The Visual Computer (2022) 38:1871–1882
https://doi.org/10.1007/s00371-021-02246-8

ORIGINAL ARTICLE

Source-free unsupervised multi-source domain adaptation via proxy task for person re-identification

Yi Ding[1] · Zhikui Duan[2] · Shiren Li[3]

Accepted: 6 July 2021 / Published online: 31 July 2021

Abstract
Most of existing unsupervised domain adaptation methods focus on aligning the feature discrepancy between labeled source and unlabeled target data. However, in practice, the source data may not be accessible due to transfer issue, privacy problem, etc. In such case, transferring knowledge through only the trained source model without the labeled source data remains a challenging problem. In this paper, we propose a novel method for source-free (without accessing any source domain data) multi-source domain adaptation in person re-identification (Re-ID). Two proxy tasks including proxy label learning and domain discriminative learning are designed to transfer knowledge from source models to the target domain with the inspiration of self-supervised learning. In the proxy label learning process, a subset of the unlabeled data in the target domain is randomly selected to trained a proxy label generator for measuring the similarity between each sample and the selected subset. With the proxy label as input, in the domain discriminative learning process, a domain discriminator is learnt to assign weights for measuring similarity between each source domain and the target one. With the combination of these two proxy tasks, knowledge from multiple source models can be properly aggregated and adaptively transferred to the target domain without any source data. Extensive evaluations on benchmark datasets, which are DukeMTMC and Market-1501, demonstrate the superior performance of the proposed method.

Keywords Multi-source domain adaptation · Person re-identification · Transfer knowledge

1 Introduction

Person re-identification (Re-ID) aims at matching images of the same identity across different cameras views, which plays an important role in the intelligent surveillance systems. Despite many researches [5,17,22,36,53,54] have made great progress in supervised learning manner, it is still difficult to learn a Re-ID model that generalizes well on a target domain without annotations. One main reason is that the data distribution discrepancy between source and target domains, caused by the variations such as body pose, camera view, illumination, image resolution, occlusion and background.

To address the unsupervised domain adaptation (UDA) problem on Re-ID, some recent works [6,38,42,47,56,57] focus on transferring the knowledge from labeled source dataset or clustering on the target unlabeled dataset. When extending to a new target domain, most of UDA Re-ID approaches require source domain data for pre-training or joint training. However, in some practical scenario, source domain data cannot be obtained because of privacy problem or transfer problem, etc. This requirement of source data reduces scalability and usability of these UDA Re-ID methods. In such case, directly transferring the knowledge from learned models trained on labeled source domains to unlabeled target domains remains a meaningful challenge, as illustrated in Fig. 1.

There are a few studies [43] on source-free (without any source data) multi-source domain adaptation on unsupervised Re-ID. In Distill [43], this problem was considered as a knowledge distillation with multi-teacher and developed sample pairwise similarity matrix for target model to imitate the source models. However, using sample pairwise similarity matrix to transfer knowledge suffers from two limitations. On the one hand, the knowledge from source models depend on the quality of sampled batch data. When there are some noisy samples in training batch, the sample pairwise simi-

✉ Yi Ding
mrtbs99@163.com

1 Hunan University of Arts and Science, Changde, China

2 Foshan University, Foshan, China

3 Sun Yat-Sen University, Guangzhou, China

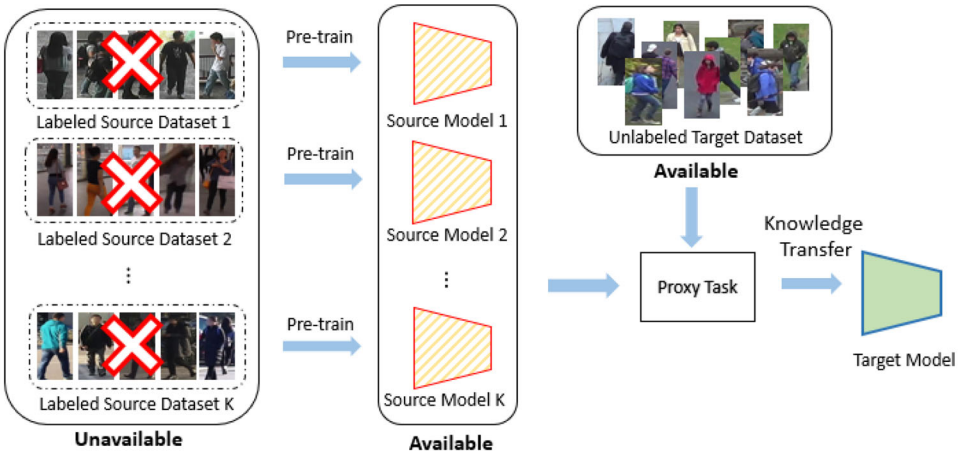


Fig. 1 Illustration of source-free multi-source domain adaptation problem setting. Using pre-trained source models and unlabeled data, the knowledge from labeled source domain can be transferred to target domain via proxy task, which improve the scalability and reusability of Re-ID system

larity matrix may not provide effective guiding information for target model learning. On the other hand, the weighting strategy for source models based on sample pairwise similarity matrix is coarse-grained, because of the limited number of batch data.

To tackle the above problems, we introduce a novel proxy task learning framework to transfer knowledge from source models. The proposed proxy task is inspired by recent self-supervised learning methods [3,13,49] which construct pretext tasks by discovering supervisory signals directly from the input data itself and learn useful visual representation from pretext task. Similarly, our proxy task also defines the supervisory signals from unlabeled data itself. Specifically, our proxy task includes two parts, which are proxy label learning and domain discriminative learning, shown in Fig. 2.

For each image, the features extracted by different source models contain various source information in proxy label learning. We train a classifier (proxy label generator) to distinguish from which image is the input feature. In this way, the supervisory signal for proxy label generator is the data itself, since each image for proxy task is regarded as an individual identity (category). When we train process converge, the output probability distribution vectors from proxy label generator are used as proxy labels. Intuitively, each entry of proxy label generator can be viewed as a prototype of image feature, since the proxy label generator is constructed by a simple full connection layer and optimized by cross-entropy loss. Thus, each element of proxy label vector represents the similarity between the input data and the corresponding image sampled for proxy label learning. By learning the proxy label, the target model can learn the knowledge embedded in source models, which relates to the similarity among unlabeled data.

Meanwhile, in domain discriminative learning, a domain discriminator is trained to distinguish from which source model is the input proxy label. The domain discriminative learning aims to estimate the discrepancy among target and different source models through proxy label, since we expect the target model to learn from the more relevant source model. To integrate the knowledge from different source models, we use weighting strategy over different source proxy labels to generate the aggregated proxy label. The weighting strategy is chosen by the domain discriminator, which emphasizes the more similar source model and suppresses the dissimilar one. Finally, the knowledge can be adaptively transferred from multiple source models into target model by learning the aggregated proxy label.

The main contributions of this paper are summarized as follows: First, we propose a novel proxy label learning method to embed the knowledge from multiple source models into proxy label. Second, we propose to use domain discriminator to automatically choose a weighting strategy for effectively aggregating knowledge from different source models by domain discriminative learning. Third, experimental results on DukeMTMC and Market-1501 show that our model outperforms the state-of-the-art source-free multi-source unsupervised Re-ID methods.

2 Related work

2.1 Unsupervised domain adaptation

Unsupervised domain adaptation (UDA) aims at tackling the domain shift [30] problem by transferring knowledge from labeled source domain to unlabeled target domain. The common idea of most single-source UDA methods is aligning the source and target domains in different level. Discrepancy-based methods explicitly measure the feature discrepancy between the source and target domains, such as the variant of maximum mean discrepancies (MMD) [20,21], correlation alignment (CORAL) [34], and adversarial discriminative loss [12,37]. GAN-based (Generative Adversarial Networks) approaches focus on learning the image-to-image transfor-

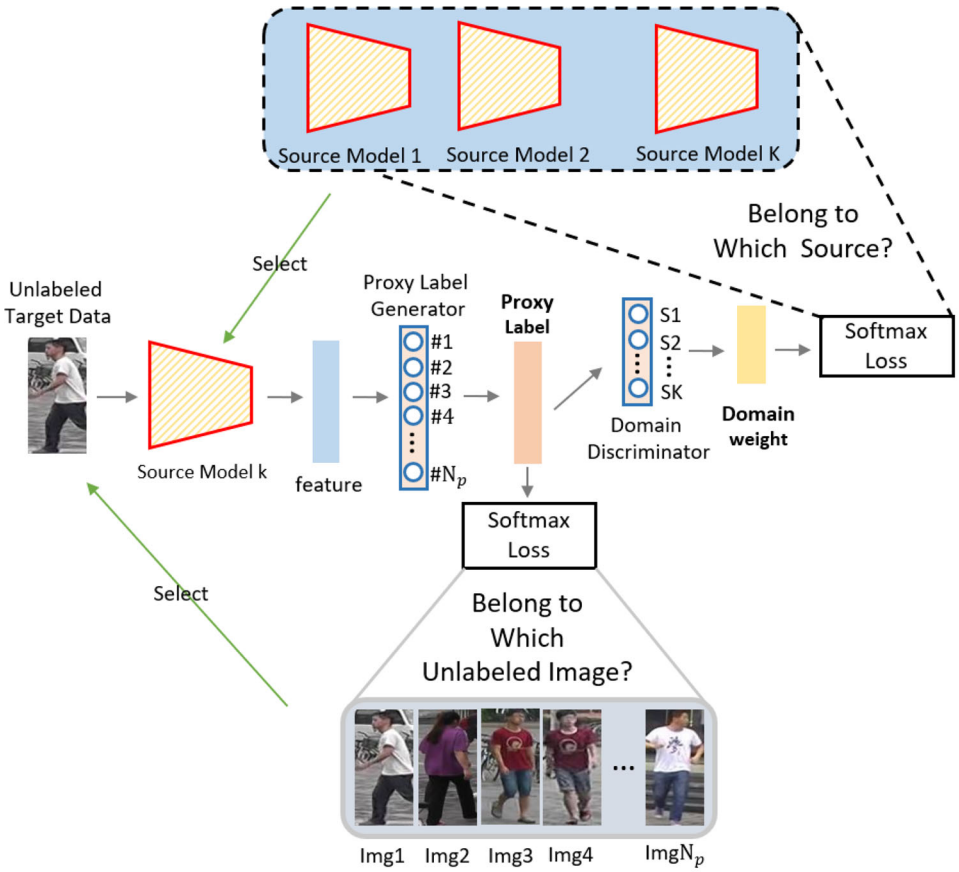


Fig. 2 Illustration of the proposed proxy task which includes proxy label learning and domain discriminative learning. The proxy label learning focuses on learning a classifier, proxy label generator, to distinguish which unlabeled images is the input feature from. The proxy label is the output probability distribution of classifier, while the domain discriminative learning attempts to distinguish which source models is the proxy label from

mation to align the discrepancies in the pixel space, such as PixelDA [4], CyCADA [15]. Multi-source methods assume that the training data are collected from multiple sources [35]. Recent multi-source methods [28,44,50,51] focus on extracting knowledge from source domains data. MDAN [50] and DCTN [44] use domain discriminator to adversarial learning domain. MMN [28] transfers the learned knowledge from multiple sources to the target by dynamically aligning moments of their feature distributions. MDDA [51] proposes a sample selection framework to distill effective knowledge from source data to target domain. SK et al. [1] propose a novel algorithm to find the optimal combinations of source models and this algorithm leads to superior results than source model. Song et al. [33] propose a theoretical foundation about domain adaptive classification theories, which is the first to extend this theory to re-ID tasks. Further, SSG focuses on how to harness the similar natural characteristics in samples from target domain to conduct person re-ID. However, the approaches mentioned above fail to directly apply to person re-ID task. Most of them assume that source and target domains share the same classes, while the person identities (classes) are totally different among different datasets.

2.2 Self-supervised learning

Self-supervised learning (SSL) constructs pretext tasks by discovering supervisory signals directly from the input data itself. A pretext task is designed for solving the problem which requires to learn a useful visual representations. These methods use various cues and pretext tasks, such as in-painting [26], jigsaw puzzles of patch context [8,25], noise-as-targets [3], colorization [16,49], predicting transformations [13,48]. Recently, contrastive learning-based methods also lead to superior result on image classification [14] and video action recognition [39]. Inspired by the idea that the pretext task can be constructed automatically and easily from images alone, we introduce proxy task to extract knowledge from source model by utilizing the supervisory signals from unlabeled images alone. The major difference is we have no available source data and we use the source model as a pseudolabel generator. In addition, the pretext for this work is based on dynamic pseudolabel which is changed with different input image.

2.3 Person Re-ID

2.3.1 Supervised learning methods

Most related works in person re-identification are based on supervised learning. Among them, MPN [7] proposes a novel robust part-aware model that driven by part-level representations and achieves significant improvement on Market-1501. Song et al. focus on adversarial attack with human-imperceptible noise to galley images [2] and further improve the model robustness. Ahmed et al. [1] propose an interesting setting: They focus on exploring the dynamic nature of a camera network and minimizing the additional effort when adapting the existing re-identification models. Besides utilizing the information from the whole image, PAUL. Yang et al. [45] introduce a patch-based unsupervised learning framework to learn discriminative feature from patches.

2.3.2 Unsupervised learning methods

Recent unsupervised person re-ID methods can fall into three categories. The clustering-based methods [9,10,19] focus on clustering images of the same identity to train, which is similar to the supervised method. BUC [9] proposes a bottom-up clustering framework with a diversity regularization. SSG [10] clusters on global and local features and assigns hard pseudolabels. With the aid of Generative Adversarial Networks (GAN), GAN-based methods [6,42,56] aim at learning a image-to-image transformation to close the gap between domains in pixel-level. PTGAN [42] and SPGAN [6] transform source images into target domain style without changing the original person identities label. HHL [56] focuses on camera-invariant learning with camera style transferred images among different domains. The third category of methods [29,47,57] attempts to explore the knowledge of source and target domains itself and proposes the designed constraints to learn a generalized model. Wang et al. [41] focus on the higher-order relationships across the entire camera network and propose a consistent cross-view matching framework. ENC [57] utilizes invariant properties to generalize the model with exemplar memory module. MAR [57] conducts soft label learning with reference persons from source data. UCDA [29] reduces the discrepancy not only between source and target domains but also among camera-aware sub-domains. Previous works solved the unsupervised person Re-ID problem focus on learning from data, while in this work we focus on learning knowledge from pre-trained models, i.e., on "model-level." However, unsupervised methods still perform poorly compared with the supervised alternatives and Wang et al. [40] propose to learn models from weak supervision.

Our work is most closely related to Distill [43] using sample similarity matrix to transfer knowledge from source models under source-free setting. However, the knowledge embedded in sample similarity matrix is limited by the sampled batch data, which might not be effective enough to transfer knowledge from source model to target domain.

3 Proposed method

Problem definition To study source-free multi-source domain adaptation for Re-ID, the problem is formulated as follows. Given K labeled source domains $\{S_1, S_2, \ldots, S_k\}$ and one fully unlabeled target domain T. Suppose we have trained models $\{F_1^s, F_2^s, \ldots, F_k^s\}$ from source domains $\{S_1, S_2, \ldots, S_k\}$, respectively. In our setting, we aim to learn a model F^t from trained source models $\{F_1^s, F_2^s, \ldots, F_k^s\}$ and unlabeled data $X^t = \{x_i^t\}_{i=1}^{N_t}$ without any source data.

3.1 Framework overview

In this section, we introduce the proposed source-free multi-source domain adaptation framework. Our framework includes 2 steps, which are proxy task learning and multi-source domain adaptation via aggregated proxy label. For proxy task learning, we first sample N^p unlabeled images from target domain to train a proxy label generator and domain discriminator. In the multi-source domain adaptation stage, the aggregated proxy labels are generated from two branches. First, we feed the features extracted from different source models of input image into proxy label generator and obtain different proxy labels which are corresponding to different source models. Second, the domain discriminator takes the proxy label of target model saved from last iteration to choose a weighting strategy over multiple proxy labels. We obtain the aggregated proxy label of input image by combining all the proxy labels from different source models with different weighting. Finally, we fine-tune the target network by the aggregated proxy labels to transfer knowledge. The details of each step will be explained in the following subsections.

3.2 Proxy task learning

The proxy task includes two sub-task, which are proxy label learning and domain discriminative learning, shown at Fig. 2. By learning the proxy task, the knowledge from source models are embedded in proxy label.

Proxy label learning In brief, proxy label learning aims at training a classifier on input feature to identify which image is from and each image x_i^t is regarded as individual identity (category) y_i^t. Suppose we sample N_p images from target

dataset for training, the proxy label is the output probability distribution $\tilde{y} \in \mathbb{R}^{N_p}$ of classifier. It is note that using K source models as feature extractors can obtain K different features for each individual image, while the classifier only discriminates the feature of one source model at each time. The classifier G is optimized by cross-entropy loss, written as:

$$\mathcal{L}_{pl}(G) = \frac{1}{N^p}\frac{1}{K}\sum_{i=1}^{N^p}\sum_{j=1}^{K}\mathcal{L}_{ce}\left(G\left(F_j^s\left(x_i^t\right)\right), y_i^t\right) \tag{1}$$

where each target image x_i^t is corresponding to a unique y_i^t. When the training process is converged, the classifier G is used as proxy label generator. The proxy label is formally defined as:

$$\tilde{y}^{(z)} = G(F(x^t))^{(z)} = \frac{\exp\left(W_{g,z}^{\mathrm{T}}F(x^t)\right)}{\Sigma_a \exp\left(W_{g,a}^{\mathrm{T}}F(x^t)\right)} \tag{2}$$

where $\tilde{y}^{(z)}$ is the z-th entry of proxy label $\tilde{y}$. $F(\cdot)$ denotes any network for feature extraction and $W_{g,a}$ is the a-th entry from the proxy label generator G.

Intuitively, the proxy label $\tilde{y}$ represents the similarity among the input image x^t and N_p images sampled for proxy task. Each entry $W_{g,a}$ of proxy label generator G can be viewed as a prototype of image x_a^t. Even though the source models suffer from the domain shift problem, which leads to the degradation of performance, they still preserve the ability of capturing the common low-level feature for Re-ID task. By optimizing the cross-entropy with learnable parameters, each prototype of N_p image is more discriminative than the original feature. Compared with modeling the similarity directly utilizing the original feature, our proxy task not only tackles the discriminative ability problem of raw feature but also integrates the knowledge contained in multiple source models. The knowledge contained in proxy label are not depend on any specific data, instead it is constructed by the similarity among a set data. Our proxy label provides another approach to characterize the similarity among samples instead of using identities which is concordant with the fact that common Re-ID task is based on the similarity metric among identities.

Domain discriminative learning Another key problem in multi-source domain adaptation is how to select effective knowledge from multiple source domain. To address this problem, we introduce another sub-task, which is domain discriminative learning, into the proxy task. In domain discriminative learning, the domain discriminator D aims to distinguish which source domain is the proxy label from, which can measure the discrepancy among different source. We can optimize D by minimizing the following cross-entropy loss:

$$\mathcal{L}_{dd}(D) = \frac{1}{N^p}\frac{1}{K}\sum_{i=1}^{N^p}\sum_{j=1}^{K}\mathcal{L}_{ce}\left(D\left(G\left(F_j^s\left(x_i^t\right)\right)\right), y_j^s\right) \tag{3}$$

where y_j^s denotes the source S_j of the model F_j^s. The output of the domain discriminator D represents the distance among the input model and multiple source models, which can provide guiding information for selecting appropriate source model to transfer knowledge.

Domain discriminator is widely used in adversarial-based unsupervised domain methods [12,15,37] to distinguish source and target domains. And it differs from the proposed approach that our proposed domain discriminative learning is based on proxy label instead of features. Compared with image classification, Re-ID task is an open set problem, which means that the identities (categories) are not shared among different datasets. Hence, the feature across different identities might not characterize the discrepancy among different source well, while the proxy labels are based on the similarity among samples, which contain more information than the specific feature.

3.3 Multi-source domain adaptation via proxy label

We show an overall illustration of our multi-source domain adaptation method in Fig. 3. Recall that the proxy label generator can generate K proxy labels using K different source models for any input image. To learn from these proxy labels, target model needs to integrate the knowledge embedded in those proxy labels. Since the discrepancy among source and target domains are various, the contributions of multiple source models need to be adjusted adaptively. To achieve this goal, we estimate the target distribution with target network $F_{\tau-1}^t$ from last iteration. The domain discriminator D generates weighting strategy over source domains by the proxy label of the target network $F_{\tau-1}^t$. The weighting strategy of different source models reflects the distance between source and target distribution through proxy label, which emphasizes more relevant sources and suppresses the irrelevant ones. The aggregated proxy label $\tilde{y}_i^*$ and weighting strategy $\{\omega_{i,j}\}_{j=1}^K$ of image x_i^t can be written as:

$$\tilde{y}_i^* = \frac{1}{K}\sum_{j=1}^{K}\omega_{i,j} * G(F_j^s(x_i^t)) \tag{4}$$

$$\omega_{i,j} = \frac{\exp\left(W_{d,j}^{\mathrm{T}}G\left(F_{\tau-1}^t\left(x_i^t\right)\right)\right)}{\Sigma_a \exp\left(W_{d,a}^{\mathrm{T}}G\left(F_{\tau-1}^t(x_i^t)\right)\right)} \tag{5}$$

Next, in order to transfer the knowledge embedded in the aggregated proxy label $\tilde{y}_i^*$, we minimize the Kullback–Leibler divergence between the proxy label from target network F_τ^t and the aggregated proxy label $\tilde{y}_i^*$ of input x_i^t,

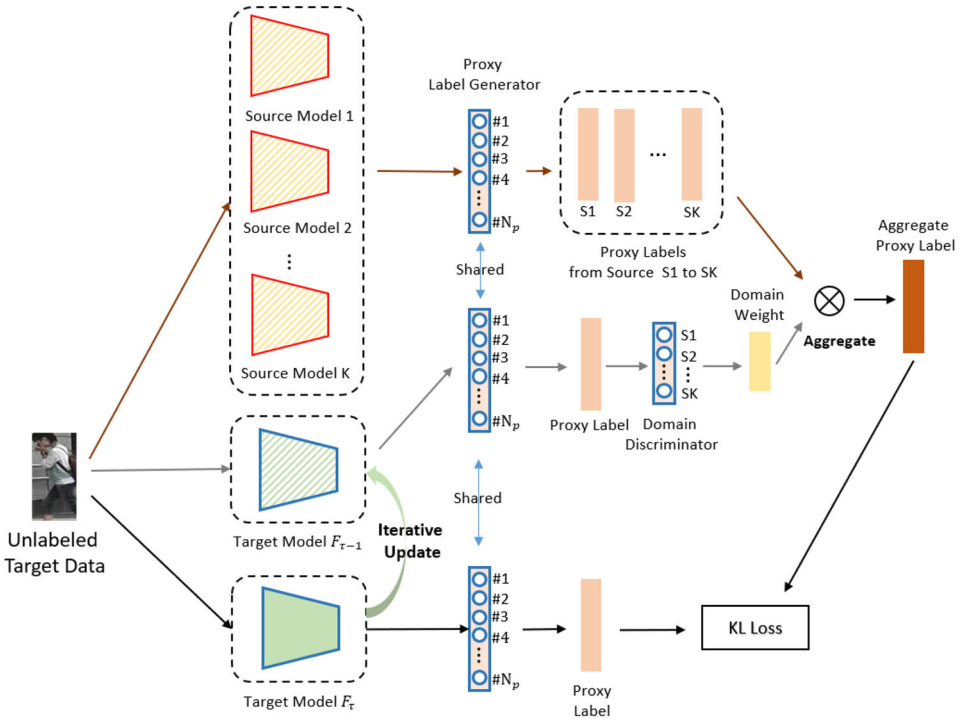


Fig. 3 Illustration of the proposed multi-source domain adaptation via proxy label. It consists of three branches: (1) providing proxy labels of input image from multiple source models; (2) estimating the target domain distribution by last iteration target model for generating domain weighting strategy; (3) updating target model by learning the aggregated proxy label

formulated as follow:

$$\mathcal{L}_{ada}(F_\tau^t) = \frac{1}{N^t}\sum_{i=1}^{N^t}\mathcal{L}_{kl}\left(G\left(F_\tau^t\left(x_i^t\right)\right), \tilde{y}_i^*\right) \tag{6}$$

With the objective of minimizing the L_{ada}, the proxy label of target model can better estimate the target distribution, promoting domain discriminator to select weighting strategy to provide better guidance for target model.

Algorithm 1 Multi-source Domain Adaptation via Proxy label

Require: pre-train source models $\{F_k^s\}_{k=1}^K$; unlabeled target training dataset $\{x_i^t\}_{i=1}^{N_t}$; target model F^t; proxy label generator G; domain discriminator D;
Ensure: well-trained target model $\hat{F}^t$
1: **function** PROXYTASKLEARNING:
2: Sample training data $\{x_i^t\}_{i=1}^{N_p}$ from $\{x_i^t\}_{i=1}^{N_t}$ training set;
3: **for** $i = 1$ to *max-iteration* **do**
4: Updating the proxy label generator G by Eq. (1)
5: Updating the domain discriminator D by Eq. (3)
6: **end for**
7: **end function**
8: **function** MULTI- SOURCEDOMAINADAPTATION:
9: **for** $i = 1$ to *max-iteration* **do**
10: Generating the proxy label from source models $\{F_k^s\}_{k=1}^K$ by Eq. (2);
11: Generating the weighting strategy from domain discriminator D by Eq. (4)
12: Updating the target model F^t by Eq. (6)
13: **end for**
14: **end function**

4 Experiment

4.1 Experimental settings and datasets

We conduct extensive experiments on two large person re-identification benchmark datasets Market-1501 [52] and DukeMTMC [55].

Datasets Source models are trained with labeled data from the training sets: CUHK03 [17], MSMT17 [42], LPW [32], Market-1501 [52], DukeMTMC [55]. *CUHK03* [17] contains 14,096 images, including 1467 identities, from two cameras. The dataset has two settings, which are labeled bounding boxes and DPM detected bounding boxes. And we use the labeled setting for source model training. *MSMT17* [42] is the current largest publicly available person Re-ID dataset. It has 126,441 images, which have 4101 identities, captured by a 15 cameras. *LPW* [32] consists of 2731 different pedestrians collected from three different crowed scenes. A total of 7694 image sequences are generated with an average of 77 frames per sequence. We random selected 2 frames from each training sequences to construct our training set. There are 11,732 images, which have 1975 identities, which are used for training. *Market-1501* [52] contains 32,668 images, including 1501 persons, from six cameras. There are 12,936 images, which contain 751 identities, used for training. And 3,368 images are in the query and 19,732 in gallery sets. *DukeMTMC* [55] has 1404 persons from eight cameras, with 16,522 training images of 702 identities, 2228 queries, and 17,661 gallery images. Basic information of the datasets is in

Table 1. In evaluation, similarities between query and gallery samples were determined by the target model. We use cumulative matching characteristic (CMC) [52] and mean Average Precision (mAP) [55] as performance metrics (Fig. 4).

Implementation details We adopt ResNet-50 as backbone with extra batch normalization layer [22] after global average pooling layer in all experiments unless otherwise indicated. The source models are only trained by ID-discriminative embedding (IDE) [53] for 80 epochs. The target model was initialized by ImageNet pre-train, without training on any Re-ID dataset. The Adam optimizer is used to train all the model with learning rate 0.00035. We resize each image into 384×128 pixels and pad 10 pixels with zero values. Then we randomly crop it into a 384×128 image and adopt random flip with 0.5 probability. We set the mini-batch size to 64. The feature maps extracted through the additional batch normalization layer were used as feature vectors with 2048 dimensions. During the proxy label learning, we random select 75% of unlabeled images from target dataset to train the proxy label generator.

4.2 Comparison with the state of the art

Table 2 presents the comparison with recent state-of-the-art unsupervised learning methods on Market-1501 and DukeMTMC. The compared methods can be divided into 4 types based on the usage of source data. (1) N-N: fully unsupervised learning methods without any source data, including LOMO [18] BoW [52], PUL [9] and CAMEL [46]; (2) S-N: single-source domain adaptation methods only using source data for pre-train, including TJ-AIDL [38] and T-Fusion [23]; (3) S-P: single-source domain adaptation methods with source data used for joint learning during adaptation, including PTGAN [42] , SPGAN [6], HHL [56] and UCDA [29]; (4) M-N: source-free multi-source domain adaptation, including Distill [43].

As seen, our M-N approach achieves competitive performances to the compared unsupervised Re-ID, while the performance on Market-1501 is not largely beyond the best N-N methods. It should be noticed that our method does not use source model for training the second phase and the target model is learning from random initialization. Compared with the state-of-the-art method Distill [43] with the same experimental setting, our model achieves 9.8 and 9.3% improvement in rank-1 accuracy and mAP on DukeMTMC, respectively. It indicates that proposed proxy task framework can more effectively transfer knowledge from multiple source models to target domain. The main advantage of our method is that we can utilize information from multiple source models even with different architectures and learn the common information from multiple teachers, which is beneficial for practical online models. Besides, we do not need to access the specific architecture of the source models and the target model can transfer knowledge just from the output of source model. Our method is more practical due to privacy ad security.

4.3 Further evaluation and ablation study

In this section, we first evaluate the baseline performance by directly deploying source model on target dataset. Then we further conduct ablation studies on the number of images for proxy task learning, the effectiveness of our weighting strategy, the loss function for adaptation and the architecture of target model.

Baseline evaluation We evaluate the performance of all source models on target datasets individually. The results are reported in Table 3. When trained and tested both on same dataset, the model can get high performance. However,

Table 1 Statistics information of datasets

Dataset	Identities	Cameras	Training images	Probe	Gallery	Total images
CUHK03 [17]	1467	2	7368	1400	5328	28,192
LPW [32]	2731	11	11,732	3024	4288	19,044
MSMT17 [42]	4101	15	32,621	11,659	82,161	126,441
Duke [55]	1812	8	16,522	2228	17,661	36,411
Market [52]	1501	6	12,936	3368	15,913	32,668

Fig. 4 Examples of the CUHK03, MSMT17, LPW, Market-1501 and DukeMTMC datasets. Images in each column represent the same identity

Table 2 Comparison with the state-of-the-art methods of unsupervised Re-ID on Market1501 and DukeMTMC-reID

Methods	Setting	DukeMTMC				Market1501			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [18]	N-N	12.3	21.3	26.6	4.8	27.2	41.6	49.1	8.0
BOW [52]		17.1	28.8	34.9	8.3	35.8	52.4	60.3	14.8
PUL [9]		30.0	43.4	48.5	16.4	45.5	60.7	66.7	20.5
Song et al. [33]		75.1	88.7	92.4	52.5	68.4	80.1	83.5	49.0
CAMEL [46]		40.3	–	–	19.8	54.5	–	–	26.3
SSG [11]		73.0	80.6	83.2	40.4	80.0	90.0	92.4	58.3
PAUL [45]		72.0	82.7	86.0	53.2	68.5	82.4	87.4	40.1
TJ-AIDL [38]	S-N	44.3	59.6	65.0	23.0	58.2	74.8	81.1	26.5
T-Fusion [23]		–	–	–	–	60.8	74.4	79.3	–
UMDL [27]	S-Y	18.5	31.4	37.6	7.3	34.5	52.6	59.6	12.4
PTGAN [42]		27.2	–	50.7	–	38.6	–	66.1	–
SPGAN [6]		46.4	62.3	68.0	26.2	57.7	75.8	82.4	26.7
HHL [56]		46.9	61.0	66.7	27.2	62.2	**78.8**	**84.0**	31.4
UCDA [29]		55.4	–	–	36.7	**64.3**	–	–	34.5
Distill [43]	M-N	48.4	–	–	29.4	61.5	–	–	33.5
Ours		**58.2**	**71.6**	**76.8**	**38.7**	63.4	78.7	83.5	**38.1**

The bold indicates the best result

Table 3 Performance (%) of source models directly tested on target dataset DukeMTMC and Market-1501

Training set	DukeMTMC				Market-1501			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
CUHK03 [17]	19.9	32.7	37.7	10.2	42.6	59.1	64.9	19.2
MSMT17 [42]	49.4	64.7	71.1	29.9	46.7	64.4	72.3	22.3
LPW [32]	32.7	49.9	56.6	18.2	54.6	71.5	77.5	28.8
DukeMTMC [55]	**82.8**	**91.7**	**94.4**	**68.8**	41.8	60.0	67.1	18.0
Market-1501 [52]	28.1	43.3	49.4	14.8	**92.1**	**97.0**	**98.1**	**80.1**

The bold indicates the best result

performance drops significantly when the model is directly deployed on the target dataset which is different from training set. For example, the baseline model trained and tested on DukeMTMC achieve 82.8% in rank-1 accuracy, but drops to 41.8% when tested on Market-1501. The domain shift among datasets is the causes of performance degradation.

Number of images N_p for proxy task In this experiment, we analyze the impact of the ratio of samples that used for our self-supervised learning pretext task. Specifically, we conduct experiments with two default setting: (1) transfer the trained model from Market-1501 to DukeMTMC. (2) Transfer the trained model from DukeMTMC to Market-1501. The direct transfer results with no unlabeled images available are also reported for reference. We show the results of sampling different number of images for proxy label learning in Table 4. We can observe that the more training samples used for proxy task learning, the higher performance of target model in general. It illustrates that the number of images used for training relates to the efficiency and robustness of knowledge transfer. We also observe when using 75% unlabeled images, our method leads to the best result on DukeMTMC. As DukeMTMC dataset may contain multiple persons and serious occlusion problem, this dataset is more challenge. The reason behind this may be the training unstable and the left 25% unlabeled images are easy samples and they contribute less to the gradient optimization. Empirically, we set N_c^t to 75% of unlabeled dataset in the following experiment.

Benefit of the multi-source aggregated adaptation We study the benefit of our propose multi-source domain adaptation in Table 5. Firstly, we conduct experiments on different combinations of source models. Results show that even using single source for proxy task, our method outperforms the baseline, i.e., directly tested on target domain. A case in point is that our method obtains rank-1 accuracy in 38.4% by using CUHK03 as source model on DukeMTMC dataset, surpassing the baseline by +18.5%. Also, the proxy label becomes more robust with the increasing number of source models, so that it boosts the performance of adaptation. Furthermore, we compare the proposed weighting strategy with the baseline using average weighting to generate aggregated proxy label.

Table 4 Performance (%) of using different numbers of unlabeled data for proxy label learning

Unlabeled images	DukeMTMC				Market-1501			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
0% (Direct transfer)	28.1	43.3	49.4	14.8	41.8	60.0	67.1	18.0
10%	51.4	65.7	72.1	32.9	53.7	69.2	75.5	29.4
25%	51.8	67.5	73.6	34.2	55.7	71.1	76.9	30.7
50%	56.0	69.0	74.5	36.3	61.5	77.0	82.2	36.7
75%	**58.2**	**71.6**	**76.8**	**38.7**	63.4	78.7	83.5	38.1
100%	57.9	70.1	75.7	38.2	**63.9**	**79.6**	**84.5**	**39.4**

The bold indicates the best result

Table 5 Performance (%) of using different combination of source models and weighting strategy under proposed adaptation framework

Methods	DukeMTMC				Market1501			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
CUHK03	38.4	53.9	59.6	23.3	56.2	71.0	76.4	31.5
MSMT17+CUHK03	55.1	69.3	74.4	36.5	58.0	73.6	78.8	31.5
MSMT17+CUHK03+LPW	56.1	69.8	75.7	37.0	60.2	76.5	82.4	35.1
Average Weight	50.9	65.9	72.0	33.5	62.2	77.7	83.1	37.3
Ours	**58.2**	**71.6**	**76.8**	**38.7**	**63.4**	**78.7**	**83.5**	**38.1**

The name of dataset corresponding to the combination of source model used for adaptation
The bold indicates the best result

As seen in Table 5, our proposed weighting strategy outperforms the average weighting. This is reasonable because the average weighting does not reveal the importance of different sources; therefore the proxy label aggregated by average weighting may not fit the target distribution well.

Different loss functions for target model adaptation As shown in Table 6, we evaluate how different loss functions affect our adaptation. The performance of using L1 and L2 loss function drops significantly compared with Kullback–Leibler divergence. This indicates that the Kullback–Leibler divergence is more suitable for characterizing the discrepancy of distribution among source and target models by proxy label.

The variations of source/target model architectures In this experiment, we explore the transfer ability of our method with different source/target architecture. To this end, we adopt three widely used networks: ResNet50, ShuffleNetV2 [24] and MobileNetV2 [31] for inference. The results are reported in Table 7 and we make following observations. (i) With less parameter and lower computation costs, the MobileNetV2 [31] achieves best result for most cases. It shows the flexibility and generalization of our methods, and the knowledge can be effectively transferred across heterogeneous model architectures. (ii) Comparing with target model, the performance is more robust to the selection of source model, e.g., the R-1 result on DukeMTMC with different source model but same MobileNetV2 varies from 59.2 to 61.7% (+2.5%). On the contrary, with same source model, the R-1 result on DukeMTMC with different target models varies from 49.7 to 59.2% (+9.5%).

Table 6 Performance (%) of using different loss function for adaptation

Loss	DukeMTMC				Market-1501			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
L1	47.7	64.0	70.3	29.0	49.4	67.9	74.5	24.2
L2	38.5	53.4	60.2	22.1	41.7	59.8	67.7	20.1
KL	**58.2**	**71.6**	**76.8**	**38.7**	**63.4**	**78.7**	**83.5**	**38.1**

The bold indicates the best result

4.4 Visualization of proxy label

In order to show the interpretability of our proposed proxy task, we use the heat map to visualize the cosine similarity of proxy labels from different models on target samples. As illustrated in Fig. 5, the coordinate axis represents different samples sorted by identity index in the same order. We can observe that the brightness in square is very low along the diagonal when using pre-trained source model (CUHK03) to generate the proxy label. Meanwhile, there are some very bright square with clear boundary along the diagonal when using the model trained and tested on same dataset. There also exits a large performance gap between these two models, which implies that our proxy label can identify image like actual annotation in some way. It verifies that our proposed

Table 7 Performance(%) of using different backbone as source and target model. R50 is short for ResNet-50

Source → Target	DukeMTMC				Market1501			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
SNV2 → SNV2 [24]	49.7	64.1	70.1	30.2	54.8	70.0	75.9	26.8
SNV2 → MNV2 [31]	**59.2**	**73.0**	**77.9**	**41.1**	**62.5**	**78.2**	**84.3**	**38.5**
SNV2 → R50	58.4	72.0	77.1	39.1	61.8	77.0	83.9	36.2
MNV2 → SNV2 [24]	52.3	68.9	72.4	33.9	58.4	72.4	80.5	30.2
MNV2 → MNV2 [31]	**61.7**	**74.2**	**79.5**	**41.8**	**63.9**	**79.9**	**86.2**	**38.6**
MNV2 → R50	58.6	72.6	77.8	39.7	63.1	78.7	83.5	37.2
R50 → SNV2 [24]	50.3	64.9	70.8	30.9	55.6	70.6	77.2	28.2
R50 → MNV2 [31]	**59.8**	**73.5**	**78.6**	**40.8**	62.6	**79.1**	**85.1**	37.7
R50 → R50	58.2	71.6	76.8	38.7	**63.4**	78.7	83.5	**38.1**

SNV2 is short for ShuffleNetV2 and MNV2 is short for MobileNetV2
The bold indicates the best result

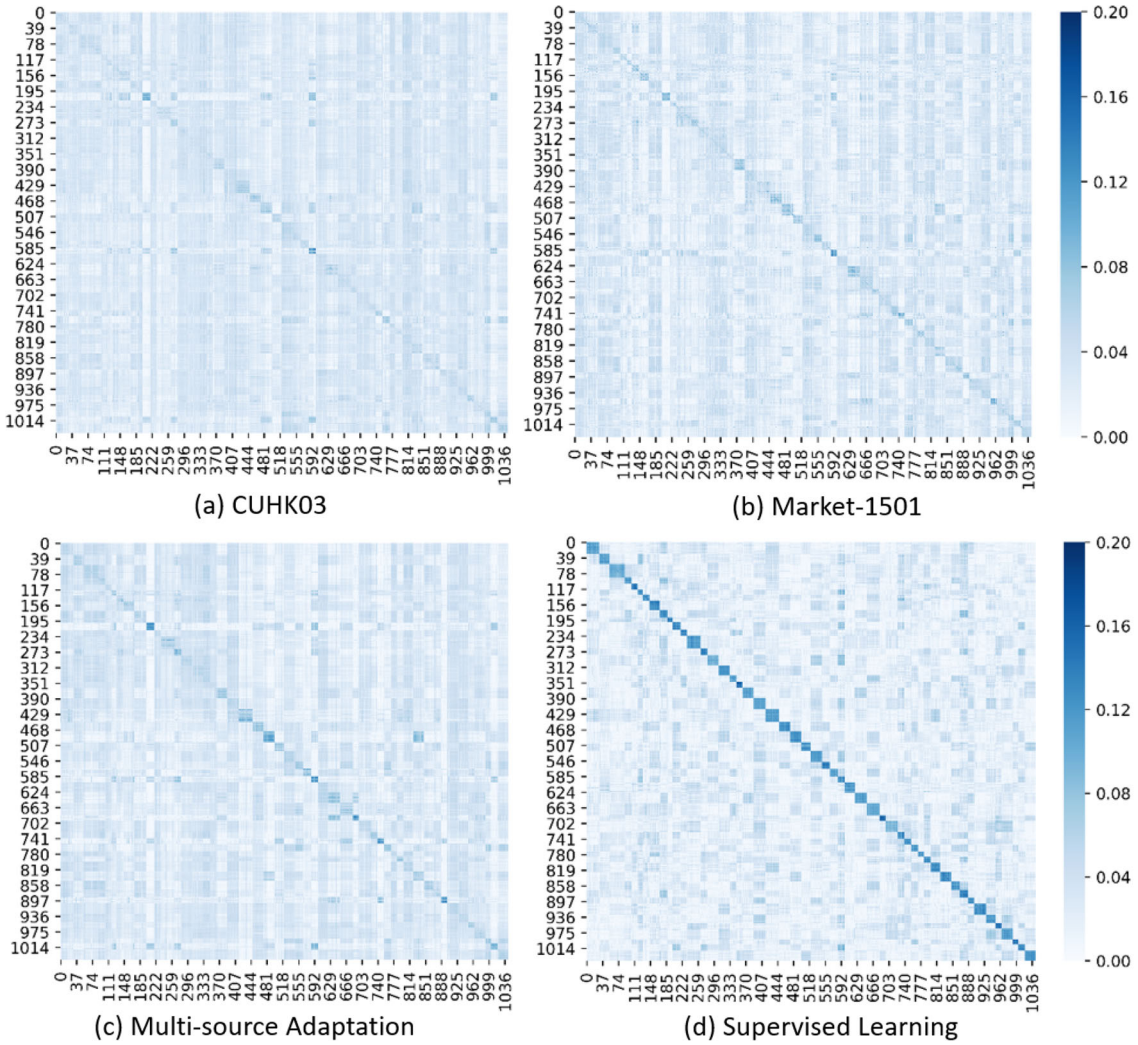


Fig. 5 Visualization of the proxy label of 50 identities randomly selected from DukeMTMC dataset. For **a**–**d** we use pre-trained source model from CUHK03, Market-1501, target model after multi-source adaptation, supervised learning on DukeMTMC, respectively. The coordinate axis represents different samples sorted by identity index in same order and the element at heatmap corresponding to cosine similarity of proxy label between the x- and y-images

proxy label indeed embeds useful knowledge from source models for Re-ID task via proxy task.

5 Conclusion

In this paper, we present a new proxy task learning framework for unsupervised multi-source domain adaptation on person re-identification, which does not require any source data. By introducing proxy label learning into proxy task, the knowledge from source models can be embedded into proxy label. To integrate the knowledge from different source models, we also propose domain discriminative learning for proxy task which aims at generating weighting strategy over proxy labels from different source. Experiments conducted on DukeMTMC and Market-1501 verify that our approach achieves competitive performance compared with the state of the art. In the future work, we will further combine the metric learning and meta learning into proxy task.

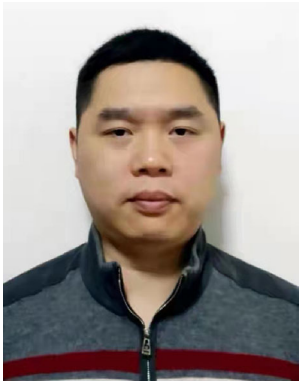
Acknowledgements This work was supported in part by the funding called Analyzing and hardening single event effects in LDO (2019A1515110127).

Funding The funded was grant by Foshan University (2019A1515110127).

References

1. Ahmed, S.M., Lejbolle, A.R., Panda, R., Roy-Chowdhury, A.K.: Camera on-boarding for person re-identification using hypothesis transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12144–12153 (2020)
2. Bai, S., Li, Y., Zhou, Y., Li, Q., Torr, P.H.: Adversarial metric attack and defense for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 2219–2126 (2020)
3. Bojanowski, P., Joulin, A.: Unsupervised learning by predicting noise. In: ICML (2017)
4. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR (2017)
5. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: attentive but diverse person re-identification. In: ICCV (2019)
6. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
7. Ding, C., Wang, K., Wang, P., Tao, D.: Multi-task learning with coarse priors for robust part-aware person re-identification. IEEE Trans. Pattern Anal. Mach, Intell (2020)
8. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
9. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: clustering and fine-tuning. In: TOMM (2018)
10. Fu, Y., Wei, Y., Wang, G., Zhou, X., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: ICCV (2018)
11. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6112–6121 (2019)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2031 (2017)
13. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
15. Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: cycle-consistent adversarial domain adaptation. In: ICML (2018)
16. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR (2017)
17. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: CVPR (2014)
18. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
19. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI (2019)
20. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: ICML (2015)
21. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML (2017)
22. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPRW (2019)
23. Lv, J., Chen, W., Li, Q., Yang, C.: Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: CVPR (2018)
24. Ma, N., Zhang, X., Zheng, H., Sun, J.: Shufflenet V2: practical guidelines for efficient CNN architecture design. In: ECCV (2018)
25. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
26. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: CVPR (2016)
27. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: CVPR (2016)
28. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
29. Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., Gao, Y.: A novel unsupervised camera-aware domain adaptation framework for person re-identification. In: ICCV (2019)
30. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.: Dataset shift in machine learning. The MIT Press. ISBN 0262170051, 9780262170055 (2009)
31. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: inverted residuals and linear bottlenecks. In: CVPR (2018)
32. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: AAAI (2018)
33. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: theory and practice. Pattern Recogn. **102**, 107173 (2020)
34. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. In: ECCVW (2016)
35. Sun, S., Shi, H., Wu, Y.: A survey of multi-source domain adaptation. Inf. Fusion. **24**, 84–92 (2015)

36. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and A strong convolutional baseline). In: ECCV (2018)
37. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
38. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
39. Wang, J., Gao, Y., Li, K., Lin, Y., Ma, A.J., Sun, X.: Removing the background by adding the background: Towards background robust self-supervised video representation learning. In: CVPR (2021)
40. Wang, X., Liu, M., Raychaudhuri, D.S., Paul, S., Wang, Y., Roy-Chowdhury, A.K.: Learning person re-identification models from videos with weak supervision. IEEE Trans. Image Process. **30**, 3017–3028 (2021)
41. Wang, X., Panda, R., Liu, M., Wang, Y., Roy-Chowdhury, A.K.: Exploiting global camera network constraints for unsupervised video person re-identification. IEEE Trans. Circuits Syst, Video Technol (2020)
42. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: CVPR (2018)
43. Wu, A., Zheng, W., Guo, X., Lai, J.: Distilled person re-identification: Towards a more scalable system. In: CVPR (2019)
44. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: multi-source unsupervised domain adaptation with category shift. In: CVPR (2018)
45. Yang, Q., Yu, H.X., Wu, A., Zheng, W.S.: Patch-based discriminative feature learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3633–3642 (2019)
46. Yu, H., Wu, A., Zheng, W.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV (2017)
47. Yu, H., Zheng, W., Wu, A., Guo, X., Gong, S., Lai, J.: Unsupervised person re-identification by soft multilabel learning. In: CVPR (2019)
48. Zhang, L., Qi, G., Wang, L., Luo, J.: AET vs. AED: unsupervised representation learning by auto-encoding transformations rather than data. In: CVPR (2019)
49. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
50. Zhao, H., Zhang, S., Wu, G., Moura, J.M.F., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: NeurIPS (2018)
51. Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., Keutzer, K.: Multi-source distilling domain adaptation. In: AAAI (2020)
52. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)
53. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. ArXiv preprint arXiv:1610.02984 (2016)
54. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: CVPR (2019)
55. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV (2017)
56. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: ECCV (2018)
57. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: exemplar memory for domain adaptive person re-identification. In: CVPR (2019)



Yi Ding received his Ph.D. degree in Communication and Information System from Sun Yat-sen University in 2015. Since July 2015, he has been an assistant professor in the College of Computer and Electrical Engineering at Human University of Arts and Science. His research interests include multimedia security, computer vision, and deep learning.

Zhikui Duan received his Ph.D. degree in Communication and Information System from the Sun Yat-sen University in 2015. Since July 2016, he has been an associate professor in the school of electronics and information technology at Foshan University. His research interests include analogue integrated circuit, computer vision, automatic speech recognition, and RSA algorithm.

Shiren Li received masters degree in Sun Yat-Sen University, Guangdong, China, in 2016. His current research interests are computer vision, machine learning, data mining and automatic speech recognition.