Information Sciences 559 (2021) 46–60

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Cross adversarial consistency self-prediction learning for unsupervised domain adaptation person re-identification

Huafeng Li [a,b,1], Jian Pang [a,b,1], Dapeng Tao [c,*], Zhengtao Yu [a,b,*]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500 Yunnan, PR China
[b] Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming, 650500 Yunnan, PR China
[c] FIST LAB, School of Information Science and Engineering, Yunnan University, Kunming, 650091 Yunnan, PR China

ARTICLE INFO

Article history:
Received 20 August 2020
Received in revised form 1 January 2021
Accepted 6 January 2021
Available online 27 January 2021

Keywords:
Person re-identification
Domain adaptation
Consistency self-prediction
Cross adversarial consistency learning

ABSTRACT

Domain-invariant feature-extraction has become very popular for unsupervised domain adaptation (UDA) person re-identification (Re-ID). However, most methods using it are limited by weak discrimination of learned domain-invariant features. To solve this problem, we develop a new approach: cross-adversarial consistency self-prediction learning. Cross-adversarial consistency is used to endow the learned feature with domain invariance and discrimination; consistency self-prediction fine-tunes the pre-trained model by selecting non-paired samples from target data. First, the camera views of source domain are randomly divided into two groups with their samples. Then, the two identifiers are used crosswise on both groups, forcing consistent results through adversarial learning between the identifiers and the feature encoder. To refine the model, a self-prediction mechanism is introduced that conservatively selects target domain samples with high identity similarities to labeled source domain samples. This practical design helps to alleviate domain bias between the source and target domains. The results of experiments conducted on five benchmark datasets verify that the proposed method is effective and outperforms state-of-the-art competitors. The source code of our method is available at https://github.com/PangJian123/CAC-CSP.

1. Introduction

Person re-identification (Re-ID) involves finding a pedestrian with the same identity as the query image from non-overlapping camera views. In contrast to common image retrieval [1–3], in which the label of the query image is the same as that of the training images, for person Re-ID, training and testing images have entirely different identities. It is thus much more challenging to retrieve, e.g., a specific pedestrian image from a large-scale pedestrian image library. Although, with the advent of deep learning [4,5], supervised person-Re-ID methods have emerged and achieved significant progress [6,7], such methods require a large number of manually labeled samples to train the model, which is extremely costly. Paired positive samples are also scarce in real scenarios. Furthermore, applying these trained models from a labeled source domain to other datasets can entail significant performance degradation owing to domain bias between the source and target data.

* Corresponding authors.
E-mail addresses: dapeng.tao@gmail.com (D. Tao), ztyu@hotmail.com (Z. Yu).
[1] Equal contribution.

https://doi.org/10.1016/j.ins.2021.01.016
0020-0255/

At present, one of the most popular solutions to these problems is unsupervised domain adaptation (UDA) person Re-ID. The existing UDA person Re-ID methods can be roughly divided into three categories: camera-style-transfer (CST) methods [8–11], pseudo-label prediction (PLP) methods [12–14], and domain-invariant feature-learning (DIFL) methods [15,16]. CST typically transfers the image style from the labeled source domain to the target domain, and then trains the model in a supervised manner with the transferred source data. The performance depends heavily on the quality of the transferred images. If the visual cues associated with the ID label cannot be well preserved before and after translation, recognition performance will suffer greatly. Although some researchers have found ways to mitigate this [17,18], the problem is still far from being solved.

PLP assigns pseudo-labels to the target samples, selecting those with high confidence to fine-tune the pre-trained model with supervised training. Such methods are effective on public datasets and their performance surpasses that of other UDA person-Re-ID methods by a large margin. However, they may not be practical in real-world deployment, as it is extremely rare for pairwise samples to appear in different camera views in real scenarios; this makes it difficult for the self-training method to select the correct matched sample pair. Furthermore, even if a few sample pairs can be selected, the amount is not sufficient to optimize a deep-learning model as such models rely on large-scale parameters.

DIFL-based person Re-ID is gaining increasing attention for practical reasons, and some effective implementations have recently been proposed [15,16,19]. However, the performance of these methods is still not ideal, and they especially lack discrimination and robustness. To solve these problems, we propose cross-adversarial consistency (CAC) learning and consistency self-prediction (CSP) learning for UDA person Re-ID (Fig. 1). CAC is used to improve the domain invariance of the learned features, and CSP is used to increase the discrimination. In CAC, first, the camera views of source domain are randomly divided into two groups with their samples; then, an identifier is trained for each group and used crosswise to obtain consistent classification results in adversarial learning. The learned features are thus endowed with domain invariance.

In CSP, we propose using identity-consistency-based discriminative feature learning to improve the discrimination and robustness of the learned features. In particular, we apply the two identifiers learned by CAC to each sample of the target domain simultaneously. Thereafter, the feature encoder is updated under the guidance of the identity-consistency output loss of the two identifiers to make the learned features more discriminant, robust, and domain-invariant. In self-prediction learning, we apply to samples in the target domain the labels of the source domain samples to which they are most similar. The assigned label for the target sample is called a soft-label, and such a labeled target sample is called a soft-labeled sample. We combine these soft-labeled samples with the corresponding labeled source samples to form pseudo-positive sample pairs and use these to optimize the feature encoder.

This design is essentially different from the existing method of pseudo-label prediction based on self-training, which selects pseudo-labeled samples from the target domain to fine-tune the pre-trained model. Although such an approach

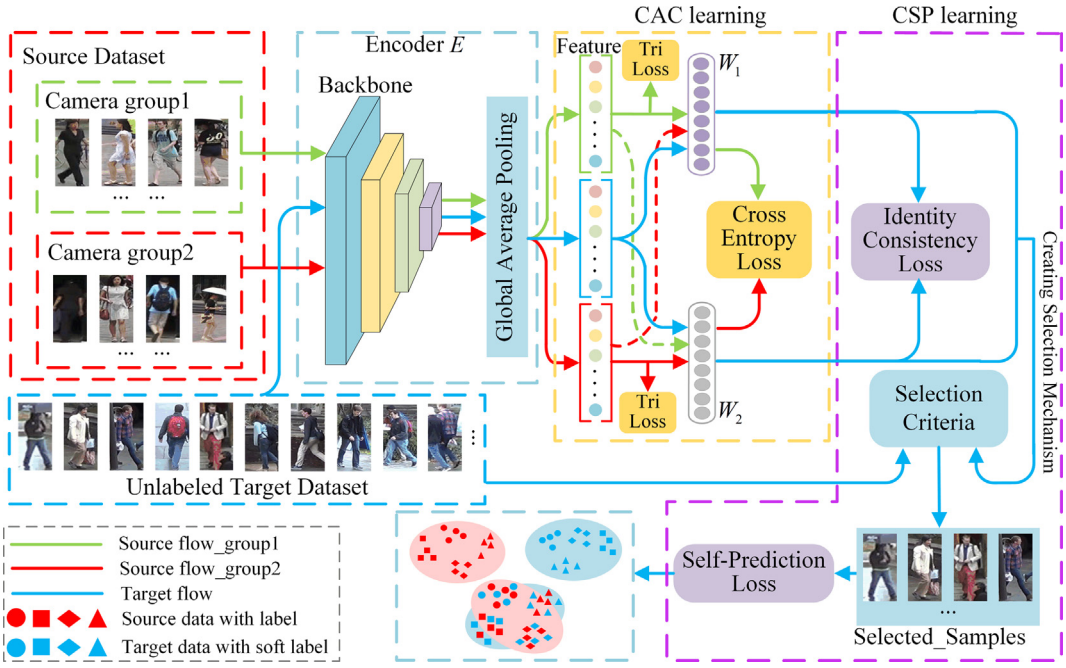

Fig. 1. Overview of cross-adversarial consistency self-prediction learning. Our proposed model consists of two parts: cross-adversarial consistency (CAC) learning and consistency self-prediction (CSP) learning. Camera groups 1 and 2 are generated by randomly dividing the source dataset into two groups according to their camera views. The backbone is ResNet-50 and pre-trained on ImageNet. W_1 and W_2 denote two different classifiers. The dotted lines entering the two classifiers represent the adversarial training.

has better performance than the simple DIFL method on public datasets, it may not be practical, because positive sample pairs are extremely scarce in real scenarios. Our method assigns the labels of the source dataset as soft-labels to the target domain samples, so that a soft-labeled sample pair including the target sample can be formed to optimize the model. This is more practical than the existing method, because there is no need to select labeled pairwise samples from the target domain to fine-tune the model. The main contributions and advantages of our work are as follows:

- We present a novel CAC learning algorithm to achieve the domain-invariant feature discrimination. The source datasets are randomly divided into two groups according to the camera views and an identifier is trained for each group. The two ID classifiers are then used on both groups. Driven by adversarial learning, the two different identifiers can output consistent recognition results for the same pedestrian image, thus endowing the feature encoder with the ability to extract domain-invariant features.
- To learn more discriminative features from the unlabeled target domain, a novel CSP learning algorithm is developed. In this method, the label of the source sample is assigned to the target sample as a soft-label. The soft-labeled target samples and the labeled source samples are then combined to form soft-hard sample pairs, improving the discrimination and robustness of the learned features.
- Our method is an end-to-end model involving no auxiliary network or additional computing cost. With the assistance of the source domain samples and the fine-tuning with the soft-labeled target samples, it can be deployed directly on other target datasets. To evaluate our approach, extensive experiments were conducted on five popular datasets. The results demonstrate that our proposed method is effective and superior to other methods.

The reminder of this paper is organized as follows: Section 2 reviews relevant related work. Section 3 describes the proposed method in detail. Section 4 discusses the experimental settings and analyzes each part of the proposed method. Section 5 presents concluding remarks.

2. Related work

2.1. DIFL for UDA person re-ID

DIFL has recently become one of the most commonly used practical methods of UDA person Re-ID. Dictionary learning, transfer learning (TL), and adversarial learning (AL) have all been used for domain-invariant feature extraction in this context [10,19]. Because of its excellent performance in computer-vision tasks [20–22], dictionary learning (the use of a learned dictionary to produce new, domain-invariant features) has attracted much attention in unsupervised person Re-ID [23,24]. TL-based UDA person Re-ID aims to transfer domain-invariant features from the labeled source domain to the unlabeled target domain. In particular, Liu et al. [10] proposed an adaptive transfer network for UDA person Re-ID, decomposing the cross-domain transfer into a set of intermediate sub-tasks corresponding to camera views, resolution, and illumination. Wang *et al.* [19] developed a transferable joint attribute-identity learning model that transfers the labeled information from the source domain to the target domain for UDA person Re-ID. However, these methods cannot optimize the model well, as they select the unlabeled samples from the target domain, thus limiting further improvement in identification accuracy.

AL-based methods achieve domain invariance by eliminating the distribution discrepancy between the target and source domains. In this regard, Qi et al. [15] developed a camera-aware domain-adaptation method based on AL. In their method, the temporal continuity of the target domain in each camera is exploited to create domain-invariant discriminative features. In order to make large-scale person Re-ID highly applicable, Yuan *et al.* [25] proposed an adversarial DIFL model to separate identity information from challenging variations. In addition, local-patch-based feature-learning methods [16,26] have attracted the attention of researchers as they outperform global feature-learning methods[27]. In the context of UDA person Re-ID, Yang *et al.* [16] developed PatchNet, which learns discriminative features for each patch. However, such methods do not exploit fully the unlabeled target domain data to learn more discriminative and robust features; moreover, models using them are more sensitive to variation between the source and target domains.

2.2. Style transfer and pseudo-label prediction for UDA person re-ID

With the development of generative adversarial networks, person Re-ID based on camera-style transfer has attracted the attention of researchers, and many effective methods have emerged[8–11]. Generative adversarial networks transfer the labeled source domain images to the target domain so that the two domains are consistent in camera style; then, the transferred source domain data are used to train the model in a supervised manner. However, such methods require high-quality transferred images, and it is difficult to guarantee that the visual features associated with the identity of an image will not change after transfer. Although Deng et al. [18] noted this issue and proposed a method to retain self-similarity and domain-dissimilarity, the problem still remains unsolved.

Recently, researchers have turned to methods based on pseudo-label prediction[12,13,28,14]. Such algorithms usually involve two steps: pseudo-label prediction for target domain samples and supervised fine-tuning of a pre-trained model with assigned pseudo-label samples. Most of these methods outperform [12,13,28] style-transfer and DIFL methods, and

even, in some recent cases, supervised-learning methods. However, they may not be practical: there are frequently not enough positive sample pairs available for self-training, which we will prove in the experiment. This contrasts sharply with methods based on domain-invariant feature extraction and camera-style transfer.

Our approach is closely related to domain-invariant feature extraction. Unlike existing methods, our proposed method can acquire discriminative information from the unlabeled target domain, and does not need any pseudo-labeled sample pairs in the target domain to fine-tune the pre-trained model.

3. Proposed method

3.1. Overview of framework

As shown in Fig. 1, the developed method consists of CAC learning and CSP learning. CAC employs two identifiers under adversarial training to make the learned features domain-invariant. By conservatively selecting non-paired samples from unlabeled target dataset in training process, CSP can fully exploit the discrimination information from unlabeled target domain and make the learned feature more discriminative and robust. In our model, the ResNet-50 [29] with the instance normalization added before the batch normalization like [30] is used as the backbone and pre-trained on ImageNet [31]. More specifically, two fully connected (FC) layers are adopted and used as classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ after the backbone. In CAC learning, $\boldsymbol{W}_1, \boldsymbol{W}_2$ and feature encoder $\boldsymbol{E}$ are first optimized so that both $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ can correctly identify the samples in their corresponding groups. After that, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are fixed to further optimize the feature encoder $\boldsymbol{E}$ so that $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ can correctly identity the samples as before when the samples between the two camera groups are exchanged. CSP consists of identity consistency learning and identity self-prediction learning. The former makes the two classifiers have identical output, and the latter aims to pull the unlabeled sample closer to the most similar labeled samples.

3.2. Cross adversarial consistency learning

Given a labeled source domain $\{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$ including N_s person images $\boldsymbol{X}_s$ and the corresponding label set $\boldsymbol{Y}_s$, each image $x_{s,i}$ in $\boldsymbol{X}_s$ is associated with an identity label $\boldsymbol{y}_{s,i} \in \boldsymbol{Y}_s$. We also have an unlabeled target domain $\boldsymbol{X}_t = \{x_{t,i}\}_{i=1}^{n_t}$ consisting of n_t images. Our goal is to develop an effective deep Re-ID model to acquire domain invariant discriminative features for UDA person re-ID. First, the camera views of source domain are randomly divided into two groups with their samples. Each group has approximately the same number of cameras. Specifically, assuming that the samples in the source domain are taken from n_s camera views, we divide them into two groups, one contains $[\frac{n_s}{2}]$ camera views, and the other contains $n_s - [\frac{n_s}{2}]$ camera views. There is no overlap of camera views between the two groups, $[\cdot]$ denotes the "round down". As shown in Fig. 2, the developed CAC learning consists of two optimization processes: simultaneous optimization of $\boldsymbol{E}, \boldsymbol{W}_1$ and $\boldsymbol{W}_2$ to make them correctly identify pedestrians of the labeled source domain; optimizing $\boldsymbol{E}$ with the learned $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ to let encoder $\boldsymbol{E}$ extract the domain invariant features.

Specifically, with the image $x_s^{l,i}$ of the l-th group $(l = 1, 2)$ and the i-th image, we train encoder $\boldsymbol{E}$ and identity classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, to make $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ correctly identify the 2048-dimensional features $\boldsymbol{f}_s^{l,i} = \boldsymbol{E}(x_s^{l,i})$. To this end, the cross-entropy loss formulated in Eq. (1) is used:

$$\boldsymbol{L}_{ID_1}(\boldsymbol{E}, \boldsymbol{W}_1, \boldsymbol{W}_2) = \left(\left(\sum_{c=1}^{N_s} -I(c, y_s^{1,i}) \log \left(p_c(\boldsymbol{W}_1(\boldsymbol{f}_s^{1,i})) \right) \right) + \left(\sum_{c=1}^{N_s} -I(c, y_s^{2,i}) \log \left(p_c(\boldsymbol{W}_2(\boldsymbol{f}_s^{2,i})) \right) \right) \right), \tag{1}$$

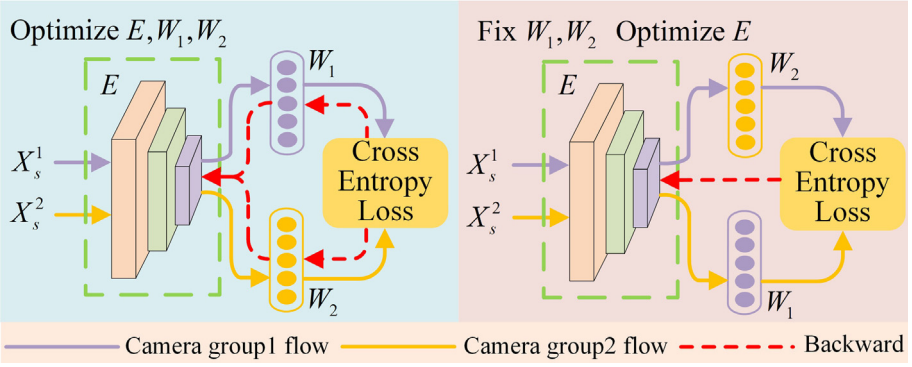

Fig. 2. Scheme of CAC learning. The samples of labeled source are divided into two parts randomly. First train encoder $\boldsymbol{E}$ and identifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ to make $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ identify the input samples, then fix $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ to train $\boldsymbol{E}$ to let $\boldsymbol{E}$ extract domain invariant features.

where p_c represents the predicted logits of class c, and $y_s^{l,i}$ represents the true label corresponding to the input image $x_s^{l,i}$. To prevent over-fitting, a smooth function $I(c,y)$ defined in Eq. (2) is utilized in Eq. (1):

$$I(c,y_s^{l,i}) = \begin{cases} 1-\frac{N_s-1}{N_s}\varepsilon, & \text{if} \quad c = y_s^{l,i} \\ \frac{\varepsilon}{N_s}, & \text{otherwise} \end{cases} \tag{2}$$

where ε is a smoothing parameter and it value is set to 0.1 in this paper. To further improve the discrimination of the learned features, the triplet loss defined in Eq. (3) is used to optimize encoder $\boldsymbol{E}$. For each mini-batch, P identities are selected randomly and K images are sampled randomly for each identity, and the triplet loss (Tri Loss) can be formulated as:

$$\begin{aligned} \boldsymbol{L}_{Tri}(\boldsymbol{E}) = \sum_{i=1}^{P}\sum_{j=1}^{K}[\alpha + \max_{j^*=1,\cdots,K} \|\boldsymbol{E}(x_{s,j}^i) - \boldsymbol{E}(x_{s,j^*}^i)\|_2 \\ - \min_{\substack{i^*=1,\cdots,P \\ j^*=1,\cdots,K \\ i\neq i^*}} \|\boldsymbol{E}(x_{s,j}^i) - \boldsymbol{E}(x_{s,j^*}^{i^*})\|_2]_+ \end{aligned}, \tag{3}$$

where $x_{s,j}^i, x_{s,j^*}^i$ and $x_{s,j^*}^{i^*}$ denote the anchor, positive and negative images of the i-th person, respectively. $[z]_+ = \max\{z,0\}$, α is a margin hyper-parameter of the triplet loss, with its value set empirically to be 0.3. In our method, we treat this model regularized by $\boldsymbol{L}_{ID_1}(\boldsymbol{E},\boldsymbol{W}_1,\boldsymbol{W}_2)$ and $\boldsymbol{L}_{Tri}(\boldsymbol{E})$ as "*baseline*".

To promote the learned encoder $\boldsymbol{E}$ to extract the domain invariant features for UDA person Re-ID, we develop a novel CAC learning method. Particularly, once encoder $\boldsymbol{E}$ and identity classifiers $\boldsymbol{W}_1,\boldsymbol{W}_2$ are learned, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are fixed to update encoder $\boldsymbol{E}$ to make $\boldsymbol{W}_1(\boldsymbol{E}(x_s^{2,i}))$ and $\boldsymbol{W}_2(\boldsymbol{E}(x_s^{1,i}))$ generate correct identity prediction, so as to let encoder $\boldsymbol{E}$ extract domain invariant features. This can be achieved by minimizing the following cross-entropy loss:

$$\boldsymbol{L}_{ID_2}(\boldsymbol{E}) = \left(\left(\sum_{c=1}^{N_s} -I(c,y_s^{1,i})\log\left(p_c(\boldsymbol{W}_2(\boldsymbol{f}_s^{1,i}))\right)\right) + \left(\sum_{c=1}^{N_s} -I(c,y_s^{2,i})\log\left(p_c(\boldsymbol{W}_1(\boldsymbol{f}_s^{2,i}))\right)\right)\right), \tag{4}$$

By conducting adversarial learning between the loss functions Eqs. (1) and (4), the discrimination ability of two identifiers will be gradually improved, forcing the encoder to extract the domain invariant features and securing consistency between the output of identifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ after cross-use on the same image and that before cross-use. Thus encoder $\boldsymbol{E}$ is endowed with the capability of extracting the domain invariant features. This design is different from the traditional domain invariant feature extraction method in that the traditional approach extracts invariant features by adversarial learning between features. The method proposed in this paper is to carry out adversarial learning between two different classifiers. They are cross-used to make encoder $\boldsymbol{E}$ able to extract domain invariant features. Accordingly, our method avoids the loss of discrimination information and shows superiority over other methods.

3.3. Consistency self-prediction learning

3.3.1. Identity consistency learning

This section involves the training of the model to let it learn robust and discriminative feature representation by exploiting unlabeled samples from target domain. As stated above, the two identifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with supervised learning can correctly identify the identities of their corresponding image groups. If used crosswise, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ may lose their original ability to recognize new data due to large appearance variations caused by the changes in illumination, camera views and background. To solve this problem, and make full use of the discrimination information of the unlabeled target domain samples, the following identity consistency loss should be minimized:

$$\boldsymbol{L}_c(\boldsymbol{E}) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_s}\|\boldsymbol{W}_1(\boldsymbol{E}(x_{t,i})) - \boldsymbol{W}_2(\boldsymbol{E}(x_{t,i}))\|_1 \tag{5}$$

where $\|\cdot\|_1$ denotes L_1-norm, K the number of batch size images, and N_s the output dimension of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. Minimizing Eq. (5) can make encoder $\boldsymbol{E}$ extract domain invariant and robust features from the target domain.

3.3.2. Identity self-prediction learning

There are a large number of unlabeled samples in the target domain. If some samples can be selected from it to optimize the model, improvement on the scalability of the model will benefit. The method based on pseudo-label prediction can select pairwise samples from target domain by assigning pseudo-labels to them and optimize the pre-trained model, so the recognition performance will be improved [12–14]. However, such methods are not practical due to the scarcity of positive sample pairs in many real scenarios. Different from these existing methods, we propose a novel identity self-prediction learning to mine discriminative information for unlabeled target domain. As shown in Fig. 3we leverage the similarity between source samples and the batch target samples to create a matching criteria for conservatively selecting non-paired target samples.

The samples of high similarity with the source domain are selected from the target domain and the labels of the source domain samples are treated as their soft-labels to optimize the model.

Let $[x_{t,1}, x_{t,2}, \cdots, x_{t,K}]$ be the images of the target domain contained in one batch size. The output of identifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ on image $x_{t,i}$ are respectively denoted as:

$$\begin{aligned} \boldsymbol{w}_i^1 &= [w_{i,1}^1, w_{i,2}^1, \cdots, w_{i,N_s}^1] \\ \boldsymbol{w}_i^2 &= [w_{i,1}^2, w_{i,2}^2, \cdots, w_{i,N_s}^2] \end{aligned} \tag{6}$$

For two recognition results $\boldsymbol{w}_i^1$ and $\boldsymbol{w}_i^2$ of image $x_{t,i}$, select the label with the maximum probability to store, and it can be formulate as

$$\begin{aligned} q_i^1 &= \arg\max_m \{w_{i,1}^1, w_{i,2}^1, \cdots, w_{i,m}^1, \cdots, w_{i,N_s}^1\} \\ q_i^2 &= \arg\max_n \{w_{i,1}^2, w_{i,2}^2, \cdots, w_{i,n}^2, \cdots, w_{i,N_s}^2\} \end{aligned} \tag{7}$$

After that, we obtain two sets of labels denoted as $\boldsymbol{q}^1 = [q_1^1, q_2^1, \cdots, q_K^1]$ and $\boldsymbol{q}^2 = [q_1^2, q_2^2, \cdots, q_K^2]$, where $q_i^l (l = 1, 2)$ indicates that $x_{t,i}$ and the q_i^l-th person images of source domain share the same label according to the decision of classifier $\boldsymbol{W}_l (l = 1, 2)$. In this way, the labels of the source domain are assigned to the sample of the target domain. For the data of the target domain, these labels are called soft-labels. This process can be achieved by the follow equation:

$$soft_{id}(x_{t,i}) = q_i^1, if q_i^1 = q_i^2 and \frac{w_{i,q_i^1}^1 + w_{i,q_i^2}^2}{2} > \eta \tag{8}$$

where $soft_{id}(x_{t,i})$ represents the soft-label prediction of $x_{t,i}$, and η is the average similarity probability empirically set to be 0.8 in this work. With the selection criteria defined in Eq. (8), we can select some samples from unlabeled target domain to further optimize the model. This can be achieved by minimizing the following identity self-prediction loss:

$$\boldsymbol{L}_{SP}(\boldsymbol{E}, \boldsymbol{W}_3) = \sum_{c=1}^{N_s} -I(c, q_i^1) \log(p_c(\boldsymbol{W}_3(\boldsymbol{E}(x_{t,i})))) \tag{9}$$

where $\boldsymbol{W}_3$ is an initialized fully connected layer, to resize the feature $\boldsymbol{E}(x_{t,i})$ to N_s dimension.

Algorithm 1: Cross Adversarial Consistency Self-Prediction Learning Model

Input: Labeled source samples $\boldsymbol{X}_s = \{x_{s,i}, y_{s,i}\}_{i=1}^{n_s}$, unlabeled target samples $\boldsymbol{X}_t = \{x_{t,i}\}_{i=1}^{n_t}$, and the number of identities N_s in source domain.
Output: The trained encoder $\boldsymbol{E}$.
Step I: CAC learning(Sec. 3.2)
1: Randomly divide images of $\boldsymbol{X}_s$ into two groups according to their camera views.
2: Sample a batch of labeled source data to $\boldsymbol{E}$.
3: **for** *iter* = 1, $\cdots$, $Iteration_1$ **do**
4: Update $\boldsymbol{E}, \boldsymbol{W}_1, \boldsymbol{W}_2$ by Eq. (1) and (3).
5: **end for**
6: **for** *iter* = 1, $\cdots$, $Iteration_2$ **do**
7: Fix $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, update $\boldsymbol{E}$ by Eq. (4) and (3).
8: **end for**
Step II: CSP learning(Sec. 3.3)
9: Load the learned encoder $\boldsymbol{E}$, classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$.
10: **for** *iter* = 1, $\cdots$, $Iteration_3$ **do**
11: Sample a batch of unlabeled data to the model.
12: Assign soft label to unlabeled target samples by Eq. (8)
13: Update $\boldsymbol{E}$ by Eqs. (5) and (9).
14: **end for**

3.4. Final loss for model

Merging the losses of source and target domains into final loss can be expressed as

$$\boldsymbol{L} = \boldsymbol{L}_{ID_1} + \boldsymbol{L}_{Tri} + \lambda_1 \boldsymbol{L}_{ID_2} + \lambda_2 \boldsymbol{L}_c + \lambda_3 \boldsymbol{L}_{SP} \tag{10}$$

where λ_1, λ_2, and λ_3 are hyper-parameters used to control the importance of the CAC learning, identity consistency learning and identity self-prediction learning, respectively. For the target images, $\boldsymbol{L}_c$ and $\boldsymbol{L}_{SP}$ are used to optimize the model exploiting

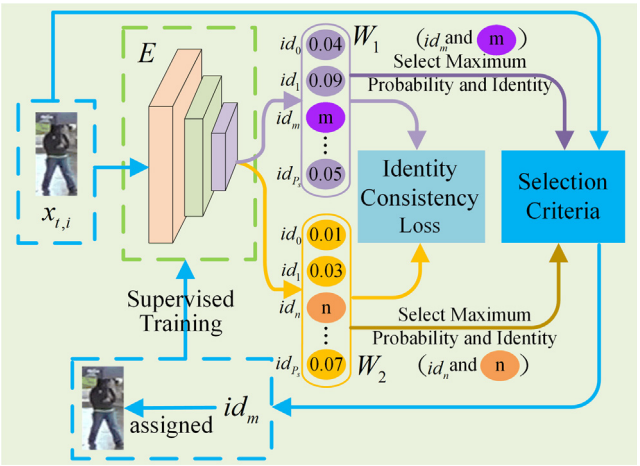

Fig. 3. The proposed CSP learning method. The unlabeled target image is fed to the model to obtain two sets of output. The selected maximum probability and identity are combined with the input image sent to the selection criteria to determine whether it is used for fine-tuning the model.

the unlabeled samples of target domain. The selection and analysis of these hyper-parameters will be discussed later in the section of parameter selection and analysis. A novel framework for UDA person Re-ID is introduced to our method. It consists of CAC learning and CSP learning. The losses of the former are combined with the differences of the same person under different views to the model to learn domain invariant representation. The losses of the latter are used to extract discriminative features from unlabeled target domain utilizing the output discrepancy between the two different identifiers in extracting process. To facilitate understanding, the above process is summarized in Algorithm 1.

4. Experiments

4.1. Datasets and evaluation protocol

We evaluated our model on three large-scale benchmark datasets (Market1501 [32], DukeMTMC-reID [33], and MSMT17 [8]) and two small-scale benchmark datasets (PRID2011 [34] and GRID [35]), comparing the results with those of state-of-the-art methods. Fig. 4 shows some randomly chosen images from each evaluation dataset.

Market1501 comprises 32,668 labeled images of 1,501 pedestrians captured by six cameras. All pedestrians were automatically detected and cropped with bounding-boxes from the video sequences. In agreement with the standard settings [32], all the images in this dataset were divided into two parts: 12,936 labeled images of 751 pedestrians for training, and 19,732 images of 750 pedestrians for testing.

DukeMTMC-reID contains 36,411 images of 1,404 pedestrians collected from eight cameras in winter. Following the split setting in [36], 16,522 images of 702 pedestrians randomly selected from the dataset were used for training. All images of the remaining 702 pedestrians were used for testing: 2,228 as query images and 17,661 as gallery images (For simplicity, this dataset is hereafter called Duke.).

MSMT17 is currently the largest dataset for person Re-ID. It contains 126,441 bounding boxes with 4101 identities, captured by 15 cameras over four days. The bounding boxes are predicted by Faster RCNN [36]. To improve the efficiency of the algorithm, according to the setting in [8], the data were divided into training images and test images in a 1:3 ratio. The training set contained 32,621 bounding boxes with 1,041 identities; the reminder belonged to the test set. **MSMT17** is one of the most challenging Re-ID datasets because the images it contains were collected over a long period (four days) and cover complicated scenarios, varied illumination, and different background conditions.

PRID2011 contains 934 identities captured by two different static surveillance cameras. One camera captured one image each of 385 identities; the other did the same for 749 identities. There are 200 identities recorded in both cameras. These 200 identities were randomly divided into two groups. A group of 100 identities with 200 images was used for training, and the remaining group of 100 identities with 100 images recorded by a single camera was used as a probe set. The 649 additional identities in the other camera view were used as a gallery set.

GRID comprises 250 person-image pairs extracted from six disjoint camera views of a busy underground station. Each pair contains two images captured by two different cameras. According to the setting in [37], 125 identities with 250 images were randomly sampled for training, and 250 images of the remaining 125 people together with 775 interference images were used for testing. More specifically, the query set included 125 images of 125 identities, while the gallery set contained the remaining 125 images of 125 pairable individuals, along with 775 interfering images. In our experiments, all of the interfering images were considered a single identity because they do not correspond to any of the query images. It should be

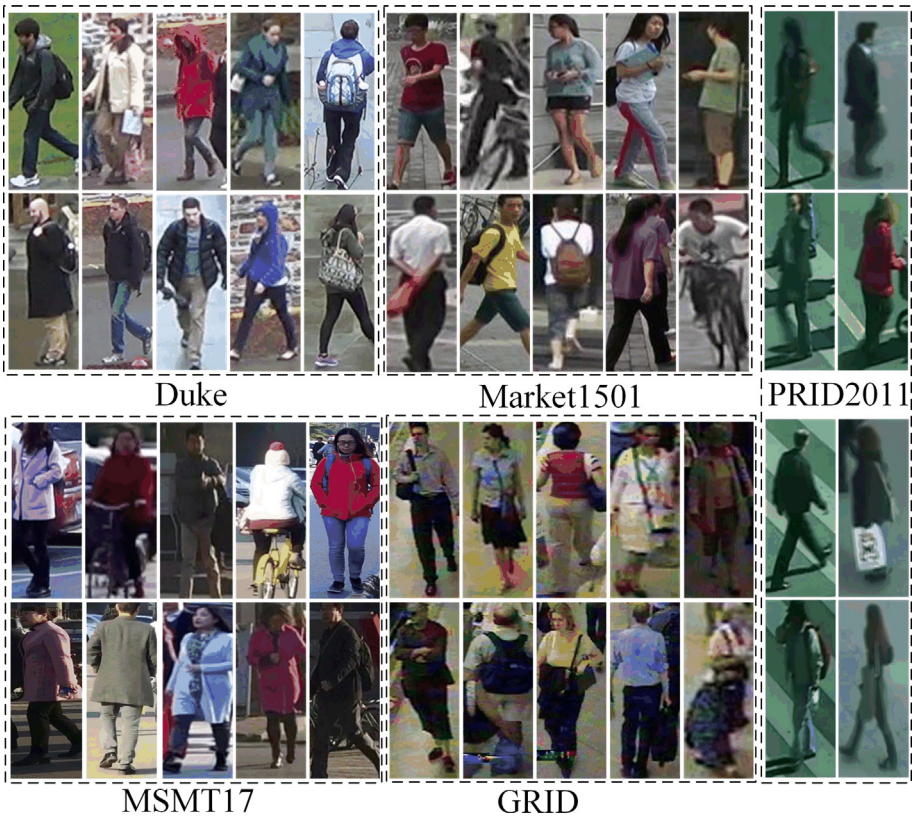

Fig. 4. Randomly sampled person images from different datasets: Duke [33], Market1501 [32], MSMT17 [8], GRID [35], PRID2011 [34].

noted that the labels of target domain samples during training are not used. Table 1 summarizes the settings used for the various datasets.

4.1.1. Evaluation protocol

In our experiment, we employed as evaluation protocols the Cumulative Match Characteristic (CMC), reported by rank-1, rank-5, and rank-10 accuracy, and the mean Average Precision (mAP). For all datasets, we used the single-query setting to retrieve person images from the gallery set. Our method does not need any post-processing such as re-ranking or fine-tuning with predicted pseudo-labels.

4.2. Implementation details

4.2.1. Network settings

We resized the input images to 256×128 and applied random cropping and horizontal flipping to all images for data augmentation. Our algorithm was implemented on the Pytorch platform with an NVIDIA Tesla P100 GPU with 16 GB memory. ResNet-50 pre-trained on ImageNet was used as the backbone in our method; it was followed by a Global Average Pooling (GAP) to resize the features into a 2,048-dimensional vector. In addition, a pair of Batch Normalization (BN) layers were added after the GAP, with the eps and momentum set to 0.00005 and 0.1, respectively. After each BN layer, we added Fully Connected (FC) layers with no bias as identifiers: their output dimension was the number of identities in the source domain. Following the procedure in [30], we added instance-normalization layers before the batch normalization so as to make the learned features more stable. In testing, we extracted the feature after the GAP and used Euclidean distance for similarity measurement.

4.2.2. Optimization

In the training process, we set the batch size to 16 for both source and target datasets. Note that we set $\boldsymbol{P} = 4$ and $\boldsymbol{K} = 4$ to meet the requirements of the triplet loss for the source dataset. We trained the model over 240 epochs in total. In CAC learning, the training first lasts 80 epochs for minimizing the Eq. (1). Subsequently, only the trained encoder $\boldsymbol{E}$ was further optimized by minimizing Eq. (4); this process lasted for 30 epochs. We repeated this adversarial training process for 60 epochs to optimize Eq. (1), and then for 40 epochs to optimize Eq. (4). The first 10 of the last 30 epochs were used to optimize the

Table 1
Settings of person Re-ID datasets in performance evaluation. #Cams denotes the number of cameras.

Datasets	#ID	Train		Gallery (Test)		Query (Test)		#Cams
		#ID	#Img	#ID	#Img	#ID	#Img	
Market1501	1501	751	12936	750	19732	750	3368	6
Duke	1404	702	16522	702	17661	702	2228	8
MSMT17	4101	1041	32621	3060	93820	3060	11659	15
PRID2011	934	100	200	649	649	100	100	2
GRID	250	125	250	126	900	125	125	6

model by minimizing Eq. (5); the remaining 20 were used for CSP learning, so that Eqs. (5) and (9) were combined to fine-tune the model. We used the Adam optimizer [38] with initial learning rate 1.17×10^{-4} and weight decay 0.0005 to optimize the model. Following the settings in [39], we employed the warmup strategy [40] to adjust the learning rate linearly. Specifically, the learning rate was increased linearly from 1.17×10^{-4} to 1.3×10^{-4} during the first 30 epochs and lowered to 1.3×10^{-5} at the 31st epoch. From 31 to 55 epochs, the learning rate increased from 1.3×10^{-5} to 1.4×10^{-5}; it was lowered to 1.4×10^{-6} after 55 epochs. Finally, the learning rate was increased to 2.3×10^{-6} at the 239th epoch. These settings are suitable for all experiments on the datasets described above.

4.3. Comparison with state-of-the-art

4.3.1. Experiments on larger scale re-ID datasets

In this section, we compare the proposed method with some state-of-the-art UDA person-Re-ID methods on six tasks: Duke→Market1501, Market1501→Duke,

MSMT17→Market1501, MSMT17→Duke, Market1501→MSMT17, and Duke→MSMT17. In these settings, A→B means that dataset A is used as the source domain and dataset B as the target domain. In the experiments Duke→Market1501 and Market1501→Duke, the methods compared included both domain-invariant feature-representation methods and style-transfer-based methods. The former include TJ-AIDL [19], ATNet [10], CAMEL[41], CFSM [42], FMC [43], PAUL [16], CDIL[25], and CAL-CCE [15], while the latter include PT-GAN [8], CamStyle [44], HHL[45], SSAE [46], CSGLP [9], and SBSGAN [47]. The performance of the different methods are listed in Table 2.

From Table 2, we can observe that the rank-1 accuracy and mAP achieved by our method are superior to those of other methods on Duke→ Market1501 and Market1501→Duke. The rank-1/mAP accuracy of our method reaches 69.4%/36.9% (57.5%/37.0%) for the task of Duke→Market1501(Market1501→Duke), surpassing the current best result by 2.7%/0.1% (1.4%/0.3%). The results shown in Table 2 demonstrate that the method proposed in this paper outperforms all the style-transfer-based methods by a large margin. Compared with the best style-transfer-based method (SBSGAN [47]), our method improved the recognition rate on rank-1/mAP from 58.5%/27.3% to 69.4%/36.9% for Duke→Market1501, and from 53.5%/30.8% to 57.5%/37.0% for Market1501→Duke. The performances of other GAN-based methods, such as PTGAN [8], TJ-AIDL [19], and ATNet [10], are generally lower than that of our method, possibly because they need an additional network to generate images with different styles. However, the generated images are usually noisy, and the ID information contained in the source images may be changed in the transferred results, leading to unsatisfactory performance. Unlike these methods, ours is an end-to-end model needs no auxiliary network and, thus, has no additional computing cost.

Methods such as UDA-TP [49], ACT [12], and MMT [28] adjust the performance of the model by using a clustering algorithm that predicts the pseudo-label for target samples. We do not do this, and therefore, we cannot directly compare our method with these algorithms. Later, we will evaluate the applicability of the proposed method and clustering-based domain-adaptation methods to interference datasets.

To evaluate the effectiveness of the proposed algorithm more comprehensively, the larger dataset MSMT17 was used as the source domain in the second experiment, and Market1501 and Duke as the target domains. In this experiment, we evaluated the performance of our method against three strong competitors (CASCL [48], MAR [14], and CDIL [25]) on the tasks of MSMT17→Market1501 and MSMT17→Duke. The recognition accuracy of the different methods in rank-1, rank-5, rank-10, and mAP with these settings is shown in Table 3. The table shows that our method surpasses CASCL [48] and CDIL [25] by 7.3% and 13.6% (8.7% and 7.3%), respectively, in rank-1 accuracy for the task of MSMT17→Market1501 (MSMT17→Duke). Compared with the second-best method, ours improves the rank-1 accuracy from 67.7% to 72.7% and mAP from 40% to 41.0% when tested on Market1501. For MSMT17→Duke, the proposed approach also outperforms CASCL [48] and CDIL [25] by a large margin and surpasses MAR [14] in Rank-1 accuracy. In the comparison of the mAP in task MSMT17→Duke, MAR [14] is 0.6% higher. In the two cross-domain tasks, the mAP of the proposed approach is slightly lower on MSMT17→Duke than that of MAR [14]; this is because we used fixed parameters in all experiments, which may not be optimal for a specific dataset.

To validate further advantages of the proposed algorithm, in the third group of experiments, we set MSMT17 as the target dataset, Market1501 and Duke as the source domains. Our method was compared with two state-of-the-art methods: PTGAN [8] and ECN [50]. This group of experiments was more challenging because the scale of the training sets (i.e. Mar-

Table 2
Experimental results of the proposed methods and state-of-the-art methods with the settings of Duke→Market1501 and Market1501→ Duke. Here, mAP denotes mean average precision and "–" denotes not reported. Bold face indicates the optimal value.

Methods	Duke→Market1501				Market1501→Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
PTGAN [8]	38.6	–	66.1	–	27.4	–	50.7	–
CAMEL [41]	54.5	–	–	26.3	42.0	–	–	21.0
TJ-AIDL [19]	58.2	–	–	26.5	44.3	–	–	23.0
ATNet [10]	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
CamStyle [44]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
SBSGAN [47]	58.5	–	–	27.3	53.5	–	–	30.8
CFSM [42]	61.2	–	–	28.3	49.8	–	–	27.3
HHL [45]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
FMC [43]	63.4	79.5	84.8	32.4	48.0	62.3	68.1	27.8
CAL-CCE [15]	64.3	–	–	34.5	55.4	–	–	36.7
CASCL [48]	64.7	80.2	85.6	35.6	51.5	66.7	71.7	30.5
PAUL [16]	66.7	–	–	36.8	56.1	–	–	35.7
SSAE [46]	60.7	–	–	26.6	50.2	–	–	28.1
CDIL [25]	57.2	73	80.0	27.4	–	–	–	–
CSGLP [9]	61.2	77.5	83.2	31.5	47.8	62.3	68.3	27.1
Ours	69.4	82.8	87.3	36.9	57.5	71.2	75.3	37.0

Table 3
Performance (%) comparison of the proposed method and various state-of-the-art methods with the settings of MSMT17→Market1501 and MSMT17→Duke. "–" denotes not reported. Bold face indicates the optimal value.

Methods	MSMT17→Market1501			
	Rank-1	Rank-5	Rank-10	mAP
CASCL (ICCV'19) [48]	65.4	80.6	86.2	35.5
MAR (CVPR'19) [14]	67.7	81.9	–	40.0
CDIL (WACV'20) [25]	59.1	75.4	–	30.3
Ours	72.7	85.1	88.8	41.0
Methods	MSMT17→Duke			
	Rank-1	Rank-5	Rank-10	mAP
CASCL (ICCV'19) [48]	59.3	73.2	77.8	37.8
MAR (CVPR'19) [14]	67.1	79.8	–	48.0
CDIL (WACV'20) [25]	60.7	74.7	–	39.1
Ours	68.0	80.3	84.3	47.4

ket1501 and Duke) is smaller than that of the target domain MSMT17, and therefore, there are only a few methods to report the recognition results on the settings of Market1501→MSMT17 and Duke→MSMT17, except for PTGAN [8] and ECN [50]. From Table 4, we can see that our method outperforms both these methods by a wide margin. On Market1501→MSMT17, our method improves the recognition rate of ECN [50] for rank-1/mAP from 25.3%/8.5% to 28.0%/10.7%, and on Duke→MSMT17 from 30.2%/10.0% to 36.6%/13.2%. These results demonstrate the consistent superiority and robustness of our method over its competitors.

4.3.2. Experiments on small-scale re-ID datasets

The experiments above demonstrate the validity and superiority of our method on large-scale Re-ID datasets. However, with small-scale datasets, does our method still perform well? To answer this question, two challenging small-scale datasets (PRID2011 and GRID) were used as the target domains, and Market1501, Duke, and MSMT17 as the labeled source domains. The proposed method was again compared to various state-of-the-art methods. Table 5 reports the results of the different methods on the PRID2011 and GRID datasets.

It can be seen that our method clearly surpasses the others in rank-1 and mAP accuracies on PRID2011 and GRID. Specifically, on PRID2011, our method outperformed the handcrafted feature-based methods SSAE [46] and AIESL [24] by at least 11.9% and 5.5% in rank-1 accuracy. On GRID, our method is superior to the other methods in rank-1 and mAP accuracies. As shown in Table 5, when Market1501 is used as the source domain, the rank-1 accuracy of our method reaches 42.4% and mAP reaches 49.2%, showing improvements in recognition rate and mAP of 21.6% and 20.3% compared with the second best method. In addition, as Table 5 indicates, the improvement of our method on rank-1 accuracy also reaches 15.2% (versus 16.8% and 32.0%) and 14.0% (versus 18.0% and 32.0%) compared with the newly released methods AIESL [24] and SSAE [46].

In recent years, the pseudo-label prediction algorithm based on clustering has been very extensively studied [49,12,28]. This kind of algorithm has excellent performance on the Market1501 and Duke datasets, and some approaches even achieve the recognition accuracy of supervised methods. For example, MMT [28]achieved a rank-1 accuracy of more than 87.7% on Market1501 and 78.0% on Duke. However, this extremely high performance was achieved under the assumption that there

Table 4
Performance (%) comparison of the proposed method and the state-of-the-art methods on Market1501→MSMT17 and Duke→MSMT17. "–" denotes not reported. Bold face indicates the optimal value.

Methods	Market1501→MSMT17			
	Rank-1	Rank-5	Rank-10	mAP
PTGAN (CVPR'18) [8]	10.2	–	24.4	2.9
ECN (CVPR'19) [50]	25.3	36.3	42.1	8.5
Ours	29.3	40.2	45.9	10.5
Methods	Duke→MSMT17			
	Rank-1	Rank-5	Rank-10	mAP
PTGAN (CVPR'18) [8]	11.8	–	27.4	3.3
ECN (CVPR'19) [50]	30.2	41.5	46.8	10.0
Ours	37.0	49.9	55.6	13.3

Table 5
Performance comparison with state-of-the-art methods on PRID2011 and GRID. Ours(xx) means "xx" is used as the source dataset. "–" denotes not reported. Bold face indicates the optimal value.

Methods	Target:PRID2011		Target:GRID	
	Rank-1	mAP	Rank-1	mAP
TJ-AIDL (CVPR'18) [19]	34.8	–	–	–
JSLAM (TPAMI'18) [23]	25.6	–	–	–
PTGAN (CVPR'18) [8]	33.5	–	–	–
ATNet (CVPR'19) [10]	24.0	–	–	–
SSAE (PR'20)[46]	29.1	–	18.0	–
AIESL (TCSVT'20) [24]	35.5	–	16.8	–
UDA-TP (PR'20)[49]	22.(22.0)	31.3(33.3)	15.2(27.2)	24.5(35.6)
ACT (AAAI'20) [12]	24.0(13.0)	37.2(21.9)	20.0(13.6)	26.3(20.4)
MMT (ICLR'20) [28]	25.0(25.0)	33.9(33.9)	20.8(32.8)	28.9(41.3)
Ours (MSMT17)	47.0(62.0)	55.1(70.6)	34.4(43.2)	41.9(51.0)
Ours (Duke)	48.0 (54.0)	56.7(65.5)	32.0(41.6)	39.1(49.3)
Ours (Market1501)	41.0(45.0)	48.5(54.6)	42.4(43.2)	49.2(52.8)

are a large number of positive samples in the target datasets. This assumption is obviously inconsistent with the actual situation, in which the scarcity of positive samples in real-world scenarios makes it very challenging for this kind of clustering algorithm to predict pseudo-labels correctly. Even if these algorithms can select the positive samples, they cannot enough to improve significantly the performance of a deep learning model with large-scale parameters. In order to prove that the proposed method is more practical than clustering-based algorithms, we additionally took out 300 and 400 interference images from the gallery of PRID and GRID respectively and added them to their training set to increase the diversity of negative samples. The results are shown in brackets in Table 5, it can be seen that UDA-TP [49], ACT [12], and MMT [28], although they have high performance in Market1501 and Duke, can only achieve a Rank-1 accuracy of approximately 20% on the interference dataset, while the proposed method can reach a Rank-1 accuracy of about 40%. From the analysis, we can conclude that the proposed method is more scalable, and not only performs well on large datasets, but also competitively on interference datasets.

4.3.3. Discussion of training efficiency

The time consumed by the proposed approach was compared with some classical methods such as PAUL [16], CAMEL [41] and HHL [45] on the tasks of Duke→Market1501 and Market1501→Duke. The different experiments were all run on the same device using the codes published by their respective authors. The experimental results are shown in Table 6. Compared with PAUL [16], CAMEL [41], and HHL [45], the proposed approach has higher computing efficiency. This is mainly because the proposed method does not measure the similarity of the local blocks of the image like PAUL, nor does it use the image after style transfer to train the model like HHL. Thus, it can significantly reduce the computing burden of the model.

4.4. Ablation study

As mentioned previously, the method proposed in this paper consists of two modules: CAC learning and CSP learning. CSP learning includes identity-consistency learning and identity self-prediction learning. In our method, we train the baseline only with $\boldsymbol{L}_{ID_1}$ and $\boldsymbol{L}_{Tri}$, and then evaluate the effectiveness of each component of the proposed method by adding them to the baseline model one by one. Market1501 and Duke are utilized for testing. The experimental results for different combinations are reported in Table 7.

Table 6
Comparison of time cost of different methods on Duke→Market1501 and Market1501→Duke. "–" denotes not reported.

Methods	Duke→Market1501		Market1501→Duke	
	Training epoch	Training time	Training epoch	Training time
PAUL [16]	120	≈12.3 h	120	≈18.5 h
CAMEL [41]	–	≈13.7 h	–	≈13.7 h
HHL [45]	60	≈21.0 h	60	≈21.0 h
Ours	240	≈6.1 h	240	≈5.7 h

Table 7
Ablation study of the proposed model on Market1501→Duke and Duke→Market1501 (%). Baseline trained on labeled source dataset with $\boldsymbol{L}_{ID_1}$ and $\boldsymbol{L}_{Tri}$. $\boldsymbol{L}_{ID_2}$ loss function used in cross adversarial consistency learning. $\boldsymbol{L}_c$ loss function used in ientity-consistency Learning. $\boldsymbol{L}_{SP}$ loss function used in identity self-prediction learning. Bold face indicates the optimal value.

Methods	Market1501→Duke				Duke→Market1501			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Baseline ($\boldsymbol{L}_{ID_1}$ + $\boldsymbol{L}_{Tri}$)	52.3	66.4	71.8	32.8	66.3	81.4	86.2	34.3
Baseline + $\boldsymbol{L}_{ID_2}$	55.4	68.3	73.6	34.9	68.3	82.3	86.8	35.7
Baseline + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_c$	55.9	69.6	74.1	35.8	69.0	82.7	87.1	36.8
Baseline + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_{SP}$	56.6	70.9	75.6	36.7	68.6	81.6	86.5	36.5
Baseline + $\boldsymbol{L}_c$ + $\boldsymbol{L}_{SP}$	55.1	68.7	73.1	34.7	68.1	82.1	86.5	36.2
Baseline + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_c$ + $\boldsymbol{L}_{SP}$	57.5	71.2	75.3	37.0	69.4	82.8	87.3	36.9

4.4.1. Effectiveness of CAC learning

To evaluate the effectiveness of CAC learning, either Market1501 is used as the labeled source domain and Duke as the target domain, or vice versa. We cross-use the two classifiers on the trained baseline model for CAC learning. From the comparison between "*Baseline* ($\boldsymbol{L}_{ID_1}$ + $\boldsymbol{L}_{Tri}$)" and "*Baseline* + $\boldsymbol{L}_{ID_2}$", we observe that the improvement of the rank-1 accuracy reaches 3.1% and 2.0% when tested on Duke and Market1501 datasets. The mAP is improved from 32.8% to 34.9% and 34.3% to 35.7% for the settings of Market1501→Duke and Duke→Market1501. These results verify that the proposed CAC learning method is very effective in learning domain-invariant features for UDA person Re-ID.

4.4.2. Effectiveness of identity consistency learning

The effectiveness of identity-consistency learning is evaluated by adding the loss function $\boldsymbol{L}_c$ into "*Baseline* + $\boldsymbol{L}_{ID_2}$". As shown in Table 7, the recognition accuracy improves from 68.3% to 69.0% in rank-1 and from 35.7% to 36.8% on mAP when Duke is used as the source domain and Market1501 as the target domain. For the task of Market1501→Duke, the improvements in rank-1 accuracy and mAP reach 0.5% and 0.9%, respectively. The improvement resulting from identity-consistency learning can ensure that the learned features of the same person predict the same identity. Compared with CAC learning, the improvement achieved by identity-consistency learning is not significant, but nevertheless it can improve the robustness of the learned domain-invariant feature.

4.4.3. Effectiveness of identity self-prediction learning

Identity-self-prediction learning is also integrated into "Baseline + $\boldsymbol{L}_{ID_2}$". We evaluated its effect by comparing the performance of "*Baseline* + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_{SP}$" with that of "*Baseline* + $\boldsymbol{L}_{ID_2}$". As reported in Table 7, after adding the self-prediction learning for training, the rank-1 accuracy and mAP of the model are improved by 0.3% and 0.8% on Duke→Market1501 and 1.2% and 2.2% on "Market1501→Duke". This can be explained by the fact that the self-prediction learning effectively exploits the discriminative information from unlabeled target domains to further refine feature learning on unlabeled datasets.

As reported in Table 7, the combination of CAC learning, identity-consistency learning and identity-self-prediction learning (i.e., "*Baseline* + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_c$ + $\boldsymbol{L}_{SP}$") is more effective than "*Baseline* + $\boldsymbol{L}_{ID_2}$", "*Baseline* + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_c$", or "*Baseline* + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_{SP}$" separately, demonstrating that the three submodules are complementary to each other. This is easy to understand. CAC learning guarantees that the learned features are domain invariant, rendering the model generalizable; identity-consistency learning ensures that the domain-invariant features are discriminative; identity-self-prediction learning exploits the discriminative information from the target domain, further refining the learned feature. Therefore, the concatenated sub-modules "*Baseline* + $\boldsymbol{L}_{ID_2}$ + $\boldsymbol{L}_c$ + $\boldsymbol{L}_{SP}$" are more effective for discrimination feature-learning than any submodule individually.

4.5. Parameter selection and analysis

We conducted a series of experiments to investigate the effects of the hyper-parameters λ_1, λ_2 and λ_3 on our model. By default, the value of one parameter was altered and the others remained unchanged to reveal the effect of this parameter on the performance. The experiments with different parameters were conducted on the Market1501 and Duke datasets.

Once the parameters are selected for both datasets, the parameter settings remain the same throughout the remainder of the paper.

4.5.1. Effect of the parameter λ_1 of $\boldsymbol{L}_{ID_2}$

In Eq. (9), λ_1 is used to adjust $\boldsymbol{L}_{ID_2}$. Figs. 5 (a) and (b) show the effects of λ_1 on our framework. When $\lambda_1 = 0$, the rank-1 accuracy and mAP of our method are lowered by approximately 55% and 35%, respectively, for the task of M→D; they are lowered by around 67% and 36% for the task of D→M. As shown in Fig. 5(a) and (b), the proposed method achieves better performance on rank-1 accuracy for these two tasks when CAC learning ($\lambda_1 > 0$) is added to our model. This demonstrates the effectiveness of CAC learning and the importance of domain-invariant feature extraction for UDA person Re-ID. More specifically, when $\lambda_1 = 1$, our approach achieves the best results for rank-1 and mAP on both datasets, as presented in Figs. 5 (a) and (b).

4.5.2. Effect of the parameter λ_2 of $\boldsymbol{L}_c$

Figs. 5(c) and (d) show the effect as λ_2 varies from 0 to 10. The rank-1 accuracy is elevated greatly when $\lambda_2 \in [0.1, 0.5]$ for both M→D and D→M, and reaches its highest value when $\lambda_2 = 0.5$. It can be seen that when λ_2 is larger than 2, the performance of our model suffers significantly, especially in terms of mAP, which drops from 37% to 27% as λ_2 changes from 2 to 10 on the Duke dataset. Considering the accuracy of both rank 1 and mAP, the most acceptable value for λ_2 is 0.5, although this does not have the best rank-1 accuracy.

4.5.3. Effect of the parameter λ_3 of $\boldsymbol{L}_{SP}$

Another experiment was conducted to evaluate the effect of λ_3. Figs. 5(e) and (f) show the rank-1 accuracy and mAP when different values of λ_3 were applied. It can be seen that our approach achieved its best rank-1 accuracy and mAP when λ_3 is approximately 0.08. Moreover, the decrease of mAP is greater than that of the rank-1 when λ_3 is greater than 0.5. This phenomenon is probably caused by over-fitting. A smaller λ_3 means a smaller role played by self-prediction learning in feature

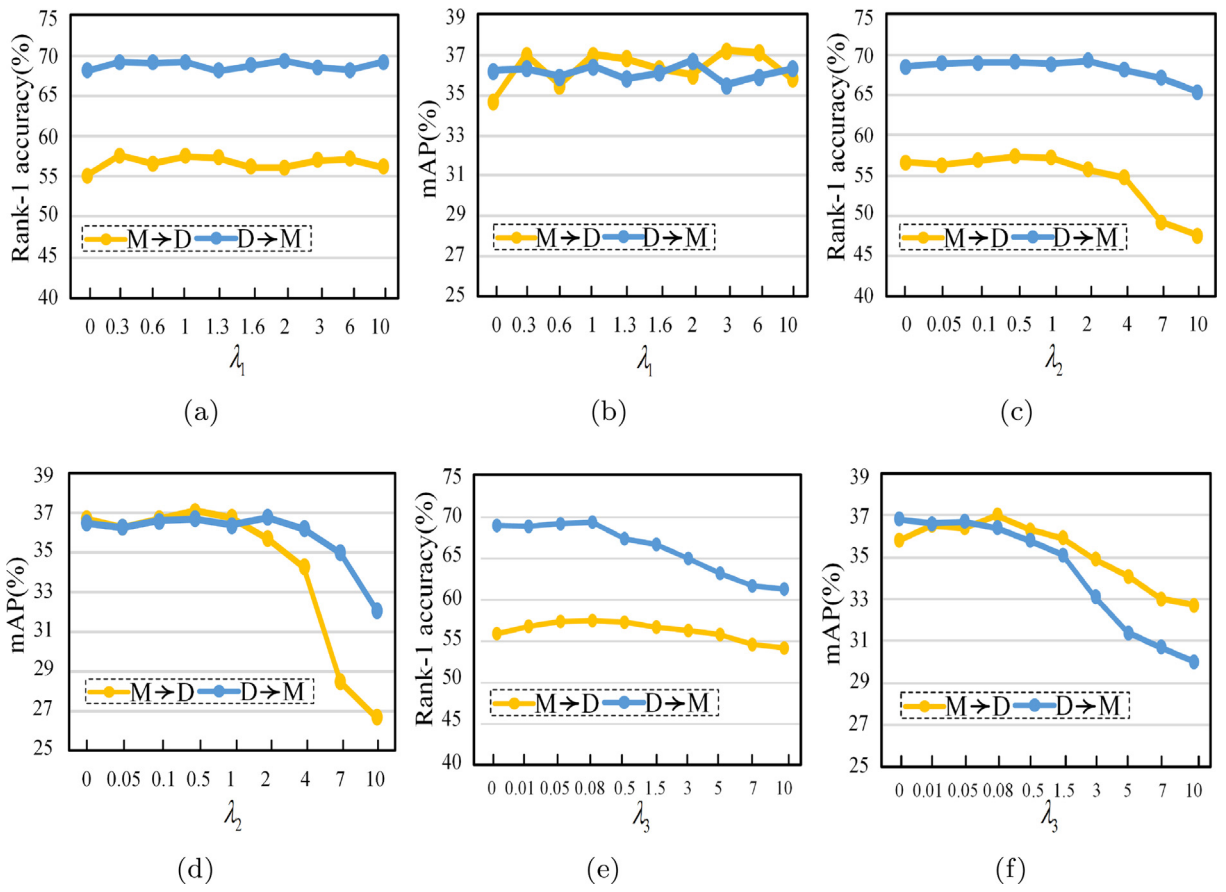

Fig. 5. Performance analysis under different values of hyper-parameters. (a) Rank-1 accuracy under varying λ_1, (b) mAP accuracy under varying λ_1, (c) Rank-1 accuracy under varying λ_2, (d) mAP accuracy under varying λ_2, (e) Rank-1 accuracy under varying λ_3, (f) mAP accuracy under varying λ_3. M denotes the Market1501 dataset, and D the Duke dataset.

learning and vice versa. Note that, in self-prediction learning, to exploit the discriminative information from target samples, samples are selected from the target domain and the source domain to form pseudo-positive sample pairs. Such pairs are used to fine tune the pre-trained model. Because the target sample and the source-domain sample are not from the same person, a pseudo-positive sample pair is not a real positive sample pair, but simply a combination of two very similar pedestrian images. Therefore, self-prediction learning cannot be assigned a large weight. Nevertheless, experiments have proved that self-prediction learning can help to improve the performance of the algorithm to some extent. This is because some discriminative information of the target domain is exploited and used in the learning of features.

5. Conclusion

This paper focused on the problems of domain-invariant discriminative feature learning for UDA person-Re-ID tasks and presented a novel method that includes both CAC and CSP learning. CAC learning can exploit fully the robust representation through adversarial training, while CSP learning is able to enhance discrimination of the features of unlabeled data by exploiting the similarities between source data and target data. The detailed structure of this method and its associated loss functions were presented. The results of the experiments conducted on five person Re-ID benchmark datasets indicate that our model can obtain state-of-the-art performance. An ablation study verified the effectiveness of each sub-module. The method is highly effective and superior to most of its competitors. It is also very practical by virtue of its availability design in real scenarios. In the future, further effort should be made to improve the recognition performance for practical applications.

CRediT authorship contribution statement

Huafeng Li: Methodology, Conceptualization, Formal analysis, Writing - review & editing. **Jian Pang:** Methodology, Software, Data curation, Validation, Writing - original draft. **Dapeng Tao:** Investigation, Resources, Writing - review & editing. **Zhengtao Yu:** Investigation, Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61966021, Grant 61772455, Grant 61562053, National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100, Yunnan provincial major science and technology special plan projects: digitization research and application demonstration of Yunnan characteristic industry, under Grant: 202002AD080001, Yunnan Natural Science Funds under Grant 2018FY001(-013) and Grant 2019FA-045, Yunnan University Natural Science Funds under Grant 2018YDJQ004.

References

[1] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, D. Tao, Unsupervised semantic-preserving adversarial hashing for image search, IEEE Trans. Image Process. 28 (8) (2019) 4032–4044.
[2] E. Yang, C. Deng, C. Li, W. Liu, J. Li, D. Tao, Shared predictive cross-modal deep quantization, IEEE Trans. Neural Networks Learn. Syst. 29 (11) (2018) 5292–5303.
[3] J. Yu, D. Tao, M. Wang, Y. Rui, Learning to rank using user clicks and visual features for image retrieval, IEEE Trans. Cybern. 45 (4) (2014) 767–779.
[4] C. Hong, J. Yu, J. Zhang, X. Jin, K.-H. Lee, Multimodal face-pose estimation with multitask manifold deep learning, IEEE Trans. Industr. Inf. 15 (7) (2018) 3952–3961.
[5] J. Yu, M. Tan, H. Zhang, D. Tao, Y. Rui, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Trans. Pattern Anal. Mach. Intell..
[6] H. Li, J. Xu, Z. Yu, J. Luo, Jointly learning commonality and specificity dictionaries for person re-identification, IEEE Trans. Image Process. 29 (2020) 7345–7358.
[7] C. Gao, Y. Chen, J.-G. Yu, N. Sang, Pose-guided spatiotemporal alignment for video-based person re-identification, Inf. Sci. 527 (2020) 176–190.
[8] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 79–88..
[9] C. Ren, B. Liang, P. Ge, Y. Zhai, Z. Lei, Domain adaptive person re-identification via camera style generation and label propagation, IEEE Trans. Inf. Forensics Secur. 15 (2020) 1290–1302.
[10] J. Liu, Z. Zha, D. Chen, R. Hong, M. Wang, Adaptive transfer network for cross-domain person re-identification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 7202–7211.
[11] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 5157–5166.
[12] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, S. Li, Asymmetric co-teaching for unsupervised cross-domain person re-identification, in: The AAAI Conference on Artificial Intelligence (AAAI), 2020.
[13] X. Zhang, J. Cao, C. Shen, M. You, Self-training with progressive augmentation for unsupervised cross-domain person re-identification, The IEEE International Conference on Computer Vision (ICCV) (2019) 8222–8231.

[14] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, J. Lai, Unsupervised person re-identification by soft multilabel learning, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2148–2157.
[15] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, Y. Gao, A novel unsupervised camera-aware domain adaptation framework for person re-identification, The IEEE International Conference on Computer Vision (ICCV) (2019) 8080–8089.
[16] Q. Yang, H. Yu, A. Wu, W. Zheng, Patch-based discriminative feature learning for unsupervised person re-identification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3633–3642.
[17] J. Liu, W. Li, H. Pei, Identity preserving generative adversarial network for cross domain person re-identification, arXiv:1811.11510v1. 2018..
[18] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, 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, 2018, pp. 994–1003.
[19] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2275–2284.
[20] H. Li, X. He, Z. Yu, J. Luo, Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior, Inf. Sci. 523 (2020) 14–37.
[21] Z. Zhu, H. Yin, Y. Chai, Y. Li, G. Qi, A novel multi-modality image fusion method based on image decomposition and sparse representation, Inf. Sci. 432 (2018) 516–529.
[22] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, D. Tao, Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion, IEEE Trans. Instrum. Meas. 69 (4) (2020) 1082–1102.
[23] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, T. Huang, Joint semantic and latent attribute modelling for cross-class transfer learning, IEEE Trans. Pattern Anal. Mach. Intell. 40 (7) (2018) 1625–1638.
[24] H. Li, S. Yan, Z. Yu, D. Tao, Attribute-identity embedding and self-supervised learning for scalable person re-identification, IEEE Trans. Circuits Syst. Video Technol. 30 (10) (2020) 3472–3485.
[25] Y. Yuan, W. Chen, T. Chen, Y., Z. Ren, Z. Wang, G. Hua, Calibrated domain-invariant learning for highly generalizable large scale re-identification, in: The IEEE Winter Conference of Applications on Computer Vision (WACV 2020), 2020, pp. 3589–3598..
[26] C. Tay, S. Roy, K. Yap, Aanet, Attribute attention network for person re-identifications, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 7134–7143.
[27] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, The IEEE International Conference on Computer Vision (ICCV) (2017) 3800–3808.
[28] Y. Ge, D. Chen, H. Li, Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification, The International Conference on Learning Representations (ICLR) (2020) 5157–5166.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.
[30] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: enhancing learning and generalization capacities via ibn-net, The The European Conference on Computer Vision (ECCV) (2018) 464–479.
[31] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imageget, A larage-scale hierarchical image database, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 79–88.
[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, The IEEE International Conference on Computer Vision (ICCV) (2015) 1116–1124.
[33] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, European Conference on Computer Vision Workshops (ECCVW) (2016) 17–35.
[34] M. Hirzer, C. Beleznai, P. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian Conference on Image Analysis, Springer, 2011, pp. 91–102.
[35] C. Loy, T. Xiang, S. Gong, Time-delayed correlation analysis for multi-camera activity understanding, Int. J. Comput. Vision 90 (1) (2010) 106–129.
[36] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99..
[37] H. Li, J. Xu, J. Zhu, D. Tao, Z. Yu, Top distance regularized projection and dictionary learning for person re-identification, Inf. Sci. 502 (2019) 472–491.
[38] D. Kingma, J. Ba, Adam,, A method for stochastic optimization, in: The International Conference for Learning Representations (ICLR), 2015, pp. 1–15.
[39] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019).
[40] X. Fan, W. Jiang, H. Luo, M. Fei, Spherereid: deep hypersphere manifold embedding for person re-identification, J. Vis. Commun. Image Represent. 60 (2019) 51–58.
[41] H.-X. Yu, A. Wu, W.-S. Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 994–1002.
[42] X. Chang, Y. Yang, T. Xiang, T. Hospedales, Disjoint label space transfer learning with common factorised space, in: The AAAI Conference on Artificial Intelligence (AAAI), 2019, pp. 3288–3295.
[43] Z. Zhang, M. Huang, S. Liu, B. Xiao, T. Durrani, Fuzzy multilayer clustering and fuzzy label regularization for unsupervised person re-identification, IEEE Trans. Fuzzy Syst. 28 (7) (2020) 1356–1368.
[44] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: a novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2019) 1176–1190.
[45] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero-and homogeneously, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 172–188.
[46] H. Li, Z. Kuang, Z. Yu, J. Luo, Structure alignment of attributes and visual features for cross-dataset person re-identification, Pattern Recogn. 106 (2020) 107414.
[47] Y. Huang, Q. Wu, J. Xu, Y. Zhong, Sbsgan: suppression of inter-domain background shift for person re-identification, in: The The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9527–9536.
[48] A. Wu, W. Zheng, J. Lai, Unsupervised person re-identification by camera-aware similarity consistency learning, The IEEE International Conference on Computer Vision (ICCV) (2019) 6922–6931.
[49] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, X. Wang, Unsupervised domain adaptive re-identification: Theory and practice, Pattern Recogn. 102 (2020) 107173.
[50] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: exemplar memory for domain adaptive person re-identification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 598–607.