



Domain adaptive person re-identification with memory-based circular ranking

Honggang Chen¹ · Xing Cheng¹ · Nan Guo¹ · Xiaochun Ye¹ · Xiao Xiao¹ · Dongrui Fan¹

Accepted: 8 April 2022 / Published online: 13 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Despite the impressive achievement in supervised person re-identification (re-id), existing supervised approaches mainly focus on exhaustive identity annotations of each image. Their performance will degrade significantly when the test dataset's angles, specifications, and illumination degrees are different from that of the training dataset. In this paper, we propose a novel domain adaptive re-id deep learning method with Memory-Based Circular Ranking (MBCR) to assign labels to each sample in the target domain adaptively. We put forward a reciprocal neighbors label smoothing loss calculated from the generated pseudo-labels to optimize the target domain in a supervised learning manner. Since most person re-id datasets have multiple camera perspectives, the cross-camera invariance loss is proposed to make the model adapt to the variations of images. Massive experiments are enforced to prove the superiority of the proposed model over the baseline. It increases the R@1 of 13.5%, 11.0%, and 5.0% in Market-1501, DukeMTMC-reID, and MSMT17 than baseline separately and reaches State-Of-The-Art (SOTA).

Keywords Person Re-identification · Reciprocal neighbors · Cross-Camera · Domain adaptation

1 Introduction

Person re-identification, also named re-id, [1–10] can be considered as a single-modal retrieval task, which requires to search the underlying suspect of the query person from

the gallery filmed from multi-view cameras. It is a technology that employs computer vision to determine whether a particular suspect is in a video. Because of the broad distribution of the cameras, their angles, specifications, and illumination degrees of the scene where the camera is located are entirely different, which may lead to the appearance characteristics of the same pedestrian vary a lot under different cameras [11]. In addition, the researchers also encounter many other challenges such as low image resolution, pedestrian posture change, and occlusion. In detail, present difficulties can be concluded as:

These authors contributed equally to this work.

✉ Nan Guo
guonan@ict.ac.cn

Honggang Chen
chenhonggang18s@ict.ac.cn

Xing Cheng
chengxing20s@ict.ac.cn

Xiaochun Ye
yexiaochun@ict.ac.cn

Xiao Xiao
xiaoxiao@ict.ac.cn

Dongrui Fan
fandr@ict.ac.cn

- In most cases, images from surveillance cameras are fuzzy. There may be only part of the pedestrian shown in an image. Even with advanced object detection algorithms, it is still at low resolution, so it is impossible to compare the embedding of extracted face features. The only solution is to judge from the aspects of clothes and postures.
- Pedestrian re-id images may be filmed in different periods, and pedestrians' posture and appearance may change to a certain extent. For instance, the images taken in the day and at night will be quite different. Moreover, in many monitoring environments, there is a large flow of people, which is prone to overlap and occlusion of pedestrians.

¹ Chinese Academy of Sciences, Institute of Computing Technology, Street, Beijing, 100190, Zhongguancun Academy of Sciences South Road, China

- The acquisition of pedestrian re-identification dataset involves the issue of personal privacy and security. Pedestrian detection should be conducted among the whole video frame sequence, which is time-consuming. It is crucial to judge whether pedestrians detected in frames are the same identities and then mark them. The whole process is tedious and inefficient. However, the existent supervised re-id models need to be supported by large-scale datasets, which results in the contradiction between model performance and model training cost.

In a word, the above situations have brought significant challenges to the re-id task. Researchers still need to focus practical issues and make long-term efforts.

1.1 Motivation

Due to its important application in security monitoring, person re-id has been widely concerned by academia. However, present algorithms commonly depend on enormous labeled datasets, limiting their availability in practical applications. Even though many traditional supervised algorithms show excellent performance on the benchmarks, they perform poorly on real-world datasets [12–15]. This is because when there is a slight gap between data distributions, the performance of the supervised model will decrease significantly. As the specification of different cameras varies greatly, the appearance of pedestrians is easily affected by factors such as clothing, shelter, posture, and lighting. All the factors mentioned above make pedestrian re-id a hot and challenging task.

This paper solves person re-id in an unsupervised way. It needs to train the neural network on labeled source domain in a supervised way and unlabeled target domain in a unsupervised way. The object is to optimize the performance in the unlabeled domain. The self-adaptive pedestrian re-id algorithm aims to improve the generalization capability and reduce the cost of manual labeling. The regular unsupervised domain adaptation (UDA) bases on the assumption that both the two domains share a identical category distribution [12, 16]. However, it is not correct given to the inherent open set peculiarity [17]. Recently, most UDA methods [18–20] are intending to decrease the distributional discrepancy between the two domains. In addition, some UDA methods leverage generative adversarial networks (GAN) [21] to realize image-to-image domain adaptation, which is essential to augment the pedestrian dataset. Nevertheless, these methods do not focus on the potential label information in the target domain, nor do they address the diversity of camera styles.

In this work, a reciprocal neighbors label is formulated smoothing loss (RNLSL), which is based on the observation

that when the model is trained on the labeled dataset, the highest-ranked returns are more likely to be the same identity as the query. In a word, the matching image of a pedestrian's similar image is more likely to be the person's similar image. Therefore, we can guide the model to be aware of the potential invariance in the target domain by decreasing the discrepancy among each sample and its neighbors. This situation is likely to deviate from the best case: the mismatch is contained in k-nearest neighbors and even has a high ranking [3]. RNLSL is designed to solve the aforesaid problem by pulling the true matches in the k-nearest neighbors and pushing the hard negative samples.

Simultaneously, this paper regards the camera in the unlabeled dataset as a domain and employ style transfer GAN to train a domain adaptive transfer model. Then, MBCR augment the unlabeled domain images with the pre-trained model, which is illustrated in Fig. 1. These generated images can help the model grasp the variations in the appearance of persons due to differences in camera styles. Finally, a cross-camera invariance loss (CCIL) is proposed to further eliminate the interference of camera style on personal identities.

1.2 Contributions

The contributions can be concluded as:

- We demonstrate how to employ the memory-based circular ranking mechanism to generate reliable smooth labels in an unsupervised way and optimize the object with RNLSL. The proposed method is named as MBCR. It does not label the total dataset by clustering the entire target domain, but label the individual samples with the guidance of neighborhood information. MBCR boosts the performance by improving the efficiency of model optimization and avoiding the noise generated by a clustering algorithm.
- A loss optimization named CCIL is proposed to direct the model to focus on cross-camera invariance in the target domain, which increases the model's robustness.
- We have verified the method on three benchmarks, including Market-1501 [22], DukeMTMC-reID [23] and MSMT17 [20]. Experiments on these datasets demonstrate that the proposed approach surpasses the SOTA method.
- It is simple to implement the proposed method and it brings no massive extra parameters and calculation overhead to the training process, so it is suitable for practical application. There is still a long way on how to design appropriate loss based on training samples and their adjacent images. The proposed method provides a wonderful reference for future work.



Fig. 1 Examples of style-transferred images in Market-1501. The Market-1501 benchmark contains six cameras. Every pedestrian filmed by one camera will also be captured by other five cameras

2 Related work

Over the past few decades, most existing supervised re-id works have focused on learning distance metric or subspace [24–28]. With the rise of neural networks, deep learning methods [29–32] are fully applied in person re-id. In this section, we primarily introduce the traditional metric learning methods and the deep learning methods.

2.1 Metric learning Person Re-identification

LMNN [33] constructs a triplet in terms of anchor sample, positive sample, and negative sample. It requires the distance among the embedding of the same category be close enough, and the ones of distinct categories should be distant enough, where the distance is measured by Euclidean distance or Cosine distance. Hinge loss is employed to optimize the objective function, which is a classical convex optimization problem with low complexity, and many scholars later improve on this basis. Chen et al. [34] apply the

quaternion loss for pedestrian re-id for the first time, where the distance among the negative samples is constrained.

KISSME [35] judges whether two pedestrians are identical or not according to the logarithmic probability value. It hypothesizes that the image features of the same pedestrian follow a Gaussian distribution, which means that the mean of the same pedestrian's image features should be 0, and the covariance should be of gaussian distribution. Hao et al. [36] proposed a local similarity metric method based on KISSME to discriminate local regional similarity of pedestrian images, avoiding the similarity conflicts between positive and negative samples, which exist in previous distance measurement.

2.2 Supervised Person Re-identification

He et al. [37] proposed SPP-Net. The traditional neural network utilizes a single-size convolution kernel, which is not capable of capturing the information of different scales in images. The author employs spatial pyramid structure to

extract samples' features, and pyramid pooling structure to make the extracted feature dimensions no longer rely on the size of input images.

Subsequently, a deeper convolutional neural network, VGG-NET, was proposed by Simonyan et al. [38]. Considering that deepening the neural network can boost the fitting capacity of the model. The author proposes to replace the previous 5×5 convolutional kernels with multiple 3×3 convolutional kernels, which increases the depth and improves the performance while maintaining the feature map's receptive field.

However, the methods mentioned above are universal feature extractors of images. Re-id owns its particularity, the challenge described in Section 1 prevents these models from being better at generalization. So OSNet [39] is proposed to attention to the full-scaled feature and advance the re-id to a new level. It designs a convolution block by introducing multiple features flows of different scales, and the feature scale concerned by each feature flow can be adjusted through hyperparameters. Moreover, features of different scales will be uniformly fed into the aggregation module to generate dynamic weights for feature flows of different scales through the full connection layer and conduct the multi-scale feature fusion. For the input images, the feature aggregation module can adaptively focus on an appropriate scale or choose to mix with features from different scales to produce heterogeneous feature scales. In addition to realizing multi-scale feature fusion learning, OSNet is designed with the principle of lightweight and employs deep separable convolution to replace the original 3×3 convolution.

In [32], image pairs are divided into three overlapped image pairs. The cosine measuring function is utilized to jointly perform feature extraction and metric learning through a siamese CNN network. Recently, deep attention mechanisms [40, 41] has been proposed to solve the problems of lighting, occlusions, and back-ground variations. In [40], a dual attention matching network is designed to search the implicit context representations and then compares them simultaneously. [41] adopts the pose-guided part attention mechanism to reduce noise interference. In addition, [42] uses full convolutional siamese networks to calculate visual similarity at different levels and combines multiple levels of information to improve the robustness of matching. However, these methods lack effective guidance for unlabeled datasets, which leads to poor scalability model in realistic deployment.

2.3 Unsupervised domain adaptation

UDA can effectively solve the learning problem of the distribution inconsistency between the two domains. When their categories are the same, an alignment operation is

realized by reducing the maximum mean difference (MMD) [43] in Reproductive Kernel Hilbert Space (RKHS) [44, 45]. However, in most scenarios, unknown categories exist in the target domain. Aiming to solve this, Busto and Grall [17] propose the open set domain adaptation. They project the estimate feature from the source labeled domain to target unlabeled domain by assigning images within the target domain to certain categories in the source domain. Recently an adversarial learning framework [46] is proposed to achieve the style transformation of the source domain by use of cycle-consistency loss. This paper also utilizes the domain style transfer GAN to bridge the gap between domains.

2.4 Domain adaptive person re-identification

Several unsupervised approaches utilize source domain to initialize a pre-trained model and mine the potential label information by unsupervised clustering on the unlabeled target domain [12–15]. Nevertheless, they do not use labeled source images to continue to refine the pre-trained model. Recently, unsupervised methods [6, 18, 20, 47] is proposed to leverage inter-domain style conversion. SPGAN [18] and PTGAN [20] utilize an image-to-image conversion network to preprocess the source dataset and then conduct the supervised learning. An iterative pseudo-label framework is proposed in [47], which significantly boosts the accuracy. However, this framework is very sensitive to initialization. Tzeng et al. [16] utilizes the style transfer GAN to acquire initialization and then generate pseudo-labels by unsupervised clustering for all target images. Nevertheless, the algorithms based on clustering works poorly on similar images. The pseudo-labels they assigned to similar images from different categories can be the same, which indicates their weakness in distinguishing confusing samples [45]. ECN [11] constructs a continuously updated feature memory to estimate the similarity between the entire target images. Compared with ECN, the proposed method does not treat every neighbor of the training sample equally because there is often contamination of mismatches in one-way ranking. In contrast, the proposed method adopts RNLSL to effectively distinguish true matches from hard negative samples, which greatly improves the reliability of similarity estimation. Meanwhile, CCRL is adopted to brings a new idea for solving camera style variations.

2.5 K-Reciprocal encoding

K-Reciprocal Encoding [3], which applies the novelty of set intersection to re-rank sample similarity for the first time, bases on the hypothesis that for an optimal match, they should each other's top- k nearest neighbors. Specifically, Mahalanobis distance is firstly employed to

obtain the query’s primary k -nearest gallery lists, and then Jaccard distance is used to get the k -reciprocal nearest list. The method sequentially calculates the k -reciprocal nearest list for each sample in the query’s k -reciprocal nearest list. Then some positive samples, which are ignored because of illumination and perspective variances, can be recalled according to some restricted conditions. Jaccard distance represents the distance between the query and the recalled image. The author encodes the k -reciprocal nearest information into an equal but more simple vector to reduce the complexity with a higher weight for nearer samples.

2.6 Regular loss functions

In the field of re-id, the commonly used loss function mainly follows the one in classification and retrieval tasks.

cross-entropy loss [48] is mainly adopted in classification. It firstly converts the model output into predicted probability by softmax function and then calculate the \log value to obtain each sample’s corresponding loss, which can be expressed as:

$$L_{CE} = -\frac{1}{B} \sum_{i=1}^B \log \frac{e^{s_i}}{\sum_{j=1}^n e^{s_j}} \tag{1}$$

where B denotes the sample amount in the mini-batch, n denotes the category number.

Center loss [49] is an improvement on cross-entropy loss. It assumes a centroid for each category and realizes the convergence by narrowing the gap of the input sample and its corresponding category centroid.

$$L_{CL} = -\frac{1}{B} \sum_{i=1}^B \log \frac{e^{s_i}}{\sum_{j=1}^n e^{s_j}} + \frac{\lambda}{2} \sum_{i=1}^B \|x_i - c_i\|^2 \tag{2}$$

where x_i denotes the embedding of sample i , c_i represents its category.

Triplet Loss, which is put forward in FaceNet [50], bases on the most crucial idea that the samples in the same category should be nearer in the embedding space. Considering such one simple regulation will result in that the cluster centroids of different categories get closer, too, a margin constant m is added:

$$L_{triplet} = \|x_i^a - x_i^p\|^2 - \|x_i^a - x_i^n\|^2 + m \tag{3}$$

where x_i^a , x_i^p , and x_i^n denote the anchor sample, positive sample, and negative sample separately.

3 Method

Preparatory Work To facilitate the following analysis, this paper first defines the mathematical symbols to be used. The domain adaptive person re-id task provides two datasets, a

source domain $\{P_s, L_s\}$ with person identities and a target domain $\{P_t, C_t\}$ with camera identities. As many UDA person re-id methods, we adopt the assumption that each target image’s camera-id is obtained in advance, which is easy to be known when gathering target images from frame sequences in a video. The source domain contains M pedestrian images and these pedestrians belong to X categories. Each $p_{s,i}$ in the source domain belongs to a label $l_{s,i}$ representing its identification. In addition, there are N pedestrian images and Y cameras in the target domain. Each target image $p_{t,i}$ corresponds to a camera identity annotation $c_{t,i}$.

3.1 Overview of network

The overall architecture of the proposed method is shown in Fig. 2. In the first step, MBCR regard each single camera as a unique domain and utilize Cycle GAN [46] to train a camera-style transfer model. The specific implementation can be found in [51]. The model trained by the first step is then employed to augment the target images and generate the corresponding fake version for each original image. These fake images are added to the original dataset to participate in model training together. All images are fed to the pre-trained CNN backbone network, followed by an embedding module that consists of 512-dimensional feed forward network (FFN), one-dimensional batch normalization, and ReLU. The 512-dimensional features of each image is extracted through the embedding module. For source images, the extracted features are sent to an X -dimensional FFN (named FC *P-id), followed by softmax. The cross-entropy function is employed for supervised training. Simultaneously, MBCR maintain a feature memory module to save the latest output of the embedding module for each target image. To more accurately estimate the similarity between the target samples in the mini-batch and the ones in the memory module, this paper proposes a reciprocal neighbors label smoothing loss (RNLSL) based on memory-based circular ranking. Since RNLSL and memory-based circular ranking mechanisms are closely related, they are introduced in Section 3.2.

In addition to RNLSL, we also add a Y -dimensional FC layer (FC *C-id) after the embedding module and formulate a cross-camera invariance loss function (CCIL) to guide the model to discern the discrepancy of pedestrians in the unlabeled domain. The factors of occlusion, illumination, pose and background clustering all will result in such variations among different domains.

3.2 Reciprocal neighbors label smoothing loss

In the view of supervised learning, we hope that the identifications of the same category are close enough in the

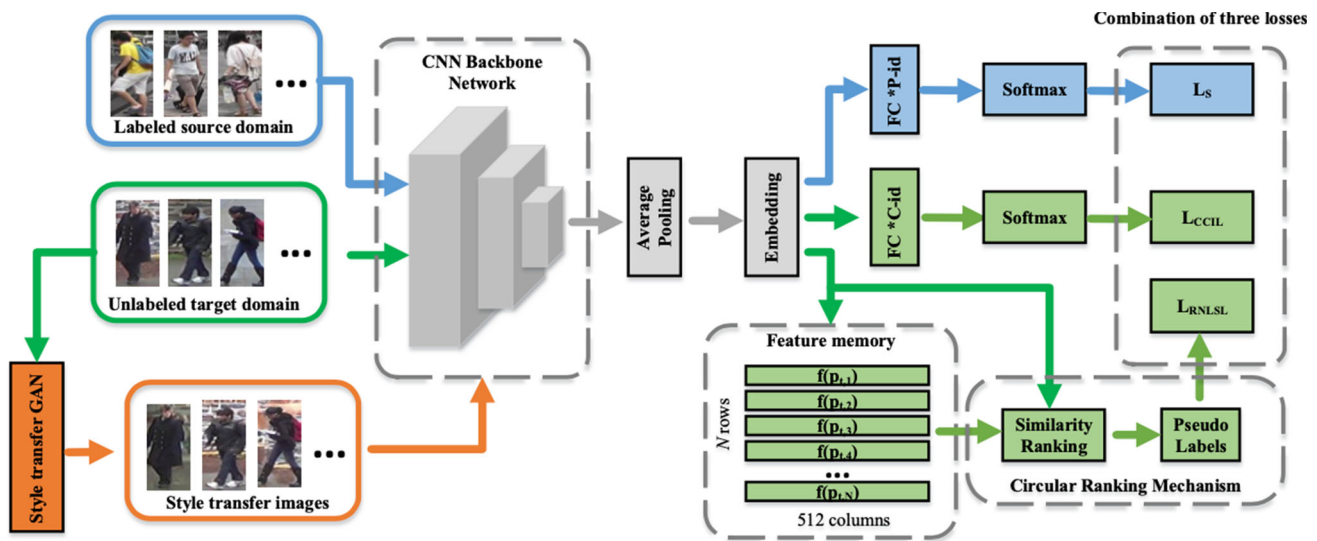


Fig. 2 The overall architecture of MBCR. In the training process, style transfer images, original target images, and source images are sent to the deep re-id network to acquire updated features. For the source

embedding space, while maintaining a certain distance from other identities. With the help of feature memory module T , reciprocal neighbors label smoothing loss can effectively mine the potential identity information in the target domain. In the beginning, each t in the target domain denotes a separate category, and assign an index to it. Each row of the feature memory module is used to store the 512 dimensional features corresponding to the index. During the iterative training process, the feature $T[i]$ corresponding to $p_{t,i}$ with the L2 normalized feature $f(p_{t,i})$ is updated by,

$$T_t[i] = (1 - \lambda)T_{t-1}[i] + \lambda f(p_{t,i}) \quad (4)$$

where t denotes the epoch numbers and λ controls the updating rate. The original k -nearest neighbors of $p_{t,i}$ can be obtained by the pairwise cosine distance function between $f(p_{t,i})$ and the feature memory module. We define the indexes of these neighbors as $S(p_{t,i})$. It is fallacious to directly pull $p_{t,i}$ and its neighbors because there are often mismatches in one-way ranking. Thus, a memory-based circular ranking mechanism is proposed to excavate the confusing samples in a batch, as shown in Fig. 3. We process the feature memory module with k -reciprocal encoding [3] and obtain an aggregate similarity matrix D_{recode} which contains the k -reciprocal encoding distance among all the embeddings saved in the memory module. It is worth noting that even though cosine function is widely employed in similarity metric, k -reciprocal encoding distance has recently exhibited better generalization.

For each image x_t in the original k -nearest ranking list $S(p_{t,i})$, its k -nearest ranking list $S(x_t)$ is obtained by sorting the similarity matrix D_{recode} . If $p_{t,i}$ also exists in $S(x_t)$, x_t denotes a positive sample of $p_{t,i}$, otherwise x_t is regarded

domain, the cross-entropy loss function is utilized to optimize it. For the target domain, the loss functions turn to RNLSL and CCIL

as a hard negative sample with a great probability. By traversing $S(p_{t,i})$, the original ranking list is divided into $S_{pos}(p_{t,i})$ and $S_{neg}(p_{t,i})$. Finally, we treat N unlabeled pedestrians as N categories and assign the pseudo-label $W_{t,i} = \{w_{i,1}, w_{i,2}, w_{i,3}, \dots, w_{i,N}\}$ to $p_{t,i}$ as,

$$w_{i,j} = \begin{cases} 1 & p_{i,j} \in S_{pos}(p_{t,i}) \\ -\frac{1}{k} & p_{i,j} \in S_{neg}(p_{t,i}) \\ 0 & otherwise \end{cases}, \forall p_{t,j} \in \{P_t\} \quad (5)$$

where, k denotes $S(p_{t,i})$'s size. The estimated probability that $p_{t,i}$ belongs to i -th class is obtained through,

$$p(i|p_{t,i}) = \frac{\exp(\alpha f(p_{t,i})) \cdot T[i]}{\sum_{j=1}^N \exp(\alpha f(p_{t,i})) \cdot T[j]} \quad (6)$$

where α is the scaling number. n_t denoted the amount of unlabeled images in a training batch. Finally, the reciprocal neighbors' label smoothing loss is formulated as,

$$\mathcal{L}_{RNLSL} = -\frac{1}{n_t} \sum_i \sum_{j=1}^N w_{i,j} \log p(j|p_{t,i}) \quad (7)$$

During the progress of RNLSL calculating, an important step is to assign pseudo-labels to target samples with memory-based circular ranking mechanism, and there are two principles of invariance. In (5), relatively large weights are assigned to the sample itself and its reciprocal neighbor samples. They are called sample invariance and neighborhood invariance in this paper.

Sample invariance allows the characteristics of the sample itself to be pulled in, which is a conservative approach in the absence of labels. But this keeps the different sample instances far from each other. However,

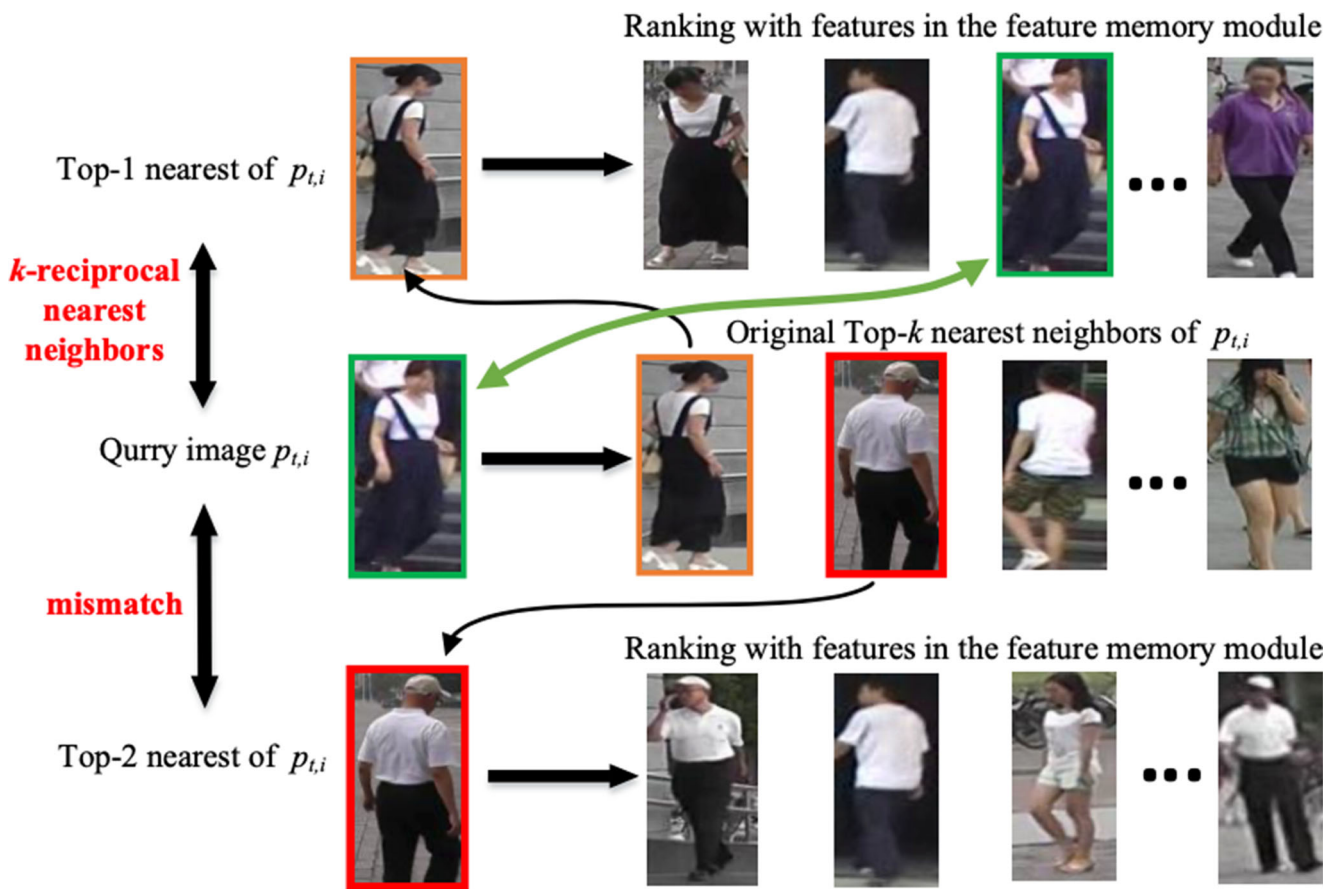


Fig. 3 Illustration of the memory-based circular ranking mechanism. First, we calculate the k -nearest neighbors list of $p_{t,i}$. For each top- k neighbor $p_{t,j}$ in this list, if $p_{t,i}$ also exists in the top- k nearest

neighbors of $p_{t,j}$, $p_{t,j}$ are called the k -reciprocal nearest neighbors of $p_{t,i}$. Otherwise, $p_{t,j}$ represents a hard negative sample of $p_{t,i}$

it widens the gap between instances of different categories. However, in this paper, each target domain image is regarded as a category, which will lead to the features of images with the same identity being pulled far away, which decreases the performance.

Neighborhood invariance can guide each pedestrian image instance and its candidate nearest neighbor sample to converge with each other. This helps to reduce the distance of similar pedestrian images in the embedding space. However, the pseudo-labels generated by circular sorting are not accurate, and neighborhood invariance is likely to shorten the embedding of two pedestrians with different identities. Even after the circular sorting and filtering, it is still not guaranteed that the query sample owns the same label as the candidate samples in the screened positive set.

Considering the limitations of these two Invariances, the cross-camera Invariance Loss is proposed as following.

3.3 Cross-camera invariance loss

Camera style variations might significantly change the appearance of person, which makes it difficult for re-id

model to find persons of the same identity in different cameras. Although we employ the camera style transfer model to reduce the difference among the camera styles, the inferred results from the network are still sensitive to image transformations. To reduce such correlation between features and camera styles, we propose a cross-camera invariance loss as,

$$\mathcal{L}_{CCIL} = \frac{1}{n_t} \sum_i \log p(c_{t,i} | p_{t,i}) \quad (8)$$

where, n_t denotes the batch size and $p(c_{t,i} | p_{t,i})$ represents the scores that the target image $p_{t,i}$ is filmed by its true camera id $c_{t,i}$, which can be obtained by the classification network. For each style transfer image, its camera identity is annotated according to its transferred style domain. Cross-camera invariance loss is a reverse form of the original cross-entropy loss. The traditional classification task strengthens the divergences between different categories by reducing the cross-entropy loss for the sake of obtaining more discriminative features. When we optimize the traditional cross-entropy loss in the opposite

direction, videlicet, the value of loss is increased. In this way, the camera style domains that were originally relatively independent are disrupted and the inter-domain distance would decrease. With the guidance of CCIL, the extracted features would be more robust to various camera styles.

3.4 Final loss for network

During training, we collaboratively optimize the source and target domains. The traditional cross-entropy loss function is leveraged to optimize the source domain as,

$$\mathcal{L}_s = -\frac{1}{n_s} \sum_i \log p(l_{s,i}|p_{s,i}) \quad (9)$$

where n_s denotes the images numbers in a mini-batch and $l_{s,i}$ is $p_{s,i}$'s person identity. Finally, the source domain loss and the target domain losses are added into the following formula,

$$\mathcal{L} = \beta \mathcal{L}_s + \mathcal{L}_{RNLSL} - \log(-\mathcal{L}_{CCIL}) \quad (10)$$

where β controls the proportion of \mathcal{L}_s . It's worth noting that the value of \mathcal{L}_{CCIL} is always negative. In experiments, we found that \mathcal{L}_{CCIL} rapidly decreased in the first epoch of training, which caused the overall loss to be a negative value and severely disrupted the optimization process. Therefore a logarithmic function is added to limit the weight of \mathcal{L}_{CCIL} in the overall loss, making the entire optimization process smoother.

4 Experiments and analysis

We perform experiments on three popular academic benchmarks, Market-1501 [22], DukeMTMC-reID [23] and MSMT17 [20]. These three datasets include abundant variations in viewpoint, occlusion, illumination, pose, and background, which exactly conform to the investigated issues,

4.1 Datasets

Market-1501 training set contains about 13k images with about 1.5k pedestrians, while the test set contains approximate 20k images with 1.5k identities and the query set contains about 3k images. Six cameras are used to capture this dataset.

DukeMTMC-reID is obtained from a multi-camera tracking dataset DukeMTMC by sampling manually bounding boxes, which results in different sizes of images in the dataset. There are 8 cameras and about 36k labeled pictures with about 1.4k pedestrians in DukeMTMC-reID. The

training set and the test set each contain half the person identities.

MSMT17 is released recently. It uses Faster RCNN [52] as the pedestrian detector and screens out 126,441 bounding boxes of 4,101 pedestrians from video sequences with different weather conditions. MSMT17 is randomly divided according to the training-test ratio of 1:3, rather than equally divided like other datasets. The purpose is to encourage efficient training strategies. Finally, the training set contains 10,421 pedestrians with 32,621 bounding boxes, while the test contains 3,060 identities with 93,820 bounding boxes.

4.2 Metrics

As to the evaluation metric, we choose the following ones:

- **Rank-n.** It denotes the probability that positive samples are shown in the recalled top-n results.
- **Mean Average Precision (mAP).** It is calculated from the proportion under the PR curve, P refers to precision and R refers to recall. In a robust re-id system, It is hoped that the matched returns of the query one can be recalled as much as possible, and the relatively more convinced images should be positive ones. mAP metric urges the model to balance the precision and recall, which is a significant means when measuring the performance.

4.3 Implementation details

ResNet-50 [53] is employed as the backbone to extract the base features. MBCR only reserve the layers before the last average pooling layer, and add an embedding module. During training, the random transformations used for data augmentation is the same as he2016deep. Finally, the input images are resized to 256x128. Each batch contains 128 source domain samples and 128 target domain samples. The target domain samples are randomly selected from the original images and the camera-style transfer images. The scaling number α is set to 20. The updating rate of feature memory is set as $\lambda = 0.1$. The weight β of L_s is set to 4. For the reciprocal neighbors label smoothing loss (RNLSL), the size of neighbor candidates k is set to 14. SGD [54] is employed as the optimizer and the learning rate is set to 0.01 for backbone and 0.1 for the others. During inference, we utilize the L2 normalized output of the average pooling layer as the features. The Euclidean distance is adopted to measure the similarity between the query image and the gallery images.

Baseline The proposed method is based on the mechanism of ECN [11], so ECN is selected as the baseline for experimental analysis.

4.4 Comparisons with the previous SOTA method

We compare MBCR with hand-crafted ones (including LOMO [31], BoW [22]), and the other excellent unsupervised learning methods. Table 1 reports the experimental comparisons on Market-1501 and DukeMTMC-reID. We choose one of them as the source domain and the other as the target domain. Concluding from the results, MBCR shows superiority on these two large-scale datasets conspicuously. As shown in the results, LOMO and BoW perform poorly on both datasets. Even the mAP of these two methods is less than 10% when tested on DukeMTMC-reID. The reason is that these two hand-crafted methods neither utilize the supervised information in the source domain nor mine the potential invariance in the target domain. CAMEL [14] significantly improve the rank-1 accuracy through unsupervised clustering methods. However, the label pollution generated by clustering limits their performance. Compared with previous excellent methods in view of domain adaptation (including PTGAN [20], SPGAN [18], CamStyle [51], HHL [55], OSNet-AIN [56], CCSE [57], PGS [6], PREST [58], and MMCL [4]), MBCR outperforms these methods significantly on these two datasets. Specifically, MBCR attains rank-1 accuracy=81.3% and mAP=53% when DukeMTMC-reID and Market-1501 are used as the source dataset and test dataset, respectively. Simultaneously, MBCR reaches rank-1 accuracy=69.2% and mAP=48.5% when using Market-1501 as source dataset and tested on DukeMTMC-reID. Compared to the baseline method ECN,

MBCR achieves rank-1 accuracy gain of 6.2% and 5.9% when tested on Market-1501 and DukeMTMC-reID respectively.

Finally, to verify the generalization performance, we also adopt a novel dedicated backbone network OSnet [39] for supplementary experiments. OSnet can dynamically capture multi-scale features and aggregate them with flexible weights. To effectively obtain the correlation among spatial channels and alleviate overfitting, OSnet employs both point convolution and depth convolution, which enables the model to achieve better performance with fewer parameters. It can be seen that with OSnet, the rank-1 accuracy and mAP of MBCR are increased by 7.3% and 14.3% when using DukeMTMC-reID as source dataset and tested on Market 1501, while ECN has only improved 3.6% and 7.9% on the two evaluation metrics. In MBCR, the performance gains from the high-performance backbone network are even more significant, as is the case when verified on DukeMTMC reID. The proposed method can be better integrated with the backbone network and give full play to the backbone network performance. Meanwhile, comparing with the previous SOTA PREST and MMCL, MBCR still shows mighty competitiveness.

We also demonstrated the scalability of the proposed method on a larger dataset MSMT17. Compared with previous two datasets, MSMT17 consists of more pedestrian images, bounding boxes, and cameras. Not only that, MSMT17 also contains more complicated scenes and backgrounds, with a longer period of time and intricate

Table 1 Performance (%) comparison with previous SOTA method on Market-1501 and DukeMTMC-reID

Method	Duke → Market				Market → Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [31]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
BoW [22]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
CAMEL [14]	54.5	-	-	26.3	-	-	-	-
PTGAN [20]	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN [18]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
CamStyle [51]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
OSNet-AIN [56]	61.0	77.0	82.5	30.6	52.4	66.1	71.2	30.5
HHL [55]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
CCSE [57]	73.7	84.0	87.9	38.0	56.1	66.7	71.5	30.6
UDAP [47]	80.9	-	-	60.0	75.0	-	-	57.1
PGS [6]	82.2	-	-	58.6	75.7	-	-	58.0
PREST [58]	82.5	92.1	94.9	62.4	74.4	83.7	85.9	56.1
MMCL [4]	84.4	92.8	95.0	60.4	72.4	82.9	85.0	51.4
ECN [11] (baseline)	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
MBCR (Ours)	81.3	81.3	91.0	53.0	69.2	79.9	83.3	48.5
ECN+OSnet [39]	78.7	89.9	93.3	50.9	65.6	77.4	81.5	44.2
MBCR (Ours) + OSnet	88.6	94.2	96.2	67.3	74.3	82.9	86.0	53.4

lighting variations. Since MSMT17 is released recently, few unsupervised methods have published experimental results on it. Two unsupervised methods (PTGAN and ECN) are selected for comparative experiments. As shown in Table 2, the proposed approach significantly precedes PTGAN and ECN, whether using Market-1501 or DukeMTMC-reID as the source domain. Specifically, MBCR attains rank-1 accuracy=35.2% and mAP=12.2% when using DukeMTMC-reID as source dataset. Compared to the baseline method ECN, MBCR boosts the performance of rank-1 and mAP by 5% and 2% separately.

All in all, MBCR can utilize memory-based circular ranking mechanism to produce smooth labels for the unlabeled dataset. With the guidance of RNLSL and CCIL, MBCR can mine the identity information hidden in the neighborhood and make the model less sensitive to various variations in the target domain.

4.5 Ablation study

To demonstrate that the performance improvement described in this paper is due to the proposed components, massive ablation experiments are performed and reported in Table 3.

First, RNLSL is added to the baseline network to demonstrate its effectiveness. As shown in Table 3, RNLSL improves the rank-1 accuracy from 75.1% to 78.7% and 63.3% to 66.7% when regarding Market-1501 and DukeMTMC-reID as the unlabeled domain. This proves that RNLSL not only can effectively narrow the gap between real matching images but also can mine hard negative samples in the neighborhood.

Next we validate the performance improvements of CCIL. In Table 3, CCIL achieves rank-1 accuracy increments of 3.2% on Market-1501 and 3.1% on DukeMTMC-reID. This demonstrates that CCIL can instruct the model to extract camera-independent features, effectively alleviating the influence of the camera style diversity. Moreover, the combination of RNLSL and CCIL further improves the performance. This indicates that the two-loss functions can coordinate with each other to guide the optimization of the model from two different aspects. Specifically, CCIL enables the network more robust to the image variations

in the unlabeled domain, narrowing the gap among images of the same identity with different camera styles. This allows us to use memory-based circular ranking mechanism to accurately distinguish the positive samples from the negative samples in the neighborhood and generate more accurate smooth labels. With the guidance of smooth labels, RNLSL prompts the model to extract more discriminative features.

To further prove that the proposed RNLSL, which is some kind of improvement based on k -Reciprocal Encoding, performs better than k -Reciprocal Encoding. The ablation study is conducted on then as Table 3. Pure k -Reciprocal Encoding indeed brings increments of 1.9% and 2.8% towards R-1 and mAP, but they are still less than the ones from RNLSL. And the combination of ECN, k -Reciprocal Encoding, and CCIL performs more poorly than that of ECN, k -Reciprocal Encoding, and RNLSL.

4.6 Further analysis

To further understand the effectiveness of the proposed memory-based circular ranking mechanism, we demonstrate how the model uses the memory-based circular ranking mechanism to screen the candidates in the sample neighborhood when training referring to Fig. 4. Simultaneously, we carry out experiments on two important hyperparameters of MBCR, including the weight of the source domain loss β and the number of neighbor candidates k .

4.6.1 Analysis of memory-based circular ranking

As shown in Fig. 4, true positive samples (marked with a green border) and true negative samples (marked with a red border) are doped with each other in the initial neighborhood. ECN mines neighborhood invariance by directly reducing the gap between the sample and all neighbors. MBCR uses feature memory to perform circular ranking, which can further examine the similarity between the two images and effectively reduce the noise in the original neighborhood. In Fig. 4, four example results are shown. As a result, most candidates in the sample neighborhood are effectively distinguished, and only a few are classified into the wrong set.

Table 2 Performance evaluation when tested on MSMT17

Method	Market → MSMT17				Duke → MSMT17			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PTGAN [20]	10.2	-	24.4	2.9	11.8	-	27.4	3.3
ECN [11]	25.3	36.3	42.1	8.5	30.2	41.5	46.8	10.2
MBCR (Ours)	30.2	41.9	47.6	10.7	35.2	46.4	51.5	12.2

Table 3 Ablation study(%) on Market-1501 and DukeMTMC-reID

Method	Market → MSMT17				Duke → MSMT17			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
ECN (baseline) [11]	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
ECN+k-Reciprocal Encoding [3]	77.0	88.6	92.0	45.8	64.6	76.4	80.8	43.1
ECN+RNLSL	78.7	89.5	92.4	47.2	66.7	77.4	81.2	44.8
ECN+CCIL	78.3	88.4	91.6	46.9	66.4	77.5	81.5	44.2
ECN+k-Reciprocal Encoding+CCIL	79.8	89.9	92.1	50.1	67.5	78.4	82.1	46.7
ECN+RNLSL+CCIL	81.3	91.0	93.3	53.0	69.2	79.9	83.3	48.5



Fig. 4 Example results of four images on the Market-1501 dataset. For each probe, its initial k-nearest neighbors are listed. The two rows after the initial list correspond to the positive and negative sample sets divided by memory-based circular ranking, respectively

Fig. 5 Evaluation with different values of β

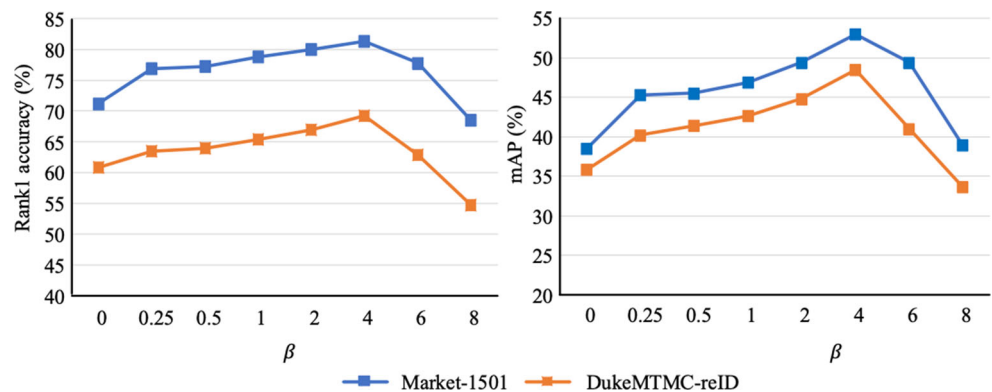
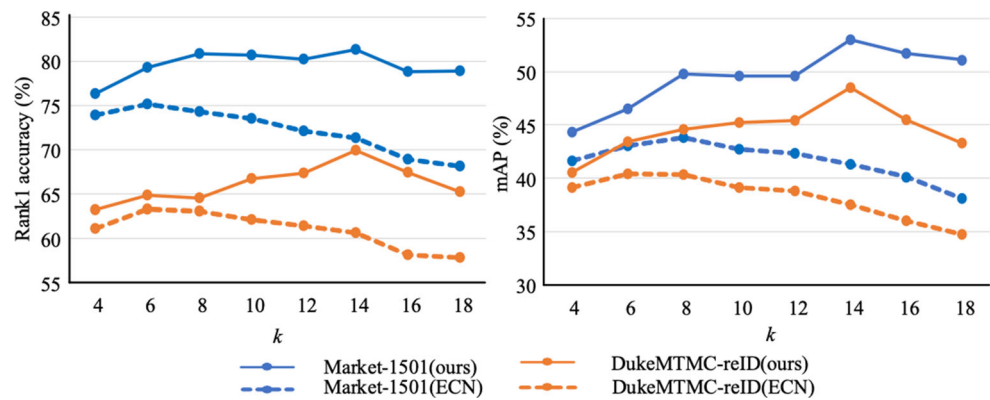


Fig. 6 Analysis of the number of candidate neighbors together with ECN



4.6.2 The weight of loss: β .

The analysis on the weight of the source domain loss β is reported in Fig. 5. When β is fixed to 0, the network is optimized only by unlabeled images of the target domain. As the value of β increases, the source domain and the target domain jointly guide the training process and improve the experimental results. Combined with Fig. 5, MBCR leverages the labels in the source domain as useful guidance in model training. When setting $\beta=4$, the model achieves the best results. As β continues to increase, the performance of MBCR begins to degrade. When setting $\beta=8$, the proportion of target domain losses is too small, so the re-id model focuses on the distribution of the source domain and ignores the invariance in the target domain.

4.6.3 The number of candidate neighbors: k .

In Fig. 6, we show the experimental results of comparison with ECN when k takes different values. As the value of k increases, the experimental results continue to improve, eventually reaching an optimal value when $k=14$. Compared with ECN, MBCR extends the size of the reliable neighborhood from 6 to 14. The reason is that MBCR can utilize a memory-based circular ranking mechanism to more accurately distinguish between images that look similar but have different identities. When k is assigned a large value, the rank-1 accuracy of MBCR does not decay as quickly as ECN. We want to point out that MBCR outperforms the best results of ECN at all k values.

Table 4 Computational cost analysis of the exemplar memory

Method	Duke \rightarrow Market			
	R-1	mAP	Time (min)	Memory (MB)
Mini-batch	73.9	44	59.3	5000
Memory	81.3	53.0	60.9	5780

4.6.4 Computational cost analysis

In this paper, feature memory and mini-batch are utilized to train and optimize the model with the proposed loss function respectively. As for the method based on mini-batch, the input samples are comprised of the target sample, the corresponding camera style transfer sample, and the corresponding k -nearest neighbor candidate sample. Referring to Table 4, memory-based approach is significantly superior to the mini-batch based one. It is worth noting that memory-based method will introduce limited additional training time cost (+1.6 minutes) and GPU memory (+780 MB), though they are negligible compared to the total cost.

Feature memory modules are frequently employed in recent self-supervised training models. Most of the existing unsupervised models are based on contrast learning. Traditional neural network training is carried out in the form of mini-batch. It is unscientific to compare positive and negative samples within a small batch due to the limitation of samples number. There may be no pedestrians with the same identity as the target sample in the mini-batch, so it is meaningless to conduct circular ordering in it. By adding a feature memory module, the model can sort globally and the performance can be improved significantly.

5 Conclusion

In this work, we present a novel unsupervised domain adaptation method for person re-identification using memory-based circular ranking mechanism to adaptively assign pseudo-labels for the target domain. It's worth noting that MBCR does not use clustering to generate pseudo-labels for the entire target images like the existing unsupervised re-id methods because the clustering algorithm involves heavy CPU calculations and long training time. The memory-based circular ranking mechanism can iteratively generate smooth labels for samples in each mini-batch, which significantly decreases the time cost and avoids the noise

caused by clustering. Different from previous unsupervised approaches, MBCR also has better scalability in the large-scale pedestrian benchmark of the real world. The labeled source images are employed to supervise the training process while realizing unsupervised optimization of the target domain by training jointly both the reciprocal neighbors' label smoothing loss (RNLSL) and the cross-camera invariance loss (CCIL). RNLSL aims to screen out positive samples in the neighborhood and mine hard negative samples with a similar appearance. Simultaneously, CCIL is designed to ensure that the deep re-id network is robust to various variations of the camera styles. Abundant experiments demonstrate the superiority of the elaborate components proposed in this paper.

Acknowledgements The Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDC05000000), this work was supported by the National Natural Science Foundation of China (Grant No. 61732018, 61872335, and 61802367), and the Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing (2019A07)

References

- Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
- Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: Past, present and future. arXiv:1610.02984
- Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1318–1327
- Wang D, Zhang S (2020) Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10981–10990
- Su J, He X, Qing L, Cheng Y, Peng Y (2021) An enhanced siamese angular softmax network with dual joint-attention for person re-identification. Appl Intell:1–19
- Chong Y, Peng C, Zhang C, Wang Y, Feng W, Pan S (2021) Learning domain invariant and specific representation for cross-domain person re-identification. Appl Intell:1–14
- Zhang T, Sun X, Li X, Yi Z (2021) Image generation and constrained two-stage feature fusion for person re-identification. Appl Intell:1–11
- Wang H, Peng J, Chen D, Jiang G, Zhao T, Fu X (2020) Attribute-guided feature learning network for vehicle reidentification. IEEE MultiMedia 27(4):112–121
- Wang H, Peng J, Zhao Y, Fu X (2020) Multi-path deep cnns for fine-grained car recognition. IEEE Trans Veh Technol 69(10):10484–10493
- Wang H, Wang Y, Zhang Z, Fu X, Zhuo L, Xu M, Wang M (2020) Kernelized multiview subspace analysis by self-weighted learning. IEEE Trans Multimed 23:3828–3840
- Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 598–607
- Fan H, Zheng L, Yan C, Yang Y (2018) Unsupervised person re-identification: Clustering and fine-tuning. ACM Trans Multimed Comput Commun Appl (TOMM) 14(4):1–18
- Liu Z, Wang D, Lu H (2017) Stepwise metric promotion for unsupervised video person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 2429–2438
- Yu H-X, Wu A, Zheng W-S (2017) Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 994–1002
- Yu H-X, Wu A, Zheng W-S (2018) Unsupervised person re-identification by deep asymmetric metric embedding. IEEE Trans Pattern Anal Mach Intell 42(4):956–973
- Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7167–7176
- Panareda Busto P, Gall J (2017) Open set domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 754–763
- Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003
- Lin S, Li H, Li C-T, Kot AC (2018) Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. arXiv:1807.01440
- Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
- Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA (2018) Generative adversarial networks: An overview. IEEE Signal Proc Mag 35(1):53–65
- Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
- Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer, pp 17–35
- Chen Y-C, Zhu X, Zheng W-S, Lai J-H (2017) Person re-identification by camera correlation aware feature augmentation. IEEE Trans Pattern Anal Mach Intell 40(2):392–408
- Koestinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 2288–2295
- Wang H, Gong S, Zhu X, Xiang T (2016) Human-in-the-loop person re-identification. In: European conference on computer vision. Springer, pp 405–422
- Wang T, Gong S, Zhu X, Wang S (2016) Person re-identification by discriminative selection in video ranking. IEEE Trans Pattern Anal Mach Intell 38(12):2501–2514
- Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: European conference on computer vision. Springer, pp 688–703
- Bazzani L, Cristani M, Murino V (2013) Symmetry-driven accumulation of local features for human characterization and re-identification. Comput Vis Image Underst 117(2):130–144
- Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision. Springer, pp 262–275

31. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2197–2206
32. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition. IEEE, pp 34–39
33. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. *J Mach Learn Res* 10(1):207–244
34. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification
35. Köstinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 2288–2295
36. Desai MR, Patel SA, Peerzade M, Chawhan G (2020) Person re-identification via deep metric learning. In: 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAIECC)
37. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916
38. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
39. Zhou K, Yang Y, Cavallaro A, Xiang T (2019) Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3702–3712
40. Si J, Zhang H, Li C-G, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5363–5372
41. Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2119–2128
42. Guo Y, Cheung N-M (2018) Efficient and deep person re-identification using multi-level similarity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2335–2344
43. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A (2006) A kernel method for the two-sample-problem. *Adv Neural Inf Process Syst* 19:513–520
44. Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. In: International conference on machine learning. PMLR, pp 97–105
45. Yan H, Ding Y, Li P, Wang Q, Xu Y, Zuo W (2017) Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2272–2281
46. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
47. Song L, Wang C, Zhang L, Du B, Zhang Q, Huang C, Wang X (2020) Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recogn* 102:107173
48. De Boer P-T, Kroese DP, Mannor S, Rubinstein RY (2005) A tutorial on the cross-entropy method. *Ann Oper Res* 134(1):19–67
49. Ranjan R, Castillo CD, Chellappa R (2017) L2-constrained softmax loss for discriminative face verification. arXiv:1703.09507
50. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
51. Zhong Z, Zheng L, Zheng Z, Li S, Yang Y (2018) Camstyle: A novel data augmentation method for person re-identification. *IEEE Trans Image Process* 28(3):1176–1190
52. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* 28:91–99
53. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
54. Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. In: International conference on machine learning. PMLR, pp 1139–1147
55. Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 172–188
56. Zhou K, Yang Y, Cavallaro A, Xiang T (2021) Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
57. Lin Y, Wu Y, Yan C, Xu M, Yang Y (2020) Unsupervised person re-identification via cross-camera similarity exploration. *IEEE Trans Image Process* 29:5481–5490
58. Zhang H, Cao H, Yang X, Deng C, Tao D (2021) Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification. *IEEE Trans Image Process*

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



Honggang Chen born in 1997, Alibaba advertising algorithm engineer. He received his master's degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2021. His research interest covers image and video retrieval, person re-id, and domain adaptation.



Xing Cheng born in 1998, a postgraduate student at the Institute of Computing Technology, Chinese Academy of Sciences. His research interest covers video comprehension, multi-modal retrieval, and other essential visual tasks.



Nan Guo born in 1990, associate researcher at the Institute of Computing Technology, Chinese Academy of Sciences. In 2017, she received her Ph.D. degree from the Beijing University of Posts and Telecommunications. Her main research interests include object detection, multi-modal retrieval, and efficient AI system design.



Xiao Xiao born in 1994, MA.Eng, Engineer in Institute of Computing Technology, Chinese Academy of Sciences. Mainly interested in Computer Vision and Cross modal learning.



Xiaochun Ye born in 1981, Ph.D., professor in Institute of Computing Technology, Chinese Academy of Sciences. His main research interests include many-core architecture and high-performance computing.



Dongrui Fan born in 1979, Ph.D, professor in Institute of Computing Technology, Chinese Academy of Sciences. His main research interests include high throughput many-core processor design and AI algorithm.