Journal of New Media
DOI:10.32604/jnm.2021.023981
Review

Review of Unsupervised Person Re-Identification

Yang Dai* and Zhiyuan Luo

School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Yang Dai. Email: 20191220014@nuist.edu.cn
Received: 30 September 2021; Accepted: 15 October 2021

Abstract: Person re-identification (re-ID) aims to match images of the same pedestrian across different cameras. It plays an important role in the field of security and surveillance. Although it has been studied for many years, it is still considered as an unsolved problem. Since the rise of deep learning, the accuracy of supervised person re-ID on public datasets has reached the highest level. However, these methods are difficult to apply to real-life scenarios because a large number of labeled training data is required in this situation. Pedestrian identity labeling, especially cross-camera pedestrian identity labeling, is heavy and expensive. Why we cannot apply the pre-trained model directly to the unseen camera network? Due to the existence of domain bias between source and target environment, the accuracy on target dataset is always low. For example, the model trained on the mall needs to adapt to the new environment of airport obviously. Recently, some researches have been proposed to solve this problem, including clustering-based methods, GAN-based methods, co-training methods and unsupervised domain adaptation methods.

Keywords: Unsupervised person re-identification; review; deep learning

1 Introduction

With the rapid development of modernization, many rural populations have poured into the cities, and the pressure on public security is increasing. In order to prevent and deal with such incidents in time, the government has installed a large number of surveillance cameras in public places. However, human cannot handle such massive amounts of video data that generated by the surveillance network. Almost all surveillance systems rely on computer vision related technologies. At present, face recognition has been widely used in various fields of life, but in many scenarios, the cameras usually cannot capture high-resolution face images, but lower-resolution holistic person images or partial pedestrian images which caused by occlusion are always collected, where face recognition is useless. The re-identification (re-ID) technology of how to identify the identity based on pedestrian images has gradually become a research hotspot and has attracted widespread attention. The purpose of person re-ID is to quickly identify a person of interest in the view of multiple uncrossed surveillance cameras. Although it has been researched for many years, it is still considered as an open problem.

Today more and more researchers try to solve person re-ID problem based on deep learning which has become the core technology of artificial intelligence. Despite impressive progress has been achieved by the deep learning methods, especially these methods based on convolutional neural network (CNN) under supervised setting, but they cannot apply to real-life scenarios because a large number of labelled image pairs are needed in the training stage which is hard to obtain in reality. Hence many researchers start to explore unsupervised re-ID methods, existing solutions can be divided into four fine-grained categories, including clustering-based pseudo labels generation [1–3], GAN-based style transformation [4–8], co-training [9–12] and unsupervised domain adaptation [13–16].

2 Clustering-Based Pseudo Labels Generation

Generally, clustering-based methods is the simplest and most efficient solution. It can be adopted in the following three-stage training scheme: (1) Pre-training on the source domain with pedestrian identities, (2) Inferencing on target domain and clustering with the extracted features, and (3) Training on the target domain which is called fine-tuning with pseudo person labels generated by clustering. Fan et al. propose PUL [1] to progressively improve the discrimination power of model with a selection operation between the clustering and fine-tuning. The goal of selection operation is to select high-confidence samples that belong to the same class. At the beginning of training, when the model is weak, only a small amount of reliable samples are selected that locate closely to the centroids in the feature space, and more samples are selected for training with iterations increase when model becomes stronger. Due to the large differences between holistic person images, it is easy to divide the same person images coming from various cameras into different clusters which can harm the model severely, therefore, Yang et al. propose SSG [2] to adopt the body part based clustering strategy where one person image is split into three parts, including the upper part, the lower part and the whole body, and develop three branches based on the CNN backbone to group different person body parts into clusters, respectively. You can see the progress in Fig. 1.

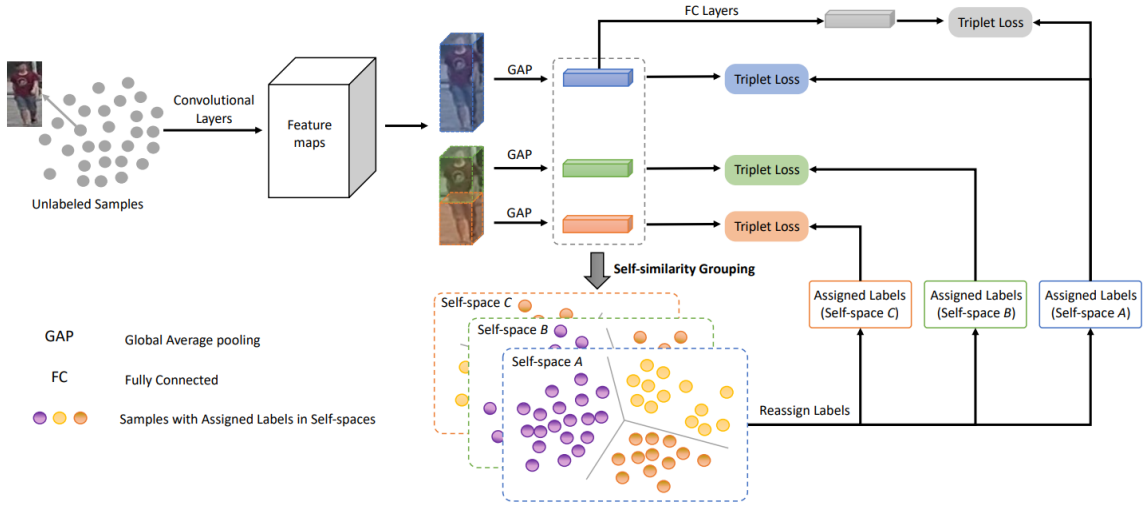

Figure 1: The architecture of SSG [2]

Inspired by label smoothing regularization, some newest clustering-guided methods adopt multi-label classification instead of traditional single-label classification to enhance robustness of model. Lin et al. propose SSL [3] to discuss the similarity relationship between unlabeled images and optimize the model with multi-label information which is evaluated by similarity estimation. As shown in Fig. 2, the model is initialized with hard-label classification where each image is seen as one particular class, then evaluate a set of samples that have similar appearance with the target image based on distance in target feature space, also a regularization called CCE is induced to penalty the discrepancy of samples between different cameras to find more cross-camera positive pairs, finally, the model is fine-tuned by a multi-label classification loss which has the similar formulation with label smoothing loss.

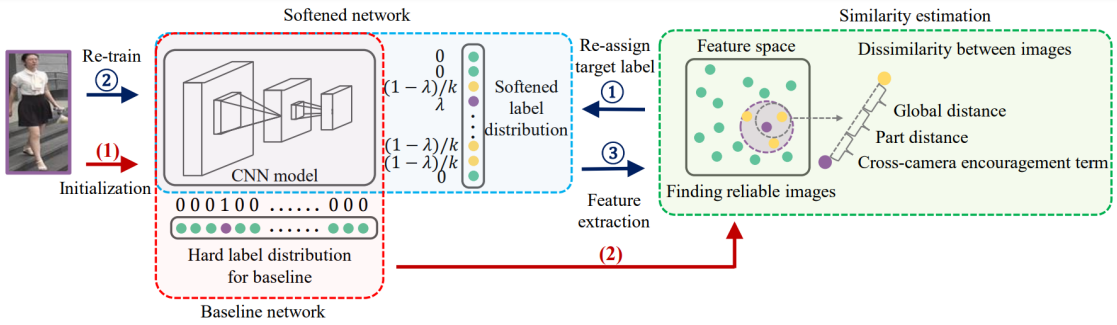

Figure 2: The architecture of SSL [3]

3 GAN-Based Style Transformation

The image-to-image style transformation typical of CycleGAN [5] has been a very important technology in person re-ID, it aims to learn a pair of mapping between two domains in the absence of labelled training image pairs. By this technology, we can convert one image from domain A to domain B that are indistinguishable in data distribution. Applied into unsupervised person re-ID, we can alleviate domain bias between source and target domain, thus we can use source-target translated images with target domain style that preserve the same identities in source domain to train the model. You can even apply this model directly to real target dataset and get a good result. Deng et al. [6] propose SPGAN to integrate an ideal that the translated image should have the same identity with source image which is called self-similarity and the translated image should have different identity with target image which is called domain-dissimilarity into the traditional CycleGAN. Ge et al. [7] propose SDA to generate source-target translated images by CycleGAN and preserve the inter-sample relations as in source domain. As shown in Fig. 3, the architecture of SDA is consisted of a domain translation (DT) which is made of a pair of generations and a pair of adversarial discriminators in two directions as in the original CycleGAN, a source-domain encoder (SE) and a target-domain encoder (TE). Specifically, SE is pre-trained with the source-domain data to provide structure information as supervision signal to guide the training of DT and TE. TE which has the same components with SE is designed to encode target-domain data and calculate the distances of them in embedded feature space which should be consistent with the distances calculated in source-domain embedded space. Thus a better target-domain style images with identities can be generated to further train the TE module that be adaptable to original target-domain data.

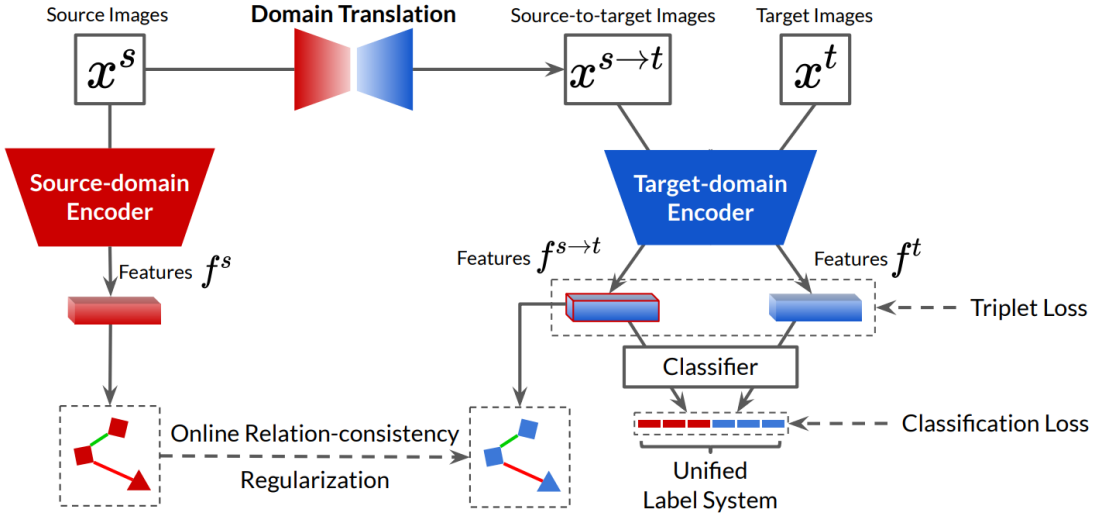

Figure 3: The architecture of SDA [7]

Generally, you do not even need to improve the performance of CycleGAN, just apply it in our task. Zhong et al. [8] propose to learn discriminative power in target domain from three perspectives, exemplar-invariance, camera-invariance and neighborhood-invariance. Exemplar-invariance is same as optimization goal of baseline network in [3]. Camera-invariance tries to generate other camera style images for each target-domain training image and prompt the network to study the invariance between them who has the same identity. Neighborhood-invariance is similar to the clustering-based methods, it aims to find the k-nearest neighbours for each target image who has the analogous visual appearance that is thought to be the same person. You can see the architecture of the proposed method in Fig. 4.

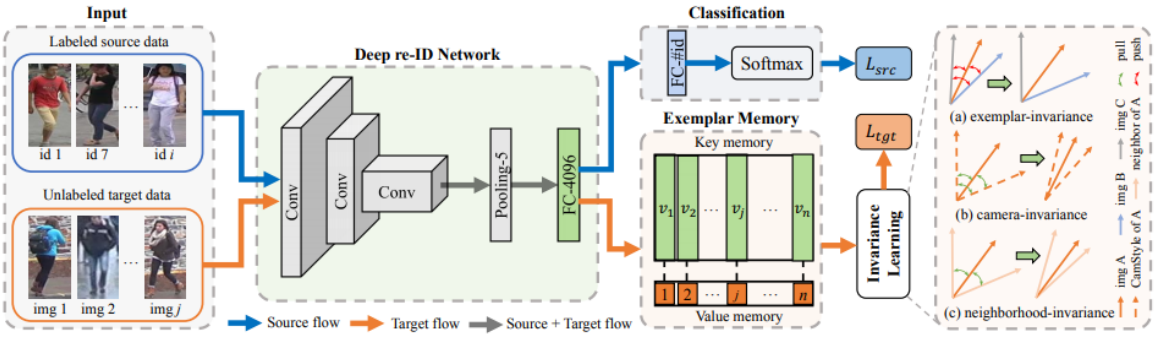

Figure 4: The architecture of the proposed method in [8]

4 Co-Training

Co-training represents the latest progress in unsupervised person re-ID, the typical case is deep mutual learning and teacher-student network. In fact, the earliest method based on the idea of co-training is proposed by Geng et al. [9] in which a CNN model and a dictionary learning model is listed simultaneously and trained iteratively to provide person coding for each other. Recently, co-training is mined again and applied in clustering-based pseudo label generation method to suppress the label noise which can harm the model severely. Tang et al. [10] propose an extremely simple way to reduce the impact of noisy labels. Specifically, as shown in Fig. 5, it uses resnet50 as backbone and replaces the original output layer with a GAP layer and a FC layer subsequently, then clustering is applied on the last two layers and generates two sets of pseudo labels, each of which is utilized for two layers output to optimize the network. Note that the pseudo labels generated by GAP is different from that generated by FC, errors in each set of labels can be neutralized by another one.

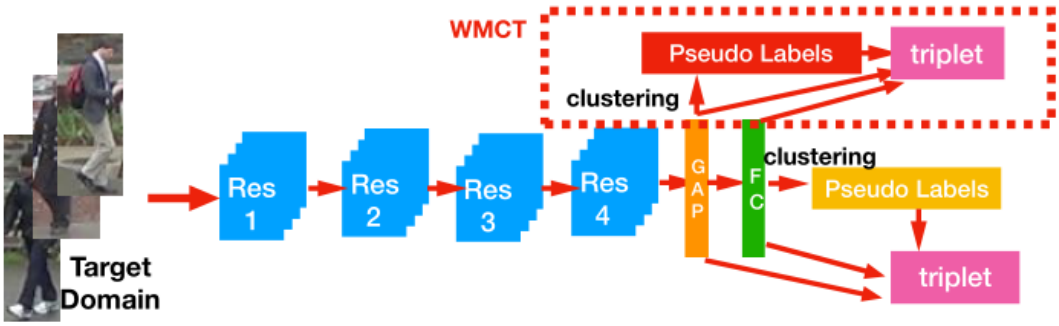

Figure 5: The architecture of the proposed method in [10]

Inspired by deep mutual learning (DML) [17], Ge et al. [11] propose MMT in which two collaborative networks that have the same architecture but with different initialization are designed to generate refined soft pseudo labels distribution for each other to go hand in hand. As illustrated in Fig. 6, MMT also innovatively proposes temporally average model which is called mean net to generate soft labels distribution for promising the stability of model and further inhibiting the impact of label noise. In

addition, the input of two collaborative networks are the same but flipped images, it not only achieves the goal of image enhancement, but also avoids the network from performing completely consistent learning.

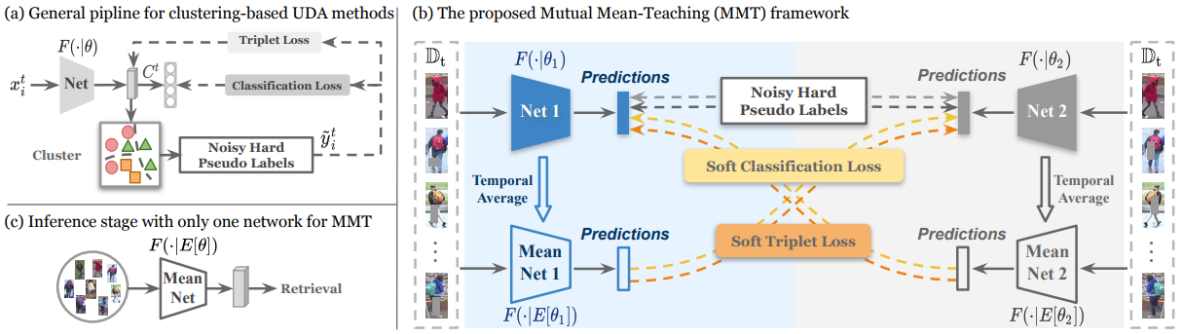

Figure 6: The architecture of MMT [11]

Different from other papers that all use two peer-to-peer networks with the same architecture for mutual learning, Zhai et al. [12] propose MEB-net assembled with at least three expert networks that constructed of diverse architectures to make brainstorming among each two experts. As shown in Fig. 7, firstly, the expert networks are pre-trained with source domain labelled images respectively, then enter the second stage of brainstorming which can be divided into two sub-steps, (1) One pedestrian mean vector can be obtained with the features that extracted by all expert networks for each target training image based on which a set of pseudo labels can be generated as supervision signal to train all expert networks in turn. Because all expert networks learn from one set of pseudo labels, therefore, the predicted probabilities provided by different experts will tend to be consistent. On the one hand, this may reduce the independence between experts and cause errors to be transmitted and amplified among them. On the other hand, experts cannot learn effectively from the same indistinguishable information, consequently, in order to keep the original knowledge of expert networks in the soft label information as much as possible that the experts can better exchange opinions, (2) similar to [11], MEB-net also introduces mean net which can keep history knowledge of experts well that may be different from each other.

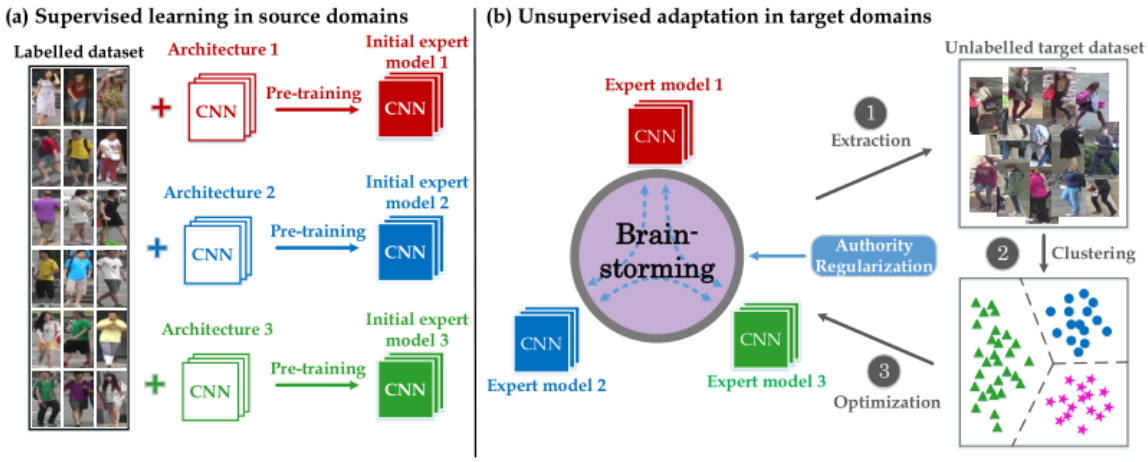

Figure 7: The architecture of MEB-net [12]

5 Unsupervised Domain Adaptation

Li et al. [13] propose ARN to decompose the pedestrian features of different datasets into domain-invariant features and domain-specific features. As shown in Fig. 8, the architecture of ARN is consisted of encoder in which E_T, E_C and E_S are designed to extract target-domain specific features, domain-invariant features and source-domain specific features respectively, and decoder in which target-domain

specific features and domain-invariant features are synthesized as original target-domain features and source-domain specific features and domain-invariant features are synthesized as original source-domain features. The supervised information from source-domain is used to train the sub-network E_C for extracting domain invariant features to learn human discrimination ability with identity classification loss. Otherwise, the sub-network E_S and E_T are trained with reconstruction loss based on orthogonality decomposition. In the testing stage, target-domain invariant features are inferenced to do matching.

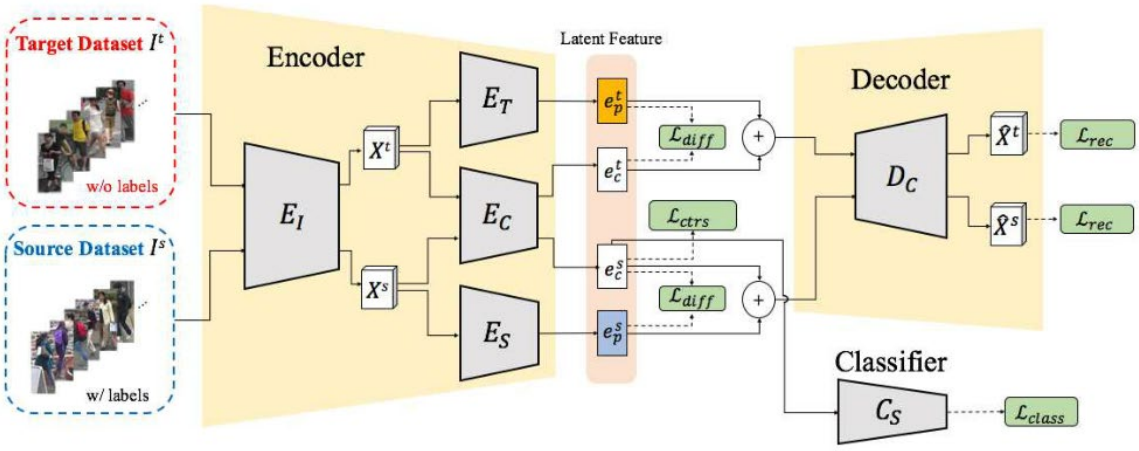

Figure 8: The architecture of ARN [13]

In addition to the above-mentioned reconstruction-based domain adaptation method, another classic method [15] for learning domain-invariant features that domain gap between features has been alleviated is based on the idea of adversarial. Specifically, a discriminator is added into the conventional person re-ID backbone of which the goal is to distinguish which domain the input feature comes from. When the output of backbone, i.e., feature extractor successfully confuse the discriminator, it means that the features extracted from source or target domain have the same distribution. Also from the angel of data distribution, Mekhazni et al. [14] propose D-MMD which is to minimize the MMD loss in dissimilarity space. Note that MMD is used to measure the distance between two data distribution. Now what is dissimilarity space? It involves two concepts, feature representation and dissimilarity representation, the former is the person feature extracted by the network, and the latter is the distance (i.e., difference or dissimilarity) between two person features, because the two person features may be of the same class or not, so the dissimilarity representation can also be divided into two types, i.e., within-class (WC) dissimilarity representation and between-class (BC) dissimilarity representation. Accordingly, the dissimilarity space can be divided by WC dissimilarity space and BC dissimilarity space and we need to align feature distribution (i.e., minimize MMD loss) in these two spaces separately.

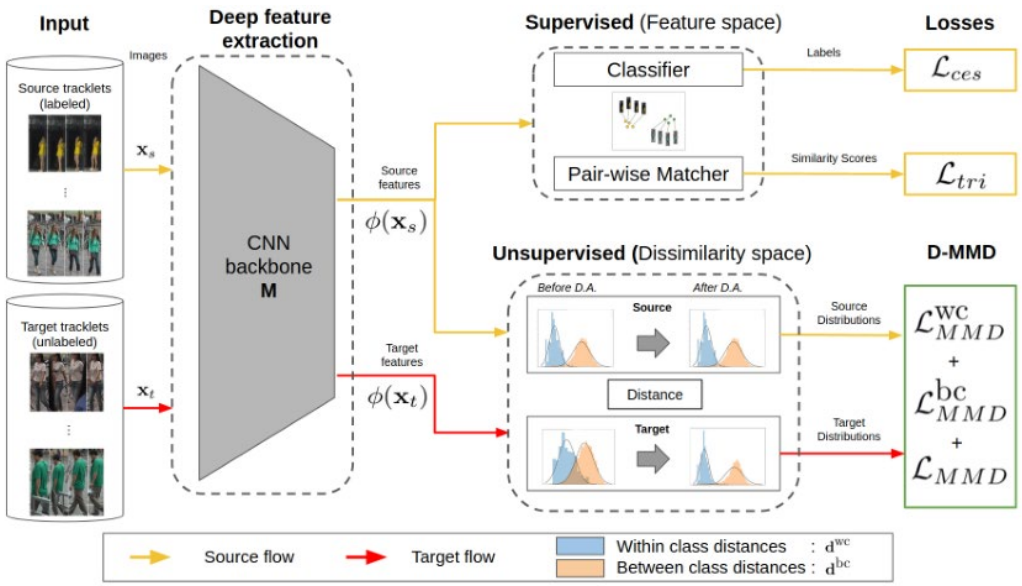

Figure 9: The architecture of D-MMD [14]

Wu et al. [16] find that the intra-camera pairwise similarity distribution (ICPSD) is inconsistent with cross-camera pairwise similarity distribution (CCPSD) that the cross-camera matching is always behind intra-camera matching in the ranking list because the huge discrepancy between different cameras. This problem is called camera-aware similarity inconsistency problem. The author tries to minimize the difference of the mean and the variance between ICPSD and CCPSD. At the same time, the intra-camera similarity is preserved as the benchmark after camera-aware similarity consistency learning. The whole architecture can be found in Fig. 10.

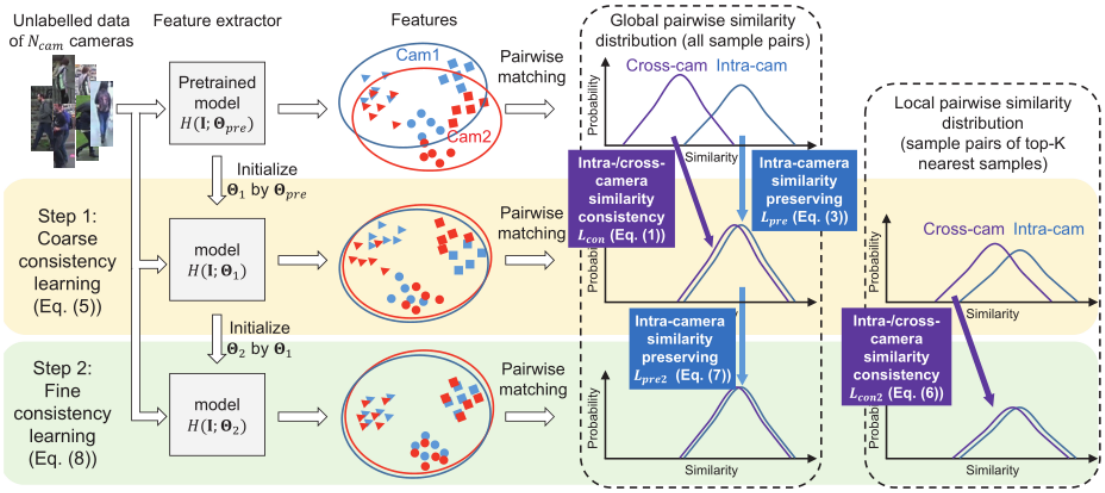

Figure 10: The architecture of the proposed method in [16]

6 Conclusion

Unsupervised person re-ID is a crucial problem where labelled pedestrian images for target-domain is not required that can truly apply into real-life. In this review, we list most of the current research methods, the difficulty of clustering-based methods lies in that the impact of pseudo label noise cannot be alleviated completely, the difficulty of GAN-based generation and other domain adaptation methods lie in that the performance improvement brought by the distribution fitting is limited in the absence of labels. There is still a lot of work worth researching in these aspects mentioned above.

Acknowledgement: I would like to show my deepest gratitude to my teacher. In addition, my abilities are limited, any suggestions would be appreciated if anyone can point out my shortcomings of the review for unsupervised person re-ID.

Funding Statement: The authors did not receive any specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. H. Fan, L. Zheng and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing*, vol. 14, no. 4, pp. 1–18, 2018.

[2] Y. Fu, Y. C. Wei, G. S. Wang, Y. Q. Zhou, H. H. Shi *et al.,* "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. CVPR,* pp. 6112–6121, 2019.

[3] Y. T. Lin, L. X. Xie, Y. Wu, C. G. Yan and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. CVPR,* pp. 3390–3399, 2020.

[4] Y. P. Zhai, S. J. Lu, Q. X. Ye, X. B. Shan, J. Chen *et al.,* "Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification," in *Proc. CVPR,* pp. 9021–9030, 2020.

[5] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV,* pp. 2223–2232, 2017.

[6] W. J. Deng, L. Zheng, Q. X. Ye, G. L. Kang, Y. Yang *et al.,* "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. CVPR,* pp. 994–1003, 2018.

[7] Y. X. Ge, F. Zhu, R. Zhao and H. S. Li, "Structured domain adaptation for unsupervised person re-identification," arXiv preprint arXiv: 2003.06650, 2020.

[8] Z. Zhong, L, Zheng, Z. M. Luo, S. Z. Li and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. CVPR,* pp. 598–607, 2019.

[9] M. Y. Geng, Y. W. Wang, T. Xiang and Y. H. Tian, "Deep transfer learning for person re-identification," arXiv preprint arXiv: 1611.05244, 2016.

[10] H. T. Tang, Y. R. Zhao and H. T. Lu. "Unsupervised person re-identification with iterative self-supervised domain adaptation," in *Proc. CVPR,* 2019.

[11] Y. X. Ge, D. P. Chen and H. S. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," arXiv preprint arXiv: 2001.01526, 2020.

[12] Y. P. Zhai, Q. X. Ye, S. J. Lu, M. X. Jia, R. R. Ji *et al.,* "Multiple expert brainstorming for domain adaptive person re-identification," in *Proc. ECCV,* pp. 594–611, 2020.

[13] Y. J. Li, F. E. Yang, Y. C. Liu, Y. Y. Yeh, X. F. Du *et al.,* "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification," in *Proc. CVPR,* pp. 172–178, 2018.

[14] D. Mekhazni, A. Bhuiyan, G. Ekladious and E. Granger, "Unsupervised domain adaptation in the dissimilarity space for person re-identification," in *Proc. ECCV,* pp. 159–174, 2020.

[15] X. B. Liu and S. L. Zhang, "Domain adaptive person re-identification via coupling optimization," in *Proc. ACM MM,* pp. 547–555, 2020.

[16] A. Wu, W. S. Zheng and J. H. Lai, "Unsupervised person re-identification by camera-aware similarity consistency learning," in *Proc. ICCV,* pp. 6922–6931, 2019.

[17] Y. Zhang, T. Xiang, T. M. Hospedales and H. C. Lu, "Deep mutual learning," in *Proc. CVPR,* pp. 4320–4328, 2018.