Generalizing Person Re-Identification by Camera-Aware Invariance Learning and Cross-Domain Mixup

Chuanchen Luo[1,2], Chunfeng Song[1,2], and Zhaoxiang Zhang[1,2,3]

[1] University of Chinese Academy of Sciences
[2] Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS
{luochuanchen2017,chunfeng.song,zhaoxiang.zhang}@ia.ac.cn

Abstract. Despite the impressive performance under the single-domain setup, current fully-supervised models for person re-identification (re-ID) degrade significantly when deployed to an unseen domain. According to the characteristics of cross-domain re-ID, such degradation is mainly attributed to the dramatic variation within the target domain and the severe shift between the source and target domain. To achieve a model that generalizes well to the target domain, it is desirable to take both issues into account. In terms of the former issue, one of the most successful solutions is to enforce consistency between nearest-neighbors in the embedding space. However, we find that the search of neighbors is highly biased due to the discrepancy across cameras. To this end, we improve the vanilla neighborhood invariance approach by imposing the constraint in a camera-aware manner. As for the latter issue, we propose a novel cross-domain mixup scheme. It alleviates the abrupt transfer by introducing the interpolation between the two domains as a transition state. Extensive experiments on three public benchmarks demonstrate the superiority of our method. Without any auxiliary data or models, it outperforms existing state-of-the-arts by a large margin. The code is available at https://github.com/LuckyDC/generalizing-reid.

Keywords: Domain Adaptation, Person Re-Identification, Camera-Aware Invariance Learning, Cross-Domain Mixup

1 Introduction

Person re-identification (re-ID) aims to associate images of the same person across non-overlapping camera views. As the fundamental component of intelligent surveillance systems, it has drawn wide attention both in the industry and academia. With the surge of deep learning techniques, recent years have witnessed great progress in fully-supervised person re-ID [34,38,52,13,33,20,51]. However, the success of this paradigm relies heavily on enormous annotated data in the target domain, which is usually prohibitive to acquire in practice. To bypass the scarcity of annotations, one can train the model with a relevant labeled

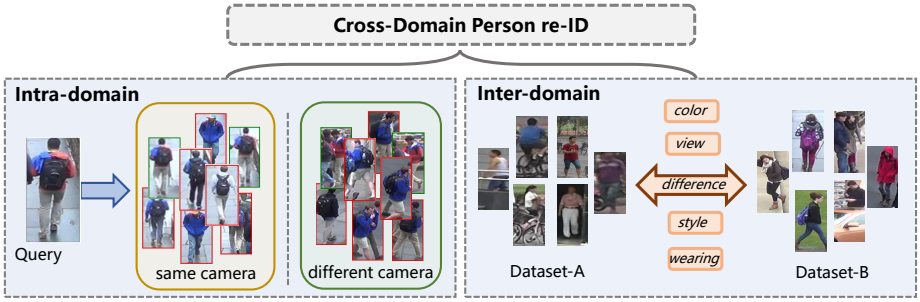


Fig. 1. The illustration cross-domain person re-ID. We consider intra-domain variation and inter-domain shift simultaneously. In terms of the former, cross-camera variations lead to a biased retrieval. As for the latter, the discrepancy between the source domain and the target domain hiders the effective adaptation.

dataset, a.k.a. the source domain. Unfortunately, due to the dramatic shift in data distribution, such a model would suffer a severe degradation in performance when directly deployed to the target domain. For this reason, it is desirable to investigate the problem of cross-domain person re-ID.

Given labeled source data and unlabeled target data, cross-domain re-ID dedicates to learn a model that generalizes well to the target domain. Compared with conventional unsupervised domain adaptation (UDA), it is characterized by the open-set setup and the domain hierarchy. The former implies the disjoint label space between the source domain and the target domain, which breaks the underlying assumption of most UDA methods. As for the latter, each domain can be further divided into multiple camera sub-domains, since the style of images is distinct across different cameras. According to such a hierarchy of domains, we impute the poor transfer performance to two factors, *intra-domain variation* and *inter-domain shift*. Wherein, the first factor is mainly derived from camera divergence. These issues are illustrated in Fig. 1. To achieve superior transfer performance, it is desirable to take both issues into account.

Recently, some studies [57,45,10,58] have verified the effectiveness of neighborhood invariance in coping with the intra-domain variation of the target domain. Equipped with a memory bank, these methods search neighbors of each probe throughout the whole dataset and impose a consistency constraint between them. However, due to the lack of supervision in the target domain, the model cannot suppress well the impact of inter-camera variation (including illumination, viewpoint, and background). In this case, the neighbor search is easily biased towards the candidates from the same camera as the probe. To be more specific, positive inter-camera matches are more likely to be arranged behind many negative intra-camera matches in the ranking list, which confuses the model learning. To address the issue, we improve the neighborhood invariance by imposing the constraint separately for intra-camera matching and inter-camera

matching. Despite the simplicity, this proposal leads to considerable improvement over its vanilla counterpart.

To alleviate the adverse effect of inter-domain shift, early works [40,9] employ extra generative models to transfer the image style across domains, which is essentially an advanced interpolation between the source and target manifold. By introducing stylized images as an intermediate domain, these methods expect to avoid the issues caused by the abrupt transfer between two very different domains. Along this insight, we explore to achieve the same goal by interpolating the samples from the two domains directly. Different from style transfer, the direct mixture in the pixel level leads to the change of content. Therefore, the identity label should also be mixed accordingly. This is exactly a mixup [49] process. However, it is nontrivial to employ vanilla mixup [49] in our case since it is initially customized for the *closed-set* classification problem. To make it applicable to *open-set* cross-domain re-ID, we augment mixup with a dynamic classifier. It can cover the label space of the input source-target pairs adaptively without the access to the exact label space of the target domain.

In summary, the contribution of this work is three-fold:

- To bypass the bias in the neighbor search, we impose the neighborhood invariance in a camera-aware way. Despite the simplicity, this approach leads to a significant improvement over its camera-agnostic counterpart.
- We propose a novel cross-domain mixup scheme to smooth the transition between the source domain and the target domain. It improves the transfer performance significantly with negligible overhead.
- Extensive experiments validate the effectiveness of our method. It achieves state-of-the-art performance on Market-1501, DukeMTMC-reID and MSMT17 datasets.

2 Related Work

Supervised person re-identification has made significant progress in recent years, thanks to the advent of deep neural networks [17,15,5] and large scale datasets [53,54,29,40]. The research in this field mainly focuses on the development of discriminative loss functions [48,16,22] or network architectures [61,34,28,37]. In term of the former direction, a series of deep metric learning methods [31,6,16,3,47] been proposed to enhance intra-class compactness and inter-class separability in the manifold. As for the customization of architecture, PCB [34] and its follow-ups [34,38,52,13] dominate the trend. Apart from the two directions mentioned above, some methods [33,20,51] attempt to involve auxiliary data for the fine-grained alignment of the human body. Despite their success in the single domain, these fully-supervised methods suffer from poor generalization ability, which prevents them from the practical application.

Cross-domain person re-identification pursues high performance in the target domain with the access to labeled source data and unlabeled target data. Early works [40,9,25,19] focus on reducing the domain gap between the two domains at the image level. They perform the image-to-image translation [62,7]

from the source domain to the target domain and then train the model with translated images. Besides, some methods [39,18] attempt to connect the two domains with common auxiliary tasks. Wang *et al*. [39] share knowledge across domains through attribute classification. Huang *et al*. [18] perform human parsing and pose estimation on both domains simultaneously to enhance alignment and model generalization. Recently, some studies [46,57,45,10] recognize the importance of mining discriminative cues in the target domain. Yu *et al*. [46] mine underlying pairwise relationships according to the discrepancy between feature similarity and class probability. They then use a contrastive loss to enforce the mined relationships. Zhong *et al*. [57] investigate the impact of intra-domain variations and impose three types of invariance constraints on target samples, *i.e*. exemplar-invariance, camera-invariance [59], and neighborhood-invariance. Yang *et al*. [45] further introduce the idea of neighborhood-invariance to the patch level. Current leading methods [11,42,12,32,50,14] adopt a pseudo label estimation scheme. They label target samples by a clustering algorithm and then train the model accordingly. Such an operation will be performed repeatedly until the model converges, which results in a heavy computational burden.

MixUp [49] is a data augmentation technique initially proposed for the supervised classification problem. Afterwards, it was extended to random hidden layers by Verma *et al*. [35]. MixUp enhances the smoothness of the learned manifold by applying convex combinations of labeled samples for training. It has demonstrated its effectiveness on several classification benchmarks. Recently, MixUp has been successfully adapted to the field of semi-supervised learning [36,1] and domain adaptation [26,30]. Without the access to the ground-truths of unlabeled/target data, these methods conduct MixUp based on the prediction of original samples. Unfortunately, all of them focus on the closed-set scenario and cannot be applied to cross-domain re-ID directly. In parallel with our work, Zhong *et al*. [60] extend MixUp scheme to the open-set scenario where the number of target classes is given. They explain the insight from the viewpoint of the label reliability and achieve very positive results on CIFAR [21] as well as ImageNet [8].

3 Method

In the context of cross-domain person re-ID, we have access to a labeled source domain $\mathcal{S} = \{X_s, Y_s\}$ and an unlabeled target domain $\mathcal{T} = \{X_t\}$. The source domain contains N_s images of P persons. Each sample $x_i^s \in X_s$ is associated with an identity label y_i^s. The target domain consists of N_t images $\{x_i^t\}_{i=1}^{N_t}$ whose identity annotations are absent. In addition, the camera indices of images (*i.e*. $C_s = \{c_i^s\}_{i=1}^{N_s}$ and $C_t = \{c_i^t\}_{i=1}^{N_t}$) are also available in both domains. Given such information, the goal is to learn a model that generalizes well to the target domain.

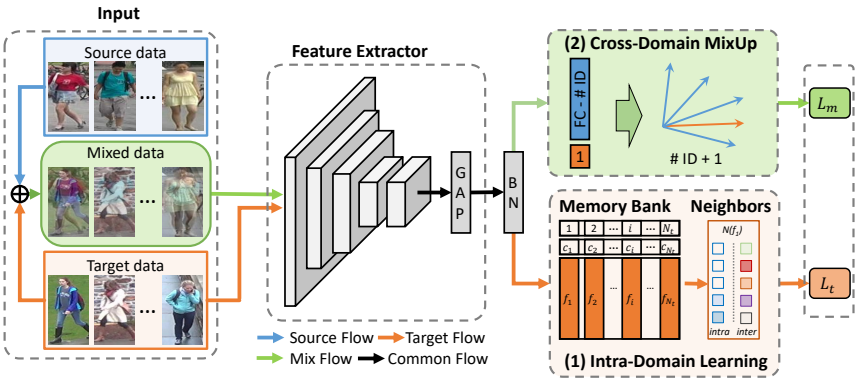


Fig. 2. The framework of our method. Firstly, mixed data is generated by the convex combination between source-target pairs. Then, it is fed into the network together with target data to acquire image embeddings. After the normalization by a BN layer, each type of embeddings is assigned to its corresponding component. (1) With the help of an augmented memory, the learning of target embeddings is supervised by intra-camera and inter-camera neighborhood consistency. (2) As for mixed embeddings, we maintain a dynamic classifier to cover the label space of each source-target pair adaptively.

3.1 Overview

As illustrated in Fig. 2, we feed-forward target samples and mixed samples into the network simultaneously. Wherein, each mixed sample is generated by interpolating between a source-target pair. For target data, we maintain a memory bank $\boldsymbol{M} \in \mathbb{R}^{N_t \times d}$, where each slot $\boldsymbol{m}_i \in \mathbb{R}^d$ stores the feature of the corresponding sample x_i^t. The memory is updated in a running-average manner during training:

$$\boldsymbol{m}_i \leftarrow \sigma \boldsymbol{m}_i + (1-\sigma)\, f\left(x_i^t\right), \quad \boldsymbol{m}_i \leftarrow \boldsymbol{m}_i / \|\boldsymbol{m}_i\|_2, \tag{1}$$

where σ denotes the momentum of the update, $f\left(x_i^t\right) \in \mathbb{R}^d$ represents the l_2-normalized feature of x_i^t extracted by the current model. In practice, the memory bank behaves as a non-parametric inner-product layer [44], by which we can obtain pairwise similarities between each input sample and all target instances on the fly. On the basis of such pairwise similarities, we can retrieve nearest-neighbors of each input image and impose a consistency constraint between them. As for mixed data, we compose a dynamic classifier to cover the label space of each source-target pair adaptively. It is built upon the source prototypes and the feature of the target instance. In the sequel, we will elaborate on the learning tasks customized for target data and mixed data.

3.2 Camera-Aware Neighborhood Invariance

Without the knowledge of the label space (*i.e.* identity annotations and the number of identities), it is infeasible to figure out the class assignment of target

Fig. 3. The visualization of ranking lists in intra-camera matching and inter-camera matching. We perform retrieval on DukeMTMC-reID using a model pre-trained on Market-1501. The green frame indicates positive matches, while the red frame indicates negative matches. The score on the top of each gallery image represents its cosine similarity with the probe.

samples directly. In this case, the pairwise relationship is a potential cue to guide the feature learning in the target domain. In representation learning, it is generally assumed that each sample shares the same underlying label with its nearest-neighbors at a high probability. Equipped with the memory bank mentioned above, we can obtain the probability that x_i^t share the same identity with x_j^t on the fly:

$$p_{ij} = \frac{\exp\left(s \cdot \boldsymbol{m}_j^{\mathrm{T}} f\left(x_i^t\right)\right)}{\sum_{k=1}^{N_t} \exp\left(s \cdot \boldsymbol{m}_k^{\mathrm{T}} f\left(x_i^t\right)\right)}, \tag{2}$$

where s is a scaling factor that modulates the sharpness of the probability distribution. According to the above assumption, ECN [57] proposes to maximize such probabilities between each probe image and its nearest-neighbors in the whole dataset:

$$\mathcal{L}_{ag} = -\sum_j w_{i,j} \log p_{ij}, \qquad w_{i,j} = \begin{cases} \frac{1}{|\Omega_i|}, & j \neq i \\ 1, & j = i \end{cases}, \qquad \forall j \in \Omega_i, \tag{3}$$

where Ω_i represents the nearest-neighbors of x_i^t throughout the whole target domain. $|\Omega(x_i^t)|$ denotes the size of the neighbor set. For convenience, we term this loss function as camera-agnostic neighborhood loss, since it treats all candidates equally regardless of their camera indices while searching neighbors.

Due to the scene variation across cameras, there is a significant discrepancy in similarity distribution between inter-camera matching and intra-camera matching [41]. The average pairwise similarity of inter-camera matching is smaller than that of intra-camera matching. As a result, intra-camera candidates can easily dominate the top ranking list, whether or not they are positive matches. In this case, it is problematic to employ Eq. (3), since it would push inter-camera positive matches away from the probe. For clarity, we visualize an example in Fig. 3. From the figure, we observe that even the first positive inter-camera match has a lower similarity score than many negative intra-camera matches. When sorting

all candidates in a camera-agnostic manner, positive inter-camera matches can be easily excluded from a pre-defined neighborhood range. An intuitive solution is to choose a larger neighborhood range. However, such a practice would involve more negative matches inevitably, which is detrimental to feature learning.

To bypass this dilemma, we propose to enforce neighborhood invariance separately for intra-camera matching and inter-camera matching. Suppose O_i^{intra} denotes the set of instances that share the same camera as x_i^t and O_i^{inter} represents the set of instances whose camera indexes are different from x_i^t. For sample x_i^t, intra-camera matching and inter-camera matching only have access to the instances in O_i^{intra} and O_i^{inter}, respectively. Therefore, the probability that x_i^t shares the same identity with an intra-camera candidate x_j^t is formulated as follows:

$$p_{i,j}^{intra} = \frac{\exp\left(s \cdot \boldsymbol{m}_j^{\mathrm{T}} f\left(x_i^t\right)\right)}{\sum_{k \in \mathrm{O}_i^{intra}} \exp\left(s \cdot \boldsymbol{m}_k^{\mathrm{T}} f\left(x_i^t\right)\right)} \tag{4}$$

The definition of the probability that x_i^t shares the same identity with an inter-camera candidate is similar:

$$p_{i,j}^{inter} = \frac{\exp\left(s \cdot \boldsymbol{m}_j^{\mathrm{T}} f\left(x_i^t\right)\right)}{\sum_{k \in \mathrm{O}_i^{inter}} \exp\left(s \cdot \boldsymbol{m}_k^{\mathrm{T}} f\left(x_i^t\right)\right)} \tag{5}$$

Accordingly, we replace the original camera-agnostic loss function Eq. (3) with the following two camera-aware loss functions:

$$\begin{aligned} \mathcal{L}_{intra} &= -\sum_j w_{i,j} \log p_{i,j}^{intra}, \quad \forall j \in \Omega_i^{intra}. \\ \mathcal{L}_{inter} &= -\sum_j w_{i,j} \log p_{i,j}^{inter}, \quad \forall j \in \Omega_i^{inter}. \end{aligned} \tag{6}$$

where Ω_i^{intra} and Ω_i^{inter} denote the neighbor sets of x_i^t throughout O_i^{intra} and O_i^{inter}, respectively. Different from ECN [57] that adopts fixed top-k nearest-neighbors, we define the neighborhood based on the relative similarity ratio to the top-1 nearest neighbors:

$$\Omega_i = \{j | \mathrm{sim}(x_i, x_j) > \epsilon \cdot \mathrm{sim}(x_i, \text{top-1 neighbor of } x_i)\} \tag{7}$$

Moreover, without the disturbance of cross-camera variations, the mined neighborhood for intra-camera matching is much more reliable than that for inter-camera matching. Thus, it is much easier to learn a discriminative intra-camera representation first, which can encourage accurate inter-camera matching. For this reason, we propose to employ $\mathcal{L}_{intra}$ before the involvement of $\mathcal{L}_{inter}$ in practice.

Remarks. Some related works [23,4,63,41,43] adopt a similar two-stage learning scheme. They focus on spreading the given local association (*i.e.* tracklet [23,4] or identity [63,41] within the same camera) to the global. They do not pay attention to the discrepancy in similarity distribution between the intra-camera matching and inter-camera matching.

3.3 Cross-Domain Mixup

In order to push the transfer performance ahead, it is desirable to handle the shift between the source domain and the target domain. Early efforts in this direction perform image-to-image translation from the source style to the target style. They introduce the stylized domain as an intermediate state to mitigate the performance loss of direct transfer. However, the style transfer process demands cumbersome generative models. Can we achieve the same goal in a more concise fashion?

Essentially, style transfe is an advanced interpolation between the source and target manifold. Considering that mixup [49] also conducts interpolation on the data manifold, we explore to employ it as the substitute for style transfer. According to the formulation of mixup [49], we mix samples and their labels simultaneously:

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha) \tag{8}$$

$$x_m = \lambda x_s + (1-\lambda) x_t \qquad y_m = \lambda y_s + (1-\lambda) y_t \tag{9}$$

where α is a hyper-parameter of Beta distribution. $(x_s, y_s) \in \mathcal{S}$ and $(x_t, y_t) \in \mathcal{T}$ denote samples from the source domain and the target domain, respectively. However, we have no access to the target annotation y_t in the context of cross-domain re-ID. Besides, the label space is disjoint between the two domains. Thus, it is infeasible to apply mixup operation directly. To address the issue, we propose to maintain a dynamic classifier that covers the label space of source-target pair adaptively. With the knowledge of the source label space, we can first define a classifier to identify P persons in the source domain. Then, we append a virtual prototype vector $\boldsymbol{w}_{virt} \in \mathbb{R}^d$ to the source classifier $\boldsymbol{W} \in \mathbb{R}^{P \times d}$:

$$\boldsymbol{W}' \leftarrow [\boldsymbol{W}, \boldsymbol{w}_{virt}] \qquad \boldsymbol{w}_{virt} = \frac{\|\boldsymbol{w}_{y_s}\|_2 \cdot f(x_t)}{\|f(x_t)\|_2}, \tag{10}$$

where $[\cdot]$ denotes the concatenate operation, $\boldsymbol{w}_{y_s}$ denotes the prototype vector of the y_s-th identity, $\boldsymbol{W}' \in \mathbb{R}^{(P+1) \times d}$ is the parameter matrix of the composed classifier. As expressed in the above equation, the dynamically created virtual prototype vector is derived from the feature of the target instance of the mixed pair. It has the same angular as the target feature and the same norm as the source prototype vector. The composed classifier can distinguish $(P+1)$ identities apart. Wherein, the $(P+1)$-th identity corresponds to the target individual of the source-target pair. To make the labels compatible with the composed classifier, we pad one-hot labels of source samples to $(P+1)$-d with the zero value. As for the labels of target samples, the final element of this $(P+1)$-d one-hot vector is always activated. Since without specific class assignment, target samples should always be identified as themselves. The feature learning of the mixed data is constrained by the cross-entropy loss between the prediction of the newly composed classifier and the mixed label:

$$\mathcal{L}_m = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{P+1} y_{i,j}^m \log p(j | x_i^m; \boldsymbol{W}'), \tag{11}$$

where $y^m_{i,j}$ denotes the j-th element of the mixed label y^m_i, $p(j|x^m_i; \boldsymbol{W}')$ denotes the probability predicted by the compose classifier that x^m_i belongs to the j-th identity. Empirically, we find that replacing the up-to-date feature $f(x^t)$ in Eq. (10) with its counterpart in the memory can benefit the stability of the training. Therefore, we adopt this practice in the following experiments. Our supplementary material provides detailed experimental results.

Remarks. Both our method and Virtual Softmax [2] introduce the concept of the virtual prototype. However, they are different in motivation and implementation. In terms of motivation, Virtual Softmax introduces the virtual prototype to enhance the discrimination of learned features under the fully-supervised setup. By contrast, our method uses it to adjust the label space of the classifier dynamically according to the input source-target pair. As for implementation, the direction of the virtual prototype is equal to that of the input feature in Virtual Softmax. Whereas in our method, the classifier operates on mixed samples. The virtual prototype has the same direction as the target instance of the input mixed pair.

3.4 Overall Loss Function

Neighborhood invariance for intra-camera matching and inter-camera-matching composes the supervision for the target domain, *i.e.*, $\mathcal{L}_t = \mathcal{L}_{intra} + \mathcal{L}_{inter}$. By combining it with the proposed constraint on the mixed data, we can obtain the final loss function for the model training:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_m. \tag{12}$$

One may ask why not impose a constraint (*e.g.* classification loss or triplet loss) on the source domain, just as other methods do. We remind that L_m already contains moderate supervision for the source domain. When the interpolation coefficient λ in Eq. (8) is sampled close to 1, L_m degrades to the classification loss on the source data. For experimental results, see our supplementary material.

4 Experiment

4.1 Dataset and Evaluation Protocol

We evaluate the performance of the proposed method on three public benchmarks, *i.e.* Market-1501 [53], DukeMTMC-reID [54,29], MSMT17 [40]. During training, we adopt two of the three datasets as the source domain and the target domain, respectively. During testing, we evaluate Cumulated Matching Characteristics (CMC) at rank-1, rank-5, rank-10 and mean average precision (mAP) in the testing set of the target domain.

4.2 Implementation Details

We adopt ResNet-50 [15] pre-trained on ImageNet [8] as the backbone of our model. The last downsampling operation of ResNet is discarded, which leads to an overall stride of 16. $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ are involved into the training at 10^{th} and 30^{th} epoch, respectively. In terms of optimizer, we employ Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of $1e-5$. The learning rate is set to 0.01 and 0.05 for the backbone layers and newly added layers, respectively. It is divided by 10 at 60^{th} epoch. The whole training process lasts for 70 epochs. As for data, each mini-batch contains 128 source images and 128 target images. All input images are resized to 256×128. Random horizontal flip, random crop and random erasing [55] are utilized for data augmentation. Unless otherwise specified, we follow the setting of scaling factor $s = 10$, neighborhood range $\epsilon = 0.8$, momentum of memory updating $\sigma = 0.6$, and parameter of Beta distribution $\alpha = 0.6$. During testing, we adopt the output of the final Batch Normalization layer as the image embedding. Cosine similarity is used as the measure for retrieval. All experiments are conducted on two NVIDIA TITAN V GPUs using Pytorch [27] platform.

4.3 Parameter Analysis

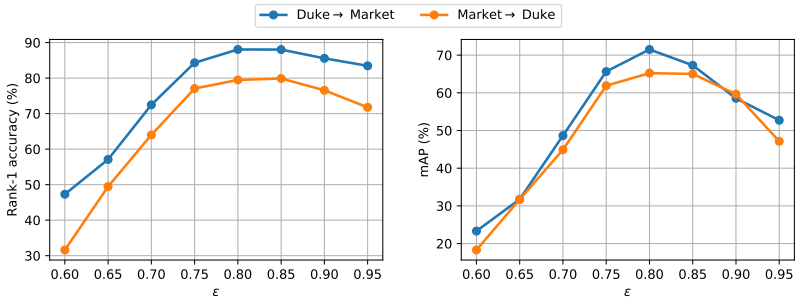
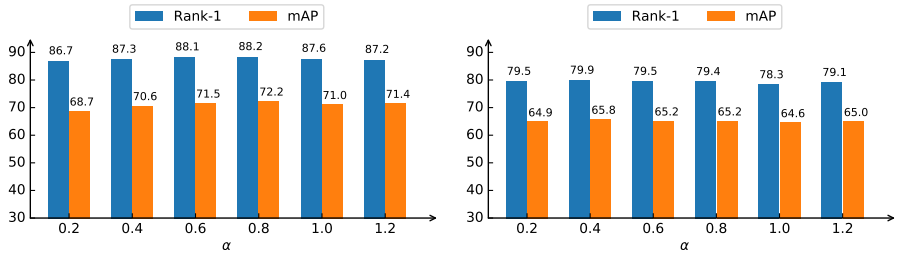


Fig. 4. Evaluation with different values of ϵ in Eq. (7).

Neighborhood range ϵ. To analyze the effect of ϵ, we vary its value in a reasonable scope and evaluate the performance under these settings. As illustrated in Fig. 4, both rank-1 accuracy and mAP first improve as ϵ decreases. Assigning too small value to ϵ may introduce considerable false positives, which is harmful to the learning of discriminative features. We obtain the optimal performance around $\epsilon = 0.8$. Our method is somewhat sensitive to the setting of neighborhood range. For detailed analysis, see our supplementary material.

Beta distribution parameter α. The parameter α determines the distribution of interpolation coefficient λ. Assigning a larger value to α leads to an stronger regularization. To investigate its effect, we vary the parameter α to five different



(a) Duke→Market (b) Market→Duke

Fig. 5. Evaluation with different values of the Beta distribution parameter α in Eq. (8).

s	Duke→ Market				Market→ Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
6	80.3	90.1	92.9	60.1	75.3	84.0	86.7	58.5
8	86.4	93.3	95.4	67.5	78.9	88.0	90.9	64.1
10	**88.1**	**94.4**	**96.2**	**71.5**	79.5	88.3	91.4	**65.2**
12	84.6	92.7	94.9	67.8	**79.7**	**89.1**	**91.6**	**65.2**
14	74.4	89.1	93.0	57.1	76.1	87.0	89.6	61.0

Table 1. Evaluation with different values of the scaling factor s in Eq. (2).

values and evaluate the performance under these settings. As shown in Fig. 5, both rank-1 accuracy and mAP fluctuate very slightly with the variation of α. This indicates that our method is relatively robust to the setting of cross-domain mixup.

Scaling factor s. The scaling factor s in Eq. (2) is crucial to the final performance. Large s can sharpen the probability distribution and ease the optimization. However, assigning too large value may make the task too trivial to learn discriminative features. We train the model under five different values of s and report their results in Tab. 1. As shown in Tab. 1, we obtain the optimal performance at $s = 10$ on Market-1501 and $s = 12$ on DukeMTMC-reID. The performance degrades dramatically when s gets too large or too small.

4.4 Ablation Study

In this section, we conduct extensive ablation studies on the adaptation between Market-1501 and DukeMTMC-reID. For the variants that do not involve cross-domain mixup loss $\mathcal{L}_m$, the supervision on the source data is necessary to ensure meaningful representations. Thus, we perform a classification task in the source domain just as ECN [57] and its follow-ups [58,10] do. Suppose $p(y_i^s|x_i^s)$ denotes the predicted probability that x_i^s belongs to the identity y_i^s. The loss function for the source data is defined as $\mathcal{L}_s = -\frac{1}{N_s}\sum_{i=1}^{N_s} \log p(y_i^s|x_i^s)$. See our supplementary material for ablation studies on other datasets.

Methods	Duke→Market				Market→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-1	mAP
Supervised Learning	90.7	96.6	98.0	74.8	82.7	91.0	93.7	66.4
Direct Transfer	48.9	65.1	71.8	19.8	30.0	44.9	50.9	15.0
$\mathcal{L}_s + \mathcal{L}_{ag}$	60.9	73.1	77.8	35.3	49.8	63.0	68.1	34.5
$\mathcal{L}_s + \mathcal{L}_{intra}$	70.6	83.9	88.6	44.6	70.0	82.4	86.0	52.1
$\mathcal{L}_m + \mathcal{L}_{intra}$	76.8	89.0	92.4	54.9	73.7	84.2	88.1	57.3
$\mathcal{L}_s + \mathcal{L}_{intra} + \mathcal{L}_{inter}$	81.2	91.7	94.2	59.2	76.2	87.5	90.4	59.6
$\mathcal{L}_m + \mathcal{L}_{intra} + \mathcal{L}_{inter}$	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2

Table 2. Ablation studies on Market-1501 and DukeMTMC-reID. **Supervised Learning**: Model trained with labeled target data. **Direct Transfer**: Model trained with only labeled source data.

Performance bound. As reported in the first two rows of Tab. 2, the model achieves promising performance when trained and tested in the same domain (termed *Supervised Learning*). However, such a model performs poorly when directly deployed to an unseen domain (termed *Direct Transfer*). Specifically, the model trained on DukeMTMC-reID achieves only 48.9% rank-1 accuracy on Market-1501, which is 41.8% lower than its single-domain counterpart. *Supervised Learning* and *Direct Transfer* behave as the upper-bound and lower-bound of the transfer performance, respectively.

Effect of camera-aware invariance learning. To investigate the effect of camera-aware invariance learning, we impose neighborhood invariance separately for intra-camera matching and inter-camera matching. From Row 3-4 in Tab. 2, we observe a considerable improvement when replacing camera-agnostic neighborhood loss $\mathcal{L}_{ag}$ with its intra-camera counterpart $\mathcal{L}_{intra}$. To be specific, rank-1 accuracy improves from 60.9% to 70.6% and 49.8% to 70.0% on Market-1501 and DukeMTMC-reID, respectively. This is interesting since $\mathcal{L}_{intra}$ even omits massive inter-camera candidates during the optimization. Such a phenomenon verifies our hypothesis mentioned in Sec. 3.2. That is, the discrepancy between intra-camera matching and inter-camera matching makes camera-agnostic neighborhood constraint ambiguous for the optimization. Besides, a discriminative intra-camera representation is beneficial for the cross-camera association. Furthermore, we add inter-camera neighborhood loss $\mathcal{L}_{inter}$ to the supervisory signal to validate its effectiveness. As shown in Row 6 of Tab. 2, the injection of inter-camera neighborhood loss improves the performance significantly. It leads to a 10.6% and 6.2% gain in rank-1 accuracy on Market-1501 and DukeMTMC-reID, respectively. Without the proposed cross-domain mixup component, such a concise variant is already on par with existing state-of-the-art methods.

Effect of cross-domain MixUp. We further investigate the effect of cross-domain mixup by incorporating it into the training. As shown in Tab. 2, the involvement of $\mathcal{L}_m$ improves the variant "$\mathcal{L}_s + \mathcal{L}_{intra}$" by 6.2% and 3.7% in terms of rank-1 accuracy on Market-1501 and DukeMTMC-reID, respectively. The gain is 6.9% and 3.3% when applying $\mathcal{L}_m$ to the variant "$\mathcal{L}_s + \mathcal{L}_{intra} + \mathcal{L}_{inter}$". Such a

Methods	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PTGAN [40]	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN [9]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
TJ-AIDL [39]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle [59]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL [56]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
MAR [46]	67.7	81.9	-	40.0	67.1	79.8	-	48.0
PAUL [45]	68.5	82.4	87.4	40.1	72.0	82.7	86.0	53.2
ARN [24]	70.3	80.4	86.3	39.4	60.2	73.9	79.5	33.4
ECN [57]	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
UDA [32]	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49.0
PAST [50]	78.4	-	-	54.6	72.4	-	-	54.3
SSG [12]	80.0	90.0	92.4	58.3	73.0	80.6	83.2	53.4
AE [10]	81.6	91.9	94.6	58.0	67.9	79.2	83.6	46.7
ECN++ [58]	84.1	92.8	95.4	63.8	74.0	83.7	87.4	54.4
MMT [14]	87.7	94.9	96.9	71.2	78.0	88.8	92.5	65.1
Ours	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2

Table 3. Comparison with state-of-the-art cross-domain methods on Market-1501 and DukeMTMC-reID. **Red** indicates the best and **Blue** the runner-up.

significant improvement validates the effectiveness of cross-domain mixup. It is noteworthy that the rank-1 accuracy of our final variant is only **2.6%** and **3.2%** lower than the supervised counterpart on Market-1501 and DukeMTMC-reID, respectively.

4.5 Comparison with State-of-the-art Methods

Results on Market-1501 dataset. We evaluate the performance of our method on Market-1501 using DukeMTMC-reID as the source domain. We compare the result with representative works of different directions, including the methods based on style transfer [40,9,25], the methods based on pseudo-label estimation [32,12,14], and those mining intra-domain cues [45,46,57,10]. As reported in Tab. 3, our method performs favorably against current leading methods in rank-1 accuracy and mAP. Note that both ECN [57] and its follow-ups [10,58] benefit a lot from CamStyle [59] augmentation, which requires an extra StarGAN [7]. MMT [14], the nearest rival, employs computationally intensive clustering operation and four ResNet-50 models in total (2 students and 2 teachers) to achieve the similar performance.

Results on DukeMTMC-reID dataset. We adopt Market-1501 as the source domain and evaluate the performance of the proposed approach on DukeMTMC-reID. As shown in the right part of Tab. 3, our methods is competitive against other state-of-the-arts. Both SSG [12] and PAUL [45] mine discriminative cues at the part level, which is orthogonal to our concerns. It has been widely validated in the field of supervised re-ID that part models are more powerful than their

vanilla counterparts in discrimination. Even so, our method achieves much higher performance than the two competitors. Compared with MMT [14], our method is superior in rank-1 and mAP with much less computational overhead.

Results on MSMT17 dataset. We further evaluate the transfer performance of our method on MSMT17. MSMT17 is characterized by large scale and abundant variations, which makes it much more challenging. As shown in Tab. 4, the proposed approach outperforms MMT [14] while using DukeMTMC-reID as the source domain. However, the performance is far inferior to MMT when the source domain is Market-1501. This is mainly attributed to the training instability induced by the unreliable neighborhood search at the early stage. We will investigate this issue in the future work.

Methods	Market→MSMT				Duke→MSMT			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PTGAN [40]	10.2	-	24.4	2.9	11.8	-	27.4	3.3
ECN [57]	25.3	36.3	42.1	8.5	30.2	41.5	46.8	10.2
AE [10]	25.5	37.3	42.6	9.2	32.3	44.4	50.1	11.7
SSG [12]	31.6	-	49.6	13.2	32.2	-	51.2	13.3
ECN++ [58]	40.4	53.1	58.7	15.2	42.5	55.9	61.5	16.0
MMT [14]	49.2	63.1	68.8	22.9	50.1	63.9	69.8	23.3
Ours	43.7	56.1	61.9	20.4	51.7	64.0	68.9	24.3

Table 4. Comparison with state-of-the-art cross-domain methods on MSMT17. **Red** indicates the best and **Blue** the runner-up.

5 Conclusion

In this paper, we propose a superior model for cross-domain person re-identification that takes both intra-domain variation and inter-domain shift into account. We adopt a neighborhood invariance approach to supervise feature learning in the target domain. However, we find that the neighbor search is highly biased due to the dramatic discrepancy across cameras. To avoid this issue, we propose to impose the constraint in a camera-aware manner. Furthermore, we devise a novel cross-domain mixup scheme to bridge the gap between the source domain and the target domain. To be more specific, it introduces the interpolation between the two domains as an intermediate state of the transfer. Extensive experiments validate the effectiveness of each proposal. By taking the two proposals together, our method outperforms existing state-of-the-arts by a large margin.

Acknowledgement This work was supported in part by the National Key R&D Program of China (No. 2018YFB1004602), the National Natural Science Foundation of China (No. 61836014, No. 61761146004, No. 61773375).

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv:1905.02249 (2019)
2. Chen, B., Deng, W., Shen, H.: Virtual class enhanced discriminative embedding learning. In: NeurIPS (2018)
3. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR (2017)
4. Chen, Y., Zhu, X., Gong, S.: Deep association learning for unsupervised video person re-identification. In: BMVC (2018)
5. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. In: NeurIPS (2017)
6. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR (2016)
7. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
9. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
10. Ding, Y., Fan, H., Xu, M., Yang, Y.: Adaptive exploration for unsupervised person re-identification. arXiv:1907.04194 (2019)
11. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications **14**, 83 (2018)
12. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: ICCV (2019)
13. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.: Horizontal pyramid matching for person re-identification. In: AAAI (2019)
14. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: ICLR (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
17. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
18. Huang, H., Yang, W., Chen, X., Zhao, X., Huang, K., Lin, J., Huang, G., Du, D.: Eanet: Enhancing alignment for cross-domain person re-identification. arXiv:1812.11369 (2018)
19. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Sbsgan: Suppression of inter-domain background shift for person re-identification. In: ICCV (2019)
20. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: CVPR (2018)
21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

22. Li, K., Ding, Z., Li, K., Zhang, Y., Fu, Y.: Support neighbor loss for person re-identification. In: ACM MM (2018)
23. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. In: ECCV (2018)
24. Li, Y.J., Yang, F.E., Liu, Y.C., Yeh, Y.Y., Du, X., Frank Wang, Y.C.: Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In: CVPR Workshop (2018)
25. Liu, J., Zha, Z.J., Chen, D., Hong, R., Wang, M.: Adaptive transfer network for cross-domain person re-identification. In: CVPR (2019)
26. Mao, X., Ma, Y., Yang, Z., Chen, Y., Li, Q.: Virtual mixup training for unsupervised domain adaptation. arXiv:1905.04215 (2019)
27. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NeurIPS Workshop (2017)
28. Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y.: Auto-reid: Searching for a part-aware convnet for person re-identification. In: ICCV (2019)
29. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshop (2016)
30. Rukhovich, D., Galeev, D.: Mixmatch domain adaptaion: Prize-winning solution for both tracks of visda 2019 challenge. arXiv:1910.03903 (2019)
31. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
32. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. arXiv:1807.11334 (2018)
33. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: ECCV (2018)
34. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV (2018)
35. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: ICML (2019)
36. Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. arXiv:1903.03825 (2019)
37. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: ECCV (2018)
38. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM (2018)
39. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
40. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: CVPR (2018)
41. Wu, A., Zheng, W.S., Lai, J.H.: Unsupervised person re-identification by camera-aware similarity consistency learning. In: ICCV (2019)
42. Wu, J., Liao, S., Wang, X., Yang, Y., Li, S.Z., et al.: Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In: ICME
43. Wu, J., Yang, Y., Liu, H., Liao, S., Lei, Z., Li, S.Z.: Unsupervised graph association for person re-identification. In: ICCV (2019)

44. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
45. Yang, Q., Yu, H.X., Wu, A., Zheng, W.S.: Patch-based discriminative feature learning for unsupervised person re-identification. In: CVPR (2019)
46. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: CVPR (2019)
47. Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., Bai, X.: Hard-aware point-to-set deep metric for person re-identification. In: ECCV (2018)
48. Zhai, Y., Guo, X., Lu, Y., Li, H.: In defense of the classification loss for person re-identification. In: CVPR Workshops (2019)
49. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
50. Zhang, X., Cao, J., Shen, C., You, M.: Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: ICCV (2019)
51. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: CVPR (2019)
52. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: CVPR (2019)
53. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
54. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV (2017)
55. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv:1708.04896 (2017)
56. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: ECCV (2018)
57. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: CVPR (2019)
58. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Learning to adapt invariance in memory for person re-identification. arXiv:1908.00485 (2019)
59. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camstyle: a novel data augmentation method for person re-identification. IEEE Transactions on Image Processing **28**, 1176–1190 (2019)
60. Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N.: Openmix: Reviving known knowledge for discovering novel visual categories in an open world. arXiv:2004.05551 (2020)
61. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV (2019)
62. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
63. Zhu, X., Zhu, X., Li, M., Murino, V., Gong, S.: Intra-camera supervised person re-identification: A new benchmark. In: ICCV Workshop (2019)