Multimedia Tools and Applications (2020) 79:19769–19786
https://doi.org/10.1007/s11042-020-08723-x

Unsupervised person re-identification by hierarchical cluster and domain transfer

Suncheng Xiang[1] · Yuzhuo Fu[1] · Mingye Xie[1] · Zefang Yu[1] · Ting Liu[1]

Received: 9 May 2019 / Revised: 15 December 2019 / Accepted: 31 January 2020 /
Published online: 30 March 2020

Abstract
Person re-identification (re-ID) has recently been tremendously boosted due to the advancement of deep convolutional neural networks. Unfortunately, the majority of deep re-ID methods focus on supervised, single-domain re-ID task, while less attention is paid on unsupervised domain adaptation. Therefore, these methods always fail to generalize well to real-world scenarios, which have attracted much attention from academia. To address this challenge, we propose a joint unsupervised domain adaptive re-ID method, named HCTL, which is aided by Hierarchical Clustering and Transfer Learning. Specifically, our method performs camera invariance learning using iStarGAN by transferring style of reliable images, which is mined by hierarchical clustering, to the style of other cameras in target domain. During training stage, HCTL integrates TriHard loss on top of ResNet-50 to reduce intra-class variance among dataset and enforce connectedness simultaneously between source domain and target domain. Comprehensive experiments based on Market-1501, DukeMTMC-reID and CUHK03 are conducted, results indicate that our method robustly achieves state-of-the-art performances with only a few reliable samples in target domain and outperform any existing approaches by a large margin.

Keywords re-ID · Unsupervised domain · Hierarchical clustering · TriHard loss

1 Introduction

Person Re-Identification (re-ID) has been attached with increasing attention in the computer vision community recently, due to its importance for many vision-related application,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11042-020-08723-x) contains supplementary material, which is available to authorized users.

✉ Suncheng Xiang
xiangsuncheng17@sjtu.edu.cn

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

including content-based video retrieval, robotics, human computer interaction, and so forth. Given an image of the person we are interested, the goal of the re-ID is to find the other images of the same person captured by different cameras or same cameras in a different time, even if the pedestrian identities are unseen before.

Encouraged by the remarkable success of deep learning methods and the availability of large scale re-ID datasets, performance of person re-identification has been significantly boosted. For example, the rank-1 accuracy on DukeMTMC-reID [17] has been improved from 25.13% [27] to 85.95% [8], the rank-1 accuracy of single query on Market-1501 [27] have been improved from 43.8% [12] to 93.68% [8], so deep feed-forward architectures has brought impressive advances to the state-of-the-art across a wide variety of re-ID tasks, which leads to a significant improvement in pedestrian retrieval system.

Although the performance is pleasing with current datasets, there still remain some issues hindering the application of unsupervised domain adaptive re-ID task. In fact, these leaps in performance come only when a large amount of labeled training data is available, in addition, training and testing data is in the same feature space or follow the similar distribution. However, this assumption may not hold for many practical situations, where the training data and testing data are in different domain. It may be still possible to obtain training sets which are big enough for training large-scale deep models, but that suffer from the *shirt* in data distribution from the actual data encountered at "test time", which is named as "domain gap" in cross domain re-ID task. To be more specific, training and testing on different domain (or datasets) results in severe performance drop, even below the level of hand-crafted features, e.g., the model trained on Market-1501 only achieves the rank-1 accuracy of 18.5% [16] when tested on DukeMTMC-reID, in such cases, supervised, single-domain re-ID methods may lose the generalization ability in real-world scenarios, where domain-specific labels are not available, so person re-ID is still very challenging. As shown in Fig. 1, domain gap could be caused by dataset bias, such as image resolution, human pose changes, scene variations and seasons, even background in different datasets. These challenges highlight the inevitable need for powerful and robust feature for person re-ID, because available training samples can not be effectively leveraged for new target domains since the feature spaces and the marginal probability between training and testing datasets are different. In our setting, the source domain is fully annotated while the unknown target domain does not have ID labels since annotatingperson ID labels is expensive. To address this problem, some transfer learning methods [2, 24] have been used for narrow-down or eliminate the domain gap, other works [11, 13, 15, 25, 29] take advantage of GAN model for sample generation, and focus on the development of effective features to describe visual appearance of person or an appropriate metric to measure the similarity among person images. Not surprisingly, these approaches have expressed their power in single-shot re-ID problem for domain adaption. However, these works either require large auxiliary information about target domain or employ both the attribute and identity labels to learn atransferable model for discriminative representation, which is not realistic in real-world scenarios. So, further progress in this task is needed.

In this paper, we focus on the unsupervised domain adaptation in person re-ID task as it is more common in real-life applications, and propose a joint unsupervised learning framework HCTL through exploiting more reliable unlabeled samples in target domain to perform camera invariance and domain connectedness learning. Compared with existing domain adaptation techniques, our method can achieve better performance with less

Fig. 1 Illustration of the domain gap in Market-1501(*left*), DukeMTMC-reID (*middle*) and CUHK03(*right*). The differences among different datasets in lighting or background, and large variance among images of the same person make person re-ID challenging

unlabeled data in target domain, and free human from heavy data annotations. On the other hand, aiming to facilitate the research towards applications in realistic scenarios, the returned rank-M images are then re-ranked with k-reciprocal encoding algorithm [31], which can significantly boost the performance of unsupervised domain adaptive re-ID task. To the best of our knowledge, this is the first attempt to explicitly take the limited reliable images into account to perform camera invariance learning in the target set, which is instructive to construct a high-quality re-ID dataset and enable the interaction of fine-grained feature among different datasets.

In summary, the main contributions of the proposed method are as follows:

- We introduce a joint unsupervised learning framework HCTL, which can exploit the reliable unlabeled samples to boost the performance of re-ID in a completely unsupervised manner.
- A novel iStarGAN network is designed to perform camera invariance learning, which is capable of achieving the camera invariance property in an unsupervised fashion and progressively enhance the diversity of the generated images.
- The extensive experiments compared with the state-of-the-art methods on benchmark datasets verify the effectiveness of the proposed method.

In the rest of the paper. we first review some related works of person re-ID in Section 2. Then in Section 3, we give the learning procedure of the proposed joint unsupervised domain adaptation model HCTL. Extensive evaluations compared with state-of-the-art and comprehensive analyses of the proposed approach are reported in Section 4. Conclusions are given in Section 5.

2 Related work

In this section, we have a brief review on related work about unsupervised domain adaptive re-ID task. The mainstream idea of the existing unsupervised re-ID methods is to learn a discriminative model for target domain. These methods can be roughly divided into hand-crafted based approaches and deep learning based approaches.

Technically, there are mainly two fundamental components we need to consider for a conventional person re-ID framework: discriminative feature extraction and similarity metrics by computing distance. Traditional studies [4, 18, 21] related to hand-crafted systems for person re-ID aim to design or learn robust feature for person re-ID, e.g., Farenzena et al. [4] presented an appearance-based method with overall and local chromatic content. Rui et al. [18] proposed an approach of learning mid-level filters from automatically discovered patch clusters for person re-identification, which reaches the balance between discriminative power and generalization ability. Besides directly using mid-level color and texture features, some methods also strive to utilize pedestrian attributes which aremore robust to image transformations [14]. Unfortunately, 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.

Recently, deep learning based methods have tremendously pushed forward the boundary of re-ID, as well as bike-person re-ID [26]. These methods focus on designing various deep CNN structures to learn discriminative feature embeddings and/or strive to devise better loss functions for training the networks, but for contrastive loss [30], since it may leads to over-fitting when the number of images is limited. Another choice in unsupervised domain adaptive re-identification consists in deploying clustering [3, 22] to exploit the similarity among identities in target domain. For example, a self-weighted multiview clustering method [22] is proposed to cluster feature points by incorporating their motion and context similarities in crowd scenes. Fan et al. [3] propose a progressive unsupervised learning method PUL consisting of clustering and fine-tuning the network, which is similar to our self-training scheme. While, PUL only focuses on unsupervised learning, not unsupervised domain adaptation. In addition, their iteration framework is not guided by specific assumptions thus having no theoretical derived loss functions. Recently, Zhong et al. [32] first propose a HHL method to learn camera-invariant network for the target domain. However, HHL overlooks the latent positive pairs in the target domain, this might lead the re-ID model to be sensitive to other variations in the target domain, such as pose and background variations. Additionally, it requires large auxiliary information about target domain to learn a camera-invariant model, which is not practical in real-world scenarios.

It is known that Generative Adversarial Networks (GAN) [2, 9, 24, 29] have been shown to be effective to produce state-of-the-art results for person re-identification in unsupervised domain adaptation and several related methods [23, 32] have been proposed to progressively narrow-down the domain gap between source and target dataset, for example, SPGAN [2] learns a similarity preserving GAN model by using the negative pairs to improve the image-image translation model, Wei et al. [24] proposed a Person Transfer GAN network to bridge the domain gap between two different style of datasets and migrate pedestrian style of one dataset to another one. Similarly, Wang et al. presented a SSIM Embedding Cycle GAN to transform the synthetic image to the photo-realistic image. However, most of existing unsupervised domain adaptation methods assume that class labels are the same across domain, while the person identities of different re-ID datasets are totally different. Hence, the approaches mentioned above failed to be utilized directly for the problem of unsupervised domain adaptive re-identification.

Motivated by these works, we explicitly consider the intra-domain image variations caused by cameras and propose to generate unlabeled images to enhance the performance of cross-domain re-ID task. Specifically, a pre-trained re-ID model is used to guide the training of GAN

and make the generated samples more adaptive to person re-ID task in target domain. In this paper, we address the unsupervised domain adaptive re-ID task with proposed joint learning method HCTL to learn the domain invariant feature fordescribing the human identity across distinct domain.

3 Proposed approach

In this section, we will introduce our approach to single-shot re-ID problem. First, we start with a formal formulation of this problem, and then briefly look back on the baseline network structure for feature representation learning, each component of our method are presented alternately. In total, the framework of the proposed approach is shown in Fig. 2.

3.1 Formulation

In single-shot re-ID task, we are given a labeled source set $\{X_s, Y_s\}$ consisting of N_s person images. Each image x_s corresponds to a label y_s, where $y_s \in \{1, 2, \ldots, M_s\}$, and M_s is the number of identities. We also have N_t unlabeled target images from unlabeled target set X_t. The identity of each target image X_t is unknown. We compute the distance (or similarity) of these two images $x_1, x_2 \in \{X_s, Y_s\}$ by:

$$\mathrm{D}(\mathrm{x}_1, x_2) = \|g(f(x_1)) - g(f(x_2))\|_2^2 \tag{1}$$

Here $f(\cdot)$ is a feature extractor that extracts discriminative feature for each image, and $g(\cdot)$ is an aggregation function that aggregates image level features to sequence level feature. We then use it to rank all the queries. In unsupervised domain adaptive re-ID task, which means that between source domain and target domain, either the term features are different between the two sets, or their marginal distributions are different. On the whole,

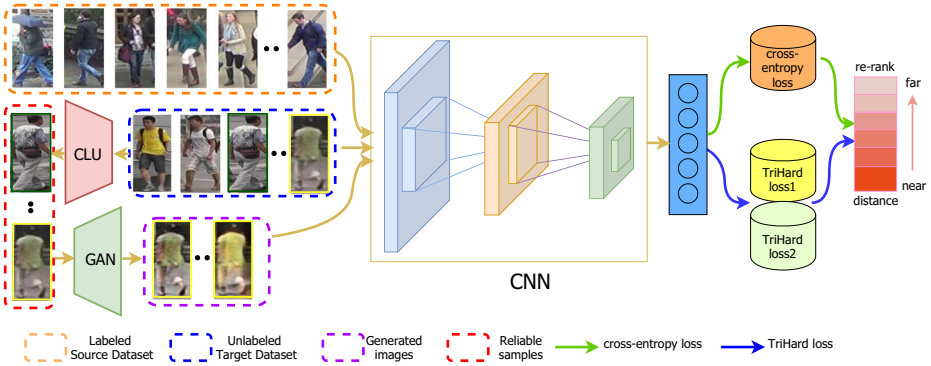


Fig. 2 The framework of the proposed approach, HCTL, integrated with two kinds of loss during training stage: cross-entropy loss for classification and TriHard loss for similarity learning, which imposes camera invariance to the model, one TriHard loss is learned through source domain and target domain, another TriHard loss is constructed by unlabeled target samples collected by clustering strategy and their corresponding cameras style transferred samples. Specifically, the re-ranking strategy is deployed to further improve re-ID accuracy in real-world scenarios. Best viewed in color

the goal of this paper is to leverage both labeled source training images and unlabeled target training images to learn discriminative embeddings of target test-set. In addition, when there exists some relationship, explicit or implicit, between the feature spaces of the two domains, we say that the source and target domains are closely related in some degree.

3.2 Baseline overview

ResNet is a well-known network structure to extract feature from lots of images dataset, which is widely used in the field of detection, segmentation and identification. We use ResNet-50 [6] as our baseline to extract feature in re-ID dataset. Assuming that there is a lot of labeled images of source domain and unlabeled images of target domain, we can obtain a ID-discriminative embedding (IDE) [28] for person re-ID task. At the same time, in order to create more fake images to train our model, we use GAN network to generate more re-ID image with different camera styles in target domain. Following the baseline training strategy [34], we obtain a strong baseline on Market-1501 and DukeMTMC-reID datasets. Specifically, we can achieve rank-1 accuracy of 32.9% on DukeMTMC-reID, training on Market-1501, which is serving as a backbone in our experiment.

Metric learning is also a method widely used in the field, unlike feature learning, metric learning aims to learn the similarity of two images through the network. On the issue of pedestrian recognition, the similarity of different pictures related the same pedestrian is greater than the different pictures of different pedestrians. In this paper, we apply the cross-entropy loss to cast the training process as a classification problem. The cross-entropy loss is written as,

$$\mathscr{L}_{\text{cross}} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\left[y^{(i)}\log\widehat{y}^{(i)} + \left(1-y^{(i)}\right)\log\left(1-\widehat{y}^{(i)}\right)\right] \tag{2}$$

where n_s is the number of labeled training images in a batch, $\widehat{y}^{(i)}$ is the predicted probability of the input belonging to ground-truth class $y^{(i)}$. In this work, if not specified, this model is regarded as our baseline to learn the ID-discriminative embedding throughout this paper.

3.3 Camera style transfer in target domain

There is a very intuitive understanding of camera invariance that the same person should be recognized as long as this person is taken a picture by camera. Unfortunately, most of re-ID models fail to perform well when facing cross-camera problems due to the disturbing factors. Considering the fact that variation of image style between different cameras in target domain is a critical influencing factor during person re-ID testing procedure [32]. In this paper, we propose a novel StarGAN-based architecture iStarGAN to generate images between multiple cameras in target domain, which is capable of achieving the camera invariance property in an unsupervised fashion and progressively enhance the diversity of the generated images. To this end, we train a deep residual network to transform the feature of images captured by one camera to appear as if they were sampled from other different cameras in target domain while maintaining the self-similarity with constraints. We emphasize that such information should not be the background or image style, but should be underlying and latent. To fulfill this goal, we integrate a SiameseNet with StarGAN, as

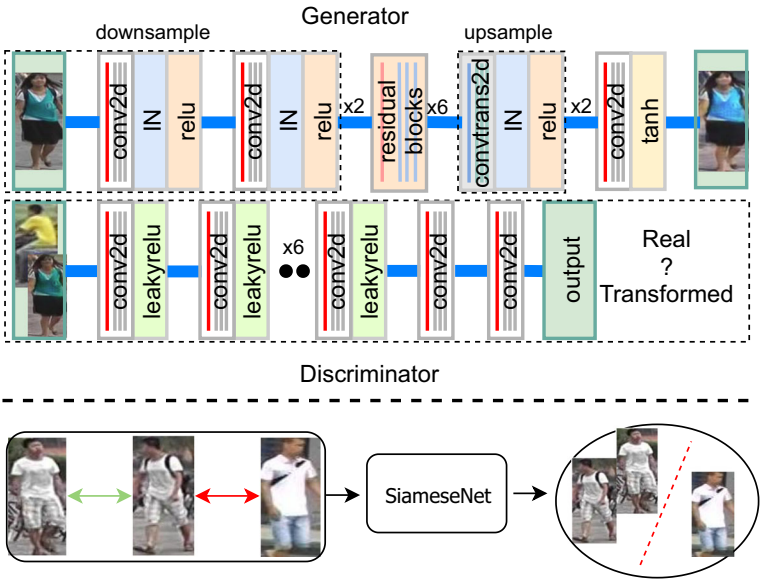


Fig. 3 iStarGAN consists two components: a StarGAN (TOP) and an SiameseNet (BOTTOM). The residual blocks in generator are employed to learn the residual representation between the pooled features from "conv2d". The discriminator takes the features of different images and the enhanced representation of domain features as inputs and tries to distinguish them, and the SiameseNet learns a latent space that constrains the learning procedure of mapping functions

shown in Fig. 3. During training, StarGAN is to learn $[N*(N-1)]/2$ mapping functions between N domains. To make the generated images indistinguishable from real images, we adopt an adversarial loss in Eq. 3.

$$\mathscr{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log (1-D_{src}(G(x,c)))] \tag{3}$$

where G generates an image $G(x, c)$ conditioned on both the input image x and the target domain label c, while D tries to distinguish between real and fake images.

Technically speaking, an n-dimensional one-hot vector $\widetilde{c}$ is deployed to allow StarGAN to learn $[N*(N-1)]/2$ mapping functions. When training StarGAN for camera transferring, we use the domain label $\widetilde{c}$ defined in Eq. 4 as input to the generator, with n being the number of cameras of specific dataset. When performing camera transferring in multiple domain, the generator learns to ignore the unspecified labels for the remaining n-1 unknown labels with help of mask vector m, and focuses on the explicitly given label. For the remaining n-1 unknown labels we simply assign zero values. In our experiments, we utilize the Market-1501 and DukeMTMC-reID datasets, where n is 6 and 8, respectively.

$$\widetilde{c} = [c_1, \ldots, c_n, m] \tag{4}$$

Especially, to enhance the identity consistency and self-similarity of an image before and after translation, SiameseNet is deployed to learn a latent space that constrains the learning of mapping function. Note that, we select training pairs in an unsupervised manner, and do not require corresponding image pairs which is the same person captured from different cameras in

target domain. We utilize the contrastive loss [5] to train the SiameseNet without additional annotations.

$$\mathscr{L}_{con}(i, x_1, x_2) = (1-i)\{\max(0, M-d)\}^2 + id^2 \quad (5)$$

where x_1 and x_2 are a pair of input vectors, d denotes the Euclidean distance between normalized embeddings of two input vectors, and i represents the binary label of the pair. $i = 1$ if x_1 and x_2 are positive pair; $i = 0$ if x_1 and x_2 are negative pair. $M \in [0, 2]$ is the margin that defines the separability of the negative pair in the embedding space. When $M = 0$, the loss of negative training pair is not back-propagated in the system. When $M > 0$, both positive and negative sample pairs are considered. A larger M means that the loss of negative training samples has a higher weight in back propagation.

In the experiment, our improved iStarGAN model learns with both unlabeled target images and their counterparts with same ID by imposing the camera invariance constraint to achieve the camera invariance property in target domain. With the learned iStarGAN model to perform camera invariance learning, we can generate new target images that preserve the person identity and reflect the style of another camera, e.g., we have real unlabeled images $x_{t,j}$ collected by camera j ($j \in \{1, 2, \ldots, m\}$ in the target domain, iStarGAN generates m fake images, denotes as $x'_{t,1}$, $x'_{t,2}$, $x'_{t,3}$, ..., $x'_{t,m}$, which more or less contain the same person with $x_{t,j}$ but whose camera styles are similar to camera 1, 2, 3, …, m, respectively. It is worth noting that the m images include the one transferred to the style of camera j, that is, the style of the real images $x_{t,j}$. One example of the fake images generated by iStarGAN are shown in Fig. 4. Quantitative analyses of camera style transfer are to be viewed in Section 4.4.

3.4 Cross-domain learning

In this section, we propose principled approach aided by TriHard loss to build the stronger connectedness between the source domain and the target domain in the feature space. Generally speaking, TriHard loss can randomly sample three images from the training data, and optimize the embedding space such that data points with the same identity are closer to each other than those with different identities. It has been shown extensively that mining hard negatives is more important in learning a discriminative embedding than naively learning from all visual samples [7, 19, 20]. Selecting hard samples for the training model through hard mining has been shown to be effective, if trained sample pairs are not hard training pairs, then this is not conductive to being better representation of network learning. In fact, many pioneering works have been found that mining of hard triplets when using TriHardloss can improve the generalization ability of the network. To the best of our knowledge, TriHard loss

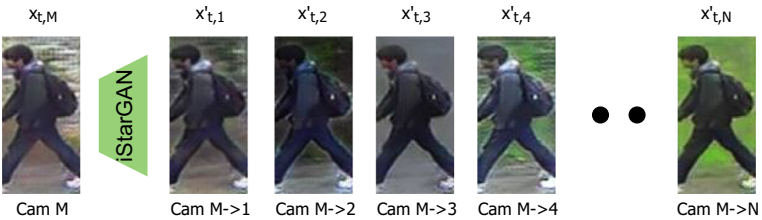


Fig. 4 Example of multi-camera image-to-image translation using our proposed iStarGAN

is an effective method to select the hardest positive and negative training pairs in person re-ID training. In this paper, $\mathscr{L}_{triplet1}$ and $\mathscr{L}_{triplet2}$ can be defined as:

$$\mathscr{L}_{triplet1} = \frac{1}{N_1} \sum_{a \in batch} \left(\max_{p \in \{x_s^i\}_{i=1}^{n_s}} d_{a,p} - \min_{n \in \{x_t^i\}_{i=1}^{n_t}} d_{a,n} + m \right)_+ \tag{6}$$

$$\mathscr{L}_{triplet2} = \frac{1}{N_2} \sum_{a \in batch} \left(\max_{p \in \{x_t^i\}_{i=1}^{n_t}} d_{a,p} - \min_{n \in \{x'^i_t\}_{i=1}^{n'_t}} d_{a,n} + m \right)_+ \tag{7}$$

where m denotes margin of TriHard loss, N_1, N_2 demote the number of samples in the batch; p and n are the collections of positive pairs and negative pairs. $\{x_s^i\}_{i=1}^{n_s}$, $\{x_t^i\}_{i=1}^{n_t}$ and $\{x'^i_t\}_{i=1}^{n'_t}$ represent labeled source domain samples, unlabeled target domain samples and their corresponding camera transferred samples, respectively.

In this paper, we propose to jointly learn camera invariance and cross-domain connectedness using cross-entropy loss and TriHard loss in a training batch. To be more specific, there are labeled source images, unlabeled target images and their corresponding camera-style-transferring images. Finally, the over all objective loss function (Fig. 2) in a training batch is expressed as:

$$\mathscr{L}_{our} = \mathscr{L}_{cross} + \alpha * \mathscr{L}_{triplet1} + (1-\alpha) * \mathscr{L}_{triplet2} \tag{8}$$

The proposed network is trained to minimize the three loss functions jointly. To figure out which objective function contributes more, we train the identification model with different loss functions separately, following the learning rate setting in Section 4.2. First, we train the models with cross-entropy loss function. Then, TriHard loss function is employed with cross-entropy to train the network jointly until two objectives both reaching convergence. The experiment details can be accessed in Section 4, our re-ID model with two kinds of objective loss outperforms the one trained individually. This result has been confirmed on the backbone ResNet network structures.

3.5 Data reducing in target domain

Section 3.3 describes that we can generate more fake images in target domain by using iStarGAN network. In real scenarios, it is not practical to use all samples to perform camera style transfer since latent feature of generated images will overlap with real images. That is to say, selecting some reliable samples from target domain to perform camera style transfer is needed. Existing method such as k-means algorithm [3] fails to work because we can not get the identity information of target domain, or how many identities in unknown domain. To address this problem, an improved Hierarchical Clustering Algorithm is deployed to solve this problem in unsupervised domain adaptive re-identification. The optimization procedure of Hierarchical Clustering is presented in Alg. 1, which combines two closed clusters by calculating the distance between different categories. It is an effective way for us to obtain some reliable and representative images in an unknown target domain, which is combined with images in source

domain to construct TriHard loss. Illustration of the learning procedure can be shown in Fig. 5.

Algorithm 1 Procedure of improved hierarchical clustering algorithm.
Input: Unlabeled data $\{x_i\}_{i=1}^{N}$; Number of clusters N_t; Distance Metric learning function d.
Output: N_t clusters of unlabeled data.
1: Initializing each sample in $\{x_i\}_{i=1}^{N}$ as one cluster;
2: **repeat**
3: Finding two nearest clusters C_{i*} , C_{j*} with metric function d;
4: Combining two clusters C_{i*} , C_{j*} : C=C_{i*}∪C_{j*};
5: **until** Number of clusters = N_t
6: **return** $\mathscr{G}$ = {C_1;C_2,…,C_{Nt}}

At the beginning, hierarchical algorithm is deployed to cluster target images into N_t clusters based on currently setting, and then randomly sample one image from each cluster to compose training data of target domain. In essence, it depends on the Euclidean distance in clustering. Clustering is performed to select some reliable instances and generate a reliable training set, which can be used to perform camera invariance learning in Section 3.3.

In this work, we propose to learn camera invariance and domain connectedness simultaneously to obtain more generalized person embeddings on the target domain. In essence, it depends on the camera transfer model iStarGAN to degrade to variation of image style between different cameras, additionally, a hierarchical clustering algorithm is deployed to exploit more reliable and representative samples to provide auxiliary information for our discriminative feature learning model.

4 Experiments

4.1 Datasets

To test the accuracy and validate the effectiveness of the proposed method, three large-scale public datasets are employed in our experiments.

Market-1501 [27] contains 6 cameras from campus in Tsinghua University. Specially, one of cameras is a low pixel, it has 32,668 labeled images of 1501 identities. At the same time, Market-1501 is divided into train-set and test-set. The training set contains 12,936 images from 751 identities and the test set contains 19,732 images from 750 identities, meanwhile, there are some distractors in the dataset.

DukeMTMC-reID [17] is collected from campus with 8 cameras. Similar to the division of Market-1501 dataset, it has 36,411 labeled images belonging to 1404 identities, which contains 16,522 training images from 702 identities, 2228 query images from another 702 identities and 17,661 gallery images.

CUHK03 [10] contains 14,096 images which are collected by the Chinese University of Hong Kong, the images are taken from only 2 cameras. In fact, CUHK03 dataset has two train / test settings: using labeled bounding boxes and using DPM detected bounding boxes. Detected setting is used in our experiment because it is more closer to practical scenarios. Considering that images in CUHK03 do not have labels of camera ID, so we use CUHK03 only as source domain instead of the target domain.

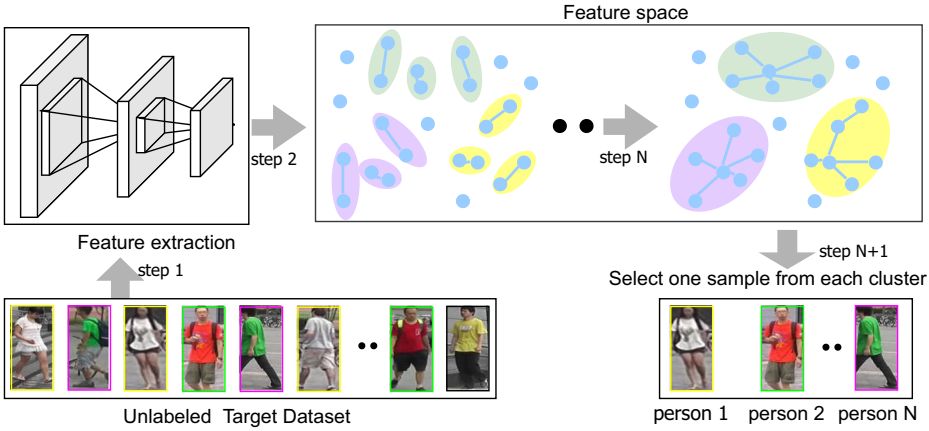


Fig. 5 An illustration of unlabeled data clustering procedure with hierarchical clustering in the feature space. As the cluster model goes on, more challenging data can be mined. Eventually, we randomly select only one reliable sample from each cluster to perform camera invariance learning

4.2 Implementation details

In experiment, we adopted PyTorch to implement and train our cross-domain learning network. Our network is modified based on the ResNet-50 [6], following the baseline training strategy introduced in [33] on 4 Tesla P100 GPU. Specifically, we keep the aspect ratio of input images and resize them to 256×128. In training, random cropping and random flipping are applied. In testing, we extract the output of the 1024-dim Pool-5 layer as the image descriptor and use the Euclidean distance to compute the similarity between thequery and database images. All experiments are conducted on a server equipped with a Intel Xeon E5-2690 V4 CPU.

In order to make our method more robust to parameters changes. In the following experiment, we set $M = 2$ in Eq. 5 as the margin of the contrastive loss, $m = 0.3$ in Eqs. 6 and 7 as the margin of the TriHard loss, $\alpha = 0.6$ in Eq. 8, learning rate $\eta = 0.1$ and training times $epochs = 70$. When real target images $n_t = 0$, only source images are used for training the re-ID model with IDE and triplet loss, this is called strongly unsupervised learning. In fact, as shown in Fig. 7(c), compared with using all the unlabeled image in target domain, our re-ID model can achieve good performance with only a few reliable images on Market and Duke datasets, which is called weakly supervised learning.

4.3 Important parameters

We evaluate two important parameters, *i. e.*, the weight of the TriHard loss α and the number of real target image n_t for camera invariance learning. We fix one parameter when evaluating the other parameter. Results are shown in Fig. 6.

Weight of the TriHard loss When $\alpha = 0$, our method comes back to the baseline (without target samples for training, Section 3.2); when $\alpha = 1$, our model only takes the cross-domain learning into consideration (without camera invariance information for training, Section 3.4). It is clearly shown that, our approach significantly improves the baseline at all values. The rank-1 accuracy and mAP improve with the increase of α and achieve the best performance when α is

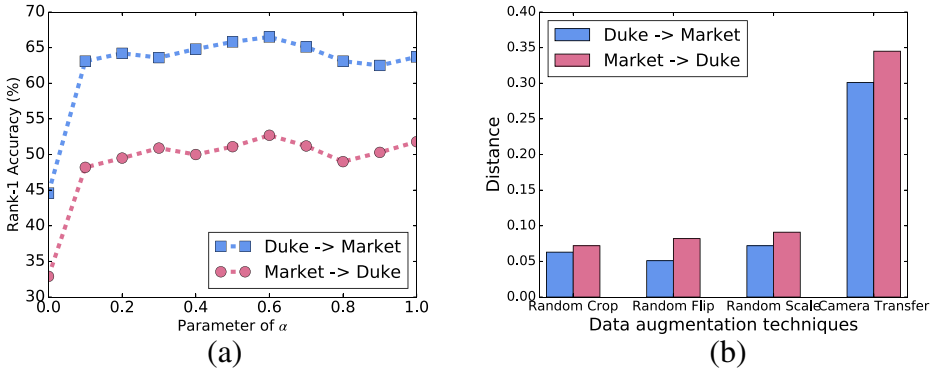


Fig. 6 **a**: Sensitivity of parameter α; **b**: The average distance between two images that undergo different data augmentation techniques

set to 0.6 for the weighting factors of first TriHard loss. In our experiment, we set $\alpha = 0.6$ since it gets best performance compared with other setting, as shown in Fig. 6(a).

Number of the real target images to perform transfer learning When $n_t = 0$, only source images are used for training the re-ID model with IDE, so our method reduces to baseline performance. On the contrary, if we use all the target images to perform camera transfer learning, number of generated images will grow to6 and 8 times on Market-1501 and DukeMTMC-reID datasets, respectively, which will leads to a large time cost for model training. Inevitably, some noises can also influence the performance of re-ID model. Consequently, considering the tradeoff between performance and time cost, we select about 750 unlabeled images to perform camera invariance learning each camera, and the number of generated images may increase to 6×750 for Market-1501 and 8×750 for DukeMTMC-reID, which is less stressful compared with the baseline task.

Based on the above analysis, our method is robust to parameters changes. In the following experiment, if not specified, we set $\alpha = 0.6$ and $n_t = 750$.

4.4 Evaluation

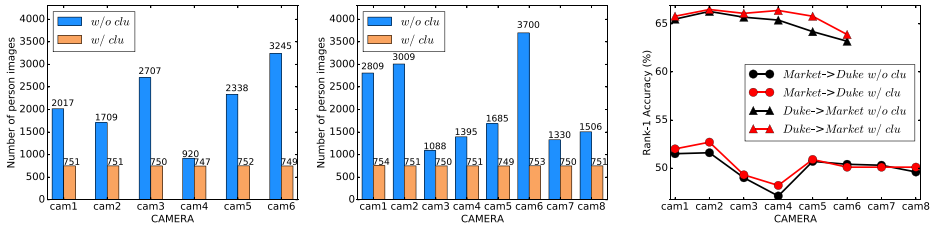
In this section, we conduct ablation study on components of the proposed model and investigate the effect of augmented data. In addition, a quantitative analysis is performed to further explain the necessity of our proposed method HCTL.

Effectiveness of camera style transfer In unsupervised domain adaptive re-ID task, transfer learning has been widely used to narrow-down or eliminate the domain gap. To prove our assumption, we compute the distance between images that undergoes different data augmentation method, *i. e.*, random cropping, random flipping, random scale and camera transfer. As shown in Fig. 6(b), we can clearly observe that the re-ID model trained on source set is robust to random cropping and random flipping, as well as random scale on target set, but it is sensitive to image variations caused by camera transfer. Therefore, the change of image style captured by different cameras on target set is a key influencing factor that should be explicitly considered in person re-ID task.

Effectiveness of hierarchical clustering In this paper, hierarchical clustering algorithm is deployed to get some representive images on the unlabeled target dataset, as shown in Fig. 7(a) and (b), this practice, to a great extent, reduces the amount of unlabeled data in the target domain on Market-1501 and DukeMTMC-reID datasets, respectively. Not surprisingly, as shown in Fig. 7(c), compared with all images used in target domain(*w/o clu*), we can still achieve good performance with less unlabeled data(*w/ clu*). To be more specific, as shown in Table 1, when tested on DukeMTMC-reID, our model yields a rank-1 accuracy of 48.0% and 42.2% when trained on Market-1501 and CUHK03, respectively, leads to +2.2% and +0.8% improvement in mAP accuracy. On the other hand, our model gets rank-1 accuracy of 65.1% and 57.1% with 750 unlabeled images of Market-1501 when using DukeMTMC-reID and CUHK03 as the source set, achieving +1.7% and +0.6% improvement in mAP accuracy, respectively. Most importantly, It is noteworthy that clustering can reduce the scale of samples for training, which leads to considerable benefit of train time cost without affecting the performance of re-ID model. This proves once again that the effectiveness of clustering with HCTL.

Effectiveness of cross-domain learning In this section, we study the benefit of cross-domain learning in Table 1, which help endow the domain connectedness to the re-ID system. When adding TriHard loss to cross-entropy loss, which aimed for metric and similarity learning, as shown in Table 1, "$L_c + L_t$" is consistently improved in all settings. Specially, the rank-1 accuracy of "$L_c + L_t$" on DukeMTMC-reID is increased from 32.9% to 45.0% trained on Market-1501. When tested on Market-1501, we can get rank-1 accuracy of 58.6% and 56.7% when using DukeMTMC-reID and CUHK03 as source set, respectively. Especially, when applying hierarchical clustering and re-ranking on DukeMTMC-reID dataset, our method "$L_c + L_t + clu + rk$" can get rank-1 accuracy of 52.7% and 46.4% trained on Market-1501 and CUHK03, which surpass the baseline (L_c) by +22.2% and +21.0% in mAP accuracy, respectively. During the ablation study, significant improvement is also observed in other setting, so the use of cross-domain learning has a prominent contribution to the performance of our model.

Correlation diagnosing Till now, we have not provided an accurate metric or principle how to evaluate correlation between diversity of train-set and performance of re-ID model. In our experiment, it is surprising to find that there exists positive correlation between number of



(a)Market-1501 (b)DukeMTMC-reID (c)Performance

Fig. 7 Performance evaluation on Market-1501 and DukeMTMC-reID with our proposed method. Meanwhile, Market-1501 has 6 cameras and DukeMTMC-reID has 8 cameras. Number of person images in each camera *without clu* and *with* textitclu are shown in Fig. 7(**a**) and (**b**). To be more specific, Market->Duke means that our model tested on DukeMTMC-reID, trained on Market-1501, *w/o clu* and *w/ clu* represent *without clustering*, *with clustering*, respectively. Rank-1 accuracy of HCTL (rk) on DukeMTMC-reID and Market-1501 are shown clearly in Fig. 7(**c**). HCTL (rk) represents our method with re-ranking

Table 1 Performance comparison under different experimental settings, L_c, L_t, *clu* and *rk* note cross-entropy loss, TriHard loss, hierarchical clustering and re-rank, respectively

Method	Duke->Market			Market->Duke			CUHK03->Market			CUHK03->Duke		
	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP
L_c	44.6	62.5	20.6	32.9	49.5	16.9	41.9	57.3	18.3	23.0	34.0	12.5
$L_c + L_t$	58.6	76.0	30.5	45.0	61.0	25.1	56.7	74.1	29.3	41.8	56.7	21.3
$L_c + L_t$ + clu	65.1	81.4	32.2	48.0	63.1	27.3	57.1	74.9	29.9	42.2	57.2	22.1
$L_c + L_t$ + clu + rk	67.0	78.6	43.9	52.7	64.3	39.1	58.8	71.5	40.1	46.4	58.8	33.5

identities and rank-1 accuracy when scale of unlabeled images is fixed, e.g., correlation coefficient R between number of identities and rank-1 accuracy: $R_{(Market->Duke)} = 0.77$; $R_{(CUHK03->Duke)} = 0.58$, when tested on DukeMTMC-reID, trained on Market-1501 and CUHK03, respectively. More detailed comparison can be assessed in our supplementary material. So there comes a meaningful conclusion: the stronger diversity of training datasets are, the better performance model will achieve when total number of images is fixed. That is to say, diversity in train-set is a key factor to influence performance of re-ID model instead of scale. This conclusion is very meaningful since it will provide guidance for us to construct a high-quality re-ID dataset. Furthermore, we can improve our performance by enhancing the diversity of person re-ID train-set in the further research.

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

In this section, we compare our proposed method with the state-of-the-art unsupervised re-ID models including: (1) the hand-crafted feature representation based models LOMO [12], UMDL [16] and BOW [27]; (2) unsupervised domain adaptation based models PUL [3], CAMEL [25], SPGAN [2], HHL [32] and CFSM [1]. Specific comparisons on Market-1501 and DukeMTMC-reID are shown in Tables 2 and 3, respectively. In the tables, HCTL (rk) represents our method with re-ranking [31].

Table 2 Performance comparison with state-of-the-art methods when trained on DukeMTMC-reID and Market-1501, separately. HCTL (rk) represents our method with re-ranking. Best performances are marked with bold emphasis

Method	DukeMTMC-reID->Market-1501				Market-1501->DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [12]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
UMDL [16]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
BOW [27]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
PUL [3]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
CAMEL [25]	54.5	–	–	26.3	–	–	–	–
SPGAN [2]	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
HHL [32]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
CFSM [1]	61.2	–	–	28.3	49.8	–	–	27.3
HCTL	65.1	**81.4**	**86.2**	32.2	48.0	63.1	**69.4**	27.3
HCTL (rk)	**67.0**	78.6	83.0	**43.9**	**52.7**	**64.3**	69.0	**39.1**

Table 3 Performance comparison with state-of-the-art methods when trained on CUHK03. HCTL(rk) represents our method with re-ranking. Best performances are marked with bold emphasis

Method	CUHK03->Market-1501				CUHK03->DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PTGAN [24]	31.5	–	60.2	–	17.6	–	38.5	–
PUL [3]	41.9	57.3	64.3	18.0	23.0	34.0	39.5	12.0
SPGAN [2]	42.3	–	–	19.0	–	–	–	–
HHL [32]	56.8	74.7	81.4	29.8	42.7	57.5	**64.2**	23.4
HCTL	57.1	**74.9**	**81.8**	29.9	42.2	57.2	63.6	22.1
HCTL(rk)	**58.8**	71.5	76.4	**40.1**	**46.4**	**58.8**	63.8	**33.5**

Market-1501 On Market-1501 dataset, we first compare our model with three hand-crafted features based models, LOMO, UMDL and BOW, these methods have inferiority and fail to produce competitive results on large-scale datasets because of their limited discriminative feature learning ability. For example, the rank-1 accuracy of LOMO and BOW on Market-1501 are only 27.2% and 35.8%, respectively. This is much lower than deep learning based methods. In order to make the comparison more comprehensive, we also compare our results with some unsupervised domain adaptation based methods, which are proposed recently, such as Progressive Unsupervised Learning (PUL), Clustering-based Asymmetric Metric Learning (CAMEL), Similarity Preserving Generative Adversarial Network with domain adaptation (SPGAN), HHL and CFSM (AAAI'19). In the multiple-query setting, our method achieves rank-1 accuracy of **65.1%** and **57.1%** when trained on DukeMTMC-reID and CUHK03, outperforming the second best method HHL by +2.9% and +0.3%, respectively. With the help of re-ranking, the rank-1 accuracy are further improved to **67.0%** and **58.8%** in our HCTL(rk), outperforming the best of pervious work HHL by +4.8% and +2.0% in rank-1 accuracy, +12.5% and +10.3% improvement in mAP separately. Again, our HCTL surpasses recently proposed CFSM by a large margin when trained on DukeMTMC-reID, which again verifies the effectiveness of our proposed method.

DukeMTMC-reID On DukeMTMC-reID, it can be easily found that HCTL surpasses person transfer PTGAN, Progresive Unsupervised Learning (PUL), domain adaptation SPGAN by a large margin. Surprisingly, our HCTL is slightly inferior to Person Retrieval Model HHL when trained on CUHK03, probably because the lighting conditions in this dataset are so extreme that the domain gap between CUHK03 and DukeMTMC-reID is still significant. Fortunately, with the help of re-ranking, our HCTL(rk) yields a competitive rank-1 accuracy of **52.7%** and **46.4%** when trained on Market-1501 and CUHK03, respectively, leads to a improvement of +5.8% and +3.7% in rank-1 accuracy, +11.9% and +10.1% in mAP accuracy comparing with second best method [32]. This confirms the effectiveness of the proposed solution, which does not require any human supervision and thus scales to large camera networks.

Discussion In essence, both HHL and HCTL focus on unsupervised domain adaptation for person re-identification, the performance of HCTL is really closed to HHL when tested on DukeMTMC-reID or Market-1501. However, HHL uses unlimited samples in target domain

to perform discriminative representation learning, which leads to a time-consuming and high-complexity training process, especially on large-scale person re-ID dataset. For instance, it will take an extra two hours during training by employingHHL instead of HCTL. On the contrary, our proposed HCTL can still achieve a competitive performance with limited reliable samples (about 750) in target domain to learn a discriminative model, it does not require any specially designed deep architecture for feature representation learning, this allows HCTL to be highly flexible. Considering the analyzing above, we firmly believe that our joint learning method HCTL is significantly better than recently proposed HHL approach.

5 Conclusion

In this paper, we have presented HCTL as a high performance approach for unsupervised domain adaptive re-ID task, which takes advantage of iStarGAN and hierarchical clustering to obtain more generalized embeddings on target domain with limited scale reliable images, then enhance the connectedness between source and target domain. To the best of our knowledge, our work is the first attempt to explicitly take the limited reliable samples into account to perform camera invariance learning in target domain. Extensive comparative evaluations were conducted on various challenging benchmarks, comprehensive experiments show that HCTL outperforms the state-of-the-art unsupervised domain adaptive re-ID approaches by a big margin. In addition, we also show that our method is complementary to other data augmentation techniques. In future work, we will extend HCTL to video-based person re-ID task with spatial-temporal patterns and attention mechanism.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Project(Grant No.61472244), the National Defense Pre-Research Foundation of China(Grant No.513110501) The authors would like to thank the anonymous reviewers for their valuable suggestions and constructive criticism.

References

1. Chang X, Yang Y, Xiang T, Hospedales TM (2019) Disjoint label space transfer learning with common factorised space. In: The Thirty-Third AAAI Conference on Artificial Intelligence, pp 3288–3295
2. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (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
3. Fan H, Zheng L, Yan C, Yang Y (2018) Unsupervised person re-identification: clustering and fine-tuning. ACM Trans Multimed Comput Commun Appl 14(4):83
4. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2360–2367
5. Hadsell R, Chopra S, Lecun Y (2006) Dimensionality reduction by learning an invariant mapping. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1735–1742
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
7. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv preprint arXiv:170307737
8. Kalayeh MM, Basaran E, Gokmen M, Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1062–1071

9. Lee HY, Tseng HY, Huang JB, Singh MK, Yang MH (2018) Diverse image-to-image translation via disentangled representations. In: European Conference on Computer Vision, pp 35–51
10. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 152–159
11. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 384–393
12. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2197–2206
13. Lin S, Li H, Li CT, Kot AC (2018) Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. arXiv preprint arXiv:180701440
14. Liu X, Song M, Zhao Q, Tao D, Chen C, Bu J (2012) Attribute-restricted latent topic model for person re-identification. Pattern Recogn 45(12):4204–4213
15. Lv J, Chen W, Li Q, Yang C (2018) Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 7948–7956
16. Peng P, Xiang T, Wang Y, Pontil M, Gong S, Huang T, Tian Y (2016) Unsupervised cross-dataset transfer learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1306–1315
17. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, pp 17–35
18. Rui Z, Ouyang W, Wang X (2014) Learning mid-level filters for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 144–151
19. Shi H, Yang Y, Zhu X, Liao S, Zhen L, Zheng W, Li SZ (2016) Embedding deep metric for person re-identification: a study against large variations. In: European conference on computer vision, pp 732–748
20. Song HO, Yu X, Jegelka S, Savarese S (2016) Deep metric learning via lifted structured feature embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4004–4012
21. Wang X, Doretto G, Sebastian T, Rittscher J, Tu PH (2007) Shape and appearance context modeling. In: IEEE International Conference on Computer Vision, pp 1–8
22. Wang Q, Chen M, Nie F, Li X (2018) Detecting coherent groups in crowd scenes by multiview clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–1
23. Wang Q, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd counting in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 8198–8207
24. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer Gan to bridge domain gap for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 79–88
25. Yu HX, Wu A, Zheng WS (2017) Cross-view asymmetric metric learning for unsupervised person re-identification. In: IEEE International Conference on Computer Vision, pp 994–1002
26. Yuan Y, Zhang J, Wang Q (2018) Bike-person re-identification: a benchmark and a comprehensive evaluation. IEEE Access pp 56059–56068
27. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: IEEE International Conference on Computer Vision, pp 1116–1124
28. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv preprint arXiv:161002984
29. 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 3774–3782
30. Zheng Z, Zheng L, Yang Y (2018) A discriminatively learned cnn embedding for person re-identification. ACM Trans Multimed Comput Commun Appl 14(1):13
31. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3652–3661
32. Zhong Z, Zheng L, Li S, Yang Y (2018a) Generalizing a person retrieval model hetero-and homogeneously. In: European Conference on Computer Vision, pp 172–188
33. Zhong Z, Zheng L, Zheng Z, Li S, Yang Y (2018b) Camera style adaptation for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 5157–5166
34. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, pp 2242–2251

Suncheng Xiang received his M.S. degree in software engineering from National University of Defense Technology, Changsha, China, in 2017. He is currently pursuing the PhD. degree in computer science and technology from Shanghai Jiao Tong University, Shanghai, China. His research interests include deep learning, image retrieval and representation learning.