International Journal of Machine Learning and Cybernetics (2021) 12:2281–2295
https://doi.org/10.1007/s13042-021-01308-6

ORIGINAL ARTICLE

Imitating targets from all sides: an unsupervised transfer learning method for person re-identification

Jiajie Tian[1] · Zhu Teng[1] · Baopeng Zhang[1] · Yanxue Wang[2] · Jianping Fan[3]

Received: 13 March 2020 / Accepted: 13 March 2021 / Published online: 2 April 2021

Abstract

Person re-identification (Re-ID) models usually present a limited performance when they are trained on one dataset and tested on another dataset due to the inter-dataset bias (e.g. completely different identities and backgrounds) and the intra-dataset difference (e.g. camera and pose changes). In other words, the absence of identity labels (who the person is) and pairwise labels (whether a pair of images belongs to the same person or not) leads to failures in unsupervised person Re-ID problem. We argue that synchronous consideration of these two aspects can improve the performance of unsupervised person Re-ID model. In this work, we introduce a Classification and Latent Commonality (CLC) method based on transfer learning for the unsupervised person Re-ID problem. Our method has three characteristics: (1) proposing an imitate model to generate an imitated target domain with estimated identity labels and create a pseudo target domain to compensate the pairwise labels across camera views; (2) formulating a dual classification loss on both the source domain and imitated target domain to learn a discriminative representation and diminish the inter-domain bias; (3) investigating latent commonality and reducing the intra-domain difference by constraining triplet loss on the source domain, imitated target domain and pairwise label target domain (composed of pseudo target domain and target domain). Extensive experiments are conducted on three widely employed benchmarks, including Market-1501, DukeMTMC-reID and MSMT17, and experimental results demonstrate that the proposed method can achieve a competitive performance against other state-of-the-art unsupervised Re-ID approaches.

Keywords Person re-identification · Unsupervised transfer learning · Dual classification

1 Introduction

The person re-identification (Re-ID) [69] targets at matching images of people in a large-scale dataset collected by non-overlapping camera views. In this task, images of person undergoes large variations in illumination, appearance, pose, viewpoints and background in different cameras. It attracts significant attentions from both academia and industry due to its great potential applications in video surveillance and security. Thanks to the development of deep learning [47, 59], the person Re-ID performance has been significantly improved in recent years. For example, the Rank-1 accuracy of single query on Market-1501 [68] has been upgraded to 93.8% [43]. The Rank-1 accuracy on DukeMTMC-reID [38, 70] has been improved to 83.3% [43]. However, the achieved high performance for person Re-ID is only restricted to supervised learning frameworks [47, 67, 73] since the database consists of a large number of manually labeled images. While in a practical person Re-ID deployment, such manual labeling is not only expensive to obtain

✉ Baopeng Zhang
bpzhang@bjtu.edu.cn

Jiajie Tian
17120413@bjtu.edu.cn

Zhu Teng
zteng@bjtu.edu.cn

Yanxue Wang
snow875@163.com

Jianping Fan
jfan1@lenovo.com

[1] The School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

[2] Baotou Railway Vocational and Technical College, Inner Mongolia, China

[3] AI Lab, Lenovo Research, Beijing, China

as the number of cameras increases, but also improbable in many cases because it requires the same person appearing in every pair of existing cameras. When models trained on a supervised dataset are directly used on another dataset, the Re-ID performance declines precipitously due to the inter-dataset bias [10, 37]. Therefore, learning an unsupervised person Re-ID model that generalizes well on a target domain is important and also of great relevance to applications [64, 74].

One solution to learn an unsupervised person Re-ID model is unsupervised domain adaption (UDA) where models are trained on a source domain consisting of labeled images and adapted on the target domain composed of unlabeled images. Recently, numerous unsupervised methods for person Re-ID [22, 49] have been proposed to extract view-invariant features. But these methods only achieve a limited Re-ID performance compared to the supervised counterparts. The main reason is that the inter-domain bias between the labeled source domain and the unlabeled target domain is not reduced effectively. Different domain images are taken under different views in different backgrounds, and even people who appear in these images might come from different nations. We consider these differences as the domain gap or inter-domain bias. In the unsupervised setting, no labeled images in the target domain are provided such that it is more important to exploit label information in the source domain to shrink the inter-domain bias. In other words, without labels in the target domain (i.e. identity label) as a learning guidance, it is hard to utilize the supervised learning framework to learn identity discriminative feature through the classification task. Meantime, another factor that influences the performance of person Re-ID is the intra-domain difference which is caused by different camera configurations in the target domain. Even in the same domain, images captured by different cameras have distinctive styles due to various lighting condition, shooting angle, background, etc. Since the goal of the Re-ID test procedure is to identify pedestrians across cameras, pairwise labels (whether a pair of images belong to the same person or not) across cameras could be a great advice to exploit camera-invariance in the training of Re-ID models.

In this work, we propose a method to explicitly address issues mentioned above. On the one hand, persons in the target domain are imitated from the labeled source domain in order to compensate for the absence of identity labels in the target domain. On the other hand, a content preserved pseudo target domain is derived in order to fill the vacancy of pairwise labels across camera views in the target domain. Furthermore, we leverage a dual classification loss on both source domain and imitated target domain to strengthen the discriminative ability of the proposed person Re-ID model and bridge inter-domain bias. There are some works [10, 54] that focus on similarity-preserving source-target translation models to bridge domain gaps, and also some methods [72, 73] that concentrate on camera style adaptation to generate new datasets in the style of other cameras. But none of them takes into account both factors. We argue that a transfer model is impacted by the overall data gap between two domains during a training period and at the same time influenced by the camera styles of the target domain in the test phase. Moreover, to enhance the generalization ability of the proposed person Re-ID model, a latent commonality of domains beyond source and target is exploited, i.e., the margin of two images originated from the same person should be smaller than that from different persons across any camera in any domain. To this end, we develop a novel unsupervised transfer learning method, named Classification and Latent Commonality method (CLC), to train a cross dataset person Re-ID model. CLC does not require any manual annotations for images in the target domain, but requires identity labels for source dataset and camera IDs for images in the source and target dataset. Note that the camera ID for each image can be easily obtained along with raw videos.

To sum up, we propose the CLC method for the unsupervised person Re-ID task. Three contributions are made: (**I**) In order to simultaneously compensate for the absence of identity labels and pairwise labels across camera views in the target domain, we design an imitate method to generate an imitated target domain and a pseudo target domain. (**II**) For the purpose of decreasing the inter-domain bias, we introduce a dual classification loss on both source domain and imitated target domain to learn a discriminative representation. (**III**) We utilize a triplet loss constrained on the source domain, imitated target domain and pairwise label target domain (composed of pseudo target domain and target domain) to investigate camera-invariance and reduce the intra-domain difference.

2 Related work

Supervised person Re-ID. Most existing person Re-ID models are based on supervised learning, i.e. trained on a large number of labeled images across cameras. They focus on feature engineering [20, 53, 55, 65, 66], distance metric learning [7, 18, 48, 60], creating new deep learning architectures [1, 24, 31] and re-ranking methods [2, 14, 23, 61, 71]. For example, Kalayeh et al. [20] learned both local features from human body parts and global features from entire human image to integrate human semantic parsing. Chen et al. [7] proposed a quadruplet loss to handle the weakness of the triplet loss on person Re-ID. Li et al. [24] proposed a new person Re-ID network with a joint learning of soft and hard attentions, which took advantage of both joint learning attention selection and feature representation. Geng et al. [14] designed a Perspective Distance Model (PDM) to further reduce the intra-class variations and increase the distance of inter-class variations. Although these models offer a promising performance on recent person Re-ID datasets (e.g., Market-1501 [68] and DukeMTMC-reID [38, 70]), it is hard to utilize in practical applications due to the demand of tremendous labeled data.

Unsupervised domain adaptation. Our work is also closely related to the unsupervised domain adaptation (UDA) [34, 42, 51, 52], where during training only the labeled source dataset and unlabeled target dataset are available. In this community, most of previous methods aimed to align the feature distributions between the two domains [33, 45], which has been justified theoretically by Ben-David et al. [5]. For example, Tzeng et al. [46] proposed a domain classifier to encourage the features of two domains to be indistinguishable. There were also some methods [39, 41] aimed at predicting pseudo-labels on the unlabeled target domain. Sener et al. [41] proposed to utilize the k-nearest neighbors to provide the labels of unlabeled samples. Most of the UDA methods assume that class labels are the same across domains, which forms a close set problem. However, in practice, there are many scenarios that source domain and target domain have different labels, which is a open set domain adaptation [3, 26, 30, 36, 40, 44]. For instance, Panareda et al. [36] proposed a method to learn a mapping from the source to the target domain and then jointly solve an assignment problem for labels between source and target domain. These methods are limited on the assumption that source domain and target domain share labels. In this paper, we study the problem of UDA in person Re-ID, where source domain and target domain contain entirely different identities (classes). It is more challenging open set problem.

Unsupervised person Re-ID. The supervised methods obtain a remarkable performance thanks to the large amount of labeled data and the deep networks [24]. However, the performance drops dramatically when employed on an unseen dataset. Hand-craft features [4, 12, 15, 27, 35] could be directly employed in unsupervised cross-domain person Re-ID. But the cross domain data is not fully exploited by these features because they neglect the inter-domain bias. In the unsupervised person Re-ID community, some works [11, 13, 29, 32, 57, 57, 58, 62, 64] used labeled source data to initialize model and attempted to predict pseudo-labels of unlabeled target images. For instance, Fan et al. [11] proposed a method that iteratively applied data clustering, instance selection and fine-tune techniques to estimate labels of images in target domain. Liu et al. [32] predicted reliable labels with k-reciprocal nearest neighbors. Yu et al. [64] proposed to learn a soft multilabel for each unlabeled target person image to learn deep soft multi-label reference learning (MAR). Wu et al. [58] proposed a progressive framework, which consists of CNN model jointed training step and label estimation step. Feng et al. [13] proposed an unsupervised cross-view metric learning method based on assumption that person samples of different views taken from different distribution. Lin et al. [29] proposed a bottom-up clustering (BUC) approach to update CNN model and the relationship among the individual samples. Although these works make efforts on the prediction of pseudo-labels, they do not take full advantage of the labeled source data as supervised learning during training. To address this problem, many methods proposed to refine model with both labeled source data and unlabeled target data. These works [25, 28, 37, 50] aimed at learning domain-invariant features. Peng et al. [37] presented a multi-task dictionary to learn a view-invariant representation and Li et al. [25] proposed to learn a share space between the source domain and the target domain under a deep learning framework. Lin et al. [28] tried to align the mid-level feature across datasets in the task of attribute learning, while Wang et al. [50] presented a deep Re-ID model to represent an attribute-semantic and identity-discriminative feature space. The above methods aim at bridge domain gap between source domain and target domain, while overlook the intra-domain difference caused by different cameras in the source domain and target domain. Different from these models, we propose the CLC to diminish both the inter-domain bias and intra-domain difference, and design a dual classification loss to strengthen the supervision from transferred knowledge.

Image-Image Domain Adaptation for Re-ID. Image-image domain adaptation aims at generating a new dataset that connects the source domain and the target domain on image-level. A number of methods [10, 54, 72–74] have studied image-image translation based on domain gap for person Re-ID. Deng et al. [10] proposed a Similarity Preserving cycle consistent Generative Adversarial Network (SPGAN) to create a new dataset which preserved the underlying ID information during image-image translation and proposed a "learning via translation" framework for cross-domain which has been widely used. Wei et al. [54] presented a Person Transfer Generative Adversarial Network (PTGAN) to translate the foreground of image in order to preserve person ID better. The objective of these methods is to narrow down the domain gap in the cross-dataset person Re-ID model, but they ignore the intra-domain variations in the target domain.

To reduce the intra-domain difference, both [73] and [72] trained several style transfer models between different cameras in a dataset, while the former employed Label Smoothing Regularization (LSR) loss to train a person Re-ID model and the latter utilized a triplet loss to train a cross-dataset person Re-ID model. Zhong et al. [74] proposed to investigate the intra-domain variations, i.e. three types of the underlying invariance on target domain. These methods make an improvement on the re-ID performance. However, they only focus on the intra-domain difference in the target domain. In contrast, the proposed CLC develops an imitated target domain transferred from the source dataset and a pseudo target domain transferred from the target dataset, based on which both the inter-domain bias and the intra-domain difference between the source and target domain are addressed.

3 Proposed method

Problem definition. For unsupervised domain adaptation in person Re-ID, a source dataset $\mathcal{S} = (X^s, Y^s, C^s)$ with labeled image-camera pairs and another unlabeled dataset $\mathcal{T} = (X^t, C^t)$ from the target domain are provided. The source dataset consists of N^s images denoted by $X^s = \{x_i^s\}_{i=1}^{N^s}$, and each image x^s corresponds to an identity label $y^s \in \{1, 2, \dots, P^s\}$ (i.e. a total of P^s different persons) and a camera ID $c^s \in \{1, 2, \dots, M^s\}$ (i.e. a total of M^s different cameras). The target dataset contains N^t images represented by $X^t = \{x_i^t\}_{i=1}^{N^t}$, and they are captured by a total of M^t cameras. Our goal is to leverage on both labeled source training images and unlabeled target training images to learn a Re-ID model that generalizes well during the test process in the target domain.

The pipeline of the proposed CLC is described in Fig. 1, which is a two-stage method. In the first stage, we employ StarGAN as the backbone of the imitate model because it can learn the image-image translation from two-domain with multiple cameras through one time training. In the second stage, we employ ResNet-50 [17] pre-trained on the ImageNet [9] as the backbone of our Re-ID model due to its rich feature representations. We discard the last 1000-dim fully connected(FC) layer and add three more FC layers to learn the representations. The output of the first FC layer is 1024-dim named as "FC-1024", followed by the second FC layer with a dimension of P^s (named as "FC-#ID"), where P^s is the number of identities in the labeled source training set. The third layer is connected with "FC-1024" as well but yields a 128-dim feature map (named as "FC-128").

As shown in Fig. 1, the first stage is to train an imitate model between the source domain and the target domain across camera views, by which two new domains including imitated target domain and pseudo target domain are generated in order to make up the identity labels and pairwise labels. The second stage consists of two branches. The first branch "FC-#ID" is to learn discriminative representations across domains and bridge the inter-domain bias based on classification task constrained by a dual classification loss, and the second branch "FC-128" is to mine the latent commonality and lessen intra-domain difference across different domains in the class-style space restricted by a triplet loss.

3.1 Supervised learning for person Re-ID

To obtain a good performance in person Re-ID task, the prime goal is to learn discriminative representations to distinguish person identities. With labeled images $\mathcal{S} = \{X^s, Y^s\}$, an effective strategy is to adopt the ID-discriminative embedding (IDE) borrowed from the classification task [69]. The first branch "FC-#ID" of the proposed person Re-ID model is treated as a classification task and employs the cross-entropy loss $\mathcal{L}^{S}_{Class}$ on the source dataset as described in Eq. (1). The IDE-based model [69] does achieve a very

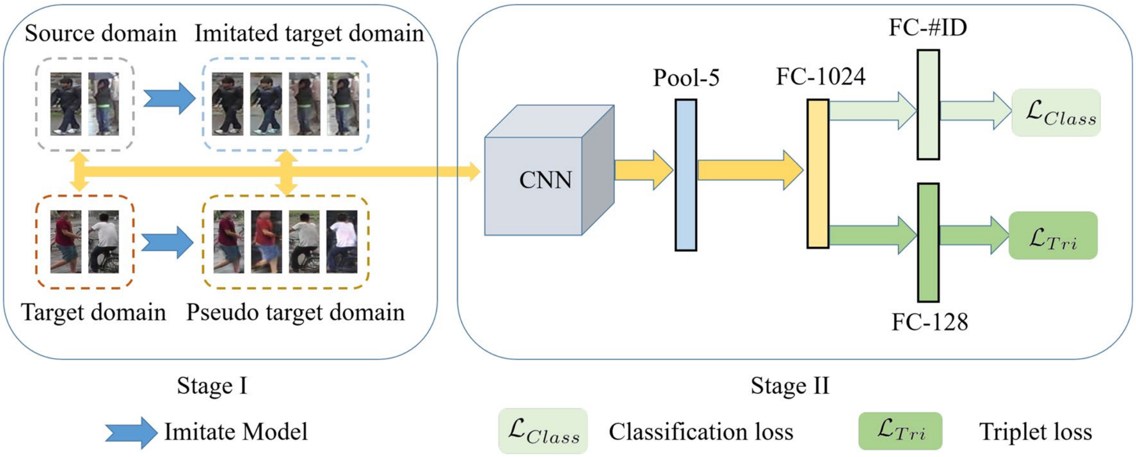

Fig. 1 The pipeline of CLC. It is divided into two stage. The first stage is to train an imitate model between the source domain and the target domain across camera views, which generates labeled imitated target domain to make up the identity labels and unlabeled pseudo target domain to compensates the pairwise labels. The second stage is to learn discriminative representations across domains and reduce the inter-domain bias by formulating a dual classification task on source domain and labeled imitated target domain, and to exploit the latent commonality and lessen the intra-domain difference across source domain, labeled imitated target domain and pairwise label target domain (composed of unlabeled pseudo target domain and unlabeled target domain) in class-style space

good performance on the single person Re-ID dataset, but it fails in the cross-domain person Re-ID problem [72]. In terms of this issue, we propose an imitate model to enhance the generalization ability.

$$\mathcal{L}^{\mathcal{S}}_{Class} = -\sum_{i=1}^{N^s} y_i^s \cdot \log \hat{y}_i^s, \tag{1}$$

where $\hat{y}^s$ is a predicted label on image $x^s \in X^s$ with ground truth y^s.

3.2 Imitate model: inter-dataset bias and intra-dataset difference

The inter-dataset bias caused by different domains is a critical factor that declines the generalization ability of unsupervised person Re-ID models. In other words, without the identity label information on the target domain as the learning guidance, it is very challenging to learn the identity discriminative information as supervised methods did. As we have none information on the target dataset, e.g., identities of people and styles of images, how to transfer information from dataset with labels is the key. On the other hand, the intra-dataset difference induced by different cameras in the target dataset is also a crucial factor, because in the test procedure images of the same person usually come from different cameras of the target domain. That is to say, without the pairwise labels on target domain, it is hard to learn commonality of cameras. Consequently, if transfer learning is employed in the unsupervised person Re-ID problem, how to narrow down the inter-dataset bias and reduce the intra-dataset difference simultaneously is a significant issue, and we propose the imitate model to address this issue.

To bridge the inter-domain gap, we propose to generate imitated target dataset by the learned imitate model, which is denoted by $\mathcal{ST} = \mathcal{S} \rightarrow \mathcal{T}$. Specifically, images in the source domain are adapted to imitate all camera views of the target domain, and thus images of the generated imitated target domain preserve the person identity of the source domain and reflect the style of different cameras in the target domain. This compensates for the absence of identity labels on the target domain. The imitated target domain $\mathcal{ST}$ is further elaborated in Sect. 3.3. To diminish the intra-domain difference, with learned imitate model, we also propose to develop a pseudo target dataset, denoted by $\mathcal{TT} = \mathcal{T} \rightarrow \mathcal{T}$, which diversifies camera styles for each image in the target domain to make up for the lack of pairwise labels. In particular, images in the target domain are transferred to all the camera styles of the target domain. And we clarify more details about $\mathcal{TT}$ in Sect. 3.4.

To train the imitate model, we follow the CamStyle [73] approach to generate new images that preserve the person ID and reflect the style of other camera views across domains. Image generation models are widely employed in person Re-ID area. For example, CycleGAN [75] is utilized to do image-image domain adaptation [10], and StarGAN [8] is employed to construct camera style transfer model [72, 74]. AttGAN [19] and RelGAN [56] are utilized in image-to-image translation. As shown in Fig. 2, we build the imitate model based on StarGAN. This is because StarGAN only requires one time training for image-image translation on two-dataset with multiple cameras, while CycleGAN, AttGAN and RelGAN require multiple translation models for each pair of camera views between the source and target dataset. Compared with previous works that utilize StarGAN to learn camera variance for the target domain, our method employs StarGAN to learn both domain gaps between source and target domain and the camera variance for the target domain, which is more advanced. Examples of real images and fake images generated by the imitate model (i.e. StarGAN) are displayed in Fig. 3, which shows the preservation of the identity of source images by the imitate model.

3.3 Semi-supervised learning for person Re-ID

Denote the imitated target dataset as $\mathcal{ST} = \{X^{st}, Y^{st}, C^{st}\}$, which is constructed by image-image translation for every camera pair from the source domain to the target domain through the imitate model. The dataset consists of N^{st} images represented by x^{st} with corresponding identity label $y^{st} \in \{1, 2, \ldots, P^{st}\}$ (i.e. a total of P^{st} different persons) and

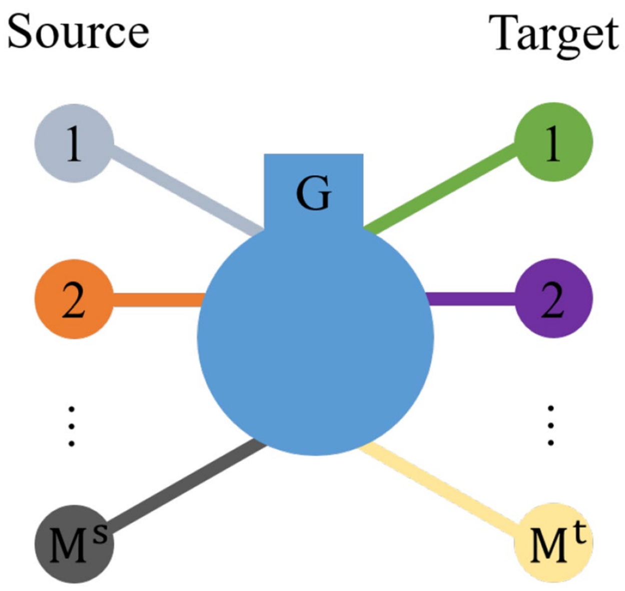

Fig. 2 Imitate model. The cameras in the source domain are represented by 1, 2,..., M_s and the cameras in the target domain are expressed by 1, 2,..., M_t. The imitate model G learns the styles of different cameras from the source domain to the target domain by only one time training

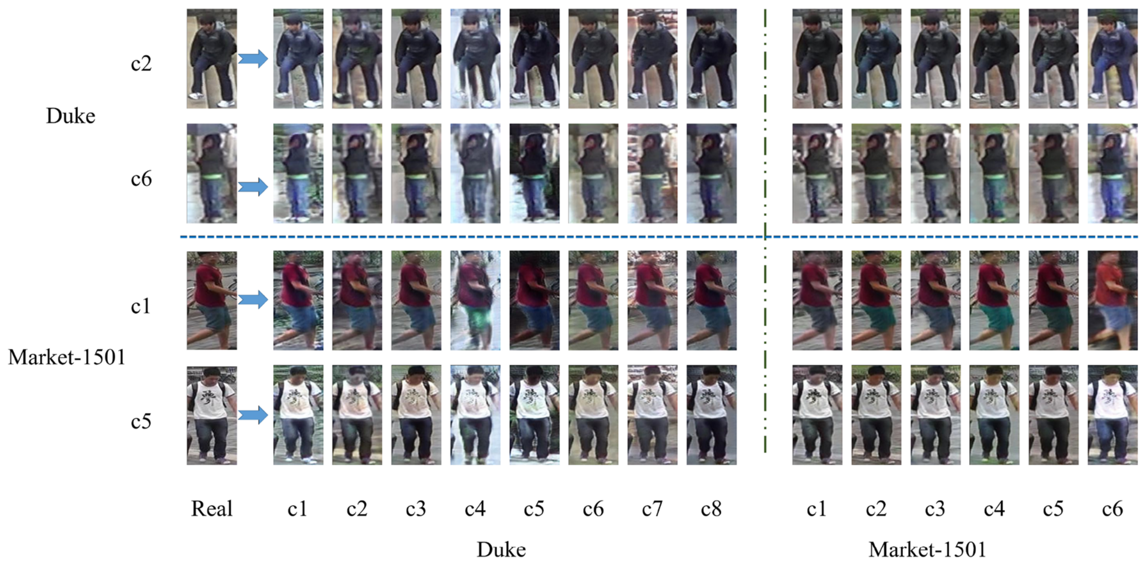

Fig. 3 Examples of image transfer on DukeMTMC-reID and Market-1501 by the imitate model. An image captured by a certain camera is transferred to views of all the cameras in both datasets. The imitate model preserves the content and identity of the source image while reflects the style of the target view

camera ID $c^{st} \in \{1, 2, \dots, M^{st}\}$ cameras (i.e. a total of M^{st} cameras), and it preserves person identities with the source dataset. Specifically, for a real image $x^{s}_{i,j}$ (i.e. the i-th image under the j-th camera) in the source dataset, we generate M^{t} imitated images $x^{st}_{i,1}, x^{st}_{i,2}, \dots, x^{st}_{i,M^t}$ via the learned imitate model. These images preserve the person identity y^{s}_{i} but their styles are similar to their corresponding target cameras $c^{t}_{1}, c^{t}_{2}, \dots, c^{t}_{M^t}$, respectively. Therefore, we have $N^{st} = N^{s} \cdot M^{t}, Y^{st} = Y^{s}, P^{st} = P^{s}, C^{st} = C^{t}, M^{st} = M^{t}$. Note that, imitated target dataset makes up for the absence of identity labels on target domain, which greatly benefits the person Re-ID performance.

Generally, an approach based on supervised learning can perform better than the corresponding unsupervised learning method for the same person Re-ID problem as the former encodes more information than the latter. Therefore, in order to boost the cross-domain person Re-ID performance, we propose to view the unsupervised person Re-ID as a semi-supervised person Re-ID task by imitating the target domain. As shown in Fig. 4, given labeled source training samples $\mathcal{S}$ and unlabeled target training samples $\mathcal{T}$, we semi-supervise the Re-ID model on the imitated target domain $\mathcal{ST}$ that is transferred from $\mathcal{S}$ to $\mathcal{T}$, and a cross-entropy loss on domain $\mathcal{ST}$ is formulated as described in Eq. (2).

$$\mathcal{L}^{\mathcal{ST}}_{Class} = -\sum_{i=1}^{N^{st}} y^{st}_{i} \cdot \log \hat{y}^{st}_{i}, \tag{2}$$

where $\hat{y}^{st}$ is a predicted label for the image $x^{st} \in X^{st}$ with the ground truth y^{st}.

Further, the dual classification loss $\mathcal{L}^{Dual}_{Class}$ is designed to bridge the inter-domain bias as follows:

$$\mathcal{L}^{Dual}_{Class} = \frac{1}{2}(\mathcal{L}^{\mathcal{S}}_{Class} + \mathcal{L}^{\mathcal{ST}}_{Class}). \tag{3}$$

3.4 Mining commonality

In Sect. 3.2, with the imitate model trained on the source domain and the target domain, we actually create two new domains, $\mathcal{ST}$ and $\mathcal{TT}$, where the former is described in Sect. 3.3. The pseudo target domain $\mathcal{TT}$ is built by the image-image translation for every camera pair from the target domain to itself. The pseudo target domain $\mathcal{TT} = \{X^{tt}, C^{tt}\}$ consists of N^{tt} images, where each image x^{tt} corresponds to a camera ID $c^{tt} \in \{1, 2, \dots, M^{tt}\}$ (i.e. a total of M^{tt} cameras), which preserves the same identity with the target domain. In particular, with the learned imitate model, for a real image $x^{t}_{i,j}$ (i.e. i-th image under the j-th camera) in the target domain, a total of M^{t} pseudo images $x^{tt}_{i,1}, x^{tt}_{i,2}, \dots, x^{tt}_{i,M^t}$ are generated. These images hold the person identity with the original images but their styles are similar to the corresponding target camera styles $c^{t}_{1}, c^{t}_{2}, \dots, c^{t}_{M^t}$, respectively, which means $N^{tt} = N^{t} \cdot M^{t}, C^{tt} = C^{t}, M^{tt} = M^{t}$. Note that the image $x^{tt}_{i,j}$ transferred from itself is also included in the M^{t} pseudo images. The pseudo target domain exactly makes up for the lack of pairwise labels on the target domain.

As mentioned above, the source domain and the target domain have totally different classes and styles, which leads

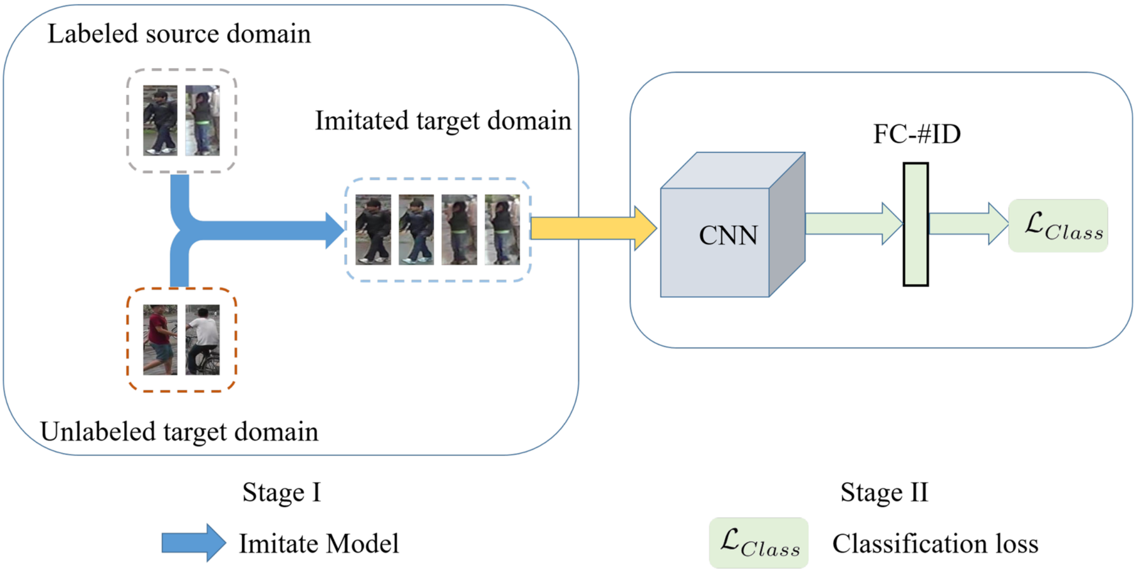

Fig. 4 Illustrations of semi-supervised learning. The proposed model is semi-supervised by the labeled source training samples and the unlabeled target training samples. Specifically, images and the corresponding labels are generated by imitating the unlabeled target domain from the labeled source domain, based on which the Re-ID model is trained by the classification loss

to a limited performance when models trained on the source domain are directly executed on the target domain. That is because models trained on the source domain only learn to extract the camera-invariance image feature in the source domain camera styles to distinguish source classes. The models are unaware of any information on the target classes or target domain camera styles. In other words, if models could exploit the latent commonalities of the source domain and the target domain, a better performance on the target domain could be achieved. Naturally, one of the latent commonalities is camera-invariance that the distance of persons with the same identity in different camera views should be smaller than that of different persons. Based on this intuition, we design a second branch, after embedding-1024 in the Re-ID network, named "FC-128" as shown in Fig. 1. The two branches have different goals: the first branch is a classification task to learn a discriminative image feature, while the second branch is a commonality mining task to acquire more common information of source and target domains and is restricted by a triplet loss as presented in Eq. (4).

$$\mathcal{L}_{Tri}(X) = \sum_{x_a,x_p,x_n} [m + D_{x_a,x_p} - D_{x_a,x_n}]_+, \tag{4}$$

where X represents images in a training batch and x_a, x_p, x_n are images from X. x_a is an anchor point, x_p is a farthest positive sample to x_a, and x_n is a closest negative sample to x_a in X. m is a margin parameter, which is set to 0.3 in our experiments, and $D(\cdot)$ is the Euclidean distance between two images in the commonality feature space. We conduct two types of triplet features: No L2-normalized triplet feature and L2-normalized triplet feature. Note that during Re-ID test process, the feature at pool-5 (2048-dim) layer is utilized as the person descriptor.

To illustrate the commonality of all domains, we view domains in the class-style space where three clusters are formed as shown in Fig. 5. $\mathcal{S}$ denotes the source domain classes and source domain styles, $\mathcal{ST}$ represents source domain classes and target domain styles, and $\mathcal{TT}\searrow\mathcal{T}$ suggests target domain classes and target domain styles. The last cluster $\mathcal{TT}\searrow\mathcal{T} = \{X^{tt\&t}, C^{tt\&t}\}$ is consisted of pseudo target domain $\mathcal{TT}$ and target domain $\mathcal{T}$ (i.e. $X^{tt\&t} = X^{tt} \cup X^t$, $C^{tt\&t} = C^{tt} \cup C^t$), which has pairwise labels of person samples. Specifically, we are not aware of the identity of the person but we do know that $x_1^{tt}, x_2^{tt}, \ldots, x_{M^t}^{tt}$ and $x_{i,j}^t$ belong to the same class, and other images from the target domain can be viewed as a different class. Clearly, such three samples share latent commonality:

$$\mathcal{L}_{Tri}^{\mathcal{S}} = \mathcal{L}_{Tri}(X^s), \tag{5}$$

$$\mathcal{L}_{Tri}^{\mathcal{ST}} = \mathcal{L}_{Tri}(X^{st}), \tag{6}$$

$$\mathcal{L}_{Tri}^{\mathcal{TT}\searrow\mathcal{T}} = \mathcal{L}_{Tri}((X^{tt\&t}), \tag{7}$$

where, in our experiment, on the $\mathcal{L}_{Tri}^{\mathcal{ST}}$, a training batch consists of $n_{st} \times M_t$ images, i.e. we randomly select n_{st} classes and corresponding M_t generated images. And on the $\mathcal{L}_{Tri}^{\mathcal{TT}\searrow\mathcal{T}}$, a training batch consists of $n_t \times (M_t + 1)$, i.e. randomly selecting n_t real images on target domain and the

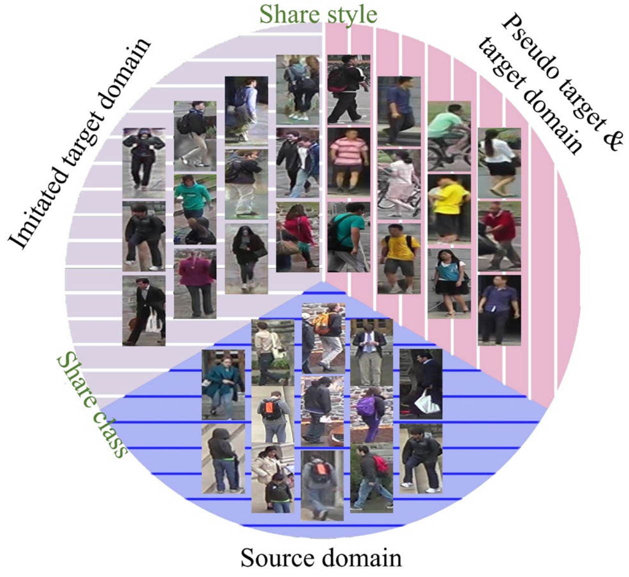

Fig. 5 Three domains in class-style space: source domain $\mathcal{S}$, imitated target domain $\mathcal{ST}$, and pseudo target and target domain $\mathcal{TT}\searrow\mathcal{T}$. Here, the source domain $\mathcal{S}$ and the imitated target domain $\mathcal{ST}$ share the same identities, and the imitated target domain $\mathcal{ST}$ and the pseudo target and target domain $\mathcal{TT}\searrow\mathcal{T}$ possess similar styles. The color of pie chart represents domain. The direction of line represents classes, i.e. identity, and the color of line delineates image styles. Among the three domains, these is a latent commonality, the distance of persons with the same identity in different camera views should be smaller than that of different persons (colour figure online)

corresponding M_t generated images, and we assume that n_t real images belong to different classes. That is because when n_t is small enough, it is a low probability event to select the same person in a training batch.

Consequently, the total loss $\mathcal{L}_{Tri}^{total}$ for the latent commonality task to lessen the intra-domain difference can be written as follows:

$$\mathcal{L}_{Tri}^{total} = \beta_1 \cdot \mathcal{L}_{Tri}^{\mathcal{S}} + \beta_2 \cdot \mathcal{L}_{Tri}^{\mathcal{ST}} + \beta_3 \cdot \mathcal{L}_{Tri}^{\mathcal{TT}\searrow\mathcal{T}}, \tag{8}$$

where $\beta_1, \beta_2, \beta_3$ are hyper-parameters that control the contribution of three clusters on the latent commonality .

Considering both the classification branch and the commonality mining branch, the total training objective $\mathcal{L}_{total}$ of CLC is formulated as follows:

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{Class}^{Dual} + \mathcal{L}_{Tri}^{total} \\ &= \frac{1}{2} \cdot (\mathcal{L}_{Class}^{\mathcal{S}} + \mathcal{L}_{Class}^{\mathcal{ST}}) \\ &\quad + \beta_1 \cdot \mathcal{L}_{Tri}^{\mathcal{S}} + \beta_2 \cdot \mathcal{L}_{Tri}^{\mathcal{ST}} + \beta_3 \cdot \mathcal{L}_{Tri}^{\mathcal{TT}\searrow\mathcal{T}}.\end{aligned} \tag{9}$$

Note that our model is trained in an end-to-end form.

Table 1 Number of training samples with respect to each camera in Market-1501, DukeMTMC-reID and MSMT17 datasets

Market-1501		DukeMTMC-reID		MSMT17	
Camera	# of images	camera	# of images	Camera	# of images
1	2017	1	2809	1	4910
2	1709	2	3009	2	203
3	2707	3	1088	3	454
4	920	4	1395	4	1614
5	2338	5	1685	5	4296
6	3245	6	3700	6	1678
		7	1330	7	3453
		8	1506	8	795
				9	1396
				10	655
				11	3154
				12	1364
				13	3635
				14	3876
				15	1138

4 Experiments

In this section, we conduct studies to examine the effectiveness of each component in the CLC and run cross-domain person Re-ID experiments against a number of state-of-the-arts.

4.1 Datasets

To evaluate the performance of the proposed method, experiments are executed on three widely used person Re-ID datasets: Market-1501 [68], DukeMTMC-reID [38, 70], and MSMT17 [54]. The details on the number of training samples under each camera are presented in Table 1.

Market-1501 [68] collects from 6 camera views, involving 32,668 labeled images of 1501 identities. The dataset consists of two non-over-lapping fixed parts: 12,936 images from 751 identities for training and 19,732 gallery images from the other 750 identities for testing. In testing, 3368 query images from 750 identities are used to retrieve the corresponding person in the gallery.

DukeMTMC-reID [38, 70] (Duke) contains 36,411 labeled images of 1404 identities captured by 8 camera. It is split into two non-over-lapping fixed parts: 16,522 images from 702 identities for training and 17,661 gallery images from the other 702 identities for testing. In testing, 2228 query images from 702 identities are used to retrieve the person in the gallery.

MSMT17 [54] has 15 cameras and 126,441 labeled images belonging to 4101 identities. Similar to the division of DukeMTMC-reID, it is split into two non-over-lapping fixed parts: 32,621 images from 1041 identities for training and 82,161 gallery images from the other 3060 identities for testing. In testing, 11,659 query images from 3060 identities are used to retrieve the person in the gallery.

Note that in all experiments we use labeled training set of the source domain and unlabeled training set of the target domain to learn our model, and examine the performance on the test set of the target domain. The conventional rank-1 accuracy and mAP are adopted as metrics for cross-domain Re-ID evaluation [68] on all three datasets.

4.2 Experimental settings

Imitate model. Given source and target training datasets with camera labels, we employ StarGAN [8] to train an imitate model to transfer images for every camera pair across two datasets. Note that no identity annotation is required during training. The input images are resized to 256×128 in our experiments and Adam optimizer [21] is employed with betas $= (0.5, 0.999)$. Following the update rule in [16], the generator is trained to optimality once after the discriminator parameter updates five times. Note that for each image in these two datasets, a total number of $M^s + M^t$ style-transferred images that preserve the identity of the original image are generated to be used in cross-domain person Re-ID. During training, we only use training samples of source domain and training samples of target domain to generate imitated target domain and pseudo target domain.

CLC. The input images are resized to 256×128, and we initialize the learning rate to 0.01 for the layer pre-trained on ImageNet and to 0.1 for the other layers. The learning rate is multiplied by a factor of 0.1 every 40 epochs and we use SGD optimizer in a total of 60 epochs. The classification task and mining commonality task are trained together. For classification task under supervised and semi-supervised framework, the mini-batch sizes of the source images, imitated target images are set to 64, $12 \times M_t$, and in the mining commonality task, the mini-batch sizes of source images, imitated target images and pseudo target images are set to $4 \times 8, 4 \times M_t, 4 \times M_t$. The involving parameters $\beta_1, \beta_2, \beta_3$ are set to 0.6, 0.6, 0.2. During training, our goal is to utilize the labeled source training samples, labeled imitated target samples, unlabeled pseudo target samples and unlabeled target training samples, to minimize the total loss $\mathcal{L}_{total}$ described on Eq. (9). In the test procedure, 2048-dim (pool-5) features are extracted to compute Euclidean distance between the query and galley images of target testing samples.

4.3 Ablation studies

To highlight the components of the proposed CLC, we conduct experiments to evaluate their contributions to the cross-domain person Re-ID performance. Table 2 reports the comparison results, where Duke→Market-1501 means that Duke is the source domain and Market-1501 is the target domain, and vise versa. Each domain contains its own training set and test set. The performance is always evaluated on the test set of the target domain. In the supervised situation, labels of training set in the target domain are utilized. In contrast, in the unsupervised situation, labels of training set in the target domain are not allowed to be used. Figure 6 shows some Re-ID results on the Market-1501 dataset when using DukeMTMC-reID as the source set. Compared with the supervised model, the person Re-ID performance of our method CLC has been significantly improved.

Comparisons between supervised learning and direct transfer. The supervised person Re-ID model (baseline) which is trained on the target training dataset is evaluated on the target test dataset, and it shows an excellent performance as reported in Table 2. However, a large performance drop can be observed when the model is trained on the source training dataset and tested on the target dataset directly. For instance, the baseline model trained and tested on Market-1501 achieves a rank-1 accuracy of 85.5% and mAP of 66.0%, but declines to 46.0% and 19.1% when it is directly tested on Market-1501. The main reason is the bias of data distributions among domains.

The effectiveness of the semi-supervised learning. Given labeled source training samples and unlabeled target training samples, an imitated target dataset is created by the imitate

Table 2 Ablation studies by using Duke/Market as the source dataset and Market/Duke as the target dataset. $\mathcal{S}$: training set with labels in the source domain. $\mathcal{T}^t$: training set with labels in the target domain. $\mathcal{T}$: training set without labels in the target domain. $\mathcal{ST}$: imitated target set with labels. $\mathcal{TT}$: pseudo target set without labels

Method	Train set	Duke→Market-1501				Market-1501→Duke			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Supervised	$\mathcal{T}^t$	85.5	94.0	96.1	66.0	73.2	84.8	88.2	52.7
$\mathcal{L}_{Class}^{\mathcal{S}}$	$\mathcal{S}$	46.0	63.0	69.7	19.1	29.9	46.2	53.4	15.6
$\mathcal{L}_{Class}^{\mathcal{ST}}$	$\mathcal{ST}$	62.9	80.1	85.7	32.0	45.7	62.3	67.2	23.4
$\mathcal{L}_{Class}^{Dual}$	$\mathcal{S}+\mathcal{ST}$	64.4	81.8	87.4	31.4	47.4	62.6	68.6	24.7
$\mathcal{L}_{Class}^{Dual}+\mathcal{L}_{Tri}^{\mathcal{S}}$	$\mathcal{S}+\mathcal{ST}$	68.1	84.3	89.1	36.1	52.6	67.2	72.4	29.9
$\mathcal{L}_{Class}^{Dual}+\mathcal{L}_{Tri}^{\mathcal{S}}+\mathcal{L}_{Tri}^{\mathcal{ST}}$	$\mathcal{S}+\mathcal{ST}$	68.2	85.0	89.7	37.8	53.1	67.1	71.8	30.0
$\mathcal{L}_{Class}^{Dual}+\mathcal{L}_{Tri}^{\mathcal{S}}+\mathcal{L}_{Tri}^{\mathcal{ST}}+\mathcal{L}_{Tri}^{\mathcal{TT}\searrow\mathcal{T}}$	$\mathcal{S}+\mathcal{ST}+\mathcal{TT}+\mathcal{T}$	72.9	86.2	90.4	40.2	55.5	68.5	73.7	31.6

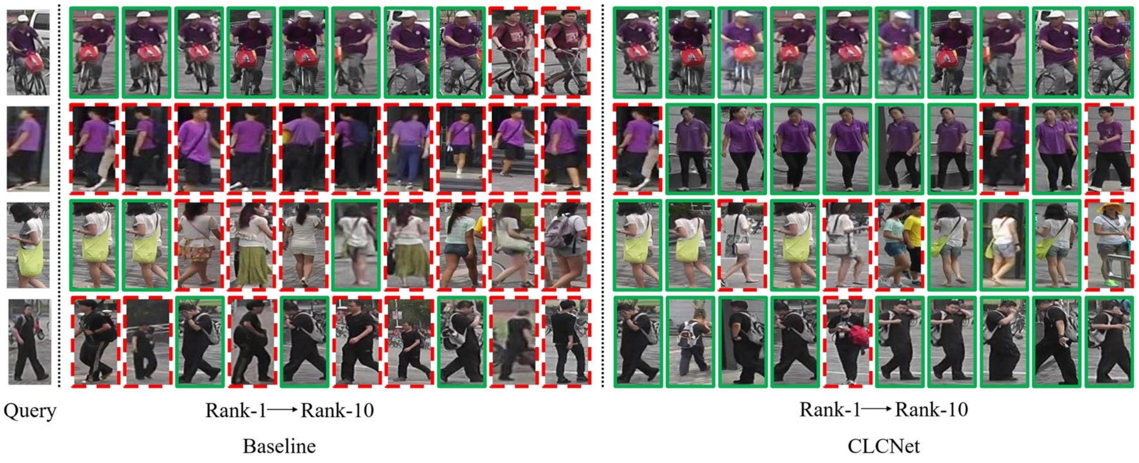

Fig. 6 Sample Re-ID results on Duke→Market-1501. Image in the first column are queries. The images in the second to sixth columns and seventh to eleventh columns are results retrieved by baseline and CLC separately, which are sorted according to their similarity to the query (high to low) from left to right. True matches and false matches are in green solid and red dashed bounding box (colour figure online)

model. It preserves the identity with the source dataset and at the same time reflects the camera style of the target dataset. And we formulate a classification loss $\mathcal{L}^{\mathcal{ST}}_{Class}$ to learn a discriminate feature on target domain and the dual classification loss $\mathcal{L}^{Dual}_{Class}$ extracts a domain-invariant feature as an open set domain adaptation to bridge inter-domain bias. As reflected in Table 2, the performances of objectives $\mathcal{L}^{\mathcal{ST}}_{Class}$ and $\mathcal{L}^{Dual}_{Class}$ are consistently improved in all settings. Compared to the direct transfer method, the proposed semi-supervised method $\mathcal{L}^{\mathcal{ST}}_{Class}$ obtains an improvement of +16.0% in rank-1 accuracy and +12.9% in mAP on Market-1501, and +15.8% in rank-1 accuracy and +7.8% in mAP on Duke. Furthermore, compared with the semi-supervised methods, the proposed dual classification loss $\mathcal{L}^{Dual}_{Class}$ obtains an improvement of +1.5% in rank-1 accuracy on Market-1501 and +1.7% in rank-1 accuracy on Duke. This demonstrates the effectiveness of the proposed semi-supervised formulation and the dual classification loss.

The effectiveness of commonality mining. A pseudo target dataset $\mathcal{TT}$ that is transferred from the target dataset to the target dataset via the imitate model is generated. The dataset $\mathcal{TT}\searrow\mathcal{T}$ is composed by the pseudo target dataset and the target dataset and the triplet loss is constrained over three datasets $\mathcal{S}, \mathcal{ST}$ and $\mathcal{TT}\searrow\mathcal{T}$ to capture the commonality over them in the class-style space. Our goal is to reduce both the inter-domain bias and the intra-domain difference.

We first compare the re-ID model with or without the commonality mining and the results are presented in Table 2, where we can see that one triplet loss largely improves the performance due to the capture of the commonality on three datasets. For example, when tested on Market-1501, the objective $\mathcal{L}^{Dual}_{Class} + \mathcal{L}^{\mathcal{S}}_{Tri}$ could improve +3.7% in rank-1 accuracy and +4.7% in mAP, and when tested on Duke, it could improve +5.2% in rank-1 accuracy and +5.2% in mAP. The consistent improvements indicate the existence of the latent commonality.

In addition, we also evaluate the impacts of the combination of two triplet losses that capture the latent commonality. As shown in Table 2, the combination of two triplet losses has little influence on the rank-1 and mAP accuracy compared with the solo triplet loss. For instance, compared with single triplet loss on source domain, when tested on Market-1501, the objective $\mathcal{L}^{Dual}_{Class} + \mathcal{L}^{\mathcal{S}}_{Tri} + \mathcal{L}^{\mathcal{ST}}_{Tri}$ achieves 68.2% (+0.1%) at rank-1 accuracy and 37.8% (+1.7%) in mAP, and when tested on Duke, it obtains 53.1% (+0.5%) in rank-1 accuracy and 30.0% (+0.1%) in mAP. For little improvements, we argue that to some extent, $\mathcal{S}$ and $\mathcal{ST}$ has share common information (i.e. identities) so that there is some overlap on two domain's latent commonality.

Finally, we verify the effectiveness of our hypothesis that the latent commonality of three datasets can be captured in the form of triplet loss. It is clear that " $\mathcal{L}_{total} = \mathcal{L}^{Dual}_{Class} + \mathcal{L}^{\mathcal{S}}_{Tri} + \mathcal{L}^{\mathcal{ST}}_{Tri} + \mathcal{L}^{\mathcal{TT}\searrow\mathcal{T}}_{Tri}$ " significantly improved. For instance, compared with two triplet losses, when tested on Market-1501, "$\mathcal{L}_{total}$" obtains a rank-1 accuracy of 72.9% (+4.7%) and mAP of 40.2% (+2.4%) when using Duke as the source dataset. Similar improvements can be observed when tested on DukeMTMC-reID, it obtains a rank-1 accuracy of 55.5% (+2.4%) and mAP of 31.6% (+1.6%). The consistent improvements indicate that the latent commonality is critical to enhance the generalization ability of models.

Normalization and margin of the triplet feature. We further analyze the influences of different margins m in Eq. (4) and types (whether or not normalized by L2) of triplet feature and the results are reported in Table 3 where NoNormalize means no L2-normalized triplet feature and Normalize means L2-normalized triplet feature in Eq. (8).

Table 3 Evaluation on different margins m and types of triplet feature

Margin	Duke→Market-1501				Market-1501→Duke			
	NoNormalize		Normalize		NoNormalize		Normalize	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
0.1	70.8	38.9	67.3	35.6	53.0	29.9	51.6	28.4
0.3	**72.9**	**40.2**	70.6	38.7	**55.5**	**31.6**	**54.6**	**30.4**
0.5	71.1	40.0	71.5	**38.8**	54.1	30.4	53.7	30.3
0.7	70.9	39.6	**71.6**	38.7	53.1	30.1	53.1	30.5
0.9	70.5	39.0	70.1	36.1	53.1	30.3	52.5	29.6

The bold number denotes the best result

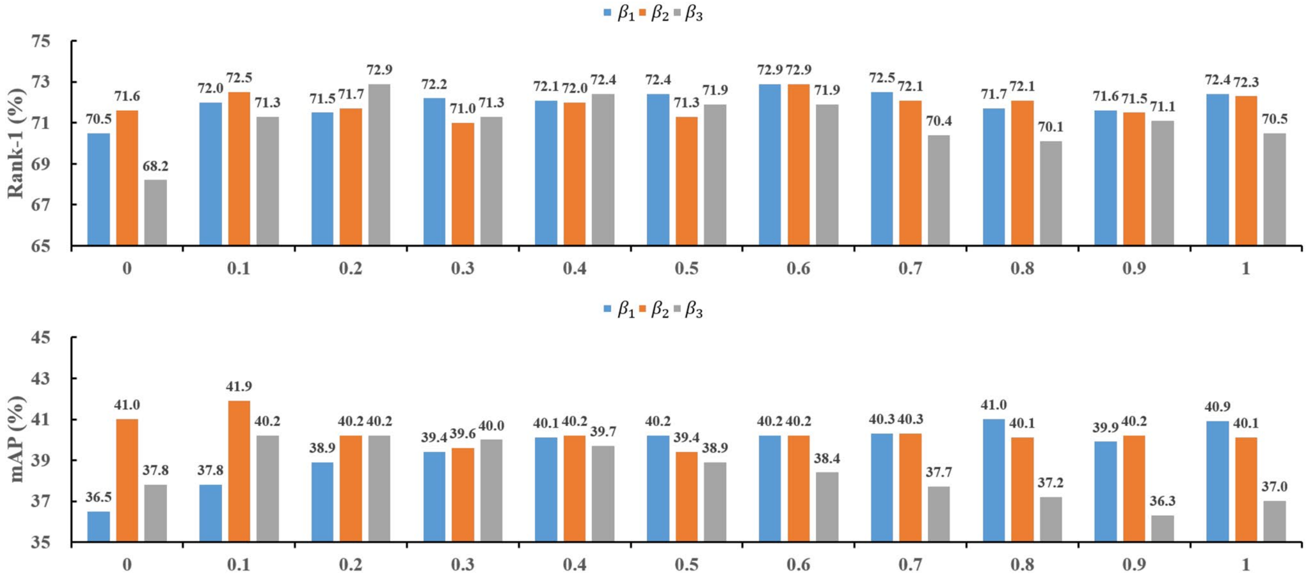

Fig. 7 Sensitivity to parameter $\beta_1, \beta_2, \beta_3$ in Eq. (9) in setting of Duke→Market-1501. When evaluating one parameter, we fix the other two. Specifically, when we evaluate β_1, we fix $\beta_2 = 0.6, \beta_3 = 0.2$. When we evaluate β_2, we fix $\beta_1 = 0.6, \beta_3 = 0.2$. when we evaluate β_3, we fix $\beta_1 = 0.6, \beta_2 = 0.6$

In our experiments, in each type for triplet feature, moderate margin m is best for results, and given margin m, no L2-normalized is always better for L2-normalized in rank-1 accuracy and mAP. We argue that from the perspective of dimensionality reduction, triplet feature that is reduced from 1024-dim to 128-dim is no longer distributed on the sphere, so we should not apply L2-normalized on triplet feature. The best results are produced when margin $m = 0.3$ and using no L2-normalized for triplet feature.

Weights of the triplet loss $\beta_1, \beta_2, \beta_3$**.** We evaluate three important parameters, i.e. the weights of the triplet loss $\beta_1, \beta_2, \beta_3$ in Eq. (9), as shown in Fig. 7. When evaluating one parameter, we fix the other two. The rank-1 accuracy and mAP of model with the dual classification loss $\mathcal{L}^{Dual}_{Class}$ is 64.4% and 31.4% in setting of Duke→Market-1501. It is clearly shown that, our approach significantly improves the model with the dual classification loss at all values. And the performance becomes best when $\beta_1 = 0.6, \beta_2 = 0.6, \beta_3 = 0.2$.

Table 4 computational cost analysis of the triplet feature

Method	Duke→Market-1501		
	R-1	Time (mins)	Memory (MB)
$\mathcal{L}^{S}_{Class}$	46.0	≈ 60	≈ 108.25
$\mathcal{L}_{total}$	72.9	≈ 200	≈ 108.75

The benefit of the triplet feature. As shown in Table 4, the method based on $\mathcal{L}_{total}$ clearly outperforms the method based $\mathcal{L}^{S}_{Class}$, and it is noteworthy that $\mathcal{L}_{total}$ introduces limited additional training time (≈ 140 mins) and GPU memory (≈ 0.5 MB) compared to $\mathcal{L}^{S}_{Class}$.

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

We compare our method against a number of state-of-the-art unsupervised learning methods on Market-1501 and DukeMTMC-reID in Table 5, which reports the results of evaluation when using these two datasets as the source and target domains respectively. The compared methods are categorized into four groups, two hand-crafted methods including LOMO [27] and Bow [68], three unsupervised methods that use a labeled source data to initialize the model and then use a target dataset to fine-tune model including UMDL [37], PUL [11], CAMEL [63], three unsupervised domain adaptation approaches without GAN including TJ-AIDL [50], MMFA [28], CFSM [6], three unsupervised domain adaptation approaches with GAN including PTGAN [54], SPGAN [10] and HHL [72].

The two hand-crafted methods [27, 68] acquire a relative worse accuracy because both of them are directly employed to the target test dataset, and between the dataset used in training and the target test dataset there is a large inter-domain bias. For example, the rank-1 accuracy of LOMO [27] is 27.2% when tested on Market-1501, which is much lower than transfer learning based methods.

In order to overcome this problem, some unsupervised methods [11, 37, 63] that train the model on the labeled source set are proposed and achieve much higher results than the hand-crafted methods. For instance, CAMEL [63] gives 54.5% rank-1 accuracy when trained on DukeMTMC-reID and tested on Market-1501, surpassing LOMO [27] by a large margin of 27.3%.

Comparing with unsupervised domain adaptation methods without GAN [6, 28, 50], the proposed method is preferable. Specifically, when tested on Market-1501, our results outperforms all the other methods, achieving rank-1 accuracy of 72.9% and mAP of 40.2%, which outperforms recent published CFSM [6] by a gain of +11.7% in rank-1 accuracy and +11.6% in mAP. When tested on DukeMTMC-reID, our method achieves a boost of +5.7% in rank-1 accuracy and +4.3% in mAP, which is superior to all the other methods as well.

Lastly, we further compare the proposed method with unsupervised domain adaptation methods using GAN, and the results show that our method is also superior. For instance, when tested on Market-1501, comparing with the recently published HHL [72], we obtain a better performance by a margin of +10.7% in rank-1 accuracy and +8.8% in mAP. When tested on DukeMTMC-reID, the proposed method upgrades the performance by a margin of +8.6% in rank-1 accuracy and +4.4% in mAP.

We also evaluate our approach on a larger and more challenging dataset, i.e. MSMT17 [54]. As shown in Table 6, our approach clearly surpasses PTGAN [54] when using

Table 6 Performance comparisons with state-of-the-art person Re-ID methods using Duke/Market as the source dataset and MSMT17 as the target dataset

Method	Src.	MSMT17			
		R-1	R-5	R-10	mAP
PTGAN [54]	Market-1501	10.2	–	24.4	2.9
Ours	Market-1501	**19.1**	**29.8**	**34.7**	**6.2**
PTGAN [54]	Duke	11.8	–	27.4	3.3
Ours	Duke	**24.0**	**35.2**	**40.7**	**7.8**

The bold number denotes the best result

Table 5 Performance comparisons with state-of-the-art person Re-ID methods using Duke/Market as the source dataset and Market/Duke as the target dataset

Method	Duke→Market-1501				Market-1501→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [27]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
UMDL [37]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
Bow [68]	35.8	52.4	60.3	14.8	17.2	28.8	34.9	8.3
PTGAN [54]	38.6	–	66.1	–	27.4	–	50.7	–
PUL [11]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
SPGAN [10]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
CAMEL [63]	54.5	–	–	26.3	–	–	–	–
MMFA [28]	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
SPGAN+LMP [10]	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
TJ-AIDL [50]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CFSM [6]	61.2	–	–	28.3	49.8	–	–	27.3
HHL [72]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
Ours	**72.9**	**86.2**	**90.4**	**40.2**	**55.5**	**68.5**	**73.7**	**31.6**

The bold number denotes the best result

Market-1501 and DukeMTMC-reID as source domains. For example, our method achieves rank-1 accuracy = 24.0% and mAP = 7.8% when using DukeMTMC-reID as source set, which get a boost of +12.2% in rank-1 accuracy and +4.5% in mAP.

5 Conclusion

In this work, we present the Classification and Latent Commonality method (CLC) method to solve the unsupervised person Re-ID problem. To make up the absence of identity labels, we generate an imitated target domain by an imitate model, and to compensate the pairwise labels across camera views, a pseudo target domain is created. We further propose a dual classification loss on both the source domain and the imitated target domain to learn a discriminative representation and bridge the inter-domain bias. To investigate the camera-invariance and diminish the intra-domain difference, triplet loss constrained on the source domain, imitated target domain and pairwise label target domain (composed of pseudo target domain and target domain) is exploited. Experiments are conducted on Market-1501 and DukeMTMC-reID, and experimental results demonstrate that the proposed architecture outperforms numerous state-of-the-art approaches.

Acknowledgements This work was supported the Natural Science Foundation of China (61972027) and the Beijing Municipal Natural Science Foundation (Grant no. 4212041).

References

1. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3908–3916
2. Bai S, Bai X (2016) Sparse contextual activation for efficient visual re-ranking. IEEE Trans Image Process 25(3):1056–1069
3. Baktashmotlagh M, Faraki M, Drummond T, Salzmann M (2018) Learning factorized representations for open-set domain adaptation. arXiv preprint arXiv:180512277
4. Bazzani L, Cristani M, Murino V (2013) Symmetry-driven accumulation of local features for human characterization and re-identification. Comput Vis Image Underst 117(2):130–144
5. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. Mach Learn 79(1–2):151–175
6. Chang X, Yang Y, Xiang T, Hospedales TM (2018) Disjoint label space transfer learning with common factorised space. arXiv preprint arXiv:181202605
7. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 403–412
8. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
10. 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: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003
11. Fan H, Zheng L, Yan C, Yang Y (2018) Unsupervised person re-identification: clustering and fine-tuning. ACM Trans Multimed Comput (TOMM) 14(4):83
12. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer society conference on computer vision and pattern recognition, IEEE, pp 2360–2367
13. Feng Y, Yuan Y, Lu X (2021) Person re-identification via unsupervised cross-view metric learning. In: IEEE Transactions on Cybernetics, vol 51, pp 1849–1859. https://doi.org/10.1109/TCYB.2019.2909480
14. Geng S, Yu M, Liu Y, Yu Y, Bai J (2019) Re-ranking pedestrian re-identification with multiple metrics. Multimed Tools Appl 78(9):11631–11653
15. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision, Springer, pp 262–275
16. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. In Advances in neural information processing systems. Springer, New York, pp 5767–5777
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
18. He Z, Cheolkon J, Qingtao F, Zhendong Z (2018) Deep feature embedding learning for person re-identification based on lifted structured loss. Multimedia tools and applications. Springer, New York, pp 1–18
19. He Z, Zuo W, Kan M, Shan S, Chen X (2019) Attgan: Facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478
20. Kalayeh MM, Basaran E, Gökmen M, Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1062–1071
21. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
22. Kodirov E, Xiang T, Gong S (2015) Dictionary learning with iterative Laplacian regularisation for unsupervised person re-identification. In: BMVC, vol 3, p 8
23. Leng Q, Hu R, Liang C, Wang Y, Chen J (2015) Person re-identification with content and context re-ranking. Multimed Tools Appl 74(17):6989–7014
24. Li W, Zhu X, Gong S (2018a) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
25. Li YJ, Yang FE, Liu YC, Yeh YY, Du X, Frank Wang YC (2018b) Adaptation and re-identification network: an unsupervised deep transfer learning approach to person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 172–178
26. Lian Q, Li W, Chen L, Duan L (2019) Known-class aware self-ensemble for open set domain adaptation. arXiv preprint arXiv:190501068

27. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2197–2206
28. 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
29. Lin Y, Dong X, Zheng L, Yan Y, Yang Y (2019) A bottom-up clustering approach to unsupervised person re-identification. Proc AAAI Conf Artif Intell 2:1–8
30. Liu H, Cao Z, Long M, Wang J, Yang Q (2019) Separate to adapt: open set domain adaptation via progressive separation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2927–2936
31. Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017a) Hydraplus-net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision, pp 350–359
32. Liu Z, Wang D, Lu H (2017b) Stepwise metric promotion for unsupervised video person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 2429–2438
33. Long M, Cao Y, Wang J, Jordan MI (2015) Learning transferable features with deep adaptation networks. arXiv preprint arXiv:150202791
34. Long M, Zhu H, Wang J, Jordan MI (2017) Deep transfer learning with joint adaptation networks. In: Proceedings of the 34th International conference on machine learning, volume 70, JMLR.org, pp 2208–2217
35. Ma B, Su Y, Jurie F (2014) Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis Comput 32(6–7):379–390
36. Panareda Busto P, Gall J (2017) Open set domain adaptation. In: Proceedings of the IEEE international conference on computer vision, pp 754–763
37. 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: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1306–1315
38. 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, Springer, pp 17–35
39. Rohrbach M, Ebert S, Schiele B (2013) Transfer learning in a transductive setting. In Advances in neural information processing systems. Springer, New York, pp 46–54
40. Saito K, Yamamoto S, Ushiku Y, Harada T (2018) Open set domain adaptation by backpropagation. In: Proceedings of the European conference on computer vision (ECCV), pp 153–168
41. Sener O, Song HO, Saxena A, Savarese S (2016) Learning transferrable representations for unsupervised domain adaptation. In Advances in neural information processing systems. Springer, New York, pp 2110–2118
42. Shu R, Bui HH, Narui H, Ermon S (2018) A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv:180208735
43. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: The European conference on computer vision (ECCV)
44. Tan S, Jiao J, Zheng WS (2019) Weakly supervised open-set domain adaptation by dual-domain collaboration. arXiv preprint arXiv:190413179
45. Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T (2014) Deep domain confusion: maximizing for domain invariance. arXiv preprint arXiv:14123474
46. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7167–7176
47. Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016a) Joint learning of single-image and cross-image representations for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1288–1296
48. Wang G, Lin L, Ding S, Li Y, Wang Q (2016b) Dari: distance metric and representation integration for person verification. In: Thirtieth AAAI conference on artificial intelligence
49. Wang H, Gong S, Xiang T (2014a) Unsupervised learning of generative topic saliency for person re-identification. In: Proceedings of the British machine vision conference (BMVC)
50. Wang J, Zhu X, Gong S, Li W (2018) Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2275–2284
51. Wang Q, Gao J, Li X (2019a) Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. IEEE Trans Image Process 28(9):4376–4386
52. Wang Q, Gao J, Lin W, Yuan Y (2019b) Learning from synthetic data for crowd counting in the wild. In: The IEEE conference on computer vision and pattern recognition (CVPR)
53. Wang T, Gong S, Zhu X, Wang S (2014b) Person re-identification by video ranking. In: European conference on computer vision, Springer, pp 688–703
54. Wei L, Zhang S, Gao W, Tian Q (2018a) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
55. Wei L, Zhang S, Yao H, Gao W, Tian Q (2018b) Glad: Global-local-alignment descriptor for scalable person re-identification. IEEE Trans Multimed 21(4):986–999
56. Wu PW, Lin YJ, Chang CH, Chang EY, Liao SW (2019a) Relgan: Multi-domain image-to-image translation via relative attributes. In: Proceedings of the IEEE international conference on computer vision, pp 5914–5922
57. Wu Y, Lin Y, Dong X, Yan Y, Ouyang W, Yang Y (2018) Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5177–5186
58. Wu Y, Lin Y, Dong X, Yan Y, Bian W, Yang Y (2019b) Progressive learning for person re-identification with one example. IEEE Trans Image Process 28(6):2872–2881
59. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1249–1258
60. Xu X, Li W, Xu D (2015) Distance metric learning using privileged information for face verification and person re-identification. IEEE Trans Neural Netw Learn Syst 26(12):3150–3162
61. Ye M, Liang C, Yu Y, Wang Z, Leng Q, Xiao C, Chen J, Hu R (2016) Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Trans Multimed 18(12):2553–2566
62. Ye M, Ma AJ, Zheng L, Li J, Yuen PC (2017) Dynamic label graph matching for unsupervised video re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 5142–5150
63. Yu HX, Wu A, Zheng WS (2017) Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 994–1002
64. Yu HX, Zheng WS, Wu A, Guo X, Gong S, Lai JH (2019) Unsupervised person re-identification by soft multilabel learning. In:

Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2148–2157

65. Yuan Y, Zhang J, Wang Q (2020) Deep Gabor convolution network for person re-identification. Neurocomputing 378:387–398
66. Zhao R, Ouyang W, Wang X (2014) Learning mid-level filters for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 144–151
67. Zhao R, Oyang W, Wang X (2017) Person re-identification by saliency learning. IEEE Trans Pattern Anal Mach Intell 39(2):356–370. https://doi.org/10.1109/TPAMI.2016.2544310
68. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
69. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv preprint arXiv:161002984
70. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision, pp 3754–3762
71. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1318–1327
72. Zhong Z, Zheng L, Li S, Yang Y (2018a) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European conference on computer vision (ECCV), pp 172–188
73. Zhong Z, Zheng L, Zheng Z, Li S, Yang Y (2018b) Camera style adaptation for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5157–5166
74. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 598–607
75. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.