to the involvement of more effective features as well as more discriminative networks. Additionally, the model trained with the proposed schemes outperforms both baseline models and ISE on all metrics. The improvements validate the effectiveness of the proposed schemes in clustering quality boosting.

Identity retrieval results. To validate the effectiveness of the proposed method, we present retrieval results on benchmark datasets in Fig. 3.12.

As can be seen, MSMT17 is more challenging due to the diversity in backgrounds, illuminations, poses, and occlusions. Compared to baseline models, models trained with the proposed schemes have better person re-identification accuracy, presenting as more matched images ranking at the top. Additionally, identity features extracted by the proposed method have more identity-related information. Taking the second row in Fig. 3.12 as an example, a girl in pink dress riding a bicycle is given as the query image. Among the top-10 images proposed by the baseline model, 6 images containing bicycle yet with different identities are wrongly selected. Such incorrect proposals indicate that the baseline models are prone to noises, such as cluttered backgrounds. However, the proposed confidence-guided centroids (CGC) and confidence-guided pseudo labels (CGL) encourage samples to approach not only better centroids where low-confidence samples are excluded, but also multiple potentially correct clusters. With the aid of the proposed schemes, more identity-related information is embedded in extracted features. Therefore, 9/10 images match the query image correctly, while the rest is highly similar to the query.

3.8 Conclusion

In this section, the author focused on the pseudo label refinement for clustering-based unsupervised person ReID, which aims to alleviate the pseudo label noise brought by imperfect clustering results. Two major contributions are made by this work: 1) Instead of relying on auxiliary information such as camera IDs, body parts, or generated samples, the author refined pseudo labels with internal characteristics, *i.e.,* the sample-wise clustering confidence. Specifically, the author proposed to use confidence-guided centroids (CGC) to provide reliable cluster-wise prototypes for feature learning, where low-confidence instances are filtered out during the formation of centroids. 2) Targeting the problem that a large proportion of samples are pushed to “wrong” centroids, the author proposed to use confidence-guided pseudo labels (CGL). Such labeling enables samples to approach not only the assigned centroid but other clusters where their identities are potentially embedded. With the aid of CGC and CGL, the proposed method yields comparable performances with or even outperforms state-of-the-art pseudo label refinement works that largely leverage auxiliary information.



(a) Market-1501

(b) MSMT17

Figure 3.12: Visualization of identity retrieval of the baseline model (Baseline) and the model trained with the proposed schemes (Ours) on (a) Market-1501 and (b) MSMT17. The performance of models (mAP) is also reported. For each group, the query image is shown at the leftmost, followed by the top 10 images of its ranking list given by different models. The green rectangles indicate correct retrieval results while the red ones denote false retrieval results. **Best viewed in color**.

Chapter 4

Fast unsupervised person re-identification

4.1 Introduction

Due to the absence of identity labels and the less informative binary codes, fast unsupervised person ReID is extremely challenging and remains an untouched field. Considering that person ReID can be regarded as image retrieval performed on person images according to identity information, some state-of-the-art image retrieval works can be adapted to fast person ReID with slight modifications. Therefore, to overcome challenges resulted from unsupervised learning manner and binary code learning, in this section, existing unsupervised binary descriptor learning methods are reviewed firstly. Targeting at the limitations, an novel unsupervised binary descriptor learning framework is proposed. Then, the proposed method is adapted to person ReID to achieve a fast unsupervised person ReID baseline.

Binary descriptors are widely applied in visual tasks like visual search [42], face recognition [236], *etc.* Therefore, learning effective binary descriptors has become an active topic in the computer vision community due to high compactness and high matching speed, which perfectly fits the demands of large-scale data. Over the past decade, numerous binary local descriptors have been proposed, including hand-crafted ones (BRISK [100], BRIEF [14], ORB [153], *etc.*), and learning-based ones (Binboost [171], LDAHash [163], *etc.*). Inspired by the advances of deep learning techniques, deep learning approaches for binary local descriptors have recently drawn increasing attention, like DeepBit [112], DBD-MQ [47], L2-Net [169], and GraphBit [46]. Depending on whether the labeled data are required, deep binary local descriptors can be further categorized as supervised [169, 133, 211] and unsupervised [112, 47, 46, 262] ones. Supervised methods generally achieve better performance with the supervision given by pairwise labels, indicating whether two patches come from the same category or not. However, such pairwise labels are too expensive to obtain in real-world applications. Therefore, unsupervised learning methods have gained more attention recently. Despite the remarkable performance improvements, there are still problems that need to be better addressed.

Firstly, an effective binary local descriptor should be robust against geometric transformations, *i.e.,* rotation, scaling, and viewpoint changes. The robustness of local descriptors will affect the matching accuracy in matching/retrieval tasks [125]. Earlier binary local descriptors [100, 14, 153] are built upon hand-crafted sampling patterns or pairwise intensity comparisons, which are vulnerable to geometric distortions due to the high sensitivity of hand-crafted features. Thus, hand-crafted binary local descriptors tend to have unstable performances [112]. On the other hand, most existing deep unsupervised binary local descriptors focus more on generating effective compact codes but pay little attention to the robustness against geometric transformations [47, 46]. A prior work, DeepBit [112], enhances the robustness of the descriptor against rotation via minimizing the Hamming distance between the descriptors of an original image and its transformed counterparts. Although it provides an intuitive way to generate the transformation-invariant local descriptors, a problem might be that such work is based on the idea that an original image and its transformed counterparts should be represented by different descriptors. However, ideally, the same object is expected to be described by exactly the same descriptor, regardless of the viewpoint or distance changes. Therefore, simply minimizing the distance between the original image and its transformed counterparts is not the optimal solution.

Secondly, an effective binary local descriptor is supposed to be informative, *i.e.,* each bit carrying distinctive information. However, previous learning-based descriptors generally follow the scheme of image hashing. Yet an image patch, as a small region around an interest point, generally contains much less information than an image. Therefore, directly employing image hashing schemes can probably lead to highly-correlated bits, which means information contained in different bits can be redundant during encoding. That would make the learned descriptor not compact enough. To explain the problem of correlated bits, the author first evaluates the average amount of information conveyed by image patches and images with *Shannon entropy.* Then, hash codes and binary local descriptors are derived from images and patches with two popular hashing methods: DeepBit [112] and Bi-half Net [109]. Later, the author compares the correlations between bits under different code length settings with mean Absolute Correlations (mAC), which indicates the average correlation between bits. A higher mAC means a higher bit correlation. Details of mAC could be found in Section 4.7.4. Specifically, the same number of images and image patches are randomly selected from an image dataset (CIFAR10 [97]) and an image patch dataset (Brown [12]) and are resized to the same size. The average Shannon entropy and mAC scores are illustrated in Table 4.1. Seen from the results, the mAC scores under 32 bits and 64 bits settings are given by DeepBit since both source code and trained models are provided. The mAC scores under 128 bits and 256 bits settings are obtained by reproducing Bi-half Net based on the provided source code. Table 4.1 clearly demonstrates that images, with a higher average Shannon entropy, generally carry more complex information than image patches. When image hashing schemes are directly employed to derive binary local descriptors, the average correlations between bits exceed that of images by 1.17%, 3.22%, 9.90% and 10.13% under 32, 64, 128 and 256 bits settings in terms of mAC scores,

Table 4.1: Comparison of the average Shannon Entropy between images and image patches, and that of mean Absolute Correlations (mAC) (%) between corresponding hash codes and binary descriptors, under different code length settings.

	Entropy	mAC			
		32 bits	64 bits	128 bits	256 bits
Images	9.21	4.16	4.95	5.67	6.04
Patches	4.28	5.39	8.17	15.57	16.17

respectively. Strong correlations between bits will undoubtedly deteriorate the representability of local descriptors [72, 211]. To mitigate the problem, most existing deep learning based works enforce the bits of binary local descriptors to be evenly distributed [49, 112, 46], which is performed on each training batch. However, such batch-based constraints generally suffer from a problem that the data distribution of a single batch cannot well represent that of the whole dataset due to the limited number of samples within a batch.

To tackle both limitations, in this work, a novel *Transformation-invariant Binary Local Descriptor* learning method (TBLD) is proposed, which is trained in an unsupervised manner. The pipeline of TBLD is illustrated in Fig. 4.1. Specifically, it takes the "Original set" and "Transformed sets" as input. The former consists of the original image patches from the dataset, while the latter is built by rotating and scaling the original image patches. The framework aims to derive transformation-invariant and low-coupling binary local descriptors.

To generate transformation-invariant binary local descriptors, visual features extracted from original image patches and their transformed counterparts are enforced to be projected into an identical Euclidean subspace and an identical Hamming subspace simultaneously. Meanwhile, the distinctiveness between binary local descriptors of dissimilar image patches is maximized. To achieve that, instead of utilizing two separate terms during the optimization, an integrated loss term, contrastive loss [25], is introduced here to propagate the neighboring structures of data from a high-dimensional feature space to a low-dimensional descriptor space. As a departure from [25], where ALL the transformed samples within a training batch are employed as negative samples, a *negative pairs selection strategy* is proposed in this work to adaptively select "Negative pairs" for each image patch during the training. By doing so, similar image patches from the same batch will form only ONE negative pair with a given image patch, instead of multiple negative pairs, thus dramatically reducing the computational costs.

In the meantime, to reduce bit correlations, instead of manually imposing deterministic regularization terms on a batch of binary local descriptors, low-coupling binary codes are introduced externally here to guide the learning of binary local descriptors. Specifically, an *Adversarial Constraint Module* (ACM), which adopts the scheme of generator-discriminator, is adopted. The Wasserstein loss employed in the *Discriminator* minimizes the distributional discrepancy between the binary local descriptors generated by the framework and the introduced low-coupling binary codes. Although the proposed bottom-up learning strategy is employed at the batch level as most correlation regularizers do, the optimization of *Discriminator* is an accumulated result of all previous batches, meaning that the adopted adversarial regularization is not restricted to the number of samples within a batch. In summary, the contributions made in the proposed work are mainly three-fold:

- An unsupervised binary local descriptor, which unites transformation-invariant and low-coupling properties, is proposed. To ensure the transformation invariance of binary local descriptors, contrastive loss is, for the first time, applied in the learning of binary local descriptors. Instead of involving a large number of negative samples, a *negative pairs selection strategy* is proposed to selectively pick up a portion of "Negative pairs" for each training batch.

- The problem of the high correlations between bits in binary local descriptors when directly applying image hashing methods is highlighted. To tackle that, a bottom-up learning strategy is proposed, termed *Adversarial Constraint Module* (ACM). Low-coupling binary codes generated externally are employed to guide the learning of binary local descriptors by minimizing their Wasserstein distances. This, by all means, is distinct from existing methods that simply use a hard threshold to enforce each bit to be evenly distributive.

- Experimental results on three benchmark datasets show that the proposed descriptor surpasses existing binary descriptors by a clear margin in various visual tasks.

4.2 Previous solutions

4.2.1 Unsupervised binary descriptors

Existing unsupervised binary local descriptors [112, 47, 46] improve the representability from mainly two aspects: 1) enhancing the robustness to geometric transformations; 2) enriching the embedded information via reducing the bit correlations.

Robustness enhancement. GraphBit [46] improves the robustness of binary descriptors by enhancing the representability of each bit. The mutual information between inputs and related bits is maximized so that the ambiguous bits could receive additional instruction for confident binarization. DBD-MQ [47] enhances the quality of binary descriptors by applying a data-dependent binarization strategy. A K-AutoEncoders network is trained along with the holistic features to classify bits into the 0/1 categories with minimal reconstruction error. Such a distribution strategy delivers stronger robustness since bits from similar holistic features are more likely to be quantized into the same binary codes. Aside from improving the binarization functions, DeepBit [112] augments patches via rotation and scaling, and employs a Siamese network to minimize the distances between the binary local descriptors of original image patches and their augmented counterparts. However, from the perspective of the essence of local descriptors, which is to describe the content in an image patch, the author argue that the same content should be described by the *same* local descriptors in spite of viewpoints, instead of similar ones.

Bit correlation reduction. Existing works, *e.g.,* DH [49], DeepBit [112], GraphBit [46], UDBD [196], simply enforce the learned local descriptors to be evenly-distributive, *i.e.,* encouraging the mean of each bit to be 0.5 with the bit value ranging in $[0, 1]$. On top of that, BinGAN [262] embeds an adjusted Binarization Representation Entropy Regularizer to increase the entropy of the particular pairs of binary vectors that are not correlated in the high-dimensional feature space. Generally, such constraints are performed within training batches. However, the number of samples within a training batch is limited, meaning that the feature distribution of each batch cannot well represent that of the whole dataset. Therefore, imposing batch-based constraints typically fails to achieve the global optimum. Instead of performing constraints directly on the derived binary local descriptors, the framework here encourages to learn the mapping from the derived binary local descriptors to the low-coupling binary codes, which are introduced externally.

In the paper, the transformation invariance of binary local descriptors is achieved by projecting original image patches and their transformed counterparts into an identical Euclidean subspace and an identical Hamming subspace with the help of contrastive loss. Additionally, a bottom-up learning strategy assisted with Wasserstein loss is proposed to reduce bit correlations, where low-coupling binary codes are introduced externally to guide the learning of binary local descriptors.

4.2.2 GAN based binary descriptor

Generative Adversarial Network (GAN) [64] has been extensively involved in unsupervised learning, where synthetic images are continuously generated to “fool” the network during training for improving the discriminability of the network. Inspired by its successful applications in feature learning [197, 26] and text-to-image generation [227], GAN has been recently introduced in the field of image hashing [15, 162]. HashGAN [15] utilizes generators to synthesize diverse images and employs a discriminator to distinguish the synthetic images and the real ones. Meanwhile, a *Hash Encoder* learns the binary hash codes with the similarity information between images being preserved. BGAN [162] employs an auto-encoder to jointly learn binary hash codes in the middle and generate synthetic images at the end. The representability of the learned binary hash codes is improved by minimizing the distances between reconstructed images and the original ones. Meanwhile, the neighboring structures of images and features are also preserved. More recently, GAN has been applied in the learning-based binary descriptors [262]. BinGAN [262] takes an intermediate layer representation of a discriminator as the compact local binary descriptor. Two regularizers are also proposed to reduce the correlation between binary local descriptors.

Contrary to the proposed work, HashGAN [15] and BGAN [162] are specifically designed for image retrieval tasks and use tanh-like activation for binarization.

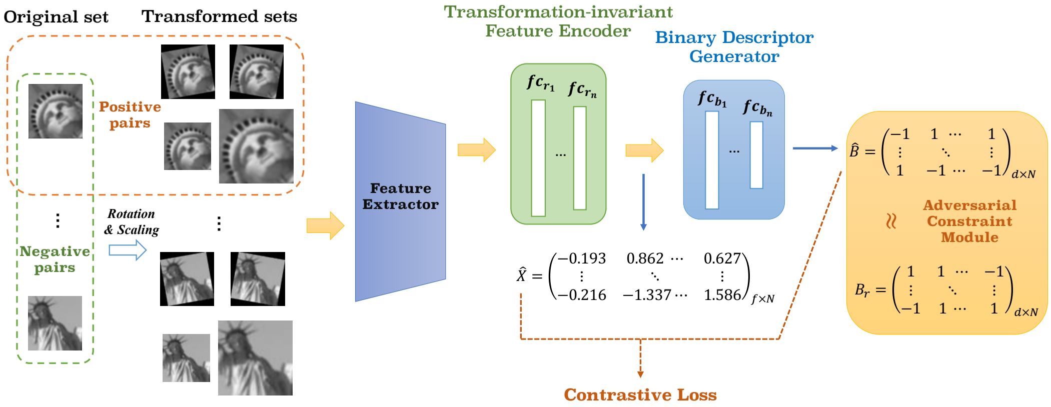

Original set
Transformed sets
Positive pairs
Negative pairs
Rotation & Scaling
Feature Extractor
Transformation-invariant Feature Encoder
fc_{r_1}
fc_{r_n}
Binary Descriptor Generator
fc_{b_1}
fc_{b_n}
$\hat{X} = \begin{pmatrix} -0.193 & 0.862 \cdots & 0.627 \\ \vdots & \ddots & \vdots \\ -0.216 & -1.337 \cdots & 1.586 \end{pmatrix}_{f \times N}$
$\hat{B} = \begin{pmatrix} -1 & 1 \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & -1 \cdots & -1 \end{pmatrix}_{d \times N}$
Adversarial Constraint Module
≈
$B_r = \begin{pmatrix} 1 & 1 \cdots & -1 \\ \vdots & \ddots & \vdots \\ -1 & 1 \cdots & 1 \end{pmatrix}_{d \times N}$
Contrastive Loss

Figure 4.1: Pipeline of the proposed TBLD. Firstly, patches from the "Original set" are augmented by rotating and scaling to build "Transformed sets". Then visual features of the image patches from the "Original set" and "Transform sets" are extracted from the VGG16 network. Subsequently, visual features are encoded by a *Transformation-invariant Feature Encoder* to obtain high-dimensional transformation-invariant features. On top of that, transformation-invariant binary local descriptors are obtained by *Binary Descriptor Generator*. f, d denote the dimension of high-dimensional transformation-invariant features and binary local descriptors, respectively. N is the number of training samples. Additionally, an *Adversarial Constraint Module* (ACM) is introduced to reduce bit correlations.

However, the proposed work focuses on patch descriptor based tasks, like patch matching. Additionally, instead of taking the intermediate representations from the discriminator, this work employs *Discriminator* along with a set of low-coupling binary codes to guide the network to directly generate low-coupling binary local descriptors from the descriptor generator.

4.3 Problem statement

To learn effective binary local descriptors, a Transformation-invariant Binary Local Descriptor learning framework (TBLD) is proposed, which improves the representability of local descriptors in terms of robustness and bit correlations. To enable binary local descriptors to be invariant to transformations, inspired by visual representation learning [25], contrastive loss is employed to preserve the neighboring structures of data. Specifically, the original image patches and their transformed counterparts are projected to an identical Euclidean subspace and an identical Hamming subspace, while the distinctiveness between binary local descriptors of dissimilar image patches is maximized. Additionally, an *Adversarial Constraint Module* (ACM) is introduced to reduce bit correlations, where low-coupling binary codes are introduced externally to guide the learning of binary local descriptors.

The pipeline of the proposed TBLD is depicted in Fig. 4.1. Specifically, given an image patch set $I^0 = \{I_i^0\}_{i=1}^c$ with c patches, I_i^0 refers to the i-th image patch.

Firstly, v transformed patch sets are built, with each containing one certain type of transformation on the original image patches, like rotation or scaling. Then, the whole training set $I = \{I^i\}_{i=0}^{v}$ is formed by I^0 and the v transformed patch sets $\{I^i\}_{i=1}^{v}$. After that, visual features of all patches, $T = \{T^i \in \mathbb{R}^{t \times c}\}_{i=0}^{v}$, are extracted via the well-known VGG16 network [160], where t refers to the feature dimension. Subsequently, visual features are encoded by a *Transformation-invariant Feature Encoder* to obtain r-dimensional transformation-invariant features $\mathbf{X} \in \mathbb{R}^{r \times c}$. On top of that, a group of b-bit transformation-invariant binary local descriptors $\mathbf{B} \in \mathbb{R}^{b \times c}$ are obtained by binarizing the output of *Binary Descriptor Generator* $\mathbf{F} \in \mathbb{R}^{b \times c}$ as follows,

$$\mathbf{B} = sign(\mathbf{F}). \tag{4.3.1}$$

In general, $r > b$. As claimed, image patches and their transformed counterparts are united to the identical high-dimensional features $\mathbf{X}$ and binary local descriptors $\mathbf{B}$.

4.4 Transformation-invariant binary descriptors

4.4.1 Pseudo positive/negative pairs selection

To preserve the neighboring structures of image patches from the feature space to the descriptor space, contrastive loss is performed after both *Transformation-invariant Feature Encoder* and *Binary Descriptor Generator*. For both modules, the feature representations of “Pseudo Positive pairs” are projected to an identical Euclidean subspace and an identical Hamming subspace. Meanwhile, the distinctiveness between the feature representations of “Pseudo Negative pairs” is maximized. Since pair-wise matching labels are not available here, the author employs the neighboring relationships of image patches to build both “Pseudo Positive pairs” and “Pseudo Negative pairs”. For simplicity, “Positive pairs” and “Negative pairs” are used to refer to “Pseudo Positive pairs” and “Pseudo Negative pairs”, respectively.

Specifically, a “Positive pair” is built by an original patch and any one of its transformed counterparts. And a “Negative pair” is formed by an original patch and a sampled “Negative set”. In the proposed scenario, there are numerous image patches in the dataset, which means exhaustively pairing the given original patch with the rest of patches seems impractical in terms of computational costs. Therefore, a *negative pairs selection strategy* is proposed, which selectively picks up a “Negative set” to form the “Negative pairs”. Concretely, given a batch with M image patches, q different image patches are selected to form a “Negative set” according to their *“clusters”*. The selection of “Negative set” is illustrated in Fig. 4.2, where the extracted visual features T are clustered into 32 clusters offline. During training, for a given batch, the cluster distribution of the image patches within the batch is analyzed. Then samples are randomly selected from the uncovered clusters to build the “Negative set”. If all the clusters are covered, samples are selected randomly and evenly from each cluster to form the “Negative set”. In the experiment, q is empirically set as 4096, considering the balance between computational cost and data diversity.

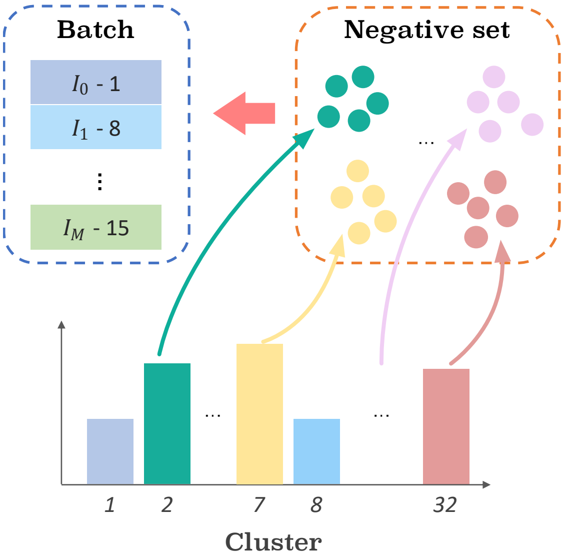


Figure 4.2: Illustration of the selection of "Negative sets" by the proposed negative pairs selection strategy. Visual features of image patches are firstly clustered into 32 clusters. For a given batch during training, the corresponding "Negative set" is formed according to the feature cluster of patches within the batch.

4.4.2 Contrastive loss

Given the "Positive pairs" and "Negative pairs", contrastive loss is performed after both *Transformation-invariant Feature Encoder* and *Binary Descriptor Generator* to propagate the neighboring structures from high-dimensional features to compact binary local descriptors. The two loss terms are represented by L_{C_r} and L_{C_b}, respectively, which enforce "Positive pairs" to have identical transformation-invariant high-dimensional features and compact binary local descriptors, respectively. Meanwhile, the distinctiveness between feature representations of "Negative pairs" is maximized. Firstly, L_{C_r} is formulated as follows.

$$L_{C_r} = -\sum_{i=0}^{v}\sum_{m=1}^{M}\frac{\alpha_i^{\gamma}}{M} log\frac{e^{-s_r Dist_E(x_m^i, x_m)}}{e^{-s_r \overline{Dist_E}(x_m^i, x_{neg})}}, \tag{4.4.1}$$

where x_m^i denotes the output of *Transformation-invariant Feature Encoder* of the m-th patch from the i-th "Transformed sets", and x_m is the transformation-invariant high-dimensional feature of the m-th patch. $Dist_E(x_m^i, x_m)$ represents the distance between x_m^i and x_m, which is formulated as follows,

$$Dist_E(x_m^i, x_m) = \|x_m^i - x_m\|_2^2. \tag{4.4.2}$$

In the denominator, the average Euclidean distance between x_m^i and the corresponding high-dimensional *"negative"* set x_{neg} formed by the real-valued representations of q *"negative"* samples is computed as,

$$\overline{Dist_E}(x_m^i, x_{neg}) = \frac{1}{q} \sum_{x_j \in x_{neg}} \|x_m^i - x_j\|_2^2. \tag{4.4.3}$$

In Eqn. (4.4.1), α_i denotes the to-be-learned non-negative weight w.r.t the i-th "Transformed sets", which sums up to 1. γ is a smoothing parameter and s_r denotes the temperature parameter for real-valued local descriptors, which are empirically set as 3 and 0.1, respectively.

Similarly, contrastive loss applied after *Binary Descriptor Generator*, *i.e.,* L_{C_b}, is defined as,

$$L_{C_b} = -\sum_{i=0}^{v} \sum_{m=1}^{M} \frac{\alpha_i^\gamma}{M} log \frac{e^{-s_b Dist_H(b_m^i, b_m)}}{e^{-s_b \overline{Dist_H}(b_m^i, b_{neg})}}, \tag{4.4.4}$$

where b_m^i denotes the binary feature representations of the m-th patch in the i-th "Transformed sets", and b_m indicates the transformation-invariant binary local descriptor of the m-th patch. $Dist_H$ denotes the Hamming distance and $\overline{Dist_H}$ represents the average Hamming distance between the binary string of the given image patch and the binary local descriptors of its counterparts from the "Negative set" b_{neg}. s_b denotes the temperature parameter for binary local descriptors, which is empirically set as 0.1.

However, directly optimizing binary values will make the back-propagation of the framework infeasible, which is known as the *ill-posed gradient* problem [162]. In the work, b_m^i and b_m in Eqn. (4.4.4) are replaced with the relaxed real-valued representations output by *Binary Descriptor Generator* before the binarization f_m^i, and the to-be-binarized transformation-invariant local descriptors f_m, respectively. Then Eqn. (4.4.4) can be rewritten as follows.

$$L_{C_b} = -\sum_{i=0}^{v} \sum_{m=1}^{M} \frac{\alpha_i^\gamma}{M} log \frac{e^{-s_b Dist_E(f_m^i, f_m)}}{e^{-s_b \overline{Dist_E}(f_m^i, f_{neg})}}. \tag{4.4.5}$$

The replacement requires a low quantization error between the binary local descriptors and the corresponding relaxed real-valued feature representations. Therefore, a quantization loss term L_Q is employed, which is denoted as,

$$L_Q = \sum_{i=0}^{v} \sum_{m=1}^{M} \frac{\alpha_i^\gamma}{M} \|f_m^i - b_m\|_2^2, \tag{4.4.6}$$

where f_m^i denotes the relaxed real-valued feature representations of the m-th patch in the i-th "Transformed sets". And b_m refers to the transformation-invariant binary local descriptor of the m-th patch, which refers to the m-th column in $\mathbf{B}$.

4.5 Low-coupling binary descriptors

Apart from enhancing the robustness of binary local descriptors against transformations, decorrelating bits of the compact descriptor is also of great importance. As revealed previously, correlated bits convey overlapped information, thus weakening the representation capacity of the binary local descriptors. According to [4], *Wasserstein distance* can measure the distance between two non-overlapped data distributions, which perfectly fits the situation where discrete and continuous distributions coexist. Inspired by this, this work advocates the use of Wasserstein loss to minimize the *Wasserstein distance* between the data distribution of low-coupling binary codes and the feature distribution of the derived binary local descriptors. Although the Wasserstein loss has been successfully employed in applications like person re-identification [251, 2], it has never been employed to learn binary local descriptors yet.

In the paper, a bottom-up learning strategy is proposed to reduce bit correlations, termed *Adversarial Constraint Module* (ACM). The structure of ACM is depicted in Fig. 4.3, which adopts the scheme of generator-discriminator in adversarial learning. Specifically, the proposed framework serves as a *Descriptor Generator* to derive binary local descriptors. Meanwhile, a sampler is employed to generate low-coupling binary codes by *randomly* and *independently* sampling 0/1 values from the Bernoulli distribution with the probability $p = 0.5$, which conforms to the principle of local descriptors in [112]. Given the input, a *Discriminator*, consisting of three fully-connected (FC) layers, is followed. The first two FC layers are followed by a ReLU activation function. In *Discriminator*, the Wasserstein loss is employed to encourage the derived binary local descriptors to mimic the distribution of the low-coupling binary codes by alternately optimizing the *Discriminator* and the *Descriptor Generator*.

Formally, given a training batch with M image patches $I = \{I_i\}_{i=1}^{M}$, a batch of binary local descriptors $\boldsymbol{B} = \{b_i\}_{i=1}^{M}$ could be learned by the *Descriptor Generator*. Similarly, to avoid the *ill-posed gradient* problem [162], the binary local descriptors $\boldsymbol{B}$ is replaced with the relaxed representations $\boldsymbol{F} = \{f_i\}_{i=1}^{M}$, which refers to the real-valued feature representations output by the *Binary Descriptor Generator* before binarization. Additionally, the low-coupling binary codes $\boldsymbol{B_r} = \{b_{r_i}\}_{i=1}^{M}$ are sampled under the Bernoulli distribution. The *Wasserstein distance* between $\boldsymbol{F}$ and $\boldsymbol{B_r}$ could be approximated by,

$$\max \mathbb{E}_{b_{r_i} \sim b}[D(b_{r_i})] - \mathbb{E}_{f_i \sim \mathbb{P}_f}[D(f_i)], \tag{4.5.1}$$

where D refers to the discriminator. $\mathbb{P}_f$ and b refer to the feature distribution of $\boldsymbol{F}$ and data distribution of $\boldsymbol{B_r}$, respectively.

According to [4], Eqn. (4.5.1) holds only when *Lipschitz constraint* is satisfied. Therefore, following [70], *Lipschitz constraint* is enforced by penalizing the p-norm of the gradient of the discriminator w.r.t. the input, *i.e.,* $\|\nabla_x D(x)\|_p \leq 1$. Following [70], instead of enforcing the gradient norm constraint being intractable in whole latent space, this work enforces it on the space that is uniformly sampled

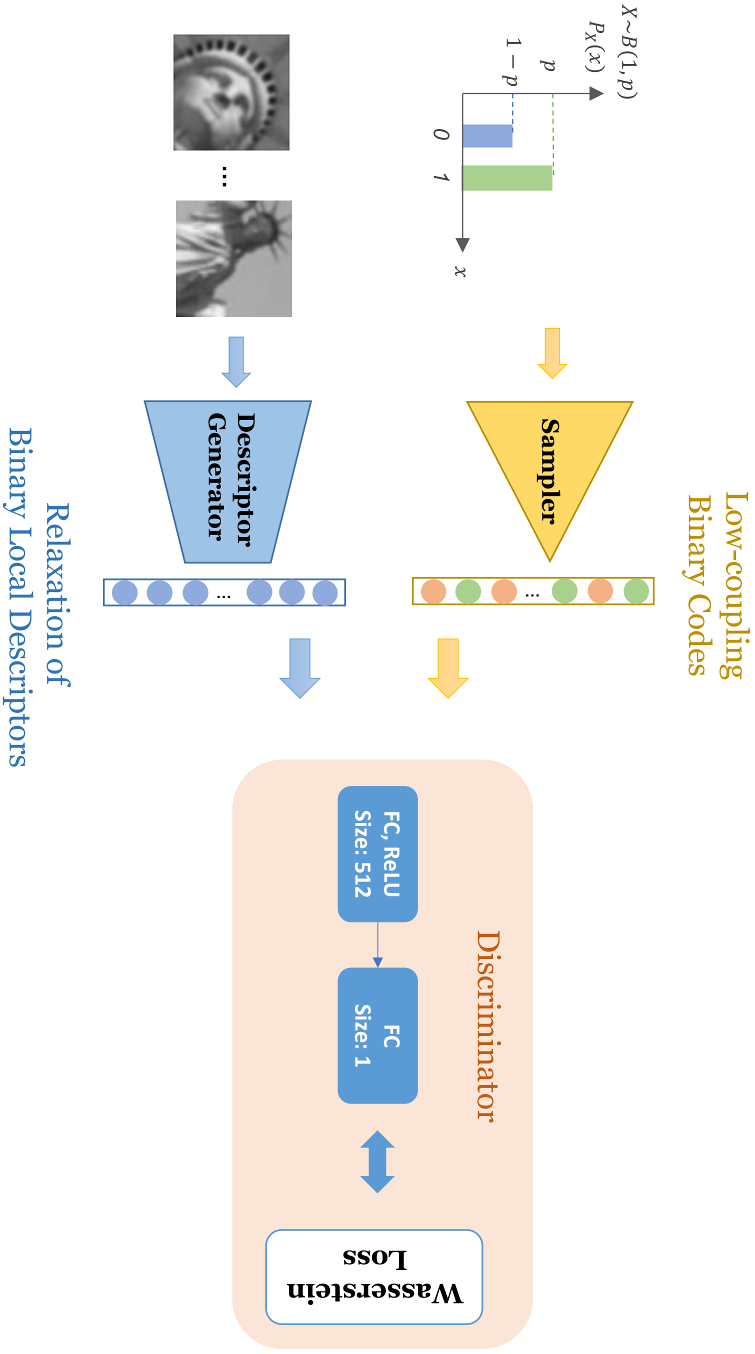


Figure 4.3: Structure of Adversarial Constraint Module (ACM). The discriminator takes the derived binary local descriptors from the *Descriptor Generator* and the low-coupling binary codes sampled from the Bernoulli distribution as input. With the help of Wasserstein loss, *Discriminator* and *Descriptor Generator* are alternately trained to learn the mapping from the derived binary local descriptors to the low-coupling binary codes.

from the feature distribution $\mathbb{P}_f$ and the data distribution b. Integrating the regularizer to the objective function, Wasserstein loss employed in *Adversarial Constraint Module* can be denoted as follows.

$$\begin{aligned} L_W = -\mathbb{E}_{b_{r_i}\sim b}[D(b_{r_i})] + \mathbb{E}_{f_i\sim\mathbb{P}_f}[D(f_i)] \\ +\eta\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}}D(\hat{x})\|_2 - 1)^2], \end{aligned} \tag{4.5.2}$$

where $\hat{x} \sim \mathbb{P}_{\hat{x}}$ is sampled from both inputs with a random sample weight $\epsilon \sim U[0,1]$, and it can be formulated as,

$$\hat{x} = \epsilon f_i + (1-\epsilon)b_{r_i}. \tag{4.5.3}$$

Notably, there are some negative values in the $\boldsymbol{F}$ since the *Binary Descriptor Generator* is trained to push $\boldsymbol{F}$ to $[-1,1]$. However, the sampled binary local descriptors $\boldsymbol{B_r} \in [0,1]$. To unify the two inputs, 0-value bits in sampled binary local descriptors $\boldsymbol{B_r}$ are replaced with -1. Additionally, to eliminate the input noise, L_2 normalization is applied on $\boldsymbol{F}$ before sending it to the Discriminator.

4.6 Objectives and optimization

4.6.1 Objectives

As the proposed method adopts the scheme of generator-discriminator, two loss terms, *i.e.,* L_G for the *Descriptor Generator* and L_D for the *Discriminator*, are employed, respectively.

Descriptor generator. Given 1) contrastive loss L_{C_r} in Euclidean space, 2) contrastive loss L_{C_b} in Hamming space, and 3) quantization loss L_Q, the objective for *Descriptor Generator* is written as follows:

$$L_G = L_{C_r} + L_{C_b} + \beta L_Q - \lambda_D \mathbb{E}_{f_i\sim\mathbb{P}_f}[D(f_i)], \tag{4.6.1}$$

where β balances the contribution of L_Q, and λ_D controls the penalty of the *Discriminator*, which are both empirically set as 1. Note that, to avoid plunging the network into the trivial solution, where all the real-valued feature representations become an all-zero or infinite matrix, the L_2-norm of the real-valued feature representations of each query is enforced to be 1, *i.e.,* $\|x_m^i\|_2 = 1$. To simplify the learning process, the constraint in the objective is integrated as L_N, which is denoted as follows,

$$L_N = \sum_{i=0}^{v}\sum_{m=1}^{M}\frac{\alpha_i^\gamma}{M}(1 - \|x_m^i\|_2), \tag{4.6.2}$$

Therefore, the objective L_G can be formulated as,

$$L_G = L_{C_r} + L_{C_b} + \beta L_Q + \lambda L_N - \lambda_D \mathbb{E}_{f_i\sim\mathbb{P}_f}[D(f_i)], \tag{4.6.3}$$

where λ is the weight for the regularizer L_N, which is set as 1e-5 empirically.

Discriminator. The *Discriminator* objective L_D is defined by,

$$L_D = \lambda_D L_W, \tag{4.6.4}$$

where λ_D is the same hyper-parameter as in Eqn. (4.6.3). In the paper, *Descriptor Generator* and *Discriminator* are trained with the SGD [10] optimizer with the initial learning rate being 5e-7 and 1e-8, respectively.

4.6.2 Optimization

The training procedure of TBLD is summarized in Algorithm 2. Firstly, the parameters of *Descriptor Generator*, including *Transformation-invariant Feature Encoder* (w_r) and *Binary Descriptor Generator* (w_b), are initialized following the Kaiming initialization [76]. The non-negative weights for the "Original set" and "Transformed sets", $\alpha = \{\alpha_i\}_{i=0}^{v}$, are all initialized as $1/v$. Given the to-be-learned variables, *i.e.,* the transformation-invariant high-dimensional features $\boldsymbol{X}$, the transformation-invariant binary local descriptors $\boldsymbol{B}$, and weights for input α, an alternating optimization method is proposed to solve the objective Eqn. (4.6.3) via conducting the following steps iteratively.

1. **Update $\boldsymbol{X}$.** With B, w_r, w_b, α fixed, the objective function w.r.t. $\boldsymbol{X}$ can be rewritten as,

   $$\psi_1 = \min_{X} L_{C_r}. \tag{4.6.5}$$

   By setting the derivation of Eqn. (4.6.5) w.r.t. $\mathbf{X}$ as 0, the closed-form solution of $\mathbf{X}$ can be formulated as:

   $$\mathbf{X} = \frac{\sum_{i=1}^{v} \alpha_i^{\gamma} \mathbf{X}^i}{\sum_{i=1}^{v} \alpha_i^{\gamma}}. \tag{4.6.6}$$

2. **Update $\boldsymbol{B}$.** Similarly, with other parameters fixed, the objective function w.r.t. $\boldsymbol{B}$ can be rewritten as follows.

   $$\psi_2 = \min_{B} L_Q. \tag{4.6.7}$$

   According to Eqn. (4.3.1), $\boldsymbol{B}$ can be obtained by binarizing $\boldsymbol{F}$ with the *sign* function. $\boldsymbol{F}$ can be obtained in a similar manner with $\boldsymbol{X}$. To conclude, $\boldsymbol{B}$ can be obtained as,

   $$\mathbf{B} = sign\Big(\frac{\sum_{i=1}^{v} \alpha_i^{\gamma} \mathbf{F}^i}{\sum_{i=1}^{v} \alpha_i^{\gamma}}\Big). \tag{4.6.8}$$

3. **Update α.** At the end of each epoch, with other parameters fixed, the objective function w.r.t. α can be rewritten as follows.

$$\psi_3 = \min_{\alpha} \sum_{i=0}^{v} \alpha_i^{\gamma} (L_{C_r}^{\tilde{i}} + L_{C_b}^{\tilde{i}} + \beta L_Q^{\tilde{i}} + \lambda L_N^{\tilde{i}}), \tag{4.6.9}$$

   where $L_{C_r}^{\tilde{i}}$, $L_{C_b}^{\tilde{i}}$, $L_Q^{\tilde{i}}$, $L_N^{\tilde{i}}$ are obtained from L_{C_r}, L_{C_b}, L_Q, L_N by factoring out α_i^{γ}, respectively. Suppose that $L^i = L_{C_r}^{\tilde{i}} + L_{C_b}^{\tilde{i}} + \beta L_Q^{\tilde{i}} + \lambda L_N^{\tilde{i}}$, the optimal α can be derived as,

$$\alpha_i = \frac{(L^i)^{\frac{1}{1-\gamma}}}{\sum_{j=0}^{v} (L^j)^{\frac{1}{1-\gamma}}}. \tag{4.6.10}$$

After alternately updating the parameters in *Descriptor Generator* and *Discriminator* until the network converges, the low-coupling binary local descriptors are derived from the *Descriptor Generator*.

Algorithm 2: The training procedure of the proposed TBLD.

```
Input: Visual features of all patches, $T = \{T^i \in \mathbb{R}^{t \times c}\}_{i=0}^{v}$; Number of
       iteration for each epoch, $N_b$;
Output: Real-valued local descriptors, $\boldsymbol{X}$; Binary local descriptors, $\boldsymbol{B}$
Initializing Descriptor Generator, Discriminator, and patch set
  weights $\alpha$;
Initializing $\boldsymbol{X_0}$ and $\boldsymbol{B_0}$ by Eqn. (4.6.6) and Eqn. (4.6.8);
while not covereged do
    for $i = 0 \to N_b$ do
        Sampling $\epsilon \sim U[0,1]$ and $\boldsymbol{B_r} \sim b(M, 0.5)$ ;
        Deriving $\boldsymbol{X}^i$, $\mathbf{F}^i$, $\boldsymbol{B}^i$ by Descriptor Generator;
        Optimizing Discriminator according to Eqn. (4.6.4) ;
        Optimizing Descriptor Generator according to Eqn. (4.6.3) ;
    Updating $\alpha_i$ by Eqn. (4.6.10);
    Updating $\boldsymbol{X}$ and $\boldsymbol{B}$ with the updated Descriptor Generator;
```

4.7 Experiments

The proposed binary local descriptor learning method is evaluated on three widely used public datasets, *i.e.,* **Brown** [12], **HPatches** [7], and **Mikolajczyk** [132]. Comparisons with the state-of-the-arts are conducted on visual analysis tasks like patch matching, patch retrieval, and patch verification. This section firstly introduces the datasets and experimental settings and then presents the comparison results and discussion.

4.7.1 Datasets

Brown dataset contains three subsets: *Liberty*, *Notre Dame* and *Yosemite*. Each subset contains 400,000 to 600,000 gray-scale image patches for training and 100,000 patch pairs for testing. The size of image patches in the dataset is 64×64. For the test sets, half of the pairs are matched and others are non-matched. This work follows the settings in [171], *i.e.,* training the network with one subset and then evaluating it on the other two subsets. There are 6 train-test combinations in total.

HPatches dataset consists of 116 image sequences, with 59 containing significant viewpoint changes and the rest containing illumination deformations. Each sequence includes a reference image and 5 target images. Image patches are detected in the reference image with Difference of Gaussians (DoG) detector and projected on the target images using the ground truth homographies. The sizes of patches are normalized to 65×65. Following the setting in [211], the to-be-evaluated model in this experiment is trained on the *Liberty* subset of the **Brown** dataset.

Mikolajczyk dataset consists of images from five scenes, which include variations in viewpoints (*Graffiti*), image quality (*Ubc*), illumination (*Leuven*), blurriness (*Trees*), and zoom & rotation (*Boat*). Each subset comprises a reference image and 5 target images, which are sorted by an increasing degree of distortion. Since TBLD is proposed to mainly handle the scale and viewpoint transformation, the evaluation is conducted on *Boat* and *Graffiti* subsets. Following the protocol [132], SIFT keypoint detector is first employed to detect 1000 interest points for each image in a reference-target image pair. Then the keypoints are matched via a brute-force search based on the Hamming distance between the corresponding binary descriptors. Following the setting of BinBoost [171], the to-be-evaluated model in this experiment is trained on the *Notre Dame* subset of the **Brown** dataset.

4.7.2 Implementation details

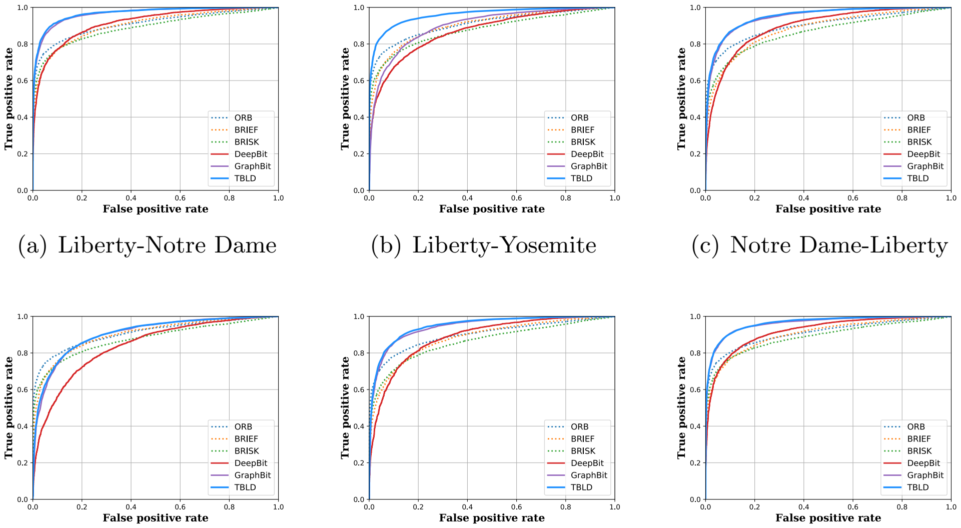
To preprocess the input data, two types of transformations, *i.e.,* rotation and scaling, are employed to derive the transformed sets. Rotation angles range in $\{-10, -5, 5, 10\}$, and scaling factors are set as 0.8 and 1.2. To obtain features from FC7 layer (4096-d) of the pre-trained VGG16 [160], the input patches are firstly resized into 256×256 and then cropped to 224×224. In this work, the length of real-valued and binary local descriptors are set as 1024 and 256, following the setting of [46]. The batch size is 32 and the maximum iteration is 10000.

4.7.3 Comparison with the state-of-the-art

Results on Brown dataset. Experiments on Brown aim to evaluate the performance of the proposed approach on the patch matching task. Following [112, 47, 46], the adopted evaluation metric is "95% error rate", which denotes the percent of incorrect matches when 95% of the ground-truth matched patches are found. *Lower represents better performance.* Comparisons are conducted with the state-of-the-art

Table 4.2: Comparisons of 95% error rate (%) with the state-of-the-art local descriptors on **Brown** [12] (*Lower is better*). The code lengths are indicated by dim and bytes for real-valued local descriptors and binary ones, respectively.

	Train	*Yosemite*		*Notre Dame*		*Liberty*		Average
	Test	*Notre Dame*	*Liberty*	*Yosemite*	*Liberty*	*Notre Dame*	*Yosemite*	Err
Real-valued	SIFT [125] (128 dim)	28.09	36.27	29.15	36.27	28.09	29.15	31.17
Binary (Supervised)	D-BRIEF [170] (4 bytes)	43.96	53.39	46.22	51.30	43.10	47.29	47.54
	BinBoost [171] (8 bytes)	14.54	21.67	18.96	20.49	16.90	22.88	19.24
Binary (Unsupervised)	BRIEF [14] (32 bytes)	54.57	59.15	54.96	59.15	54.57	54.96	56.23
	BRISK [100] (64 bytes)	74.88	79.36	73.21	79.36	74.88	73.21	75.81
	ORB [153] (32 bytes)	48.03	56.26	54.13	56.26	48.03	54.13	52.81
	DeepBit [112] (32 bytes)	28.49	34.64	54.63	33.83	20.66	56.69	38.15
	DBD-MQ [47] (32 bytes)	20.13	25.77	50.99	22.92	18.95	50.36	31.52
	BinGAN [262] (32 bytes)	16.88	26.08	40.80	25.76	27.84	47.64	30.76
	GraphBit [46] (32 bytes)	17.78	24.72	49.94	21.18	15.25	49.64	29.75
	TBLD (32 bytes)	**16.53**	**21.95**	**35.09**	**20.45**	**14.47**	**36.88**	**18.25**



(a) Liberty-Notre Dame

(b) Liberty-Yosemite

(c) Notre Dame-Liberty

(d) Notre Dame-Yosemite

(e) Yosemite-Liberty

(f) Yosemite-Notre Dame

Figure 4.4: ROC curves of the proposed TBLD and state-of-the-art methods on **Brown** [12], with all train-test combinations among three subsets.

works, including supervised descriptors (*e.g.,* D-BRIEF [170], and BinBoost [171]) and unsupervised ones (*e.g.,* BRISK [100], BRIEF [14], and GraphBit [46], *etc.*). The comparison results are reported in Table 4.2, where the results of real-valued descriptor SIFT [125] and supervised descriptors are also provided as references.

As can be seen, TBLD outperforms the state-of-the-art unsupervised binary descriptors, including both hand-crafted and deep learning based ones, on all the subsets. A decline of 11% can be found in terms of 95% error rate in contrast to the best unsupervised binary local descriptor learning method so far (GraphBit [46]). It is worth mentioning that when compared with a widely-used floating-point descriptor, *i.e.,* SIFT [125], the proposed TBLD obtains a lower 95% error rate, along with a much lower computation cost for measuring similarities.

Moreover, Receiver Operating Characteristic (ROC) curves of the state-of-the-art unsupervised binary local descriptors are plotted in Fig. 4.4. The curves illustrate the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. For a fair comparison, the author first reproduces the algorithms and then plot the curves. In terms of deep learning based binary local descriptors, the competitors include DeepBit [112] and GraphBit [46] because only their source codes are provided and GraphBit [46] still maintains the best performance until now. As can be seen, the ROC curves of TBLD rank at the top on all train-test configurations.

Results on HPatches dataset. The HPatches dataset is used to evaluate the performance of binary local descriptors on three visual tasks: patch matching, patch retrieval, and patch verification. Specifically, descriptors are compared in the match-

Table 4.3: Comparisons of mAP (%) with the state-of-the-art binary local descriptors on **HPatches** [7] (*Higher is better*).

	Method	Matching	Retrieval	Verification
Real-valued	SIFT [125] (128 dim)	25.47	31.98	65.12
Binary (Supervised)	BinBoost [171] (32 bytes)	16.97	38.68	76.27
Binary (Unsupervised)	BRIEF [14] (32 bytes)	10.50	16.03	58.07
	BRISK [100] (64 bytes)	15.97	18.10	65.65
	ORB [153] (32 bytes)	15.32	18.85	60.15
	DeepBit [112] (32 bytes)	13.05	20.61	61.27
	GraphBit [46] (32 bytes)	14.22	25.19	65.19
	TBLD (32 bytes)	**15.39**	**27.03**	**68.25**

ing task to find matched patches between the reference image and the target ones. For the patch retrieval task, local descriptors are employed to match a query patch to a pool of patches extracted from many images. In terms of patch verification, descriptors are utilized to classify whether two patches are matched or not.

Following the evaluation metrics suggested by [7], the comparison is conducted between TBLD and the state-of-the-art descriptors in terms of mean average precision (mAP). The comparison results are reported in Table 4.3. *Higher mAP means better performance.* Again, the binary local descriptors are categorized as supervised and unsupervised ones according to the training manner. Since DBD-MQ [47] and BinGAN [262] did not report the results on the HPatches dataset, they are not included in Table 4.3.

It can be seen that TBLD beats all unsupervised baselines, including both hand-crafted ones (BRISK [100], BRIEF[14], ORB [153]) and deep learning based ones (DeepBit [112], GraphBit [46]) on all the tasks. Specifically, compared to GraphBit [46], TBLD improves the mAP score by 8.2%, 6.8%, 4.5%, respectively, in the three tasks. Here the result of the real-valued SIFT [125] is also presented for reference. It can be observed that the proposed method even outperforms SIFT on the patch verification task with a 4.58% increase in terms of mAP.

Results on Mikolajczyk dataset. Experiments are conducted on “Boat” and “Graffiti” subsets of *Mikolajczyk* dataset [132] to prove the generalization of the binary local descriptors. Here, the proposed TBLD is compared with both hand-crafted binary local descriptors (BRISK [100], ORB [153]), and a state-of-the-art deep binary descriptor (GraphBit [46]). Considering the fairness, the binary local descriptors are set as 32 bytes for all the methods. Specifically, the author firstly reproduces BRISK, ORB as well as GraphBit, and then evaluate the performance in terms of *Recognition rate* following [14, 125], which can be obtained as follows,

- Extracting n_1 interest points from the reference image, and n_2 from the target image. Among them, n matching pairs are obtained from the ground-truth homograph transformation matrix.

- For each interest point in the reference point set, finding the nearest neighbour in the target point set via binary local descriptors.
- Counting the number of correct matches n_c, and calculating the recognition rate with $r = n_c/n$.

Following the previous works [14, 171], for an image, interest points are firstly detected by the SURF Hessian-based detector and patches are then cropped and normalized to the required size of each descriptor. Note that, the sizes of patches keep unchanged for BRISK [100] and ORB [153], while the patches are resized to 224×224 for feature extraction for GraphBit and the proposed method.

The recognition rates of the state-of-the-art binary local descriptors on *Boat* and *Graffiti* scenes with the challenges of zoom/rotation and viewpoint variations are shown in Fig. 4.5. As can be seen, TBLD outperforms other state-of-the-art binary local descriptors on all the reference-target configurations. Additionally, it can be found that, compared to the *Boat* scene, the proposed method performs better on the *Graffiti* scene, where a significant increase in the recognition rate can be seen for all configurations. A potential reason could be that the gap between training scenes (*Notre Dame*) and *Graffiti* is relatively smaller than *Boat*. Experimental results on *Mikolajczyk* further verify the generalization ability of the proposed binary descriptors.

Comparisons with unified metrics To show the superiority of the proposed method intuitively, the evaluation on the patch matching task is conducted on different datasets with an unified metrics: mAP. Since such comparison has not been conducted in previous works, the author firstly reproduces some representative baselines, including hand-crafted ones (BRIEF and ORB) and deep learning based one (GraphBit). Then, evaluate their performances on three benchmark datasets for fair comparisons. Due to time constraints, only state-of-the-art binary descriptor learning methods are chosen.

The mAP scores are illustrated in Table 4.4. The model used for evaluation is trained on the Liberty subset of the Brown dataset. Note that, the mAP scores of the Mikolajczyk dataset are obtained by the patch matching between the target image (img1) and the reference image with the mildest distortion (img2). As can be seen, the proposed TBLD still outperforms the state-of-the-art approaches on the three datasets when a unified metric is employed.

4.7.4 Ablation study

Two comparison experiments are designed to prove the effectiveness of the proposed *Adversarial Constraint Module* (ACM). 1) The comparison with **evenly-distributive constraint**, which is employed to reduce the bit correlations in existing works, such as DH [49] and DeepBit [112]. 2) The comparison between binary descriptors generated by models trained with (w/) and without (w/o) ACM in terms

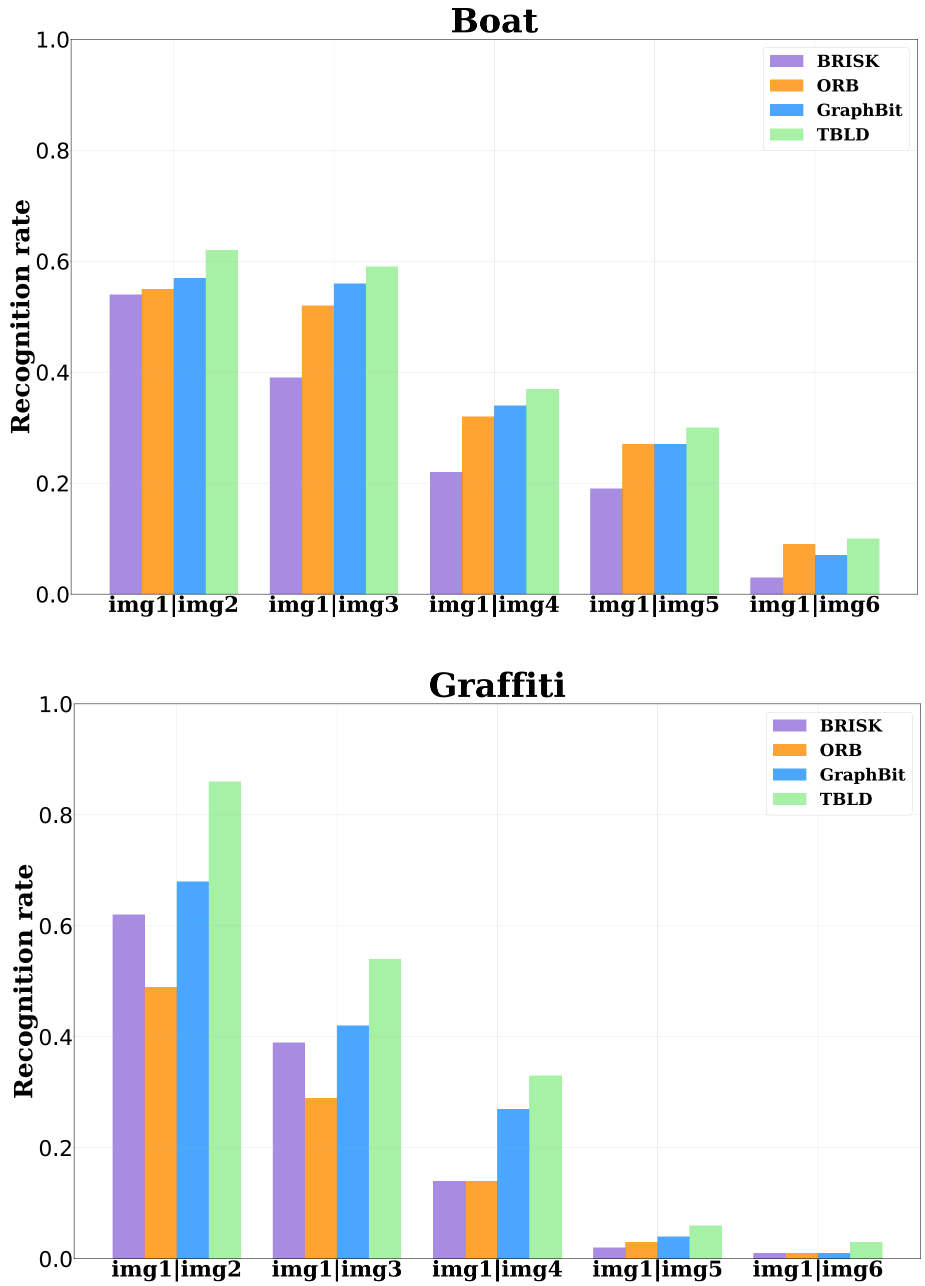


Figure 4.5: Recognition rate on of **Mikolajczyk** [132], including Boat and Graffiti subsets. (*Higher is better*). TBLD outperforms other state-of-the-art binary local descriptors learning approaches in terms of recognition rate on all reference-target configurations.

Table 4.4: Comparison of mAP (%) with the state-of-the-art binary descriptors on the three datasets (*Higher is better*).

Method	Brown (Liberty)		Hpatches	Mikolajczyk	
	Notre Dame	Yosemite		Boat	Graffiti
BRIEF	62.05	66.40	16.03	43.24	34.34
ORB	64.19	68.63	18.85	51.11	44.83
GraphBit	68.78	72.27	25.19	59.19	57.39
TBLD	**69.52**	**74.39**	**27.03**	**62.41**	**60.07**

Table 4.5: Comparison of 95% error rate (%) with the model trained with the Evenly-Distributive Constraint (EDC) and the proposed ACM on Brown (*Lower is better*).

Train	*Yosemite*		*Notre Dame*		*Liberty*	
Test	*Notre Dame*	*Liberty*	*Yosemite*	*Liberty*	*Notre Dame*	*Yosemite*
EDC	18.16	24.10	36.57	21.93	15.72	37.67
ACM	**16.53**	**21.95**	**35.09**	**20.45**	**14.47**	**36.88**

of bit correlations. Since HPatches [7] and Mikolajczyk [132] are designed only for evaluation, the comparisons are only conducted on the Brown dataset [12].

Evenly-distributive constraint. To compare with the evenly-distributive constraint, ACM in the proposed model is replaced by the evenly-distributive constraint, which is fomulated as follows,

$$L_M = \sum_{k=1}^{K} ||\mu_k - 0,5||^2, \qquad \mu_k = \frac{1}{N}\sum_{n=1}^{N} b_{nk}, \tag{4.7.1}$$

where K is the length of binary descriptors and μ_k denotes the mean value of each bit over N samples within a mini-batch. Therefore, the overall objective for the to-be-compared models can be derived as,

$$L = L_{C_r} + LC_b + \beta L_Q + \lambda L_N + L_M. \tag{4.7.2}$$

The performances of the derived binary descriptors are evaluated in terms of 95% error rates, which are reported in Table 4.5. It can be observed that binary descriptors derived by the proposed method outperform those derived by the model trained with the evenly-distributive constraint with lower 95% error rates.

Bit correlation reduction. To investigate the reduction of bit correlations brought by the proposed Adversarial Constraint Module (ACM), the author compares the average correlations between bits of binary descriptors derived by the models trained

Table 4.6: Comparison of the mean Absolute Correlations (%) of binary local descriptors derived by models trained without (w/o) and with (w/)the proposedAdversarial Constraint Module (ACM) (*Lower is better*). Δ refers to the mAC reduction.

	w/o	**w/**	Δ
Yosemite	14.25	13.73	-0.52
Notre Dame	9.64	7.43	-2.21
Liberty	10.26	9.64	-0.62

without (w/o) and with (w/) ACM, respectively. Specifically, the proposed network is trained on three subsets of the Brown dataset separately with the same experimental settings, except that ACM is removed from each model.

Specifically, the average bit correlations can be evaluated by the metric - mean Absolute Correlations (mAC). Specifically, given N to-be-evaluated image patches with corresponding k-bit binary descriptors $B = \{b_1, ...b_n\}$, the mAC score is calculated as follows,

$$mAC = \frac{1}{k(k-1)} \sum_{i,j \neq i} |P_{ij}|, \tag{4.7.3}$$

$$P_{ij} = \frac{\sum_{n=1}^{N}(b_{in} - \bar{b_i})(b_{jn} - \bar{b_j})}{\sqrt{\sum_{n=1}^{N}(b_{in} - \bar{b_i})^2}\sqrt{\sum_{n=1}^{N}(b_{jn} - \bar{b_j})^2}}, \tag{4.7.4}$$

where P_{ij} presents the *Pearson correlation coefficient* between the i-th bit and the j-th bit. Specifically, b_{in} denotes the i-th bit of the binary descriptor of the n-th image patch. $\bar{b_i}$ and $\bar{b_j}$ are the mean values of the i-th bit and the j-th bit over N image patches. According to the definition of mAC, it can be inferred that *a lower mAC score means lower bit correlations.* The mAC scores of binary local descriptors derived by models without (w/o) or with (w) ACM are reported in Table 4.6. As can be seen, the correlation between bits reduces by 0.52, 2.21, 0.62, respectively, in terms of mAC score with ACM, which proves its effectiveness on bits correlation reduction.

4.7.5 Descriptor robustness

Transformation Invariance. Firstly, the transformation invariance of binary local descriptors derived by TBLD is investigated. Since the to-be-evaluated models are trained on the Brown dataset, the analysis is conducted on the other two datasets for fair comparison. For Mikolajczyk, as discussed above, both *Boat* and *Graffiti* subsets contain a certain type of transformations, which means the transformation invariance of the derived binary local descriptors has been proved by the results in Fig. 4.5. Therefore, here the ablation study is conducted on the HPatches dataset [7] to evaluate the robustness of the derived binary local descriptors against geometric noises and viewpoints.

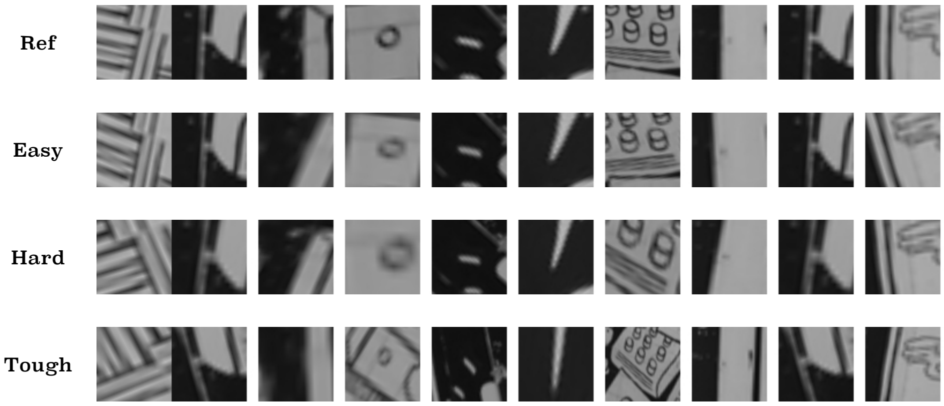


Figure 4.6: Visualization of patches from Hpatches [7]. Patches from the reference image (REF) are shown in the first row. Patches from target images with the increasing level of geometric noises: EASY, HARD and TOUGH, are shown in rows 2 to 4, respectively.

Geometric noise. Specifically, image patches in the HPatches dataset have been divided into different subsets according to the level of geometric noises, which are indicated by EASY, HARD and TOUGH. Examples of the reference and the target image patches in each subset are shown in Fig. 4.6. On each subset, the proposed TBLD is compared with other transformation-invariant binary local descriptors: BRISK [100], ORB [153] and DeepBit [112] in terms of mAP, following the settings in [7]. The results are illustrated in Fig. 4.7. As can be seen, the proposed TBLD achieves a higher mAP on all the subsets, which proves the robustness of the proposed method against multiple levels of geometric noises.

Viewpoints. For the task of patch matching, HPatches further groups the data by different levels of geometric noises into "ILLUM" and "VIEW" subsets, enabling the evaluation of the robustness of binary local descriptors against illumination and viewpoints changes. Since the proposed TBLD focuses on improving the robustness of descriptors against transformations like rotation and scaling, the comparison is specifically conducted on the "VIEW" subset. The mAP of the state-of-the-art unsupervised transformation-invariant binary local descriptors on the "VIEW" subsets are illustrated in Fig. 4.8.

As can be seen, although the proposed method underperforms BRISK [100] on the overall mAP of the patch matching task (15.39% and 15.97%, respectively), it outperforms BRISK on the three "VIEW" subsets from EASY, HARD, and TOUGH, respectively. The results prove the robustness of the binary local descriptors derived by TBLD against viewpoints changes.

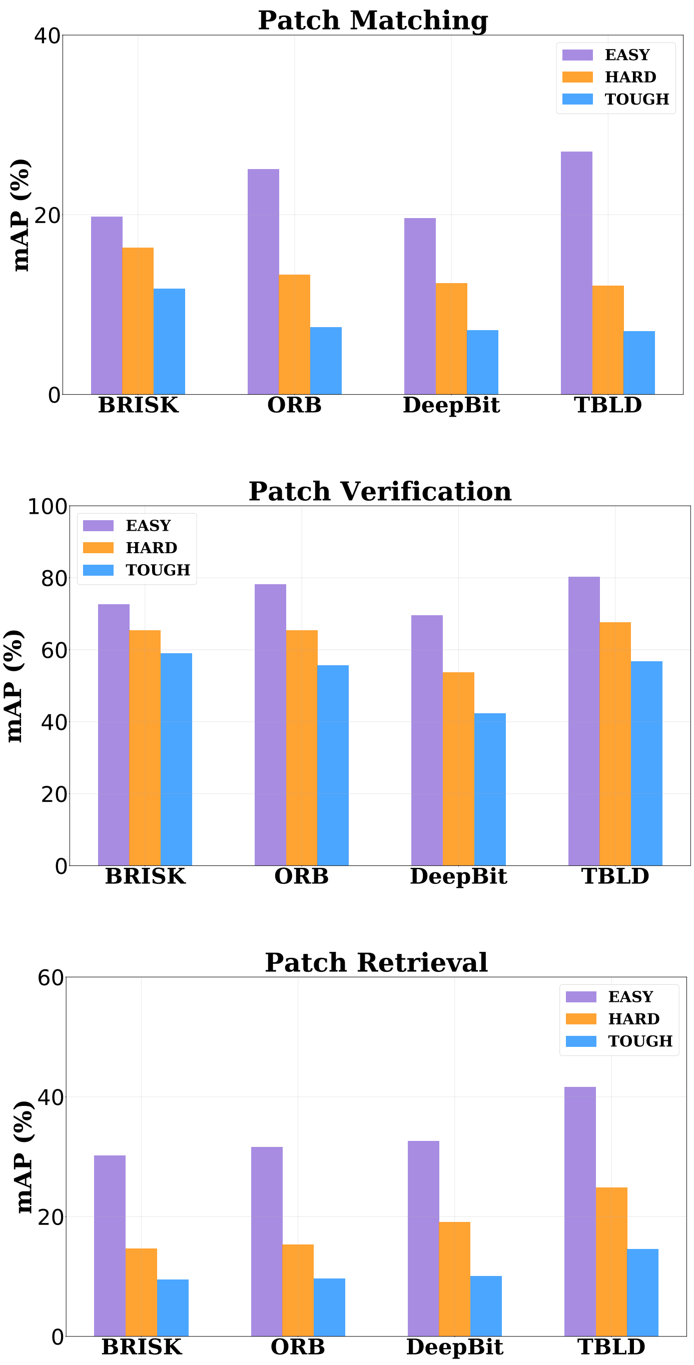


Figure 4.7: Comparisons of state-of-the-art transformation-invariant binary local descriptors on three tasks of HPatches in terms of mAP(%) (*Higher is better*). For each task, data are divided into three subsets according to the level of geometrical noises, which are indicated by EASY, HARD, and TOUGH, respectively.

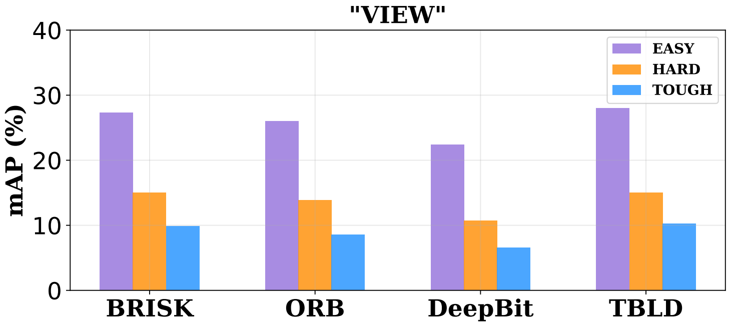


Figure 4.8: Comparison with the state-of-the-art transformation-invariant binary local descriptors on the “VIEW” subset of the patch matching task in terms of mAP(%) (*Higher is better*).

4.7.6 Out-of-sample experiment on person ReID datasets

To explore the potential of the proposed unsupervised binary descriptor learning framework in person ReID, comparison experiments are conducted on person ReID benchmark datasets - Market-1501 [246] and MSMT17 [190]. Specifically, following the proposed training scheme, given a person image, positive pairs and negative pairs are defined as its augmented images and images from clusters except that query belongs to. The inference settings follow that of the unsupervised person ReID work, which are aforementioned in Section 3.7.2. Metrics adopted to evaluate performances are top-1 accuracy and mAP.

The comparison is conducted among a traditional unsupervised binary descriptors (ITQ [63]), the proposed method (Ours), and two state-of-the-art (SOTA) supervised fast person ReID work (DLBC [23] and SIAMH [238]). As discussed in Section 2.3, iterative quantization (ITQ) derives the similarity-preserving binary descriptors for large-scale image retrieval. Deep local binary coding (DLBC) focuses on subtle details of person ReID, where local features are firstly extracted from discriminative local regions and are then projected to robust binary codes by a hash layer. Similarly, salience-guided iterative asymmetric mutual hashing (SIAMH) learns compact binary descriptors from salient regions. Additionally, SIAMH handles the drawbacks of mutual learning via an iterative asymmetric learning strategy.

The performances on Market-1501 and MSMT17 are illustrated in Table 4.7 and Table 4.8, respectively. As can be seen, for all methods, as the feature length increases from 64 bits to 512 bits, the identity matching accuracy gradually increases due to the fact that more id-related information is embedded in extended bits. However, the identity matching accuracy varies dramatically from method to method. Taking the performances on Market-1501 as an example, the traditional binary descriptor learning method - ITQ can barely achieve 20% top-1 accuracy even with 512 bits. With training on person images, the proposed data-driven unsupervised binary descriptor learning method significantly outperforms ITQ, where the top-1 accuracy and mAP score achieve 51.6% and 38.4%, respectively. Note that, the backbone network (VGG-19) is replaced by ResNet50 following DLBC [23]

Table 4.7: Comparisons of top-1 accuracy (%) / mAP (%) among ITQ [63] (unsupervised), the proposed method (unsupervised), DLBC [23] (supervised), and SIAMH [238] (supervised) on **Market-1501** (*Higher is better*).

Methods	64 bits	128 bits	512 bits
ITQ [63]	9.7 / 3.0	14.1 / 4.4	20.3 / 7.8
Ours	40.3 / 25.1	48.9 / 32.7	51.6 / 38.4
DLBC [23]	82.7 / 62.5	89.7 / 72.8	92.1 / 81.2
SIAMH [238]	83.0 / 65.6	90.6 / 78.4	94.8 / 86.7

Table 4.8: Comparisons of top-1 accuracy (%) / mAP (%) among ITQ [63] (unsupervised), the proposed work (unsupervised), and SIAMH [238] (supervised) on **MSMT17** (*Higher is better*).

Methods	64 bits	128 bits	512 bits
ITQ [63]	1.3 / 0.6	1.6 / 0.9	2.0 / 1.3
Ours	24.1 / 13.2	31.2 / 18.5	36.2 / 24.1
SIAMH [238]	54.2 / 29.8	69.2 / 42.5	78.7 / 54.5

and SIAMH [238] for a fair comparison.

However, due to difficulties caused by the absence of identity labels as well as the less informative binary bits, the performance gap between the proposed unsupervised binary descriptors and SOTA supervised ones, which leverage identity labels during binary code learning, is quite large. Specifically, with the same binary code length (512 bits), DLBC achieves the top-1 accuracy 92.1%, and the improved work SIAMH achieves 94.8% on Market-1501. Similarly, the large performance gap can also be found in MSMT17, which highlights the potential of improving fast unsupervised person ReID. Considering the large performance gap, how to improve the performance of fast unsupervised person ReID methods will be a main focus of the future work.

4.8 Conclusion

In this section, an unsupervised transformation-invariant binary local descriptor learning method (TBLD) is proposed, which can be adapted to fast unsupervised person ReID. Two major contributions are made by the proposed work: 1) A framework that derives transformation-invariant binary local descriptors is proposed. Based on the assumption that the same object should be described by the same local descriptors, the original image patches and their transformed counterparts are projected to an identical Euclidean subspace and an identical Hamming subspace with the help of contrastive loss. 2) In order to solve the problem brought by directly applying the scheme of image hashing in local descriptor learning, which refers to the high correlations between bits, an *Adversarial Constraint*

Module (ACM) is proposed. A set of low-coupling binary codes are introduced to guide the learning of binary local descriptors. By means of Wasserstein loss, the framework is optimized to transfer the distribution of the learned binary local descriptors to the low-coupling ones, thereby making the learned ones as low-coupling as possible. Experimental results on three benchmark datasets well demonstrate the superiority of the proposed approach over the state-of-the-art unsupervised binary descriptors. Additionally, evaluation is conducted on person ReID benchmark datasets to validate the effectiveness of the proposed unsupervised binary descriptor learning scheme to fast unsupervised person ReID. Although not comparable to supervised methods that leverage identity labels, the proposed method significantly outperforms existing unsupervised binary descriptors on unsupervised person ReID.

Chapter 5

Cross-modality person re-identification

5.1 Introduction

Person re-identification (ReID) aims at retrieving the same identity across multiple disjoint cameras, which has gained much attention from the recent computer vision community [99, 220]. Most existing ReID methods focus on the matching between visible images, which are generally collected under good illumination conditions [103, 24, 71, 251, 181, 150, 81]. However, those systems seem impractical because visible images cannot provide sufficient discriminatory information in poor lighting environments, *e.g.,* at night. To this end, Visible-Infrared person Re-identification (VI-ReID) emerges as an alternative to performing the retrieval between visible (RGB) images and infrared (IR) counterparts, thus enabling the day-to-night person re-identification.

However, VI-ReID is a challenging problem due to large intra-class variations and modality discrepancies across different cameras. The former refers to identity's appearance differences within a modality caused by poses, clothes, viewpoints, *etc.* While, the latter denotes intrinsic differences between visible and infrared images caused by the spectrum of cameras. To reduce both discrepancies, the central research question in this field has always been seeking better ways to extract discriminative features for identity retrieval, which are modality-invariant and ID-related [194, 220, 261, 217].

Despite the vigorous development of this field, an observation can be found that most algorithms extract identity features in a heuristic manner with NO common view of what sort of features are specifically helpful for VI-ReID. To tackle this problem, the author visualizes the features extracted by several representative methods in Fig. 5.1, aiming to investigate how visual feature extractions evolved over the years to improve the performance of VI-ReID systems. Concretely, ZeroPad [194], as the first work in VI-ReID, extracts features from random regions in an image. Additionally, for a specific identity, the features derived from the two modalities share NO commonality. As a result, the mean Average Precision (mAP) score is

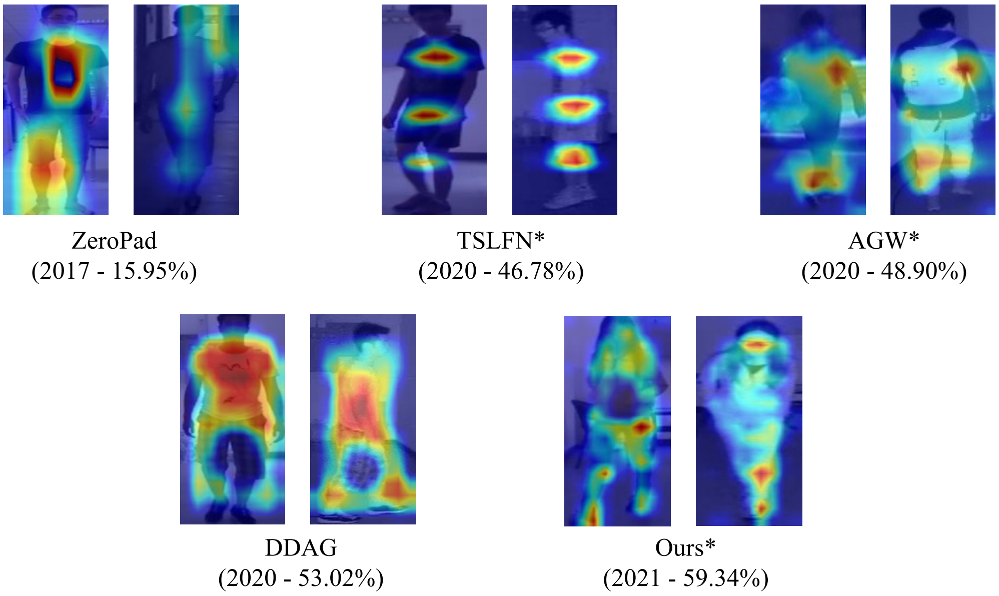


Figure 5.1: Visualization of features derived by ZeroPad [194], TSLFN [261], AGW [220], DDAG [217], and the proposed method on SYSU-MM01 [194]. For each method, the RGB image of identity is shown on the left and the IR image is on the right. The publication year and mAP score (%) at *All-search* setting of each method are also reported. For a fair comparison, the mAP scores of TSLFN, AGW and the proposed method are given by models trained merely with plain identity loss and triplet loss.

far from satisfactory. Later, TSLFN [261] horizontally partitions the backbone feature maps (global) into several stripes (local) and employs local-level constraints on each of them. Clearly, features derived by TSLFN cover more parts of the human body, compared to the global-level constraint-based method, *e.g.,* ZeroPad. That might be the reason why its performance is significantly superior to that of ZeroPad. In the meantime, a baseline for VI-ReID (AGW [220]) inserts non-local attention blocks during the feature extraction, which enforces the features to be extracted from identity's body instead of backgrounds. It reveals that *attention-aware* features help to increase performance. Recently, a state-of-the-art approach (DDAG [217]) integrates both local-level and global-level constraints into an end-to-end framework. Compared to previous works, the features derived by DDAG are more fine-grained, which are not only shared by two modalities but distinguishable for different identities. Based on the above observations, a conclusion can be drawn: *as more features from the human region (modality-shared) and attentive body part (ID-related) features are extracted, the retrieval performance improves consistently.* Benefiting from this conclusion, the proposed method extracts more features from body skeleton joints, which are not only ID-related but immune to modality changes. Therefore, this work achieves over 10% mAP improvements against DDAG on the challenging SYSU-MM01 dataset at *All-search* setting.

Having depicted the visual features that are conducive to VI-ReID, the next question is: how to extract them effectively? It is noted that body skeleton points are explicit modality-shared cues and the features describing certain skeleton points are ID-related. In light of this, in the work, the author aims to facilitate the extraction of discriminative features for identity retrieval with the aid of the pose estimation task. However, making effective use of the pose information for cross-modality ReID does not seem easy, though it has recently appeared to be exploited in some single-modality ReID works [239, 164, 248, 240, 165, 131], where only visible images are involved. Earlier methods [239, 164, 248] utilize detected body joints to segment [239] or align [164, 248] body regions in order to cope with human pose changes. After the calibration of body parts, different body parts need to be stitched, which usually yields unrealistic transformed visual features. If moving to the cross-modality setting, the transformed errors would be further magnified due to the huge discrepancy between the two modalities. Alternatively, another group of methods employ the body keypoints information to refine ID-related feature maps either by means of highlighting discriminative body regions [240, 131] or complementing human appearance features [165]. Although pose-assisted features are proven to improve feature discriminability under the single-modality setting, they are rigidly based on the outputs of off-the-shelf pose estimators. However, such blind trust in the pre-trained pose estimators will lead to poor re-identification performance if the gap between the source domain and the target domain of the pose estimator is huge. Therefore, employing pose information in the VI-ReID task is extremely challenging due to the massive gap between visible images (source domain) and infrared images (target domain). To highlight the problem, two state-of-the-art pose-guided single-modality person ReID methods [165, 131] are adapted to VI-ReID and the results turn out that the best performance [165] in the mAP score only reaches 42.19%, which is far from satisfaction due to the inadequate usage of the pose information (more results and comparisons are provided in Section 5.7.6).

To solve the above problems, a two-stream VI-ReID framework is proposed, where modality-shared and ID-related features for identity retrieval are extracted by means of learning an auxiliary task (pose estimation) and the main task (person ReID) jointly. Unlike previous works, which rely dramatically on off-the-shelf pose estimators, pose features are adaptively adjusted to facilitate the ReID task in the proposed work. Additionally, apart from ID-related constraints, an extra constraint is imposed on the pose estimation branch, ensuring that not only body skeleton points are precisely estimated but also the ID-related information is fully embedded in feature maps, *i.e.,* at both local and global levels. Despite the significant improvements obtained by the horizontal-divided feature constraints [166] in the VI-ReID task [261, 217], the learning of individual striped features is generally independent and its discriminability consistency with global backbone features is neglected. To this end, a Hierarchical Feature Constraint (HFC) is proposed, in the paper, which bonds the learnings of global features and local ones via the knowledge distillation strategy. Concretely, predictions of backbone features serve as "soft-target" to supervise the learning of partitioned feature stripes, hence preserving the discriminability consistency of global features and local ones. In summary, the contributions made in the proposed work are mainly three-fold:

- A novel two-stream framework for VI-ReID is proposed, where the pose estimation acts as an auxiliary learning task to assist the identity feature learning in VI-ReID. To learn fine-grained pose features embedded ID-related information, both pose and ReID constraints are imposed on the pose estimation branch.
- Instead of imposing feature constraints on local feature stripes only, Hierarchical Feature Constraint (HFC) is proposed to ensure the discriminability consistency of global features and local ones via the knowledge distillation strategy.
- The proposed method performs far better than the state-of-the-art methods on two benchmark datasets: SYSU-MM01 [194] and RegDB [136].

5.2 Previous solutions

5.2.1 Pose-guided person ReID

As discussed in Section 2.1.1, different kinds of auxiliary information have been employed to facilitate identity feature learning. However, most of them, such as attributes and human segmentations, require extensive additional annotations, which are expensive to obtain in real-world scenarios. Additionally, since some color-related cues are unavailable for infrared images, the usage of color-related attributes is restricted in VI-ReID, thereby limiting performance improvement. Therefore, facilitating VI-ReID with body keypoint information seems to be feasible.

Considering existing pose-guided person ReID approaches aforementioned in Section 2.1.1, they are rigidly based on the outputs of off-the-shelf pose estimators, which may provide unreliable information due to the gap between the source domain and the target domain. To handle the problem, PABR [165] takes an on-the-fly pose estimator as an individual branch to derive pose features, which are then aggregated with the appearance features from the other branch via a bilinear pooling layer. Such a joint training scheme encourages the pose branch to learn features that are beneficial to the person ReID task. However, only ID-related constraints are imposed on the fused features in this work, which ignores the quality of pose features from the individual branch.

Given the above concerns, the auxiliary pose estimation task designed for single-modality ReID cannot perform well in the VI-ReID task. Recently, Chen *et al.* [19] exploit structure information to reduce the background noise and misalignments in VI-ReID. Specifically, key point heatmaps output by an off-the-shelf pose estimator serve as the inputs of a structure representation module. Then, the structure information is used to refine appearance features. Although body key points features are considered, the method is also rigidly based on the outputs of the pose estimators.

As a departure from existing pose-guided VI-ReID works, the pose estimation serves as an auxiliary learning task in the proposed work, where body skeleton points cues are learned under both pose and identity guidance, to enhance the discriminative identity feature extraction for the cross-modality person re-identification. The comparison results in Section 5.7.3 demonstrate that using pose estimation is more sophisticated, thus enabling to improve the performance of identification significantly.

5.2.2 Feature constraints in VI-ReID

To improve the discriminability of learned features, most existing works impose either global-level feature constraints on backbone convolutional features or local-level feature constraints on partitioned feature stripes.

Global-level feature constraints. Ye *et al*. [212] propose a two-stream network, which jointly optimizes modality-specific and modality-shared metrics. Based on the idea, a bi-directional top-ranking loss is then introduced in [215] to incorporate the above two constraints. Alternatively, AGW [220] presents a weighted regularized triplet loss to embed the neighboring relationship of images from two modalities in a common feature space. Considering the large gap between visible modality and infrared modality, HC loss [261] is proposed to pull the centers of RGB features and IR features of a given identity closer. Then Liu *et al*. [116] extends HC loss into the triplet version, where apart from pulling modality centers of positive image pairs, feature centers of different identities are pushed away at the same time.

Part-level feature constraints. Inspired by the competitive performance of the part-based convolutional baseline (PCB) [166] in single-modality person ReID, recent VI-ReID studies start imposing local-level feature constraints on feature stripes obtained by partitioning backbone convolutional features. For example, DDAG [217] simultaneously mines both cross-modality global-level and intra-modality part-level contextual cues, which achieves state-of-the-art performance. Although both global-level and local-level constraints are considered in DDAG, those constraints are imposed independently without taking their discriminability consistency of them into account. To this end, a Hierarchical Feature Constraint (HFC) is proposed to bond the global feature learning with the local ones, where predictions of global features supervise the learning of part features via the knowledge distillation strategy.

5.3 Problem statement

As can be seen in Fig. 2.23(b), VI-ReID conducts the identity matching across the visible modality and the infrared modality, where the query image and gallery images are from different modalities. Formally, taking visible-infrared matching as an example, given the query q and N gallery images $\mathcal{G} = \{g_i\}_{i=1}^{N}$, the VI-ReID model maps the images to identity representations f_q and $\{f_{g_i}\}_{i=1}^{N}$. Then, a query-gallery distance matrix is computed to rank the gallery images according to their similarity

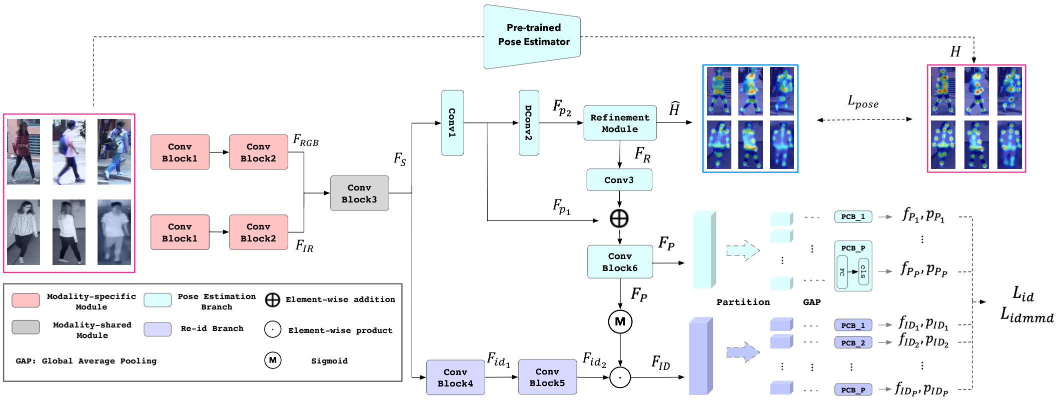


Figure 5.2: Framework of the proposed method, which contains four key components: Modality-specific module, Modality-shared module, Pose Estimation branch, and ReID branch. The Modality-specific module and Modality-shared module learn the modality-specific and modality-shared backbone features, respectively. The Pose Estimation branch (Section 5.4.2) extracts the keypoint-aware features to refine the identity features derived by the ReID branch (Section 5.4.3).

with the query as follows,

$$i^* = \underset{i=1,2,\ldots,N}{argmin}\ Dist(q, g_i), \tag{5.3.1}$$

where i^* refers to the top-1 candidate, which ideally should belong to the same identity as a query. Therefore, as the core of VI-ReID, a pose estimation assisted VI-ReID framework is proposed to learn representative identity representations.

5.4 Pose estimation assisted VI-ReID framework

The pose estimation assisted framework aims to learn modality-shared and ID-related features for the cross-modality identity retrieval, which is illustrated in Fig. 5.2. As can be seen, the framework mainly consists of four components: Modality-specific module, Modality-shared module, Pose Estimation branch, and ReID branch. Details about these components will be discussed in the following content.

5.4.1 Modality-specific module and Modality-shared module

Following previous works [215, 116, 261], ResNet50 [77] is exploited as a backbone feature extractor to provide discriminative features for both pose estimation and ReID tasks. Specifically, Modality-specific module consists of two blocks (*"Conv Block1 ˜2"*), which adopt the structures of shallow convolution block (*layer0*) and the first res-convolution block (*layer1*) of ResNet50, respectively. Note that the parameters of Modality-specific module for RGB and IR modalities are separately updated. Then, modality-specific features of two modalities are projected into a shared feature space by the Modality-shared module (*"Conv Block3"*), which adopts the structure of the second res-convolution block (*layer2*) of ResNet50.

Mathematically, given a RGB image I_{RGB} and an IR image I_{IR}, modality-specific features $\mathbf{F}_m$ and modality-shared features $\mathbf{F}_S$ can be obtained by,

$$\begin{aligned}&\mathbf{F}_m = ConvB2(ConvB1(I_m, \theta_m^1), \theta_m^2), \quad m \in \{RGB, IR\},\\&\mathbf{F}_S = ConvB3([\mathbf{F}_{RGB}, \mathbf{F}_{IR}]_b, \theta^3),\end{aligned} \tag{5.4.1}$$

where $[\cdot,\cdot]_b$ represents the feature concatenation along the data dimension. Modules $ConvB1(*, \theta_m^1)$, $ConvB2(*, \theta_m^2)$, and $ConvB3(*, \theta^3)$ denote the convolution blocks of Modality-specific module and Modality-shared module with corresponding parameters θ_m^1, θ_m^2 and θ^3, respectively.

5.4.2 Pose Estimation branch

Body keypoint features extraction. Given the observation that modality-shared features are beneficial to the VI-ReID task, a Pose Estimation branch is integrated as an auxiliary to extract modality-shared features. The structure of the proposed Pose Estimation branch is shown in Fig. 5.2. Specifically, a convolutional layer (*"Conv1"*) and a deconvolutional layer (*"DConv2"*) are employed to extract high-level features, *i.e.,* $\mathbf{F}_{p_1}$ and $\mathbf{F}_{p_2}$, and restore the resolution of feature maps to that of ground-truth body keypoint heatmaps, which are denoted as follows,

$$\mathbf{F}_{p_1} = Conv1(\mathbf{F}_S, \theta_P^1), \quad \mathbf{F}_{p_2} = DConv2(\mathbf{F}_{p_1}, \theta_P^2), \tag{5.4.2}$$

where $Conv1(*, \theta_P^1)$ and $DConv2(*, \theta_P^2)$ represents a 3×3 convolutional layer with parameters θ_P^1 and θ_P^2, respectively. Note that both layers are followed by a ReLU activation function, which are omitted in equations for simplicity.

Subsequently, a Refinement Module [137] is used to extract refined body keypoint features and predict body keypoint heatmaps. Specifically, the Refinement Module consists of a U-Shaped Block, three Refine Blocks and two convolutional layers. The U-Shaped Block and Refine Blocks are employed to extract refined features. On top of it, the convolutional layers are applied for the heatmap estimation. Mathematically, given high-level features $\mathbf{F}_{p_2}$, refined keypoint features $\mathbf{F}_R$ and predicted heatmaps $\hat{\mathbf{H}}$ can be respectively obtained by,

$$\mathbf{F}_R = RM_F\big(\mathbf{F}_{p_2}, \theta_{RM}^F\big), \tag{5.4.3}$$

$$\hat{\mathbf{H}} = RM_H(\mathbf{F}_R, \theta_{RM}^H), \tag{5.4.4}$$

where $RM_F(*, \theta_{RM}^F)$ and $RM_H(*, \theta_{RM}^H)$ denote the refined feature extraction stage with parameters θ_{RM}^F, and the heatmap estimation stage with parameters θ_{RM}^H in the Refinement Module, respectively.

Body keypoint features transferring. To exploit body keypoint features derived by Pose Estimation branch in the ReID branch, a convolutional layer (*"Conv3"*) and a convolutional block (*"Conv Block6"*) are employed to deal with the mismatch in terms of the resolution and the channel number of feature maps.

Specifically, the refined keypoint features $\mathbf{F}_R$ are firstly downsampled by *"Conv3"* so that the derived $\mathbf{F}'_R$ has the identical resolution as $\mathbf{F}_{p_1}$, which is denoted as,

$$\mathbf{F}_{R'} = Conv3(\mathbf{F}_R, \theta_P^3), \tag{5.4.5}$$

where $Conv3(*, \theta_P^3)$ represents the 3×3 convolutional layer with a stride of 2 and parameters θ_P^3. On top of that, *"Conv Block6"* aligns the channel number of feature maps from the Pose Estimation branch with that from the ReID branch. Therefore, the keypoint-aware features $\mathbf{F}_P$ can be derived as follows,

$$\mathbf{F}_P = ConvB6(\mathbf{F}_{R'} + \mathbf{F}_{p_1}, \theta_P^4), \tag{5.4.6}$$

where $ConvB6(*, \theta_P^4)$ denotes the convolutional block with the parameters θ_P^4, consisting of a 3×3 convolutional layer with a stride of 2 and a 1×1 convolutional layer. Note that all convolutional layers are followed by a ReLU activation function.

Body keypoint features integration. To highlight the body keypoint regions in the features output by the ReID branch, the keypoint-aware features $\mathbf{F}_P$ are employed to generate the body keypoint masks $\mathbf{M}$, *i.e.,*

$$\mathbf{M} = sigmoid(\mathbf{F}_P). \tag{5.4.7}$$

Keypoint-aware features are regularized by the sigmoid function to (0, 1), which serve as soft attention masks to refine the identity features from ReID branch.

5.4.3 ReID branch

Global-level feature extraction. Apart from the Pose Estimation branch that aids modality-shared features extraction, a ReID branch is employed to extract ID-related features. As can be seen from Fig. 5.2, the ReID branch mainly consists of 2 convolutional blocks (*"Conv Block4˜5"*). Following previous works [215, 220, 261], *"Conv Block4"*, *"Conv Block5"* follow the structures of the third and fourth res-convolution block (*layer3*, *layer4*) of ResNet50 [77], respectively. Mathematically, identity features $\mathbf{F}_{id_2}$ are extracted by,

$$\mathbf{F}_{id_2} = ConvB5(ConvB4(I_m, \theta_{ID}^1), \theta_{ID}^2), m \in \{RGB, IR\}, \tag{5.4.8}$$

where $ConvB4(*, \theta_{ID}^1)$ and $ConvB5(*, \theta_{ID}^2)$ denote a convolutional block with the parameters θ_{ID}^1 and θ_{ID}^2, respectively. Then, identity features $\mathbf{F}_{id_2}$ are refined by the body keypoint masks $\mathbf{M}$ derived by Pose Estimation branch by performing the element-wise product operation $\odot$ to obtain the final identity features $\mathbf{F}_{ID}$, *i.e.,*

$$\mathbf{F}_{ID} = \mathbf{F}_{id_2} \odot \mathbf{M}. \tag{5.4.9}$$

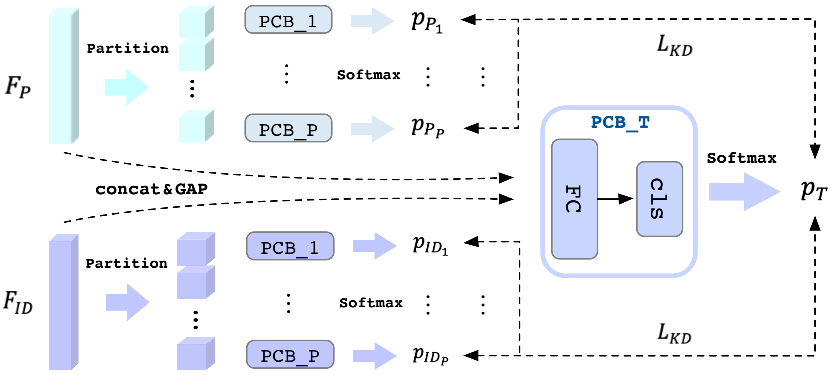


Figure 5.3: Illustration of Hierarchical Feature Constraint (HFC). GAP represents Global Average Pooling. The predictions of F_P and F_{ID} are employed as "soft-targets" to provide extra supervision for PCB models of the Pose Estimation branch and the ReID branch, respectively.

Local-level feature partition. Since part features can offer fine-grained information for identity identification, PCB models [166] are exploited in the proposed framework for local feature learning. Following [217, 261], convolutional features $\mathbf{F}_{ID}$ from ReID branch are firstly partitioned into P horizontal stripes and then transferred to feature vectors via Global Average Pooling (GAP) before being sent to the corresponding PCB model, which can be formulated as follow,

$$\mathbf{F}_{ID_1}, ..., \mathbf{F}_{ID_P} = GAP\big(Part(\mathbf{F}_{ID})\big), \tag{5.4.10}$$

where $Part(\cdot)$, $GAP(\cdot)$ denote the horizontal partition and GAP, respectively.

As can be seen from Fig. 5.2, a PCB model consists of a Fully-Connected (FC) layer and a classifier. The former reduces the dimensions of feature vectors from 2048 to 512, and the latter is employed for identity prediction. For the i-th PCB model, the fine-grained part features f_{ID_i} are obtained by,

$$f_{ID_i} = FC_{ID_i}(\mathbf{F}_{ID_i}, \theta^i_{fc_{id}}), \quad i = \{1, ..., P\}, \tag{5.4.11}$$

where $FC_{ID_i}(*, \theta^i_{fc_{id}})$ denotes the FC layer with the parameters $\theta^i_{fc_{id}}$. P is the number of PCB models, which is empirically set as 6 in the paper.

The local feature learning is also performed on convolutional features $\mathbf{F}_P$ from Pose Estimation branch to obtain the corresponding fine-grained part features f_{P_i}. During inference, fine-grained part features are concatenated for identity retrieval, *i.e.*,

$$f_{ID} = [f_{ID_1}, ..., f_{ID_P}]_c, \quad f_P = [f_{P_1}, ..., f_{P_P}]_c, \tag{5.4.12}$$

$$f_{ALL} = [f_{ID}, f_P]_c, \tag{5.4.13}$$

where $[\cdot, ..., \cdot]_c$ represents the feature concatenation along the channel dimension.

5.5 Hierarchical feature constraint

To ensure the discriminability consistency of global and local features, Hierarchical Feature Constraint (HFC) is proposed to bond the learnings of global features and local ones. The structure of HFC is illustrated in Fig. 5.3, which is inspired by the Teacher-Student learning spirit in Knowledge Distillation (KD) [84]. As can be seen, instead of introducing an additional pre-trained teacher model, the predictions of convolutional features are employed as "soft-targets" to provide an extra supervision for *"Student"* models, *i.e.,* PCB models of Pose Estimation branch and ReID branch.

Specifically, convolutional features of Pose Estimation branch and ReID branch are firstly concatenated along channel dimension and then transferred to feature vectors by Global Average Pooling. Then, a PCB model (*"PCB_T"*) is employed to obtain "soft-targets", *i.e.,* $P_T = \{p_T^i\}_{i=1}^N$. N refers to the number of training images. Formally, given an image I_i with the identity label y_i, p_T^i can be obtained as follows,

$$p_T^i = p(y_T^i = y_i|I_i) = \frac{\exp(y_T^i)}{\sum_{k=1}^{N_{id}} \exp(y_T^k)}, \tag{5.5.1}$$

where $y_T^k = w_{t_k}^T f_T^i$, f_T^i is the fine-grained feature of the i-th image output by the FC layer in *"PCB_T"*. w_{t_k} indicates the parameter of the classifier in *"PCB_T"* for the k-th identity. N_{id} is the number of identities in the whole training set. For the j-th PCB model of Pose Estimation branch and ReID branch, the corresponding probability predictions, *i.e.,* $P_{P_j} = \{p_{P_j}^i\}_{i=1}^N$ and $P_{ID_j} = \{p_{ID_j}^i\}_{i=1}^N$, can be calculated in the same way, respectively.

In order to supervise the local feature learning with the global one, KD loss L_{KD} is employed to reduce the distance between two prediction distributions, *i.e.,* P_T and $P_{ID(P)_j}$. Given a mini-batch with M images, L_{KD} is formulated as follows,

$$L_{KD} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{P}\big(KL(p_T^i, p_{ID_j}^i) + KL(p_T^i, p_{P_j}^i)\big), \tag{5.5.2}$$

where $KL(p, q)$ measures the Kullback-Leibler divergence between distribution p and distribution q. P denotes the number of PCB models of each branch.

5.6 Objectives

Batch sampling method Following [261, 217], an online sampling strategy is adopted during training to build mini-batches. Specifically, for each mini-batch, D identities are randomly selected. For each identity, K RGB images and K infrared (IR) images are randomly selected. Therefore, the batch size M is $2 * DK$. In the work, $D = 8$, $K = 4$, and $M = 64$ are adopted during training.

Pose estimation loss. To encourage Pose Estimation branch to learn modality-shared features, pose estimation loss L_{pose} is introduced to minimize pixel-wise Euclidean distances between ground-truth body keypoint heatmaps and the predicted ones. In the paper, ground-truth heatmaps are derived by a pose estimation model [137], which is pre-trained on a pose estimation dataset - LIP [111]. Formally, L_{pose} over a mini-batch is defined by,

$$L_{pose} = \frac{1}{M}\sum_{i=1}^{M}\sum_{x,y}(H_i(x,y) - \hat{H}_i(x,y))^2, \tag{5.6.1}$$

where $H_i(x,y)$, $\hat{H}_i(x,y)$ represent the pixel value at the position (x,y) of i-th ground truth and the predicted body keypoint heatmap, respectively.

Identity loss. To extract ID-related features, identity loss is performed on each PCB model of both Pose Estimation branch and ReID branch. For i-th image I_i whose identity label is y_i, the probability of being correctly classified given by j-th PCB model of Pose Estimation branch can be presented as $p^i_{P_j}(y_i|I_i)$, which is denoted as $p^i_{P_j}$ in the following content for simplicity. Similarly, the probability given by j-th PCB model of ReID branch can be presented as $p^i_{ID_j}$. Given probabilities, identity loss L_{id} over a mini-batch is formulated as,

$$L_{id} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{P}(\log p^i_{P_j} + \log p^i_{ID_j}), \tag{5.6.2}$$

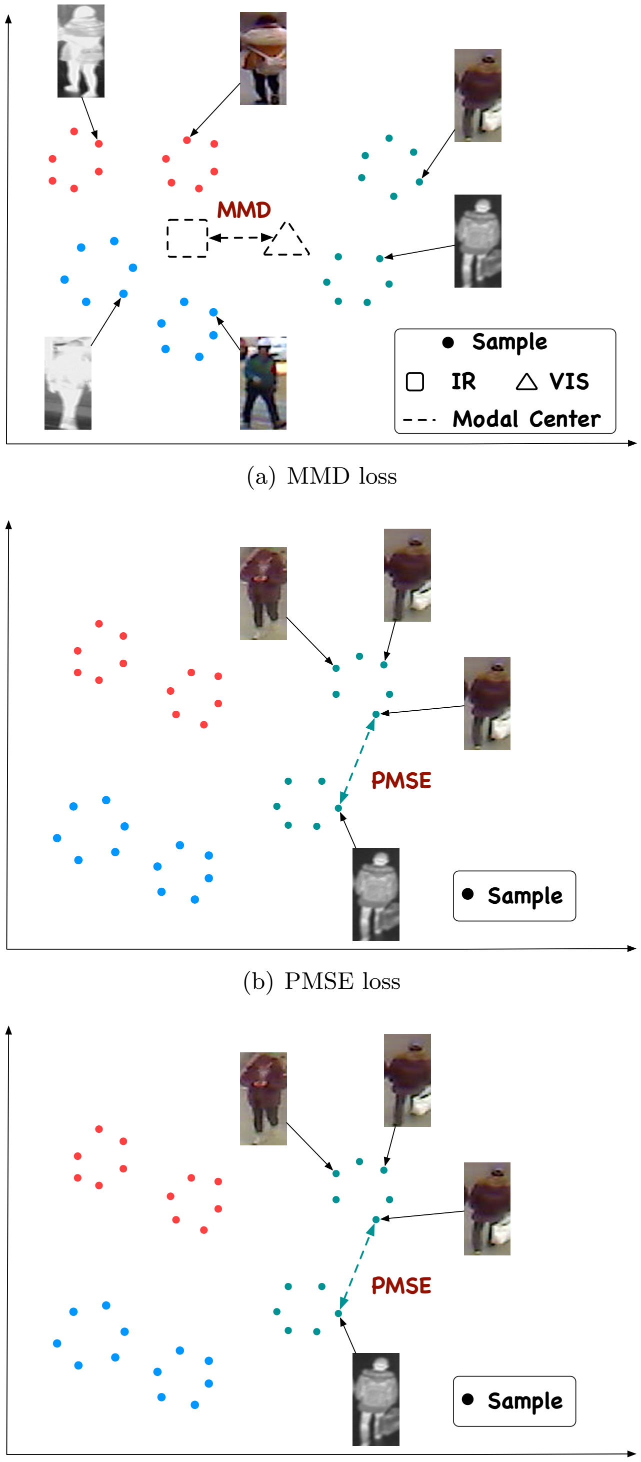
where P denotes the number of PCB models of each branch.

ID-MMD Loss. To alleviate the cross-modality discrepancy of VI-ReID, a distribution alignment metric - Maximum Mean Discrepancy (MMD) loss [66] is adopted. Supposing that a mini-batch contains M infrared (IR) images and N visible (VIS) images, the MMD loss $\mathcal{L}_{mmd}$ is formulated as follows,

$$\mathcal{L}_{mmd} = \left\| \frac{1}{M}\sum_{i=1}^{M}\phi(x_i^{ir}) - \frac{1}{N}\sum_{j=1}^{N}\phi(x_j^{vis}) \right\|^2_{\mathcal{H}}, \tag{5.6.3}$$

where $\phi(\cdot)$ represents the kernel function that maps the original data to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$. Although the MMD loss reduces the domain discrepancy by aligning IR-VIS feature distributions, rigidly adopting such domain-level constraint to VI-ReID is sub-optimal as it considers each modality as a whole and ignores the identity feature distribution within the modality, as shown in Fig 5.4.

To solve the problem, an explicit solution is reducing the distance between each IR-VIS image pair of the same identity in the latent space, *i.e.,* minimizing the Pairwise Mean Square Error (PMSE) $\mathcal{L}_{pmse}$ between IR-VIS pairs. Concretely,



(a) MMD loss

(b) PMSE loss

(c) IDMMD loss

Figure 5.4: Comparisons between the traditional MMD loss, the Pairwise Mean Square Error (PMSE) and the proposed IDentity-based MMD (ID-MMD) loss. Different identities are marked by different colors (best viewed in color).

following the PK-sampling setting discussed above, the $\mathcal{L}_{pmse}$ is denoted as follows,

$$\mathcal{L}_{pmse} = \frac{1}{D}\frac{1}{K}\sum_{p=1}^{D}\sum_{k=1}^{K}\left\| f_{p,k}^{ir} - f_{p,k}^{vis} \right\|^2, \tag{5.6.4}$$

where $f_{p,k}^{ir/vis}$ denotes the normalized feature of k-th IR/VIS image of p-th identity. Although the identity distribution is considered, such pairwise loss reduces the modality discrepancy at the instance level, where the network is highly likely to focus on identity-irrelevant details, such as poses and backgrounds, rather than identity features. Taking the man with the handbag in Fig. 5.4 as an example, the PMSE loss may reduce the feature distance between the IR-VIS image pair by encouraging the network to attend to the back viewpoint or the handbag, instead of the identity features.

To address the problems of the domain-based, *i.e.,* $\mathcal{L}_{mmd}$, and the instance-based modality discrepancy reduction loss, *i.e.,* $\mathcal{L}_{pmse}$, an ID-based MMD loss $\mathcal{L}_{idmmd}$ is employed, which bridges the modality gap by reducing the distance between the IR-VIS feature centroids of the same identity in the RKHS. Formally, for the given mini-batch in the z-th PCB model of Pose Estimation branch, the $\mathcal{L}_{idmmd}^{z}$ is computed as follows,

$$\mathcal{L}_{idmmd}^{z} = \frac{1}{D}\sum_{p=1}^{D}\left\| \phi(\frac{1}{K}\sum_{k=1}^{K} f_{p,k}^{ir}) - \phi(\frac{1}{K}\sum_{k=1}^{K} f_{p,k}^{vis}) \right\|_{\mathcal{H}}^{2}. \tag{5.6.5}$$

Therefore, the overall ID-MMD loss can be derived as follows,

$$\mathcal{L}_{idmmd} = \frac{1}{P}\sum_{z=1}^{P}\mathcal{L}_{idmmd}^{z} \tag{5.6.6}$$

Overall, the objective for training is defined as,

$$L = L_{id} + \beta L_{idmmd} + \lambda L_{pose} + \gamma L_{KD}, \tag{5.6.7}$$

where β, λ, and γ are the weighting factors to balance each loss term, which are empirically set as 0.1, 5, 1 in the paper, respectively.

5.7 Experiments

5.7.1 Datasets

Since SYSU-MM01 [194] and RegDB [136] are the only cross-modality person ReID datasets so far, the evaluation of this work is conducted on the two benchmark datasets. Details of benchmark datasets are reported in Table 5.1.

SYSU-MM01 [194] consists of images captured by 6 cameras, including 2 IR cameras and 4 RGB ones (2 outdoors and 2 indoors). The training set contains 395

Dataset	RGB images	IR images	Cameras	Persons
SYSU-MM01 [194]	287,628	15,792	6	491
RegDB [136]	4120	4120	2	412

Table 5.1: Details of VI-ReID benchmark datasets.

persons, with 22,258 RGB images and 11,909 IR images. The test set contains 96 persons, with 3,803 IR images for query and 301 randomly selected RGB images as the gallery. Following [194], two evaluation modes are conducted: *All-search* and *Indoor-search*. For *Indoor-search* mode, images collected by indoor RGB cameras are exclusively selected to build the gallery set. For *All-search* mode, images are randomly selected from all RGB cameras to form the gallery set.

RegDB [136] contains 412 identities, with 206 for training and 206 for testing. Each identity has 10 RGB and 10 IR images. Two evaluation modes are employed: *Visible-Thermal* and *Thermal-Visible*. The former refers to searching for corresponding IR images with an RGB image and vice versa. The dataset is randomly split into 10 training/testing trials. The evaluation results are given by averaging the performances over the 10 trials.

5.7.2 Evaluation metrics and implementation details

Evaluation metrics. Following the standard evaluation protocol given by [212, 261, 220], Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) are adopted as evaluation metrics. Here, CMC reports the probabilities of the targeted identity occurring at top-r in the ranking list, *i.e.,* “Rank-r” accuracy. mAP measures the overall retrieval performance when multiple matching cases occur in the gallery set.

Implementation details. The experiments are deployed on an NVIDIA GeForce 2080Ti GPU with Pytorch. Following most existing works [217, 261, 220], all input images are resized to 288×144. Random cropping, random erasing, and horizontal flipping are adopted for data augmentation. The parameters of Modality-specific module, Modality-shared module, and ReID branch are initialized by ResNet50 [77] pre-trained on ImageNet. Other parameters are initialized by Xavier initialization [62]. The SGD optimizer with a weight decay of 0.0005 is adopted for optimization. The learning rate is initialized as 0.01 and decays by 0.5 at every 20 epochs. The training process iterates for 100 epochs in total.

5.7.3 Comparison with the state-of-the-art

The author extensively compares the proposed algorithm with the current State-Of-The-Art (SOTA) methods on both SYSU-MM01 [194] and RegDB [136] datasets. The SOTA methods include pioneering ones (Zero-Pad [194] and HCML [212]), GAN-based ones (cmGAN [35], AlignGAN [180], and D^{2}RL [188]), middle

Table 5.2: Comparison with the state-of-the-art methods on **SYSU-MM01** [194] (***All-search*** setting). Rank-r ($r = 1, 10, 20$) accuracy(%) and mAP(%) are reported (*Higher is better*). $Ours_{ID}$ and $Ours_{ALL}$ denote the features used for evaluation are obtained from ReID branch and both branches, respectively.

Method	Venue	All-search			
		Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [194]	ICCV2017	14.80	54.12	71.33	15.95
HCML [212]	AAAI2018	14.32	53.16	69.17	16.16
cmGAN [35]	IJCAI2018	26.97	67.51	80.56	31.49
eDBTR [215]	TIFS2019	27.82	67.34	81.34	28.42
HSME [74]	AAAI2019	20.68	32.74	77.95	23.12
D^{2}RL [188]	CVPR2019	28.90	70.60	82.40	29.20
MSR [54]	TIP2019	37.35	83.40	93.34	38.11
AlignGAN [180]	ICCV2019	42.40	85.00	93.70	40.70
TSLFN [261]	Neuro2020	56.96	91.50	96.82	54.95
AGW [220]	Arxiv2020	47.50	-	-	47.65
X-Modal [101]	AAAI2020	49.92	89.79	95.96	50.73
MACE [216]	TIP2020	51.64	87.25	94.44	50.11
DDAG [217]	ECCV2020	54.75	90.36	95.81	53.02
cm-SSFT [126]	CVPR2020	61.60	89.20	93.90	63.20
NFS [28]	CVPR2021	56.91	91.34	96.52	55.45
CICL [242]	AAAI2021	57.2	94.3	98.4	59.3
GLMC [229]	TNNLS2021	64.37	93.90	97.53	63.43
LbA [138]	ICCV2021	55.41	-	-	54.14
SPOT [19]	TIP2022	65.34	92.73	97.04	62.25
$Ours_{ID}$	-	65.82	94.53	98.23	64.52
$Ours_{ALL}$	-	**71.21**	**95.35**	**98.81**	**67.15**

modality based ones (X-Modal [101] and cm-SSFT [126]), feature constraints based ones (eBDTR [215], HSME [74], MSR [54], TSLFN [261], AGW [220], and GLMC [229]), dual-level feature alignment based ones (DDAG [217], MACE [216], CICL [242], LbA [138] and SPOT [19]).

Evaluations on SYSU-MM01. Table 5.2 and Table 5.3 report the performance of the proposed model and State-Of-The-Art (SOTA) approaches on the SYSU-MM01 [194] dataset at *All-search* setting and *Indoor-search* setting, respectively. "$Ours_{ID}$" and "$Ours_{ALL}$" refer to the retrieval performance with features f_{ID} (Eqn. 5.4.12) and f_{ALL} (Eqn. 5.4.13), respectively. It can be seen that both "$Ours_{ID}$" and "$Ours_{ALL}$" outperform SOTA approaches on ALL evaluation metrics in both *All-search* and *Indoor-search* evaluation modes. Specifically, when retrieving only with the identity features from ReID branch, *i.e.,* f_{ID}, the proposed method outperforms the SOTA performance (GLMC [229]) by 1.45% and 1.09% in terms of the rank-1 accuracy and the mAP score, respectively. Compared to SPOT [19], which

Table 5.3: Comparison with the state-of-the-art methods on **SYSU-MM01** [194] (***Indoor-search*** setting). Rank-r ($r = 1, 10, 20$) accuracy(%) and mAP(%) are reported (*Higher is better*). $Ours_{ID}$ and $Ours_{ALL}$ denote the features used for evaluation are obtained from ReID branch and both branches, respectively.

Method	Venue	Indoor-search			
		Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [194]	ICCV2017	20.58	68.38	85.79	26.92
HCML [212]	AAAI2018	24.52	73.25	86.73	30.08
cmGAN [35]	IJCAI2018	31.63	77.23	89.18	42.19
eDBTR [215]	TIFS2019	32.46	77.42	89.62	42.46
MSR [54]	TIP2019	39.64	89.29	97.66	50.88
AlignGAN [180]	ICCV2019	45.90	87.60	94.40	54.30
TSLFN [261]	Neuro2020	59.74	92.07	96.22	64.91
AGW [220]	Arxiv2020	54.17	-	-	62.97
MACE [216]	TIP2020	57.35	93.02	97.47	64.79
DDAG [217]	ECCV2020	61.02	94.06	98.41	67.98
cm-SSFT [126]	CVPR2020	70.50	94.90	97.70	72.60
NFS [28]	CVPR2021	62.69	96.53	99.07	69.79
CICL [242]	AAAI2021	66.6	98.8	99.7	74.7
GLMC [229]	TNNLS2021	67.35	98.10	99.77	74.02
LbA [138]	ICCV2021	58.46	-	-	66.33
SPOT [19]	TIP2022	69.42	96.22	99.12	74.63
$Ours_{ID}$	-	71.74	94.57	97.60	74.54
$Ours_{ALL}$	-	**72.55**	**97.15**	**98.60**	**77.05**

also employs the pose information, the proposed method improves the rank-1 accuracy by 0.48% and the mAP score by 2.27%. When the keypoint-aware features f_P are also involved for evaluation, the performances are further improved by 5.39% and 2.63% in terms of the rank-1 accuracy and the mAP score, respectively.

Evaluations on RegDB. The evaluation results on RegDB [136] are shown in Table 5.4. It can be observed that the proposed model obtains surprisingly good results in both *"Visible-Thermal"* and *"Thermal-Visible"* modes. Specifically, when retrieving merely with f_{ID}, the proposed method exceeds the SOTA one (GLMC [229]) by 6.46% and 5.64% on the mAP score in two evaluation modes, respectively. Further, the improvements achieve 7.56% and 6.77% when features from both pose estimation branch and ReID branch (f_{ALL}) are used for identity retrieval. Additionally, compared to the pose-guided method SPOT [19], the proposed method improves the mAP score by 15.42% and 16.52% when retrieving with features f_{ID} and f_{ALL}, respectively.

Table 5.4: Comparison with the state-of-the-art methods on **RegDB**. Rank-r ($r = 1, 10, 20$) accuracy(%) and mAP(%) are reported (*Higher is better*). $Ours_{ID}$ and $Ours_{ALL}$ denote the features used for evaluation are obtained from ReID branch and both branches, respectively.

Method	Visible-Thermal				Thermal-Visible			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad [194]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML [212]	24.44	47.53	56.78	20.80	21.70	45.02	55.58	22.24
eDBTR [215]	34.62	58.96	68.72	33.46	34.21	58.74	68.64	32.49
HSME [74]	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
D^{2}RL [188]	43.40	66.10	76.30	44.10	-	-	-	-
MSR [54]	48.43	70.32	79.95	48.67	-	-	-	-
AlignGAN [180]	57.9	-	-	53.6	56.3	-	-	53.4
X-Modal [101]	62.21	83.13	91.72	60.18	-	-	-	-
DDAG [217]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
AGW [220]	70.05	-	-	66.37	-	-	-	-
MACE [216]	72.37	88.40	93.59	69.09	72.12	88.07	93.07	68.57
cm-SSFT [126]	72.3	-	-	72.9	71.0	-	-	71.7
NFS [28]	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
CICL [242]	78.8	-	-	69.4	77.9	-	-	69.4
GLMC [229]	91.84	97.86	98.98	81.42	91.12	97.86	98.69	81.06
LbA [138]	74.17	-	-	67.64	72.43	-	-	65.46
SPOT [19]	80.35	93.48	96.44	72.46	79.37	92.79	96.01	72.26
$Ours_{ID}$	92.14	98.16	99.22	87.88	91.36	97.57	98.88	86.70
$Ours_{ALL}$	**93.35**	**98.61**	**99.42**	**88.98**	**92.72**	**98.79**	**99.36**	**87.83**

5.7.4 Ablation study

Ablation studies are conducted on SYSU-MM01 [194] and RegDB [136] to verify the effectiveness of the proposed method. To prove the effectiveness of the proposed components, *i.e.,* Pose Estimation branch (PEB) and Hierarchical Feature Constraint (HFC), a "Baseline" model is trained under the supervision of identity loss L_{id} and HC-tri loss L_{idmmd}, which only consists of Modality-specific module, Modality-shared module, and ReID branch. The evaluation results are shown in the first row in Table 5.5. Based on "Baseline", Pose Estimation branch, pose estimation loss (L_{pose}), and Hierarchical Feature Constraint are gradually applied, which results are illustrated in the following rows in Table 5.5. Note that the reported results are in the *All-search* mode for the SYSU-MM01 dataset while in the *"Visible-Thermal"* mode for the RegDB dataset.

Effectiveness of Pose Estimation branch (PEB). As can be seen from the $2nd$ row in Table 5.5, by integrating the PEB branch that is pre-trained on a pose estimation dataset, the framework yields an increase of approximately 2% and 4% in terms of the mAP score on the SYSU-MM01 dataset and the RegDB dataset, respectively, when f_{ID} is applied for identity retrieval. Likewise, the improvements exceed 6.5% when employing f_{ALL}, which clearly demonstrates the effectiveness of the proposed PEB branch.

Apart from L_{id} and L_{idmmd}, L_{pose} is further employed on such a structure to ensure the body skeleton points are precisely estimated. The results for f_{ID} and f_{ALL} are shown in the $3rd$ and $6th$ rows, respectively. Specifically, an increase of 3.83% and 2.41% can be found in terms of the mAP score on two benchmark datasets for f_{ID}. A similar increase, *i.e.,* 3.16% and 2.96%, can be obtained if using f_{ALL}.

Additionally, the comparison of the identity retrieval performance when adopting different features are conducted, *i.e.,* f_{ID}, f_P in Eqn. (5.4.12), and f_{ALL} in Eqn. (5.4.13). The results are illustrated in Table 5.6. As can be seen, retrieving identity with features derived from the ReID branch (f_{ID}) achieves a better performance than that from the Pose Estimation branch (f_P). Specifically, "$Ours_{ID}$" surpasses "$Ours_P$" by 4.33% and 6.66% in terms of the mAP score on SYSU-MM01 and RegDB, respectively. The results prove that, compared to the modality-invariant pose features, the identity-related features play a dominant role in VI-ReID. Additionally, when combining the identity-related features and the modality-invariant pose features, the method achieves the best performance. Concretely, "$Ours_{ALL}$" improves the mAP score of "$Ours_{ID}$" by 2.63% and 1.10% on SYSU-MM01 and RegDB, respectively.

Effectiveness of Hierarchical Feature Constraint (HFC). It can be observed from the $4th$ row that HFC yields a satisfactory improvement for f_{ID} in terms of the mAP score, which are 1.73% and 3.72% on the SYSU-MM01 dataset and the RegDB dataset, respectively. A similar improvement can also be found for f_{ALL} (the $7th$ row), which are around 2% for both datasets. The boost in performance

Table 5.5: Ablation studies on SYSU-MM01 [194] and RegDB [136]. "PEB", "HFC" refer to Pose Estimation branch, and Hierarchical Feature Constraint, respectively.

	Components			SYSU-MM01		RegDB	
	PEB	L_{pose}	HFC	Rank-1	mAP	Rank-1	mAP
Baseline				57.03	56.21	85.76	77.21
$Ours_{ID}$	✓			60.19	58.96	87.39	81.75
	✓	✓		63.90	62.79	89.25	84.16
	✓	✓	✓	**65.82**	**64.52**	**92.14**	**87.88**
$Ours_{ALL}$	✓			65.05	62.78	89.53	84.12
	✓	✓		68.84	65.94	91.96	87.08
	✓	✓	✓	**71.21**	**67.15**	**93.35**	**88.98**

Table 5.6: Comparisons between identity retrieval performances with different features, *i.e.,* f_{ID}, f_P in Eqn. (5.4.12), and f_{ALL} in Eqn. (5.4.13), which are denoted by $Ours_{ID}$, $Ours_P$, and $Ours_{ALL}$, respectively.

Methods	SYSU-MM01		RegDB	
	Rank-1	mAP	Rank-1	mAP
$Ours_{ID}$	65.82	64.52	92.14	87.88
$Ours_P$	62.84	60.19	87.38	81.22
$Ours_{ALL}$	71.21	67.15	93.35	88.98

proves that the proposed HFC encourages the information flow between global and local features by introducing extra supervision on part-level features. Additionally, the performance boost from "$Ours_{ID}$" to "$Ours_{ALL}$" in Table 5.6 also proves that, with the proposed HFC, each feature stripe is embedded with modality-shared and ID-related information. Such an advantage leads to satisfactory identity accuracy even when features with less dimensions are used for identity retrieval.

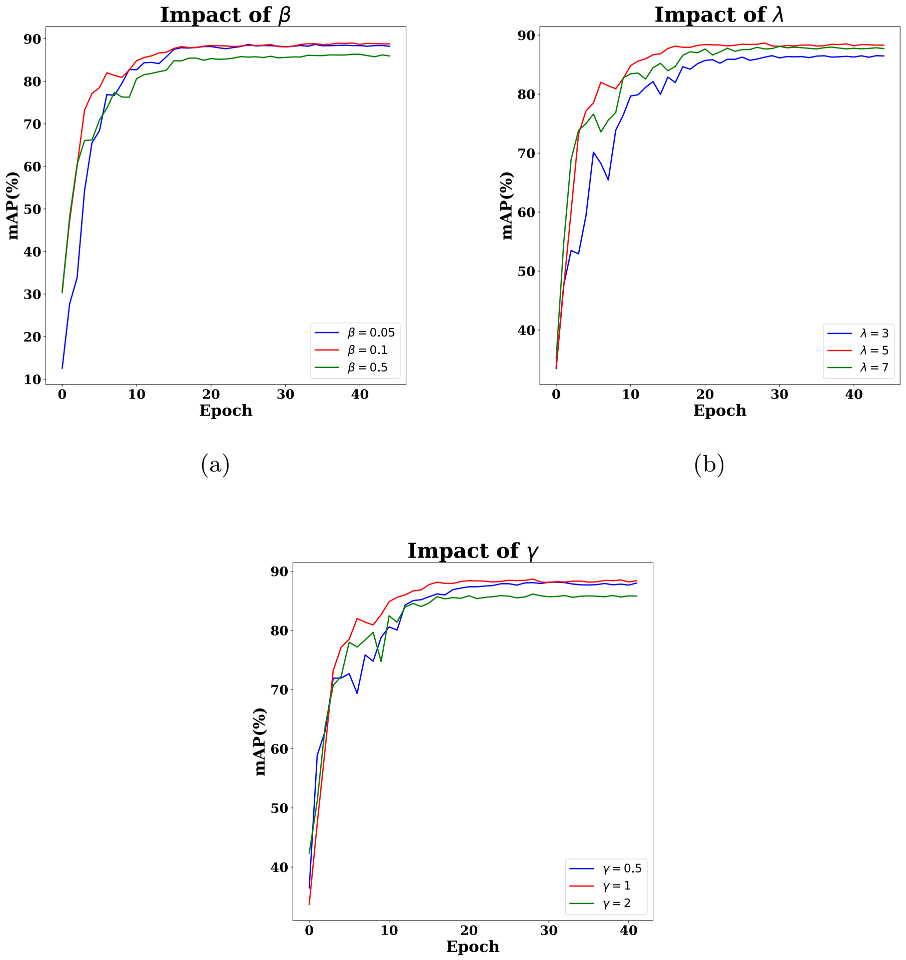
5.7.5 Further analysis and discussions

The impact of loss coefficients are evaluated and an out-of-sample experiment of the domain reduction loss L_{idmmd} on the infrared-visible face recognition task is conducted.

Impact of loss coefficients. Firstly, the author aims to investigate an optimal combination of weight factors to balance multiple loss terms. According to the paper, the overall objective L for training is defined as,

$$L = L_{id} + \beta L_{idmmd} + \lambda L_{pose} + \gamma L_{KD}. \tag{5.7.1}$$

β, λ, and γ adjust the contribution of Hetero-center triplet loss, Pose estimation loss, and KD loss, respectively. For a fair comparison, the author changes one factor at a time while keeping the rest fixed. Subsequently, models are trained by objectives



(a)

(b)

(c)

Figure 5.5: Parameter analysis on RegDB [136]. **Best viewed in color.**

with different combinations of weight factors. The corresponding mAP scores are depicted in Fig. 5.5.

It can be observed that no matter which combination is employed, the proposed framework could achieve a stable performance within 50 epochs. Additionally, according to the experimental results, the optimal combination $\{\beta, \lambda, \gamma\} = \{0.1, 5, 1\}$ is chosen during the training.

Out-of-sample experiment. To further demonstrate the effectiveness of the proposed ID-MMD, an out-of-sample experiment is conducted in the IR-VIS face recognition task. Similar to VI-ReID, IR-VIS face recognition performs facial image retrieval across the IR modality and the VIS modality.

The framework of IR-VIS face recognition is illustrated in Fig. 5.6. Following

	FAR=0.01%	FAR=0.1%	Rank-1
$\mathcal{L}_{\text{idmmd}}$	$\mathbf{91.97 \pm 1.5}$	$\mathbf{97.96 \pm 0.3}$	$\mathbf{98.87 \pm 0.3}$
$\mathcal{L}_{\text{mmd}}$	90.93 ± 1.6	97.27 ± 0.3	98.04 ± 0.3
$\mathcal{L}_{\text{pmse}}$	90.41 ± 1.5	96.88 ± 0.2	97.85 ± 0.3

Table 5.7: Comparisons between IR-VIS face recognition performances with network (LightCNN-29) trained with MMD, PMSE, and the proposed ID-MMD loss on LAMP-HQ [222] (*Higher is better*).

a SOTA IR-VIS face recognition work [56], the encoder adopts the structure of a SOTA face recognition network, *i.e.,* LightCNN29 [199]. During training, two commonly used IR-VIS face recognition losses are adopted for the network optimization, including an identity loss L_{id} and a modality gap reduction loss L_{idmmd}. To compare with other SOTA modality gap reduction losses discussed in Section 5.6, the networks are trained with MMD loss, PMSE loss, and the proposed ID-MMD loss, respectively, while keeping other settings fixed.

The comparison experiments are conducted on an IR-VIS face recognition benchmark dataset, *i.e.,* LAMP-HQ [222]. LAMP-HQ consists of 73616 images of 573 identities, which are diverse in poses, illumination, and ages. Following the standard protocol [222], 50% identities are randomly selected as the training set while the rest serve as the testing set. The verification rate (VR)@false accept rate (FAR)=0.01%, VR@FAR=0.1%, and Rank-1 accuracy are employed for evaluation. *Higher is better* for all metrics. The comparison between performances of models trained with different losses is reported in Table 5.7. As can be seen, a model trained with the proposed ID-MMD outperforms the other two in terms of ALL metrics.

To better understand the advantage of the proposed ID-MMD, the identity feature distributions of LAMP-HQ [222] is visualized. Specifically, 10 identities are randomly selected from the testing set and for each identity, 10 IR images and 10 VIS images are randomly selected. The distribution of features derived by models

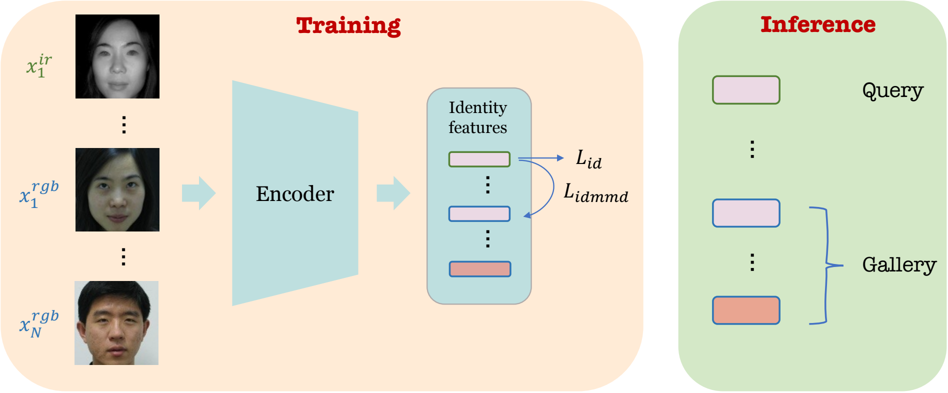
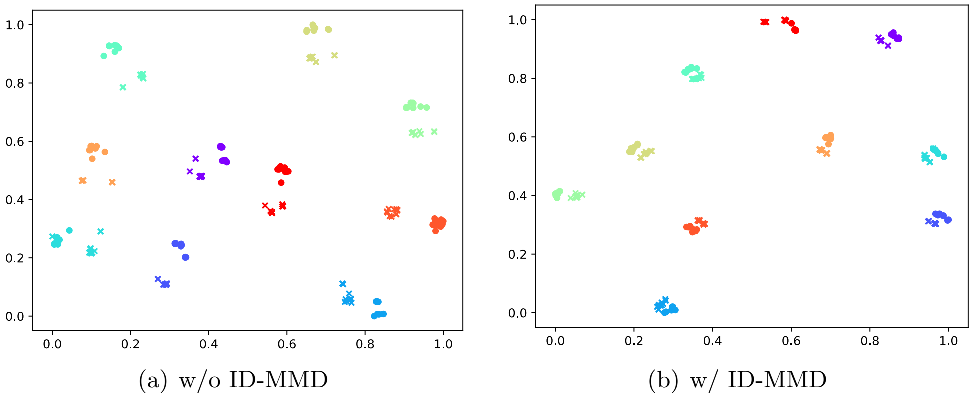


Figure 5.6: Pipeline of IR-VIS face recognition. Different identities are denoted by different colors. **Best viewed in color.**



(a) w/o ID-MMD

(b) w/ ID-MMD

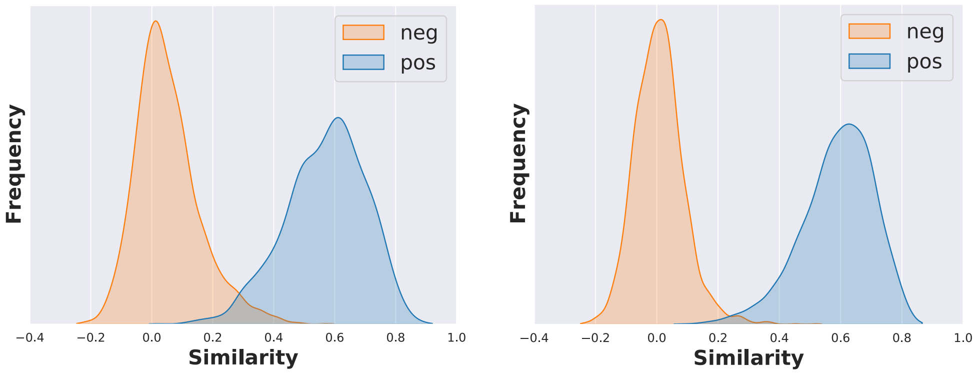
Figure 5.7: Visualization of identity features derived by models trained without (w/o) and with(w/) ID-MMD on LAMP-HQ via t-SNE [173]. LC-29 refers to LightCNN29 [199]. Different identities are denoted by different colors. ●: IR images; ×: VIS images. **Best viewed in color.**

trained without and with the proposed ID-MMD is visualized in Fig. 5.7, which are denoted as w/o ID-MMD and w/ ID-MMD, respectively. As can be seen, after involving the ID-MMD loss, features of the same identity of two modalities are pulled closer. Meanwhile, for each identity, the features within the IR/VIS domain distribute more compactly. Such visualization results suggest that the proposed method can effectively reduce both intra-modality and inter-modality discrepancies. Additionally, the feature similarity distribution of positive pairs (belonging to the same identity) and negative pairs (belonging to different identities) of LAMP-HQ are visualized in Fig. 5.8. Benefiting from the proposed ID-MMD loss, the similarities between positive pairs increase while the similarities between negative pairs decrease.

5.7.6 Comparison with pose-guided person ReID

As aforementioned, employing pose information derived by an RGB-based pose estimator in the VI-ReID task is extremely challenging due to the massive gap between visible images (source domain) and infrared images (target domain). To reveal how to enable pose information to facilitate the VI-ReID task, several exploratory experiments are conducted on the SYSU-MM01 [194] dataset.

Firstly, the VI-ReID task is considered as the single-modality person ReID task with two different types of images. For each modality, a modality-specific pose-assisted feature extractor is trained. During the inference, features of the query set (IR images) and gallery set (RGB images) are derived by the corresponding feature extractor to conduct the cross-modality person retrieval. Taking a state-of-the-art (SOTA) pose-guided single-modality person ReID work, *i.e.,* PGFA [131], as an example, where human landmarks obtained by an off-the-shelf pose estimator are rigidly used to generate attention maps to highlight body regions, the performance is reported in the first row of Table 5.8. As can be seen, although the same pose



(a) w/o ID-MMD

(b) w/ ID-MMD

Figure 5.8: Feature similarity distribution of positive/negative pairs of LAMP-HQ. w/o and w/ refer to models trained without and with ID-MMD, respectively.

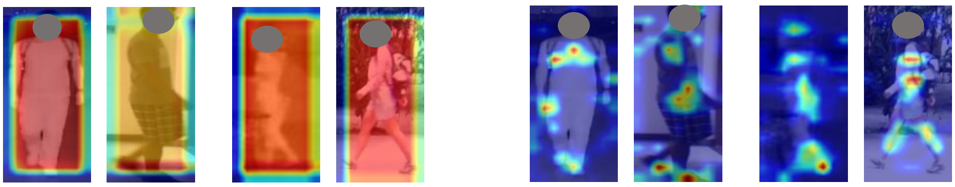
Table 5.8: Comparison with the state-of-the-art pose-guided single-modality person ReID methods on SYSU-MM01 [194] in terms of Rank-1 accuracy(%) and mAP(%) at *All-search* setting (*Higher is better*). The dimension of features used for the evaluation is set as 2560 for all experiments. $Ours_{ID}$*: P in Eqn. (5.4.11) is set as 5. Impl.: experiments are implemented with the official source code provided.

Methods	Rank-1	mAP
PGFA(Impl.) [131]	10.04	11.45
PABR(Impl.) [165]	40.73	42.19
$Ours_{ID}$*	62.82	61.90

estimator is used for two modalities, retrieving identities across modalities with modality-specific pose-assisted features can only achieve 11.45% in terms of the mAP score. The poor performance demonstrates that the off-the-shelf pose estimators cannot handle the huge gap between the data distribution of the two modalities. In this case, the pose information cannot serve as effective modality-shared cues in the VI-ReID task.

To make the best of the pose information in VI-ReID, the two-stream network is adopted by a single-modality person ReID approach, *i.e.*, PABR [165], where the ReID branch and the pose branch extract appearance and pose features, respectively. Then two kinds of features are fused for identity retrieval. To adapt PABR to the VI-ReID task, the cross-modality hard triplet loss [241] is used to replace the single-modality one, where the hardest cross-modality triplets are also considered. The performance of PABR in the VI-ReID task is shown in the second row of Table 5.8.

As can be seen, compared to PGFA, PABR achieves a higher Rank-1 accuracy (40.73%) and mAP score (42.19%). In addition to numerical results, the attentive feature maps output by the pose branch of PABR [165] are visualized by means of Grad-CAM [156] in Fig. 5.9(a), to have a better understanding of where pose



(a) PABR [165]

(b) *Ours**

Figure 5.9: Visualization of attentive feature maps output by the pose branch of PABR [165] and the proposed method on SYSU-MM01 [194]. For each person (group), the RGB image is shown on the left and the IR image is on the right. *Ours** means the proposed model is trained with identity loss and cross-modality hard triplet loss only.

features are extracted. As can be seen, due to the lack of adequate guidance on the pose branch during training, the extracted pose features only locate the whole body coarsely.

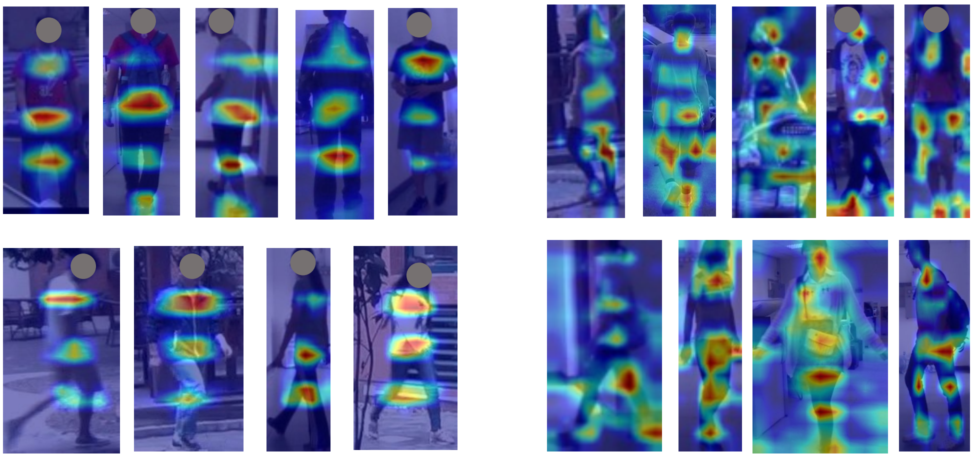
To solve the problem, as shown in Fig. 5.2, the body keypoints generated by a pre-trained pose estimator serve only as the guidance of Pose Estimation branch in the proposed method. By applying the pose estimation loss during training, more fine-grained pose features can be obtained. The performance of the proposed method is shown in the third row of Table 5.8. It can be seen that, with the above improvements, the proposed method outperforms PGFA and PABR by a large margin in terms of the Rank-1 accuracy (62.92%) and the mAP score (61.90%). Additionally, the visualization of the attentive feature maps output by the proposed Pose Estimation branch is shown Fig. 5.9(b). In comparison to PABR, the proposed method focuses on more body details, such as shoulders and feet, which can serve as distinctive cues for VI-ReID.

5.7.7 Visualization

Apart from quantitative results, the gradient feature maps output by ReID branch is also visualized by Grad-CAM [156], in order to examine where the features are extracted. Visualization results of "Baseline" and the proposed framework ("Ours") are illustrated in Fig. 5.10(a) and Fig. 5.10(b), respectively. As can be seen, instead of rigidly extracting features from several vertical regions, the proposed method extracts features from body skeleton joints, which are not only ID-related but highly immune to viewpoint changes and modality changes. The visualization results not only intuitively reveal the reason why the proposed method performs better, but also show the potential of pose estimation tasks in the field of VI-ReID.

5.8 Conclusion

In this section, a novel two-stream VI-ReID framework is proposed, where modality-shared and ID-related features for identity retrieval are extracted by means



(a) Baseline

(b) Ours

Figure 5.10: Visualization of gradient feature maps output by Baseline and ours on SYSU-MM01.

of learning an auxiliary task (pose estimation) and the main task (person ReID) simultaneously. Two major contributions are made by the proposed work: 1) By imposing pose estimation and ReID constraints on the Pose Estimation branch at the same time, both modality-shared and ID-related information are fully embedded on each feature stripe. 2) Apart from learning discriminative features at the local level, the author also proposes a Hierarchical Feature Constraint to bond the learning of global features with local ones by employing the knowledge distillation strategy to ensure discriminability consistency. The proposed framework achieves new state-of-the-art VI-ReID benchmarks in terms of rank-1 accuracy and the mAP score.

Chapter 6

Conclusions and future work

6.1 Conclusions

In this thesis, three key challenges of person ReID: unsupervised person ReID, fast unsupervised person ReID, and cross-modality person ReID were targeted. For each challenge, existing solutions were firstly reviewed, including methodologies and limitations. To address the unsolved problems, the author proposed her own solutions. Subsequently, in order to compare with state-of-the-art methods, extensive experiments were conducted on benchmark datasets. Moreover, for each task, parameter analysis was conducted to ensure optimal performance and ablation studies were conducted to validate the effectiveness of each component of the proposed methods. Additionally, more qualitative results were provided to intuitively demonstrate the advantages of the proposed methods.

Specifically, in Chapter 3, a pseudo label refinement scheme was proposed for clustering-based unsupervised person ReID, which alleviates the pseudo label noise brought by imperfect clustering results under the unsupervised setting. Two major contributions were made: 1) Pseudo labels were refined with internal characteristics instead of auxiliary information such as camera IDs, body parts, or generated samples. Specifically, confidence-guided centroids (CGC) were proposed to provide reliable cluster-wise prototypes for feature learning, where low-confidence instances are filtered out during the formation of centroids. 2) Targeting the problem whereby a large proportion of samples are pushed to "wrong" centroids, the author proposed to use confidence-guided pseudo labels (CGL), which enables samples to approach not only the assigned centroid but other clusters where their identities are potentially embedded. With the aid of CGC and CGL, the proposed method yielded comparable and superior performance to state-of-the-art pseudo label refinement works that largely leverage auxiliary information.

In Chapter 4, an unsupervised transformation-invariant binary local descriptor learning method (TBLD) was proposed, which can be adapted to fast unsupervised person ReID. Two major contributions were made: 1) A framework that derives transformation-invariant binary local descriptors was proposed. Based on the assumption that the same object should be described by the same descriptors, the original image patches and their transformed counterparts were projected to an identical Euclidean subspace and an identical Hamming subspace with the help of contrastive loss. 2) An *Adversarial Constraint Module* (ACM) was proposed to solve the problem of high bit correlation of binary descriptors, where a set of low-coupling binary codes were introduced to guide the learning of binary local descriptors. By means of Wasserstein loss, the framework was optimized to transfer the distribution of the learned binary descriptors to the low-coupling ones, thereby making the learned ones as low-coupling as possible. Experimental results on three descriptor evaluation benchmark datasets demonstrate the superiority of the proposed approach over state-of-the-art unsupervised binary descriptors. Additionally, experiments were conducted on person ReID benchmark datasets to evaluate the effectiveness of the proposed unsupervised binary descriptor learning scheme to fast unsupervised person ReID. Although not to comparable to supervised methods that leverage identity labels, the proposed method significantly outperformed existing unsupervised binary descriptors on unsupervised person ReID.

In Chapter 5, a novel two-stream framework was proposed for person ReID cross visible modality and infrared modality, where ID-related features for identity retrieval were extracted by means of learning an auxiliary task (pose estimation) and the main task (person ReID) simultaneously. Two contributions were made: 1) Both keypoint-aware and ID-related features were enforced to be fully embedded on each feature stripe by imposing pose estimation and ReID constraints at the same time. 2) Hierarchical Feature Constraint was proposed to bond the learning of global features with local features by employing the knowledge distillation strategy to ensure discriminability consistency. The proposed framework achieved a new state-of-the-art performance on VI-ReID benchmark datasets in terms of the Rank-1 accuracy and the mAP score.

6.2 Future Work

In Chapter 2, existing solutions for challenges in person ReID were reviewed. Despite satisfactory performances achieved on public benchmark datasets, there are still some unsolved issues. In the section, unsolved problems and several future research directions of each challenge will be elaborately discussed.

6.2.1 Unsupervised person ReID

Several potential future research topics of unsupervised person ReID are listed as follows:

- **Generalizable unsupervised ReID.** As mentioned earlier, in Section 2.2, unsupervised ReID methods are categorized as unsupervised domain adaptation (UDA) ones and purely unsupervised (USL) ones. Generally, UDA methods perform well if the gap between the source domain and the target domain is slight, yet fail when the gap is significant. Therefore, the generalizability of UDA methods in real-world applications can be a big issue. Although

USL methods avoid the problem by discarding the knowledge from the source domain, their performances are far from satisfactory, especially on challenging datasets such as MSMT17 [190]. Therefore, an investigation in generalizable unsupervised ReID models can be a promising direction, which can achieve satisfactory performances on images with various illumination conditions as well as different image quality.

- **Unsupervised text-image person ReID.** Apart from VI-ReID, identity retrieval across text and image is also an emerging and promising cross-modality person ReID topic thanks to advances in natural language processing (NLP). Currently, the knowledge leveraged by UDA methods is collected from an image-based pretrained model. Following this concept, unsupervised text-image person ReID can be achieved by replacing the image-based pretrained model with text-based ones, which are trained with descriptions of people and corresponding people images. Such high-level semantic knowledge can benefit identity feature learning in the target domain.

6.2.2 Fast unsupervised person ReID

Several potential future research topics of fast person ReID are listed as follows:

- **Ranking strategy.** Due to the fact that binary bits carry far more less information than real-valued ones, the representability of binary descriptors cannot match real-valued ones, which results in unsatisfactory retrieval accuracy. Therefore, designing effective ranking strategies that take the advantage of both binary codes and real-valued representations is a feasible solution to balance the searching efficiency and accuracy.

- **Fast unsupervised person ReID.** Although many efforts have been made to fast person ReID [231, 22, 123, 182, 221], unsupervised binary identity feature learning remains an untouched topic. In the thesis, the author attempts to adapt the proposed unsupervised binary descriptor learning framework to fast person ReID to extract identity features. However, as demonstrated in the out-of-sample experiments of Section 4.7.6, the performances of unsupervised binary descriptors on benchmark person ReID datasets are far from satisfactory. Such failure can be attributed to two potential reasons: 1) Inferior backbones and training schemes are adopted, and 2) the significant gap between patch-based datasets and person-based datasets makes the knowledge transfer challenging.

  To verify the first argument, an additional experiment is conducted on Market-1501 [246]. Concretely, the proposed unsupervised person ReID framework in Chapter 3 (Fig. 3.5), termed CC, is based on a novel contrastive SSL method - MoCo [78]. To evaluate the potential of CC at fast person ReID setting, the author embeds the quantization loss L_Q (Eqn. (4.4.6)) into the objective of CC (Eqn. (3.6.3)) during training, and binarizes the real-valued features output by

Backbone	Dims / Bits	Type	mAP (%)
CC	2048	Real-valued	85.3
		Binary	72.4
	128	Real-valued	75.2
		Binary	54.5
TBLD	128	Binary	32.7

Table 6.1: Comparison between fast unsupervised person ReID performances with different backbones and code lengths. CC and TBLD refer to the person ReID backbone adopted in Chapter 3 and Chapter 4, respectively.

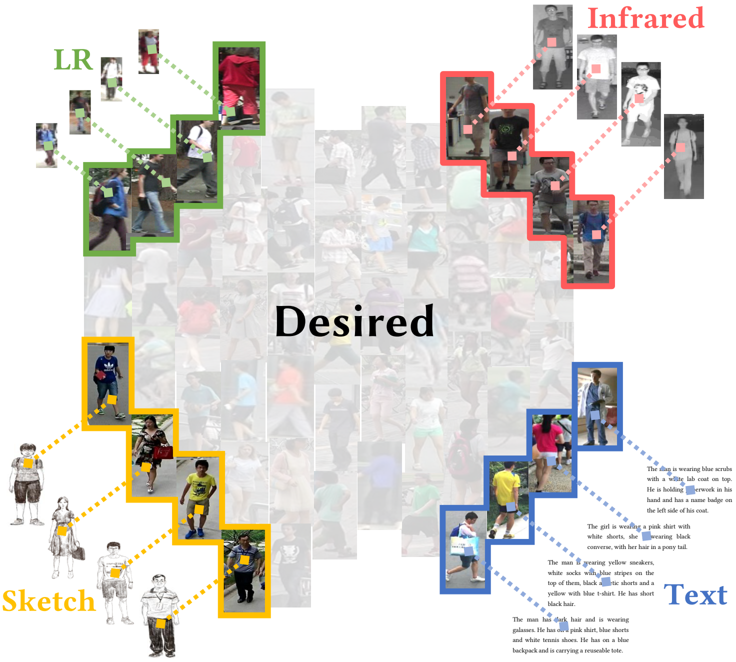
Encoder for the identity retrieval during inference. To make fair comparisons, multiple Encoders that can derive features of different dimensions/bits are trained. The identity retrieval performance in terms of mAP score are reported in Table 6.1. As can be seen, with binary features of the same dimension (128 bit), the mAP improves by over 20% when a more powerful backbone (CC) is adopted. Moreover, it can be observed that the retrieving with high-dimension binary features can achieve a better performance than low-dimension ones since more ID-related cues are embedded. Despite its superiority, CC has a vital problem: the clustering algorithm used to assign pseudo labels, *i.e.,* DBSCAN, generally fail to give adequate clustering results when adopting low-dimension features, such as 64, thereby leading to invalid identity features. Therefore, how to design powerful backbone frameworks and effective training schemes can be a promising future work.

Additionally, the significant gap between patch-based datasets and person-based ones can also leads to the performance gap of fast person ReID between unsupervised and supervised setting. As can be seen from Section 4.7, although the proposed method has achieved the state-of-the-art performances on patch-based datasets and has been re-trained on the corresponding person-based datasets when adapting to fast person ReID, the performances are not comparable with supervised ones. Such gap might be caused by the significant differences in the image content as well as the data acquisition of patch-based datasets and person-based ones. Therefore, how to eliminate the gap between two types of datasets and how to transfer the knowledge learned from patch-based datasets to person-based ones can also be promising future works.

6.2.3 Cross-modality person ReID

Several potential future research topics of cross-modality person ReID are listed as follows:

- **More challenging datasets.** As demonstrated in Chapter 5, there are only two VI-ReID benchmarks, *i.e.,* SYSU-MM01 [194] and RegDB [136]. Additionally, the size of both datasets is insufficient for the more powerful ReID



The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

The girl is wearing a pink shirt with white shorts, she is wearing black converse, with her hair in a pony tail.

The man is wearing yellow sneakers, white socks with blue stripes on the top of them, black athletic shorts and a yellow with blue t-shirt. He has short black hair.

The man has dark hair and is wearing galasses. He has on a pink shirt, blue shorts and white tennis shoes. He has on a blue backpack and is carrying a reuseable tote.

Figure 6.1: Various types of cross-modality person ReID. Taken from [187].

model training. Especially, the number of images in RegDB is less than 10,000, and meanwhile, the diversity of RGB-IR image pairs is not enormous, which fails to mimic real-world scenarios. Therefore, the emergence of challenging large-scale VI-ReID datasets will undoubtedly benefit the investigation of more powerful and adaptable VI-ReID models as well as more effective training schemes.

- **Lightweight VI-ReID models.** Since the gallery scale can be extremely large in real-world scenarios. For example, gallery can be presented as all staff in the workplace or all registered students on campus. Real-time search in such large-scale gallery requires lightweight ReID models as well as fast feature matching. The former enables the real-time identity feature extraction while the latter enables the real-time identity retrieval. Binary feature has been introduced to facilitate the fast feature matching and achieved satisfactory performances at the supervised setting, *i.e.*, with the help of ID labels. However, the research on lightweight VI-ReID models has been neglected in the past. Therefore, the development of lightweight VI-ReID models should be one of the main focuses of the future work.

- **Semi-supervised and unsupervised VI-ReID.** Deep learning powered ReID models generally require large datasets for training. However, manually labelling can be very burdensome, expensive, and even impossible for some frequently changing cases, such as retailers or shopping malls. Therefore, how to train a VI-ReID model with limited annotated data or even without identity labels has become an important research topic.

An attempt has been made by Liang *et al.* [110], where a two-stage unsupervised VI-ReID method is proposed. In the first stage, two self-learning ReID models are learned separately within each modality to extract intra-modality feature embeddings and generate intra-modality pseudo labels for the next stage. Then, in the second stage, the model is trained to learn modality-shared features under the supervision of distilled knowledge from pseudo labels. The work provides a baseline for unsupervised VI-ReID. However, the learning process is complex and lacks practical value, making semi-supervised and unsupervised VI-ReID remain challenging.

- **Various types of cross-modality person ReID.** Apart from the day to night identity retrieval, a wide range of challenges have been posed by real-world person search scenarios, which are shown in Fig. 6.1. For example, due to the significant differences in the type, setting, and deployment of cameras, the collected image data generally vary in image resolutions, leading to large variations in the image quality as well as the appearance information. Additionally, for challenging scenarios such as suspect identification or lost people searching, acquiring a clear frontal person image as query seems intractable. Instead, the identification is required to be conducted only with a verbal description or a sketch given by profilers. To handle above challenging scenarios, person ReID conducted between image of multiple resolutions, and between text or sketch and photo, should be put more efforts in the future work.

Bibliography

[1] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. FREAK: Fast retina keypoint. In *IEEE conference on computer vision and pattern recognition*, pages 510–517, 2012.

[2] Marwan Ali Albahar, Abdulaziz Ali Albahr, and Muhammad Hatim Binsawad. An efficient person re-identification model based on new regularization technique. *IEEE access*, 8:171049–171057, 2020.

[3] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[5] Avasecurity. Introducing: Ava aware video management system. https://docs.video.avasecurity.com/Products/aware/aware.htm, 2020.

[6] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *IEEE conference on computer vision and pattern recognition*, pages 2530–2539, 2017.

[7] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krys Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017.

[8] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[9] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Number 4. Springer, 2006.

[10] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Proceedings of the international conference on neural information processing systems*, pages 161–168, 2008.

[11] David Breitkreutz and Kate Casey. Clusterers: a comparison of partitioning and density-based algorithms and a discussion of optimisations. 2008.

[12] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):43–57, 2010.

[13] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018.

[14] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792, 2010.

[15] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Hashgan: Deep learning to hash with pair conditional wasserstein gan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1287–1296, 2018.

[16] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[17] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the european conference on computer vision*, pages 132–149, 2018.

[18] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the international conference on neural information processing systems*, volume 33, pages 9912–9924, 2020.

[19] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE transactions on image processing*, 31:2352–2364, 2022.

[20] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 14960–14969, 2021.

[21] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 2004–2013, 2021.

[22] Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, and Ling Shao. Fast person re-identification via cross-camera semantic binary transformation. In *IEEE conference on computer vision and pattern recognition*, pages 3873–3882, 2017.

[23] Jiaxin Chen, Jie Qin, Yichao Yan, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. Deep local binary coding for person re-identification by delving into the

details. In *Proceedings of the ACM international conference on multimedia*, pages 3034–3043, 2020.

[24] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *IEEE international conference on computer vision*, pages 8351–8361, 2019.

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[26] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the international conference on neural information processing systems*, pages 2172–2180, 2016.

[27] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *IEEE international conference on computer vision*, pages 232–242, 2019.

[28] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for RGB-infrared person re-identification. In *IEEE international conference on computer vision*, pages 587–597, 2021.

[29] De Cheng, Jingyu Zhou, Nannan Wang, and Xinbo Gao. Hybrid dynamic contrast and probability distillation for unsupervised person re-id. *IEEE transactions on image processing*, 31:3334–3346, 2022.

[30] Ding Cheng, Xiaohong Li, Meibin Qi, Xueliang Liu, Cuiqun Chen, and Dawei Niu. Exploring cross-modality commonalities via dual-stream multi-branch network for infrared-visible person re-identification. *IEEE access*, 8:12824–12834, 2020.

[31] Yunzhou Cheng, Xinyi Li, Guoqiang Xiao, Wenzhuo Ma, and Xinye Gou. Dual-path deep supervision network with self-attention for visible-infrared person re-identification. In *IEEE international symposium on circuits and systems*, pages 1–5, 2021.

[32] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 7308–7318, 2022.

[33] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 10257–10266, 2020.

[34] Huangpeng Dai, Qing Xie, Yanchun Ma, Yongjian Liu, and Shengwu Xiong. RGB-infrared person re-identification via image modality conversion. In *IEEE international conference on pattern recognition*, pages 592–598, 2021.

[35] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the international joint conference on artificial intelligence*, page 2, 2018.

[36] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *IEEE conference on computer vision and pattern recognition*, pages 11864–11874, 2021.

[37] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. *arXiv preprint arXiv:2103.11568*, 2021.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[39] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[41] Guodong Ding, Salman Khan, Zhenmin Tang, Jian Zhang, and Fatih Porikli. Towards better validity: dispersion based clustering for unsupervised person re-identification. *arXiv preprint arXiv:1906.01308*, 2019.

[42] Lin Ding, Yonghong Tian, Hongfei Fan, Changhuai Chen, and Tiejun Huang. Joint coding of local and global deep features in videos for visual search. *IEEE transactions on image processing*, 29:3734–3749, 2020.

[43] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *IEEE conference on computer vision and pattern recognition*, pages 304–311, 2009.

[44] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[46] Yueqi Duan, Ziwei Wang, Jiwen Lu, Xudong Lin, and Jie Zhou. Graphbit: Bitwise interaction mining via deep reinforcement learning. In *IEEE conference on computer vision and pattern recognition*, pages 8270–8279, 2018.

[47] Yueqi Duan, Jiwen Lu, Ziwei Wang, Jianjiang Feng, and Jie Zhou. Learning deep binary descriptor with multi-quantization. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1924–1938, 2019.

[48] Fabian Dubourvieux, Romaric Audigier, Angelique Loesch, Samia Ainouz, and Stephane Canu. Unsupervised domain adaptation for person re-identification through source-guided pseudo-labeling. In *IEEE international conference on pattern recognition*, pages 4957–4964, 2021.

[49] Venice Erin Liong, Jiwen Lu, Gang Wang, Pierre Moulin, and Jie Zhou. Deep hashing for compact binary codes learning. In *IEEE conference on computer vision and pattern recognition*, pages 2475–2483, 2015.

[50] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996.

[51] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: clustering and fine-tuning. *ACM transactions on multimedia computing, communications, and applications*, 14(4):1–18, 2018.

[52] Xing Fan, Hao Luo, Chi Zhang, and Wei Jiang. Cross-spectrum dual-subspace pairing for rgb-infrared cross-modality person re-identification. *arXiv preprint arXiv:2003.00213*, 2020.

[53] Yujian Feng, Feng Chen, Yi-mu Ji, Fei Wu, and Jing Sun. Efficient cross-modality graph reasoning for rgb-infrared person re-identification. *IEEE signal processing letters*, 28:1425–1429, 2021.

[54] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE transactions on image processing*, 29:579–590, 2019.

[55] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *IEEE international conference on computer vision*, pages 11823–11832, 2021.

[56] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2938–2952, 2021.

[57] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *IEEE international conference on computer vision*, pages 6112–6121, 2019.

[58] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *arXiv preprint arXiv:1810.02936*, 2018.

[59] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020.

[60] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Proceedings of the international conference on neural information processing systems*, pages 11309–11321, 2020.

[61] IFSEC Global. The video surveillance reeport 2022. https://www.ifsecglobal.com/resources-1/the-video-surveillance-report-2022/, 2022.

[62] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[63] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2916–2929, 2012.

[64] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the international conference on neural information processing systems*, pages 2672–2680, 2014.

[65] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE international workshop on performance evaluation for tracking and surveillance*, 2007.

[66] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[67] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the international*

conference on neural information processing systems, volume 33, pages 21271–21284, 2020.

[68] Hongyang Gu, Jianmin Li, Guangyuan Fu, Chifong Wong, Xinghao Chen, and Jun Zhu. Autoloss-gms: searching generalized margin-based softmax loss function for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 4744–4753, 2022.

[69] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: An efficient clustering algorithm for large databases. *ACM sigmod record*, 27(2):73–84, 1998.

[70] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Proceedings of the international conference on neural information processing systems*, pages 5767–5777, 2017.

[71] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: dual part-aligned representations for person re-identification. In *IEEE international conference on computer vision*, pages 3642–3651, 2019.

[72] Yuchen Guo, Xin Zhao, Guiguang Ding, and Jungong Han. On trivial solution and high correlation problems in deep supervised hashing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[73] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE conference on computer vision and pattern recognition*, volume 2, pages 1735–1742, 2006.

[74] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. HSME: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8385–8392, 2019.

[75] Kaiming He, Fang Wen, and Jian Sun. K-means hashing: An affinity-preserving quantization method for learning binary compact codes. In *IEEE conference on computer vision and pattern recognition*, pages 2938–2945, 2013.

[76] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE international conference on computer vision*, pages 1026–1034, 2015.

[77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[78] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[79] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[80] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *IEEE conference on computer vision and pattern recognition*, pages 596–605, 2018.

[81] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *IEEE international conference on computer vision*, pages 15013–15022, 2021.

[82] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 879–887, 2022.

[83] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[84] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[85] Bingyu Hu, Jiawei Liu, and Zheng-jun Zha. Adversarial disentanglement and correlation network for rgb-infrared person re-identification. In *IEEE international conference on multimedia and expo*, pages 1–6, 2021.

[86] Nianchang Huang, Jianan Liu, Yunqi Miao, Qiang Zhang, and Jungong Han. Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review. *Information Fusion*, 91:396–411, 2023.

[87] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.

[88] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 8526–8536, 2021.

[89] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[90] Taeho Jo. *Machine Learning Foundations.* Springer, 2021.

[91] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 1062–1071, 2018.

[92] Kajal Kansal, A Venkata Subramanyam, Zheng Wang, and Shinichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE transactions on circuits and systems for video technology*, 30(10):3422–3432, 2020.

[93] George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[94] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[95] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[96] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(3):231–240, 2011.

[97] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, department of computer science, University of Toronto*, 2009.

[98] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. Simultaneous feature learning and hash coding with deep neural networks. In *IEEE conference on computer vision and pattern recognition*, pages 3270–3278, 2015.

[99] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE transactions on circuits and systems for video technology*, 30(4):1092–1108, 2019.

[100] Stefan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: Binary robust invariant scalable keypoints. In *IEEE international conference on computer vision*, pages 2548–2555, 2011.

[101] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4610–4617, 2020.

[102] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Proceedings of the european conference on computer vision*, pages 483–499, 2020.

[103] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[104] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[105] Wenkang Li, Ke Qi, Wenbin Chen, and Yicong Zhou. Bridging the distribution gap of visible-infrared person re-identification with modality batch normalization. In *IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 23–28, 2021.

[106] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, pages 2359–2367, 2017.

[107] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *IEEE international conference on computer vision*, pages 8090–8099, 2019.

[108] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *IEEE international conference on computer vision*, pages 7919–7929, 2019.

[109] Yunqiang Li and Jan van Gemert. Deep unsupervised image hashing by maximizing bit entropy. *arXiv preprint arXiv:2012.12334*, 2020.

[110] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE transactions on image processing*, 30:6392–6407, 2021.

[111] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

[112] Kevin Lin, Jiwen Lu, Chu-Song Chen, Jie Zhou, and Ming-Ting Sun. Unsupervised deep learning of compact binary descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1501–1514, 2018.

[113] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8738–8745, 2019.

[114] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *IEEE conference on computer vision and pattern recognition*, pages 3390–3399, 2020.

[115] Hefei Ling, Ziyang Wang, Ping Li, Yuxuan Shi, Jiazhong Chen, and Fuhao Zou. Improving person re-identification by multi-task learning. *Neurocomputing*, 347:109–118, 2019.

[116] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE transactions on multimedia*, 23:4414–4425, 2021.

[117] Jiachang Liu, Wanru Song, Changhong Chen, and Feng Liu. Cross-modality person re-identification via channel-based partition network. *Applied intelligence*, 52(3):2423–2435, 2022.

[118] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 4099–4108, 2018.

[119] Li Liu and Ling Shao. Sequential compact code learning for unsupervised image hashing. *IEEE transactions on neural networks and learning systems*, 27(12):2526–2536, 2015.

[120] Ran Liu, Mehdi Azabou, Max Dabagia, Chi-Heng Lin, Mohammad Gheshlaghi Azar, Keith Hengen, Michal Valko, and Eva Dyer. Drop, swap, and generate: A self-supervised approach for generating neural activity. In *Proceedings of the international conference on neural information processing systems*, volume 34, pages 10587–10599, 2021.

[121] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *IEEE conference on computer vision and pattern recognition*, pages 5187–5196, 2019.

[122] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2023.

[123] Zheng Liu, Jie Qin, Annan Li, Yunhong Wang, and Luc Van Gool. Adversarial binary coding for efficient person re-identification. In *IEEE international conference on multimedia and expo*, pages 700–705, 2019.

[124] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *IEEE transactions on image processing*, 28(8):3794–3807, 2019.

[125] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[126] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *IEEE conference on computer vision and pattern recognition*, pages 13379–13389, 2020.

[127] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the european conference on computer vision*, pages 282–297, 2018.

[128] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE transactions on multimedia*, 22(10):2597–2609, 2019.

[129] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, 1967.

[130] Djebril Mekhazni, Amran Bhuiyan, George Ekladious, and Eric Granger. Unsupervised domain adaptation in the dissimilarity space for person re-identification. In *Proceedings of the european conference on computer vision*, pages 159–174, 2020.

[131] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *IEEE international conference on computer vision*, pages 542–551, 2019.

[132] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65 (1-2):43–72, 2005.

[133] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Proceedings of the international conference on neural information processing systems*, pages 4826–4837, 2017.

[134] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

[135] James Moore. Use of automated facial recognition by south wales police deemed unlawful, court rules. https://www.ifsecglobal.com/video-surveillance/use-of-automated-facial-recognition-by-south-wales-police-deemed-unlawful-court-rules/, 2020.

[136] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[137] Daniil Osokin. Global context for convolutional pose machines. *arXiv preprint arXiv:1906.04104*, 2019.

[138] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *IEEE international conference on computer vision*, pages 12046–12055, 2021.

[139] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[140] Chris Price. Is AI really necessary for video surveillance? https://www.ifsecglobal.com/video-surveillance/is-ai-really-necessary-for-video-surveillance/, 2022.

[141] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *Proceedings of the ACM international conference on multimedia*, pages 2149–2158, 2020.

[142] Meibin Qi, Suzhi Wang, Guanghong Huang, Jianguo Jiang, Jingjing Wu, and Cuiqun Chen. Mask-guided dual attention-aware network for visible-infrared person re-identification. *Multimedia tools and applications*, 80(12): 17645–17666, 2021.

[143] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *European conference on computer vision*, pages 650–667, 2018.

[144] Zexuan Qiu, Qinliang Su, Zijing Ou, Jianxing Yu, and Changyou Chen. Unsupervised hashing with contrastive information bottleneck. In *Proceedings of the international joint conference on artificial intelligence*, pages 959–965, 2021.

[145] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[146] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[147] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[148] LKPJ Rdusseeun and P Kaufman. Clustering by means of medoids. In *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*, volume 31, 1987.

[149] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[150] Min Ren, Lingxiao He, Xingyu Liao, Wu Liu, Yunlong Wang, and Tieniu Tan. Learning instance-level spatial-temporal patterns for person re-identification.

In *IEEE international conference on computer vision*, pages 14930–14939, 2021.

[151] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018.

[152] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[153] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. ORB: An efficient alternative to sift or surf. In *IEEE international conference on computer vision*, page 2, 2011.

[154] SafeTrolley. How CCTV works. https://www.safetrolley.com/how-cctv-works/, 2020.

[155] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International journal of approximate reasoning*, 50(7):969–978, 2009.

[156] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE international conference on computer vision*, pages 618–626, 2017.

[157] Rhianna Sexton. Automatic facial recognition the debate in 2022. https://www.ifsecglobal.com/surveillance/automatic-facial-recognition-the-debate-in-2022/, 2022.

[158] Yuming Shen, Jie Qin, Jiaxin Chen, Mengyang Yu, Li Liu, Fan Zhu, Fumin Shen, and Ling Shao. Auto-encoding twin-bottleneck hashing. In *IEEE conference on computer vision and pattern recognition*, pages 2818–2827, 2020.

[159] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International conference on machine learning*, pages 20026–20040, 2022.

[160] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[161] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 1179–1188, 2018.

[162] Jingkuan Song, Tao He, Lianli Gao, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Binary generative adversarial networks for image retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[163] Christoph Strecha, Alex Bronstein, Michael Bronstein, and Pascal Fua. LDA-Hash: Improved matching with smaller descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):66–78, 2011.

[164] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE international conference on computer vision*, pages 3960–3969, 2017.

[165] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *European conference on computer vision*, pages 402–419, 2018.

[166] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European conference on computer vision*, pages 480–496, 2018.

[167] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[168] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *IEEE conference on computer vision and pattern recognition*, pages 7134–7143, 2019.

[169] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *IEEE conference on computer vision and pattern recognition*, pages 661–669, 2017.

[170] Tomasz Trzcinski and Vincent Lepetit. Efficient discriminative projections for compact binary descriptors. In *European conference on computer vision*, pages 228–242, 2012.

[171] Tomasz Trzcinski, Mario Christoudias, and Vincent Lepetit. Learning image descriptors with boosting. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):597–610, 2014.

[172] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Proceedings of the international conference on neural information processing systems*, volume 30, 2017.

[173] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[174] Pawan Vashisht. Role of artificial intelligence. *Security link india*, pages 66–69, 2020.

[175] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the international conference on neural information processing systems*, page 60006010, 2017.

[176] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the international conference on machine learning*, pages 1096–1103, 2008.

[177] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[178] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *IEEE conference on computer vision and pattern recognition*, pages 10981–10990, 2020.

[179] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE signal processing letters*, 25(7):926–930, 2018.

[180] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *IEEE international conference on computer vision*, pages 3623–3632, 2019.

[181] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8933–8940, 2019.

[182] Guanan Wang, Shaogang Gong, Jian Cheng, and Zengguang Hou. Faster person re-identification. In *European conference on computer vision*, pages 275–292, 2020.

[183] Jinpeng Wang, Ziyun Zeng, Bin Chen, Tao Dai, and Shu-Tao Xia. Contrastive quantization with code memory for unsupervised image retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2468–2476, 2022.

[184] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2764–2772, 2021.

[185] Pingyu Wang, Fei Su, Zhicheng Zhao, Yanyun Zhao, Lei Yang, and Yang Li. Deep hard modality alignment for visible thermal person re-identification. *Pattern pecognition letters*, 133:195–201, 2020.

[186] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *IEEE conference on computer vision and pattern recognition*, pages 8042–8051, 2018.

[187] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 2019.

[188] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 618–626, 2019.

[189] Longhui Wei, Xiaobin Liu, Jianing Li, and Shiliang Zhang. Vp-reid: Vehicle and person re-identification system. In *Proceedings of the ACM on international conference on multimedia retrieval*, pages 501–504, 2018.

[190] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[191] Xing Wei, Diangang Li, Xiaopeng Hong, Wei Ke, and Yihong Gong. Co-attentive lifting for infrared-visible person re-identification. In *Proceedings of the ACM international conference on multimedia*, pages 1028–1037, 2020.

[192] Ziyu Wei, Xi Yang, Nannan Wang, Bin Song, and Xinbo Gao. ABP: Adaptive body partition model for visible infrared person re-identification. In *IEEE international conference on multimedia and expo*, pages 1–6, 2020.

[193] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE transactions on neural networks and learning systems*, 33(9):4676–4687, 2022.

[194] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *IEEE international conference on computer vision*, pages 5380–5389, 2017.

[195] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. RGB-IR person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128(6):1765–1785, 2020.

[196] Gengshen Wu, Zijia Lin, Guiguang Ding, Qiang Ni, and Jungong Han. On aggregation of unsupervised deep binary descriptor with weak bits. *IEEE transactions on image processing*, 29:9266–9278, 2020.

[197] Lin Wu, Yang Wang, Hongzhi Yin, Meng Wang, and Ling Shao. Few-shot deep adversarial learning for video-based person re-identification. *IEEE transactions on image processing*, 29:1233–1245, 2019.

[198] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 4330–4339, 2021.

[199] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE transactions on information forensics and security*, 13(11):2884–2896, 2018.

[200] Yong Wu, Sizhe Wan, Di Wu, Chao Wang, Changan Yuan, Xiao Qin, Hongjie Wu, and Xingming Zhao. Position attention-guided learning for infrared-visible person re-identification. In *Intelligent computing theories and application*, pages 387–397, 2020.

[201] Yuhang Wu, Tengteng Huang, Haotian Yao, Chi Zhang, Yuanjie Shao, Chuchu Han, Changxin Gao, and Nong Sang. Multi-centroid representation network for domain adaptive person re-id. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2750–2758, 2022.

[202] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[203] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2014.

[204] Xuezhi Xiang, Ning Lv, Zeting Yu, Mingliang Zhai, and Abdulmotaleb El Saddik. Cross-modality person re-identification based on dual-path multi-branch network. *IEEE sensors journal*, 19(23):11706–11713, 2019.

[205] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016.

[206] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[207] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 11926–11935, 2021.

[208] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 4855–4864, 2021.

[209] Hantao Yao, Shiliang Zhang, Dongming Zhang, Yongdong Zhang, Jintao Li, Yu Wang, and Qi Tian. Large-scale person re-identification as retrieval. In *IEEE international conference on multimedia and expo*, pages 1440–1445, 2017.

[210] Hanrong Ye, Hong Liu, Fanyang Meng, and Xia Li. Bi-directional exponential angular triplet loss for rgb-infrared person re-identification. *IEEE transactions on image processing*, 30:1583–1595, 2020.

[211] Jianming Ye, Shiliang Zhang, Tiejun Huang, and Yong Rui. CDbin: Compact discriminative binary descriptor learned with efficient neural network. *IEEE transactions on circuits and systems for video technology*, 30(3):862–874, 2020.

[212] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[213] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the international joint conference on artificial intelligence*, pages 1092–1099, 2018.

[214] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *Proceedings of the ACM international conference on multimedia*, pages 347–355, 2019.

[215] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE transactions on information forensics and security*, 15:407–419, 2019.

[216] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE transactions on image processing*, 29:9387–9399, 2020.

[217] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the european conference on computer vision*, pages 229–247, 2020.

[218] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE transactions on information forensics and security*, 16:728–739, 2020.

[219] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *IEEE international conference on computer vision*, pages 13567–13576, 2021.

[220] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[221] Qingze Yin, Guanan Wang, Guodong Ding, Qilei Li, Shaogang Gong, and Zhenmin Tang. Rapid person re-identification via sub-space consistency regularization. *Neural processing letters*, pages 1–20, 2022.

[222] Aijing Yu, Haoxue Wu, Huaibo Huang, Zhen Lei, and Ran He. LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for nir-vis face recognition. *International journal of computer vision*, 129(5):1467–1483, 2021.

[223] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 13657–13665, 2020.

[224] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: augmented discriminative clustering for domain adaptive person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 9021–9030, 2020.

[225] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *Proceedings of the european conference on computer vision*, pages 594–611, 2020.

[226] Can Zhang, Hong Liu, Wei Guo, and Mang Ye. Multi-scale cascading network with compact feature learning for rgb-infrared person re-identification. In *IEEE international conference on pattern recognition*, pages 8679–8686, 2021.

[227] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE international conference on computer vision*, pages 5907–5915, 2017.

[228] Jingjing Zhang, Xiaohong Li, Cuiqun Chen, Meibin Qi, Jingjing Wu, and Jianguo Jiang. Global-local graph convolutional network for cross-modality person re-identification. *Neurocomputing*, 452:137–146, 2021.

[229] Liyan Zhang, Guodong Du, Fan Liu, Huawei Tu, and Xiangbo Shu. Global-local multiple granularity learning for cross-modality visible-infrared person re-identification. *IEEE transactions on neural networks and learning systems*, pages 1–11, 2021.

[230] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: feature-level modality compensation for visible-infrared person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 7349–7358, 2022.

[231] Ruimao Zhang, Liang Lin, Rui Zhang, Wangmeng Zuo, and Lei Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE transactions on image processing*, 24(12): 4766–4779, 2015.

[232] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1(2):141–182, 1997.

[233] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 3436–3445, 2021.

[234] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 8222–8231, 2019.

[235] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 7369–7378, 2022.

[236] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *IEEE transactions on image processing*, 29:1001–1015, 2019.

[237] Ziyue Zhang, Shuai Jiang, Congzhentao Huang, Yang Li, and Richard Yi Da Xu. RGB-IR cross-modality person reid based on teacher-student gan model. *Pattern recognition letters*, 150:155–161, 2021.

[238] Cairong Zhao, Yuanpeng Tu, Zhihui Lai, Fumin Shen, Heng Tao Shen, and Duoqian Miao. Salience-guided iterative asymmetric mutual hashing for fast person re-identification. *IEEE transactions on image processing*, 30:7776–7789, 2021.

[239] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017.

[240] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE international conference on computer vision*, pages 3219–3228, 2017.

[241] Yun-Bo Zhao, Jian-Wu Lin, Qi Xuan, and Xugang Xi. Hpiln: a feature learning framework for cross-modality person re-identification. *IET image processing*, 13:2897–2904, 2019.

[242] Zhiwei Zhao, Bin Liu, Qi Chu, Yan Lu, and Nenghai Yu. Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3520–3528, 2021.

[243] Feng Zheng and Ling Shao. Learning cross-view binary identities for fast person re-identification. In *Proceedings of the international joint conference on artificial intelligence*, pages 2399–2406, 2016.

[244] Huantao Zheng, Xian Zhong, Wenxin Huang, Kui Jiang, Wenxuan Liu, and Zheng Wang. Visible-infrared person re-identification: a comprehensive survey and a new setting. *Electronics*, 11(3):454, 2022.

[245] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 5310–5319, 2021.

[246] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE international conference on computer vision*, pages 1116–1124, 2015.

[247] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017.

[248] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE transactions on image processing*, 28(9):4500–4509, 2019.

[249] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 8371–8381, 2021.

[250] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE international conference on computer vision*, pages 3754–3762, 2017.

[251] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE international conference on computer vision*, pages 2138–2147, 2019.

[252] Xian Zhong, Tianyou Lu, Wenxin Huang, Mang Ye, Xuemei Jia, and Chia-Wen Lin. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE transactions on circuits and systems for video technology*, 32(3):1418–1430, 2021.

[253] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017.

[254] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE conference on computer vision and pattern recognition*, pages 598–607, 2019.

[255] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020.

[256] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2723–2738, 2020.

[257] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *IEEE international conference on computer vision*, pages 3702–3712, 2019.

[258] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision*, pages 2223–2232, 2017.

[259] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Proceedings of the european conference on computer vision*, pages 346–363, 2020.

[260] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *IEEE conference on computer vision and pattern recognition*, pages 6538–6547, 2020.

[261] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *arXiv preprint arXiv:1910.09830*, 2019.

[262] Maciej Zieba, Piotr Semberecki, Tarek El-Gaaly, and Tomasz Trzcinski. Bingan: Learning compact binary descriptors with a regularized gan. In *Proceedings of the international conference on neural information processing systems*, pages 3608–3618, 2018.