Machine Learning (2023) 112:1923–1940
https://doi.org/10.1007/s10994-022-06184-x

Learning from self-discrepancy via multiple co-teaching for cross-domain person re-identification

Suncheng Xiang[1] · Yuzhuo Fu[1] · Mengyuan Guan[1] · Ting Liu[1]

Received: 14 September 2021 / Revised: 7 April 2022 / Accepted: 9 May 2022 /
Published online: 14 June 2022

Abstract

Employing clustering strategy to assign unlabeled target images with pseudo labels has become a trend for person re-identification (re-ID) algorithms in domain adaptation. A potential limitation of these clustering-based methods is that they always tend to introduce noisy labels, which will undoubtedly hamper the performance of our re-ID system. To handle this limitation, an intuitive solution is to utilize collaborative training to purify the pseudo label quality. However, there exists a challenge that the complementarity of two networks, which inevitably share a high similarity, becomes weakened gradually as training process goes on; worse still, these approaches typically ignore to consider the self-discrepancy of intra-class relations. To address this issue, in this paper, we propose a multiple co-teaching framework for domain adaptive person re-ID, opening up a promising direction about self-discrepancy problem under unsupervised condition. On top of that, a mean-teaching mechanism is leveraged to enlarge the difference and discover more complementary features in target domain. Comprehensive experiments conducted on several large-scale datasets show that our method achieves competitive performance compared with the state-of-the-arts.

Keywords Re-identification · Self-discrepancy · Multiple co-teaching · Mean-teaching

Editors: Bo Han, Tongliang Liu, Quanming Yao, MingmingGong, Gang Niu, Ivor W. Tsang, Masashi Sugiyama.

✉ Suncheng Xiang
xiangsuncheng17@sjtu.edu.cn

Yuzhuo Fu
yzfu@sjtu.edu.cn

Mengyuan Guan
gemini.my@sjtu.edu.cn

Ting Liu
louisa_liu@sjtu.edu.cn

[1] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

1 Introduction

Given a query image, person re-identification (re-ID) aims to match the person-of-interest across multiple non-overlapped cameras distributed in different places. Encouraged by the remarkable success of deep learning methods and the availability of large-scale datasets, re-ID research community has achieved significant progress during the past few years (Zheng and Yang 2016; Ye et al. 2021). However, as for pedestrian images from an unseen domain, even with a large diversity of training data, person re-ID model generally experiences catastrophic performance drops because of the huge domain gaps or scene shifts, which cannot satisfy the need of application in real scenarios. To alleviate this problem, unsupervised domain adaptation (UDA) (Ganin & Lempitsky 2015; Xiang et al. 2020; Saito et al. 2018) is therefore proposed to employ the model trained on source dataset with identity labels to perform inference on the target domain. Nevertheless, it still remains an open research challenge in industry and academia due to the lack of identity annotations.

Currently, there are two main categories of UDA methods in re-ID community. The first category of image-level adaptation aims to eliminate the data distribution discrepancy across source and target domain, such as PTGAN (Wei et al. 2018) and SPGAN (Deng et al. 2018). Although these approaches achieve promising progress, their performance deeply relies on the images generation quality. The second category of clustering-based adaptation (Song et al. 2020; Fu et al. 2019; Fan et al. 2018) deploys clustering algorithm to generate pseudo-labels for unsupervised target images during training period. Unfortunately, their abilities are substantially hindered by the inevitable label noises caused by imperfect clustering algorithms. To alleviate this problem, some co-teaching based re-ID approaches (Yang et al. 2020; Ge et al. 2020; Zhao et al. 2020; Zhai et al. 2020 have been introduced for combating with noisy labels after clustering. Even though their optimal performance is often achieved by sub-network's discrimination ability, the self-discrepancy of intra-class relation (as shown in Fig. 1) in target domain still remains unexplored. So

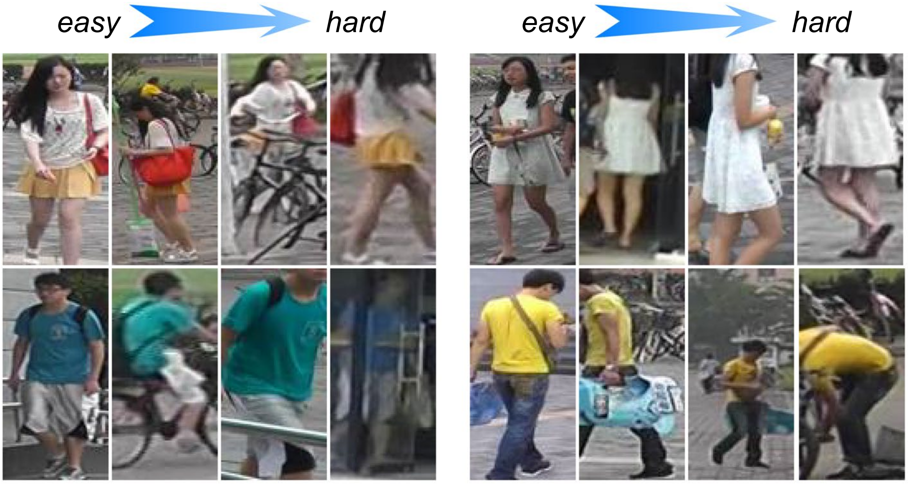


Fig. 1 Illustration of the self-discrepancy of intra-class relations for UDA person re-ID tasks, which is caused by variations in pose, viewpoint and occlusion, etc. For each identity, some ***easy*** samples can be assigned with reliable pseudo labels. However, most of ***hard*** samples are always given with noisy pseudo labels

a natural question then comes to our attention: *how to leverage self-discrepancy features of multiple sub-networks, and then optically adapt them to unlabelled domain?* which has to be fully elaborated. Another challenge we observe is that, as the training process goes on, two neural networks in traditional co-teaching (Han et al. 2018 tend to converge and unavoidably share a high similarity, which weakens their complementarity and further improvement in terms of performance.

To solve the challenges mentioned above, we propose a simple yet powerful **M**ultiple **C**o-teaching **N**etwork **MCN**[1] that considerably explores the self-discrepancy of intra-class relation in target domain, consequently, person re-ID can be more effectively performed to resist with noisy labels in domain adaptation. In addition, we introduce a mean-teaching mechanism to greatly enhance the complementarity and independence of collaborative networks, which, in turn, further improves the discriminability of learned representations in a progressive fashion. To the best of our knowledge, this is the first research effort to exploit the potential of *self-discrepancy* among intra-class relations to address the UDA problem. Compared with existing co-teaching based method (Ge et al. 2020; Zhao et al. 2020; Zhai et al. 2020), our MCN is different from them in terms of two perspective: **data input** and **model structure** **(1)** Our work proposes to adopt samples with different discrepancy granularity ($T_1 \sim T_n$) as asymmetric inputs to multiple networks, while previous methods applied same dataset as symmetric inputs during collaborative training; **(2)** MEB-Net (Zhai et al. 2020) used DenseNet-121 (Huang et al. 2017), ResNet-50 (He et al. 2016) and Inception-v3 (Szegedy et al. 2016) as backbone to collaboratively learn from each other during the same time, MMT (Ge et al. 2020) and NRMT (Zhao et al. 2020) utilized random erasing (Zhong et al. 2020) or random seeds for creating a difference, MMT (Ge et al. 2020) and MEB-Net (Zhai et al. 2020) also adopted symmetrical architecture with soft pseudo labels as well as hard pseudo labels in UDA re-ID tasks, which requires more computation costs and model parameters during training. In contrast, our MCN is only trained based on vanilla ResNet-50 backbone with hard pseudo labels, which makes it more flexible and adaptable. In addition, our proposed method can significantly mine the self-discrepancy features of samples in target domain, and a novel mean-teaching mechanism is also adopted to enhance the independence and complementarity between teacher network and several student networks, while previous asymmetric co-teaching approaches (*e.g.* ACT (Yang et al. 2020) and MEB-Net (Zhai et al. 2020)) fail to satisfy these needs in real-world application.

In summary, our main contributions of this paper can be summarized as follows:

- We propose a multiple co-teaching network MCN to mine the self-discrepancy of intra-class relations in target domain for solving noisy labels.
- A **M**ean-**T**eaching mechanism is introduced to further enhance the output complementarity in a progressive manner based on proposed MCN method ("MCN-MT" for short).
- Experimental results conducted on several benchmark datasets demonstrate the effectiveness of our proposed method.

In the rest of the paper. we first review some related works of unsupervised domain adaptive person re-ID in Sect. 2. Then in Sect. 3, we give the learning procedure of the

[1] The code of this work is publicly available at https://github.com/JeremyXSC/MCN-MT.

proposed multiple co-teaching network MCN, as well as its mean-teaching version MCN-MT. Extensive evaluations compared with state-of-the-art methods and comprehensive analyses of the proposed approach are reported in Sect. 4. Conclusion and future work are given in Sect. 5.

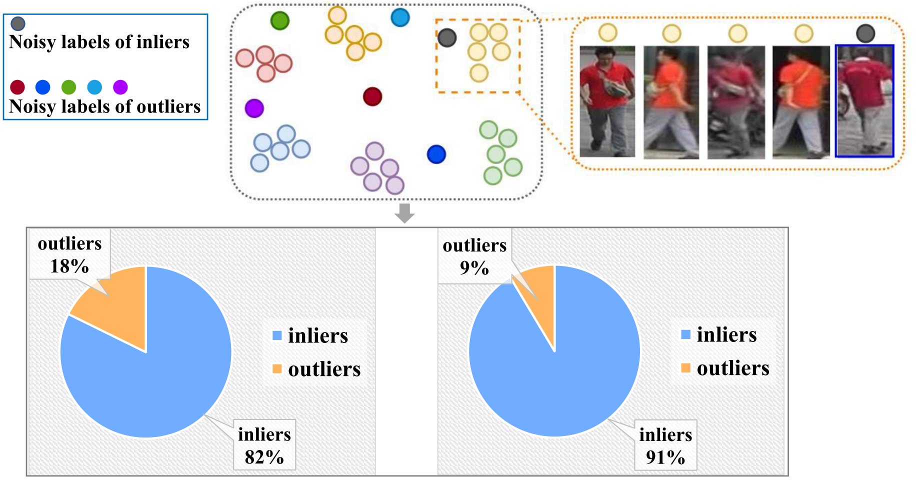
2 Related works

2.1 Unsupervised domain adaptation (UDA)

UDA aims to generalize the model learned from labeled source domain to the other unlabeled target domain, and the labeled and unlabeled examples are from non-overlapping classes, which makes it difficult to build the learning relationship between them (Xiang et al. 2020; Peng et al. 2016). Traditional studies (Zheng et al. 2015; Farenzena et al. 2010) related to hand-crafted systems for person re-ID aim to design or learn robust features for person re-ID, *e.g.*, (Farenzena et al. 2010) presented an appearance-based method with overall and local chromatic content. However, these hand-crafted feature based models always fail to produce competitive results on large-scale datasets. The main reason is that these early works are mostly based on heuristic design, and thus they could not learn optimal discriminative features on current large-scale dataset. Benefited from the success of deep learning, some recent works (Deng et al. 2018; Wei et al. 2018) attempt to address unsupervised domain adaptation based on deep learning framework. For instance, SPGAN (Deng et al. 2018) learns a similarity preserving GAN model by using the negative pairs to improve the image-image translation performance, Wei et al. (2018) proposed a Person Transfer GAN network to bridge the domain gap between different styles of two datasets and migrate pedestrian styles from one dataset to another. Recently, Zhong et al. (2018) firstly proposed a HHL method to learn camera-invariant networks for the target domain, Xiang et al. (2021) took a first attempt to explicitly dissect person re-ID from the aspect of attribute on synthetic datasets. Although these unsupervised domain adaptive approaches achieve promising progress, their performance is still very far from satisfactory compared with the fully supervised approaches (Xiang et al. 2020).

2.2 Clustering-based person re-ID

In recent years, training deep model with clustering has been widely studied, and these approaches mainly focus on estimating pseudo identity labels on the target domain so as to learn deep models in a supervised manner. Usually, clustering-based methods are used to generate a series of clusters in the feature space, so pseudo labels assigned by clustering are employed to update networks with an embedding loss. For example, Fan et al. (2018) proposes a progressive unsupervised learning method consisting of clustering and fine-tuning the network. Fu et al. (2019) proposes a self-similarity grouping method which exploits the potential similarity (from the global body to local parts) of unlabeled samples to build multiple clusters from different views automatically. In addition, a self-training augmentation method PAST (Zhang et al. 2019) is proposed to promote the performance on target dataset progressively. However, pseudo labels assigned by clustering algorithm can be very noisy as clustering accuracy on hard samples is not always satisfactory. To be more specific, there are two main noises for clustering-based method: **Noisy labels of inliers** and **Noisy labels of outliers**. As illustrated in the Fig. 2, as for the Noisy labels of inliers, the



(a) Proportion of noisy labels on Market **(b)** Proportion of noisy labels on Duke

Fig. 2 Some visualization results of noisy labels with traditional clustering method on Market-1501 and DukeMTMC-reID dataset, respectively

pseudo labels generated by clustering algorithm always contains some noisy labels due to its limited clustering performance, which will undoubtedly hamper the performance of our re-ID system. As for the Noisy labels of outliers, previous clustering methods tend to leave low confidence samples as outliers and do not assign cluster labels to them. These outliers are usually hard samples that encounter high image variations. Consequently, most of clustering-based methods suffer from the limitation of noisy clustering results and slow convergence since classification performance extremely depends on the quality of pseudo labels in target domain, which is not practical in real-world scenarios.

2.3 Learning with noisy labels

Deep learning with noisy labels is practically challenging, which has been widely studied in recent years. One of popular deep learning paradigm is co-teaching (Han et al. 2018). During the co-teaching, the key idea of teacher-student models is to create consistent training supervision for labeled or unlabeled data via different models' predictions, then allow teacher and student networks to teach other in mutual perspective. Recently, this idea has been applied to distill powerful and easy-to-train large networks into small but harder-to-train networks (Romero et al. 2014) that can even outperform their teacher. However, the output of teacher network and student networks might converge to equal each other and the two networks tend to loss their output independence when inputs of different branches share a high similarity or repeat with each others. Importantly, existing teacher-student models could not be directly utilized on unsupervised domain adaptation (UDA) tasks of person re-ID since they are mostly designed for close-set recognition problems, which hinders the further improvement of UDA person re-ID task. Aiming to address these challenges mentioned above, in this work, we develop a multiple co-teaching network with mean-teaching mechanism for unsupervised domain adaptive re-ID task.

Although MCN inherits (and extends) the structure of asymmetric co-teaching approach ACT (Yang et al. 2020), there exists some significant new designs in MCN to allow it to work for a very different manner. (1) MCN can significantly mine the self-discrepancy feature of intra-class relations in target domain, while ACT is mainly constructed to find more outliers to perform asymmetric co-teaching. (2) MCN seamlessly integrates the mean-teaching mechanism to enhance the independence and complementarity between teacher network and student networks. In contrast, ACT did not consider the complementarity problems of two networks during collaborative training. (3) MCN substantially outperforms ACT for unsupervised cross-domain re-ID tasks on several benchmark datasets. (4) This is the first time as far as we know, to comprehensively explore the self-discrepancy of intra-class relations in target domain for UDA re-ID task.

3 Methodology

3.1 Problem formulation

In the UDA re-ID task, we are given a labeled source dataset $S = \{x_1, x_2, \cdots, x_{N_s}\}$, consisting of N_s person images with manually annotated labels $Y = \{y_1, y_2, \cdots, y_{N_s}\}$. We also have an unlabeled target dataset $T = \{t_1, t_2, \cdots, t_M\}$. Note that there is non-overlapping in terms of identity between source domain and target domain in open set domain adaptation. Our goal is to learn a feature embedding function that can be applied to test set $X^t = \left\{x_1^t, x_2^t, \ldots x_{N_t}^t\right\}$ of N_t person images and query set $X^q = \left\{x_1^q, x_2^q, \ldots x_{N_q}^q\right\}$ of N_q person images during the evaluation stage. By leveraging both labeled source images from S and unlabeled target samples in T, we can learn a discriminative re-ID model that generalized well in the target domain for adaptive re-ID task.

3.2 Multiple co-teaching network

Without the supervised signal for cross-domain re-ID task, it is important to design a framework that can be used to mine the self-discrepancy features of target unlabeled samples. In this work, we creatively develop a multiple co-teaching network MCN to considerably explore the self-discrepancy of intra-class relations in target domain, and then combat with noisy labels in domain adaptation. Furthermore, a mean-teaching mechanism is introduced to enhance the complementarity and avoid error amplification of different collaborative networks.

3.2.1 Multiple co-teaching

To learn self-discrepancy of intra-class relations in target domain, we propose a multiple co-teaching network MCN which trains several networks progressively with samples at different granularity levels. As shown in Fig. 3, MCN consists one teacher network M_1 and several student networks $M_2 \sim M_n$. Specifically, we firstly train CNN on the source labeled data and fine-tune it on target data with pseudo labels to get initial weights for $M_1 \sim M_n$, then we perform multiple co-teaching paradigms between ***teacher*** network and several ***student*** networks. In particular, teacher network receives highest confidence samples (T_1) as much as possible while student networks take in samples with lower confidence levels (T_n, T_{n-1}, $\cdots$,T_2) as diverse as possible. In the first paradigm of co-teaching, ***teacher*** network M_1 performs co-teaching with ***student*** network M_n until they reach convergence, followed by

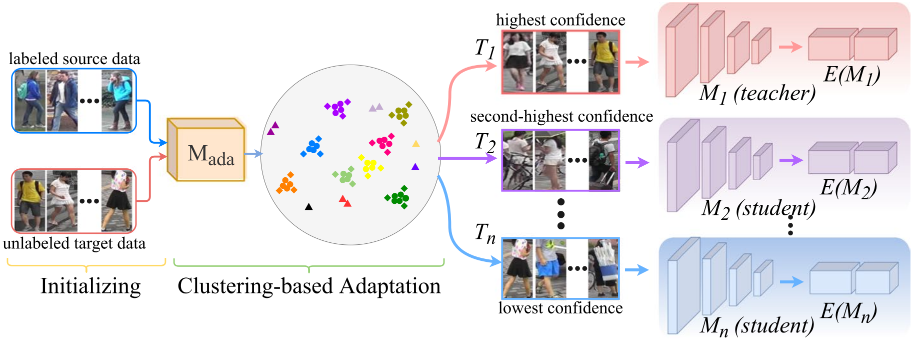
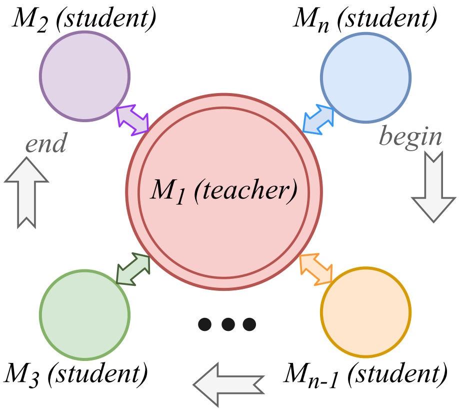


Fig. 3 Overall framework of our multiple co-teaching method. $T_1 \sim T_n$ denote target samples in different granularity levels of clustering confidences (highest → lowest). $E\left[M_1\right] \sim E\left[M_n\right]$ represent temporal average models of $M_1 \sim M_n$, which represent multiple co-teaching networks corresponding to $T_1 \sim T_n$, respectively

Fig. 4 The training manner of our propose MCN network. To be more specific, multiple student networks are organized to learn from teacher network in a progressive fashion ($M_n \rightarrow M_2$)



the second co-teaching paradigm between teacher network M_1 and student network M_{n-1}. More detailed training manner of our propose MCN network is illustrated in Fig. 4. Note that there are n-1 co-teaching paradigms in total needed to be performed between teacher network M_1 and several student networks $M_2 \sim M_n$, respectively. To be more specific, the student networks select diversified samples from lower confidence set to train teacher network when multiple students are involved progressively, which encourages the teacher network to have a basic discriminability for representation learning. The critical idea behind MCN is to create a difference between several collaborative networks for enhancing their complementarity. After several iterations in multiple co-teaching network, we use final fine-tuned teacher network M_1 to perform inference.

3.2.2 Mean-teaching mechanism

In traditional co-teaching, a popular strategy is to employ the predictions of teacher model for training other student networks. However, most of researchers always neglect that directly using the current predictions to train student models degrades the complementarity of teacher models' outputs Tarvainen and Valpola (2017). To address this issue, we introduce a mean-teaching mechanism to greatly enhance the independence and complementarity of teacher network and student networks. To be more specific, mean-teaching leverages the temporally average models of networks to generate pseudo labels for supervising each other. During the training iteration T, the parameters of the temporally average models are denoted as $\Theta\left(E^{T}[M]\right)$, which can be calculated as

$$\Theta\left(E^{T}[M]\right) = \alpha * \Theta\left(E^{T-1}[M]\right) + (1-\alpha) * \Theta(M) \tag{1}$$

where $\Theta\left(E^{T-1}[M]\right)$ indicates the temporal average parameters of the networks in the previous iteration T-1, the initial temporal average parameters are $\Theta\left(E^{0}[M]\right) = \Theta(M)$, α is the hyper-parameter within the range [0,1). Different from Tarvainen and Valpola (2017) whose weight is temporal average of the student network parameters, our teacher model is trained with diverse samples mined by student networks, which encourages the teacher network to receive samples as diverse as possible, so the weights of MCN-MT can be dynamic updated as training goes on. The pseudo hard labels of both average model and its peer network are utilized jointly to train the several collaborative networks. During the evaluation period, we adopt the past average model of teacher network for down-stream re-ID task.

3.3 Dynamic network updating

As shown in the Algorithm 1, we firstly use pre-trained ResNet-50 (He et al. 2016) on ImageNet (Deng et al. 2009) for initializing with source dataset S, then adopt source model M_{src} to extract pooling-5 features of target images T, which assigns reliable pseudo hard labels for exemplars in high-density area while noisy pseudo labels for samples in low-density area. We set a hyper-parameter n to control the granularity of discrepancy and a hyper-parameter r to represent maximum iteration round during training. Consequently, target images T can be divided into n granularity levels ($T_1 \sim T_n$ sets) based on the clustering results. In particular, M_{ada} is a fine-tuned model over diverse set $T_1, T_2, ..., T_n$, which acts as a warm start for training multiple student networks, then we perform several co-teaching paradigms between teacher network and student networks progressively. In this paper, the noisy pseudo labels caused by clustering, which result in a decline in performance, can be alleviated by our MCN framework with mean-teaching induction, this gives rise to our MCN-MT method.

In fact, many previous works (Hermans et al. 2017; Fu et al. 2019) have been found that performing training with triplet loss has great potential to learn a robust and discriminative model in person re-ID tasks. The triplet loss optimizes the embedding space such that data points with the same identity are closer to each other those with different identities. Motivated by this, in this work, we use triplet loss to mine the relationship of training samples during training, which can minimize the distance among positive pairs and maximize the distance between negative pairs. And our loss is defined as:

$$L_{triplet} = \left(d_{a,p} - d_{a,n} + m\right)_{+} \tag{2}$$

where $d_{a,p}$, $d_{a,n}$ denote the feature distances of positive pair and negative pars, respectively, m represents the margin of our triplet loss, $(z)_{+}$ denotes *max(z,0).*

Algorithm 1 The training procedure of our method

```
Input:
  Labeled source dataset $\mathcal{S}$, unlabeled target dataset $\mathcal{T}$;
  CNN model $M$, maximum iteration round $r$;
  Granularity level $n$ of self-discrepancy;
Output:
  Best model $M_1$ & $E^{T+1}[M_1]$.
▷ Baseline Initialization ***
$M_{src} \leftarrow$ Initialize $M$ on $\mathcal{S}$
▷ Clustering-based Adaptation ***
Divide $\mathcal{T}$ into inliers $\mathcal{T}_{in}$ and outliers $\mathcal{T}_{out}$ by DBSCAN clustering results
$\mathcal{T}_n \leftarrow \mathcal{T}_{out}$; k = 1;
repeat
    Divide $\mathcal{T}_{in}$ into inliers $\mathcal{T}_{in}^k$ and outliers $\mathcal{T}_{out}^k$ by clustering;
    $\mathcal{T}_{in} \leftarrow \mathcal{T}_{in}^k$, $\mathcal{T}_{n-k} \leftarrow \mathcal{T}_{out}^k$; k ++;
until n = k + 1
$\mathcal{T}_1 \leftarrow \mathcal{T}_{in}$;
$M_{ada} \leftarrow$ Fine-tune $M_{src}$ with $\mathcal{T}_1 \cup \mathcal{T}_2 \cup \cdots \cup \mathcal{T}_{n-1}$;
$M_1 \leftarrow M_{ada}$, $M_2 \leftarrow M_{ada}$, $\cdots$, $M_n \leftarrow M_{ada}$;
▷ Multiple Co-teaching Paradigm ***
for i = n → 2 do
    for T = 1 → r do
        if T % 2 == 0 then
            Deploy $M_i$ to select reliable instances from $\mathcal{T}_i$;
            Then employ samples to optimize $M_1$;
            Finally update $E^{T+1}[M_1]$;
        else
            Deploy $M_1$ to select reliable instances from $\mathcal{T}_1$;
            Then employ reliable samples to optimize $M_i$;
            Finally update $E^{T+1}[M_i]$;
        end if
    end for
end for
return best model $M_1$ & $E^{T+1}[M_1]$
```

Therefore, through the triplet loss via multiple co-teaching, the discriminative ability of teacher network is enhanced in the domain adaptation process, and more reliable or hard samples are involved during training, eventually leading to the improvement of the multiple co-teaching network.

4 Experiment

4.1 Datasets and protocols

We evaluate our method on three benchmark datasets, including Market-1501 (Zheng et al. 2015), DukeMTMC-reID (Ristani et al. 2016; Zheng et al. 2017) and CUHK03 (Li et al. 2014). The evaluation protocols are also elaborated before showing our experiments results.

Market-1501 Market-1501 has 1,501 identities in 32,668 images. 12,936 images of 751 identities are used for training, the query has 3,368 images and gallery has 19,732 images. All the samples are collected from 6 cameras in the summer of Tsinghua University. This dataset is automatically detected and cropped by the DPM detector. Therefore, there may be some noise samples with incomplete pedestrian components.

DukeMTMC-reID DukeMTMC-reID is collected in the winter of Duke University from 8 different cameras, which contains 16,522 images of 702 identities for training, and the remaining images of 702 identities for testing, including 2,228 images as query and 17,661 images as gallery, the bounding-box of this dataset are all manually annotated.

CUHK03 CUHK03 consists of 14,097 images with a total 1,467 identities, which are collected in the Chinese University of Hong Kong with only 2 cameras. This datasets has two settings of labelling: human labeled bounding boxes and DPM detected bounding boxes. In this work, we use the DPM detected setting since it is more challenging and closer to real-world scenarios.

Protocols In this experiment, we follow the standard evaluation protocol used in Zhong et al. (2017) and adopt mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-1, Rank-5 and Rank-10 for performance evaluation on all the candidate datasets.

4.2 Implementation details

In this paper, we follow the training procedure in Yang et al. (2020) and empirically set $\alpha = 0.999$ in Eq. 1. ImageNet pretrained model is adopted for supervised learning during baseline initializing period. Specifically, we keep the size of input images and resize them to 128×64. The batch size of training samples is set as 64. As for triplet selection, we randomly selected 16 persons and sampled 4 images for each identity, m is set as 0.5 in Eq. 2. For data augmentation, we employ random cropping, flipping and random erasing (Zhong et al. 2020). DBSCAN clustering (Ester et al. 1996) is employed to mine the samples with different granularity of self-discrepancy, the minimum size of a cluster is constrained to 4 and the density radius is set to 1.6×10^{-3}. Additionally, we set maximum iteration rounds r = 30 until it reaches convergence state.

4.3 Ablation study

To further validate the effectiveness of the our proposed method, we perform several ablation studies on the individual component of our proposed multiple co-teaching method.

The effectiveness of proposed MCN To argue the effectiveness of our proposed method MCN, we conduct extensive experiments under another setting of simple fine-tuning with single network. As depicted in Table 1, it can be easily observed that our multiple co-teaching network can achieve more competitive performance than fine-tuning with

Table 1 Ablation study. We evaluate the performance (%) of the simple fine-tuning and our proposed MCN respectively

Method	DukeMTMC → Market-1501		Market-1501 → DukeMTMC	
	Rank-1↑	mAP↑	Rank-1↑	mAP↑
Direct transfer	57.6	20.6	28.3	15.2
Fine-tuning	72.8	51.6	64.3	46.8
MCN (Ours)	82.6	63.2	72.5	53.5

"Direct transfer" means a model trained on source dataset is directly adopted for evaluation on target dataset

target images in cross-domain re-ID task, *e.g.*, our MCN method can achieve a remarkable performance of **82.6%** in rank-1 accuracy and **63.2%** in mAP on Market-1501 dataset, however, it can only obtain 72.8% in rank-1 accuracy and 51.6% in mAP by simple fine-tuning strategy with single network. Not surprisingly, mAP accuracy is also significantly reduced from 53.5% to 46.8% if directly applying fine-tuning on DukeMTMC-reID benchmark, which demonstrates the superiority of our proposed MCN method.

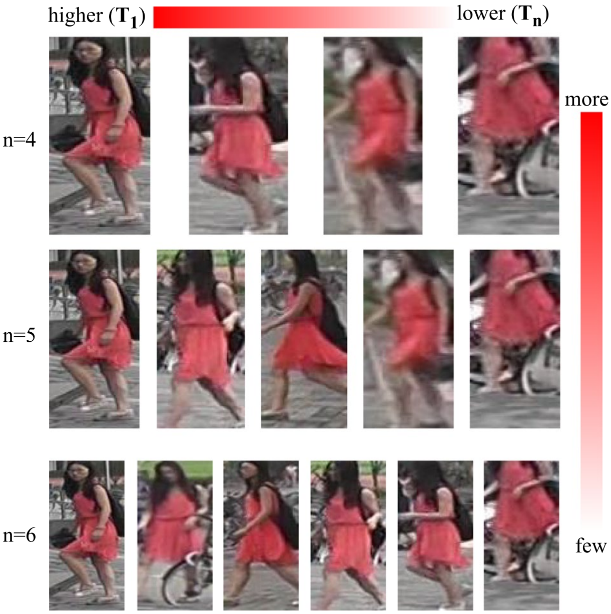
The effectiveness of mean-teaching mechanism We evaluate the mean-teaching component proposed in Sect. 3.2.2. As illustrated in Table 2 and Fig. 7, when n=3, results show that mAP accuracy drops significantly from **64.9%** to **63.2%** on Market-1501 and from **57.8%** to **53.5%** on DukeMTMC-reID respectively without adopting mean-teaching induction. Additionally, similar drops can also be observed no matter which self-discrepancy parameter is employed in multiple co-teaching framework. The effectiveness of the mean-teaching can be largely attributed to that it enhances the discrimination capability of all collaborative networks during multiple co-teaching, which is vital for domain adaptation in cross-domain where the target supervision is not available.

The impact of discrepancy granularity Intuitively, n determines the granularity of the self-discrepancy relations, when n=2, only two models are trained collaboratively. As n increases, retrieval accuracy improves at first. However, mAP accuracy does not always increase with confidence granularity level n. As illustrated in Fig. 7, when n=4, 5 or 6, the performance drops dramatically, regardless of using mean-teaching induction. To go even further, we gave an explanation about this phenomenon from two aspects: Qualitative perspective and Quantitative perspective.

First, from the qualitative perspective, a visualization of the self-discrepancy offers explanations to this phenomenon, as shown in Fig. 5. When n equals to 4, 5 or 6, some images of same pedestrian in different confidences are very similar to each other, so the inputs of different branches share a high similarity, which may degrade the complementary capacity of MCN-MT. Additionally, some image pairs when n=3 are shown in Fig. 6, which has a significant difference among different image pairs.

Second, from the quantitative perspective, our qualitative observations above are confirmed by the quantitative evaluations. On the one hand, we use F-score (Otto et al. 2017) to measure the pseudo label generation quality of our proposed MCN-MT method. As depicted in Fig. 7, the quality of pseudo labels will be negatively affected with an over-increased n. When n equals to 3, we can obtain the highest F-score on Market-1501 and DukeMTMC-reID datasets respectively, suggesting the high quality of our pseudo labels, which may be the main reason that the proposed MCN-MT can achieve the best performance when n=3. In real-world applications, we would recommend to use n=3. On the

Fig. 5 Visualization of the self-discrepancy under different n values. When n=4 or n=5 or n=6, some samples of same pedestrian in different self-discrepancy granularity repeat with others or share a high similarity



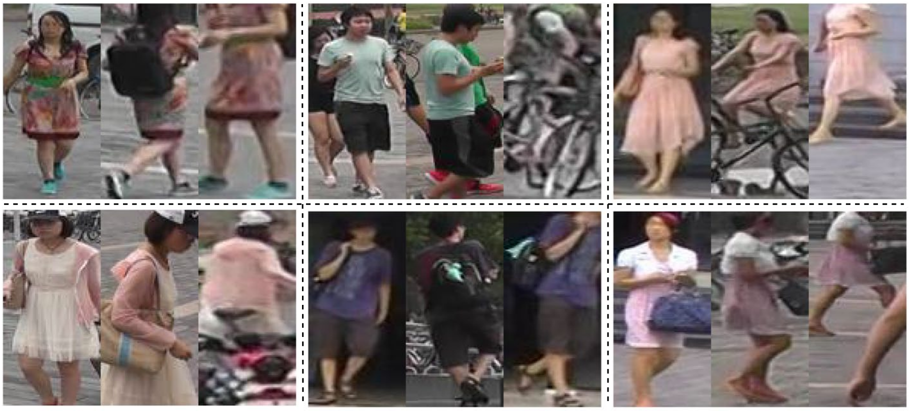
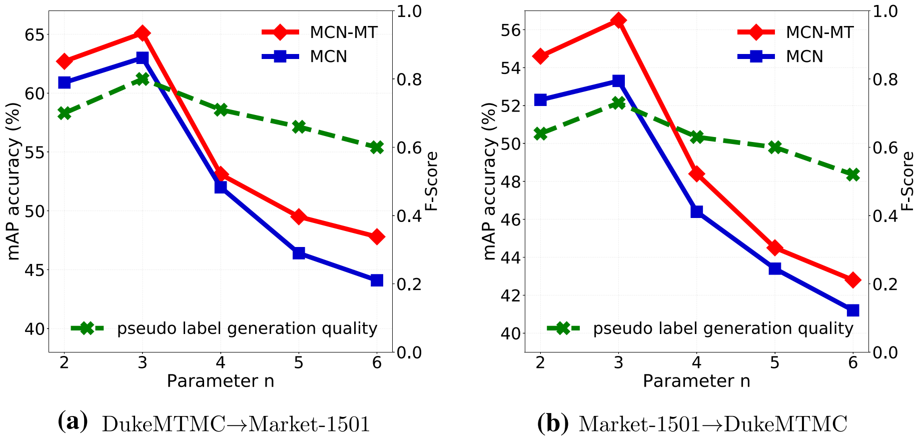


Fig. 6 Examples of instances with self-discrepancy obtained by the MCN-MT in a mini-batch when n=3

other hand, we adopt the metric of Structural Similarity[2] (SSIM) (Wang et al. 2004) to measure the variation of intra-class relation with different self-discrepancy values. According to the Fig. 8, we found that the images of same pedestrian in a large n values (*e.g.* n=4, 5 or 6) share a high similarity (lower SSIM scores), which may degrade the complementary capacity of MCN-MT. In contrast, the input images with a smaller value (*e.g.* n=3) tend to have a higher SSIM score, indicating that the data inputs of teacher network and

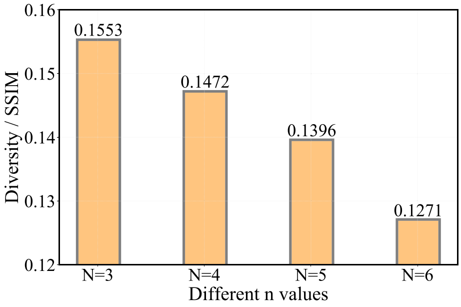
[2] A image pair with lower SSIM score means two images are much more similar visually.



(a) DukeMTMC→Market-1501 **(b)** Market-1501→DukeMTMC

Fig. 7 Impact of *n*. mAP accuracy are compared. We adopt average F-score of MCN-MT (higher is better; marked in **green** dash-line) to measure the quality of pseudo label generation on Market-1501 and DukeMTMC-reID datasets, respectively

Fig. 8 SSIM (higher is better) scores to evaluate the diversity of intra-class relation with different n values. Please note that we did not mesure the SSIM score when n=2 since our MCN method with two networks (n=2) maybe degrade into the asymmetric co-teaching approach ACT method



student networks have a large variation among each other, which can significantly enhance the difference and complementarity of multiple co-teaching networks.

4.4 Comparison with the state-of-the-art methods

In this section, we compare our proposed method with the hand-crafted feature approaches: LOMO (Liao et al. 2015), Bow (Zheng et al. 2015) and UMDL (Peng et al. 2016), GAN-based re- ID models: PTGAN (Wei et al. 2018), SPGAN (Deng et al. 2018) and HHL (Zhong et al. 2018); clustering-based methods PUL (Fan et al. 2018), SSG (Fu et al. 2019) and PCB-PAST (Zhang et al. 20190, as well as hybrid-based approaches, ECN (Zhong et al. 2019), MAR (Yu et al. 2019), EANet (Huang et al. 2018) and co-teaching based methods ACT (Yang et al. 2020) and MMT (Ge et al. 2020). More detailed comparison results of these methods are demonstrated as follows:

Evaluation on Market-1501 According to the results from Tables 2 and 3, it can be seen clearly that our MCN-MT (*w/* Mean-Teaching) achieves remarkable rank-1 accuracy of 84.3 and 84.8% when trained on DukeMTMC-reID and CUHK03 respectively, outperforming the second-best methods MMT (Ge et al. 2020) and ACT (Yang et al. 2020) by

Table 2 Performance (%) comparisons with the state-of-the-art methods

Method	DukeMTMC-reID→Market-1501				Market-1501→DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
LOMO Liao et al. (2015)	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
Bow Zheng et al. (2015)	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
UMDL Peng et al. (2016)	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PTGAN Wei et al. (2018)	38.6	57.3	66.1	15.7	27.4	43.6	50.7	13.5
SPGAN Deng et al. (2018)	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
HHL Zhong et al. (2018)	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
PUL Fan et al. (2018)	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
SSG Fu et al. (2019)	80.0	90.0	92.4	58.3	73.0	80.6	83.2	53.4
PCB-PAST Zhang et al. (2019)	78.4	–	–	54.6	72.4	–	–	54.3
ECN Zhong et al. (2019)	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
MAR Yu et al. (2019)	67.7	81.9	–	40.0	67.1	79.8	–	48.0
ACT Yang et al. (2020)	80.5	–	–	60.6	72.4	–	–	54.5
MMT($\mathcal{L}^{t}_{stri}$) Ge et al. (2020)	***84.0***	***93.4***	***95.4***	***62.6***	**74.9**	**85.2**	**89.5**	**58.1**
MCN (Ours)	82.6	90.7	94.1	63.2	72.5	81.8	84.6	53.5
MCN-MT (Ours)	**84.3**	**93.6**	**95.9**	**64.9**	***74.7***	***83.8***	***86.3***	***57.8***

Bold indicates the best and **BoldItalic** indicates the second best

Table 3 Performance (%) comparisons with the state-of-the-art methods

Method	CUHK03→Market-1501				CUHK03→DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
PTGAN Wei et al. (2018)	31.5	–	60.2	–	17.6	–	38.5	–
SPGAN Deng et al. (2018)	42.3	–	–	19.0	–	–	–	–
HHL Zhong et al. (2018)	56.8	74.7	81.4	29.8	42.7	57.5	64.2	23.4
EANet Huang et al. (2018)	66.4	–	–	40.6	45.0	–	–	26.4
ACT Yang et al. (2020)	***81.2***	–	–	***64.1***	***52.8***	–	–	***35.4***
MCN (Ours)	82.2	92.4	95.5	66.1	53.3	66.3	71.3	37.2
MCN-MT (Ours)	**84.8**	**93.1**	**95.7**	**68.7**	**56.3**	**67.3**	**73.3**	**40.2**

Bold indicates the best and **BoldItalic** the second best

+2.3% and **+4.6%** in mAP accuracy, demonstrating the priority of our proposed multiple co-teaching network. The superiority of our proposed method can be largely contributed to the self-discrepancy of intra-class relations mined by MCN-MT during multiple collaborative training, as well as mean-teaching mechanism, which is beneficial to learn a more robust and discriminative model in UDA re-ID tasks.

Evaluation on DukeMTMC-reID When performing evaluation on DukeMTMC-reID[3] dataset, our approach has also achieved superior results than state-of-the-art methods on

[3] Note that the DukeMTMC and its derived datasets have been officially removed due to some ethic concerns. Here we include it only for the sake of comparison to some existing results. We discourage further usage of DukeMTMC datasets in the future.

Table 4 Ablation study

Method	DukeMTMC → Market-1501		Market-1501 → DukeMTMC	
	Rank-1↓	mAP↓	Rank-1↓	mAP↓
MCN-MT (hard labels)	84.3	64.9	74.7	57.8
MCN-MT (soft labels)	77.8	59.8	71.2	52.7

We evaluate the performance (%) of our proposed MCN-MT with hard pseudo labels and soft pseudo labels respectively

this dataset, even DukeMTMC-reID is the most challenging dataset currently with some body occlusion and overlap. To be more specific, when comparing to MMT (Ge et al. 2020), our model obtains nearly similar mAP score when trained on Market-1501, but achieving a higher rank-1 accuracy leading by **+3.5%** improvement comparing to asymmetric co-teaching method ACT (Yang et al. 2020) when trained on CUHK03. It is worth mentioning that MMT utilizes soft softmax-triplet loss with soft triplet labels, and performance of MMT-500 (*w/* L^t_{stri}) & (*w/o* L^t_{sid}) is reported in Table 2, which indicates that MMT is more complex than our proposed method and this may be the main reason leading to the better performance when trained on Market-1501.

4.5 Discussion

As shown in Table 2, when tested on Market-1501→DukeMTMC-reID, we find an interesting phenomenon that performance of MCN-MT is slightly inferior and less competitive compared with MMT (Ge et al. 2020) (*w/* L^t_{stri}). Generally speaking, soft pseudo-labels perform relatively better than hard labels in general application with symmetric networks, which motivates us to adopt soft pseudo-labels on MCN-MT methods. So a natural question then comes to our attention: *how good performance will our method achieve with soft pseudo-labels?* To find the answer to this question, we have performed some experiments by adopting soft pseudo-labels in our MCN-MT method, which is generated by the past temporally average model of teacher and student networks. The detailed results are shown in Table 4.

Unfortunately, we found that our model is extremely difficult to reach a convergence state with soft pseudo-labels during training process, which leads to significantly performance degradation on DukeMTMC-reID and Market-1501 dataset, *e.g.*, with soft pseudo labels, we can only achieve a mAP accuracy of **59.8%** on DukeMTMC-reID→Market-1501, and **52.7%** on Market-1501→DukeMTMC-reID respectively. We suspect this is due to the inputs of multiple student networks MCN-MT are asymmetric and have a large difference in terms of self-discrepancy relations. As a result, the soft pseudo-labels of same identity generated by the past temporally average model cannot maintain their consistency, which undoubtedly brings some negative impacts on the training of multiple co-teaching networks.

In addition, we also report the training time, memory cost, and the parameters of the proposed MCN-MT method in the Table 5. Specifically, the experiments are conducted on a server equipped with four RTX 2080 Ti GPUs. According to the Table 5, it can be

Table 5 The training time, GPU memory cost and the parameters of the proposed MCN-MT method on Market-1501 and DukeMTMC dataset respectively

Method	DukeMTMC → Market-1501			Market-1501 → DukeMTMC		
	#time	#GPU_memory	#params	#time	#GPU_memory	#params
MCN-MT	17h	19G	118M	19h	19G	119M

easily observed that our method can perform multiple co-teaching paradigms with acceptable computational or memory cost, which allows our method more flexible and adaptable in practical scenarios.

5 Conclusion and future work

In this paper, we firstly present a simple yet effective multiple co-teaching network MCN to mine the self-discrepancy in target domain for UDA re-ID task, which trains several neural networks simultaneously with unlabeled samples in coarse-grained discrepancy. Furthermore, a novel mean-teaching induction is introduced to further enlarge the difference and learn discriminative features on the basis of MCN. By plugging our mean-teaching mechanism into MCN, the complementarity of the teacher network and student network is significantly enhanced. Comprehensive experiments conducted on benchmark datasets show that our method outperforms the state-of-the-art UDA methods by a clear margin. As a future direction, we will extend our method to handle with other challenging computer vision tasks, such as and vehicle re-identification and fine-grained image retrieval.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Project (Grant No. 61977045). The authors would like to thank the anonymous reviewers for their valuable suggestions and constructive criticisms.

Funding This work was supported by the National Natural Science Foundation of China under Project (Grant No. 61977045).

Data availability statement The data used for the experiments in this paper are available online, see Sect. 4.1 for more details.

Code availability The code is available at https://github.com/JeremyXSC/MCN-MT.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Deng, J., Dong, W., & Socher, R. et al. (2009). Imagenet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition,* IEEE, pp. 248–255.
Deng, W., Zheng, L., & Ye, Q. et al. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 994–1003.

Ester, M., Kriegel, H.P., & Sander, J. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* pp. 226–231.

Fan, H., Zheng, L., Yan, C. et al. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications and Applications, 14*(4), 1–18.

Farenzena, M., Bazzani, L., & Perina, A. et al. (2010). Person re-identification by symmetry-driven accumulation of local features. In: *IEEE Conference on Computer Vision and Pattern Recognition,* IEEE, pp. 2360–2367.

Fu, Y., Wei, Y., & Wang, G. et al. (2019). Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: *IEEE International Conference on Computer Vision,* pp. 6112–6121.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In: *International conference on machine learning,* PMLR, pp. 1180–1189.

Ge, Y., Chen, D., & Li, H. (2020). Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526.

Han, B., Yao, Q., & Yu, X. et al. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. arXiv preprint arXiv:1804.06872.

He, K., Zhang, X., & Ren, S. et al. (2016). Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 770–778.

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

Huang, G., Liu, Z., Van Der Maaten, L. et al. (2017). Densely connected convolutional networks. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4700–4708.

Huang, H., Yang, W., & Chen, X. et al. (2018). Eanet: Enhancing alignment for cross-domain person re-identification. arXiv preprint arXiv:1812.11369.

Li, W., Zhao, R., & Xiao, T. et al. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 152–159.

Liao, S., Hu, Y., & Zhu, X. et al. (2015). Person re-identification by local maximal occurrence representation and metric learning. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2197–2206.

Otto, C., Wang, D., & Jain, A. K. (2017). Clustering millions of faces by identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(2), 289–303.

Peng, P., Xiang, T., & Wang, Y. et al. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1306–1315.

Ristani, E., Solera, F., & Zou, R. et al. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In: *European Conference on Computer Vision,* Springer, pp. 17–35.

Romero, A., Ballas, N., Kahou, S.E. et al. (2014). Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550

Saito, K., Yamamoto, S., & Ushiku, Y. et al. (2018). Open set domain adaptation by backpropagation. In: *European Conference on Computer Vision,* pp. 153–168.

Song, L., Wang, C., Zhang, L. et al. (2020). Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition, 102*(107), 173.

Szegedy, C., Vanhoucke, V., & Ioffe, S. et al. (2016). Rethinking the inception architecture for computer vision. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2818–2826.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780

Wang, Z., Bovik, A. C., Sheikh, H. R. et al. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Wei, L., Zhang, S., & Gao, W. et al. (2018). Person transfer gan to bridge domain gap for person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 79–88.

Xiang, S., Fu, Y., & Chen, H. et al. (2020a). Multi-level feature learning with attention for person re-identification. *Multimedia Tools and Applications, 79*(43), 32,079–32,093.

Xiang, S., Fu, Y., & You, G. et al. (2020b). Unsupervised domain adaptation through synthesis for person re-identification. In: *IEEE International Conference on Multimedia and Expo,* IEEE, pp. 1–6.

Xiang, S., Fu, Y., You, G. et al. (2021). Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification. In: *IEEE International Conference on Acoustics, Speech and Signal Processing,* IEEE, pp. 3765–3769.

Yang, F., Li, K., & Zhong, Z. et al. (2020). Asymmetric co-teaching for unsupervised cross-domain person re-identification. In: *AAAI Conference on Artificial Intelligence,* pp. 12,597–12,604.

Ye, M., Shen, J., & Lin, G. et al. (2021) .Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, H.X., Zheng, W.S., & Wu, A. et al. (2019). Unsupervised person re-identification by soft multilabel learning. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2148–2157.

Zhai, Y., Ye, Q., & Lu, S. et al. (2020). Multiple expert brainstorming for domain adaptive person re-identification. arXiv preprint arXiv:2007.01546.

Zhang, X., Cao, J., & Shen, C. et al. (2019). Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: *IEEE International Conference on Computer Vision,* pp. 8222–8231.

Zhao, F., Liao, S., & Xie, G.S. et al. (2020). Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In: *European Conference on Computer Vision,* Springer, pp. 526–544.

Zheng, L., Shen, L., & Tian, L. et al. (2015). Scalable person re-identification: A benchmark. In: *IEEE International Conference on Computer Vision,* pp. 1116–1124.

Zheng, L., & Yang, Y., Hauptmann, A.G. (2016). Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: *IEEE International Conference on Computer Vision,* pp 3754–3762

Zhong, Z., Zheng, L., & Cao, D. et al. (2017). Re-ranking person re-identification with k-reciprocal encoding. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1318–1327.

Zhong, Z., Zheng, L., & Li, S. et al. (2018). Generalizing a person retrieval model hetero-and homogeneously. In: *European Conference on Computer Vision,* pp. 172–188.

Zhong, Z., Zheng, L., & Luo, Z. et al. (2019). Invariance matters: Exemplar memory for domain adaptive person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 598–607.

Zhong, Z., Zheng, L., & Kang, G. et al. (2020). Random erasing data augmentation. In: *AAAI Conference on Artificial Intelligence,* pp. 13,001–13,008.