Towards unsupervised person re-identification

Hao Chen

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Vers la Ré-identification de Personnes Non-supervisée

Hao CHEN

INRIA Sophia Antipolis

Présentée en vue de l'obtention du grade de docteur en Informatique **d'**Université Côte d'Azur
Dirigée par : François Brémond
Co-encadrée par : Benoit Lagadec
Soutenue le : 12 Mai 2022

Devant le jury, composé de :
Président du jury :
Frédéric Jurie, Professeur, Université de Caen Normandie, SAFRAN, France
Rapporteurs :
Vittorio Murino, Professeur, Université de Vérone, IIT, Italie
Shiliang Zhang, Professeur Associé, Université de Pékin, Chine
Examinateurs :
Rémi Munos, Directeur de Recherche, INRIA Lille, Google DeepMind, France

Inria

Vers la Ré-identification de Personnes Non-supervisée

Towards Unsupervised Person Re-identification

Jury :

Président du jury

Frédéric Jurie, Professeur, Université de Caen Normandie, SAFRAN, France

Rapporteurs

Vittorio Murino, Professeur, Université de Vérone, IIT, Italie

Shiliang Zhang, Professeur Associé, Université de Pékin, Chine

Examinateurs

Rémi Munos, Directeur de Recherche, INRIA Lille, Google DeepMind, France

François Brémond, Directeur de Recherche, INRIA Sophia Antipolis, France

Vers la Ré-identification de Personnes Non-supervisée

Résumé

En tant que composant central des systèmes de vidéo-surveillance intelligents, la ré-identification de personnes (ReID) vise à rechercher une personne d'intérêt à travers des caméras qui ne se chevauchent pas. Malgré des améliorations significatives de la ReID supervisée, le processus d'annotation encombrant le rend moins évolutif dans les déploiements réels. De plus, comme les représentations d'apparence peuvent être affectées par des facteurs bruyants, tels que le niveau d'éclairage et les propriétés de la caméra, entre différents domaines, les modèles ReID de personnes subissent une baisse de performances importante en présence d'écarts de domaine. Nous sommes particulièrement intéressés par la conception d'algorithmes capables d'adapter un modèle ReID de personnes à un domaine cible sans supervision humaine. Dans un tel contexte, nous nous concentrons principalement sur la conception de méthodes d'adaptation de domaine non-supervisée et d'apprentissage de représentation non-supervisée pour le ReID de personnes.

Dans cette thèse, nous explorons d'abord comment construire des représentations robustes en combinant à la fois des caractéristiques globales et locales sous la condition supervisée. Ensuite, vers un système ReID adaptatif au domaine non-supervisé, nous proposons trois méthodes non-supervisées pour la ReID de personnes, notamment 1) la distillation des connaissances enseignant-étudiant avec des structures de réseau asymétriques pour encourager la diversité des caractéristiques, 2) un cadre d'apprentissage conjoint génératif et contrastif qui génère des vues augmentées avec un réseau génératif pour l'apprentissage contrastif, et 3) explorer les relations inter-instances et concevoir des fonctions de perte conscientes des relations pour une meilleure ReID de personnes basée sur l'apprentissage contrastif.

Nos méthodes ont été largement évaluées sur des benchmarks de ReID, tels que Market-1501, DukeMTMC-reID et MSMT17. Les méthodes proposées surpassent considérablement les méthodes précédentes sur les benchmarks de ReID, poussant considérablement la ReID de personnes vers des déploiements dans le monde réel.

Mots clés : ré-identification, adaptation de domaine non-supervisée, apprentissage non-supervisé de représentations

Towards Unsupervised Person Re-identification

Abstract

As a core component of intelligent video surveillance systems, person re-identification (ReID) targets at retrieving a person of interest across non-overlapping cameras. Despite significant improvements in supervised ReID, cumbersome annotation process makes it less scalable in real-world deployments. Moreover, as appearance representations can be affected by noisy factors, such as illumination level and camera properties, between different domains, person ReID models suffer a large performance drop in the presence of domain gaps. We are particularly interested in designing algorithms that can adapt a person ReID model to a target domain without human supervision. In such context, we mainly focus on designing unsupervised domain adaptation and unsupervised representation learning methods for person ReID.

In this thesis, we first explore how to build robust representations by combining both global and local features under the supervised condition. Then, towards an unsupervised domain adaptive ReID system, we propose three unsupervised methods for person ReID, including 1) teacher-student knowledge distillation with asymmetric network structures for feature diversity encouragement, 2) joint generative and contrastive learning framework that generates augmented views with a generative adversarial network for contrastive learning, and 3) exploring inter-instance relations and designing relation-aware loss functions for better contrastive learning based person ReID.

Our methods have been extensively evaluated on main-stream ReID datasets, such as Market-1501, DukeMTMC-reID and MSMT17. The proposed methods significantly outperform previous methods on the ReID datasets, significantly pushing person ReID to real-world deployments.

Keywords : re-identification, unsupervised domain adaptation, unsupervised representation learning

Acknowledgements

I have received countless help and encouragement throughout the three years of preparing this PhD dissertation.

First of all, I would like to thank my supervisor Francois Bremond, whose scientific expertise was invaluable in formulating research questions. Every time I raised an idea, his insightful feedback always helped me to deepen my understanding on the research topic. Without his guidance, I would not have been able to complete this dissertation. I am also thankful to my co-supervisor, Benoit Lagadec, for his consistent encouragement and support in the whole work.

I would like to acknowledge my colleagues from STARS team. I have received a lot of help during my on-boarding days at STARS team. The collaboration and discussion with my colleagues also brought my works to a higher level.

I would like to thank Université Côte d'Azur and Inria for providing resources and technical support for my research. The Nef computational cluster has been the main platform for conducting all my experiments. I would also like to thank European Systems Integration (ESI) and ANRT for funding my PhD. The support from ESI has significantly facilitated my research.

Last but not the least, I would like to thank my family. My parents and grandparents have always been supported my life decisions. Without their support, I probably would not have been able to study in France and complete my PhD research.

Contents

1 Introduction **1**
- 1.1 Definition . . . 2
- 1.2 Applications . . . 3
- 1.3 Challenges . . . 4
- 1.4 Motivation and Contributions . . . 6
- 1.5 Publication list . . . 7
- 1.6 Broader impact . . . 8
- 1.7 Thesis outline . . . 8

2 Literature review **10**
- 2.1 Overview . . . 10
- 2.2 Datasets and Evaluation metrics . . . 10
  - 2.2.1 Image person ReID datasets . . . 10
  - 2.2.2 Video person ReID datasets . . . 11
  - 2.2.3 Evaluation metrics . . . 12
- 2.3 Feature extractors . . . 12
  - 2.3.1 Handcrafted feature extractors . . . 12
  - 2.3.2 Deep neural networks . . . 14
- 2.4 Loss functions . . . 18
- 2.5 Data augmentation . . . 20
  - 2.5.1 Image-level data augmentation . . . 20
  - 2.5.2 Feature-level data augmentation . . . 21
  - 2.5.3 GAN-based data augmentation . . . 22
- 2.6 Unsupervised person ReID . . . 23
  - 2.6.1 Unsupervised domain adaptive ReID . . . 23
  - 2.6.2 Fully unsupervised ReID . . . 24
- 2.7 Conclusion . . . 25

3 Spatial-channel partition for supervised object ReID **27**
- 3.1 Introduction . . . 27
- 3.2 Study of Appearance Representations . . . 29
  - 3.2.1 State-of-the-art . . . 29
  - 3.2.2 Study of Partitioned Representations . . . 31
- 3.3 Proposed Framework . . . 34
  - 3.3.1 Spatial and Channel Partition Representation Network (SCR) . . . 34

3.3.2 Loss Functions . . . 34
3.4 Experiments . . . 35
3.4.1 Implementation Details . . . 35
3.4.2 Datasets and Protocols . . . 36
3.4.3 Ablation Studies . . . 36
3.4.4 Comparison with State-of-the-art . . . 39
3.5 Conclusion . . . 40

4 Asymmetric branches for unsupervised person ReID 43
4.1 Introduction . . . 43
4.2 Related Work . . . 44
4.3 Proposed Method . . . 46
4.3.1 Overview . . . 46
4.3.2 Asymmetric branches . . . 47
4.3.3 Asymmetric Branched Mean Teaching . . . 47
4.4 Coupling Problem in Mean Teacher Based Methods . . . 50
4.5 Experiments . . . 51
4.5.1 Datasets and Evaluation Protocols . . . 51
4.5.2 Implementation details . . . 51
4.5.3 Comparison with State-of-the-Art Methods . . . 52
4.5.4 Ablation Studies . . . 56
4.6 Conclusion . . . 56

5 Joint generative and contrastive learning for unsupervised person ReID 58
5.1 Introduction . . . 58
5.2 Related Work . . . 60
5.3 Proposed Method . . . 62
5.3.1 View Generator (Generative Module) . . . 62
5.3.2 View Contrast (Contrastive Module) . . . 64
5.3.3 Joint Training . . . 65
5.4 Experiments . . . 65
5.4.1 Datasets and Evaluation Protocols . . . 65
5.4.2 Implementation Details . . . 66
5.4.3 Unsupervised ReID Evaluation . . . 68
5.4.4 Generation Quality Evaluation . . . 69
5.5 Conclusions . . . 74

6 Inter-instance Contrastive Encoding for unsupervised person ReID 78
6.1 Introduction . . . 78
6.2 Related Work . . . 79
6.3 Proposed Method . . . 80
6.3.1 Overview . . . 80
6.3.2 Proxy Centroid Contrastive Baseline . . . 82
6.3.3 Hard Instance Contrastive Loss . . . 84
6.3.4 Soft Instance Consistency Loss . . . 85

6.4 Experiments 87
6.4.1 Datasets and Evaluation Protocols 87
6.4.2 Implementation details 87
6.4.3 Parameter analysis 89
6.4.4 Ablation study 90
6.4.5 Comparison with state-of-the-art methods 91
6.5 Conclusion 94

7 Conclusion and Perspective **95**
7.1 Contributions 95
7.2 Limitations 96
7.3 Perspectives 98

List of Figures

1.1 Data source: CCTV cameras and social media. Images are free images from Pixabay [115]. 1
1.2 A person ReID system includes a) Camera network, b) Person detection and c) Person Re-identification. 2
1.3 Person ReID seeks the most similar images of a query person in the gallery set. 3
1.4 Challenges in Person ReID. 5

2.1 An covariance handcrafted feature space [5]. Every covariance is extracted from a region (P), distance layer (D) and three channel functions (*e.g.*, bottom covariance feature is extracted from region P3 using layers: D, I-intensity, ∇_I -gradient magnitude and θ_I -gradient orientation). 13
2.2 An example [149] of color histogram from a person bounding box and a segmented person foreground. 14
2.3 Comparison between VGG-19 [108], a 34-layer plain network and a 34-layer residual network [49]. 15
2.4 An example [93] of combining CNN and RNN for video-based person ReID. 16
2.5 Left: Transformer [123]. Right: Multi-head attention in a Transformer, where V, K, Q are feature matrix from the last layer. . 17
2.6 Comparison between (a) cross-entropy loss and (b) batch-hard triplet loss. 19
2.7 Comparison between (a) image part-based model [149] and (b) feature map part-based model [118]. 21
2.8 Three main approaches for unsupervised person ReID: (a) Adversarial learning, (b) Metric learning and (c) Pseudo labelling. 24
2.9 Left: conventional contrastive learning. Right: Our proposed joint generative and contrastive learning. 25

3.1 Example of a spatial-channel partition. H, W and C stand for respectively Height, Width and Channel in a deep feature map. In this example, we partition a whole feature map into two spatial parts (upper body and lower body) and two channel groups. 28

3.2 Comparisons of saliency maps generated by Grad-CAM [107] applied on 4 CNN models on Market-1501 test set. (a): A ResNet-50 w/o partitions nor attention mechanism. (b) to (e): A ResNet-50 w/ spatial-channel partition, where (b) and (c) are saliency maps on two spatial parts after spatial partition, (d) and (f) are saliency maps on two channel groups after channel partition. (f): Squeeze-and-Excitation Network [56]. (g): Residual Attention Network [126]. 30

3.3 Spatial and Channel Partition Representation network. For the backbone network, we duplicate layers after conv4_1 into 3 identical but independent branches that generate 3 feature maps "p1", "p2" and "p3". Then, multiple spatial-channel partitions are conducted on the feature maps. "s2" and "c2" refer to 2 spatial parts and 2 channel groups. "s3" and "c3" refer to 3 spatial parts and 3 channel groups. After global max pooling (GMP), dimensions of global (dim = 2048) and local (dim = 2048, 1024*2 and 683*2+682) features are unified by 1*1 convolution (1*1 Conv) and batch normalization (BN) to 256. Then, fully connected layers (FC) give identity predictions of input images. All the dimension unified feature vectors (dim = 256) are aggregated together as appearance representation (Rep) for testing. 33

3.4 Examples of several mismatched samples in PCB [118] on Market-1501 dataset, which are addressed by our proposed SCR. Red borders refers to mismatched samples. "#1", "#2" and "#3" correspond to top 3 retrieved gallery samples. 42

4.1 Source domain pre-training for asymmetric branched network. One ResNet bottleneck block corresponds to three convolutional layers. For UDA setting, inputs are labelled images from source training set. GAP refers Global Average Pooling, while GMP refers to Global Max Pooling. FC refers to Fully Connected layer. 46

4.2 ABMT adaptation. For UDA setting, inputs are training set images from both source and target domains. For fully unsupervised setting, inputs are unlabeled images from target training set. 48

4.3 Comparison between (a) Mean Teacher Baseline (b) Mutual Mean Teaching [39] and (c) our Mean Teacher with cross-branch supervised asymmetric branches. Teacher network is formed by exponential moving average (EMA) values of student network. 50

4.4 Distance comparison between features extracted from a ResNet50 backbone on all samples in DukeMTMC-reid training set for Market $\rightarrow$ Duke task. **Left**: Feature distance between two teacher models in MMT and between two teacher branches in our proposed method. **Right**: Feature distance between teacher and student networks. 51

4.5 Examples of retrieved most similar 5 images in Market → Duke task from MMT [39] and our proposed method. Given a query image, different identity images are highlighted by red bounding boxes, while same identity images are highlighted by green bounding boxes. 55

5.1 **Left**: Traditional self-supervised contrastive learning maximizes agreement between representations (f_1 and f_2) of augmented views from Data Augmentation (DA). **Right**: Joint generative and contrastive learning maximizes agreement between original and generated views. 59

5.2 **(a) General architecture of GCL**: Generative and contrastive modules are coupled by the shared identity encoder E_{id}. **(b) Generative module**: The decoder G combines the identity features encoded by E_{id} and structure features E_{str} to generate a novel view x'_{new} with a cycle consistency. **(c) Contrastive module**: View-invariance is enhanced by maximizing the agreement between original $E_{id}(x)$, synthesized $E_{id}(x'_{new})$ and memory f_{pos} representations. 62

5.3 Example images as generated by the View Generator via 3D mesh rotation based on left input image. 63

5.4 Cluster number curve on Market-1501. TDA denotes traditional data augmentation, including random flipping, cropping, jittering, erasing. 70

5.5 More qualitative ablation study on the view-invariant losses. For simplicity, $\mathcal{L}_{vi}$ denotes three view-invariant losses $\mathcal{L}_{vi}+\mathcal{L}'_{vi}+\mathcal{L}''_{vi}$, which helps E_{id} to extract better identity features (white shirt). 71

5.6 Qualitative ablation study on the view-invariant losses. For simplicity, $\mathcal{L}_{vi}$ denotes three view-invariant losses $\mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}$, which helps E_{id} to extract view-invariant features (red shirt). 72

5.7 Comparison of the generated images on Market-1501 dataset. $\star$ refers to methods without sharing source code, whose examples are cropped from their papers. Examples of FD-GAN, IS-GAN, DG-Net and GCL are generated from six real images shown in the figure. 72

5.8 Generated novel views on the three datasets. 73

5.9 Linear interpolation on identity features. Identity features are swapped between left and right persons. 74

5.10 Examples of generated novel views on **Market-1501** training and test sets. 75

5.11 Examples of generated novel views on **DukeMTMC-reID** training and test sets. 76

5.12 Examples of generated novel views on **MSMT17** training and test sets. 77

6.1 General architecture of ICE. We maximize the similarity between anchor and pseudo positives in both inter-class (proxy agreement between an instance representation f_1 and its cluster proxy p_1) and intra-class (instance agreement between f_1 and its pseudo positive m_2) manners. 81
6.2 Proxy contrastive loss. Inside a cluster, an instance is pulled to a cluster centroid by $L_{agnostic}$ and to cross-camera centroids by L_{cross}. 83
6.3 Comparison between triplet and hard instance contrastive loss. 85
6.4 Based on inter-instance similarity ranking between anchor (A), pseudo positives (P) and pseudo negatives (N), **Hard Instance Contrastive Loss** matches an anchor with its hardest positive in a mini-batch. **Soft Instance Consistency Loss** regularizes the inter-instance similarity before and after data augmentation. 86
6.5 Parameter analysis on Market-1501 dataset. 87
6.6 Dynamic cluster numbers during 40 training epochs on DukeMTMC-reID. “hard" and “soft" respectively denote L_{h_ins} and L_{s_ins}. A lower number denotes that clusters are more compact. 89
6.7 Dynamic KL divergence during 40 training epochs on DukeMTMC-reID. Lower KL divergence denotes that a model is more robust to augmentation perturbation. 90
6.8 Dynamic cluster numbers of **ICE(agnostic)** during 40 training epochs on DukeMTMC-reID. A lower number denotes that clusters are more compact (less intra-cluster variance). 92
6.9 T-SNE visualization of 10 random classes in DukeMTMC-reID test set between **camera-aware baseline (Left)** and **ICE (Right)**. 92
6.10 Comparison of top 5 retrieved images on Market1501 between CAP [130] and ICE. Green boxes denote correct results, while red boxes denote false results. Important visual clues are marked with red dashes. 94

List of Tables

2.1 Representative image-based person ReID datasets. PersonX and UnrealPerson are synthetic datasets built on game engines (Unity and UnrealEngine 4). 11
2.2 Representative video-based person ReID datasets. 12
2.3 Id-related and Id-unrelated factors in a person image. 23

3.1 Comparison of results (%) between attention and partition on Market-1501 dataset. SENet refers to Squeeze-and-Excitation Network [56]. Spatial-channel partition refers to the model trained with 2 spatial parts and 2 channel groups. 32
3.2 Performance comparison (%) of different partition types (spatial partition, channel partition and spatial-channel partition) on CUHK03 dataset using the new protocol [170] where the bold font denotes the best partition type. "s2" and "s3" refer that the entire feature map is partitioned into 2 and 3 spatial parts, while "c2" and "c3" refer respectively to 2 and 3 channel groups. 37
3.3 Performance comparison (%) of the proposed SCR with different number of branches where the bold font denotes the best architecture. "p1", "p2" and "p3" refer to 3 feature maps in SCR. "s" and "c" represent "spatial" and "channel" respectively, followed by the number of parts. For instance, "s2" and "c2" refer that the entire feature map is partitioned into 2 spatial parts and 2 channel parts. 37
3.4 Performance comparison (%) of training SCR with different parameter values for λ from L_{total}. The bold font denotes the best parameter. 37
3.5 Comparison of different temporal pooling strategies where the bold font denotes the best method. "R" and "E" refer respectively to Random Sampling and Even Sampling. "TP" refers to conventional Temporal Pooling. "TPP" refers to Temporal Partiton Pooling. 38
3.6 Comparison of supervised results (%) on Market-1501 and DukeMTMC-reID dataset. 39
3.7 Comparison of supervised results (%) on CUHK03 dataset using the new protocol [170] 40
3.8 Comparison of supervised results (%) on MARS dataset. . . . 40

3.9 Comparison of unsupervised cross-domain results (%). M → D refers to training on Market-1501 and testing on DukeMTMC-reID. D → M refers to training on DukeMTMC-reID and testing on Market-1501. 41

4.1 Comparison of unsupervised domain adaptation (UDA) Re-ID methods (%) on medium-to-medium datasets (Market→ Duke and Duke → Market) and medium-to-large datasets (Market → MSMT and Duke → MSMT). 52
4.2 Comparison of unsupervised Re-ID methods (%) with a ResNet50 backbone on Market and Duke datasets. * refers to our implementation where we remove the source pre-training step. DBSCAN refers to a DBSCAN clustering based on re-ranked distance. 53
4.3 Ablation studies with ResNet50 backbone. MT-Baseline corresponds to the Mean Teacher Baseline in Figure 4.3 (a) with a ResNet-50. K-Means refers to a K-Means++ clustering whose cluster number is set to 500. AB refers to asymmetric branches. DBSCAN refers to a DBSCAN clustering [31]. 53
4.4 Ablation studies on structure of asymmetric branches. 54
4.5 Ablation studies on loss functions. 54

5.1 Comparison of unsupervised ReID methods (%) with a ResNet50 backbone on Market and Duke datasets. We test our proposed method on several baselines, whose names are in brackets. * refers to our implementation based on authors' code. 66
5.2 Comparison of unsupervised Re-ID methods (%) with a ResNet50 backbone on MSMT17. * refers to our implementation based on authors' code. 67
5.3 Ablation study on loss functions used in two modules. (1). $\mathcal{L}_{gan}$ corresponds to generation w/o contrast. (2). $\mathcal{L}_{vi}^{woGAN}$ corresponds to contrast w/o generation. TDA denotes traditional data augmentation. (3). $\mathcal{L}_{gan} + \mathcal{L}_{vi}$ ($\mathcal{L}_{vi}^{'}$ and $\mathcal{L}_{vi}^{''}$) correspond to joint generative and contrastive learning. 70
5.4 Comparison of FID (lower is better) and SSIM (higher is better) on Market-1501 dataset. U denotes the fully unsupervised setting. UDA denotes Duke→Market setting. 70

6.1 Augmentation settings for 3 losses. 82
6.2 Comparison between using the hardest negative and all negatives in the denominator of $\mathcal{L}_{h\_ins}$. 85
6.3 Comparison of consistency loss. Ours refers to KL divergence between images with and without data augmentation. 87
6.4 Comparison of ResNet50 and IBN-ResNet50 backbones on Market1501, DukeMTMC-reID and MSMT17 datasets. 87
6.5 Comparison of different distance thresholds on Market1501, DukeMTMC-reID and MSMT17 datasets. 88

6.6 Comparison of different losses. Camera-aware memory occupies up to 6, 8 and 15 times memory space than camera-agnostic memory on Market1501, DukeMTMC-reID and MSMT17 datasets. 89
6.7 Comparison of ReID methods on Market1501, DukeMTMC-reID and MSMT17 datasets. The best and second best unsupervised results are marked in red and blue. 93

Liste des abréviations

AI	artificial intelligence
CCTV	closed-circuit television
ReID	re-identification
UAV	unmanned aerial vehicle
UDA	unsupervised domain adaptation
GAN	generative adversarial network
RGB	red green blue
HSV	hue saturation lightness
DNN	deep neural network
CNN	convolutional neural network
RNN	recurrent neural network
NLP	natural language processing

Chapter 1

Introduction

Computer vision is a field of Artificial Intelligence (AI) that enables machines to get high-level understanding from visual inputs, such as images and videos. Based on the acquired visual understanding, a computer vision system is supposed to give optimal responses, which allows for automating machine tasks. Explosive growth of image data volume that recorded by closed-circuit television (CCTV) cameras or unloaded by users to social media has been witnessed in recent years, as shown in Figure 1.1. For example, from the end of 2013 to the beginning of 2020, the number of CCTV cameras in the 50 most populous cities in France has multiplied by 2.4 from nearly 4,800 cameras to more than 11,400 [71]. On the other hand, more than 500 hours of video were uploaded to YouTube by more than two billion monthly users around the world every minute in 2019 [152]. How to properly use such huge amount of image and video data to serve people remains a challenge for computer vision researchers.

Recent computer vision researches cover almost every task that can be realized by a human, such as object detection, image classification, image segmentation, object tracking, image and video retrieval, action recognition and image synthesis. To realize a complex real-world project, people usually need to combine several visual tasks. For example, an intelligent video surveillance

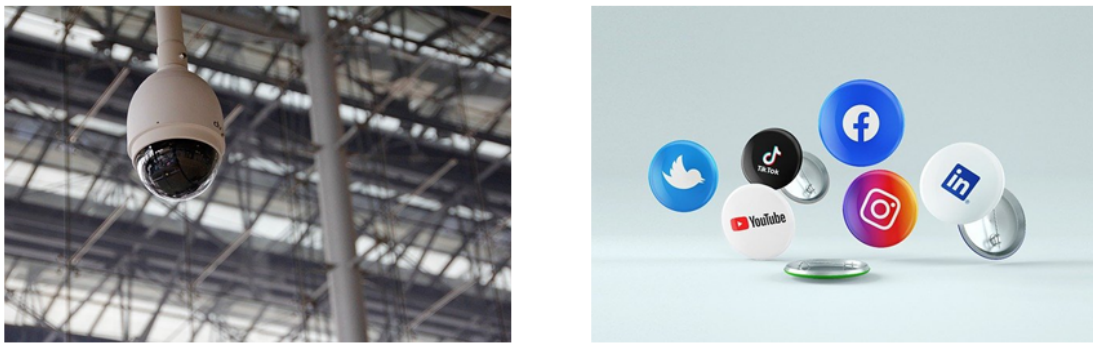
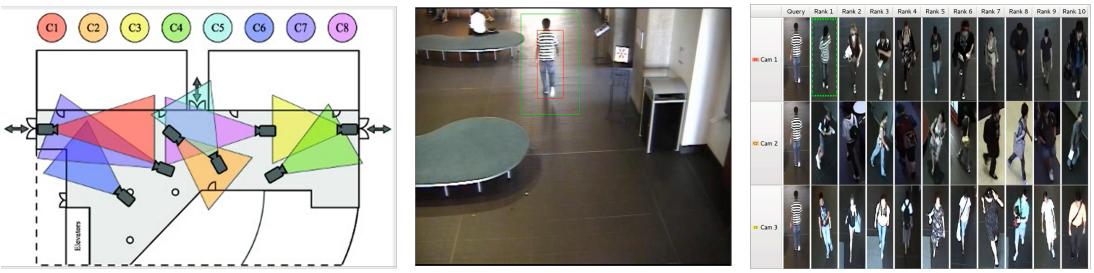


Figure 1.1: Data source: CCTV cameras and social media. Images are free images from Pixabay [115].



a) Camera network b) Person detection c) Person re-identification

Figure 1.2: A person ReID system includes a) Camera network, b) Person detection and c) Person Re-identification.

system usually relies on a human detector to tell there is a person, a tracker to know the trajectory of a person within a camera view, a person re-identification (ReID) algorithm to associate a target person across different cameras and an action recognition algorithm to tell if there is a dangerous event. As one of the most important components in computer vision systems, person ReID helps a machine to re-identify a target person, regardless of the pose of the target person and environmental noise.

In this thesis, we focus on the problem of person ReID. Our objective is to build image representations that are identity-discriminative and invariant to identity-unrelated factors, such as pose, view-point and illumination condition. We first use human-annotated labels to build such identity representations for supervised person ReID. As annotating cross-camera person images is a cumbersome task for real-world deployments, we are particularly interested in getting rid of human supervision in person ReID. We gradually move our focus to unsupervised person ReID.

In this section, we go through the definition, applications and challenges of person ReID. We proceed to present a list of our contributions in person ReID and an outline of this thesis.

1.1 Definition

Re-identification is a research topic about matching similar objects by ranking object appearance representations, which is essential for the tracking of moving objects. When applied to moving persons, person ReID aims at searching the same person across non-overlapping cameras (called the gallery) with a given query. Recent person ReID is defined as the human association task on the bounding boxes drawn by a person detection algorithm. Thus, an image-based person ReID system consists of a person detection algorithm and a person ReID algorithm inside a network of cameras, as shown in Figure 1.2. For video-based person ReID, a human tracker should be inserted between the detection part and the ReID part to generate tracklets.

In the research community, person ReID researchers usually focus on bound-

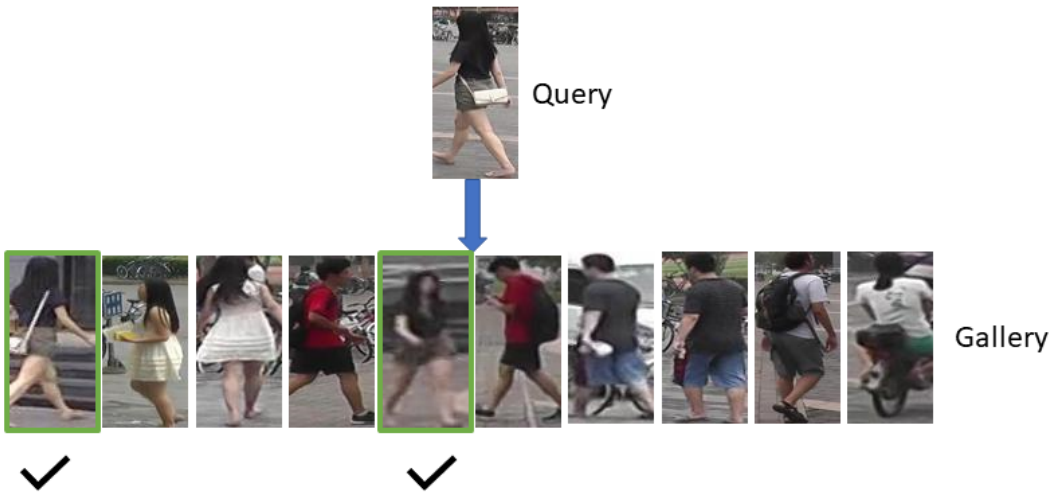


Figure 1.3: Person ReID seeks the most similar images of a query person in the gallery set.

ing box matching, while detection researchers focus on generating high-quality bounding boxes. A major assumption for person ReID research is that an individual usually keeps the same clothes in a short time period. Based on this assumption, person ReID can be defined as a fine-grained image retrieval task that ranks gallery images or tracklets based on the visual similarity between query and gallery sets, as shown in Figure 1.3. Thus, the key point for a robust person ReID system is to train a feature extractor in the training set that can build discriminative identity representations for unseen query and gallery images or tracklets.

Different from the conventional identification tasks, such as face recognition and finger print verification, person ReID does not rely on face and finger print information to identity a specific person, but to re-identify a person by matching whole-body appearances. As CCTV cameras are usually installed in some fixed positions, occlusions, low resolution and undesired camera view-point are inevitable, making face and finger print information not always accessible. In such context, person ReID becomes indispensable to permit an intelligent video surveillance system to keep running. In contrast to face recognition that only builds face representations, person ReID encodes more body parts into appearance representations, which are more robust in real-world deployments.

1.2 Applications

Person ReID shows significant impact for numerous real-world applications in security, retail, healthcare, person search and human-robot interaction.

Security. When certain persons move from one camera view to another, person ReID can establish the correspondence between different persons for cross-

camera tracking. Such system can be installed in important places, such as airports, train stations and stadiums for security monitoring.

Retail. Using a person ReID system, shopkeepers can provide personalized recommendations for customers in order to improve customers' shopping experience. Recent cashier-less stores can also benefit from person ReID systems to re-identify a customer and finalize an automatic checkout. Moreover, person ReID algorithms can be extended to general object ReID. Given a target merchandise photo, such as clothing or vehicle, ReID researches help to conduct an accurate image matching for a better merchandise recommendation system.

Healthcare. Geriatric healthcare researchers analyze elder people's behavior to make more accurate diagnoses. To automatically re-identify a certain person inside a nursing home, a person ReID system is also needed.

Person search. In a city-level video surveillance system, CCTV cameras generally cover large spatial areas. Person ReID algorithms can be applied to find missing children, as well as seniors with Alzheimer's or some other forms of dementia.

Human-robot interaction. Person ReID has gained increasing attention in robotic vision research. A personal assistant robot relies on person ReID algorithms to recognize and give optimal personalized responses to each nearby person.

Traffic modeling. Intelligent transportation systems are acknowledged as a key component for smart cities. When applied on passengers and vehicles, ReID algorithms help city planners to get a better understanding of traffic flow. Based on the acquired traffic models, city planners can improve the transportation system, such as road width and distance between two bus stations.

1.3 Challenges

To get optimal performance in previous mentioned applications, ideal identity representations are supposed to have maximal intra-identity similarities and minimal inter-identity similarities. However, there are multiple real-world constraints (see Figure 1.4) that make it challenging to build robust identity representations for deployments.

Camera view-point. Due to fixed camera location and unfixed person trajectory, person images recorded from different cameras are usually in different view-points. When people wear clothes with inconsistent front-back color or carry a bag, the front, side and back views of a same person recorded by different street cameras can be dramatically different. Furthermore, with the

Figure 1.4: Challenges in Person ReID.

popularization of unmanned aerial vehicles (UAV), the top-to-side view ReID raises a new challenge.

Person pose. Inside a network of cameras, person images can be recorded in different poses. For example, in a university campus, there are people walking in a street, sitting on benches and riding bicycles. Pose variance is a factor that affects appearance representations. Different persons in a similar pose can be more visually similar, while images of a same person in different poses can be less similar.

Occlusion. In a crowded street, a target person can be easily occluded by other persons or objects. An occluded person image only contain incomplete appearance features, leading to a less robust representation. Given representations built from occluded person images, query-to-gallery matching is prone to fail.

Imperfect bounding box. Recent person ReID research is usually defined as a image retrieval task on cropped person bounding boxes, which are generated by human detectors. The quality of person bounding boxes strongly depends on the performance of detection algorithms. An imperfect bounding box can easily result in incomplete appearances, which degrade the person ReID performance.

Low resolution. In the real-world deployments, high resolution cameras could be expensive and also cause problems for storage systems. Depending on the distance between a camera and a target person, the resolution of a person bounding box is usually unsteady. A low resolution person image can lose detailed appearance information, making it hard for a person ReID system to associate a same person.

Illumination. Illumination level can more or less affect matching appearance representations, when some of them are brighter or darker. As a network of cameras can be installed across in-door and out-door scenes, illumination conditions can vary from camera to camera, which leads to camera style variance for a dataset. In addition, illumination level also depends on the recording time of a day as well as the weather of a day.

Domain gap. A domain (dataset) is usually recorded during a short time period, which has specific cloth style and illumination level. For example, Market-1501 dataset [164] is recorded in summer, while DukeMTMC-reID dataset [105] is recorded in winter. As a consequence, person in Market-1501 usually wear summer clothes, and images are also brighter. On the contrary, person in DukeMTMC-reID usually wear winter clothes, and images are darker. Since person ReID is an appearance-sensitive task, a ReID model trained on a domain has proven to be hard to generalize to other domains.

Insufficient data. Data volume and data diversity are core problems for data-driven deep learning models. A model pretrained on a large-scale diversified dataset usually has stronger generalizability. However, due to hardware and privacy constraints, it is always difficult to have enough data that can cover all above-mentioned variations. Designing proper data augmentation techniques to virtually increase data volume can be a possible solution.

Insufficient annotation. To adapt a pretrained ReID model to a new domain, a straightforward solution is to annotate all images in the new domain and re-train the model. However, annotating cross-camera person images is a cumbersome and time-consuming task, which strongly limited the scalability of real-world deployments. On the other hand, inspired by ImageNet [26] pretraining, several large-scale unlabeled [35] and weakly labeled [127] have been recently proposed to pre-train ReID model. How to efficiently pretrain a generalizable ReID model on large-scale datasets with insufficient annotation remains a challenging problem.

All these described challenges influence the performance of a person ReID system. In the next section, we present our motivation and contributions to handle these challenges.

1.4 Motivation and Contributions

Among earlier mentioned challenges in person ReID, we are particularly interested in tackling domain gap, insufficient data and insufficient annotation issues. Domain gap and insufficient annotation are mainly related to unsupervised person ReID, including unsupervised domain adaption (UDA) and fully unsupervised settings. Insufficient data is a general problem that has been considered in both supervised and unsupervised person ReID. We argue that enhancing data diversity is always beneficial for person ReID. After

seeing enough diversified training data, the trained ReID model can be more noise-invariant for inference.

In this thesis, our motivation is to reduce human supervision and push person ReID to real-world deployments. We argue that designing proper data augmentation techniques allows for learning invariant features from mimicked variance for person ReID. The data augmentation techniques can be realized at either image level or feature level. Conventional data augmentation techniques usually introduce mimicked distortions at image level by basic positional transformations (scaling, cropping, flipping, padding, rotation, translation, *etc*.) or color transformations (color jittering, Gaussian noise, Gaussian blur, *etc*.). Differently, we design data augmentation in feature level to combine with the conventional data augmentation techniques. We also design advanced image-level augmentation with an generative adversarial network (GAN) as a substitute for conventional data augmentation techniques. Our main contributions are listed as follows:

1. Extracting local features to complete global features can enrich feature diversity in a person appearance representation. We propose spatial-channel partition to build robust representations with both global and local features for supervised person ReID.
2. We propose asymmetric branches as a feature-level augmentation technique to encourage feature diversity and alleviate teacher-student consensus for unsupervised person ReID.
3. We propose a 3D mesh guided generator as an image-level augmentation technique to provide augmented views by conditionally modifying disentangled id-unrelated features for unsupervised person ReID.
4. We propose to regularize the instance-to-instance similarity before and after applying data augmentation techniques to learn robust representations for unsupervised person ReID.

1.5 Publication list

Our contributions have been published in major computer vision conferences, as follows:

1. H. Chen, B. Lagadec and F. Bremond. Partition and Reunion: A Two-Branch Neural Network for Vehicle Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
2. H. Chen, B. Lagadec and F. Bremond. Learning discriminative and generalizable representations by spatial-channel partition for person re-identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.
3. H. Chen, B. Lagadec and F. Bremond. Enhancing Diversity in Teacher-Student Networks via Asymmetric branches for Unsupervised Person

Re-identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

4. H. Chen*, Y. Wang*, B. Lagadec, A. Dantcheva and F. Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. (* Co-first authors)
5. H. Chen, B. Lagadec and F. Bremond. ICE: Inter-instance Contrastive Encoding for Unsupervised Person Re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

1.6 Broader impact

The recent rapid development of AI algorithms, such as deepfake generation and video surveillance, has led to increasing ethical and societal concerns about the possible misuse of these algorithms.

Our contributions significantly push person ReID to real-world deployments in intelligent security protection, intelligent retail and intelligent transportation systems. In addition, our contributions on person ReID explores how to build fine-grained person identity representations in both supervised and unsupervised settings, which can be extended to more general research topics, such as fine-grained classification, unsupervised learning and unsupervised domain adaptation.

Despite the importance in earlier mentioned applications, ReID algorithms may raise privacy concerns, given that a street camera records every person and vehicle that pass in front of the camera. Most of persons can be unaware of being recorded. To prevent this situation from happening, governments and law makers should enact corresponding regulations and laws to control the deployments of surveillance cameras and the usage of ReID algorithms.

1.7 Thesis outline

This thesis is composed of 7 Chapters, in which Chapter 3, 4 , 5 and 6 correspond to details of our contributions.

Chapter 1: Introduction. In the first chapter, we begin with the definition of person ReID. Next, we highlight the significance of conducting person ReID research with numerous person ReID applications. We list the challenges in person ReID and briefly summarize our contributions that are targeted at addressing specific challenges.

Chapter 2: Literature Review. In the second chapter, a general literature review is given on person ReID and unsupervised representation learning. For

person ReID, we present representative features extractors, data augmentation, datasets and metrics. On the other hand, for unsupervised representation learning, we present representative methods in adversarial learning, metric learning, pseudo labeling, clustering and contrastive learning, which are related to unsupervised person ReID.

Chapter 3: Spatial-channel partition for supervised peron and vehicle ReID. In the third chapter, we provide details of our first contribution. We study the difference between attention mechanism and feature map partition. Based on the study, we propose to partition a feature map into spatial strips and channel groups and combine global and portioned local features as robust identity representations.

Chapter 4: Asymmetric branches for unsupervised person ReID. In the fourth chapter, we create two different versions of global feature based representations to enhance feature-level diversity. Instead of feature map partitions, we propose asymmetric branches with different depth and pooling methods to alleviate the consensus in teacher-student networks.

Chapter 5: Joint generative and contrastive learning for unsupervised person ReID. In the fifth chapter, a 3D mesh guided GAN is proposed as image-level data augmentation for contrastive learning. We incorporate the GAN and a contrastive learning module into a joint training framework in order to learn invariance from generated variance.

Chapter 6: Inter-instance Contrastive Encoding for unsupervised person ReID. In the sixth chapter, instead of direct data augmentation techniques, we propose a consistency regularization for data augmentation. We calculate the image-to-image similarity before data augmentation as soft labels to regularize that after data augmentation for unsupervised contrastive person ReID.

Chapter 7: Conclusion and Perspective. In the last chapter, we conclude our proposed methods in this thesis. Several unsolved and understudied problems in person ReID are discussed. We further raise perspectives for future person ReID research.

Chapter 2

Literature review

2.1 Overview

In this chapter, we review representative person ReID approaches proposed in recent years. Revolving around person ReID, we conduct this literature review on several aspects that affects ReID performance, including feature extractor, data augmentation, dataset and metric.

We first talk about person ReID datasets and evaluation metrics in Section 2.2. Feature extractors build representations from images, which directly influences the quality of identity representations. In Section 2.3, we go through popular features extractors for person ReID from handcrafted feature extractors to deep neural networks. Once a feature extractor is selected, a well-defined loss function is needed to train the feature extractor. In Section 2.4, we introduce popular loss functions for person ReID. Next, as data augmentation is one of the major contributions in this thesis, different categories of data augmentation are discussed in Section 2.5. We proceed to present representative unsupervised person ReID methods in Section 2.6. In the end, a conclusion for this chapter is drawn in Section 2.7.

2.2 Datasets and Evaluation metrics

Information about person ReID datasets is presented in this section. Depending on the type of inputs, person ReID approaches can be classified into image-based and video-based approaches.

2.2.1 Image person ReID datasets

Image person ReID datasets can be further categorized into single-shot and multi-shot datasets.

Single-shot datasets

In a single-shot dataset, each identity has only one query image and one correct match gallery image. Single-shot appraoches build appearance representa-

Dataset	BBoxes	Identities	Cameras	Scene
CAVIER [23]	610	72	2	indoor
VIPeR [46]	1,264	632	2	outdoor
GRID [88]	1,275	1,025	2	indoor
PRID [54]	1,134	934	2	outdoor
CUHK01 [79]	1,942	971	2	indoor
CUHK02 [78]	7,264	1,816	2	indoor
CUHK03 [80]	13,164	1,360	2	indoor
Market-1501 [164]	32,217	1,501	6	outdoor
DukeMTMC-reID [105]	36,411	1,812	8	outdoor
MSMT17 [138]	126,441	4,101	15	indoor&outdoor
PersonX [116]	273,456	1,266	6	synthetic
UnrealPerson [157]	120,000	3,000	34	synthetic
Person30K [2]	1,384,940	30,000	6,497	indoor&outdoor

Table 2.1: Representative image-based person ReID datasets. PersonX and UnrealPerson are synthetic datasets built on game engines (Unity and UnrealEngine 4).

tions using only a singe image of an individual, which use the least information among all kinds of appearance-based person ReID approaches. The representative single-shot datasets are VIPeR [46], GRID [88], PRID [54], CUHK01 [79]. These single-shot datasets usually contain hundreds of identities, each of which is recorded once from two different camera views.

Multi-shot datasets

In a multi-shot dataset, one identity may have one or multiple query images and several correct match gallery image. The representative multi-shot datasets are CAVIER [23], CUHK02 [78], CUHK03 [80], Market-1501 [164], DukeMTMC-reID [105] and MSMT17 [138], PersonX [116], UnrealPerson [157] and Person30K [2]. Recent trends in person ReID datasets emphasize more on generalizability and privacy protection. An ultra-large-scale person dataset Person30K [2] has been lately proposed to encourage domain generalization and pre-training research in person ReID. As collecting real person images may raise privacy concerns, several synthetic person datasets, such as PersonX [116] and UnrealPerson [157], are proposed to replace real person datasets for ReID research.

2.2.2 Video person ReID datasets

Video datasets and multi-shot datasets differ in the continuity between frames. A video dataset is composed of continuous images sequences, while a multi-shot dataset only has incontinuous images captured from different camera views. A person video clip usually contains more identity information than a single person image. In a crowded street, when a person in a single frame is occluded by other persons, properly using nearby frames in a video clip

Dataset	Tracklets	BBoxes	Identities	Cameras	Scene
PRID [54]	400	40,033	200	2	outdoor
iLIDS-VID [131]	600	42,460	300	2	indoor
DukeMTMC-VideoReID [105]	4,832	815,420	1,404	8	outdoor
MARS [163]	20,715	1,067,516	1,261	6	outdoor
LS-VID [74]	14,943	2,982,685	3,772	16	indoor&outdoor

Table 2.2: Representative video-based person ReID datasets.

helps to alleviate identity information missing. Representative video person ReID datasets and their statistics are presented in Table 2.2. Similar to image person ReID, handcrafted feature based methods are usually tested on small-scale datasets, such as PRID video version [54] and iLIDS-VID [131]. Recent deep learning methods are usually tested on large-scale datasets, including DukeMTMC-VideoReID [105], MARS [163] and LS-VID [74].

2.2.3 Evaluation metrics

Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are two most used evaluation metrics in person ReID. CMC judges the ranking capabilities of a ReID algorithm. When a first correct match appears at ith position, this query is considered as successful at rank i. mAP measures the mean value of average precision of all queries images, where average precision is the area under the Precision-Recall curve. For a single-shot dataset (*e.g.*, VIPeR, GRID, PRID and CUHK01), only CMC is enough, because each identity has only one query image and one gallery image. For a multi-shot dataset (*e.g.*, CUHK03, Market-1501, DukeMTMC-reID and MSMT17) where each identity has one or multiple query images and multiple gallery images, the evaluation protocol is to use both CMC and mAP.

In the deep learning era, large-scale multi-shot datasets have become main-stream benchmarks in person ReID community. Following most of recent works, our proposed methods are mainly evaluated on Market-1501, DukeMTMC-reID and MSMT17 with both CMC and mAP metrics. We plan to explore the new ultra-large-scale dataset Person30K and synthetic datasets PersonX and UnrealPerson in the future work.

2.3 Feature extractors

In this section, we focus on feature extractors that used in recent person ReID research. Feature extractors can be rough categorized into handcrafted feature extractors and deep neural networks.

2.3.1 Handcrafted feature extractors

In the pre-deep learning era, handcrafted feature extractors are main-stream approaches for person ReID. Color, texture and shape are the main visual clues to describe a person appearance in person ReID.

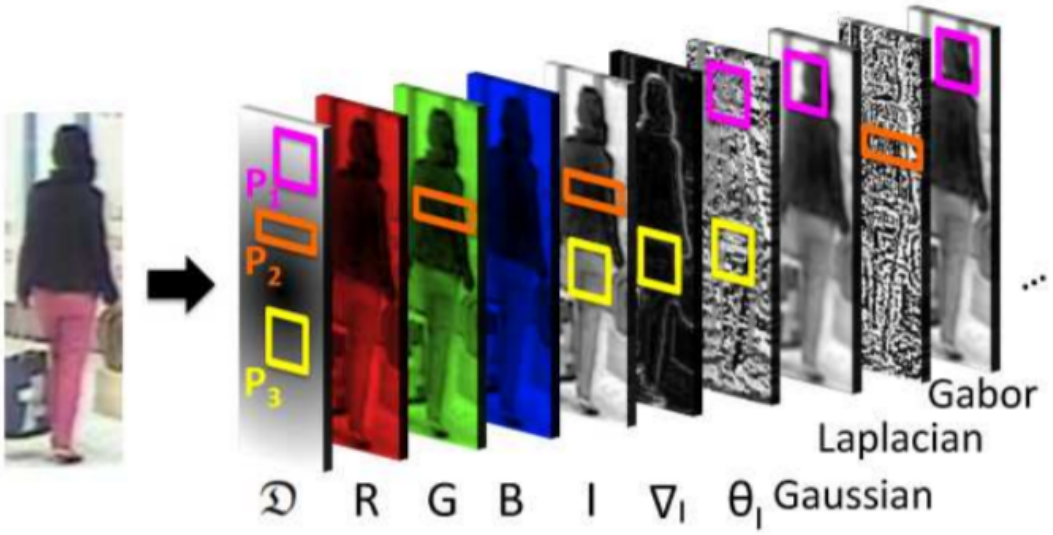


Figure 2.1: An covariance handcrafted feature space [5]. Every covariance is extracted from a region (P), distance layer (D) and three channel functions (*e.g.*, bottom covariance feature is extracted from region P3 using layers: D, I-intensity, ∇_I -gradient magnitude and θ_I -gradient orientation).

In 2006, Gheissari *et al.* [43] proposed to generate an appearance signature with an HS histogram and an edgle histogram for each local part in a segmented person foreground.

In 2008, Gray and Tao [46] totally considered eight color channels corresponding to the three separate channels of the RGB YCbCr and HSV [1] and nineteen texture channels to describe person appearances.

In 2010, Farenzena *et al.* [33] explored the symmetry of a peron body and used weighted color histograms, maximally stable color regions and recurrent high-structured patches to match a target person. Bak *et al.* [4] first used Histogram of Oriented Gradient (HOG) to detect human body parts and calculated covariance of detected body parts between two person images.

In 2012, Bak *et al.* [5] further proposed to select a feature subset from color, intensity, gradients and filter responses for learning a covariance descriptor (see Figure 2.1) in Riemannian geometry.

In 2013, Zhao *et al.* [161] proposed an unsupervised salience learning method on appearance patch representations composed of LAB color histogram and scale-invariant feature transform (SIFT) descriptor. Such salience learning was then extended into a RankSVM training to enhance ReID performance in [160].

In 2014, Das *et al.* [25] enhanced consistency between different camera pairs with appearance signatures built from an HSV color histogram on horizontal sub-regions respectively in torso and leg regions. AS shown in Figure 2.2 Yang

1. Both Y and V represent luminance. only one of them is used.

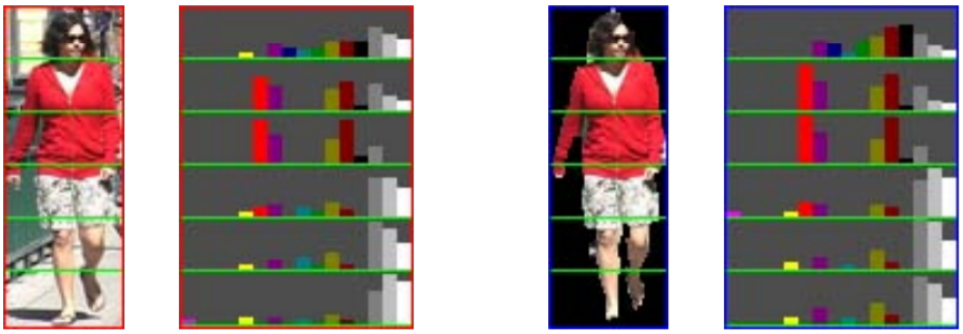


Figure 2.2: An example [149] of color histogram from a person bounding box and a segmented person foreground.

et al. [149] introduced a novel Salient Color Names Based Color Descriptor (SCNCD) for person ReID.

In 2015, Liao *et al.* [82] proposed an efficient feature representation called Local Maximal Occurrence (LOMO), which combines features from Scale Invariant Local Ternary Pattern (SILTP) histogram and HSV histogram.

In 2016, Matsukawa *et al.* [92] proposed an hierarchal Gaussian Of Gaussian (GOG) descriptor to generate features that can describe color and texture information in horizontal strips for a person image.

Several disadvantages make hand-crafted feature extractors less popular than deep neural networks in recent years: 1) Handcrafted features are difficult to engineer. Unlike general-purpose deep neural networks, *e.g.*, ResNet [49], handcrafted feature extractors require more domain-specific prior knowledge to design. 2) Handcrafted feature extractors are usually less effective than deep neural networks on large-scale datasets that contain more variance. Deep neural networks have proven to be effective on both large-scale datasets and small-scale datasets with a fine-tuning.

2.3.2 Deep neural networks

Deep neural networks (DNN) have lately become main-stream feature extractors for person ReID. In contrast to handcrafted feature extractors, deep neural networks have proven to be performant on large-scale datasets. Depending on main purposes, recent DNN can be caterorized into general-purpose neural networks and ReID-specific neural networks.

General-purpose neural networks

In computer vision community, general-purpose neural networks are usually designed for classification tasks, *i.e.*, image classification [68, 26] and video classification [113, 7]. Image classification (or image-based person ReID) usually relies on convolutional neural networks (CNN) to process 2D images. Video classification focuses on both visual spatial information and temporal informa-

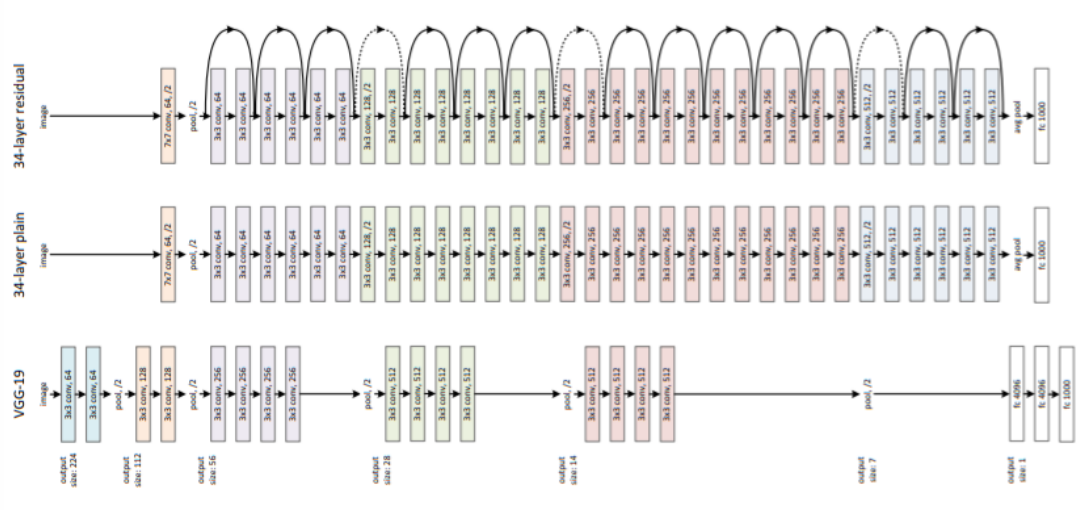


Figure 2.3: Comparison between VGG-19 [108], a 34-layer plain network and a 34-layer residual network [49].

tion, which can be processes by temporal models, *e.g.*, recurrent neural network (RNN) or 3D CNN [121].

CNN. Compared with a fully-connected neural network, a CNN avoids oversized fully-connected parameters and is good at maintaining spatial relation between nearby pixels. For image-based person ReID, most of previous works were conducted on a CNN backbone, *e.g.*, a ResNet-50.

LeNet [72] was one of the earliest CNN. Proposed in 1989 for handwritten digit recognition, LeNet contained 2 convolutional layers followed by average pooling and 3 fully connected layers, which were trained by back-propagation algorithms. LeNet defined the basic components of CNN.

AlexNet [69] was the champion model of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. Thanks to its GPU-implementation, AlexNet (convolutional layers, 3 of them followed by max pooling, and 3 fully connected layers) can be efficiently trained.

VGGNet [108] deepened the CNN architecture into 16 or 19 layers, which were respectively considered as VGG-16 and VGG-19 networks. VGGNet proposed to replace large kernel-sized filters in AlexNet (11×11 and 5×5) with multiple stacked 3×3 kernel-sized filters.

ResNet [49] proposed skip-connections to address gradient vanishing problem in deep neural networks. We compare a VGG-19 [108], a 34-layer plain network and a 34-layer residual network in Figure 2.3. Thanks to the skip-connections, authors successfully extended CNN architecture to 152 layers, which was 8 times deeper than VGGNets but still having less complexity. ResNet won 2015 ImageNet Large Scale Visual Recognition Challenge. Nowadays, ResNet and its variants, *e.g.*, ResNext [145] and DenseNet [57], remain most popular backbone networks in computer vision community. Moreover, ResNet-50 is the standard backbone for most of recent person ReID works.

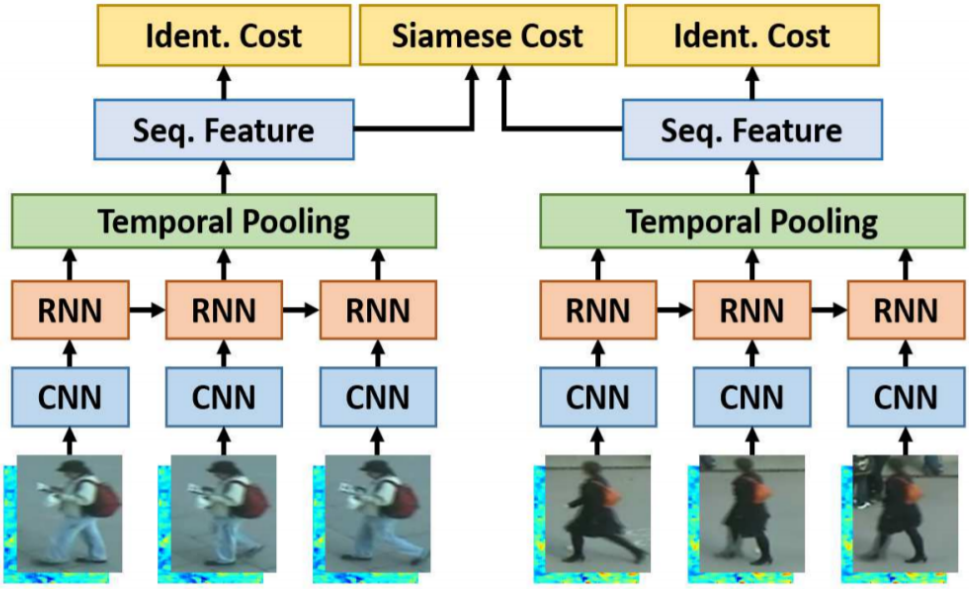


Figure 2.4: An example [93] of combining CNN and RNN for video-based person ReID.

MobileNet [55] targeted at seeking a trade-off between accuracy and complexity. Earlier mentioned networks only considered accuracy, which can not be used on certain platforms, *e.g.*, a mobile phone. Due to hardware constrains, researchers pay increasing attention to light-weight CNNs.

RNN. Conventional 2D CNN is good at handing spatial data, such as images, but is hard to handle temporal information. In contrast, RNN works better for temporal and sequential data, such as videos. To build robust appearance representations from tracklets, video-based ReID sometimes needs to analyze temporal information. In this context, several previous works [93, 1] have explored the possibility of combining both CNN and RNN into a unified framework (see Figure 2.4), where CNN mainly analyzes spatial information while RNN analyzes temporal information.

Transformers. In 2017, Transformer [123] was proposed as a novel network architecture, which was solely based on attention mechanisms illustrated in Figure 2.5. Dispensing with convolutions and recurrences, transformer provided an alternative backbone for computer vision and natural language processing (NLP). Recently, as a stronger backbone, the Vision Transformer (ViT) [29] has been successfully applied into person and vehicle ReID research [50].

ReID-specific neural networks

Other than general-purpose neural networks, several attempts have been made in designing ReID-specific networks. Compared to general-purpose neu-

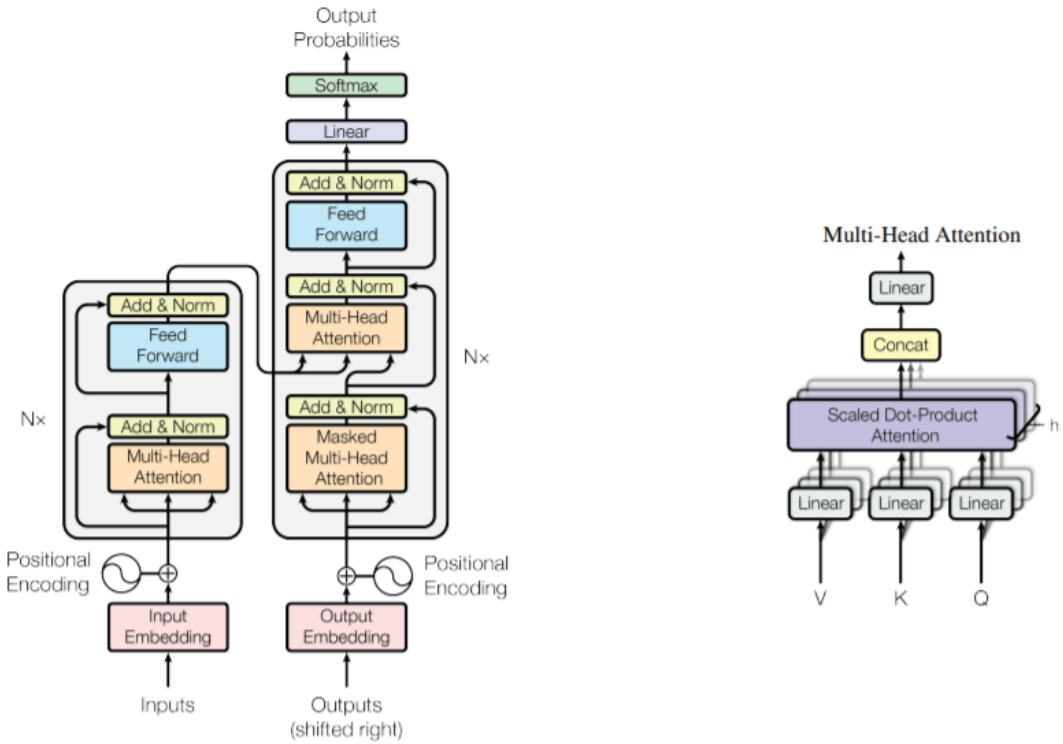


Figure 2.5: Left: Transformer [123]. Right: Multi-head attention in a Transformer, where V, K, Q are feature matrix from the last layer.

ral networks, *e.g.*, ResNet-50, ReID-specific neural networks usually have a better trade-off between accuracy and complexity.

HA-CNN [81] is a 39-layer CNN with only 2.7 million parameter numbers, which is approximately one-ninth of ResNet-50. Authors of HA-CNN used Inception [119] units as basic blocks and inserted Harmonious Attention blocks between Inception blocks. especially, an Harmonious Attention block is composed of complementary attention mechanisms, including soft spatial-channel attention, channel attention and hard regional attention, to learn discriminative attentive feature maps.

OS-Net [176] is another light-weight ReID-specific neural network with only 2.2 million parameter numbers. Inspired by MobileNet, authors of OS-Net also used depthwise separable convolutions that can be divided into a depthwise kernel and a pointwise kernel. Authors also stacked different number of 3×3 lite convolutional layers to learn multi-scale feature.

Auto-ReID [102] is the first automatically searched network for person ReID. Earlier mentioned networks are all manually designed, which is a cumbersome task and requires strong human expertise in the target task. In contrast, based on Neural Architecture Search (NAS), an optimal ReID-specific network can be formed automatically with Auto-ReID. Only using 40% parameter number of ResNet-50, Auto-ReID achieved competitive performance.

Different types of feature extractors are compared in this section. The standard feature extractor in the person ReID research community is ResNet-50. Our research direction does not focus on designing new network structures. To conduct fair comparison with previous methods, all of our proposed methods are based on ResNet-50 backbone. In the next section, we discuss representative loss functions that can train robust feature extractors.

2.4 Loss functions

A loss function measures the distance between predictions and ground truth labels. Based on the distance measure by a loss function, neural networks use the back propagation to gradually seek optimal network parameters. As a fine-grained image retrieval task, person ReID particularly relies on properly-designed loss functions to learn discriminative identity features. We present some representative loss functions for neural network training in the computer vision community, such as ℓ_1 loss, ℓ_2 loss, cross-entropy loss and triplet loss.

ℓ_1 and ℓ_2 loss. ℓ_1 and ℓ_2 (or mean squared error) losses are simplest metrics to measure the distance. Given a random person image x, we use $f(x)$ to denote the prediction that is encoded by a feature extractor f. ℓ_1 and ℓ_2 losses between the prediction $f(x)$ and the corresponding target f_+ are respectively defined as:

$$\ell_1(f(x), f_+) = \mathbb{E}[\|f(x) - f_+\|_1] \tag{2.1}$$

$$\ell_2(f(x), f_+) = \mathbb{E}[\|f(x) - f_+\|_2] \tag{2.2}$$

ℓ_1 and ℓ_2 losses are commonly used in regression tasks, as well as in auto-encoder and GAN [45] as image/feature reconstruction losses. However, ℓ_1 and ℓ_2 losses are not quite suitable for classification tasks, because they do not punish misclassifications enough.

Cross-entropy loss. Cross-entropy loss is the most used loss function in classification. Binary cross-entropy loss is usually associated with logistic function for binary classification, while softmax cross-entropy loss is usually associated with softmax function for multi-class classification. Compared with ℓ_1 and ℓ_2 losses, the cross-entropy loss uses a log function to exponentially punish misclassifications. Since the first CNN-based person ReID method [80], the cross-entropy loss has always been the most used loss function to train ReID models. Given the prediction $f(x)$ and the corresponding target f_+, a cross-entropy loss is defined as:

$$\mathcal{L}_{xentropy}(f(x), f_i) = \mathbb{E}[-\log \frac{\exp\left(f(x) \cdot f_+\right)}{\sum_{i=0}^{K} \exp\left(f(x) \cdot f_i\right)}] \tag{2.3}$$

where f_i is a set of candidates that include the positive target f_+. This loss minimizes the distance between $f(x)$ and the target f_+ (usually a class center as shown in Figure 2.6 (a)) while maximizing the distance between $f(x)$

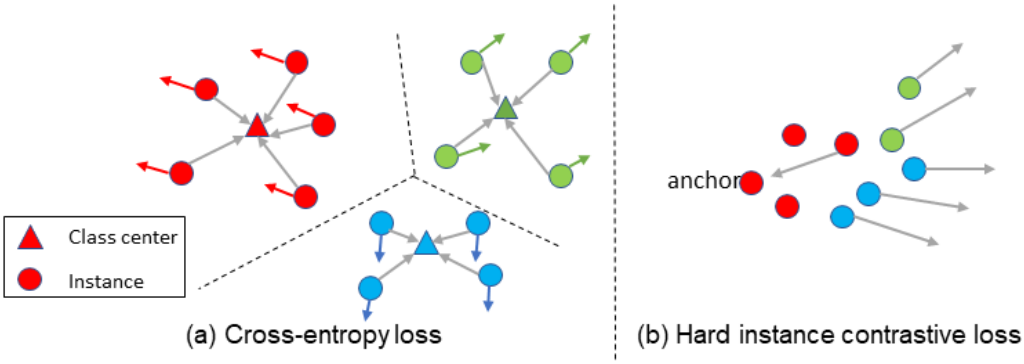


Figure 2.6: Comparison between (a) cross-entropy loss and (b) batch-hard triplet loss.

and other negative samples f_i in the candidate set, which is a perfect fit for multi-class classification. Besides multi-class classification, the cross-entropy loss has also been widely extended to other use cases, such as InfoNCE contrastive loss [96] for self-supervised learning and soft cross-entropy loss [53] for knowledge distillation.

Triplet loss. Triplet loss [9] is another commonly used loss function in person ReID. Different from earlier mentioned losses, a triplet loss is designed on three subjects, namely an anchor (A), a positive (P) and a negative (N). The objective of triplet loss is to minimize the distance between a positive pair (A, P) and maximize the distance between a negative pair (A, N). The triplet loss is defined as:

$$\mathcal{L}_{triplet}(f(A), f(P), f(N)) = \mathbb{E}[\max(\|f(A) - f(P)\|_2 - \|f(A) - f(N)\|_2 + m, 0)] \tag{2.4}$$

where m is a margin hyper-parameter that controls the distance scale between positive and negative pairs. To enhance the representation discriminability for a fine-grained task, *e.g.*, person ReID and face recognition, researchers proposed to mine hardest positive and negative from a mini-batch to form a batch-hard triplet loss [51], as shown in Figure 2.6 (b). To cope with the imbalanced data, the batch-hard triplet loss is usually combined with a random identity sampler. In the random identity sampler, for each training iteration, a mini-batch batch is formed with I identities and S samples for each identity. The batch-hard triplet loss is defined as:

$$\begin{aligned}\mathcal{L}_{triplet\_hard}(f(A), f(P), f(N)) = \mathbb{E}[\max(&\max_{P=1,\dots,S} \|f(A) - f(P)\|_2 \\ &- \min_{N=1,\dots,(I-1)\times S} \|f(A) - f(N)\|_2 + m, 0)]\end{aligned} \tag{2.5}$$

where the hardest positive P has maximal distance from the anchor A in S positives and the hardest negative N has maximal distance from the anchor A in $(I-1) \times S$ negatives. Completing cross-entropy loss that provides general

class-level discriminability with a batch-hard triplet loss that focus on hard sample discriminability has proven to be extremely effective in several person ReID methods [128, 12, 39, 14]. Moreover, there are also several variants of the triplet loss, such as hard instance contrastive loss [14] for self-supervised learning and soft triplet loss [39] for knowledge distillation.

Other losses. Other than above mentioned losses, there are other intelligently designed losses to enhance the match accuracy in person ReID. For example, Kullback–Leibler (KL) divergence is used in some works [165, 179] to reduce the distance between two distributions. Chen *et al*. introduce a quadruplet loss [19] in form of (A, P, N_1, N_2) as a substitute for the triplet loss (A, P, N), in which a second negative N_2 was used to encourage a better intra-class compactness and inter-class separateness. As triplet loss pulls the positive closer and pushes the negative further away equally, circle loss [117] is proposed to adaptively pull or push if the positive is already close or the negative is already far away. Metric learning has always been a major research direction for person ReID.

Different types of loss functions are compared in this section. Most of these loss functions are proposed for supervised learning. Based on the cross-entropy loss and KL divergence, we introduce a hard instance contrastive loss and a soft instance consistency loss in Chapter 6, which are more suitable for unsupervised person ReID.

2.5 Data augmentation

As a major technique to enrich data diversity, data augmentation plays an important role in recent data-driven deep learning methods. By introducing meaningful data augmentation mimicked distortions into neural network training, inference representations can be more invariant to distortions that encountered in the training. Data augmentation techniques can be categorized into image-level and feature-level augmentation. Moreover, GAN provides a new approach for augmentation.

2.5.1 Image-level data augmentation

Traditional image-level data augmentation is conducted directly on images. Pixel position and RGB values are two main factors that can influence the visual perception of images. Image-level data augmentation can be realized by basic positional transformations, such as scaling, cropping, flipping, padding, rotation, translation, erasing [171] *etc*. It can also be realized by color transformations, such as color jittering, Gaussian noise, Gaussian blur, *etc*.

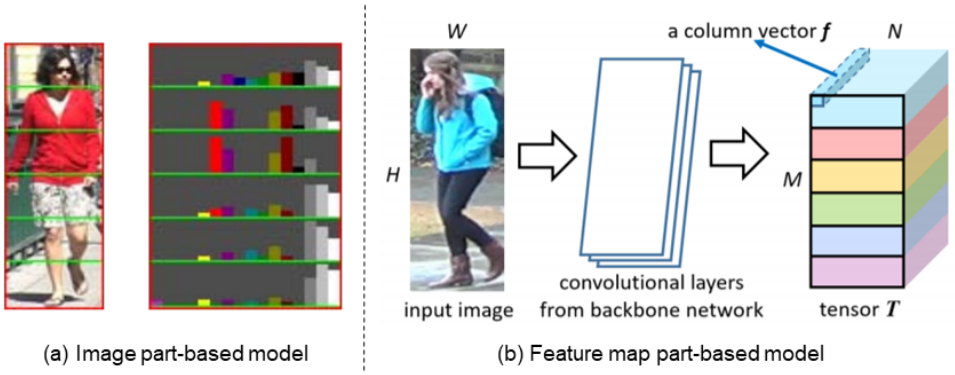


Figure 2.7: Comparison between (a) image part-based model [149] and (b) feature map part-based model [118].

Image part-based models

Other than general image-level data augmentation, several previous ReID models extract local features from image parts, which can be regarded as an image-level data augmentation technique. There are three types of image parts, *i.e.*, patches, stripes and estimated body parts. For example, in [161], a person image is densely segmented into 10×10 pixel squares with a step size of 5 pixel. Similar patches are also used in other works [87, 164, 80]. However, dense image patches are somewhat computationally expensive. As shown in Figure 2.7 (a), partitioning a person into several horizontal stripes can be a more efficient substitute. For example, Gray *et al.* [46] partitioned a person image into six equally-sized horizontal stripes and extract local features from each stripe. Similar stripes are also used in other works [149, 82, 151]. Both patches and stripes are pre-defined image parts, which are not deformable according to the position and pose of a person. To address this issue, there are several attempts [33, 4, 23] that first use a pose estimator to detect body parts and then extract local features from estimated body parts.

2.5.2 Feature-level data augmentation

Feature map part-based models

Instead of image parts, as shown in Figure 2.7 (b), partitioning a neural network encoded feature maps into several parts to extract local feature is another widely used approach in person ReID. PCB [118] partitions a ResNet-50 encoded feature map into six horizontal stripes and use six fully-connected layers as classifiers to separately process each local representation. MGN [128] duplicates last convolutional layers of ResNet-50 into three separate branches, each of which generates different sized feature map stripes. HPM [38] uses different global pooling methods to convert feature map stripes to local representations. SCR [12] further considers channel-wise partitions to complete previous spatial partitions.

Feature dropout

Dropout [114] simulates training a large number of neural networks with different architectures in parallel, which can be also regarded as a virtual data augmentation technique. By randomly dropping some neurons in the training, dropout augments the feature noise, forcing a neural network to handle feature missing situations. Following dropout, domain guided dropout [144] was proposed to learn generic feature representations from multiple domains for person ReID.

Randomly dropping separate neurons is not enough effective to bring in distortions on semantic information. Dropblock [44] thus drops continuous regions, forcing a neural network to handle semantic information missing situations. Batch dropblock [24] applies dropblock on same regions for all images in a mini-batch, which creates samples “harder" than original hardest samples for batch-hard triplet loss.

2.5.3 GAN-based data augmentation

Zheng *et al.* [169] unconditionally generated a lot of unlabeled person images with DCGAN [103] to enlarge data volume for supervised ReID. Following GAN-based methods were usually conditionally conducted on some factors from Table 2.3.

1) Pose: With the guidance of 2D poses, FD-GAN [40] and PN-GAN [100] generated a target person in new poses to learn pose-irrelevant representations for single-domain supervised ReID. Similar pose transfer [73] was then proposed to address unsupervised domain adaptive (UDA) ReID.

2) Dataset style (illumination): As a dataset is usually recorded in a uniform illumination condition, PTGAN [138] and SyRI [3] used CycleGAN [178] to minimize the domain gap between different datasets by generating person images in the style of a target domain.

3) Camera style: Instead of the general dataset style, CamStyle [175] transferred images captured from one camera into the style of another camera, in order to reduce inter-camera style gaps. Similar method [172] was then applied to UDA ReID.

4) Background: SBSGAN [59] and CR-GAN [22] respectively were targeted at removing and switching the background of a person image to mitigate background influence for UDA ReID.

5) General structure: By switching global and local level identity-unrelated features, IS-GAN [30] disentangled a representation into identity-related and identity-unrelated features without any concrete guidance. As a concrete guidance, a grey-scaled image contains multiple id-unrelated factors of a person image, including pose, background and carrying structures. By re-coloring grey-scaled person images with the color distribution of other images, DG-Net [165] and DG-Net++ [179] learned disentangled identity representations invariant to structure factors.

Id-related	Id-unrelated
cloth color, hair color, texture, body shape	pose, view-point, illumination, camera style background

Table 2.3: Id-related and Id-unrelated factors in a person image.

Over the recent years, designing novel data augmentation techniques has always been a major direction for person ReID research, especially part-based models that extract local features. However, previous part-based models focus on spatial dimensions, but neglect the channel dimension. We conduct channel-wise partition to extract more meaningful secondary features in Chapter 3. We further use different depth and different polling methods to enrich feature diversity in Chapter 4 and a novel GAN-based augmentation for contrastive learning in Chapter 5.

2.6 Unsupervised person ReID

Recently, significant improvement has been witnessed in unsupervised representation learning. In comparison with general object class classification (cat, dog, person, *etc*. in ImageNet), person ReID aims at distinguishing fine-grained person identity classes, which can be easily affected by environmental factors. A ReID model trained on a specific domain is usually difficult to keep good performance on other domains. In such context, how to quickly adapt a pre-trained model to a target domain or directly train a model in a target domain without a cumbersome annotation phase has become a major research direction in recent years. Depending on the necessity of a large-scale labeled source dataset, unsupervised person ReID methods can be roughly categorized into unsupervised domain adaptive (UDA) and fully unsupervised ReID.

2.6.1 Unsupervised domain adaptive ReID

There are usually strong domain gaps between different person ReID datasets, leading to a degraded performance for cross-domain evaluation. To bridge domain gaps, there are three main approaches for UDA person ReID as shown in Figure 2.8.

1) Adversarial learning: Adversarial learning methods include adversarial learning and generative adversarial network (GAN) methods, which target at learning a target domain distribution that can fool a domain discriminator. Most of GAN-based unsupervised ReID methods [73, 138, 3, 172, 22, 179] are based on the idea of style transfer. Transferring images from a labeled source domain into the style of an unlabeled target domain, which fells into the setting of UDA ReID. Especially, Wei *et al*. [138] directly transfer source domain images into the style of a target domain, while Zhong *et al*. [172] transfer the camera styles. Bak *et al*. [138] generate images under different illumi-

Figure 2.8: Three main approaches for unsupervised person ReID: (a) Adversarial learning, (b) Metric learning and (c) Pseudo labelling.

nation conditions, while Li *et al.* [73] generate person images with different poses. DG-Net++ [179] generates cross-domain images with the guidance of target-domain structures. Based on generation, our contribution in Chapter 5 provides a novel approach of generating person images in different view-points for contrastive learning.

2) Metric learning: Designing proper metrics to reduce distribution discrepancy between different domains also allows for better UDA. For example, several works [129, 83] form a multi-task classification on identities and semantic attributes to facilitate the domain adaptation, in which the distribution distance between a source and a target domain on both identity and attribute representations is minimized. For example, the distribution distance can be estimated with Maximum Mean Discrepancy (MMD), which represents distribution distance as the distance between mean embeddings of features. Inspired by triplet loss [51], our contribution in Chapter 6 improves contrastive learning with techniques from metric learning.

3) Pseudo labeling: Another major approach consists of assigning pseudo labels to unlabeled images and conducting pseudo label learning [153, 37, 147, 39, 89, 173, 174]. Pseudo labels can be obtained by existing clustering algorithms, *e.g.*, K-means [39], DBSCAN [179, 147] and HDBSCAN [37] or newly designed pseudo labelling algorithms [153, 174]. Our contributions in Chapter 4, Chapter 5 and Chapter 6 use DBSCAN [31] to assign pseudo labels, which allows for learning clustering-based neighborhood relationship.

2.6.2 Fully unsupervised ReID

Compared with UDA person ReID methods, fully unsupervised person ReID lifts the constraint on the labeled source domain, which thus has a better flexibility. Most of previous fully unsupervised methods [84, 85, 125, 75, 140] are based on pseudo labeling. Instead of adapting a source domain pretained model, such fully unsupervised methods directly learn robust identity representations with clustering-based pseudo labels on unlabeled target domain images.

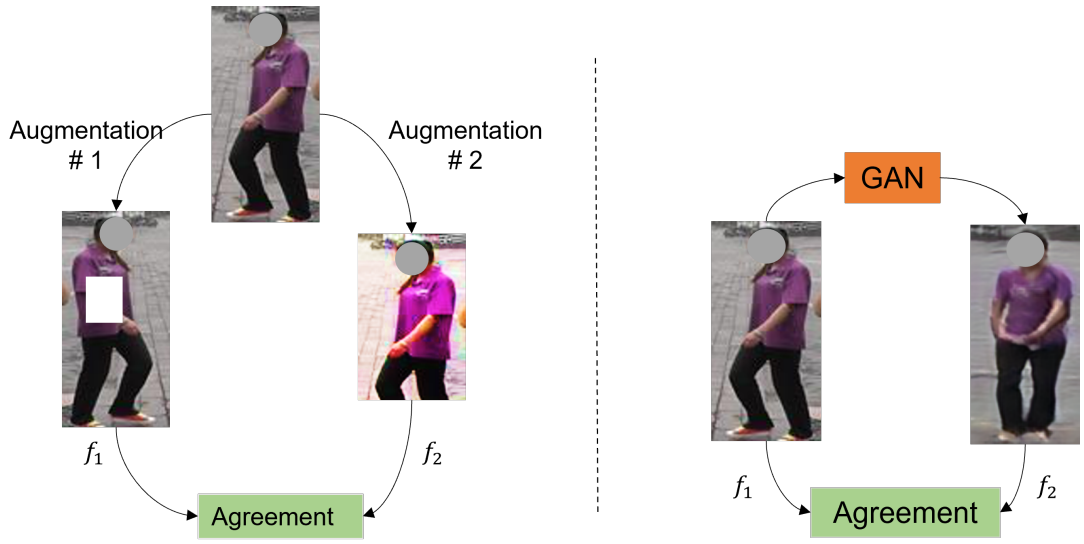


Figure 2.9: Left: conventional contrastive learning. Right: Our proposed joint generative and contrastive learning.

Contrastive learning [48, 17, 21] has been a novel approach for unsupervised representation learning. The main idea of contrastive learning is to form a positive pair with augmented views of a same image and regard other images as negative pairs, as shown in Figure 2.9 left. By maximizing the similarity between positive pairs while minimizing the similarity between negative pairs, a model should be robust to augmented distorsions. In the person ReID community, researchers combine pseudo labeling and contrastive learning to learn better unsupervised identity representations [42, 130]. Especially, Ge *et al.* [42] form a positive pair between an image and a corresponding cluster prototype. Wang *et al.* [130] form a positive pair between an image and corresponding camera sub-cluster prototypes, which further reduces camera style variance.

Our proposed methods in Chapter 5 and Chapter 6 are related to pseudo labeling and contrastive learning. We improve the performance in our methods by designing proper data augmentation and exploring inter-instance affinity that are neglected in previous methods.

2.7 Conclusion

In a nutshell, a literature review on representative methods in different aspects of person ReID research is presented in this chapter, including datasets, feature extractors, loss functions, data augmentation and unsupervised approaches. Among these aspects, we are particularly interested in insufficient data diversity and insufficient annotation problems. To enhance diversity from feature level, we present a spatial-channel partition mechanism extract meaningful local features for supervised person ReID in Chapter 3. Towards a ro-

bust unsupervised person ReID system, we mainly explore unsupervised person ReID in the following chapters. Another approach to enhance diversity from feature level by modifying network depth and pooling method is presented for unsupervised person ReID in Chapter 4. We propose a novel GAN-based augmentation technique to generate more suitable augmented views for unsupervised contrastive person ReID in Chapter 5. We further explore the inter-instance similarity affinity in unsupervised contrastive person ReID in Chapter 6.

Chapter 3

Spatial-channel partition for supervised object ReID

3.1 Introduction

Our contributions revolve around enhancing image or feature diversity with data augmentation techniques. In this chapter, we propose a new feature-level data augmentation technique that permits identity representations to contain both spatial and channel-wise local features. In contrast to attention mechanisms, our proposed spatial-channel partitions effectively enrich extracted feature diversity for person ReID. This work has been published in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2020 [12]. A vehicle ReID model [11] based on similar partition techniques was proposed in 2019 AI City challenge [95] and has been published in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

As mentioned in Chapter 1, there are still various challenging problems to be solved in a real-world Re-ID task, such as camera view point changes, illumination differences, pose variation and partial occlusion. With the rapid development of deep learning based techniques, recent research with convolutional neural networks [81, 167] get remarkable advances in person Re-ID, which surpass the performance of traditional handcrafted methods [82, 164] on large datasets.

To measure similarities between two captured images, we need to build an appearance representation for each sample in datasets. The most intuitive method for building representations is to extract directly global feature map from the entire bounding box. However, Re-ID methods that rely solely on global features of a person are prone to errors in case of occlusion and misalignment. On the other hand, key local features (carried objects and body parts, such as face and hands) can not be always observable due to low camera resolution and occlusions.

Since viewpoint change, partial occlusion and misalignment are frequent in real-world Re-ID task, complementing global features with local features addresses these issues and builds better person representations. This further improves a neural network's capacity to distinguish similar people based on

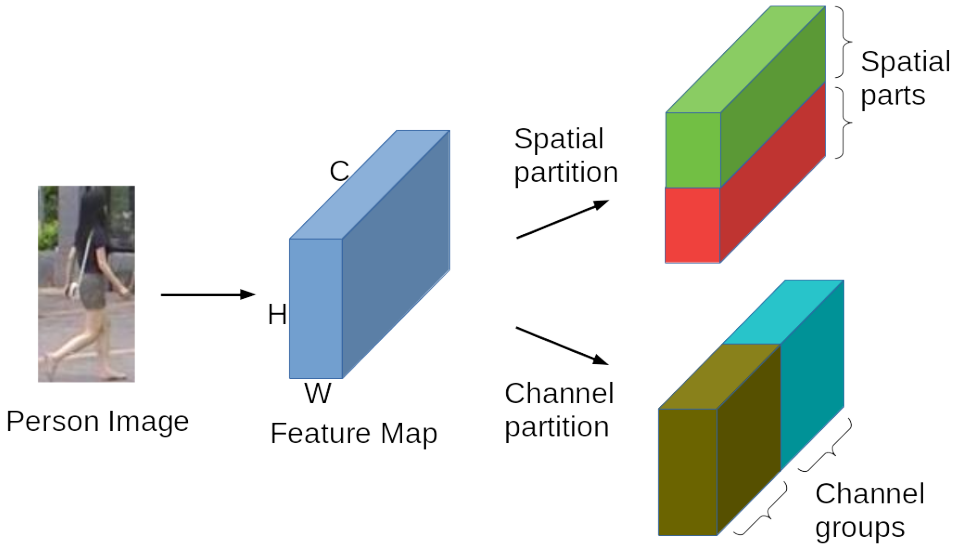


Figure 3.1: Example of a spatial-channel partition. H, W and C stand for respectively Height, Width and Channel in a deep feature map. In this example, we partition a whole feature map into two spatial parts (upper body and lower body) and two channel groups.

small part differences.

Therefore, part-based models [4, 46, 94] have attracted a lot of attention in Person Re-ID research community. Recently, several sophisticated models [38, 128] have combined multiple partitions conducted along the height dimension in a pyramidal structure. They have significantly outperformed previous state-of-the-art. A feature map extracted from an image has 3 dimensions, *i.e.*, height, width and channel (also called depth). Because height and width dimensions correspond both to the spatial coordinates of pixels in an image, the partition conducted along the height and width dimensions is called spatial partition. Independently from spatial coordinates, partition conducted along the channel dimension is called channel partition. An example is shown in Figure 3.1.

Spatial partition is a common strategy in Re-ID task, which enables part-to-part matching by extracting local features corresponding to specific body parts. Channel information comes from filters in a convolutional layer. As the CNN goes deeper, last layers are more abstract, and outputs of the channels are higher level features, corresponding potentially to concepts such as hair color, body shape, *etc*. Channel partition does not extract local features from specific body parts, but keeps features that indicate the presence of these high level concepts. By splitting channels into several groups and training separate channel groups, semantic concepts in each channel group can be decorrelated. Therefore, channel partition allows to conduct semantic concept-to-concept matching (*e.g.*, with or without a bag strap), which can complement spatially partitioned representations. As proven in [129, 83], semantic attributes show a strong generalibility in cross domain Re-ID task. By combing part-to-part and

semantic concept-to-concept matching, we are able to build discriminative and generalizable representations with spatial-channel partition.

Another well-considered strategy for extracting discriminative features is attention mechanism. Attention mechanism helps CNNs to focus on the most discriminative part (called primary information) in feature maps. Features on other parts which are salient but less discriminative (called secondary information) are then neglected. These features can be complementary clues for distinguishing people with similar appearances. Partitions enable a CNN not only to consider primary features but also to keep secondary features.

In this work, we focus on how to build robust person representations for Re-ID task by complementing global features with local features extracted through partitions.

In summary, our contribution is twofold:

1. We conduct a comparative study between spatial, channel partitions and attention mechanism. Results of this study can be summarized by 2 statements: (a) Attention mechanism may remove useful secondary information, which can be kept by partitions. (b) Compared to traditional spatial partition, channel partition shows a superior capacity of maintaining secondary local information.
2. Spatial and channel partitions are combined (called spatial-channel partition) to further enhance deep neural networks' ability to learn secondary information. By adopting multiple spatial-channel partitions in a pyramidal structure, we propose a unified end-to-end trainable framework for Person Re-ID.

Our proposed framework is exhaustively evaluated on three image-based Re-ID datasets, Market-1501, DukeMTMC-ReID, CUHK03 and one video-based dataset MARS. On the MARS dataset, partition is also applied on the temporal dimension to build a more robust representation for each tracklet. The evaluation results show that our method can build both discriminative and generalizable representations, which outperform previous state-of-the-art in both supervised single domain and unsupervised cross-domain Re-ID tasks.

3.2 Study of Appearance Representations

Building discriminative appearance representations to measure quantitatively the similarity between query and gallery images is a common approach in Re-ID task. First, we evaluate robustness of appearance representations within the state-of-the-art. Then, we explain why spatial and channel partitions should be combined together and why they can outperform attentive models.

3.2.1 State-of-the-art

The two main approaches to make appearance representations robust in the state-of-the-art consist of choosing appropriate loss function and partitioning

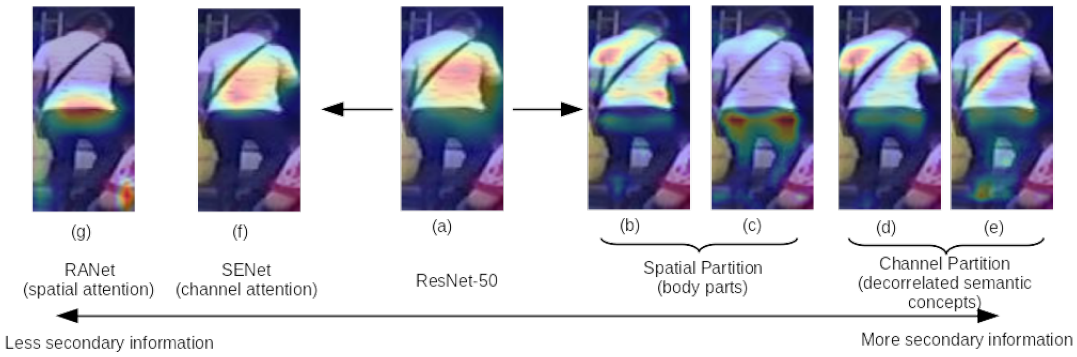


Figure 3.2: Comparisons of saliency maps generated by Grad-CAM [107] applied on 4 CNN models on Market-1501 test set. (a): A ResNet-50 w/o partitions nor attention mechanism. (b) to (e): A ResNet-50 w/ spatial-channel partition, where (b) and (c) are saliency maps on two spatial parts after spatial partition, (d) and (f) are saliency maps on two channel groups after channel partition. (f): Squeeze-and-Excitation Network [56]. (g): Residual Attention Network [126].

a person image into several spatial parts.

Loss Functions for Re-ID. A domain guided dropout is proposed in [144] to train a classification model on various datasets, which consider Re-ID as an identification task. Zheng *et al*. [167] use both pair-wise verification and classification loss to learn a more discriminative representation for Re-ID task. However, classification loss fails in cases of people wearing similar clothes, misaligned bounding boxes, *etc*. In [51], Batch Hard triplet loss is proposed to focus more on these hard samples. But the performance of triplet loss highly relies on how to select the hardest positive and negative pairs in a batch, which is difficult based on local features extracted from body parts. More details about sampling and hard pair selection are given in section 3.3.2. To get a better performance, we adopt triplet and classification losses to train our SCR model in a joint learning manner.

Spatial Part Based Models. Partitioning the entire body image into several spatial parts has always been a popular strategy in Re-ID task. Gray *et al*. [46] first propose to partition the person image into six equally-sized horizontal stripes and extract color and texture features in each stripe. Farenzena *et al*. [33] segment entire images into three salient and meaningful regions (head, torso and legs) by exploiting asymmetry and symmetry principles. These hand-crafted feature based approaches work well on small datasets, but are less robust and fail on large datasets.

Recently, part-based deep learning methods build more robust representations with deep features learned on large datasets. Yao *et al*. [150] train CNN in several maximally activated regions on feature maps. In [32], authors propose a spatial-channel loss to ensure that each channel in the representation pays attention to a dedicated partitioned part of the body. But original channel information in the feature map is replaced by spatial information, making

this loss inappropriate to maintain specifically channel information. Part-based Convolutional Baseline (PCB) proposed in [118] introduces a simple yet effective model based on six identical horizontal stripes. But same body part can be found in different stripes between different training samples, especially when bounding boxes are misaligned. To address this issue, a recent research trend is to combine multiple partitions to build a robust appearance representation.

Multi-partition Pyramidal Models. In HPM [38], Fu *et al*. partition respectively the entire body feature map into one, two, four and eight identical horizontal stripes. In MGN [128], authors split last layers of a ResNet into three branches and partition feature maps into one, two and three horizontal stripes. CPM [162] adopts multiple overlapping partitions. However, experiments show that these overlapping partitions do not increase the performance of our network. All these methods only consider multiple spatial partitions and neglect the channel ones.

Channel Group Operations. In AlexNet [69], Krizhevsky *et al*. firstly partition channels into 2 groups and introduce grouped convolutions to distribute a model into two GPUs. In ResNeXt [145], authors show that a larger number of channel groups can improve accuracy in image classification task without increasing computational complexity. MobileNet [55] adopts channel-wise convolutions where the number of groups equals the number of channels. In a similar way, by partitioning channels into several groups and computing within each channel group the mean and the variance for normalization, Group Normalization [142] outperforms Batch Normalization [60]. These studies confirm the potential of partitioned channel groups as an effective dimension in various visual tasks. In the following section, we discuss using partitioned channel groups to enhance the robustness of representations in Re-ID task.

3.2.2 Study of Partitioned Representations

In Figure 3.2 (a), the heat map generated by Grad-CAM [107] shows that a vanilla ResNet-50 trained with cross-entropy classification loss focuses on upper body especially on the region next to right arm, where other regions, *e.g.*, right arm, legs and shoes are totally neglected. When training in this way, a CNN solely considers features on some discriminative regions. In consequence, it suffers from over-fitting on these regions and becomes less robust for hard samples. To overcome this issue and build a more generalized appearance representation, we employ spatial-channel partitions and train multiple classifiers separately on partitioned maps. Specific local features are fed separately into dedicated classifiers, each of them can be regarded as a local expert. A local expert works better on a dedicated part. Combining all local experts allows to build more robust representations for Re-ID. To verify this idea, we have conducted experiments on Market-1501 dataset, whose results are shown in Figure 3.2 (b) to (e). With spatial partition, more regions in upper and lower body are highlighted respectively in Figure 3.2 (b) and (c) as compared to (a). With channel partition, the model does not train local experts on dedicated body parts but on a group of high level features. Thanks to channel partition,

in Figure 3.2 (d) and (e), the obtained saliency maps have more highlighted regions corresponding to semantic concepts, such as shoulder strap and shoes. The T-shirt is highlighted in (d), while the shoulder strap and shoes are highlighted in (e). Different activated semantic concepts in (d) and (e) show that channel partition is able to decorrelate high level features in different channel groups and conduct finer concept-to-cencept matching.

Since activated regions of channel partitions are different from those of spatial partitions, we can infer that local features extracted from both types of partitions are complementary when trained jointly. Harmonious Attention [81] is a combination of multiple attention mechanisms, such as Channel Attention [56] and Spatial Attention [126]. Inspired by Harmonious Attention, we propose to combine both types of partitions to form a spatial-channel partition. Comparison results between only spatial partition, only channel partition and spatial-channel partition are reported in Table 3.2. The performance of channel partition is better than that of spatial partition, because semantic concept-to-concept matching is more robust to misalignment than body part-to part matching. Spatial-channel partition can further enhance the performance.

Both partition and attention mechanism aim at enhancing the ability of neural networks to extract more discriminative features, but in opposite ways. Attention mechanism guides neural networks in locating the most important region in an image. As a consequence, secondary information may be neglected by attention mechanism. On the contrary, training local experts on partitioned parts enables neural networks to learn more local features. Heat maps of attention models in Figure 3.2 (g) and (f) keep less secondary information in the feature map than a ResNet-50 in Figure 3.2 (a), while partition based models in Figure 3.2 (b) to (e) keep more secondary information. Results in Table 3.1 validate that keeping secondary information by spatial-channel partition brings more improvement compared to removing secondary information by attention.

Model	Rank1	mAP
ResNet-50	89.5	73.3
ResNet-50 + channel attention (SENet)	90.8	75.6
ResNet-50 + spatial-channel partition	94.4	85.8

Table 3.1: Comparison of results (%) between attention and partition on Market-1501 dataset. SENet refers to Squeeze-and-Excitation Network [56]. Spatial-channel partition refers to the model trained with 2 spatial parts and 2 channel groups.

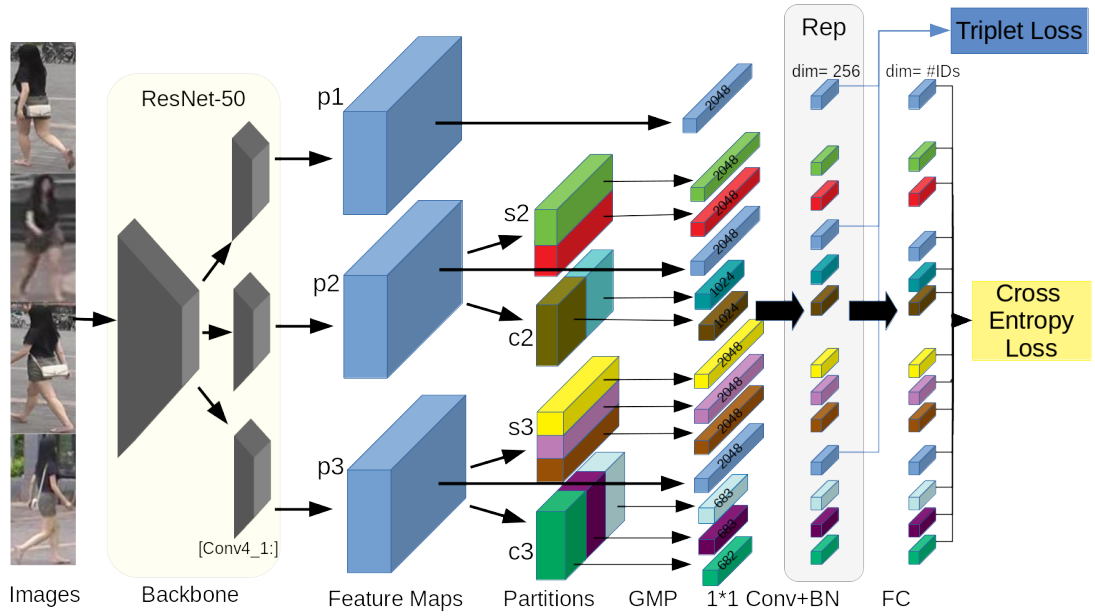


Figure 3.3: Spatial and Channel Partition Representation network. For the backbone network, we duplicate layers after conv4_1 into 3 identical but independent branches that generate 3 feature maps "p1", "p2" and "p3". Then, multiple spatial-channel partitions are conducted on the feature maps. "s2" and "c2" refer to 2 spatial parts and 2 channel groups. "s3" and "c3" refer to 3 spatial parts and 3 channel groups. After global max pooling (GMP), dimensions of global (dim = 2048) and local (dim = 2048, 1024*2 and 683*2+682) features are unified by 1*1 convolution (1*1 Conv) and batch normalization (BN) to 256. Then, fully connected layers (FC) give identity predictions of input images. All the dimension unified feature vectors (dim = 256) are aggregated together as appearance representation (Rep) for testing.

3.3 Proposed Framework

3.3.1 Spatial and Channel Partition Representation Network (SCR)

The general architecture of our proposed SCR network is represented in Figure 3.3. A batch of input images are fed into a backbone network. Last layers of backbone network are split into 3 independent branches in order to satisfy the need for a pyramidal structure, which generate 3 feature maps of equal size. A global feature map is extracted from each branch. The second and third feature maps are then partitioned into 2 and 3 spatial-channel parts respectively. A global max pooling (GMP) is used to replace global average pooling (GAP) in order to extract the most discriminative features in each part. Global and local feature maps are transferred to vectors with distinct dimensions. Next, dimensions of feature vectors are unified by 1*1 convolutional layers into 256. We train 13 fully connected layers as classifiers with 13 softmax cross-entropy losses respectively on each feature vector and 3 triplet losses on the 3 global feature vectors. More details are given in the following.

Backbone Network. Our proposed framework can take any convolutional neural network designed for image classification as backbone network, such as VGG [109] and ResNet [49]. To conduct a fair comparison, we follow previous state-of-the-art methods [128, 118] and use a ResNet-50 as our backbone. Two modifications are conducted: (1) the down-sampling with stride-2 convolution is replaced by a stride-1 convolution in the conv5_1 layer. (2) all the layers after conv4_1 layer are duplicated to form 3 independent branches. With these modifications, more high level features can be kept in the feature map.

Multiple Spatial-channel Partitions in a Pyramidal structure. To take full advantage of information contained in the feature map, global, spatial and channel partitioned features should be trained separately in the network. A pyramidal structure has proven to be beneficial for part based models in the previous state-of-the-art [162]. Thus, the second feature map is partitioned equally into 2 spatial parts and 2 channel groups. Similarly, the third feature map is partitioned into 3 spatial parts and 3 channel groups. With GMP, each partitioned map is transformed to a vector. Besides these local feature vectors, a global feature vector is extracted from each unpartitioned feature map. In total, there are 3 global vectors and 10 local vectors.

3.3.2 Loss Functions

Softmax Cross-Entropy loss. The Softmax Cross-Entropy loss in a mini-batch can be described as:

$$L_{CE} = -\sum_{i=1}^{N_i} \log\left(\frac{\exp(x[y])}{\sum_{j=1}^{N_{id}} \exp(x[j])}\right) \tag{3.1}$$

where N_i denotes the number of images in the mini-batch, N_{id} is the number of identities in the whole training set. y is the ground truth identity of input

image and $x[j]$ represents the output of fully-connected layer for jth identity.

Triplet loss. For a better performance on hard samples, the variant Batch Hard [51] is adopted. In a mini-batch which contains P identities and K images for each identity, Batch Hard triplet loss aims at pulling the hardest positive pair (a, p) together while pushing the hardest negative pair (a, n) away by a margin. A Batch Hard triplet loss can be defined as:

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K}[\max_{p=1,...,K} \|\mathbf{a_i} - \mathbf{p_i}\|_2 - \min_{\substack{n=1,...,K \\ j=1,...,P \\ j\neq i}} \|\mathbf{a_i} - \mathbf{n_j}\|_2 + \alpha]_+ \tag{3.2}$$

where $\mathbf{a_i}$, $\mathbf{p_i}$ and $\mathbf{n_i}$ are the feature vectors of anchor, positive and negative samples respectively, and α is the margin to control the distance between positive and negative pair.

Total loss. Training the SCR model on global and local features jointly helps to build more robust representations. Local features extracted from small parts are sensitive to misalignment and viewpoint changes. Searching for the hardest positive and negative pairs with local features can be challenging, for example, we can not only look at the upper body when two people wear similar white T-shirts. Thus, the triplet loss is only employed on global features. Softmax cross-entropy loss helps to estimate the presence of specific features in small parts, which makes it more suitable for local features.

$$L_{total} = \lambda\frac{1}{N_{CE}}\sum_{i=1}^{N_{CE}} L_{CE} + \frac{1}{N_{triplet}}\sum_{i=1}^{N_{triplet}} L_{triplet} \tag{3.3}$$

where N_{CE} and $N_{triplet}$ are the number of softmax cross entropy losses and triplet losses respectively. In the SCR model, we have $N_{CE} = 13$ and $N_{triplet} = 3$. Parameter λ balances the contribution of two types of loss functions. Several possibilities of λ are tested in the next section to find an optimal setting for all experiments.

3.4 Experiments

3.4.1 Implementation Details

First of all, input images are resized to $384*192$. For the backbone network, we use a ResNet-50 pretrained on ImageNet [26] to accelerate the training process. All the layers after conv4_1 are duplicated into 3 independent branches. Each 1*1 convolutional layer is followed by a Batch Normalization [60] layer and a fully connected layer. These layers do not share weights. Following previous state-of-the-art methods [128, 118], we apply a standard Random Horizontal Flip for data augmentation. The batch size is set to 32 with randomly selected 8 identities and 4 images for each identity. We train our model with an Adam optimizer with AMSGrad setting [104] for 500 epochs. The weight decay factor for L2 regularization is set to 5e-4. The initial learning rate is set to 2e-4 and

decay to 2e-5 after 300 epochs and to 2e-6 after 400 epochs. The margin α in triplet loss is set to 1.2 in all experiments and the parameter λ in total loss is set to 2. For the evaluation, we concatenate all the feature vectors after Batch Normalization layer together as the appearance representation for images in query and gallery sets. Our model is implemented on PyTorch framework and takes about 6 hours on a single NVIDIA 1080 Ti GPU for training on Market-1501 dataset.

3.4.2 Datasets and Protocols

To validate the effectiveness of our proposed SCR model, experiments are conducted on four mainstream Re-ID datasets: Market-1501 [164], DukeMTMC-reID [105, 168], CUHK03 [80] and MARS [163].

Image based datasets. Market-1501 dataset is collected in front of a supermarket in Tsinghua University. It contains 19,732 images of 751 identities in the training set and 12,936 images of 750 identities in the testing set. There are 17.2 images per identity in the training set. DukeMTMC-reID is a subset of the DukeMTMC dataset. It contains 16,522 images of 702 persons in the training set and 2,228 query images and 17,661 gallery images of 702 persons for testing. There are 23.5 images per identity in the training set. CUHK03 contains 14,096 images of 1,467 identities captured from Chinese University of Hong Kong campus. Each identity is captured from two cameras and has an average of 4.8 images in each camera. CUHK03 dataset provides both manually labeled bounding boxes and DPM [34] detected bounding boxes.

Video based dataset. MARS is an extension of the Market-1501 dataset. There are 509,914 bounding boxes for training, belonging to 8,298 tracklets of 625 identities. There are 681,089 bounding boxes for test (gallery+query), belonging to 12,180 tracklets of 636 identities.

Evaluation Protocols. Both Cumulative Matching Characteristics (CMC) and mean Average Precisions (mAP) are used in our experiments. CMC represents the matching accuracy of Person Re-ID and CMC at Rank1 is the most intuitive metric where each query has only one ground truth match. mAP is more appropriate for the case where each query has multiple gallery matches. On CUHK03 dataset, to simplify the evaluation procedure and meanwhile enhance the accuracy of the performance reflected by results, we employed the new protocol described in [170]. For MARS dataset, we conduct a tracklet-to-tracklet search by building an overall appearance representation on each tracklet instead of on single image. Re-ranking algorithm is not used to further improve mAP in all experiments.

3.4.3 Ablation Studies

To verify the effectiveness of each component in SCR and design an optimal architecture, we conduct extensive ablation studies on Market-1501, DukeMTMC-reID, CUHK03 and MARS datasets.

Partition Type	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
Spatial p1+p2(s2)+p3(s3)	75.9	72.1	75.6	71.8
Channel p1+p2(c2)+p3(c3)	81.5	77.4	77.9	73.9
Spatial-Channel p1+p2(s2c2)+p3(s3c3)	**83.8**	**80.4**	**82.2**	**77.6**

Table 3.2: Performance comparison (%) of different partition types (spatial partition, channel partition and spatial-channel partition) on CUHK03 dataset using the new protocol [170] where the bold font denotes the best partition type. "s2" and "s3" refer that the entire feature map is partitioned into 2 and 3 spatial parts, while "c2" and "c3" refer respectively to 2 and 3 channel groups.

Architecture	Number of Branches	Representation Dimension	Market-1501		DukeMTMC-reID		CUHK03-detected	
			Rank1	mAP	Rank1	mAP	Rank1	mAP
p1	1	256*1	89.5	73.3	83.3	66.8	52.4	46.3
p2(s2c2)	1	256*5	94.4	85.8	89.6	77.6	73.7	68.4
p3(s3c3)	1	256*7	94.5	86.3	89.2	78.7	75.1	70.7
p1+p2(s2c2)	2	256*(1+5)	94.9	87.3	89.5	78.6	74.8	70.2
p1+p2(s2c2)+p3(s3c3)	3	256*(1+5+7)	**95.7**	**89.0**	**91.1**	**81.4**	**82.2**	**77.6**

Table 3.3: Performance comparison (%) of the proposed SCR with different number of branches where the bold font denotes the best architecture. "p1", "p2" and "p3" refer to 3 feature maps in SCR. "s" and "c" represent "spatial" and "channel" respectively, followed by the number of parts. For instance, "s2" and "c2" refer that the entire feature map is partitioned into 2 spatial parts and 2 channel parts.

Loss Function	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
w/o $L_{triplet}$	76.9	73.5	75.1	70.5
$\lambda = 1$	**84.8**	**81.4**	79.5	75.5
$\lambda = 2$	83.8	80.4	**82.2**	**77.6**
$\lambda = 3$	82.2	78.8	80.7	76.8

Table 3.4: Performance comparison (%) of training SCR with different parameter values for λ from L_{total}. The bold font denotes the best parameter.

Temporal Pooling	MARS	
	Rank1	mAP
TP(R)	84.5	78.9
TPP(R)	86.6	80.8
TP(E)	85.7	79.5
TPP(E)	**87.3**	**81.3**

Table 3.5: Comparison of different temporal pooling strategies where the bold font denotes the best method. "R" and "E" refer respectively to Random Sampling and Even Sampling. "TP" refers to conventional Temporal Pooling. "TPP" refers to Temporal Partiton Pooling.

Partition Strategies. We conduct extensive experiments to validate the effectiveness of spatial-channel partition by comparing our proposed model with only spatial partitions, with only channel partitions and with spatial-channel partitions. These partition strategies are compared on the most challenging dataset CUHK03. Results are reported in Table 3.2. The model with spatial-channel partitions outperforms respectively the one with only channel partitions and the one with only spatial partitions by an average margin of 3% and 7% on CUHK03 dataset.

Pyramidal Multi-Branch Architectures. Each branch "p1", "p2" and "p3" is separately tested. As shown in Table 3.3, performances of "p2" and "p3" with spatial-channel partition have a significant improvement as compared to "p1" without partition. But the results are still below those of state-of-the-art. Thus, we adapt a pyramidal multi-branch architecture to our proposed SCR, which gives a boost to the performance of our model. We gradually increase the number of branch and report their performance in Table 3.3. Two phenomena are observed: 1) Spatial-channel partitions significantly increase the performance of the neural network in Re-ID task. 2) Multi-branch structure further enhances performance of the model. We also tested a similar architecture with 4 branches where the feature map was partitioned into 4 spatial strips and 4 channel groups. It did not give a further improvement. Thus, we set the number of branches as 3.

Parameters in Total Loss. To balance contributions of softmax cross-entropy and triplet losses, a weight parameter λ should be determined. Four possibilities $\lambda = 1, 2, 3$ and without triplet loss are tested on CUHK03 dataset with both labelled and detected bounding boxes. Results in Table 3.4 shows that SCR gets best performance with $\lambda = 2$ on detected bounding boxes, while it gets best performance with $\lambda = 1$ on labeled bounding boxes. To form a unified framework, we set $\lambda = 2$ for all experiments.

Temporal Partition Pooling (TPP). For video based Re-ID, a traditional approach for building a tracklet representation is to use a temporal average (or max) pooling on all sampled image representations for the tracklet. To generalize partition strategies for the video-based Re-ID task, we conduct a partition on the temporal dimension over the tracklet. Instead of adopting

Method	Market-1501		DukeMTMC-reID	
	Rank1	mAP	Rank1	mAP
HA-CNN [81]	91.2	75.7	80.5	63.8
Mancs [124]	93.1	82.3	84.9	71.8
PCB+RPP [118]	93.8	81.6	83.3	69.2
SCPNet-a [32]	94.1	81.8	84.4	68.5
HPM [38]	94.2	82.7	86.6	74.3
CAMA [148]	94.7	84.5	85.8	72.9
MGN [128]	**95.7**	86.9	88.7	78.4
CPM [162]	**95.7**	88.2	89.0	79.0
SCR(ours)	**95.7**	**89.0**	**91.1**	**81.4**

Table 3.6: Comparison of supervised results (%) on Market-1501 and DukeMTMC-reID dataset.

directly a temporal pooling on all sampled images in a tracklet, we firstly split the images into several sub-tracklets and use the temporal pooling separately on each sub-tracklet. Representations of sub-tracklets are concatenated together to form a final representation of the tracklet. To validate the performance of our proposed TPP, we fix the sample size to 15 and partition the 15 images into 3 groups (beginning, middle, end). Different sample size and group number are tested but they do not have a strong effect on results. A temporal average pooling is performed on each sub-tracklet. Results in Table 3.5 show that temporal partition can enhance the performance of our model for the video-based Re-ID task.

3.4.4 Comparison with State-of-the-art

We compare our proposed model SCR with current state-of-the-art methods on the 4 candidate datasets.

Results on Market-1501. Comparisons between SCR and state-of-the-art methods on Market-1501 are shown in Table 3.6. To get a better understanding on how our proposed SCR can outperform previous state-of-the-art, we compare some retrieved results between PCB [118] and our SCR in Figure 3.4. These results confirm the effectiveness of spatial-channel partition on keeping more salient information and that of pyramidal structure to deal with misalignment.

Results on DukeMTMC-reID. Results of SCR and previous state-of-the-art methods on DukeMTMC-reID dataset are reported in Table 3.6. Our SCR network also performs excellently on DukeMTMC-reID dataset. SCR outperforms the former state-of-the-art by 2.1% on Rank1 and 2.4% on mAP.

Results on CUHK03. Table 3.7 shows results on CUHK03 dataset. Due to less training samples per identity, algorithms tend to get lower scores on CUHK03, which makes CUHK03 the most challenging evaluation protocol. With the same parameter settings, SCR outperforms previous state-of-the-art

Method	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
HA-CNN [81]	44.4	41.0	41.7	38.6
PCB+RPP [118]	-	-	63.7	57.5
HPM [38]	-	-	63.9	57.5
MGN [128]	68.0	67.4	68.0	66.0
CAMA [148]	70.1	66.5	66.6	64.2
CPM [162]	78.9	76.9	78.9	74.8
SCR(ours)	**83.8**	**80.4**	**82.2**	**77.6**

Table 3.7: Comparison of supervised results (%) on CUHK03 dataset using the new protocol [170] .

Method	MARS	
	Rank1	mAP
IDE+Kissme [163]	68.3	49.3
TriNet [51]	79.8	67.7
DRSTA [77]	82.3	65.8
M3D [76]	84.4	74.0
SCR(ours)	**87.3**	**81.3**

Table 3.8: Comparison of supervised results (%) on MARS dataset.

CPM by a large margin.

Results on MARS. To validate the adaptability of our model in the video-based Re-ID task, we conduct experiments on MARS dataset and report results in Table 3.8. Our model is able to outperform current state-of-the-art video-based models. SCR outperforms the previous most performant model M3D [76] by a large margin.

Unsupervised cross-domain results. Our proposed method also shows a strong generalizability on unsupervised cross-domain problem, in which a model is trained on a source domain and tested on a target domain. We compare results of SCR and unsupervised cross-domain methods in Table 3.9. Without using unlabeled images in target domain like [27, 172, 86] or extra attribute annotation [129], our SCR outperforms previous state-of-the-art under a direct deployment (no re-training on target domain) setting.

3.5 Conclusion

In this chapter, we carry out a comparative study between spatial, channel partitions and attention mechanism. Based on this study, a novel end-to-end trainable Spatial and Channel partition Representation network (SCR) is proposed to maintain salient local information by spatial-channel partitions. By

Method	M → D		D → M	
	Rank1	mAP	Rank1	mAP
SPGAN [27]	41.1	22.3	51.5	22.8
TJ-AIDL [129]	44.3	23.0	58.2	26.5
ATNet [86]	45.1	24.9	55.7	25.6
HHL [172]	46.9	27.2	**62.2**	**31.4**
SCR(ours)	**53.6**	**32.4**	59.7	30.6

Table 3.9: Comparison of unsupervised cross-domain results (%). M → D refers to training on Market-1501 and testing on DukeMTMC-reID. D → M refers to training on DukeMTMC-reID and testing on Market-1501.

combining spatial-channel partitioned local features with global features, our SCR model is able to build a discriminative and generalizable representation for each sample for Re-ID task. In addition, to address the misalignment problem, we use spatial-channel partitions in a pyramidal multi-branch architecture, which can further improve the robustness of local features. To get a better performance in video based Re-ID, partition is extended to the temporal dimension. The effectiveness of each proposed component is validated in the ablation studies. Crucial components, like spatial-channel partition and temporal partition pooling, can be easily embedded into other part based models for Re-ID. By incorporating all these components, our well-designed method outperforms current state-of-the-art in both image and video based supervised Re-ID task, as well as in unsupervised cross domain task. Another approach of enhancing feature-level diversity is presented in the next chapter.

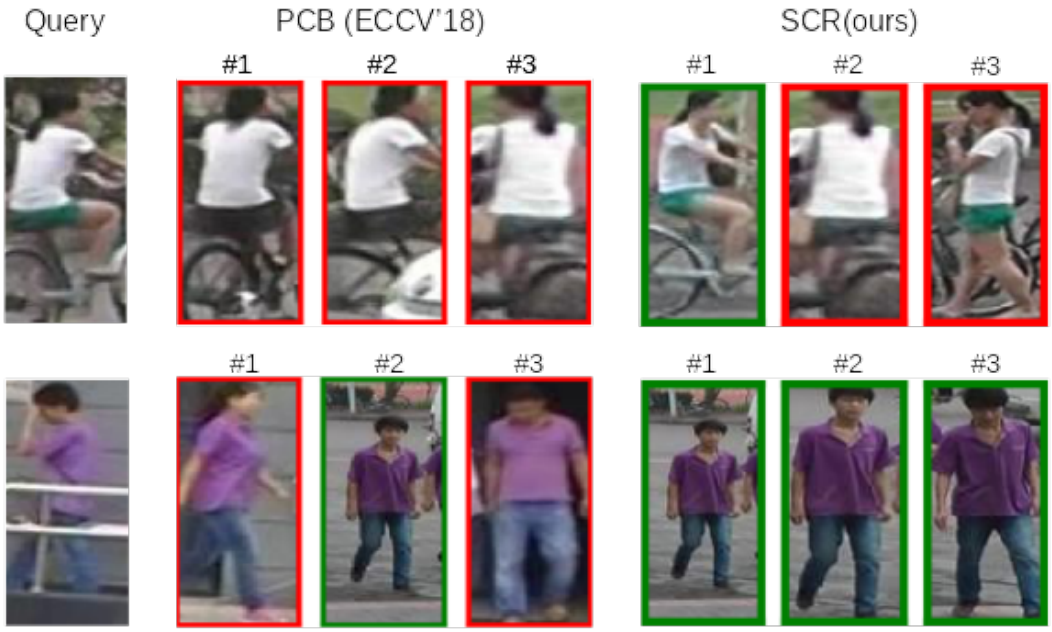


Figure 3.4: Examples of several mismatched samples in PCB [118] on Market-1501 dataset, which are addressed by our proposed SCR. Red borders refers to mismatched samples. "#1", "#2" and "#3" correspond to top 3 retrieved gallery samples.

Chapter 4

Asymmetric branches for unsupervised person ReID

4.1 Introduction

Enhancing feature-level diversity improve the performance of not only supervised person ReID, but also unsupervised person ReID. As our second contribution, we propose a new decoupling method, namely asymmetric branches, to avoid the undesired weight consensus in teacher-student networks. Asymmetric branches have different depths and global pooling methods, which encourage the feature diversity in knowledge distillation. The work presented in this chapter has been published in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2021 [13].

Since there are domain gaps resulting from illumination condition, camera property and view-point variation, a Re-ID model trained on a source domain usually shows a huge performance drop on other domains.

Unsupervised domain adaptation (UDA) targets at shifting the model trained from a source domain with identity annotation to a target domain via learning from unlabeled target images. In the real world, unlabeled images in a target domain can be easily recorded, which is almost labor-free. It is intuitive to use these images to adapt a pretrained Re-ID model to the desired domain. Fully unsupervised Re-ID further minimises the supervision by removing pre-training on the labelled source domain.

State-of-the-art UDA Person Re-ID methods [39, 146] and unsupervised methods [85] assign pseudo labels to unlabeled target images. The generated pseudo labels are generally very noisy. The noise is mainly from several inevitable factors, such as the strong domain gaps and the imperfection of clustering. In this way, an unsupervised Re-ID problem is naturally transferred into Generating pseudo labels and Learning from noisy labels problems.

To generate pseudo labels, the most intuitive way is to use a clustering algorithm, which gives a good starting point for clustering based UDA Re-ID [158, 36]. Recently, Ge *et al.* [39] propose to add a Mean Teacher [120] model as online soft pseudo label generator, which effectively reduces the error amplification during the training with noisy labels. In this chapter, we also use

both clustering-based hard labels and teacher-based soft labels in our baseline.

To handle noisy labels, one of the most popular approaches is to train paired networks so that each network helps to correct its peer, e.g., two-student networks in Co-teaching [47] and two-teacher-two-student networks in MMT [39]. However, these paired models with identical structure are prone to converge to each other and get stuck in a local minimum. There are several attempts to alleviate this problem, such as Co-teaching+ [154], ACT [146] and MMT [39]. These attempts of keeping divergence between paired models are mainly based on either different training sample selection [154, 146] or different initialization and data augmentation[39]. In this chapter, we propose a strong alternative by designing asymmetric neural network structure in the Mean Teacher Model. We use two independent branches with different depth and global pooling methods as last layers of a neural network. Features extracted from both branches are concatenated as the appearance signature, which enhances the feature diversity in the appearance signature and allows to get better clustering-based hard labels. Each branch gets supervision from its peer branch of different structure, which enhances the divergence between paired teacher-student networks. Our proposed decoupling method does not rely on different source domain initializations, which makes it more effective in the fully unsupervised scenario where the source domain is not available.

In summary, our contributions are:

1. We propose to enhance the feature diversity inside person Re-ID appearance signatures by splitting last layers of a backbone network into two asymmetric branches, which increases the quality of clustering-based hard labels.
2. We propose a novel decoupling method where asymmetric branches get cross-branch supervision, which avoids weights in paired teacher-student networks converging to each other and increases the quality of teacher-based soft labels.
3. Extensive experiments and ablation study are conducted to validate the effectiveness of each proposed component and the whole framework.

4.2 Related Work

Unsupervised domain adaptive Re-ID. Recent unsupervised cross-domain Re-ID methods can be roughly categorized into distribution alignment and pseudo label based adaptation. The objective of distribution alignment is to learn domain invariant features. Several attempts [129, 83] leverage semantic attributes to align the feature distribution in the latent space. However, these approaches strongly rely on extra attribute annotation, which requires extra labor. Another possibility is to align the feature distribution by transferring labeled source domain images into the style of target domain with generative adversarial networks [138, 172, 22]. Style transferred images are usually combined with pseudo label based adaptation to get a better performance. Pseudo label

based adaptation is a more straightforward approach for unsupervised cross-domain Re-ID, which directly assigns pseudo labels to unlabelled target images and allows to fine-tune a pre-trained model in a supervised manner. Clustering algorithms are widely used in previous unsupervised cross-domain Re-ID methods. UDAP [112] provides a good analysis on clustering based adaptation and use a k-reciprocal encoding [170] to improve the quality of clusters. PCB-PAST [158] simultaneously learns from a ranking-based and clustering-based triplet losses. SSG [36] assigns clustering-based pseudo labels to both global and local features. To mitigate the clustering-based label noise, researchers borrow ideas from how unlabeled data is used in Semi-supervised learning and Learning from noisy labels. ECN [173] uses an exemplar memory to save averaged features to assign soft labels. ACT [146] splits the training data into inliers/outliers to enhance the divergence of paired networks in Co-teaching [47]. MMT [39] adopts two student and two Mean Teacher networks. Two students are initialized differently from source pre-training in order to enhance the divergence of paired teacher-student networks. Each mean teacher network provides soft labels to supervise peer student network. However, despite different initializations and different data augmentations used in peer networks, the decoupling is not encouraged enough during the training. We directly use asymmetric neural network structure inside teacher-student networks, which encourages the decoupling at all epochs.

Teacher-Student Network for Semi-Supervised Learning. Unsupervised domain adaptation can be regarded to some extent as Semi-Supervised Learning (SSL), since both of them utilize labeled data (source domain for UDA) and large amount of unlabeled data (target domain for UDA). A teacher-student structure is commonly used in SSL. This structure allows student network to gradually exploit data with perturbations under consistency constraints. In Π model and Temporal ensembling [70], the student learns from either samples forwarded twice with different noise or exponential moving averaged (EMA) predictions under consistency constraints. Instead of EMA predictions, Mean-teacher model [120] uses directly the EMA weights from the student to supervise the student under a consistency constraint. Authors of Dual student [67] point out that the Mean Teacher converging to student along with training (coupling problem) prevents the teacher-student from exploiting more meaningful information from data. Inspired by Deep Co-training [101], they propose to train two independent students on stable samples which have same predictions and enough large feature difference. However, in unsupervised cross-domain Re-ID, labeled source domain and unlabeled target domain do not share the same identity classes, which makes traditional close-set SSL methods hard to use.

Fully unsupervised Re-ID. Recently, several fully unsupervised Re-ID methods are proposed to further minimize the supervision, which does not require any Re-ID annotation. A bottom-up clustering framework is proposed in BUC [84], which trains a network based on the clustering-based pseudo

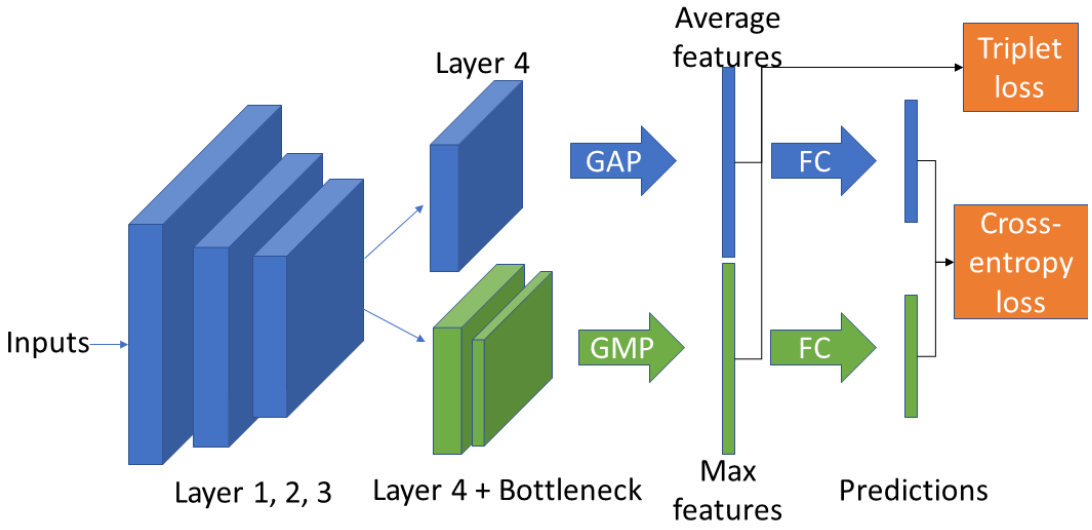


Figure 4.1: Source domain pre-training for asymmetric branched network. One ResNet bottleneck block corresponds to three convolutional layers. For UDA setting, inputs are labelled images from source training set. GAP refers Global Average Pooling, while GMP refers to Global Max Pooling. FC refers to Fully Connected layer.

labels in an iterative way. [85] replaces clustering-based pseudo labels with similarity-based softened labels. Different to image-based unsupervised Re-ID, [139] learns tacklet information with clustering-based pseudo labels. MMT [39] can be transferred into an unsupervised method by removing the pre-training in source domain. However, without different source domain initializations, divergence between peer networks can not be enough encouraged in MMT. Instead of different source domain initializations, divergence is encouraged by asymmetric network structures, which is more suitable for fully unsupervised Re-ID.

4.3 Proposed Method

4.3.1 Overview

Given two datasets: one labeled source dataset D_s and one unlabeled target dataset D_t, the objective of UDA is to adapt a source pretrained model M_{pre} to the target dataset with unlabeled target data. To achieve this goal, we propose a two-staged adaptation approach based on Mean Teacher Model. We focus on the coupling problem (teacher and student converge to each other) existing inside the original Mean Teacher. Asymmetric branches and cross-branch supervision are proposed in this chapter to address this problem and to enhance the diversity in the network, which show great effectiveness for UDA Re-ID.

4.3.2 Asymmetric branches

A multi-branch structure is widely used in the fully supervised Re-ID methods, especially in global-local feature based methods [38, 24, 12]. Such structure keeps independence between branches, which makes features extracted from different branches diversified. In the unsupervised Re-ID, we conduct clustering on appearance signatures computed from person images to generate pseudo labels. The quality of appearance signatures can be improved by extracting distinct meaningful features from different branches. Thus, we duplicate last layers of a backbone network and make them different in the structure, which we call Asymmetric Branches.

Asymmetric branches are illustrated in Figure 4.1. For a ResNet-based [49] backbone, the layer 4 is duplicated. The first branch is kept unchanged as the one used in the original backbone: 3 bottlenecks and global average pooling (GAP). The second branch is composed of 4 bottlenecks and global max pooling (GMP). The GAP perceives global information, while the GMP focuses on the most discriminative information (most distinguishable identity information, such as a red bag or a yellow t-shirt). Asymmetric branches improve appearance signature quality by enhancing the feature diversity, which is validated by source pre-training performance boost in Table 4.3 as well as examples in Figure 4.5. They further improve the quality of pseudo labels during the adaptation, which is validated by target adaptation performance in Table 4.3.

4.3.3 Asymmetric Branched Mean Teaching

We call our proposed adaptation method Asymmetric Branched Mean Teaching (ABMT). Our proposed ABMT contains two stages: Source pre-training and Target adaptation.

Source domain supervised pre-training

In the first stage, we train a network in the fully supervised way on the source domain. Thanks to this stage, the model used for adaptation obtains a basic Re-ID capacity, which helps to alleviate pseudo label noise. Given a source sample x_i^s and its ground truth identity y_i^l, the network (with weight θ) encodes x_i^s into average $F_a(x_i^s|\theta)$ and max features $F_m(x_i^s|\theta)$ and then gets two predictions $P_a(x_i^s|\theta)$ and $P_m(x_i^s|\theta)$. Cross-entropy L_{ce} and batch hard triplet [51] L_{tri} losses are used in this stage as shown in Figure 4.1.

The whole network is trained with a combination of both losses:

$$\begin{aligned} L_{scr} =& \lambda_{ce}^s L_{ce}(P_a(x_i^s|\theta), y_i^l) + \lambda_{ce}^s L_{ce}(P_m(x_i^s|\theta), y_i^l) + \\ & \lambda_{tri}^s L_{tri}(F_a(x_i^s|\theta), y_i^l) + \lambda_{tri}^s L_{tri}(F_m(x_i^s|\theta), y_i^l) \end{aligned} \tag{4.1}$$

Target domain unsupervised adaptation

The adaptation procedure is illustrated in Figure 4.2. It contains two components: Clustering-based hard label generation and Cross-branch teacher-

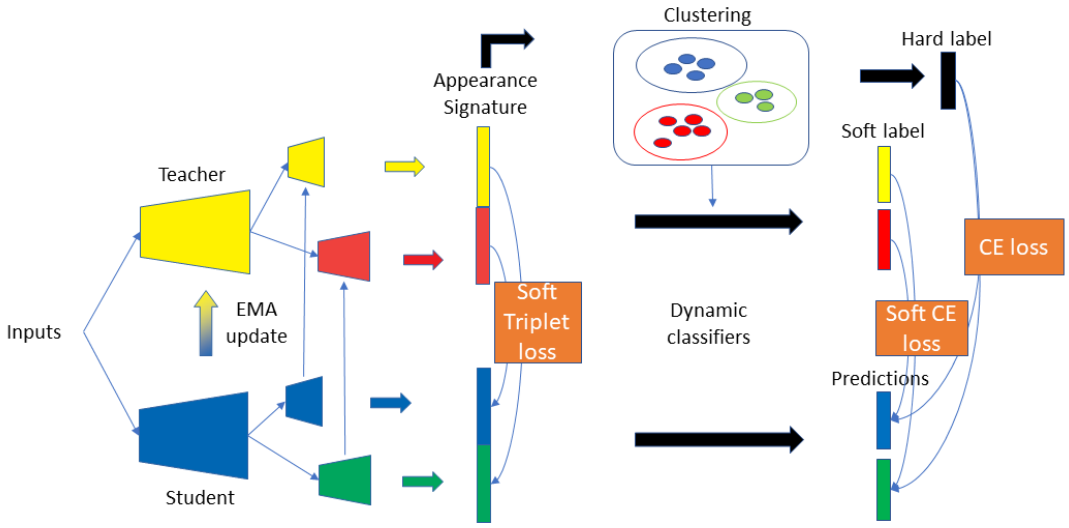


Figure 4.2: ABMT adaptation. For UDA setting, inputs are training set images from both source and target domains. For fully unsupervised setting, inputs are unlabeled images from target training set.

based soft label training. After adaptation, only teacher network is used during the inference.

Clustering-based hard label generation. In previous UDA Re-ID methods, distance-based K-Means [39] and density-based clustering DBSCAN [146, 112] are main approaches to generate pseudo labels.

We follow the state-of-the-art DBSCAN based clustering method presented in [112]. To adapt it to our proposed asymmetric branches, we concatenate the average and max features from asymmetric branches in the teacher network as appearance signatures. Images belonging to the same identity should have the same nearest neighbors in the feature space. Distance metric for DBSCAN are obtained by k-reciprocal re-ranking encoding [170] between target domain and source domain samples.

The density-based clustering generates unfixed cluster numbers at different epochs, which means old classifiers from the last epoch can not be reused after a new clustering. Thus, we simply create new classifiers depending on the number of clusters at the beginning of each epoch. We take normalized mean features of each cluster from the average branch to initialize the average branch classifiers and similarly normalized max features from max branch to initialize the max branch classifiers. We call these classifiers with flexible dimension "Dynamic Classifiers". With the help of these Dynamic Classifiers, the student is trained on cluster components (outliers are discarded) with cross-entropy loss:

$$L_{ce} = -\sum_i (y'_i \log(P_m(x_i^t|\theta))) - \sum_i (y'_i \log(P_a(x_i^t|\theta))) \tag{4.2}$$

where y'_i is the clustering based hard label and $P_a(x_i^t|\theta)$ and $P_m(x_i^t|\theta)$ are student predictions from both asymmetric branches.

Cross-branch teacher-based soft label training. Clustering algorithms generate hard pseudo labels whose confidences are 100%. Since Re-ID is a fine-grained recognition problem, people with similar clothes are not rare in the dataset. Hard pseudo labels of these similar samples can be extremely noisy. In this case, soft pseudo labels (confidences $< 100\%$) are more reliable. Learning with both hard and soft pseudo labels can effectively alleviate label noise.

The Mean Teacher Model [120] (teacher weights θ') uses the EMA weights of the student model (student weights θ). The Mean Teacher Model shows strong capacity to handle label noise and avoids error amplification along with training. We define θ'_t at training step t as the EMA of successive weights:

$$\theta'_t = \begin{cases} \theta_t, & \text{if } t = 0 \\ \alpha\theta'_{t-1} + (1-\alpha)\theta_t, & \text{otherwise} \end{cases} \tag{4.3}$$

where α is a smoothing coefficient that controls the self-ensembling speed of the Mean Teacher.

Despite these advantages of Mean Teacher, such self-ensembling teacher-student networks (the teacher is formed by EMA weights of the student, and the student is supervised by the teacher) face the coupling problem. We use the Mean Teacher soft label generator as in [39] and address the coupling problem by cross-branch supervision. Each branch in the student is supervised by a teacher branch which has different structure. Weight diversity between the paired teacher-student can be better kept. Given one target domain sample x_i^t, the teacher (teacher weights θ') encodes it into two feature vectors from two asymmetric branches, average features $F_a(x_i^t|\theta')$ and max features $F_m(x_i^t|\theta')$. The dynamic classifiers then transform these two feature vectors into two predictions respectively $P_a(x_i^t|\theta')$ and $P_m(x_i^t|\theta')$. Similarly, features of the student (student weights θ) are $F_a(x_i^t|\theta)$ and $F_m(x_i^t|\theta)$, while predictions are $P_a(x_i^t|\theta)$ and $P_m(x_i^t|\theta)$. The predictions from the teacher supervise those from the student with a soft cross-entropy loss [53] in a cross-branch manner, which can be formulated as

$$L_{sce}^{a\to m} = -\sum_i (P_a(x_i^t|\theta')\log(P_m(x_i^t|\theta))) \tag{4.4}$$

$$L_{sce}^{m\to a} = -\sum_i (P_m(x_i^t|\theta')\log(P_a(x_i^t|\theta))) \tag{4.5}$$

To further enhance the teacher-student networks' discriminative capacity, the features in the teacher supervise those of the student with a soft triplet loss [39]:

$$L_{stri}^{a\to m} = -\sum_i (T_a(x_i^t|\theta')\log(T_m(x_i^t|\theta))) \tag{4.6}$$

$$L_{stri}^{m\to a} = -\sum_i (T_m(x_i^t|\theta')\log(T_a(x_i^t|\theta))) \tag{4.7}$$

where $T(x_i^t|\theta) = \frac{exp(\left\|F(x_i^t|\theta) - F(x_p^t|\theta)\right\|_2)}{exp(\left\|F(x_i^t|\theta) - F(x_p^t|\theta)\right\|_2) + exp(\left\|F(x_i^t|\theta) - F(x_n^t|\theta)\right\|_2)}$ is the softmax triplet distance of the sample x_i^t, its hardest positive x_p^t and its hardest negative

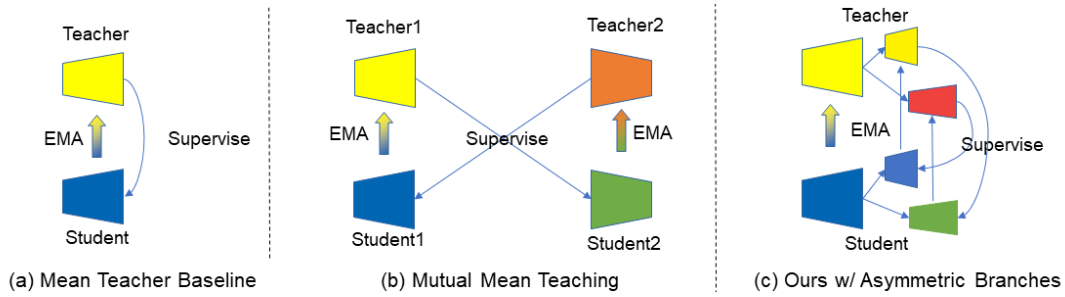


Figure 4.3: Comparison between (a) Mean Teacher Baseline (b) Mutual Mean Teaching [39] and (c) our Mean Teacher with cross-branch supervised asymmetric branches. Teacher network is formed by exponential moving average (EMA) values of student network.

x_n^t in a mini-batch. By minimizing the soft triplet loss, the softmax triplet distance in a mini-batch from the student is encouraged to get as close as possible to the distance from the teacher. The positive and negative samples within a mini-batch are decided by clustering-based hard pseudo labels. It can effectively improve the UDA Re-ID performance. The teacher-student networks are trained end-to-end with Equation (4.2), (4.4), (4.5), (4.6), (4.7).

$$\begin{aligned} L_{target} =& \lambda_{ce}^t L_{ce} + \lambda_{sce}^t (L_{sce}^{a\to m} + L_{sce}^{m\to a}) \\ &+ \lambda_{stri}^t (L_{stri}^{a\to m} + L_{stri}^{m\to a}) \end{aligned} \tag{4.8}$$

4.4 Coupling Problem in Mean Teacher Based Methods

The Mean Teacher Baseline is illustrated in Figure 4.3 (a) where the student gets supervision from its own EMA weights. In the Mean Teacher Baseline, the student and the teacher quickly converge to each other (coupling problem), which prevents them from exploring more diversified information. Authors of MMT [39] propose to pre-train 2 student networks with different seeds. As illustrated in Figure 4.3 (b), two Mean Teacher networks are formed separately from two students, which alleviates the coupling problem. However, different initializations decouple both teacher peers only at first epochs. Without a diversity encouragement during the adaptation, both teachers still converge to each other along with training. In Figure 4.3 (c), our proposed asymmetric branches provide a diversity encouragement during the adaptation, which decouples both teacher peers at all epochs.

To validate our idea, we propose to measure Euclidean distance of appearance signature features between two teacher networks or two teacher branches. We extract feature vectors after global pooling on all images in the target training set. Then, we calculate the Euclidean distance between feature vectors of both teachers and sum up the distance of every image as the final feature

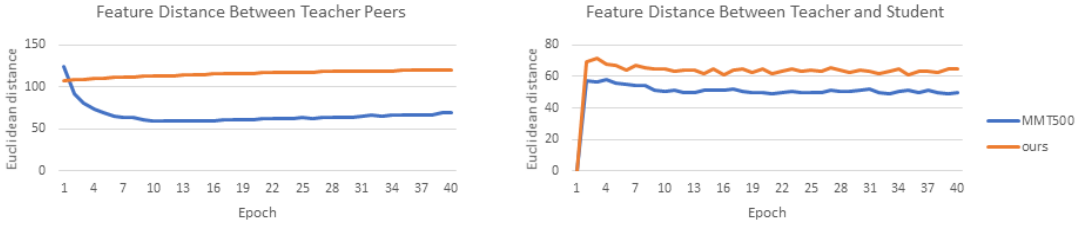


Figure 4.4: Distance comparison between features extracted from a ResNet50 backbone on all samples in DukeMTMC-reid training set for Market → Duke task. **Left**: Feature distance between two teacher models in MMT and between two teacher branches in our proposed method. **Right**: Feature distance between teacher and student networks.

distance. If the feature distance is large, we can say that both teacher peers extract diversified features. Otherwise, the teacher peers converge to each other. As we can see from the left curves in Figure 4.4, the feature distance between two teachers in MMT is large at the beginning, but it decreases and then stabilizes. Differently, the feature distance between two branches in our proposed method remains large during the training. Moreover, we visualize the Euclidean distance of appearance signature features on all target training samples between teacher and student networks in Figure 4.4 right curves. Our method can maintain a larger distance, which shows that it can better decouple teacher-student networks.

4.5 Experiments

4.5.1 Datasets and Evaluation Protocols

Our proposed adaptation method is evaluated on 3 Re-ID datasets: Market-1501, DukeMTMC-reID and MSMT17. **Market-1501** [164] dataset is collected in front of a supermarket in Tsinghua University from 6 cameras. It contains 12,936 images of 751 identities in the training set and 19,732 images of 750 identities in the testing set. **DukeMTMC-reID** [105] is a subset of the DukeMTMC dataset. It contains 16,522 images of 702 persons in the training set, 2,228 query images and 17,661 gallery images of 702 persons for testing from 8 cameras. **MSMT17** [138] is a large-scale Re-ID dataset, which contains 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras. Both Cumulative Matching Characteristics (CMC) and mean Average Precisions (mAP) are used in our experiments.

4.5.2 Implementation details

Hyper-parameters used in our proposed method are searched empirically from the Market→ Duke task and kept the same for the other tasks. To conduct fair comparison with state-of-the-arts, we use a ImageNet [26] pre-trained

UDA Methods	Market → Duke		Duke → Market		Market → MSMT		Duke → MSMT	
	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1
HHL (ECCV'18)[172]	27.2	46.9	31.4	62.2	-	-	-	-
ECN (CVPR'19)[173]	40.4	63.3	43.0	75.1	8.5	25.3	10.2	30.2
PCB-PAST (ICCV'19)[158]	54.3	72.4	54.6	78.4	-	-	-	-
SSG (ICCV'19)[36]	53.4	73.0	58.3	80.0	13.2	31.6	13.3	32.2
UDAP (PR'20)[112]	49.0	68.4	53.7	75.8	-	-	-	-
ACT (AAAI'20)[146]	54.5	72.4	60.6	80.5	-	-	-	-
ECN+ (PAMI'20) [174]	54.4	74.0	63.8	84.1	15.2	40.4	16.0	42.5
MMT500 (ICLR'20)(ResNet50)[39]	63.1	76.8	71.2	87.7	16.6	37.5	17.9	41.3
MMT700 (ICLR'20)(ResNet50)[39]	65.1	78.0	69.0	86.8	-	-	-	-
MMT1500 (ICLR'20)(ResNet50)[39]	-	-	-	-	22.9	49.2	23.3	50.1
ours (ResNet50)	**69.1**	**82.0**	**78.3**	**92.5**	**23.2**	**49.2**	**26.5**	**54.3**
MMT500 (ICLR'20)(IBN-ResNet50)[39]	65.7	79.3	76.5	90.9	19.6	43.3	23.3	50.0
MMT700 (ICLR'20)(IBN-ResNet50)[39]	68.7	81.8	74.5	91.1	-	-	-	-
MMT1500 (ICLR'20)(IBN-ResNet50)[39]	-	-	-	-	26.6	54.4	29.3	58.2
ours (IBN-ResNet50)	**70.8**	**83.3**	**80.4**	**93.0**	**27.8**	**55.5**	**33.0**	**61.8**

Table 4.1: Comparison of unsupervised domain adaptation (UDA) Re-ID methods (%) on medium-to-medium datasets (Market→ Duke and Duke → Market) and medium-to-large datasets (Market → MSMT and Duke → MSMT).

ResNet-50 [49] as our backbone network. The backbone can be extended to ResNet-based networks designed for cross domain tasks, *e.g.*, IBN-ResNet-50 [97]. An Adam optimizer with a weight decay rate of 0.0005 is used to optimize our networks. Our networks are trained on 4 Nvidia 1080Ti GPUs under Pytorch [98] framework. Detailed configurations are given in the following paragraphs.

Stage1: Source domain supervised pre-training. We set $\lambda^s_{ce} = 0.5$ and $\lambda^s_{tri} = 0.5$ in Equation 4.1. The max epoch E_{pre} is set to 80. For each epoch, the networks are trained $R_{pre} = 200$ iterations.The initial learning rate is set to 0.00035 and is multiplied by 0.1 at the 40th and 70th epoch. For each iteration, 64 images of 16 identities are resized to 256*128 and fed into networks.

Stage2: Target domain unsupervised adaptation. For the clustering, we set the minimum cluster samples to 4 and the density radius r=0.002. Re-ranking parameters for calculating distances are kept the same as in [112] for UDA setting. Re-ranking between source and target domain is not considered for fully unsupervised setting. The Mean Teacher network is initialized and updated in the way of Equation 4.3 with a smoothing coefficient $\alpha = 0.999$. We set $\lambda^t_{ce} = 0.5$, $\lambda^t_{sce} = 0.5$ and $\lambda^t_{stri} = 1$ in Equation 4.8. The adaptation epoch E_{ada} is set to 40. For each epoch, the networks are trained $R_{ada} = 400$ iterations with a fixed learning rate 0.00035. For each iteration, 64 images of 16 clustering-based pseudo identities are resized to 256*128 and fed into networks with Random erasing [171] data augmentation.

4.5.3 Comparison with State-of-the-Art Methods

We compare our proposed methods with state-of-the-art UDA methods in Table 4.1 for 4 cross-dataset Re-ID tasks: Market→ Duke, Duke → Market,

Unsupervised methods	Market		Duke	
	mAP	Rank1	mAP	Rank1
MMT500*(ICLR'20)[39]	26.9	48.0	7.3	12.7
BUC (AAAI'19)[84]	30.6	61.0	21.9	40.2
SoftSim (CVPR'20)[85]	37.8	71.7	28.6	52.5
TSSL (AAAI'20)[139]	43.3	71.2	38.5	62.2
MMT*+DBSCAN (ICLR'20)[39]	53.5	73.1	54.5	69.5
ours w/o Source pre-training	**65.1**	**82.6**	**63.1**	**77.7**

Table 4.2: Comparison of unsupervised Re-ID methods (%) with a ResNet50 backbone on Market and Duke datasets. * refers to our implementation where we remove the source pre-training step. DBSCAN refers to a DBSCAN clustering based on re-ranked distance.

Source pre-training	Market → Duke		Duke → Market	
	mAP	Rank1	mAP	Rank1
ResNet50	29.6	46.0	31.8	61.9
ResNet50+AB	31.5	49.7	33.2	63.2
Target adaptation	**Market → Duke**		**Duke → Market**	
	mAP	Rank1	mAP	Rank1
MT-Baseline+K-Means	59.9	74.8	68.9	88.2
MT-Baseline+DBSCAN	61.9	77.3	69.9	88.3
MT-Baseline+K-Means+AB	64.7	78.1	74.8	90.5
MT-Baseline+K-Means+AB+Cross-branch	66.4	79.9	76.8	91.7
MT-Baseline+DBSCAN+AB	67.8	81.1	77.3	92.0
ABMT(MT-Baseline+DBSCAN+AB+Cross-branch)	**69.1**	**82.0**	**78.3**	**92.5**
ABMT+Stochastic data augmentation	68.8	81.2	77.6	91.7
ABMT+Drop out	68.3	81.8	77.9	92.0

Table 4.3: Ablation studies with ResNet50 backbone. MT-Baseline corresponds to the Mean Teacher Baseline in Figure 4.3 (a) with a ResNet-50. K-Means refers to a K-Means++ clustering whose cluster number is set to 500. AB refers to asymmetric branches. DBSCAN refers to a DBSCAN clustering [31].

Market → MSMT and Duke → MSMT. Post-processing techniques (*e.g.*, Re-ranking [170]) are not used in the comparison. Our proposed method outperforms MMT [39] (cluster number is set to 500, 700 and 1500 respectively). We can also adjust the density radius in DBSCAN depending on target domain size to get a better performance, but we think it is hard to know the target domain size in the real world. With an IBN-ResNet50 [97] backbone, the performance on 4 tasks can be further improved. Examples of retrieved images are illustrated in Figure 4.5. Compared to MMT, embeddings from our proposed method contain more discriminative appearance information (*e.g.*, shoulder bag in the first row), which are robust to noisy information (*e.g.*, pose variation in the second row, occlusion in the third row and background variation in the fourth row). This qualitative comparison confirms that appearance signatures of our proposed method are of improved quality.

Structure	Market → Duke		Duke → Market	
	mAP	Rank1	mAP	Rank1
ABMT	**69.1**	**82.0**	**78.3**	**92.5**
ABMT w/o different pooling	65.2	79.7	74.2	90.1
ABMT w/o extra bottleneck	67.5	80.6	77.6	92.4
ABMT + one more branch	68.1	80.7	76.2	90.4

Table 4.4: Ablation studies on structure of asymmetric branches.

Loss	Market → Duke		Duke → Market	
	mAP	Rank1	mAP	Rank1
ABMT	**69.1**	**82.0**	**78.3**	**92.5**
ABMT w/o L_{ce}	52.5	69.6	57.5	79.8
ABMT w/o L_{sce}	66.7	79.8	77.7	92.2
ABMT w/o L_{stri}	64.7	78.5	75.5	91.2

Table 4.5: Ablation studies on loss functions.

We compare unsupervised Re-ID methods in Table 4.2. Since the Mean Teacher is designed for handling label noise, it is interesting to see the performance without source pre-training, which introduces more label noise during the adaptation. This setting corresponds to an unsupervised Re-ID. We use ImageNet pretained weights as initialization. Our proposed method outperforms previous unsupervised Re-ID by a large margin, which shows that ImageNet initialization can provide basic discriminative capacity for Re-ID.

MMT [39] is the first Mean Teacher based UDA Re-ID method. Authors of MMT propose to use 2 students and 2 teachers with different initialization and stochastic data augmentation to address the coupling problem. We also use Mean Teacher soft pseudo labels but propose a different decoupling solution. Features in asymmetric branches are always extracted in different manners during the adaptation. Compared to MMT, our proposed method has less parameters (approximately 10% less parameters and 20% less operations) but achieves better performance. Moreover, in the unsupervised scenario, we can not pre-train MMT with different seeds to obtain different Re-ID initializations. This decoupling strategy becomes inappropriate. Our decoupling strategy relies on structural asymmetry instead of different initializations, which is much more effective in the unsupervised scenario.

ACT [146] uses 2 networks, in which each network learns from its peer. Input data are split into inliers and ouliers after DBSCAN. Then, the first network selects small entropy inliers to train the second network, while the second selects small entropy outliers to train the first. This method enhances input asymmetry by data split. Differently, our proposed method focuses on neural network structure asymmetry.

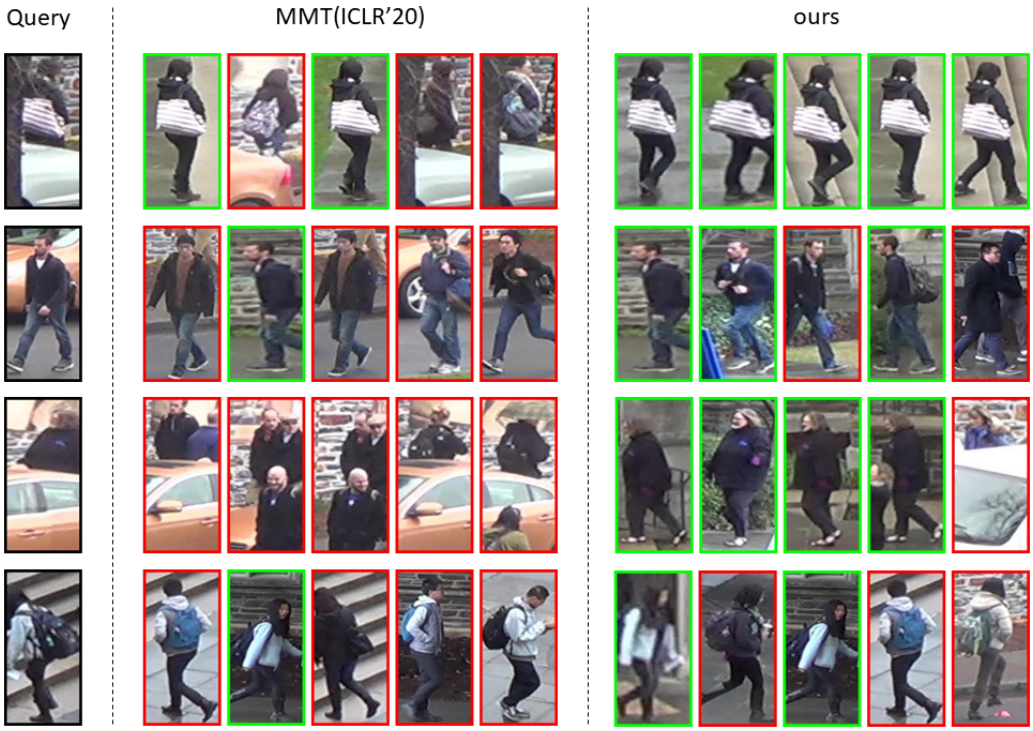


Figure 4.5: Examples of retrieved most similar 5 images in Market $\rightarrow$ Duke task from MMT [39] and our proposed method. Given a query image, different identity images are highlighted by red bounding boxes, while same identity images are highlighted by green bounding boxes.

4.5.4 Ablation Studies

Effectiveness of each component in ABMT. Compared with traditional clustering-based Re-ID methods, the performance improvement mainly comes from DBSCAN on re-ranked distance, asymmetric branches and cross-branch supervision. We use a Mean Teacher Baseline where original ResNet-50 and a K-Means++ clustering of 500 clusters are adopted. We conduct ablation studies by gradually adding one component at each time. Results are shown in Table 4.3. We can observe: (1) Our proposed asymmetric branches bring the most significant performance improvement during the adaptation. Moreover, as we can see from first two rows in Table 4.3, they can directly improve the domain generalizability of appearance signatures without target adaptation. (2) DBSCAN on re-ranked distance works better than a K-Means++ clustering of 500 clusters during the adaptation. (3) Cross-branch supervision works on asymmetric branches, which can further improve the adaptation performance.

Effectiveness of asymmetric branch structure. To validate the effectiveness of our proposed asymmetric branch structure, we compare several possible structures: (1) 2 branches with different pooling methods and different depths, (2) 2 branches with same pooling methods (global average pooling) but different depths, (3) 2 branches with different methods but same depths, (4) 3 branches where the new branch is composed of 5 bottleneck blocks and global average pooling. From results given in Table 4.4, we can conclude that different pooling methods play a more important role in asymmetric branches.

Effectiveness of loss functions. We conduct ablation studies on loss functions used in our proposed method and report results in Table 4.5. The degree of influence of 3 loss functions used in our proposed method: $L_{ce} > L_{sce} > L_{stri}$.

Can traditional decoupling methods further improve the performance? Stochastic data augmentation (teacher inputs and student inputs are under stochastic data augmentation methods) and drop out (teacher feature vectors and student feature vectors are under independent drop out operations before classifiers) are 2 widely-used methods to provide random noise, which also helps to decouple the weights between the teacher and the student. We conduct experiments with stochastic data augmentation . The results in Table 4.3 show that they can not further improve the UDA Re-ID performance. These methods are not designed for fine-grained Re-ID task. As UDA Re-ID performance is already very high, they can not contribute anymore.

4.6 Conclusion

In this chapter, we propose a novel unsupervised cross-domain Re-ID framework. Our proposed method is mainly based on learning from noisy pseudo labels generated by clustering and Mean Teacher. A self-ensembled Mean

Teacher is robust to label noise, but the coupling problem inside paired teacher-student networks leads to a performance bottleneck. To address this problem, we propose asymmetric branches and cross-branch supervision, which can effectively enhance the diversity in two aspects: appearance signature features and teacher-student weights. By enhancing the diversity in the teacher-student networks, our proposed method achieves better performance on both unsupervised domain adaptation and fully unsupervised Re-ID tasks. Our proposed decoupling method augments feature-level diversity. Augmentation can be also conducted on image-level to enhance the data diversity. In the next chapter, we use a GAN to conditionally generate more images as augmented views for contrastive learning to improve the performance of unsupervised Re-ID.

Chapter 5

Joint generative and contrastive learning for unsupervised person ReID

5.1 Introduction

The contributions described in the two previous sections enriches feature-level diversity via operations on feature maps and network architecture. A more advanced image-level data augmentation technique is explored in this chapter, which consists in using GAN to conditionally generate person images as augmentation. This work has been published in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021 [15].

Recently, self-supervised contrastive methods [48, 17] have provided an effective retrieval-based approach for unsupervised representation learning. Given an image, such methods maximize agreement between two augmented views of one instance (see Fig. 5.1). *Views* refer to transformed versions of the same input. As shown in very recent works [17, 20], data augmentation enables a network to explore view-invariant features by providing augmented views of a person, which are instrumental in building robust representations. Such and similar methods considered traditional data augmentation techniques, *e.g.*, 'random flipping', 'cropping', and 'color jittering'. Generative Adversarial Networks (GANs) [45] constitute a novel approach for data augmentation. As opposed to traditional data augmentation, GANs are able to modify id-unrelated features substantially, while preserving id-related features, which is highly beneficial in contrastive ReID.

Previous GAN-based methods [3, 28, 179, 73, 138, 172] considered unsupervised ReID as an unsupervised domain adaptation (UDA) problem. Under the UDA setting, researchers used both, a labeled source dataset, as well as an unlabeled target dataset to gradually adjust a model from a source domain into a target domain. GANs can be used in cross-domain style transfer, where labeled source domain images are generated in the style of a target domain. However, the UDA setting necessitates a large-scale labeled source dataset. Scale and quality of the source dataset strongly affect the performance of UDA methods.

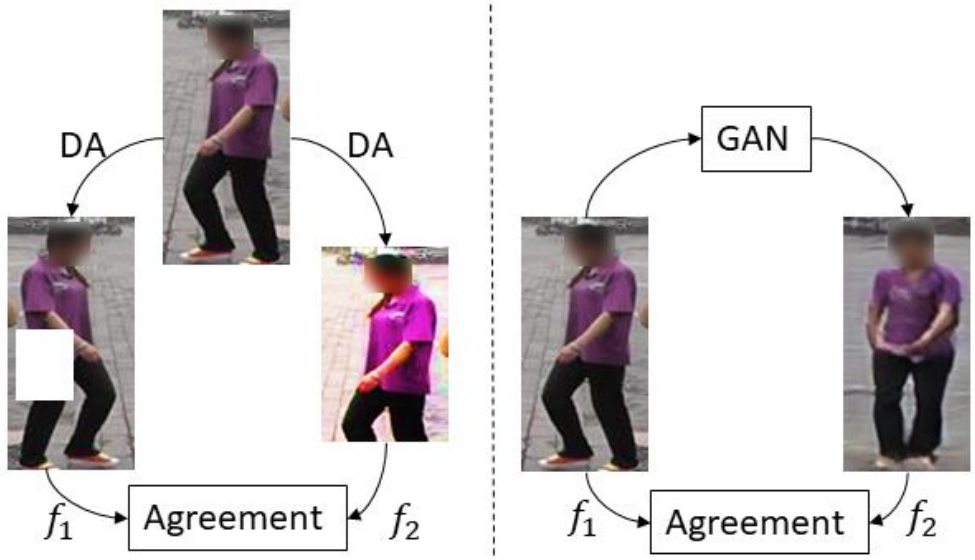


Figure 5.1: **Left**: Traditional self-supervised contrastive learning maximizes agreement between representations (f_1 and f_2) of augmented views from Data Augmentation (DA). **Right**: Joint generative and contrastive learning maximizes agreement between original and generated views.

Recent research has considered fully unsupervised ReID [125, 75], where under the fully unsupervised setting, a model directly learns from unlabeled images without any identity labels. Self-supervised contrastive methods [48, 17] belong to this category. In this work, we use a GAN as a novel view generator for contrastive learning, which does not require a labeled source dataset.

Here, we aim at enhancing view diversity for contrastive learning via generation under the fully unsupervised setting. Towards this, we introduce a mesh-based novel view generator. We explore the possibility of disentangling a person image into identity features (color distribution and body shape) and structure features (pose and view-point) under the fully unsupervised ReID setting. We estimate 3D meshes from unlabeled training images, then rotate these 3D meshes to simulate new structures. Compared to skeleton-guided pose transfer [40, 73], which neglects body shape, mesh recovery [64] jointly estimates pose and body shape. Estimated meshes preserve body shape during the training, which facilitates the generation and provides more visual clues for fine-grained ReID. Novel views can be generated by combining identity features with new structures.

Once we obtain the novel views, we design a pseudo label based contrastive learning module. With the help of our proposed view-invariant loss, we maximize representation similarity between original and generated views of a same person, whereas representation similarity of other persons is minimized.

Our proposed method incorporates generative and contrastive modules into one framework, which are trained jointly. Both modules share the same identity feature encoder. The generative module disentangles identity and structure features, then generates diversified novel views. The novel views are then used in the contrastive module to improve the capacity of the shared identity feature

encoder, which in turn improves the generation quality. Both modules work in a mutual promotion way, which significantly enhances the performance of the shared identity feature encoder in unsupervised ReID. Moreover, our method is compatible with both UDA and fully unsupervised settings. With a labeled source dataset, we obtain better performance by alleviating the pseudo label noise.

Our contributions can be summarized as follows.

1. We propose a joint generative and contrastive learning framework for unsupervised person ReID. Generative and contrastive modules mutually promote each other's performance.
2. In the generative module, we introduce a 3D mesh based novel view generator, which is more effective in body shape preservation than skeleton-guided generators.
3. In the contrastive module, a view-invariant loss is proposed to reduce intra-class variation between original and generated images, which is beneficial in building view-invariant representations under a fully unsupervised ReID setting.
4. We overcome the limitation of previous GAN-based unsupervised ReID methods that strongly rely on a labeled source dataset. Our method significantly surpasses the performance of state-of-the-art methods under both, fully unsupervised, as well as UDA settings.

5.2 Related Work

Unsupervised representation learning. Recent contrastive instance discrimination methods [143, 48, 17] have witnessed a significant progress in unsupervised representation learning. The basic idea of instance discrimination has to do with the assumption that each image is a single class. Contrastive predictive coding (CPC) [96] included an InfoNCE loss to measure the ability of a model to classify positive representation amongst a set of unrelated negative samples, which has been commonly used in following works on contrastive learning. Recent contrastive methods treated unsupervised representation learning as a retrieval task. Representations can be learnt by matching augmented views of a same instance from a memory bank [143, 48] or a large mini-batch [17]. MoCoV2 [20] constitutes the improved version of the MoCo [48] method, incorporating larger data augmentation. We note that data augmentation is pertinent in allowing a model to learn robust representations in contrastive learning. However, only traditional data augmentation was used in aforementioned methods.

Data augmentation. MoCoV2 [20] used 'random crop', 'random color jittering', 'random horizontal flip', 'random grayscale' and 'gaussian blur'. However, 'random color jittering' and 'grayscale' were not suitable for fine-grained person ReID, because such methods for data augmentation tend to change the color

distribution of original images. In addition, 'Random Erasing' [171] has been a commonly used technique in person ReID, which randomly erases a small patch from an original image. Cross-domain Mixup [89] interpolated source and target domain images, which alleviated the domain gap in UDA ReID. Recently, Generative Adversarial Networks (GANs) [45] have shown great success in image [66, 65, 6] and video synthesis [122, 132, 8, 134, 133]. GAN-based methods can serve as a method for evolved data augmentation by conditionally modifying id-unrelated features (style and structure) for supervised ReID. CamStyle [175] used the CycleGAN-architecture [177] in order to transfer images from one camera into the style of another camera. FD-GAN [40] was targeted to generate images in a pre-defined pose, so that images could be compared in the same pose. IS-GAN [30] was streamlined to disentangle id-related and id-unrelated features by switching both local and global level identity features. DG-Net [165] recolored grayscale images with a color distribution of other images, targeting to disentangle identity features. Deviating from such supervised GAN-based methods, our method generates novel views by rotating 3D meshes in an *unsupervised* manner.

Unsupervised person ReID. Recent unsupervised person ReID methods were predominantly based on UDA. Among UDA-based methods, several works [129, 83] used semantic attributes to facilitate domain adaptation. Other works [141, 37, 13, 147, 39] assigned pseudo labels to unlabeled images and proceeded to learn representations with pseudo labels. Transferring source dataset images into the style of a target dataset represents another line of research. SPGAN [28] and PTGAN [138] used CycleGAN [177] as domain style transfer-backbone. HHL [172] aims at transferring cross-dataset camera styles. ECN [173, 174] exploited invariance from camera style transferred images for UDA ReID. CR-GAN [22] employed parsing-based masks to remove noisy backgrounds. PDA [73] included skeleton estimation to generate person images with different poses and cross-domain styles. DG-Net++ [179] jointly disentangled id-related/id-unrelated features and transferred domain styles. While the latter is related to our method, we aim at training jointly a GAN-based online data augmentation, as well as a contrastive discrimination, which renders the labeled source dataset unnecessary, rather than transferring style.

Fully unsupervised methods do not require any identity labels. BUC [84] represented each image as a single class and gradually merged classes. In addition, TSSL [140] considered each tracklet as a single class to facilitate cluster merging. SoftSim [85] utilized similarity-based soft labels to alleviate label noise. MMCL [125] assigned multiple binary labels and trained a model in a multi-label classification way. JVTC and JVTC+ [75] added temporal information to refine visual similarity based pseudo labels. We note that all aforementioned fully unsupervised methods learn from pseudo labels. We show in this work that disentangling view-invariant identity features is possible in fully unsupervised ReID, which can be an add-on to boost the performance of previous pseudo label based methods.

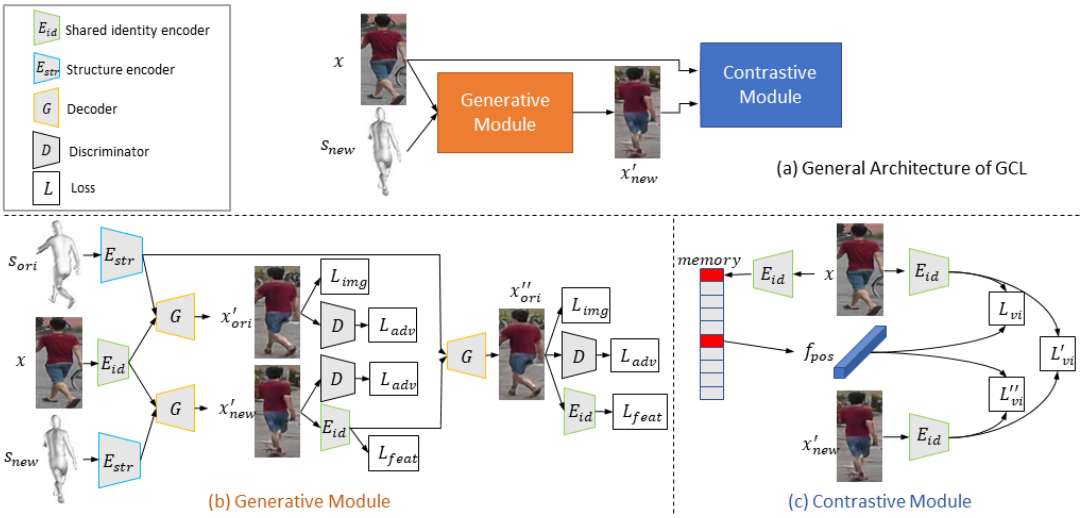


Figure 5.2: **(a) General architecture of GCL**: Generative and contrastive modules are coupled by the shared identity encoder E_{id}. **(b) Generative module**: The decoder G combines the identity features encoded by E_{id} and structure features E_{str} to generate a novel view x'_{new} with a cycle consistency. **(c) Contrastive module**: View-invariance is enhanced by maximizing the agreement between original $E_{id}(x)$, synthesized $E_{id}(x'_{new})$ and memory f_{pos} representations.

5.3 Proposed Method

We refer to our proposed method as joint *Generative and Contrastive Learning* as GCL. The general architecture of GCL comprises of two modules, namely a View Generator, as well as a View Contrast Module, see Fig. 5.2. Firstly, the View Generator uses cycle-consistency on both, image and feature reconstructions in order to disentangle identity and structure features. It combines identity features and mesh-guided structure features to generate one person in new view-points. Then, original and generated views are exploited as positive pairs in the View Contrast Module, which enables our network to learn view-invariant identity features. We proceed to elaborate on both modules in the following.

5.3.1 View Generator (Generative Module)

As shown in Fig. 5.2, the proposed View Generator incorporates 4 networks: an identity encoder E_{id}, a structure encoder E_{str}, a decoder G and an image discriminator D. Given an unlabeled person ReID dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, we generate corresponding 3D meshes with a popular 3D mesh generator Human Mesh Recovery (HMR) [64], which simultaneously estimates body shape and pose from a single RGB image. Here, we denote the 2D projection of a 3D mesh as original structure s_{ori}. Then, as depicted in Fig. 5.3, we rotate each 3D mesh by $45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ$ and 315°, respectively and proceed to randomly pick one 2D projection of these rotated meshes as a new structure

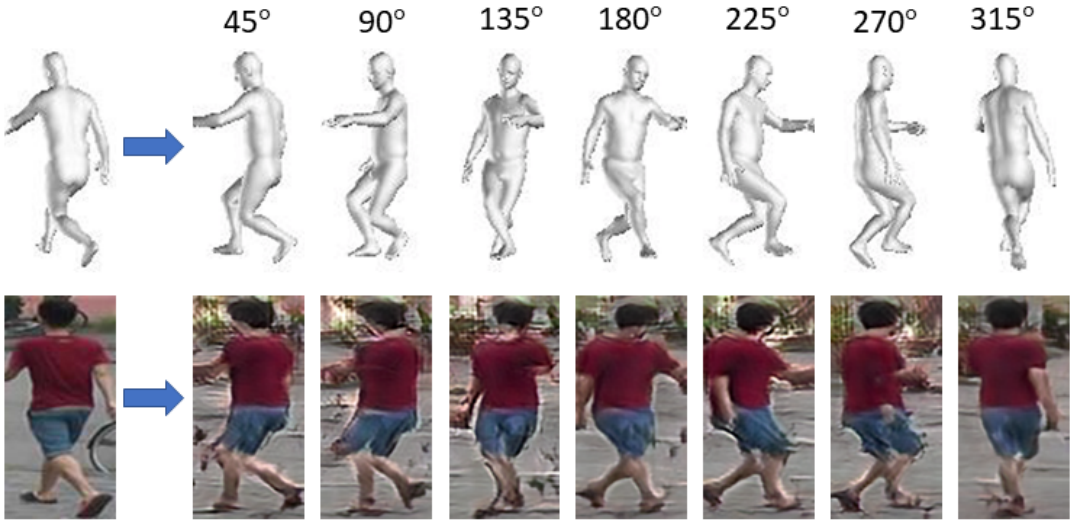


Figure 5.3: Example images as generated by the View Generator via 3D mesh rotation based on left input image.

s_{new}. We use the 3D mesh rotation to mimic view-point variance from different cameras. Next, unlabeled images are encoded to identity features by the identity encoder $E_{id} : x \rightarrow f_{id}$, while both original and new structures are encoded to structure features by the structure encoder $E_{str} : s_{ori} \rightarrow f_{str(ori)}, s_{new} \rightarrow f_{str(new)}$. Combining both, identity and structure features, the decoder generates synthesized images $G : (f_{id}, f_{str(ori)}) \rightarrow x'_{ori}, (f_{id}, f_{str(new)}) \rightarrow x'_{new}$, where a prime is used to represent generated images.

Given the lack of real images corresponding to the new structures, we consider a cycle consistency [177] to reconstruct the original image by swapping the structure features in the View Generator. We encode and decode once again to get synthesized images in original structures $G(E_{id}(x'_{new}), s_{ori}) \rightarrow x''_{ori}$. We calculate an image reconstruction loss as follows.

$$\mathcal{L}_{img} = \mathbb{E}[\|x - x'_{ori}\|_1] + \mathbb{E}[\|x - x''_{ori}\|_1] \tag{5.1}$$

In addition, we compute a feature reconstruction loss

$$\begin{aligned} \mathcal{L}_{feat} = &\mathbb{E}[\|f_{id} - E_{id}(x'_{new})\|_1] + \\ &\mathbb{E}[\|f_{id} - E_{id}(x''_{ori})\|_1]. \end{aligned} \tag{5.2}$$

The discriminator D attempts to distinguish between real and generated images with the adversarial loss

$$\begin{aligned} \mathcal{L}_{adv} = &\mathbb{E}[\log D(x) + \log(1 - D(x'_{ori}))] + \\ &\mathbb{E}[\log D(x) + \log(1 - D(x'_{new}))] + \\ &\mathbb{E}[\log D(x) + \log(1 - D(x''_{ori}))]. \end{aligned} \tag{5.3}$$

Consequently, the overall GAN loss combines the above named losses with weighting coefficients λ_{img} and λ_{feat}

$$\mathcal{L}_{gan} = \lambda_{img}\mathcal{L}_{img} + \lambda_{feat}\mathcal{L}_{feat} + \mathcal{L}_{adv}. \tag{5.4}$$

5.3.2 View Contrast (Contrastive Module)

The previous reconstruction and adversarial losses work in an unconditional manner. They only explore identity features within the original view-point, which renders appearance representations view-variant. In rotating an original mesh to a different view-point, *e.g.*, from front to side view-point, the generation is prone to fail due to lack of information pertained to the side view. This issue can be alleviated by enhancing the view-invariance of representations.

Given an anchor image x, the first step is to find positive images that belong to the same identity and negative images that belong to different identities. Here, we store all instance representations in a memory bank [143], which stabilizes pseudo labels and enlarges the number of negatives during the training with mini-batches. The memory bank $\mathcal{M}$ is updated with a momentum coefficient α.

$$\mathcal{M}[i]^t = \alpha \cdot \mathcal{M}[i]^{t-1} + (1-\alpha) \cdot f^t \tag{5.5}$$

where $\mathcal{M}[i]^t$ and $\mathcal{M}[i]^{t-1}$ respectively refer to the identity feature vector in the t and $t-1$ epochs.

We use a clustering algorithm DBSCAN [31] on all memory bank feature vectors to generate pseudo identity labels $\mathcal{Y} = \{y_1, y_2, ..., y_J\}$, which are renewed at the beginning of every epoch. Given the obtained pseudo labels, we have N_{pos} positive and N_{neg} negative instances for each training instance. N_{pos} and N_{neg} vary for different instances. For simplicity in a mini-batch training, we fix common positive and negative numbers for every training instance. Given an image x, we randomly sample K instances that have different pseudo identities and one instance representation f_{pos} that has the same pseudo identity with x from the memory bank. Note that f_{pos} is from a random positive image that usually has a pose and camera style different from x and x'_{new}. x and x'_{new} are encoded by E_{id} into identity feature vectors f and f'_{new}. Next, f, f'_{new} and f_{pos} are used in turn to form three positive pairs. The f'_{new} and K different identity instances in the memory bank are used as K negative pairs. Towards learning robust view-invariant representations, we extend the InfoNCE loss [96] into a view-invariant loss between original and generated views. We use $sim(u,v) = \frac{u}{\|u\|_2} \cdot \frac{v}{\|v\|_2}$ to denote the cosine similarity. We define the view-invariant loss as a softmax log loss of $K+1$ pairs as following.

$$\mathcal{L}_{vi} = \mathbb{E}[\log{(1 + \frac{\sum_{i=1}^{K} \exp{(sim(f'_{new}, k_i)/\tau)}}{\exp{(sim(f, f_{pos})/\tau)}})}] \tag{5.6}$$

$$\mathcal{L}'_{vi} = \mathbb{E}[\log{(1 + \frac{\sum_{i=1}^{K} \exp{(sim(f'_{new}, k_i)/\tau)}}{\exp{(sim(f'_{new}, f)/\tau)}})}] \tag{5.7}$$

$$\mathcal{L}''_{vi} = \mathbb{E}[\log{(1 + \frac{\sum_{i=1}^{K} \exp{(sim(f'_{new}, k_i)/\tau)}}{\exp{(sim(f'_{new}, f_{pos})/\tau)}})}], \tag{5.8}$$

where τ indicates a temperature coefficient that controls the scale of calculated similarities. $\mathcal{L}_{vi}$ maximizes the invariance between original and memory positive views. $\mathcal{L}'_{vi}$ maximizes the invariance between synthesized and original views. $\mathcal{L}''_{vi}$ maximizes the invariance between synthesized and memory positive

views. Meanwhile, the synthesized view is pushed away from K negative views in the latent space. Replacing $sim(f'_{new}, k_i)$ in Eq. 5.6, Eq. 5.7 and Eq. 5.8 with $sim(f, k_i)$ is another possibility, which pushes away the original view from negative instances. After testing, $sim(f'_{new}, k_i)$ works better, because pushing away the synthesized view from negative instances aid the generation of more accurate synthesized views that look different from the K negative instances.

5.3.3 Joint Training

Our proposed GCL framework is trained in a joint training way. Both GAN and contrastive instance discrimination can be trained in a self-supervised manner. While the GAN learns a data distribution via adversarial learning on each instance, contrastive instance discrimination learns representations by retrieving each instance from candidates. In our designed joint training, the two modules work as two collaborators with the same objective: enhancing the quality of representations built by the shared identity encoder E_{id}. We formulate our GCL as an approach to augment contrast for unsupervised ReID. Firstly, the generative module generates online data augmentation, which enhances the positive view diversity for contrastive module. Secondly, the contrastive module, in turn, learns view-invariant representations by matching original and generated views, which refine the generation quality. The joint training boosts both modules simultaneously. Our joint training conducts forward propagation initially on the generative module and subsequently on the contrastive module. Back-propagation is then conducted with an overall loss that combines Eq. 5.4, Eq. 5.6, Eq. 5.7 and Eq. 5.8.

$$\mathcal{L}_{all} = \mathcal{L}_{gan} + \mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi} \tag{5.9}$$

To accelerate the training process and alleviate the noise from imperfect generation quality at beginning epochs, we need to warm up the four modules used in the View Generator E_{id}, E_{str}, G and D. We firstly use a state-of-the-art unsupervised ReID method to warm up E_{id}, which is then considered as a baseline in our ablation studies. Generally speaking, any unsupervised ReID method can be used to warm up E_{id}. Before conducting the View Contrast, we freeze E_{id} and warm up E_{str}, G, and D only with GAN loss in Eq. 5.4 for 40 epochs. In the following, we bring in the memory bank and the pseudo labels to jointly train the whole framework with $\mathcal{L}_{all}$ for another 20 epochs. During the joint training, pseudo labels are updated at the beginning of every epoch.

5.4 Experiments

5.4.1 Datasets and Evaluation Protocols

Three mainstream person ReID datasets are considered in our experiments, including Market-1501 [164], DukeMTMC-reID [105] and MSMT17 [138]. Market-1501 is composed of 12,936 images of 751 identities for training and 19,732 images of 750 identities for test captured from 6 cameras.

Method	Reference	Market1501					DukeMTMC-reID				
		Source	mAP	Rank1	Rank5	Rank10	Source	mAP	Rank1	Rank5	Rank10
BUC [84]	AAAI'19	None	29.6	61.9	73.5	78.2	None	22.1	40.4	52.5	58.2
SoftSim [85]	CVPR'20	None	37.8	71.7	83.8	87.4	None	28.6	52.5	63.5	68.9
TSSL [140]	AAAI'20	None	43.3	71.2	-	-	None	38.5	62.2	-	-
MMCL [125]	CVPR'20	None	45.5	80.3	89.4	92.3	None	40.2	65.2	75.9	80.0
JVTC [75]	ECCV'20	None	41.8	72.9	84.2	88.7	None	42.2	67.6	78.0	81.6
JVTC+ [75]	ECCV'20	None	47.5	79.5	89.2	91.9	None	50.7	74.6	82.9	85.3
MMCL*	This chapter	None	45.1	79.5	89.0	91.9	None	40.9	64.8	75.2	79.8
JVTC*	This chapter	None	47.2	75.4	86.7	90.5	None	43.9	66.8	77.6	81.0
JVTC+*	This chapter	None	50.9	79.1	89.8	92.9	None	52.8	74.9	83.3	85.8
ours(MMCL*)	This chapter	None	54.9	83.7	91.6	94.0	None	49.3	69.7	79.7	82.8
ours(JVTC*)	This chapter	None	63.4	83.7	91.6	94.3	None	53.3	72.4	82.0	84.9
ours(JVTC+*)	This chapter	None	**66.8**	**87.3**	**93.5**	**95.5**	None	**62.8**	**82.9**	**87.1**	**88.5**
ECN [173]	CVPR'19	Duke	43.0	75.1	87.6	91.6	Market	40.4	63.3	75.8	80.4
PDA [73]	ICCV'19	Duke	47.6	75.2	86.3	90.2	Market	45.1	63.2	77.0	82.5
CR-GAN [22]	ICCV'19	Duke	54.0	77.7	89.7	92.7	Market	48.6	68.9	80.2	84.7
SSG [37]	ICCV'19	Duke	58.3	80.0	90.0	92.4	Market	53.4	73.0	80.6	83.2
MMCL [125]	CVPR'20	Duke	60.4	84.4	92.8	95.0	Market	51.4	72.4	82.9	85.0
ACT [147]	AAAI'20	Duke	60.6	80.5	-	-	Market	54.5	72.4	-	-
DG-Net++ [179]	ECCV'20	Duke	61.7	82.1	90.2	92.7	Market	63.8	78.9	87.8	90.4
JVTC [75]	ECCV'20	Duke	61.1	83.8	93.0	95.2	Market	56.2	75.0	85.1	88.2
ECN+ [174]	PAMI'20	Duke	63.8	84.1	92.8	95.4	Market	54.4	74.0	83.7	87.4
JVTC+ [75]	ECCV'20	Duke	67.2	86.8	95.2	97.1	Market	66.5	80.4	**89.9**	92.2
MMT [39]	ICLR'20	Duke	71.2	87.7	94.9	96.9	Market	65.1	78.0	88.8	**92.5**
CAIL [89]	ECCV'20	Duke	71.5	88.1	94.4	96.2	Market	65.2	79.5	88.3	91.4
ACT*	This chapter	Duke	59.1	78.8	88.9	91.7	Market	51.5	70.9	80.0	83.4
JVTC*	This chapter	Duke	65.0	85.7	93.6	95.6	Market	56.5	73.9	84.5	87.7
JVTC+*	This chapter	Duke	67.6	87.0	95.2	97.0	Market	66.7	81.0	**89.9**	91.5
ours(ACT*)	This chapter	Duke	66.7	83.9	91.4	93.4	Market	55.4	71.9	81.6	84.6
ours(JVTC*)	This chapter	Duke	73.4	89.1	95.0	96.6	Market	60.4	77.2	86.2	88.4
ours(JVTC+*)	This chapter	Duke	**75.4**	**90.5**	**96.2**	**97.1**	Market	**67.6**	**81.9**	88.9	90.6

Table 5.1: Comparison of unsupervised ReID methods (%) with a ResNet50 backbone on Market and Duke datasets. We test our proposed method on several baselines, whose names are in brackets. * refers to our implementation based on authors' code.

DukeMTMC-reID contains 16,522 images of 702 persons for training, 2,228 query images and 17,661 gallery images of 702 persons for test from 8 cameras. MSMT17 is a larger dataset, which contains 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras.

Following state-of-the-art unsupervised ReID methods [125, 75], we evaluate our proposed method GCL under fully unsupervised setting on the three datasets and under four UDA benchmark protocols, including Market→Duke, Duke→Market, Market→MSMT and Duke→MSMT. We report both quantitative and qualitative results for unsupervised person ReID and view generation.

5.4.2 Implementation Details

We firstly present network design details of E_{id}, E_{str}, G and D. In the following descriptions, we write the size of feature maps in channel×height×width. Our model design is mainly inspired by [165, 179]. (1) E_{id} is a ImageNet [106] pre-trained ResNet50 [49] with slight modifications. The original fully connected layer is replaced by a fully connected embedding layer, which outputs identity representations f in 512×1×1 for the View Contrast. In parallel, we add a part average pooling that outputs iden-

Method	Reference	MSMT17				
		Source	mAP	R1	R5	R10
MMCL [125]	CVPR'20	None	11.2	35.4	44.8	49.8
JVTC [75]	ECCV'20	None	15.1	39.0	50.9	56.8
JVTC+ [75]	ECCV'20	None	17.3	43.1	53.8	59.4
JVTC*	This chapter	None	13.4	36.0	48.8	54.9
JVTC+*	This chapter	None	16.3	40.4	55.6	61.8
ours(JVTC*)	This chapter	None	18.0	41.6	53.2	58.4
ours(JVTC+*)	This chapter	None	**21.3**	**45.7**	**58.6**	**64.5**
ECN [173]	CVPR'19	Market	8.5	25.3	36.3	42.1
SSG [37]	ICCV'19	Market	13.2	31.6	49.6	-
MMCL [125]	CVPR'20	Market	15.1	40.8	51.8	56.7
ECN+ [174]	PAMI'20	Market	15.2	40.4	53.1	58.7
JVTC [75]	ECCV'20	Market	19.0	42.1	53.4	58.9
DG-Net++ [179]	ECCV'20	Market	22.1	48.4	60.9	66.1
CAIL [89]	ECCV'20	Market	20.4	43.7	56.1	61.9
MMT [39]	ICLR'20	Market	22.9	49.2	63.1	68.8
JVTC+ [75]	ECCV'20	Market	25.1	48.6	65.3	68.2
JVTC*	This chapter	Market	17.1	39.6	53.3	59.3
JVTC+*	This chapter	Market	20.5	44.0	59.5	71.1
ours(JVTC*)	This chapter	Market	21.5	45.0	57.1	66.5
ours(JVTC+*)	This chapter	Market	**27.0**	**51.1**	**63.9**	**69.9**
ECN [173]	CVPR'19	Duke	10.2	30.2	41.5	46.8
SSG [37]	ICCV'19	Duke	13.3	32.2	51.2	-
MMCL [125]	CVPR'20	Duke	16.2	43.6	54.3	58.9
ECN+ [174]	PAMI'20	Duke	16.0	42.5	55.9	61.5
JVTC [75]	ECCV'20	Duke	20.3	45.4	58.4	64.3
DG-Net++ [179]	ECCV'20	Duke	22.1	48.8	60.9	65.9
MMT [39]	ICLR'20	Duke	23.3	50.1	63.9	69.8
CAIL [89]	ECCV'20	Duke	24.3	51.7	64.0	68.9
JVTC+ [75]	ECCV'20	Duke	27.5	52.9	70.5	75.9
JVTC*	This chapter	Duke	19.9	45.4	59.1	64.9
JVTC+*	This chapter	Duke	23.6	49.4	65.2	71.1
ours(JVTC*)	This chapter	Duke	24.9	50.8	63.4	68.9
ours(JVTC+*)	This chapter	Duke	**29.7**	**54.4**	**68.2**	**74.2**

Table 5.2: Comparison of unsupervised Re-ID methods (%) with a ResNet50 backbone on MSMT17. * refers to our implementation based on authors' code.

tity features f_{id} in 2048×4×1 for the View Generator. (2) E_{str} is composed of four convolutional and four residual layers, which output structure features f_{str} in 128×64×32. (3) G contains four residual and four convolutional layers. Every residual layer contains two adaptive instance normalization layers [58] that transform f_{id} into scale and bias parameters. (4) D is a multi-scale PatchGAN [61] discriminator at 64×32, 128×64 and 256×128.

Then, we present the training and testing configuration details. Our framework is implemented in Pytorch and trained with one Nvidia Titan RTX GPU. (1) For the E_{id} warm-up, we consider JVTC [75], because it is a state-of-the-art ReID method that is compatible with both fully unsupervised and UDA settings. We also test other baselines, *e.g.*, MMCL [125] and ACT [147] to demonstrate the generalizability of our method. (2) For training, inputs are resized to 256×128. We empirically set a large weight $\lambda_{img} = \lambda_{feat} = 5$ for reconstruction in Eq. 5.4. With a batch size of 16, we use SGD to train E_{id} and Adam optimizer to train E_{str}, G and D. Learning rate is set to 1×10^{-4} during the warm-up. In the joint-training, learning rate in Adam is set to 1×10^{-4} and 3.5×10^{-4} in SGD and are multiplied by 0.1 after 10 epochs. (3) In the View Contrast module, we set the momentum coefficient $\alpha = 0.2$ in Eq. 5.5 and the temperature $\tau = 0.04$ in Eq. 5.6. The number of negatives K is 8192. DBSCAN density radius is set to 2×10^{-3}. (4) For testing, only E_{id} is conserved and outputs representations f of dimension 512.

Important parameters are set by a grid search on the fully unsupervised Market-1501 benchmark. The temperature τ is searched from $\{0.03, 0.04, 0.05, 0.06, 0.07\}$ and finally is set to 0.04. A smaller τ increases the scale of similarity scores in the Eq. 5.6, Eq. 5.7 and Eq. 5.8, which makes view-invariant losses more sensitive to inter-instance difference. However, when τ is set to 0.03, these losses become too sensitive and make the training unstable. The number of negatives K is searched from $\{2048, 4096, 8192\}$. A larger K pushes away more negatives in the view-invariant losses. Since the Market-1501 dataset has only 12936 training images, we set $K = 8192$.

5.4.3 Unsupervised ReID Evaluation

Comparison with state-of-the-art methods. Tab. 5.1 shows the quantitative results on the Market-1501 and DukeMTMC-reID datasets. Tab. 5.2 shows the quantitative results on the MSMT17 dataset. Our method is mainly designed for fully unsupervised ReID. Under this setting, we test the performance of GCL with three different baselines, including MMCL, JVTC and JVTC+. Our implementation of the three baselines provides results that are slightly different from those mentioned in the corresponding papers. Thus, we firstly report results of our implementations and then add our GCL on these baselines. Our method improves the performance of the baselines by large margins. These improvements show that GANs are not limited to cross-domain style transfer for unsupervised ReID.

Under the UDA setting, we also evaluate the performance of GCL with three different baselines, including ACT, JVTC and JVTC+. The labeled source

dataset is only used to warm up our identity encoder E_{id}, but not used in our joint generative and contrastive training. Compared to fully unsupervised methods, the UDA warmed E_{id} is stronger and extracts improved identity features. Thus, the performance of UDA methods is generally higher than fully unsupervised methods. With a strong baseline JVTC+, our GCL achieves state-of-the-art performance.

Ablation Study. To better understand the contribution of generative and contrastive modules, we conduct ablation experiments on the two fully unsupervised benchmarks: Market-1501 and DukeMTMC-reID. Quantitative results with a JVTC baseline are reported in Tab. 5.3. By gradually adding loss functions on the baseline, our ablation experiments correspond to three scenarios. (1) Only Generation: with only $\mathcal{L}_{gan}$, our generation module disentangles identity and structure features. Since there is no inter-view constraint, E_{id} tends to extract view-specific identity features, which decreases the ReID performance. (2) Only Contrast: we use $\mathcal{L}_{vi}^{woGAN} = \mathbb{E}[\log\left(1 + \frac{\sum_{i=1}^{K} \exp\left(sim(f,k_i)/\tau\right)}{\exp\left(sim(f,f_{pos})/\tau\right)}\right)]$ to train our contrastive module without generation. We also add a set of traditional data augmentation, including random flipping, cropping, jittering, erasing, to train our contrastive module like a traditional memory bank based contrastive method. (3) Joint Generation and Contrast: $\mathcal{L}_{vi}$, $\mathcal{L}_{vi}'$ and $\mathcal{L}_{vi}''$ enhance the view-invariance of identity representations between original, synthesized and memory-stored positive views, while negative views are pushed away.

We minimize intra-class variance via contrasting generated images, which leads to a larger inter-class distance in latent space. Learning view-invariant representations from diversified generated data helps clustering algorithms to generate more accurate pseudo labels. With a same DBSCAN clustering, the cluster number of GCL is closer to real identity number than that of contrastive learning with traditional data augmentation. For example, Market-1501 dataset has 751 real identities. DBSCAN in GCL categorizes unlabeled images into around 520 clusters, while the contrastive learning with traditional data augmentation has around 460 clusters (see Fig. 5.4).

We also conduct a qualitative ablation study, where synthesized novel views without and with view-invariant losses are illustrated in Fig. 5.6. Results confirm that E_{id} extracts view-specific identity features (black bag), in the case that view-invariant losses are not used. Given view-invariant losses, E_{id} is able to extract view-invariant identity features (red shirt). Another example is provided in Figure 5.5.

5.4.4 Generation Quality Evaluation

Comparison with state-of-the-art methods. We compare generated images between our proposed GCL under the JVTC [75] warmed fully unsupervised setting and state-of-the-art GAN-based ReID methods in Fig. 5.7. FD-GAN [40], IS-GAN [30] and DG-Net [165] are supervised Re-ID methods.

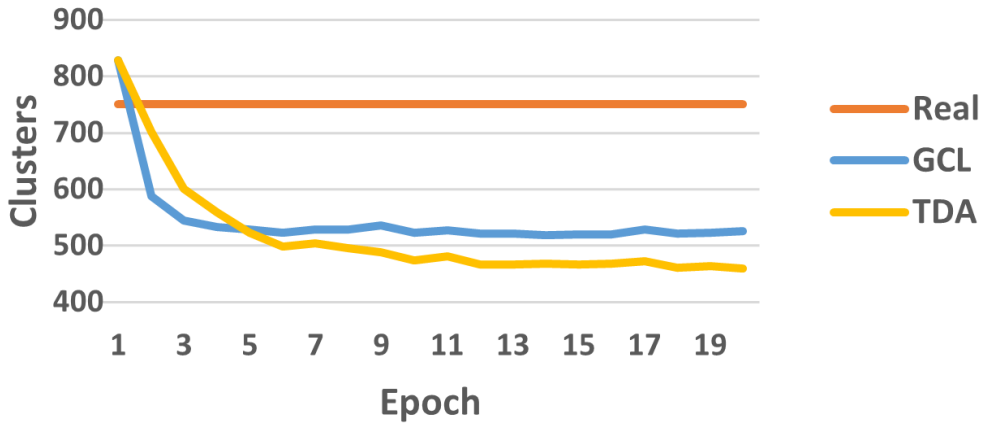


Figure 5.4: Cluster number curve on Market-1501. TDA denotes traditional data augmentation, including random flipping, cropping, jittering, erasing.

Loss	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
Baseline	47.2	75.4	43.9	66.8
$+\mathcal{L}_{gan}$	41.6	69.0	25.8	45.9
$+\mathcal{L}_{vi}^{woGAN}$	47.8	75.2	44.1	67.8
$+\mathcal{L}_{vi}^{woGAN} + TDA$	53.7	78.7	48.5	70.0
$+\mathcal{L}_{gan} + \mathcal{L}_{vi}$	54.1	79.4	47.4	68.4
$+\mathcal{L}_{gan} + \mathcal{L}_{vi} + \mathcal{L}'_{vi}$	59.2	82.2	50.5	71.0
$+\mathcal{L}_{gan} + \mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}$	**63.4**	**83.7**	**53.3**	**72.4**

Table 5.3: Ablation study on loss functions used in two modules. (1). $\mathcal{L}_{gan}$ corresponds to generation w/o contrast. (2). $\mathcal{L}_{vi}^{woGAN}$ corresponds to contrast w/o generation. TDA denotes traditional data augmentation. (3). $\mathcal{L}_{gan} + \mathcal{L}_{vi}$ ($\mathcal{L}'_{vi}$ and $\mathcal{L}''_{vi}$) correspond to joint generative and contrastive learning.

Method	FID(realism)	SSIM(diversity)
Real	**7.22**	0.350
FD-GAN [40]	216.88	0.271
IS-GAN [30]	281.63	0.165
DG-Net [165]	18.24	0.360
Ours(U)	59.86	0.367
Ours(UDA)	53.07	**0.369**

Table 5.4: Comparison of FID (lower is better) and SSIM (higher is better) on Market-1501 dataset. U denotes the fully unsupervised setting. UDA denotes Duke→Market setting.

Figure 5.5: More qualitative ablation study on the view-invariant losses. For simplicity, $\mathcal{L}_{vi}$ denotes three view-invariant losses $\mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}$, which helps E_{id} to extract better identity features (white shirt).

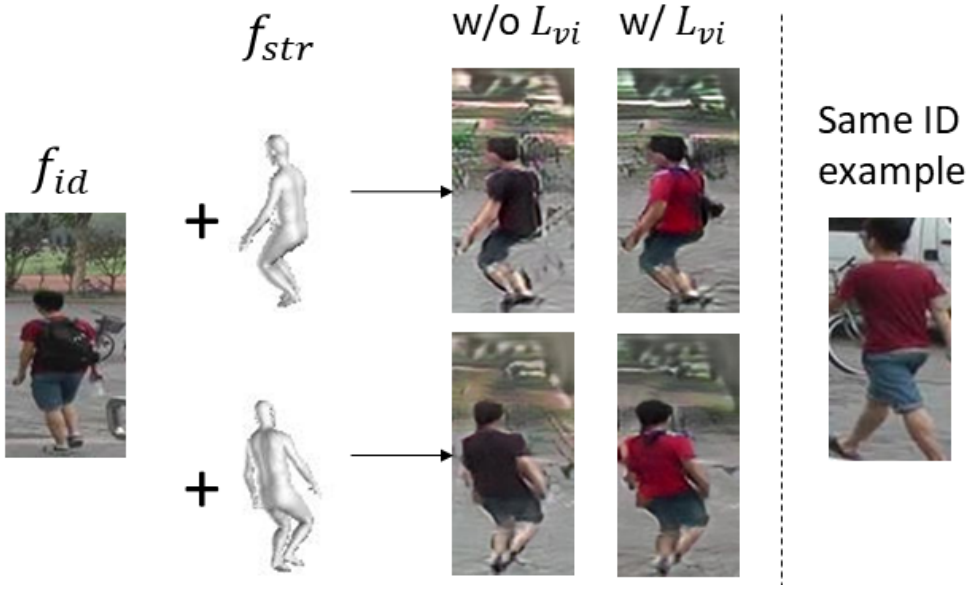


Figure 5.6: Qualitative ablation study on the view-invariant losses. For simplicity, $\mathcal{L}_{vi}$ denotes three view-invariant losses $\mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}$, which helps E_{id} to extract view-invariant features (red shirt).

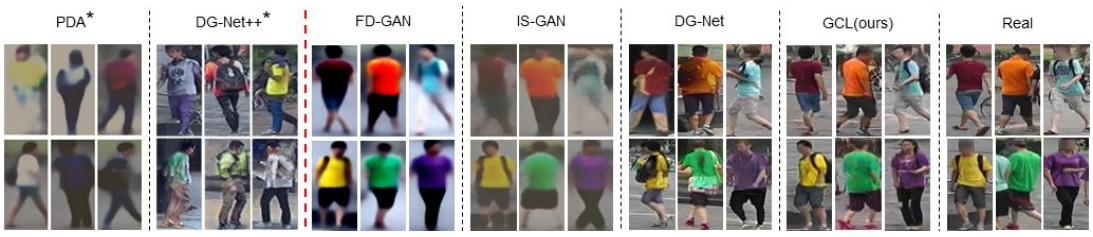


Figure 5.7: Comparison of the generated images on Market-1501 dataset. $\star$ refers to methods without sharing source code, whose examples are cropped from their papers. Examples of FD-GAN, IS-GAN, DG-Net and GCL are generated from six real images shown in the figure.

Figure 5.8: Generated novel views on the three datasets.

Since the source code of these three methods is available, we compare generated images of same identities. We observe that there exists blur in images generated by FD-GAN and IS-GAN. DG-Net generates sharper images, but different body shapes and some incoherent objects (bags and clothes) are observed. PDA [73] and DG-Net++ [179] are UDA methods, whose source code is not yet released. We can only compare several generated images with unknown identities as illustrated in their papers. PDA generates blurred cross-domain images, whose quality is similar to FD-GAN and IS-GAN. DG-Net++ extends DG-Net into cross-domain generation, which has same problems of body shape and incoherent objects. Our GCL preserves better body shape information and does not generate incoherent objects. Moreover, our GCL is a fully unsupervised method.

We use Fréchet Inception Distance (FID) [52] to measure visual quality, as well as Structural SIMilarity (SSIM) [136] to capture structure diversity of generated images. In Tab. 5.4, we compare our method with FD-GAN [40], IS-GAN [30] and DG-Net [165], whose source code is available. FID measures the distribution distance between generated and real images, where a lower FID represents the case, where generated images are similar to real ones. SSIM measures the intra-class structural similarity, where a larger SSIM represents a larger diversity. We note that DG-Net is outperforms our method w.r.t. FID, because the distribution is better maintained with ground truth identities in the supervised method DG-Net. However, our method is superior to DG-Net w.r.t. SSIM, as DG-Net swaps intra-dataset structures, whereas our rotated meshes build structures that do not exist in the original dataset.

Figure 5.9: Linear interpolation on identity features. Identity features are swapped between left and right persons.

More discussion. To validate, whether identity and structure features can be really disentangled under a fully unsupervised ReID setting, two experiments are conducted by changing firstly only structure features and then only identity features. Results in Fig. 5.8 show that changing structure features only change structures and do not affect appearances. We also fix structure features and linearly interpolate two random identity feature vectors. Results in Fig. 5.9 show that identity features only change appearances and do not affect structures in generated images. More examples are provided in Figure 5.10, 5.11 and 5.12.

5.5 Conclusions

In this chapter, we propose a joint generative and contrastive learning framework for unsupervised person ReID. Deviating from previous contrastive methods with traditional data augmentation techniques, we generate diversified views with a 3D mesh guided GAN. These generated novel views are then combined with original images in memory based contrastive learning, in order to learn view-invariant representations, which in turn improve generation quality. Our generative and contrastive modules mutually promote each other's performance in unsupervised ReID. Moreover, our framework does not rely on a source dataset, which is mandatory in style transfer based methods. Extensive experiments on three datasets validate the effectiveness of our framework in both unsupervised person ReID and multi-view person image generation. This chapter mainly focuses on how to create more suitable positive image pairs for contrastive unsupervised person ReID. In the next chapter, we regularize inter-instance affinities with different data augmentation techniques to further enhance the performance of contrastive unsupervised person ReID.

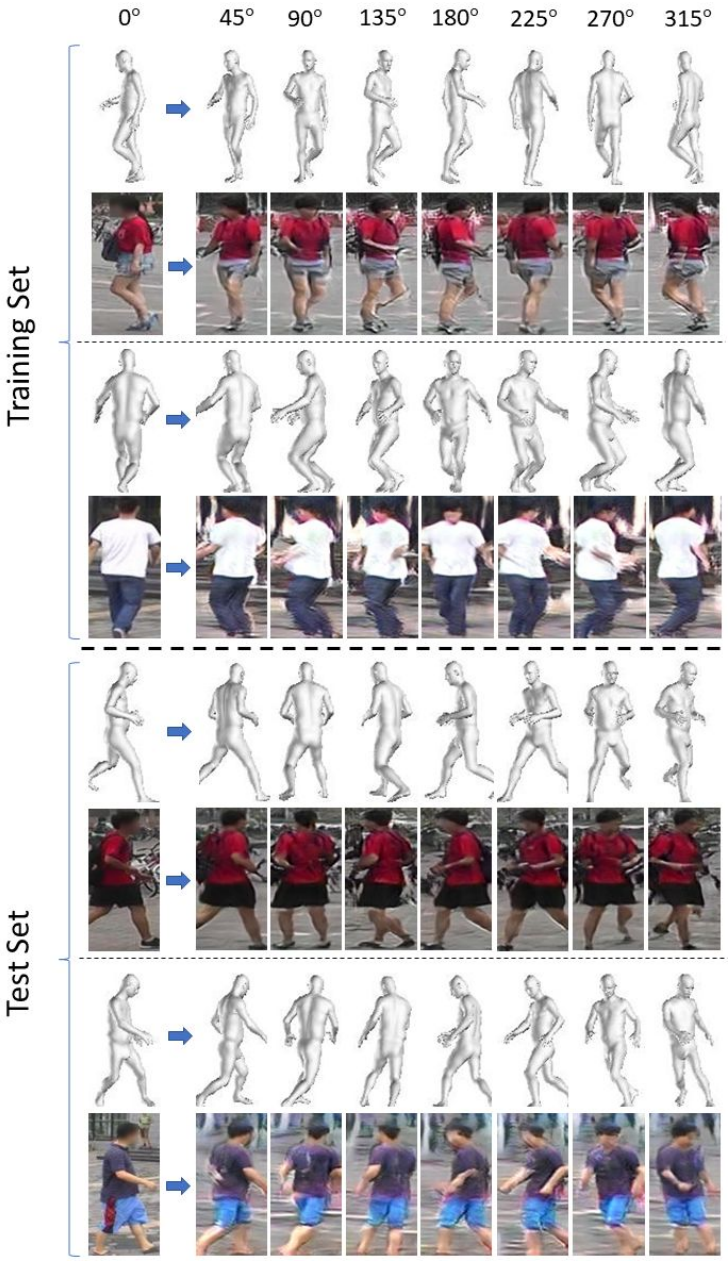


Figure 5.10: Examples of generated novel views on **Market-1501** training and test sets.

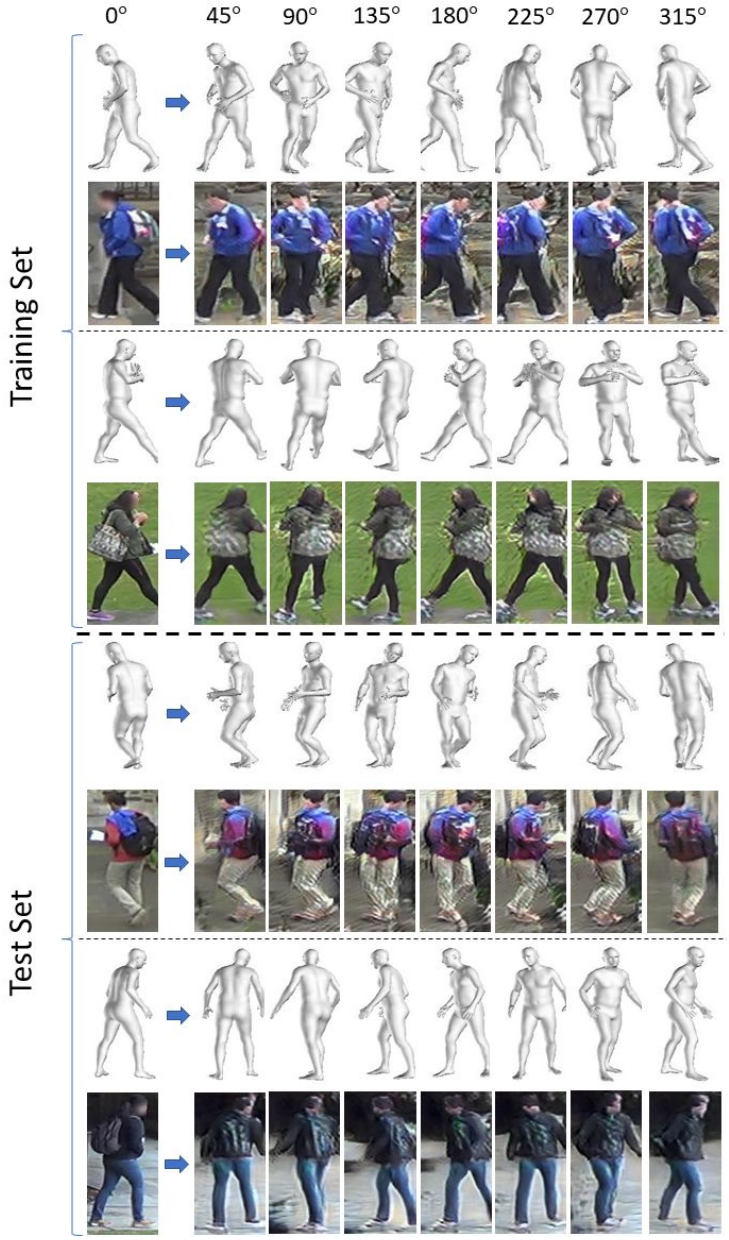


Figure 5.11: Examples of generated novel views on **DukeMTMC-reID** training and test sets.

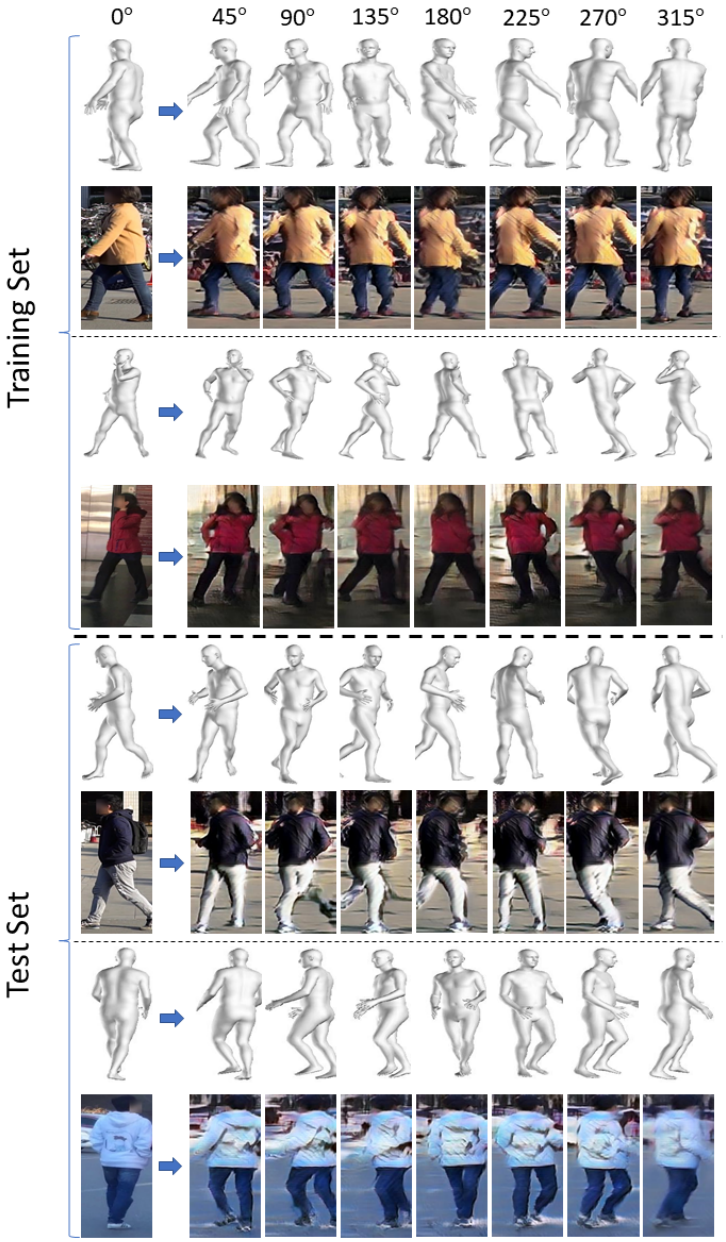


Figure 5.12: Examples of generated novel views on **MSMT17** training and test sets.

Chapter 6

Inter-instance Contrastive Encoding for unsupervised person ReID

6.1 Introduction

In this chapter, instead of direct data augmentation techniques to enhance data diversity, we propose a soft label regularization to enhance the inter-instance similarity consistency before and after data augmentation. A pre-requisite for the proposed soft label is low intra-class variance, which can be achieved by a hard instance contrastive loss. Our proposed hard instance contrastive loss and soft instance consistency loss permit our fully unsupervised ReID models to get competitive performance with recent supervised counterparts. The described work in this chapter has been published in IEEE/CVF International Conference on Computer Vision (ICCV) 2021 [14].

Similar to the last chapter, we continue the research on contrastive learning applied to unsupervised person ReID. State-of-the-art contrastive methods [143, 17, 48] consider each image instance as a class and learns representations by matching augmented views of a same instance. As a class is usually composed of multiple positive instances, it hurts the performance of fine-grained ReID tasks when different images of a same identity are considered as different classes. Self-paced Contrastive Learning (SpCL) [42] alleviates this problem by matching an instance with the centroid of the multiple positives, where each positive converges to its centroid at a uniform pace. Although SpCL has achieved impressive performance, this method does not consider inter-instance affinities, which can be leveraged to reduce intra-class variance and make clusters more compact. In supervised ReID, state-of-the-art methods [12, 90] usually adopt a hard triplet loss [51] to lay more emphasis on hard samples inside a class, so that hard samples can get closer to normal samples. In this chapter, we introduce Inter-instance Contrastive Encoding (ICE), in which we match an instance with its hardest positive in a mini-batch to make clusters more compact and improve pseudo label quality. Matching the hardest positive refers to using one-hot "hard" pseudo labels.

Since no ground truth is available, mining hardest positives within clusters is likely to introduce false positives into the training process. In addition, the one-hot label does not take the complex inter-instance relationship into consideration when multiple pseudo positives and negatives exist in a mini-batch. Contrastive methods usually use data augmentation to mimic real-world distortions, *e.g.*, occlusion, view-point and resolution variance. After data augmentation operations, certain pseudo positives may become less similar to an anchor, while certain pseudo negatives may become more similar. As a robust model should be invariant to distortions from data augmentation, we propose to use the inter-instance pairwise similarity as "soft" pseudo labels to enhance the consistency before and after augmentation.

Our proposed ICE incorporates class-level label (centroid contrast), instance pairwise hard label (hardest positive contrast) and instance pairwise soft label (augmentation consistency) into one fully unsupervised person ReID framework. Without any identity annotation, ICE significantly outperforms state-of-the-art UDA and fully unsupervised methods on main-stream person ReID datasets.

To summarize, our contributions are: (1) We propose to use pairwise similarity ranking to mine hardest samples as one-hot hard pseudo labels for hard instance contrast, which reduces intra-class variance. (2) We propose to use pairwise similarity scores as soft pseudo labels to enhance the consistency between augmented and original instances, which alleviates label noise and makes our model more robust to augmentation perturbation. (3) Extensive experiments highlight the importance of inter-instance pairwise similarity in contrastive learning. Our proposed method ICE outperforms state-of-the-art methods by a considerable margin, significantly pushing unsupervised ReID to real-world deployment.

6.2 Related Work

Unsupervised person ReID. Recent unsupervised person ReID methods can be roughly categorized into unsupervised domain adaptation (UDA) and fully unsupervised methods. Among UDA-based methods, several works [129, 83] leverage semantic attributes to reduce the domain gap between source and target domains. Several works [138, 172, 22, 173, 179, 15] use generative networks to transfer labeled source domain images into the style of target domain. Another possibility is to assign pseudo labels to unlabeled images, where pseudo labels are obtained from clustering [112, 37, 159, 13] or reference data [153]. Pseudo label noise can be reduced by selecting credible samples [10] or using a teacher network to assign soft labels [39]. All these UDA-based methods require a labeled source dataset. Fully unsupervised methods have a better flexibility for deployment. BUC [84] first treats each image as a cluster and progressively merge clusters. Lin *et al.* [85] replace clustering-based pseudo labels with similarity-based softened labels. Hierarchical Clustering is proposed in [155] to improve the quality of pseudo labels. Since each identity usually has multiple positive instances, MMCL [125] introduces a memory-based multi-label classi-

fication loss into unsupervised ReID. JVTC [75] and CycAs [135] explore temporal information to refine visual similarity. SpCL [42] considers each cluster and outlier as a single class and then conduct instance-to-centroid contrastive learning. CAP [130] calculates identity centroids for each camera and conducts intra- and inter-camera centroid contrastive learning. Both SpCL and CAP focus on instance-to-centroid contrast, but neglect inter-instance affinities.

Contrastive Learning. Recent contrastive learning methods [143, 48, 17] consider unsupervised representation learning as a dictionary look-up problem. Wu *et al.* [143] retrieve a target representation from a memory bank that stores representations of all the images in a dataset. MoCo [48] introduces a momentum encoder and a queue-like memory bank to dynamically update negatives for contrastive learning. In SimCLR [17], authors directly retrieve representations within a large batch. However, all these methods consider different instances of a same class as different classes, which is not suitable in a fine-grained ReID task. These methods learn invariance from augmented views, which can be regarded as a form of consistency regularization.

Consistency regularization. Consistency regularization refers to an assumption that model predictions should be consistent when fed perturbed versions of the same image, which is widely considered in recent semi-supervised learning [120, 110, 18]. The perturbation can come from data augmentation [110], temporal ensembling [120, 70, 41] and shallow-deep features [166, 18]. Artificial perturbations are applied in contrastive learning as strong augmentation [20, 137] and momentum encoder [48] to make a model robust to data variance. Based on temporal ensembling, Ge *et al.* [41] use inter-instance similarity to mitigate pseudo label noise between different training epochs for image localization. Wei *et al.* [137] propose to regularize inter-instance consistency between two sets of augmented views, which neglects intra-class variance problem. We simultaneously reduce intra-class variance and regularize consistency between augmented and original views, which is more suitable for fine-grained ReID tasks.

6.3 Proposed Method

6.3.1 Overview

Given a person ReID dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, our objective is to train a robust model on $\mathcal{X}$ without annotation. For inference, representations of a same person are supposed to be as close as possible. State-of-the-art contrastive methods [48, 17] consider each image as an individual class and maximize similarities between augmented views of a same instance with InfoNCE loss [96]:

$$\mathcal{L}_{InfoNCE} = \mathbb{E}[-\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}] \tag{6.1}$$

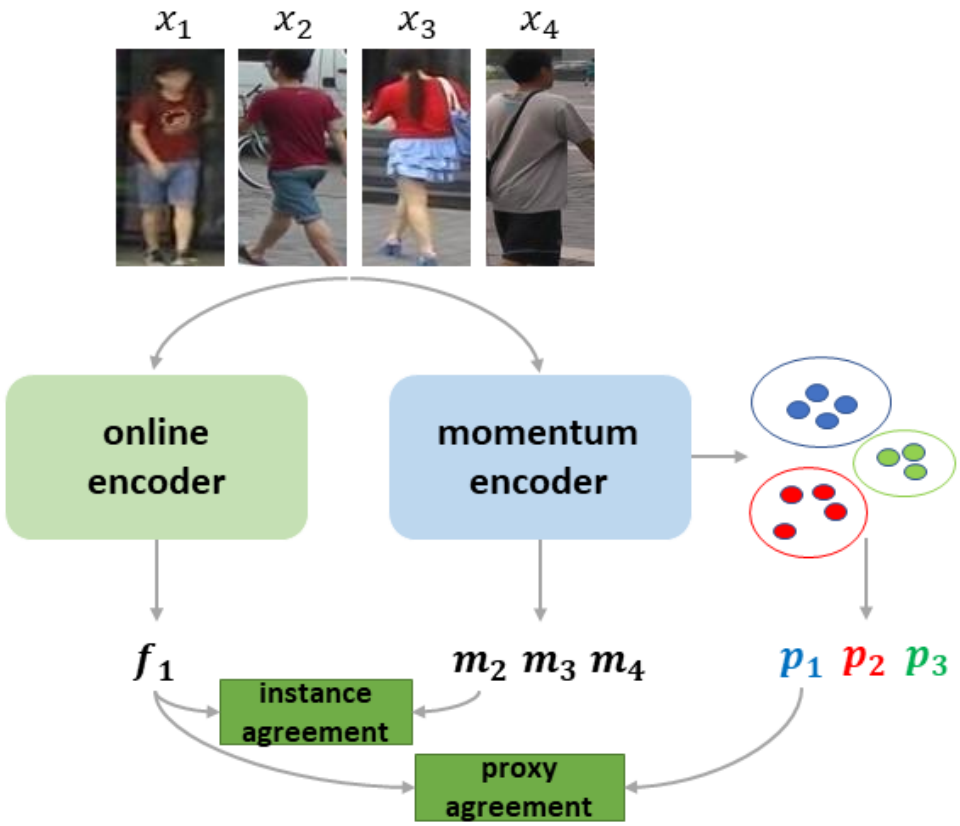


Figure 6.1: General architecture of ICE. We maximize the similarity between anchor and pseudo positives in both inter-class (proxy agreement between an instance representation f_1 and its cluster proxy p_1) and intra-class (instance agreement between f_1 and its pseudo positive m_2) manners.

where q and k_+ are two augmented views of a same instance in a set of candidates k_i. τ is a temperature hyper-parameter that controls the scale of similarities.

Following MoCo [48], we design our proposed ICE with an online encoder and a momentum encoder as shown in Fig. 6.1. The online encoder is a regular network, *e.g.*, ResNet50 [49], which is updated by back-propagation. The momentum encoder (weights noted as θ_m) has the same structure as the online encoder, but updated by accumulated weights of the online encoder (weights noted as θ_o):

$$\theta_m^t = \alpha\theta_m^{t-1} + (1-\alpha)\theta_o^t \tag{6.2}$$

where α is a momentum coefficient that controls the update speed of the momentum encoder. t and $t-1$ refer respectively to the current and last iteration. The momentum encoder builds momentum representations with the moving averaged weights, which are more stable to label noise.

At the beginning of each training epoch, we use the momentum encoder to extract appearance representations $\mathcal{M} = \{m_1, m_2, ..., m_N\}$ of all the samples in the training set $\mathcal{X}$. We use a clustering algorithm DBSCAN [31] on these appearance representations to generate pseudo identity labels $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$. We only consider clustered inliers for contrastive learning, while un-clustered outliers are discarded. We calculate proxy centroids $p_1, p_2, ...$ and store them in a memory for a proxy contrastive loss $\mathcal{L}_{proxy}$ (see Sec. 6.3.2). Note that this proxy memory can be camera-agnostic [42] or camera-aware [130].

Then, we use a random identity sampler to split the training set into mini-batches where each mini-batch contains N_P pseudo identities and each identity has N_K instances. We train the whole network by combining the $\mathcal{L}_{proxy}$ (with class-level labels), a hard instance contrastive loss $\mathcal{L}_{h\_ins}$ (with hard instance pairwise labels, see Sec. 6.3.3) and a soft instance consistency loss $\mathcal{L}_{s\_ins}$ (with soft instance pairwise labels, see Sec. 6.3.4):

$$\mathcal{L}_{total} = \mathcal{L}_{proxy} + \lambda_h\mathcal{L}_{h\_ins} + \lambda_s\mathcal{L}_{s\_ins} \tag{6.3}$$

To increase the consistency before and after data augmentation, we use different augmentation settings for prediction and target representations in the three losses (see Tab. 6.1).

Loss	Predictions (augmentation)	Targets (augmentation)
$\mathcal{L}_{proxy}$	f (Strong)	p (None)
$\mathcal{L}_{h\_ins}$	f (Strong)	m (Strong)
$\mathcal{L}_{s\_ins}$	P (Strong)	Q (None)

Table 6.1: Augmentation settings for 3 losses.

6.3.2 Proxy Centroid Contrastive Baseline

For a camera-agnostic memory, the proxy of cluster a is defined as the averaged momentum representations of all the instances belonging to this clus-

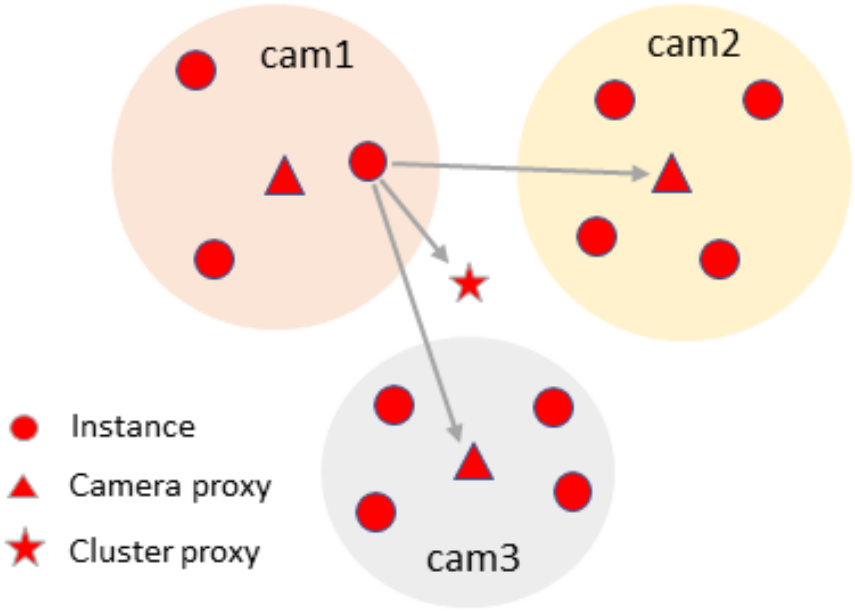


Figure 6.2: Proxy contrastive loss. Inside a cluster, an instance is pulled to a cluster centroid by $L_{agnostic}$ and to cross-camera centroids by L_{cross}.

ter:

$$p_a = \frac{1}{N_a} \sum_{m_i \in y_a} m_i \tag{6.4}$$

where N_a is the number of instances belonging to the cluster a.

We apply a set of data augmentation on $\mathcal{X}$ and feed them to the online encoder. For an online representation f_a belonging to the cluster a, the camera-agnostic proxy contrastive loss is a softmax log loss with one positive proxy p_a and all the negatives in the memory:

$$\mathcal{L}_{agnostic} = \mathbb{E}[-\log \frac{\exp\left(f_a \cdot p_a / \tau_a\right)}{\sum_{i=1}^{|p|} \exp\left(f_a \cdot p_i / \tau_a\right)}] \tag{6.5}$$

where $|p|$ is the number of clusters in a training epoch and τ_a is a temperature hyper-parameter. Different from unified contrastive loss [39], outliers are not considered as single instance clusters. In such way, outliers are not pushed away from clustered instances, which allows us to mine more hard samples for our proposed hard instance contrast. As shown in Fig. 6.2, all the clustered instances converge to a common cluster proxy centroid. However, images inside a cluster are prone to be affected by camera styles, leading to high intra-class variance. This problem can be alleviated by adding a cross-camera proxy contrastive loss [130].

For a camera-aware memory, if we have $\mathcal{C} = \{c_1, c_2, ...\}$ cameras, a camera proxy p_{ab} is defined as the averaged momentum representations of all the instances belonging to the cluster a in camera c_b:

$$p_{ab} = \frac{1}{N_{ab}} \sum_{m_i \in y_a \cap m_i \in c_b} m_i \tag{6.6}$$

where N_{ab} is the number of instances belonging to the cluster a captured by camera c_b.

Given an online representation f_{ab}, the cross-camera proxy contrastive loss is a softmax log loss with one positive cross-camera proxy p_{ai} and N_{neg} nearest negative proxies in the memory:

$$\mathcal{L}_{cross} = \mathbb{E}[-\frac{1}{|\mathcal{P}|}\sum_{i\neq b \cap i \in C}\log\frac{\exp\left(< f_{ab}\cdot p_{ai} > /\tau_c\right)}{\sum_{j=1}^{N_{neg}+1}\exp\left(< f_{ab}\cdot p_j > /\tau_c\right)}] \tag{6.7}$$

where $< \cdot >$ denotes cosine similarity and τ_c is a cross-camera temperature hyper-parameter. $|\mathcal{P}|$ is the number of cross-camera positive proxies. Thanks to this cross-camera proxy contrastive loss, instances from one camera are pulled closer to proxies of other cameras, which reduces intra-class camera style variance.

We define a proxy contrastive loss by combining cluster and camera proxies with a weighting coefficient 0.5 from [130]:

$$\mathcal{L}_{proxy} = \mathcal{L}_{agnostic} + 0.5\mathcal{L}_{cross} \tag{6.8}$$

6.3.3 Hard Instance Contrastive Loss

Although intra-class variance can be alleviated by cross-camera contrastive loss, it has two drawbacks: 1) more memory space is needed to store camera-aware proxies, 2) impossible to use when camera ids are unavailable. We propose a camera-agnostic alternative by exploring inter-instance relationship instead of using camera labels. Along with training, the encoders become strong and stronger, which helps outliers progressively enter clusters and become hard inliers. Pulling hard inliers closer to normal inliers effectively increases the compactness of clusters.

A mini-batch is composed of N_P identities, where each identity has N_K positive instances. Given an anchor instance f^i belonging to the ith class, we sample the hardest positive momentum representation m_k^i that has the lowest cosine similarity with f^i, see Fig. 6.4. For the same anchor, we have $J = (N_P - 1) \times N_K$ negative instances that do not belong to the ith class. The hard instance contrastive loss for f^i is a softmax log loss of $J + 1$ (1 positive and J negative) pairs, which is defined as:

$$\mathcal{L}_{h\_ins} = \mathbb{E}[-\log\frac{\exp\left(< f^i\cdot m_k^i > /\tau_{h\_ins}\right)}{\sum_{j=1}^{J+1}\exp\left(< f^i\cdot m_j > /\tau_{h\_ins}\right)}] \tag{6.9}$$

where $k = \arg\min_{k=1,..,N_K}(< f^i\cdot m_k^i >)$ and τ_{h_ins} is the hard instance temperature hyper-parameter. By minimizing the distance between the anchor and the hardest positive and maximizing the distance between the anchor and all negatives, $\mathcal{L}_{h\_ins}$ increases intra-class compactness and inter-class separability.

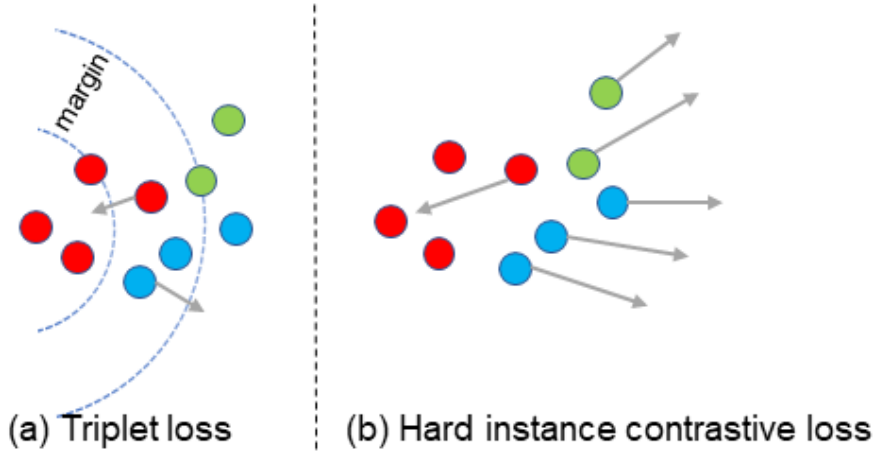


Figure 6.3: Comparison between triplet and hard instance contrastive loss.

Relation with triplet loss. Both $\mathcal{L}_{h\_ins}$ and triplet loss [51] pull an anchor closer to positive instances and away from negative instances. As shown in Fig. 6.3, the traditional triplet loss pushes away a negative pair from a positive pair by a margin. Differently, the proposed $\mathcal{L}_{h\_ins}$ pushes away all the negative instances as far as it could with a softmax. If we select one negative instance, the $\mathcal{L}_{h\_ins}$ can be transformed into the triplet loss. If we calculate pairwise distance within a mini-batch to select the hardest positive and the hardest negative instances, the $\mathcal{L}_{h\_ins}$ is equivalent to the batch-hard triplet loss[51]. We compare hard triplet loss (hardest negative) with the proposed $\mathcal{L}_{h\_ins}$ (all negatives). in Tab. 6.2.

Negative in $\mathcal{L}_{h\_ins}$	Market1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
hardest	80.1	92.8	68.2	82.5
all	**82.3**	**93.8**	**69.9**	**83.3**

Table 6.2: Comparison between using the hardest negative and all negatives in the denominator of $\mathcal{L}_{h\_ins}$.

6.3.4 Soft Instance Consistency Loss

Both proxy and hard instance contrastive losses are trained with one-hot hard pseudo labels, which can not capture the complex inter-instance similarity relationship between multiple pseudo positives and negatives. Especially, inter-instance similarity may change after data augmentation. As shown in Fig. 6.4, the anchor A becomes less similar to pseudo positives (P_1, P_2, P_3), because of the visual distortions. Meanwhile, the anchor A becomes more similar to pseudo negatives (N_1, N_2), since both of them have red shirts. By maintaining the consistency before and after augmentation, a model is supposed to be more invariant to augmentation perturbations. We use the inter-instance similarity scores without augmentation as soft labels to rectify those with augmentation.

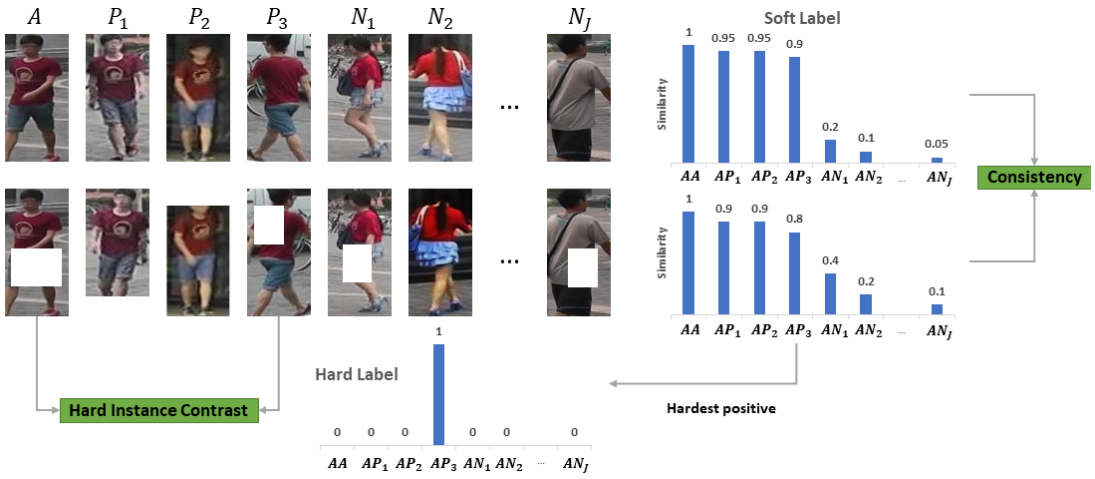


Figure 6.4: Based on inter-instance similarity ranking between anchor (A), pseudo positives (P) and pseudo negatives (N), **Hard Instance Contrastive Loss** matches an anchor with its hardest positive in a mini-batch. **Soft Instance Consistency Loss** regularizes the inter-instance similarity before and after data augmentation.

For a batch of images after data augmentation, we measure the inter-instance similarity between an anchor f_A with all the mini-batch $N_K \times N_P$ instances, as shown in Fig. 6.4. Then, the inter-instance similarity is turned into a prediction distribution P by a softmax:

$$P = \frac{\exp\left(< f_A \cdot m > / \tau_{s\_ins}\right)}{\sum_{j=1}^{N_P \times N_K} \exp\left(< f_A \cdot m_j > / \tau_{s\_ins}\right)} \tag{6.10}$$

where τ_{s_ins} is the soft instance temperature hyper-parameter. f_A is an online representation of the anchor, while m is momentum representation of each instance in a mini-batch.

For the same batch without data augmentation, we measure the inter-instance similarity between momentum representations of the same anchor with all the mini-batch $N_K \times N_P$ instances, because the momentum encoder is more stable. We get a target distribution Q:

$$Q = \frac{\exp\left(< m_A \cdot m > / \tau_{s\_ins}\right)}{\sum_{j=1}^{N_P \times N_K} \exp\left(< m_A \cdot m_j > / \tau_{s\_ins}\right)} \tag{6.11}$$

The soft instance consistency loss is Kullback-Leibler Divergence between two distributions:

$$\mathcal{L}_{s\_ins} = \mathcal{D}_{KL}(P||Q) \tag{6.12}$$

In previous methods, consistency is regularized between weakly augmented and strongly augmented images [110] or two sets of differently strong augmented images [137]. Some methods [70, 120] also adopted mean square error (MSE) as their consistency loss function. We compare our setting with other possible settings in Tab. 6.3.

Consistency	Market1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
MSE	80.0	92.7	68.4	82.1
Strong-strong Aug	80.4	92.8	68.2	82.5
ours	**82.3**	**93.8**	**69.9**	**83.3**

Table 6.3: Comparison of consistency loss. Ours refers to KL divergence between images with and without data augmentation.

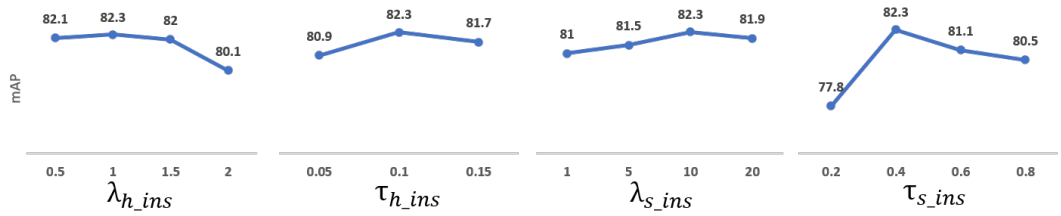


Figure 6.5: Parameter analysis on Market-1501 dataset.

6.4 Experiments

6.4.1 Datasets and Evaluation Protocols

Market-1501 [164], DukeMTMC-reID[105] and MSMT17 [138] datasets are used to evaluate our proposed method. Market-1501 dataset is collected in front of a supermarket in Tsinghua University from 6 cameras. It contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for test. DukeMTMC-reID is a subset of the DukeMTMC dataset. It contains 16,522 images of 702 persons for training, 2,228 query images and 17,661 gallery images of 702 persons for test from 8 cameras. MSMT17 is a large-scale Re-ID dataset, which contains 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras. Both Cumulative Matching Characteristics (CMC) Rank1, Rank5, Rank10 accuracies and mean Average Precision (mAP) are used in our experiments.

6.4.2 Implementation details

General training settings. To conduct a fair comparison with state-of-the-art methods, we use an ImageNet [106] pre-trained ResNet50 [49] as our backbone network. Instance-batch normalization (IBN) [97] has shown better performance than regular batch normalization in unsupervised domain adaptation

Backbone	Market1501				DukeMTMC-reID				MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
ResNet50	82.3	93.8	**97.6**	98.4	69.9	83.3	91.5	94.1	38.9	70.2	80.5	84.4
IBN-ResNet50	**82.5**	**94.2**	**97.6**	**98.5**	**70.7**	**83.6**	**91.9**	**93.9**	**40.6**	**70.7**	**81.0**	**84.6**

Table 6.4: Comparison of ResNet50 and IBN-ResNet50 backbones on Market1501, DukeMTMC-reID and MSMT17 datasets.

Threshold	Market1501				DukeMTMC-reID				MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
0.45	82.5	93.4	97.5	98.3	68.0	82.8	**91.5**	93.4	36.6	69.2	79.3	82.7
0.5	**83.0**	**94.1**	**97.7**	98.3	69.2	82.9	91.2	93.2	38.4	69.9	80.2	83.8
0.55	82.3	93.8	97.6	98.4	**69.9**	83.3	**91.5**	**94.1**	38.9	70.2	80.5	84.4
0.6	81.2	93.0	97.3	**98.5**	69.4	**83.5**	91.4	94.0	**39.4**	**70.9**	**81.0**	**84.5**

Table 6.5: Comparison of different distance thresholds on Market1501, DukeMTMC-reID and MSMT17 datasets.

[97, 42] and domain generalization [62]. We compare the performance of ICE with ResNet50 and IBN-ResNet50 backbones in Tab. 6.4. The performance of our proposed ICE can be further improved with an IBN-ResNet50 backbone network. An Adam optimizer with a weight decay rate of 0.0005 is used to optimize our networks. The learning rate is set to 0.00035 with a warm-up scheme in the first 10 epochs. No learning rate decay is used in the training. The momentum encoder is updated with a momentum coefficient $\alpha = 0.999$. We renew pseudo labels every 400 iterations and repeat this process for 40 epochs. We use a batchsize of 32 where $N_P = 8$ and $N_K = 4$. We set $\tau_a = 0.5$, $\tau_c = 0.07$ and $N_{neg} = 50$ in the proxy contrastive baseline. Our network is trained on 4 Nvidia 1080 GPUs under Pytorch framework. The total training time is around 2 hours on Market-1501. After training, only the momentum encoder is used for the inference.

Clustering settings. We calculate k-reciprocal Jaccard distance [170] for clustering, where k is set to 30. We set a minimum cluster samples to 4 and a distance threshold to 0.55 for DBSCAN.

In DBSCAN [31], the distance threshold is the maximum distance between two samples for one to be considered as in the neighborhood of the other. A smaller distance threshold is likely to make DBSCAN mark more hard positives as different classes. On the contrary, a larger distance threshold makes DBSCAN mark more hard negatives as same class.

The distance threshold for DBSCAN between same cluster neighbors is set to 0.55, which is a trade-off number for Market1501, DukeMTMC-reID and MSMT17 datasets. To get a better understanding of how ICE is sensitive to the distance threshold, we vary the threshold from 0.45 to 0.6. As shown in Tab. 6.5, a smaller threshold 0.5 is more appreciate for the relatively smaller dataset Market1501, while a larger threshold 0.6 is more appreciate for the relatively larger dataset MSMT17. State-of-the-art unsupervised ReID methods SpCL [42] and CAP [130] respectively used 0.6 and 0.5 as their distance threshold. Our proposed ICE can always outperform SpCL and CAP on the three datasets with a threshold between 0.5 and 0.6.

Data augmentation. All images are resized to 256×128. The strong data augmentation refers to random horizontal flipping, cropping, Gaussian blurring and erasing [171].

Camera-aware memory	Market1501				DukeMTMC-reID				MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Baseline $\mathcal{L}_{proxy}$	79.3	91.5	96.8	97.6	67.3	81.4	90.8	92.9	36.4	67.8	78.7	82.5
$+\mathcal{L}_{h\_ins}$	80.5	92.6	97.3	98.4	68.8	82.4	90.4	93.6	38.0	69.1	79.9	83.4
$+\mathcal{L}_{s\_ins}$	81.1	93.2	97.5	98.5	68.4	82.0	91.0	93.2	38.1	68.7	79.8	83.7
$+\mathcal{L}_{h\_ins}+\mathcal{L}_{s\_ins}$	**82.3**	**93.8**	**97.6**	**98.4**	**69.9**	**83.3**	**91.5**	**94.1**	**38.9**	**70.2**	**80.5**	**84.4**
Camera-agnostic memory	Market1501				DukeMTMC-reID				MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Baseline $\mathcal{L}_{agnostic}$	65.8	85.3	95.1	96.6	50.9	67.9	81.6	86.6	24.1	52.3	66.2	71.6
$+\mathcal{L}_{h\_ins}$	78.2	91.3	96.9	98.0	65.4	79.6	88.9	91.9	**30.3**	**60.8**	**72.9**	**77.6**
$+\mathcal{L}_{s\_ins}$	47.2	66.7	86.0	91.6	36.2	50.4	70.3	76.3	17.8	38.8	54.2	60.9
$+\mathcal{L}_{h\_ins}+\mathcal{L}_{s\_ins}$	**79.5**	**92.0**	**97.0**	**98.1**	**67.2**	**81.3**	**90.1**	**93.0**	29.8	59.0	71.7	77.0

Table 6.6: Comparison of different losses. Camera-aware memory occupies up to 6, 8 and 15 times memory space than camera-agnostic memory on Market1501, DukeMTMC-reID and MSMT17 datasets.

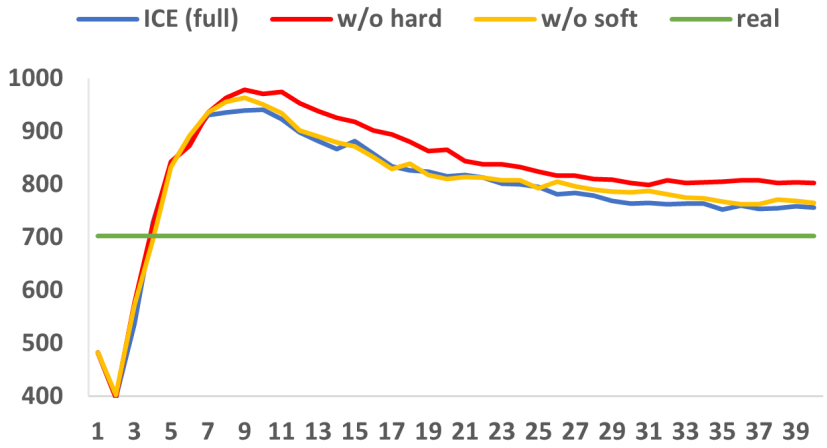


Figure 6.6: Dynamic cluster numbers during 40 training epochs on DukeMTMC-reID. "hard" and "soft" respectively denote L_{h_ins} and L_{s_ins}. A lower number denotes that clusters are more compact.

6.4.3 Parameter analysis

Compared to the proxy contrastive baseline, ICE brings in four more hyper-parameters, including λ_{h_ins}, τ_{h_ins} for hard instance contrastive loss and λ_{s_ins}, τ_{s_ins} for soft instance consistency loss. We analyze the sensitivity of each hyper-parameter on the Market-1501 dataset. The mAP results are illustrated in Fig. 6.5. As hardest positives are likely to be false positives, an overlarge λ_{h_ins} or undersized τ_{h_ins} introduce more noise. λ_{h_ins} and λ_{s_ins} balance the weight of each loss in Eq. (6.3). Given the results, we set $\lambda_{h\_ins}=1$ and $\lambda_{s\_ins}=10$. τ_{h_ins} and τ_{s_ins} control the similarity scale in hard instance contrastive loss and soft instance consistency loss. We finally set $\tau_{h\_ins}=0.1$ and $\tau_{s\_ins}=0.4$. Our hyper-parameters are tuned on Market-1501 and kept same for DukeMTMC-reID and MSMT17. Achieving state-of-the-art results simultaneously on the three datasets can validate the generalizability of these hyper-parameters.

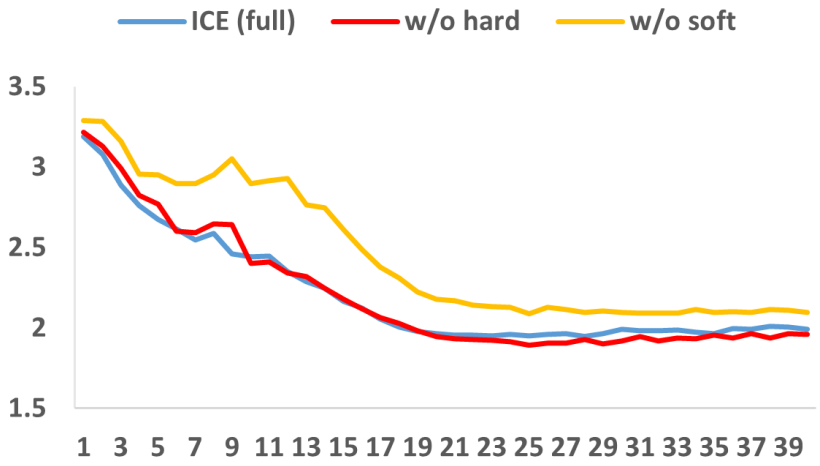


Figure 6.7: Dynamic KL divergence during 40 training epochs on DukeMTMC-reID. Lower KL divergence denotes that a model is more robust to augmentation perturbation.

6.4.4 Ablation study

The performance boost of ICE in unsupervised ReID mainly comes from the proposed hard instance contrastive loss and soft instance consistency loss. We conduct ablation experiments to validate the effectiveness of each loss, which is reported in Tab. 6.6. We illustrate the number of clusters during the training in Fig. 6.6 and t-SNE [91] after training in Fig. 6.9 to evaluate the compactness of clusters. We also illustrate the dynamic KL divergence of Eq. (6.12) to measure representation sensitivity to augmentation perturbation in Fig. 6.7 .

Hard instance contrastive loss. Our proposed $\mathcal{L}_{h\_ins}$ reduces the intra-class variance in a camera-agnostic manner, which increases the quality of pseudo labels. By reducing intra-class variance, a cluster is supposed to be more compact. With a same clustering algorithm, we expect to have less clusters when clusters are more compact. As shown in Fig. 6.6, DBSCAN generated more clusters during the training without our proposed $\mathcal{L}_{h\_ins}$. The full ICE framework has less clusters, which are closer to the real number of identities in the training set. On the other hand, as shown in Fig. 6.9, the full ICE framework has a better intra-class compactness and inter-class separability than the camera-aware baseline in the test set. The compactness contributes to better unsupervised ReID performance in Tab. 6.6.

Soft instance consistency loss. Hard instance contrastive loss reduces the intra-class variance between naturally captured views, while soft instance consistency loss mainly reduces the variance from artificially augmented perturbation. If we compare the blue (ICE full) and yellow (w/o soft) curves in Fig. 6.7, we can find that the model trained without $\mathcal{L}_{s\_ins}$ is less robust to augmentation perturbation. The quantitative results in Tab. 6.6 confirms that

the $\mathcal{L}_{s\_ins}$ improves the performance of baseline. The best performance can be obtained by applying $\mathcal{L}_{h\_ins}$ and $\mathcal{L}_{s\_ins}$ on the camera-aware baseline.

Camera-agnostic scenario. Above results are obtained with a camera-aware memory, which strongly relies on ground truth camera ids. We further validate the effectiveness of the two proposed losses with a camera-agnostic memory, whose results are also reported in Tab. 6.6. Our proposed $\mathcal{L}_{h\_ins}$ significantly improves the performance from the camera-agnostic baseline. However, $\mathcal{L}_{s\_ins}$ **should be used under low intra-class variance, which can be achieved by the variance constraints on camera styles** $\mathcal{L}_{cross}$ **and hard samples** $\mathcal{L}_{h\_ins}$. $\mathcal{L}_{h\_ins}$ reduces intra-class variance, so that $AA \approx AP_1 \approx AP_2 \approx AP_3 \approx 1$ before augmentation in Fig. 6.4. $\mathcal{L}_{s\_ins}$ permits that we still have $AA \approx AP_1 \approx AP_2 \approx AP_3 \approx 1$ after augmentation. However, when strong variance exists, *e.g.*, $AA \not\approx AP_1 \not\approx AP_2 \not\approx AP_3 \not\approx 1$, maintaining this relationship equals maintaining intra-class variance, which decreases the ReID performance. On medium datasets (*e.g.*, Market1501 and DukeMTMC-reID) without strong camera variance, our proposed camera-agnostic intra-class variance constraint $\mathcal{L}_{h\_ins}$ is enough to make $\mathcal{L}_{s\_ins}$ beneficial to ReID. On large datasets (*e.g.*, 15 cameras in MSMT17) with strong camera variance, only camera-agnostic variance constraint $\mathcal{L}_{h\_ins}$ is not enough.

We provide the dynamic cluster numbers of camera-agnostic ICE in Fig. 6.8. The red curve is trained without the hard instance contrastive loss $\mathcal{L}_{h\_ins}$ as intra-class variance constraint. In this case, the soft instance consistency loss $\mathcal{L}_{s\_ins}$ maintains high intra-class variance, *e.g.*, $AA \not\approx AP_1 \not\approx AP_2 \not\approx AP_3 \not\approx 1$, which leads to less compact clusters. The orange curve is trained without $\mathcal{L}_{s\_ins}$, which has less clusters at the beginning but more clusters at last epochs than the blue curve. The blue curve is trained with both $\mathcal{L}_{h\_ins}$ and $\mathcal{L}_{s\_ins}$, whose cluster number is most accurate among the three curves at last epochs. Fig. 6.8 confirms that combining $\mathcal{L}_{h\_ins}$ and $\mathcal{L}_{s\_ins}$ reduces naturally captured and artificially augmented view variance at the same time, which gives optimal ReID performance.

6.4.5 Comparison with state-of-the-art methods

We compare ICE with state-of-the-art ReID methods in Tab. 6.7.

Comparison with unsupervised method. Previous unsupervised methods can be categorized into unsupervised domain adaptation (UDA) and fully unsupervised methods. We first list state-of-the-art UDA methods, including MMCL [125], JVTC [75], DG-Net++ [179], ECN+ [174], MMT [39], DCML [10], MEB [156], SpCL [42] and ABMT [13]. UDA methods usually rely on source domain annotation to reduce the pseudo label noise. Without any identity annotation, our proposed ICE outperforms all of them on the three datasets.

Under the fully unsupervised setting, ICE also achieves better performance than state-of-the-art methods, including BUC [84], SSL [85], MMCL [125],

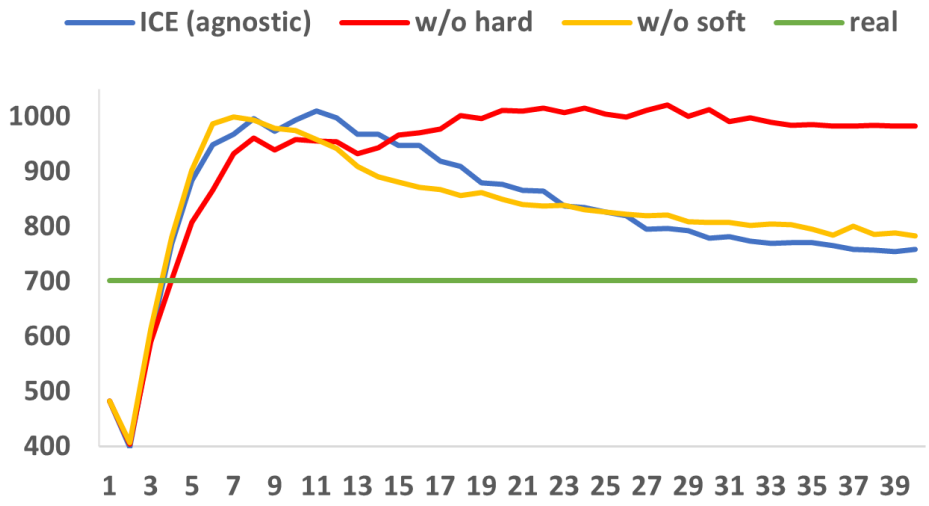


Figure 6.8: Dynamic cluster numbers of **ICE(agnostic)** during 40 training epochs on DukeMTMC-reID. A lower number denotes that clusters are more compact (less intra-cluster variance).

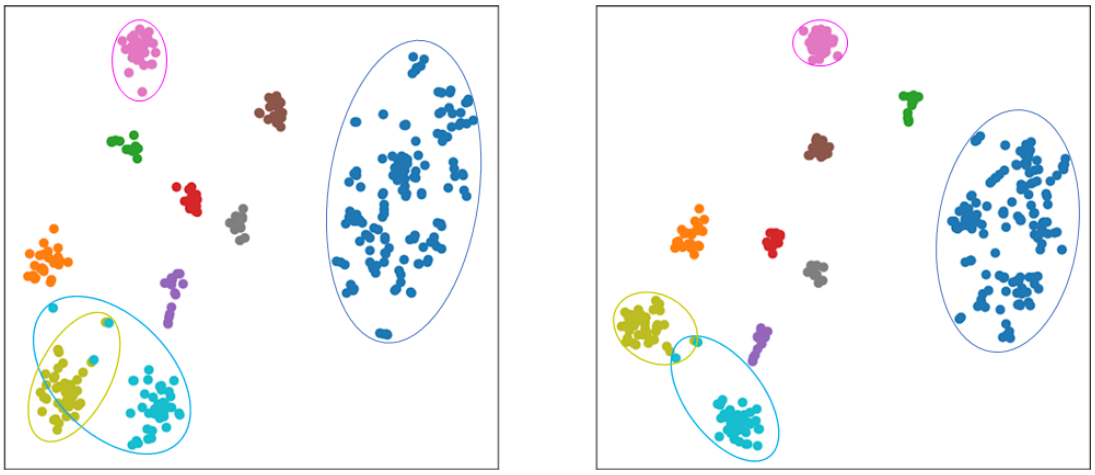


Figure 6.9: T-SNE visualization of 10 random classes in DukeMTMC-reID test set between **camera-aware baseline (Left)** and **ICE (Right)**.

Method	Reference	Market1501				DukeMTMC-reID				MSMT17			
		mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
Unsupervised Domain Adaptation													
MMCL [125]	CVPR'20	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0	16.2	43.6	54.3	58.9
JVTC [75]	ECCV'20	61.1	83.8	93.0	95.2	56.2	75.0	85.1	88.2	20.3	45.4	58.4	64.3
DG-Net++ [179]	ECCV'20	61.7	82.1	90.2	92.7	63.8	78.9	87.8	90.4	22.1	48.8	60.9	65.9
ECN+ [174]	TPAMI'20	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4	16.0	42.5	55.9	61.5
MMT [39]	ICLR'20	71.2	87.7	94.9	96.9	65.1	78.0	88.8	92.5	23.3	50.1	63.9	69.8
DCML [10]	ECCV'20	72.6	87.9	95.0	96.7	63.3	79.1	87.2	89.4	-	-	-	-
MEB [156]	ECCV'20	76.0	89.9	96.0	97.5	66.1	79.6	88.3	92.2	-	-	-	-
SpCL [42]	NeurIPS'20	76.7	90.3	96.2	97.7	68.8	82.9	90.1	92.5	26.8	53.7	65.0	69.8
ABMT [13]	WACV'21	78.3	92.5	-	-	69.1	82.0	-	-	26.5	54.3	-	-
Fully Unsupervised													
BUC [84]	AAAI'19	29.6	61.9	73.5	78.2	22.1	40.4	52.5	58.2	-	-	-	-
SSL [85]	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9	-	-	-	-
JVTC [75]	ECCV'20	41.8	72.9	84.2	88.7	42.2	67.6	78.0	81.6	15.1	39.0	50.9	56.8
MMCL [125]	CVPR'20	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0	11.2	35.4	44.8	49.8
HCT [155]	CVPR'20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4	-	-	-	-
CycAs [135]	ECCV'20	64.8	84.8	-	-	60.1	77.9	-	-	26.7	50.1	-	-
GCL [15]	CVPR'21	66.8	87.3	93.5	95.5	62.8	82.9	87.1	88.5	21.3	45.7	58.6	64.5
SpCL(agnostic) [42]	NeurIPS'20	73.1	88.1	95.1	97.0	65.3	81.2	90.3	92.2	19.1	42.3	55.6	61.2
ICE(agnostic)	This chapter	79.5	92.0	97.0	98.1	67.2	81.3	90.1	93.0	29.8	59.0	71.7	77.0
CAP(aware)[130]	AAAI'21	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8	36.9	67.4	78.0	81.4
ICE(aware)	This chapter	82.3	93.8	97.6	98.4	69.9	83.3	91.5	94.1	38.9	70.2	80.5	84.4
Supervised													
PCB [118]	ECCV'18	81.6	93.8	97.5	98.5	69.2	83.3	90.5	92.5	40.4	68.2	-	-
DG-Net [165]	CVPR'19	86.0	94.8	-	-	74.8	86.6	-	-	52.3	77.2	-	-
ICE (w/ ground truth)	This chapter	86.6	95.1	98.3	98.9	76.5	88.2	94.1	95.7	50.4	76.4	86.6	90.0

Table 6.7: Comparison of ReID methods on Market1501, DukeMTMC-reID and MSMT17 datasets. The best and second best unsupervised results are marked in red and blue.

JVTC [75], HCT [155], CycAs [135], GCL [15], SpCL [42] and CAP [130]. CycAs leveraged temporal information to assist visual matching, while our method only considers visual similarity. SpCL and CAP are based on proxy contrastive learning, which are considered respectively as camera-agnostic and camera-aware baselines in our method. With a camera-agnostic memory, the performance of ICE(agnostic) remarkably surpasses the camera-agnostic baseline SpCL, especially on Market1501 and MSMT17 datasets. With a camera-aware memory, ICE(aware) outperforms the camera-aware baseline CAP on all the three datasets. By mining hard positives to reduce intra-class variance, ICE is more robust to hard samples. We illustrate some hard examples in Fig. 6.10, where ICE succeeds to notice important visual clues, *e.g.*, characters in the shirt (1st row), blonde hair (2nd row), brown shoulder bag (3rd row) and badge (4th row).

Comparison with supervised method. We further provide two well-known supervised methods for reference, including the Part-based Convolutional Baseline (PCB) [118] and the joint Discriminative and Generative Network (DG-Net) [165]. Unsupervised ICE achieves competitive performance with PCB. If we replace the clustering generated pseudo labels with ground truth, our ICE can be transformed into a supervised method. The supervised ICE is competitive with state-of-the-art supervised ReID methods (*e.g.*, DG-Net), which shows that the supervised contrastive learning has a potential to be considered into future supervised ReID.

Figure 6.10: Comparison of top 5 retrieved images on Market1501 between CAP [130] and ICE. Green boxes denote correct results, while red boxes denote false results. Important visual clues are marked with red dashes.

6.5 Conclusion

In this chapter, we propose a novel inter-instance contrastive encoding method ICE to address unsupervised ReID. Deviated from previous proxy based contrastive ReID methods, we focus on inter-instance affinities to make a model more robust to data variance. We first mine the hardest positive with mini-batch instance pairwise similarity ranking to form a hard instance contrastive loss, which effectively reduces intra-class variance. Smaller intra-class variance contributes to the compactness of clusters. Then, we use mini-batch instance pairwise similarity scores as soft labels to enhance the consistency before and after data augmentation, which makes a model robust to artificial augmentation variance. By combining the proposed hard instance contrastive loss and soft instance consistency loss, ICE significantly outperforms previous unsupervised ReID methods on Market1501, DukeMTMC-reID and MSMT17 datasets. In the next chapter, we talk about limitations of recent person ReID works and perspectives for future person ReID research.

Chapter 7

Conclusion and Perspective

In this thesis, we first go through the general context of person ReID, including its definition, applications and major challenges. Towards a discriminative and generalizable person ReID system, we mainly concentrate on insufficient data and insufficient annotation problems. Revolving around data augmentation, we have proposed several novel methods for person ReID, involving both supervised and unsupervised methods. Each method has been extensively evaluated on main-stream person ReID datasets. We conclude this thesis by pointing out our main contributions in Section 7.1 and their limitations in Section 7.2. In the end, we discuss future perspectives of our work in Section 7.3.

7.1 Contributions

We have made four main contributions in this thesis, including Chapter 3: Spatial-channel partitions, Chapter 4: Asymmetric branches for teacher-student networks, Chapter 5: Joint generative and contrastive learning and Chapter 6: Inter-instance contrastive encoding.

Spatial-channel partitions

We propose a spatial-channel partition method to learn discriminative and generalizable representations for supervised person ReID. Previous part-based ReID models mainly focused on extracting local features from different spatial regions, but neglected channel-wise local features. We conduct a study to show the difference between spatial attention, channel attention, spatial partition and channel partition, which demonstrates that spatial partition and channel partition are complementary. We further use spatial-channel partitions in a pyramidal structure, which helps us to build robust representations with multi-granularity local features. The proposed spatial-channel partitions can be regarded as a feature-level data augmentation technique, which enhances feature diversity in representations.

Asymmetric branches for teacher-student networks

We propose another feature-level data augmentation technique by modifying neural network structures. In teacher-student networks, the teacher and the student are prone to converge to each other, preventing them from exploring diversified features. To overcome this issue, we replace original last layers of a backbone network, such as a ResNet-50, with two asymmetric branches of different depths and global pooling methods. The teacher asymmetric branches supervise the student asymmetric branches in a cross-branch manner, which alleviates the consensus in knowledge distillation process.

Joint generative and contrastive learning

We propose an image-level data augmentation technique, which consists in using GAN to generate augmented views as a replacement of traditional data augmentation. As traditional data augmentation may bring in undesired distortions on identity features, we use a GAN instead to disentangle a person image into id-related and id-unrelated features. GAN-based augmentation has the capabilities to only modify id-unrelated features while preserving id-related features. By contrasting augmented and original views from different identities, a neural network learns unsupervised representations that are invariant to augmented views.

Inter-instance contrastive encoding

As the last contribution, we propose inter-instance contrastive encoding, a unified contrative framework that combines a class-level contrastive loss, a hard instance-level contrastive loss and a soft instance-level consistency loss, for unsupervised person ReID. The hard instance-level contrastive loss aims at reducing the intra-class variance by mining and contrasting hard pseudo positive samples, while the soft instance-level consistency loss aims at making the network robust to perturbations mimicked by data augmentation.

7.2 Limitations

Although significant performance improvements have been witnessed on main-stream datasets, the earlier mentioned contributions still face some limitations in real-world deployments.

Computational complexity

To quickly match a target person, the computational speed, especially the inference speed, is an important factor to be considered in real-world deployments. The first limitation on computational complexity is related to our first two contributions on feature-level data augmentation. Our proposed spatial-channel partitions and asymmetric branches use more parameters and operations to extract more diversified features. Both of them enlarge the dimension

of representations, which eventually influences the inference speed. For example, our spatial-channel partition enlarges the ResNet-50 representation dimension from 2048 to 3328, which results in slower inference speed for large-scale person ReID scenario.

Hyper-parameter sensitivity

Most of unsupervised person ReID methods, including Asymmetric branches for teacher-student networks in Chapter 4, Joint generative and contrastive learning in Chapter 5 and Inter-instance contrastive encoding in Chapter 6, are based on clustering-generated pseudo labels. Clustering algorithms are usually sensitive to hyper-parameters for different domains. For example, we need to specify a maximal neighbor distance and a minimal neighbor number in the density-based clustering DBSCAN [31] to have the optimal ReID performance. However, as the scale and properties of different domains are quite different, it is difficult to keep these clustering hyper-parameters always optimal for different datasets.

Data augmentation quality

Our contributions are strongly related to data augmentation. However, in reality, it is difficult to always have high-quality augmented views. For example, our proposed Joint generative and contrastive learning in Chapter 5 uses a GAN to generate a same person in different poses and view-points as augmentation. However, the current GAN-based augmentation is prone to fail, for example, when we use front appearance to generate back appearance. Furthermore, the GAN-based augmentation can easily lose detailed appearance information. We hope new advances in image generation will eventually increase generated person image quality and make GAN-based augmentation more suitable for discriminative tasks.

Domain generalizability

Our proposed unsupervised person ReID methods, including Asymmetric branches for teacher-student networks in Chapter 4, Joint generative and contrastive learning in Chapter 5 and Inter-instance contrastive encoding in Chapter 6, are designed for single-domain scenario. In this case, even though human supervision is not needed, collecting enough data and re-training the model are still mandatory every time the person ReID system enters into a new domain. However, the changeable weather condition and illumination level during a day make such a re-training process impractical in real-world deployments. The domain generalizability on multiple seen domains and unseen domains remains a understudied question for person ReID.

7.3 Perspectives

Previous person ReID methods focus more on single-domain discriminability rather than multi-domain generalizability. However, an ideal person ReID system should be discriminative in each single domain while generalizable to different domains. Towards this goal, current person ReID systems can be improved from following directions.

Multi-domain generalization

Domain generalization is defined as a task of learning shared domain-invariant features from multiple domains for novel unseen domains. Due to its significance in real-world deployments, domain generalization has attracted increasing attention in person ReID community. However, recent domain generalizable person ReID methods [111, 63, 16] rely on multiple large-scale labeled datasets to learn domain-invariant features, in which annotating multiple datasets is an extremely cumbersome task. As our unsupervised methods can get competitive performance with supervised methods in a single domain, it is interesting to see if unsupervised methods can achieve similar performance in the multi-domain generalization scenario.

Pre-training

Pre-training on ultra-large scale datasets [2] is another approach that has proven to be effective in enhancing the generalizability of a person ReId method. Unsupervised pre-training [21] explores inter-instance relationship, which has shown better transferability in down-stream tasks than supervised pre-training. However, clustering can be extremely time-consuming on ultra-large scale datasets, making clustering-based unsupervised person ReID methods impossible to be unsupervised pre-training methods. Design efficient and effective unsupervised person ReID pre-training methods is worth exploring in the future.

Lifelong learning

Due to constraints on hardware and privacy protection, an ultra-large scale dataset is hard to collect at once in real world. Instead, when a camera network is deployed, data can be recorded day by day. A generalizable model can be trained without an ultra-large scale dataset, but with knowledge accumulated from daily data. Lifelong learning is a research topic that targets at adapting a model to new domains without forgetting knowledge learnt from old domains. A recent work [99] has proven that a lifelong sequential training on several domains can achieve similar generalizability with training on several domains at once. Furthermore, it will be more interesting if we can extend unsupervised person ReID methods to unsupervised lifelong person ReID methods.

To summarize, recent supervised and unsupervised person ReID algorithms show impressive performance on medium-sized datasets, such as Market-1501 and DukeMTMC-reID. Such methods can be deployed in small or medium surveillance scenarios, such as a cashier-less supermarket that has around 10 cameras. But for a larger scale surveillance scenario that contains more diversity (indoor/outdoor, different illumination levels, weather conditions, seasons, *etc*.), more robust and generalizable person ReID algorithms are needed. Towards a robust person ReID system, we will consider more on domain generalization, large-scale pre-training and lifelong learning for future person ReID research.

Bibliography

[1] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu and Y. Xu. 'Deep-Person: Learning Discriminative Deep Features for Person Re-Identification'. In: *ArXiv* abs/1711.10658 (2020) (cit. on p. 16).

[2] Y. Bai, J. Jiao, W. Ce, J. Liu, Y. Lou, X. Feng et al. 'Person30K: A Dual-Meta Generalization Network for Person Re-Identification'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2123–2132 (cit. on pp. 11, 98).

[3] S. Bak, P. Carr and J.-F. Lalonde. 'Domain adaptation through synthesis for unsupervised person re-identification'. In: *ECCV*. 2018 (cit. on pp. 22, 23, 58).

[4] S. Bak, E. Corvée, F. Brémond and M. Thonnat. 'Person Re-identification Using Spatial Covariance Regions of Human Body Parts'. In: *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance* (2010), pp. 435–440 (cit. on pp. 13, 21, 28).

[5] S. Bąk, G. Charpiat, E. Corvée, F. Brémond and M. Thonnat. 'Learning to Match Appearances by Correlations in a Covariance Metric Space'. In: *ECCV*. 2012 (cit. on p. 13).

[6] A. Brock, J. Donahue and K. Simonyan. 'Large Scale GAN Training for High Fidelity Natural Image Synthesis'. In: *ICLR*. 2019 (cit. on p. 61).

[7] J. Carreira and A. Zisserman. 'Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4724–4733 (cit. on p. 14).

[8] C. Chan, S. Ginosar, T. Zhou and A. A. Efros. 'Everybody dance now'. In: *ICCV*. 2019 (cit. on p. 61).

[9] G. Chechik, V. Sharma, U. Shalit and S. Bengio. 'Large Scale Online Learning of Image Similarity Through Ranking'. In: *J. Mach. Learn. Res.* 11 (2010), pp. 1109–1135 (cit. on p. 19).

[10] G. Chen, Y. Lu, J. Lu and J. Zhou. 'Deep Credible Metric Learning for Unsupervised Domain Adaptation Person Re-identification'. In: *ECCV*. 2020 (cit. on pp. 79, 91, 93).

[11] H. Chen, B. Lagadec and F. Bremond. 'Partition and Reunion: A Two-Branch Neural Network for Vehicle Re-identification'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019 (cit. on p. 27).

[12] H. Chen, B. Lagadec and F. Bremond. 'Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification'. In: *The IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020 (cit. on pp. 20, 21, 27, 47, 78).

[13] H. Chen, B. Lagadec and F. Bremond. 'Enhancing Diversity in Teacher-Student Networks via Asymmetric Branches for Unsupervised Person Re-Identification'. In: *WACV*. 2021 (cit. on pp. 43, 61, 79, 91, 93).

[14] H. Chen, B. Lagadec and F. Bremond. 'ICE: Inter-instance Contrastive Encoding for Unsupervised Person Re-identification'. In: *arXiv preprint arXiv:2103.16364* (2021) (cit. on pp. 20, 78).

[15] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva and F. Bremond. 'Joint Generative and Contrastive Learning for Unsupervised Person Re-Identification'. In: *CVPR*. 2021 (cit. on pp. 58, 79, 93).

[16] P. Chen, P. Dai, J. Liu, F. Zheng, Q. Tian and R. Ji. 'Dual Distribution Alignment Network for Generalizable Person Re-Identification'. In: *AAAI*. 2021 (cit. on p. 98).

[17] T. Chen, S. Kornblith, M. Norouzi and G. Hinton. 'A Simple Framework for Contrastive Learning of Visual Representations'. In: *ICML*. 2020 (cit. on pp. 25, 58–60, 78, 80).

[18] T. Chen, S. Kornblith, K. Swersky, M. Norouzi and G. Hinton. 'Big Self-Supervised Models are Strong Semi-Supervised Learners'. In: *NeurIPS*. 2020 (cit. on p. 80).

[19] W. Chen, X. Chen, J. Zhang and K. Huang. 'Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1320–1329 (cit. on p. 20).

[20] X. Chen, H. Fan, R. Girshick and K. He. 'Improved baselines with momentum contrastive learning'. In: *arXiv preprint arXiv:2003.04297* (2020) (cit. on pp. 58, 60, 80).

[21] X. Chen and K. He. 'Exploring Simple Siamese Representation Learning'. In: *CVPR*. 2021 (cit. on pp. 25, 98).

[22] Y. Chen, X. Zhu and S. Gong. 'Instance-guided context rendering for cross-domain person re-identification'. In: *ICCV*. 2019 (cit. on pp. 22, 23, 44, 61, 66, 79).

[23] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani and V. Murino. 'Custom Pictorial Structures for Re-identification'. In: *BMVC*. 2011 (cit. on pp. 11, 21).

[24] Z. Dai, M. Chen, X. Gu, S. Zhu and P. Tan. 'Batch DropBlock network for person re-identification and beyond'. In: *ICCV*. 2019 (cit. on pp. 22, 47).

[25] A. Das, A. Chakraborty and A. Roy-Chowdhury. 'Consistent Re-identification in a Camera Network'. In: *ECCV*. 2014 (cit. on p. 13).

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. 'ImageNet: A large-scale hierarchical image database'. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255 (cit. on pp. 6, 14, 35, 51).

[27] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye and J. Jiao. 'Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification'. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 994–1003 (cit. on pp. 40, 41).

[28] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang and J. Jiao. 'Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification'. In: *CVPR*. 2018 (cit. on pp. 58, 61).

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *ArXiv* abs/2010.11929 (2021) (cit. on p. 16).

[30] C. Eom and B. Ham. 'Learning Disentangled Representation for Robust Person Re-identification'. In: *NeurIPS*. 2019 (cit. on pp. 22, 61, 69, 70, 73).

[31] M. Ester, H.-P. Kriegel, J. Sander and X. Xu. 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: *KDD*. 1996 (cit. on pp. 24, 53, 64, 82, 88, 97).

[32] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang and W. Jiang. 'SCPNet: Spatial-Channel Parallelism Network for Joint Holistic and Partial Person Re-Identification'. In: *ACCV*. 2018 (cit. on pp. 30, 39).

[33] M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani. 'Person re-identification by symmetry-driven accumulation of local features'. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), pp. 2360–2367 (cit. on pp. 13, 21, 30).

[34] P. F. Felzenszwalb, D. A. McAllester and D. Ramanan. 'A discriminatively trained, multiscale, deformable part model'. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition* (2008), pp. 1–8 (cit. on p. 36).

[35] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang et al. 'Unsupervised Pre-training for Person Re-identification'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2021) (cit. on p. 6).

[36] Y. Fu, Y. Wei, G. Wang, X. Zhou, H. Shi and T. S. Huang. 'Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2018), pp. 6111–6120 (cit. on pp. 43, 45, 52).

[37] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi and T. S. Huang. 'Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification'. In: *ICCV*. 2019 (cit. on pp. 24, 61, 66, 67, 79).

[38] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang et al. 'Horizontal Pyramid Matching for Person Re-identification'. In: *AAAI*. 2019 (cit. on pp. 21, 28, 31, 39, 40, 47).

[39] Y. Ge, D. Chen and H. Li. 'Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification'. In: *International Conference on Learning Representations*. 2020 (cit. on pp. 20, 24, 43–46, 48–50, 52–55, 61, 66, 67, 79, 83, 91, 93).

[40] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang et al. 'FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification'. In: *NeurIPS*. 2018 (cit. on pp. 22, 59, 61, 69, 70, 73).

[41] Y. Ge, H. Wang, F. Zhu, R. Zhao and H. Li. 'Self-supervising Fine-grained Region Similarities for Large-scale Image Localization'. In: *ECCV*. 2020 (cit. on p. 80).

[42] Y. Ge, F. Zhu, D. Chen, R. Zhao and H. Li. 'Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID'. In: *NeurIPS*. 2020 (cit. on pp. 25, 78, 80, 82, 88, 91, 93).

[43] N. Gheissari, T. Sebastian and R. Hartley. 'Person Reidentification Using Spatiotemporal Appearance'. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 2 (2006), pp. 1528–1535 (cit. on p. 13).

[44] G. Ghiasi, T.-Y. Lin and Q. V. Le. 'DropBlock: A regularization method for convolutional networks'. In: *NeurIPS*. 2018 (cit. on p. 22).

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair et al. 'Generative adversarial nets'. In: *NeurIPS*. 2014 (cit. on pp. 18, 58, 61).

[46] D. Gray and H. Tao. 'Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features'. In: *ECCV*. 2008 (cit. on pp. 11, 13, 21, 28, 30).

[47] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu et al. 'Co-teaching: Robust training of deep neural networks with extremely noisy labels'. In: *NeurIPS*. 2018 (cit. on pp. 44, 45).

[48] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick. 'Momentum Contrast for Unsupervised Visual Representation Learning'. In: *CVPR*. 2020 (cit. on pp. 25, 58–60, 78, 80, 82).

[49] K. He, X. Zhang, S. Ren and J. Sun. 'Deep Residual Learning for Image Recognition'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778 (cit. on pp. 14, 15, 34, 47, 52, 66, 82, 87).

[50] S. He, H. Luo, P. Wang, F. Wang, H. Li and W. Jiang. 'TransReID: Transformer-based Object Re-Identification'. In: *ArXiv* abs/2102.04378 (2021) (cit. on p. 16).

[51] A. Hermans, L. Beyer and B. Leibe. 'In Defense of the Triplet Loss for Person Re-Identification'. In: *CoRR* abs/1703.07737 (2017) (cit. on pp. 19, 24, 30, 35, 40, 47, 78, 85).

[52] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter. 'Gans trained by a two time-scale update rule converge to a local nash equilibrium'. In: *NeurIPS*. 2017 (cit. on p. 73).

[53] G. E. Hinton, O. Vinyals and J. Dean. 'Distilling the Knowledge in a Neural Network'. In: *ArXiv* abs/1503.02531 (2015) (cit. on pp. 19, 49).

[54] M. Hirzer, C. Beleznai, P. Roth and H. Bischof. 'Person Re-identification by Descriptive and Discriminative Classification'. In: *SCIA*. 2011 (cit. on pp. 11, 12).

[55] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand et al. 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications'. In: *CoRR* abs/1704.04861 (2017) (cit. on pp. 16, 31).

[56] J. Hu, L. Shen and G. Sun. 'Squeeze-and-Excitation Networks'. In: *CVPR*. 2018 (cit. on pp. 30, 32).

[57] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger. 'Densely Connected Convolutional Networks'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2261–2269 (cit. on p. 15).

[58] X. Huang and S. Belongie. 'Arbitrary style transfer in real-time with adaptive instance normalization'. In: *ICCV*. 2017 (cit. on p. 68).

[59] Y. Huang, Q. Wu, J. Xu and Y. Zhong. 'SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 (cit. on p. 22).

[60] S. Ioffe and C. Szegedy. 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'. In: *ICML*. 2015 (cit. on pp. 31, 35).

[61] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros. 'Image-to-image translation with conditional adversarial networks'. In: *CVPR*. 2017 (cit. on p. 68).

[62] J. Jia, Q. Ruan and T. M. Hospedales. 'Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice'. In: *BMVC*. 2019 (cit. on p. 88).

[63] X. Jin, C. Lan, W. Zeng, Z. Chen and L. Zhang. 'Style Normalization and Restitution for Generalizable Person Re-Identification'. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 3140–3149 (cit. on p. 98).

[64] A. Kanazawa, M. J. Black, D. W. Jacobs and J. Malik. 'End-to-end Recovery of Human Shape and Pose'. In: *CVPR*. 2018 (cit. on pp. 59, 62).

[65] T. Karras, S. Laine and T. Aila. 'A style-based generator architecture for generative adversarial networks'. In: *CVPR*. 2019 (cit. on p. 61).

[66] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila. 'Analyzing and Improving the Image Quality of StyleGAN'. In: *CVPR*. 2020 (cit. on p. 61).

[67] Z. Ke, D. Wang, Q. Yan, J. Ren and R. W. H. Lau. 'Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 6727–6735 (cit. on p. 45).

[68] A. Krizhevsky. 'Learning Multiple Layers of Features from Tiny Images'. In: *University of Toronto* (May 2012) (cit. on p. 14).

[69] A. Krizhevsky, I. Sutskever and G. E. Hinton. 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Commun. ACM* 60 (2012), pp. 84–90 (cit. on pp. 15, 31).

[70] S. Laine and T. Aila. 'Temporal Ensembling for Semi-Supervised Learning'. In: *ICLR*. 2017 (cit. on pp. 45, 80, 86).

[71] *Le palmarès des 50 plus grandes villes vidéosurveillées*. https://www.lagazettedescommunes.com/660599/le-palmares-des-50-plus-grandes-villes-videosurveillees/. Accessed: 2021-08-23 (cit. on p. 1).

[72] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard et al. 'Backpropagation Applied to Handwritten Zip Code Recognition'. In: *Neural Computation* 1 (1989), pp. 541–551 (cit. on p. 15).

[73] Y.-J. Li, C.-S. Lin, Y.-B. Lin and Y.-C. F. Wang. 'Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation'. In: *ICCV*. 2019 (cit. on pp. 22–24, 58, 59, 61, 66, 73).

[74] J. Li, J. Wang, Q. Tian, W. Gao and S. Zhang. 'Global-Local Temporal Representations for Video Person Re-Identification'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 3957–3966 (cit. on p. 12).

[75] J. Li and S. Zhang. 'Joint Visual and Temporal Consistency for Unsupervised Domain Adaptive Person Re-Identification'. In: *ECCV*. 2020 (cit. on pp. 24, 59, 61, 66–69, 80, 91, 93).

[76] J. Li, S. Zhang and T. Huang. 'Multi-scale 3D Convolution Network for Video Based Person Re-Identification'. In: *AAAI*. 2019 (cit. on p. 40).

[77] S. Li, S. Bak, P. Carr and X. Wang. 'Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-Identification'. In: *CVPR*. 2018 (cit. on p. 40).

[78] W. Li and X. Wang. 'Locally Aligned Feature Transforms across Views'. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3594–3601 (cit. on p. 11).

[79] W. Li, R. Zhao and X. Wang. 'Human Reidentification with Transferred Metric Learning'. In: *ACCV*. 2012 (cit. on p. 11).

[80] W. Li, R. Zhao, T. Xiao and X. Wang. 'DeepReID: Deep Filter Pairing Neural Network for Person Re-identification'. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 152–159 (cit. on pp. 11, 18, 21, 36).

[81] W. Li, X. Zhu and S. Gong. 'Harmonious Attention Network for Person Re-Identification'. In: *CVPR*. 2018 (cit. on pp. 17, 27, 32, 39, 40).

[82] S. Liao, Y. Hu, X. Zhu and S. Z. Li. 'Person re-identification by Local Maximal Occurrence representation and metric learning'. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 2197–2206 (cit. on pp. 14, 21, 27).

[83] S. Lin, H. Li, V. Sanchez and A. C. Kot. 'Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification'. In: *BMVC*. 2018 (cit. on pp. 24, 28, 44, 61, 79).

[84] Y. Lin, X. Dong, L. Zheng, Y. Yan and Y. Yang. 'A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification'. In: *AAAI*. 2019 (cit. on pp. 24, 45, 53, 61, 66, 79, 91, 93).

[85] Y. Lin, L. Xie, Y. Wu, C. Yan and Q. Tian. 'Unsupervised Person Re-identification via Softened Similarity Learning'. In: *ArXiv* abs/2004.03547 (2020) (cit. on pp. 24, 43, 46, 53, 61, 66, 79, 91, 93).

[86] J. Liu, Z.-J. Zha, D. Chen, R. Hong and M. Wang. 'Adaptive Transfer Network for Cross-Domain Person Re-Identification'. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on pp. 40, 41).

[87] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen and J. Bu. 'Semi-supervised Coupled Dictionary Learning for Person Re-identification'. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 3550–3557 (cit. on p. 21).

[88] C. C. Loy, T. Xiang and S. Gong. 'Multi-camera activity correlation analysis'. In: *CVPR*. 2009 (cit. on p. 11).

[89] C. Luo, C. Song and Z. Zhang. 'Generalizing Person Re-Identification by Camera-Aware Invariance Learning and Cross-Domain Mixup'. In: *ECCV*. 2020 (cit. on pp. 24, 61, 66, 67).

[90] H. Luo, Y. Gu, X. Liao, S. Lai and W. Jiang. 'Bag of Tricks and a Strong Baseline for Deep Person Re-Identification'. In: *CVPR Workshops*. 2019 (cit. on p. 78).

[91] L. van der Maaten and G. Hinton. 'Visualizing Data using t-SNE'. In: *JMLR* (2008) (cit. on p. 90).

[92] T. Matsukawa, T. Okabe, E. Suzuki and Y. Sato. 'Hierarchical Gaussian Descriptor for Person Re-identification'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1363–1372 (cit. on p. 14).

[93] N. McLaughlin, J. M. D. Rincón and P. Miller. 'Recurrent Convolutional Network for Video-Based Person Re-identification'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1325–1334 (cit. on p. 16).

[94] A. Mignon and F. Jurie. 'PCCA: A new approach for distance learning from sparse pairwise constraints'. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 2666–2672 (cit. on p. 28).

[95] M. Naphade, Z. Tang, M.-C. Chang, D. C. Anastasiu, A. Sharma, R. Chellappa et al. 'The 2019 AI City Challenge.' In: *CVPR Workshops*. Vol. 8. 2019 (cit. on p. 27).

[96] A. van den Oord, Y. Li and O. Vinyals. 'Representation Learning with Contrastive Predictive Coding'. In: *ArXiv* abs/1807.03748 (2018) (cit. on pp. 19, 60, 64, 80).

[97] X. Pan, P. Luo, J. Shi and X. Tang. 'Two at once: Enhancing learning and generalization capacities via ibn-net'. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 464–479 (cit. on pp. 52, 53, 87, 88).

[98] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan et al. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *NeurIPS*. 2019 (cit. on p. 52).

[99] N. Pu, W. Chen, Y. Liu, E. M. Bakker and M. S. Lew. 'Lifelong Person Re-Identification via Adaptive Knowledge Accumulation'. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 (cit. on p. 98).

[100] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu et al. 'Pose-normalized image generation for person re-identification'. In: *ECCV*. 2018 (cit. on p. 22).

[101] S. Qiao, W. Shen, Z. Zhang, B. Wang and A. Yuille. 'Deep co-training for semi-supervised image recognition'. In: *Proceedings of the european conference on computer vision (eccv)*. 2018, pp. 135–152 (cit. on p. 45).

[102] R. Quan, X. Dong, Y. Wu, L. Zhu and Y. Yang. 'Auto-ReID: Searching for a Part-Aware ConvNet for Person Re-Identification'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 3749–3758 (cit. on p. 17).

[103] A. Radford, L. Metz and S. Chintala. 'Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks'. In: *CoRR* abs/1511.06434 (2016) (cit. on p. 22).

[104] S. J. Reddi, S. Kale and S. Kumar. 'On the Convergence of Adam and Beyond'. In: *International Conference on Learning Representations*. 2018 (cit. on p. 35).

[105] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi. 'Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking'. In: *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*. 2016 (cit. on pp. 6, 11, 12, 36, 51, 65, 87).

[106] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma et al. 'ImageNet Large Scale Visual Recognition Challenge'. In: *IJCV* (2015) (cit. on pp. 66, 87).

[107] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra. 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization'. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 618–626 (cit. on pp. 30, 31).

[108] K. Simonyan and A. Zisserman. 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. In: *CoRR* abs/1409.1556 (2015) (cit. on p. 15).

[109] K. Simonyan and A. Zisserman. 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. In: *CoRR* abs/1409.1556 (2014) (cit. on p. 34).

[110] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk et al. 'FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence'. In: *NeurIPS*. 2020 (cit. on pp. 80, 86).

[111] J. Song, Y. Yang, Y.-Z. Song, T. Xiang and T. M. Hospedales. 'Generalizable Person Re-Identification by Domain-Invariant Mapping Network'. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 719–728 (cit. on p. 98).

[112] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang et al. 'Unsupervised domain adaptive re-identification: Theory and practice'. In: *Pattern Recognition* (2020) (cit. on pp. 45, 48, 52, 79).

[113] K. Soomro, A. Zamir and M. Shah. 'UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild'. In: *ArXiv* abs/1212.0402 (2012) (cit. on p. 14).

[114] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. 'Dropout: a simple way to prevent neural networks from overfitting'. In: *J. Mach. Learn. Res.* 15 (2014), pp. 1929–1958 (cit. on p. 22).

[115] *Stunning free images and royalty free stock*. https://pixabay.com/. Accessed: 2021-08-23 (cit. on p. 1).

[116] X. Sun and L. Zheng. 'Dissecting Person Re-Identification From the Viewpoint of Viewpoint'. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 608–617 (cit. on p. 11).

[117] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang et al. 'Circle Loss: A Unified Perspective of Pair Similarity Optimization'. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6397–6406 (cit. on p. 20).

[118] Y. Sun, L. Zheng, Y. Yang, Q. Tian and S. Wang. 'Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)'. In: *ECCV*. 2018 (cit. on pp. 21, 31, 34, 35, 39, 40, 42, 93).

[119] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi. 'Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning'. In: *AAAI*. 2017 (cit. on p. 17).

[120] A. Tarvainen and H. Valpola. 'Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results'. In: *NeurIPS*. 2017 (cit. on pp. 43, 45, 49, 80, 86).

[121] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani and M. Paluri. 'Learning Spatiotemporal Features with 3D Convolutional Networks'. In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 4489–4497 (cit. on p. 15).

[122] S. Tulyakov, M.-Y. Liu, X. Yang and J. Kautz. 'MoCoGAN: Decomposing motion and content for video generation'. In: *CVPR*. 2018 (cit. on p. 61).

[123] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al. 'Attention is All you Need'. In: *ArXiv* abs/1706.03762 (2017) (cit. on pp. 16, 17).

[124] C. Wang, Q. Zhang, C. Huang, W. Liu and X. Wang. 'Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification'. In: *ECCV*. 2018 (cit. on p. 39).

[125] D. Wang and S. Zhang. 'Unsupervised Person Re-Identification via Multi-Label Classification'. In: *CVPR*. 2020 (cit. on pp. 24, 59, 61, 66–68, 79, 91, 93).

[126] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang et al. 'Residual Attention Network for Image Classification'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 6450–6458 (cit. on pp. 30, 32).

[127] G. Wang, G. Wang, X. Zhang, J. Lai, Z. Yu and L. Lin. 'Weakly Supervised Person Re-ID: Differentiable Graphical Learning and A New Benchmark'. In: *IEEE Transactions on Neural Networks and Learning Systems (T-NNLS)*. 2020 (cit. on p. 6).

[128] G. Wang, Y. Yuan, X. Chen, J. Li and X. Zhou. 'Learning Discriminative Features with Multiple Granularities for Person Re-Identification'. In: *ACM Multimedia*. 2018 (cit. on pp. 20, 21, 28, 31, 34, 35, 39, 40).

[129] J. Wang, X. Zhu, S. Gong and W. Li. 'Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification'. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 2275–2284 (cit. on pp. 24, 28, 40, 41, 44, 61, 79).

[130] M. Wang, B. Lai, J. Huang, X. Gong and X.-S. Hua. 'Camera-aware Proxies for Unsupervised Person Re-Identification'. In: *AAAI*. 2021 (cit. on pp. 25, 80, 82–84, 88, 93, 94).

[131] T. Wang, S. Gong, X. Zhu and S. Wang. 'Person Re-Identification by Discriminative Selection in Video Ranking'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (2016), pp. 2501–2514 (cit. on p. 12).

[132] Y. Wang, P. Bilinski, F. Bremond and A. Dantcheva. 'G3AN: Disentangling Appearance and Motion for Video Generation'. In: *CVPR*. 2020 (cit. on p. 61).

[133] Y. Wang, P. Bilinski, F. Bremond and A. Dantcheva. 'ImaGINator: Conditional Spatio-Temporal GAN for Video Generation'. In: *WACV*. 2020 (cit. on p. 61).

[134] Y. Wang, F. Bremond and A. Dantcheva. 'InMoDeGAN: Interpretable Motion Decomposition Generative Adversarial Network for Video Generation'. In: *arXiv preprint arXiv:2101.03049* (2021) (cit. on p. 61).

[135] Z. Wang, J. Zhang, L. Zheng, Y. Liu, Y. Sun, Y. Li et al. 'CycAs: Self-supervised Cycle Association for Learning Re-identifiable Descriptions'. In: *ECCV*. Ed. by A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm. 2020 (cit. on pp. 80, 93).

[136] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. 'Image quality assessment: from error visibility to structural similarity'. In: *TIP* (2004) (cit. on p. 73).

[137] C. Wei, H. Wang, W. Shen and A. Yuille. 'CO2: Consistent Contrast for Unsupervised Visual Representation Learning'. In: *ICLR*. 2021 (cit. on pp. 80, 86).

[138] L. Wei, S. Zhang, W. Gao and Q. Tian. 'Person transfer gan to bridge domain gap for person re-identification'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 79–88 (cit. on pp. 11, 22, 23, 44, 51, 58, 61, 65, 79, 87).

[139] G. Wu, X. Zhu and S. Gong. 'Tracklet Self-Supervised Learning for Unsupervised Person Re-Identification'. In: *AAAI*. 2020 (cit. on pp. 46, 53).

[140] G. Wu, X. Zhu and S. Gong. 'Tracklet Self-Supervised Learning for Unsupervised Person Re-Identification.' In: *AAAI*. 2020 (cit. on pp. 24, 61, 66).

[141] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian and Y. Yang. 'Progressive Learning for Person Re-Identification with One Example'. In: *TIP* (2019) (cit. on p. 61).

[142] Y. Wu and K. He. 'Group Normalization'. In: *The European Conference on Computer Vision (ECCV)*. 2018 (cit. on p. 31).

[143] Z. Wu, Y. Xiong, S. X. Yu and D. Lin. 'Unsupervised Feature Learning via Non-parametric Instance Discrimination'. In: *CVPR*. 2018 (cit. on pp. 60, 64, 78, 80).

[144] T. Xiao, H. Li, W. Ouyang and X. Wang. 'Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1249–1258 (cit. on pp. 22, 30).

[145] S. Xie, R. B. Girshick, P. Dollár, Z. Tu and K. He. 'Aggregated Residual Transformations for Deep Neural Networks'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 5987–5995 (cit. on pp. 15, 31).

[146] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng et al. 'Asymmetric Co-Teaching for Unsupervised Cross Domain Person Re-Identification'. In: (2020) (cit. on pp. 43–45, 48, 52, 54).

[147] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng et al. 'Asymmetric Co-Teaching for Unsupervised Cross-Domain Person Re-Identification.' In: *AAAI*. 2020 (cit. on pp. 24, 61, 66, 68).

[148] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang and S. Zhang. 'Towards Rich Feature Discovery With Class Activation Maps Augmentation for Person Re-Identification'. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on pp. 39, 40).

[149] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi and S. Li. 'Salient Color Names for Person Re-identification'. In: *ECCV*. 2014 (cit. on pp. 14, 21).

[150] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu and Q. Tian. 'Deep representation learning with part loss for person re-identification'. In: *IEEE Transactions on Image Processing* (2019) (cit. on p. 30).

[151] D. Yi, Z. Lei, S. Liao and S. Li. 'Deep Metric Learning for Person Re-identification'. In: *2014 22nd International Conference on Pattern Recognition* (2014), pp. 34–39 (cit. on p. 21).

[152] *YouTube at 15: My personal journey and the road ahead.* https://blog.youtube/news-and-events/youtube-at-15-my-personal-journey. Accessed: 2021-08-23 (cit. on p. 1).

[153] H.-X. Yu, W. Zheng, A. Wu, X. Guo, S. Gong and J. Lai. 'Unsupervised Person Re-Identification by Soft Multilabel Learning'. In: *CVPR* (2019) (cit. on pp. 24, 79).

[154] X. Yu, B. Han, J. Yao, G. Niu, I. W.-H. Tsang and M. Sugiyama. 'How does Disagreement Help Generalization against Label Corruption?' In: *ICML*. 2019 (cit. on p. 44).

[155] K. Zeng, M. Ning, Y. Wang and Y. Guo. 'Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification'. In: *CVPR*. 2020 (cit. on pp. 79, 93).

[156] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji and Y. Tian. 'Multiple Expert Brainstorming for Domain Adaptive Person Re-identification'. In: *ECCV*. 2020 (cit. on pp. 91, 93).

[157] T. Zhang, L. Xie, L. Wei, Z. Zhuang, Y. Zhang, B. Li et al. 'UnrealPerson: An Adaptive Pipeline Towards Costless Person Re-Identification'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11506–11515 (cit. on p. 11).

[158] X. Zhang, J. Cao, C. Shen and M. You. 'Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 8221–8230 (cit. on pp. 43, 45, 52).

[159] X. Zhang, J. Cao, C. Shen and M. You. 'Self-training with progressive augmentation for unsupervised cross-domain person re-identification'. In: *ICCV*. 2019 (cit. on p. 79).

[160] R. Zhao, W. Ouyang and X. Wang. 'Person Re-identification by Salience Matching'. In: *2013 IEEE International Conference on Computer Vision* (2013), pp. 2528–2535 (cit. on p. 13).

[161] R. Zhao, W. Ouyang and X. Wang. 'Unsupervised Salience Learning for Person Re-identification'. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3586–3593 (cit. on pp. 13, 21).

[162] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu et al. 'Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training'. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on pp. 31, 34, 39, 40).

[163] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang et al. 'MARS: A Video Benchmark for Large-Scale Person Re-Identification'. In: *ECCV 2016*. 2016 (cit. on pp. 12, 36, 40).

[164] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian. 'Scalable Person Re-identification: A Benchmark'. In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1116–1124 (cit. on pp. 6, 11, 21, 27, 36, 51, 65, 87).

[165] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang and J. Kautz. 'Joint discriminative and generative learning for person re-identification'. In: *CVPR*. 2019 (cit. on pp. 20, 22, 61, 66, 69, 70, 73, 93).

[166] Z. Zheng and Y. Yang. 'Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation'. In: *IJCV* (2021) (cit. on p. 80).

[167] Z. Zheng, L. Zheng and Y. Yang. 'A Discriminatively Learned CNN Embedding for Person Reidentification'. In: *TOMCCAP* 14 (2017), 13:1–13:20 (cit. on pp. 27, 30).

[168] Z. Zheng, L. Zheng and Y. Yang. 'Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro'. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 3774–3782 (cit. on p. 36).

[169] Z. Zheng, L. Zheng and Y. Yang. 'Unlabeled samples generated by gan improve the person re-identification baseline in vitro'. In: *ICCV*. 2017 (cit. on p. 22).

[170] Z. Zhong, L. Zheng, D. Cao and S. Li. 'Re-ranking Person Re-identification with k-Reciprocal Encoding'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 3652–3661 (cit. on pp. 36, 37, 40, 45, 48, 53, 88).

[171] Z. Zhong, L. Zheng, G. Kang, S. Li and Y. Yang. 'Random Erasing Data Augmentation'. In: *AAAI*. 2020 (cit. on pp. 20, 52, 61, 88).

[172] Z. Zhong, L. Zheng, S. Li and Y. Yang. 'Generalizing A Person Retrieval Model Hetero- and Homogeneously'. In: *The European Conference on Computer Vision (ECCV)*. 2018 (cit. on pp. 22, 23, 40, 41, 44, 52, 58, 61, 79).

[173] Z. Zhong, L. Zheng, Z. Luo, S. Li and Y. Yang. 'Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification'. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on pp. 24, 45, 52, 61, 66, 67, 79).

[174] Z. Zhong, L. Zheng, Z. Luo, S. Li and Y. Yang. 'Learning to adapt invariance in memory for person re-identification'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020) (cit. on pp. 24, 52, 61, 66, 67, 91, 93).

[175] Z. Zhong, L. Zheng, Z. Zheng, S. Li and Y. Yang. 'Camera style adaptation for person re-identification'. In: *CVPR*. 2018 (cit. on pp. 22, 61).

[176] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang. 'Omni-Scale Feature Learning for Person Re-Identification'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 3701–3711 (cit. on p. 17).

[177] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros. 'Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks'. In: *ICCV*. 2017 (cit. on pp. 61, 63).

[178] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros. 'Unpaired image-to-image translation using cycle-consistent adversarial networks'. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232 (cit. on p. 22).

[179] Y. Zou, X. Yang, Z. Yu, B. V. K. V. Kumar and J. Kautz. 'Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification'. In: *ECCV*. 2020 (cit. on pp. 20, 22–24, 58, 61, 66, 67, 73, 79, 91, 93).