DomainMix: Learning Generalizable Person Re-Identification Without Human Annotations

Wenhao Wang[1], Shengcai Liao[2*], Fang Zhao[2], Cuicui Kang[3], Ling Shao[2,3]
[1]School of Mathematical Sciences (SMS), Beihang University, Beijing, China
[2]Inception Institute of Artificial Intelligence (IIAI), Masdar City, Abu Dhabi, UAE
[3]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, UAE
wangwenhao@buaa.edu.cn, scliao@ieee.org, fang.zhao@inceptioniai.org, cuicui.kang@mbzuai.ac.ae, ling.shao@ieee.org

Abstract

Existing person re-identification methods often have low generalization capability, which is mostly due to the limited availability of large-scale labeled training data. However, labeling large-scale training data is very expensive and time-consuming. To address this, this paper presents a solution, called DomainMix, which can learn a person re-identification model from both synthetic and real-world data, for the first time, completely without human annotations. This way, the proposed method enjoys the cheap availability of large-scale training data, and benefiting from its scalability and diversity, the learned model is able to generalize well on unseen domains. Specifically, inspired from a recent work generating large-scale synthetic data for effective person re-identification training, the proposed method firstly applies unsupervised domain adaptation from labeled synthetic data to unlabeled real-world data to generate pseudo labels. Then, the two sources of data are directly mixed together for supervised training. However, a large domain gap still exists between them. To address this, a domain-invariant feature learning method is proposed, which designs an adversarial learning between domain-invariant feature learning and domain discrimination, and meanwhile learns a discriminant feature for person re-identification. This way, the domain gap between synthetic and real-world data is much reduced, and the learned feature is generalizable thanks to the large-scale and diverse training data. Experimental results show that the proposed annotation-free method is more or less comparable to the counterpart trained with full human annotations, which is quite promising. In addition, it achieves the current state of the art on several popular person re-identification datasets under direct cross-dataset evaluation.

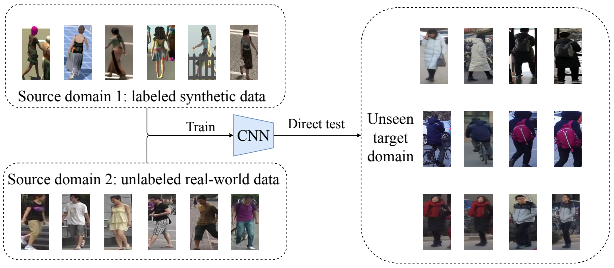
Source domain 1: labeled synthetic data
Source domain 2: unlabeled real-world data
Train
CNN
Direct test
Unseen target domain

Figure 1. The illustration of the proposed task A+B→C, *i.e.* how to use labeled synthetic data A and unlabeled real-world data B to train a model that can generalize well to an unseen target domain C.

1. Introduction

The goal of person re-identification (re-ID) is to match a given person across many gallery images captured at different times, locations, etc. With the development of deep learning, fully supervised person re-ID has been extensively investigated [25, 24, 21, 37, 3] and gained impressive progress. However, significant performance degradation can be observed when a trained model is tested on a previously unseen dataset. The generalization capability of known algorithms is hindered by two main aspects. First, the generalization capability of an algorithm is often ignored by its designer. There are only a few methods designed for domain generalization (DG). Second, the number of subjects in public datasets is limited, and their diversity is insufficient.

Labeling large-scale and diverse real-world datasets is expensive and time-consuming. For instance, labeling a dataset of the magnitude of MSMT17 [29] requires three labelers to work for two months. To address this, RandPerson [28] inspires us to use large-scale synthetic data for effective person re-identification training, which gets rid of the need of human annotations. However, if using synthetic

[*] Shengcai Liao is the Corresponding Author.

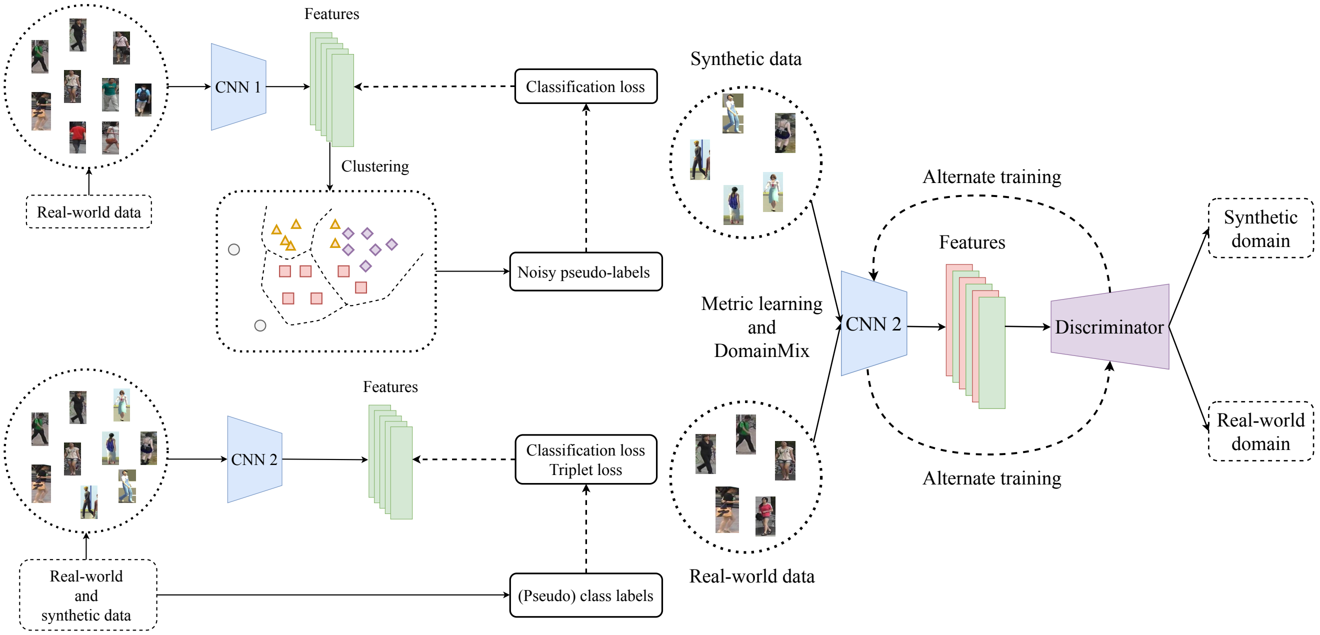

Figure 2. The pipeline of the DomainMix approach. DomainMix is trained in the following order: upper left corner, lower left corner, and right. Specifically, Stage 1 is the pseudo labels generation and refinement. Stage 2 is the backbone training and Stage 3 is the alternate training of the discriminator and backbone.

data alone, the generalization ability of the learned model is still limited due to the domain gap between the synthetic and real-world data. Therefore, a solution is provided in [28] which learns from mixed synthetic data and labeled real-world data. However, though performance is improved, this solution still relies on heavy human annotations of the real-world data, and the domain gap still exists which is suboptimal for generalization.

Therefore, the goal of this paper is to learn generalizable person re-identification completely without human annotations, so as to make use of a large amount of unlabeled real-world data. Specifically, we aim at how to combine a labeled synthetic dataset with unlabeled real-world data to learn a ready-to-use model with good generalization capability. The proposed setting is illustrated in Fig. 1, which is denoted as A (labeled) + B (unlabeled) → C (unseen target domain) with direct cross-dataset evaluation on C. The key to achieve domain generalization here is to make full use of the discriminative labels in the synthetic domain and the style and diversity of unlabeled real-world images simultaneously. A plausible method to tackle this problem would be Unsupervised Domain Adaptation (UDA) from A to B and trying to test it on C. However, the goal of UDA is different; it transfers the knowledge from the source domain A to the target domain B, and the testing is performed on the same target domain B. After the transfer, the model will learn domain-specific features from the less reliable real-world data without annotations and ignore the value of the large-scale high-quality labeled synthetic data. Therefore, directly applying UDA from A to B will have inferior generalization capability on C.

To address this problem, a solution called DomainMix is proposed, for discriminant, domain-invariant, and generalizable person re-identification feature learning. Specifically, a cluster-based UDA algorithm is adopted to label the real-world dataset with the help of the large-scale synthetic data RandPerson [28]. Then, the backbone is trained for feature extraction using the labeled synthetic data and real-world data with pseudo labels. After this, we design a discriminator to classify the feature extracted by the backbone into a synthetic or real-world domain. In addition, the backbone is trained with two tasks: class-discriminant metric learning for person re-identification, and domain-invariant adversarial learning for confusing the discriminator. The above process is illustrated in Fig. 2. Following this pipeline, the need of human annotations is completely eliminated, and the domain gap between the synthesized and real-world data is reduced, so that the generalization capability is improved thanks to the large-scale and diverse training data.

Comprehensive experiments are conducted on several popular person re-identification datasets. Experimental results show that the proposed annotation-free method DomainMix is more or less comparable to the counterpart trained with full human annotations, which is quite promising. In addition, it achieves the current state of the art on these datasets under direct cross-dataset evaluation.

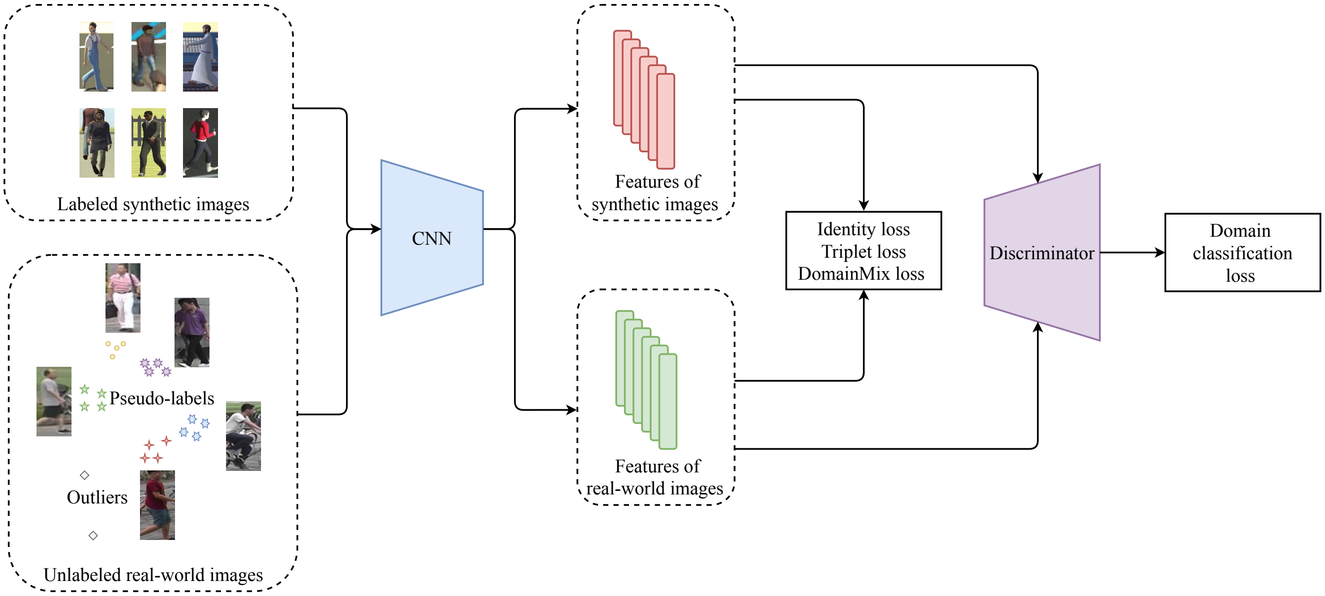
Labeled synthetic images
Pseudo-labels
Outliers
Unlabeled real-world images
CNN
Features of synthetic images
Features of real-world images
Identity loss
Triplet loss
DomainMix loss
Discriminator
Domain classification loss

Figure 3. The design of the DomainMix approach. By training the backbone and discriminator alternately, the backbone can extract domain-invariant and discriminative features, and the discriminator correctly classifies each feature into its domain.

2. Related Work

2.1. Unsupervised Domain Adaptation for Person Re-ID

With the development of deep learning, fully supervised person re-ID has gained impressive progress. However, expensive and time-consuming manual labeling is a must. Further, the domain gap between different datasets prevents a model trained on one domain from performing well on another. As a result, Unsupervised Domain Adaptation (UDA) for person re-ID has been introduced. Its goal is to learn a model on a labeled source domain and fine-tune it to an unlabeled target domain. The main UDA algorithms can be categorized into three classes. The first is image-level methods [39, 5, 29, 16], which use a generative adversarial network (GAN) [10] to translate the image style. The second class is feature-level methods [17, 2, 19], which aim to find domain-invariant features between different domains. The last category is cluster-based algorithms [20, 34, 30, 33, 7, 8, 6, 35], which generate pseudo labels to help fine-tune on the target domain.

However, a drawback of UDA is that a lot of data inevitably needs to be collected to fine-tune the trained model when facing a new scene. Although no manual labeling is required, gathering enough data to train a deep learning model is still time-consuming or even impossible. For instance, one application of person re-ID is criminal investigation, which requires suspects to be tracked across different scenes. However, fine-tuning a model to a large number of scenes is prohibitive. Therefore, developing an out-of-the-box algorithm is necessary.

2.2. Domain Generalization for Person Re-ID

Domain Generalization (DG) for person re-ID was first studied in [31], aiming to generalize a trained model to unseen scenes. In recent years, with the increasing accuracy of fully supervised person re-ID and the limitations of UDA, DG has begun to attract attention again. For instance, DualNorm [13] uses instance normalization to filter out variations in style statistic in earlier layers to increase the generalization capability. SNR [14] filters out identity-irrelevant interference and keeps discriminative features by using an attention mechanism. QAConv [18] constructs query-adaptive convolution kernels to find local correspondences in feature maps, which is more generalizable than using features.

Other works, such as RandPerson [28], focus on using synthetic data to enlarge the diversity and scale of person re-ID datasets. However, although algorithms trained on RandPerson [28] are more generalizable than most of those trained on real-world data, there is still much room for improvement because of the domain gap between the synthetic and the real-world datasets. Therefore, in their original paper [28], the authors also tried to directly combine RandPerson with a real-world dataset to further improve performance. However, two drawbacks still exist. On the one hand, directly mixing RandPerson with real-world data still requires time-consuming labeling. On the other hand, the domain gap between the synthetic and real-world datasets is still ignored. Thus, this paper discusses how to design a generalizable re-ID approach that can exploit valuable real-world unlabeled data and eliminate the domain gap between synthetic and real-world datasets.

2.3. Methods for Reducing Domain Gap

Domain gap hinders one trained model performs well on an unseen dataset. In the task of UDA for person re-ID, some methods, such as PTGAN [29], utilize GAN [10] to transfer the image style of the source domain to the target domain. The methods reduce the domain gap from the image-level. Another category is feature-level and our method belongs to it. Some methods try to train a domain-invariant model by reducing the pairwise domain discrepancy with Maximum Mean Discrepancy (MMD) [26]. However, this pipeline, which shares the same classes between domains, is not suitable for person re-ID task because the identities in two re-ID domains are different. One work similar to us is CaNE [32], and its main contribution is proposing a calibrated negative entropy loss to learn domain-invariant features. In comparison, the proposed DomainMix is different from the above method in three folds: First, at the domain level, DomainMix explores domain-invariant features from the synthetic and real-world data, while CaNE [32] learns camera-invariant features in only one dataset. Second, at the task level, CaNE [32] focuses on the imbalance of nuisance classes while DomainMix tries to address a new proposed task. At last, the process of learning domain-invariant features is different between CaNE [32] and DomainMix.

3. Proposed Method

3.1. Problem Definition

Two source domains S_1 and S_2, where S_1 is a synthetic dataset and S_2 is a real-world dataset, are given. For the synthetic dataset, the labels and images are both available. It is denoted as $D_{s_1} = \left\{ (x_i^{s_1}, y_i^{s_1}) \Big|_{i=1}^{N_{s_1}} \right\}$, where $x_i^{s_1}$ and $y_i^{s_1}$ are the i-th training sample and its corresponding person identity label, respectively, and N_{s_1} is the number of images in the synthetic dataset. For the real-world dataset, only the images are available. The N_{s_2} images in the real-world dataset are denoted as $D_{s_2} = \left\{ x_i^{s_2} \Big|_{i=1}^{N_{s_2}} \right\}$. Besides, a target domain T, which is a real-world dataset different from D_{s_2}, is given. It is denoted as $D_t = \left\{ x_i^t \Big|_{i=1}^{N_t} \right\}$, where x_i^t denotes the i-th target-domain image and N_t is the total number of target-domain images. This setting simulates the practical application scene, *i.e.* synthesizing labeled datasets is time-saving and cheap, while labeling a large-scale real-world dataset is time-consuming and expensive. Our goal is to design an algorithm that can be trained on the datasets D_{s_1} and D_{s_2}, and then directly generalized to unseen D_t without fine-tuning.

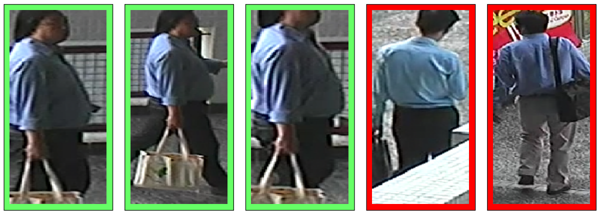

Figure 4. The above five images are given the same pseudo labels. However, the last two have different identities from the first three. Although the pseudo labels are noisy, the five people at least share similar attributes, such as the color of their clothes, gender, etc., which are also helpful for the learning process.

3.2. DomainMix

To tackle the problem mentioned above, we propose the DomainMix approach. In this approach, pseudo labels of an unlabeled real-world dataset are first generated by a UDA algorithm. Then, the backbone is trained over several epochs for feature extraction. Finally, the DomainMix loss is used to learn domain-invariant features, and thus our algorithm can generalize well to unseen target domains. The pipeline is shown in Fig. 3.

3.2.1 Pseudo labels Generation and Refinement

To make full use of the valuable real-world datasets, labeling them is essential. With the development of cluster-based UDA algorithms, given one dataset with labels and one without, it is now possible to generate pseudo labels of the latter by using the former. But there is often noise in pseudo labels as displayed in Fig. 4.

In our task, a labeled synthetic dataset and an unlabeled real-word dataset are provided. Therefore, the real-word dataset can be labeled with a cluster-based UDA algorithm and the labeled synthetic dataset. The real-word dataset with pseudo labels is denoted as $D'_{s_2} = \left\{ (x_i^{s_2}, \hat{y}_i^{s_2}) |_{i=1}^{N_{s_2}} \right\}$, where $\hat{y}_i^{s_2}$ is the i-th sample's pseudo label. However, there are many outliers, *i.e.* pseudo labels with only a few images. Furthermore, some pseudo labels with too many images may bring noise into the algorithm. Therefore, it is essential to refine pseudo labels. The pseudo labels set is denoted as $L_1 = \left\{ l_i \Big|_{i=1}^{M} \right\}$, where l_i is the i-th pseudo label, and M is the total number of pseudo labels. Given the lower bound b_l and higher bound b_h, labels with a total number of images below b_l or above b_h are both discarded. Thus the refined pseudo labels dataset is obtained by

$$L_2 = \{ l_i \mid l_i \in L_1, b_l \leq S(l_i) \leq b_h \}, \tag{1}$$

where $S(l_i)$ denotes the number of images belonging to the i-th pseudo label.

3.2.2 Backbone Training

Given the images from a labeled synthetic source domain D_{s_1} and a real-world source domain with pseudo labels D'_{s_2}, the backbone is trained to model a feature encoder function $F\left(\cdot\,|\theta\right)$, which can transform each input sample x_i^s to its feature representation $F\left(x_i^s\,|\theta\right)$. Then, the features represented are classified by an identity classifier C_s to predict which classes they belong to. A classification loss $L_{id}^s\left(\theta\right)$ and a triplet loss $L_{tri}^s\left(\theta\right)$ [12] are adopted to train the backbone. They are defined as

$$\mathcal{L}_{id}^s(\theta) = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}_{ce}\left(C_s\left(F\left(x_i^s \mid \theta\right)\right), y_i^s\right), \qquad (2)$$

and

$$\begin{aligned}\mathcal{L}_{tri}^s(\theta) = \tfrac{1}{N_s}\textstyle\sum_{i=1}^{N_s}\max\big(0, m + \left\|F\left(x_i^s \mid \theta\right) - F\left(x_{i,p}^s \mid \theta\right)\right\| \\ - \left\|F\left(x_i^s \mid \theta\right) - F\left(x_{i,n}^s \mid \theta\right)\right\|\big),\end{aligned} \qquad (3)$$

where N_s is the sum of the number of images in the two source domains, $\|\cdot\|$ denotes the L^2-norm distance, m is the triplet distance margin, $\mathcal{L}_{ce}(\cdot,\cdot)$ represents the cross-entropy loss, y_i^s is the corresponding label or pseudo label, and the subscripts ${}_{i,p}$ and ${}_{i,n}$ indicate the hardest positive and the hardest negative index for the sample x_i^s in a mini-batch. The overall loss for training the backbone is defined as

$$\mathcal{L}_1^s\left(\theta\right) = \lambda^{s_1}\mathcal{L}_{id}^s\left(\theta\right) + \mathcal{L}_{tri}^s\left(\theta\right), \qquad (4)$$

where λ^{s_1} is a parameter for weighting the two losses.

3.2.3 Alternate Training of Discriminator and backbone

Once the backbone has been trained for feature extraction, the discriminator and the backbone are trained alternately. A discriminator is used to classify a given feature into its domain. Specifically, features of the images in the synthetic and real-world datasets are extracted by the backbone. Then a discriminator is trained to judge which dataset the extracted feature comes from. When training the discriminator, the cross-entropy loss $\mathcal{L}_{ce}$ is adopted. Thus the domain classification loss is defined as

$$\mathcal{L}_d^s(\theta) = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}_{ce}\left(C_d\left(F\left(x_i^s \mid \theta\right)\right), d_i^s\right), \qquad (5)$$

where C_d denotes the discriminator and d_i^s is the domain label of the i-th image, *i.e.* if the image belongs to the synthetic dataset, $d_i^s = 0$, and if it belongs to the real-world dataset, $d_i^s = 1$.

To encourage the backbone to extract domain-invariant features, beyond discriminative metric learning for re-ID, it is also trained to confuse the domain discriminator. Therefore, a DomainMix loss is proposed. Two kinds of DomainMix loss $\mathcal{L}_{dm}$ are designed. One is the symmetrized Kullback–Leibler (KL) divergence, and the other one is named as the domain balance.

The symmetrized KL divergence can be calculated as

$$\begin{aligned}\mathcal{L}_{kl}(\theta) = \tfrac{1}{N_s}\textstyle\sum\limits_{i=1}^{N_s}\left(\mathcal{D}_{kl}\left(C_d\left(F\left(x_i^s \mid \theta\right)\right)\|\, h_i\right) + \right. \\ \left.\mathcal{D}_{kl}\left(h_i\left\|C_d\left(F\left(x_i^s \mid \theta\right)\right)\right.\right)\right),\end{aligned} \qquad (6)$$

where $\mathcal{D}_{kl}(\cdot\|\cdot)$ is the KL divergence and

$$h_i \equiv \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)^T \in \mathbb{R}^n, \qquad (7)$$

where n is the number of source domains, *i.e.* two in our setting. When the discriminator is fixed, this loss encourages the backbone to produce features that the discriminator cannot easily classify. Therefore, domain-invariant features can be extracted from images by the backbone.

For the domain balance, the definition is

$$\mathcal{L}_{db} = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\sum_{j=1}^{n}\left(x_j^i\log\left(x_j^i\right) + 0.5\right)\right), \qquad (8)$$

where x_j^i is the j-th coordinate of $C_d\left(F\left(x_i^s \mid \theta\right)\right)$. In this loss, considering the function

$$f(x) = x\log(x) + 0.5, x \in (0,1), \qquad (9)$$

the second derivative of f is

$$f''\left(x\right) = \frac{1}{x} > 0. \qquad (10)$$

Therefore, it is a convex function. Given $\sum_{j=1}^{n} x_j^i$ =1, the minimum value of the function can be achieved when $x_j^i = 1/n (j = 1, 2, \ldots, n)$, according to Jensen's inequality. In conclusion, when $\mathcal{L}_{db}$ is minimized, the backbone can extract domain-invariant features by confusing the discriminator.

Through alternate training with $\mathcal{L}_d^s(\theta)$ and $\mathcal{L}_2^s(\theta)$, which is calculated as

$$\mathcal{L}_2^s(\theta) = \lambda^m\mathcal{L}_{dm}(\theta) + \lambda^{s_2}\mathcal{L}_{id}^s(\theta) + \mathcal{L}_{tri}^s(\theta), \qquad (11)$$

where λ^m and λ^{s_2} are the balance parameters, the discriminator can classify a given feature into its domain, and the backbone can extract domain-invariant and discriminative features.

To summarize the proposed algorithm, the pseudo codes are given in Algorithm 1.

Algorithm 1: DomainMix for generalizable person re-ID

```
Require: Labeled synthetic dataset $D_{s_1}$ and
         unlabeled real-world dataset $D_{s_2}$;
Require: Weighting factors $\lambda^{s_1}$, $\lambda^{s_2}$, and $\lambda^{m}$ for
         Eqs. (4), (11);
Stage 1: Pseudo labels generation and refinement
Generate pseudo labels for data in $D_{s_2}$ with UDA;
Refine the pseudo labels;
Stage 2: Backbone training
for $n \leftarrow 1$ to $num\_epochs\_1$ do
    for each mini-batch $\{x_i^s, y_i^s\} \subset D_{s_1} \cup D'_{s_2}$ do
        Update the parameters of the backbone by
          minimizing the objective function Eq. (4);
    end
end
Stage 3: Alternate training of discriminator and
  backbone
for $n \leftarrow 1$ to $num\_epochs\_2$ do
    if $n \equiv 0 \pmod 2$ then
        for each mini-batch $\{x_i^s, y_i^s\} \subset D_{s_1} \cup D'_{s_2}$
          do
            Update the discriminator by minimizing
              the objective function Eq. (5) with
              backbone fixed;
        end
    end
    if $n \equiv 1 \pmod 2$ then
        for each mini-batch $\{x_i^s, y_i^s\} \subset D_{s_1} \cup D'_{s_2}$
          do
            Update the backbone by minimizing the
              objective function Eq. (11) with
              discriminator fixed;
        end
    end
end
```

4. Experiments

4.1. Datasets and Evaluation Metrics

the DomainMix, extensive experiments are conducted on four widely used public person re-ID datasets. Among them, RandPerson (RP) [28] is selected as the synthetic dataset. Its subset contains $8,000$ persons in $132,145$ images. Nineteen cameras were used to capture them under eleven scenes. All images in the subset are used as training data, *i.e.*, no gallery or query is available. The real-world datasets used are Market-1501 [36], CUHK03-NP [38, 15], and MSMT17 [29]. Market-1501 [36] includes $1,501$ labeled persons in $32,668$ images. The training set has $12,936$ images of 751 identities. For testing, the query has $3,368$ images and the gallery has $19,732$ images. CUHK03-NP [38, 15] contains $1,467$ persons from six cameras. In this dataset, $7,365$ images of 767 identities are used for training. For testing, there are $1,400$ queries and $5,332$ gallery images. MSMT17 [29] is the most diverse and challenging re-ID dataset, consisting of $126,441$ bounding boxes of $4,101$ identities taken by 15 cameras. There are $32,621$ images for training, while the query has $11,659$ images and the gallery has $82,161$ images.

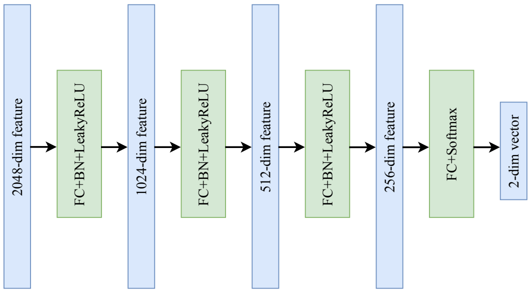

Figure 5. The design of the discriminator. Through multiple fully connected and non-linear layers, the discriminator can classify the feature of a given image into its domain.

Evaluation metrics are mean average precision (mAP) and cumulative matching characteristic (CMC) at rank-1, rank-5, and rank-10. The models trained on the source domains are directly tested on the target domain without transfer learning. Single-query evaluation protocols without post-processing methods is adopted.

4.2. Implementation Details

DomainMix is trained on four Tesla-V100 GPUs. The ImageNet-pre-trained [4] ResNet-50 [11] and IBN-ResNet-50 [22] are adopted as the backbone. All images are resized to 256×128 before being fed into the networks. Each training batch includes 64 persons of 16 actual or generated identities.

Stage 1**: Pseudo labels generation and refinement.** SpCL [8] is used to generate pseudo labels. We follow its the standard training settings, but the training iterations in each epoch are set to 800 due to the huge amount of data. For each generated label, if the number of images belonging to it is less than 5 or greater than 120, the label and images will be discarded.

Stage 2**: Backbone training.** Before alternate training, it is essential to first train the backbone to extract discriminative features. The λ^{s_1} in equation 4 is set to 10. The backbone is trained for 20 epochs. The initial learning rate is set to 3.5×10^{-4}, and it is decreased to $1/10$ of its previous value on the 10th and 15th epoch. The number of iterations in each epoch is $2,000$.

Stage 3**: Alternate training of discriminator and backbone.** In this stage, the discriminator and backbone are trained alternately. The design of the discriminator is

Table 1. Ablation studies for the DomainMix on the 'RP+MSMT → Market' and 'RP+CUHK → Market' tasks. 'Initial' denotes using the original noisy pseudo labels, and 'refined' denotes using pseudo labels refinement. 'RP' denotes RandPerson [28], and 'DB' is short for 'domain balance'.

RP+MSMT → Market	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
DirectMix (initial)	37.4	66.3	43.3	72.3
DirectMix (refined)	41.0	70.2	47.2	76.0
DomainMix (K-L)	43.1	72.1	49.0	76.5
DomainMix (DB)	42.9	72.0	48.9	76.8
Only RandPerson	36.6	65.9	41.2	70.5
Only MSMT (labeled)	31.3	62.2	39.0	70.1
DirectMix (labeled)	43.1	70.9	49.4	77.1
DomainMix (labeled)	45.2	72.2	51.4	78.0
DirectMix (unlabeled)	41.0	70.2	47.2	76.0
DomainMix (unlabeled)	42.9	72.0	48.9	76.8
RP→MSMT (SDA [9])	26.6	56.3	31.3	60.9
RP→MSMT (MMT [7])	22.7	46.5	30.0	57.5
RP→MSMT (SpCL [8])	24.2	49.8	33.5	60.4
DomainMix (Stage 2)	41.0	70.2	47.2	76.0
DomainMix (Stage 3)	42.9	72.0	48.9	76.8

RP+CUHK → Market	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
DirectMix (initial)	38.2	68.2	44.0	72.9
DirectMix (refined)	39.0	67.6	44.6	72.9
DomainMix (K-L)	40.3	68.3	46.0	74.3
DomainMix (DB)	40.5	68.5	46.4	74.2
Only RandPerson	36.6	65.9	41.2	70.5
Only CUHK (labeled)	28.3	56.5	38.7	67.3
DirectMix (labeled)	41.4	69.3	48.3	74.9
DomainMix (labeled)	43.7	72.1	50.4	77.2
DirectMix (unlabeled)	39.0	67.6	44.6	72.9
DomainMix (unlabeled)	40.5	68.5	46.4	74.2
RP→CUHK (SDA [9])	26.6	55.1	30.4	58.6
RP→CUHK (MMT [7])	24.6	51.2	29.9	56.3
RP→CUHK (SpCL [8])	9.3	24.1	18.3	39.4
DomainMix (Stage 2)	39.0	67.6	44.6	72.9
DomainMix (Stage 3)	40.5	68.5	46.4	74.2

shown in Fig. 5. The λ^{s_2} and λ^{m} in equation 11 are set to 10 and 1, respectively. The above process lasts for 20 epochs. The learning rate of the discriminator is 100 times that of the backbone.

4.3. Ablation Study

Comprehensive ablation studies are performed to prove the effectiveness of each component in the DomainMix approach. ResNet-50 [11] and IBN-ResNet-50 [22] are used as backbones. Two different domain generalization tasks are selected: labeled RandPerson [28] with unlabeled MSMT17 [29] to Market-1501 [36] and labeled RandPerson [28] with unlabeled CUHK03-NP [38, 15] to Market-1501 [36]. The experimental results and analyses are reported below.

Effectiveness of pseudo labels refinement. To investigate the necessity of refining the pseudo labels, we compare the domain generalization capability of a model trained on two different real-world datasets, *i.e.* MSMT17 [29] and CUHK03-NP [38, 15]. The baseline model performances are shown in Table 1 as "DirectMix (initial)". The pseudo labels refinement brings 3.6% and 3.9% in mAP improvement on the ResNet-50 and IBN-ResNet-50 backbones respectively for the 'RP+MSMT → Market' task. For the 'RP+CUHK → Market' task, the mAP increases by about 1%. Because MSMT17 [29] is a large-scale and challenging dataset, the labels are not easy to be predicted by the UDA algorithm, and many outliers (about 17% of the total training images) are generated. Failing to discard them may influence the gradient descent of the classification loss. Therefore, the refinement process improves the performance. However, because the scale and diversity of CUHK03-NP [38, 15] are both limited, the UDA algorithm can predict the labels precisely, and only about 8% of images are considered outliers. Therefore, though improvement is obtained, the impact is not so obvious.

Influence of DomainMix loss. To verify the necessity of using the DomainMix loss to learn domain-invariant features, results obtained with and without this loss are compared in the experiment and the results are shown in Table 1. The baselines are denoted as "DirectMix (refined)". All experiments with the use of DomainMix loss show distinct improvement on both the 'RP+MSMT → Market' and 'RP+CUHK → Market' tasks. Specifically, the mAP increases by 2.1% on ResNet-50 and 1.8% on IBN-ResNet-50 when the real-world source domain is MSMT17 [29]. As for the 'RP+ CUHK → Market' task, similar mAP improvement of 1.5% and 1.8% on the two network structures can be observed. When comparing the K-L divergence with the domain balance (DB), they are both beneficial for learning domain-invariant features, and no apparent difference is observed. For simplicity, the domain balance is used as the DomainMix loss in the following experiments. The improvement brought by DomainMix loss on all tasks (real-world datasets are all unlabeled) are displayed in Table 2.

Effectiveness of using unlabeled real-world dataset. We also verify the effectiveness of using the unlabeled real-world dataset. The baselines are denoted as "Only RP/MSMT (labeled)/CUHK (labeled)" in Table 1. On the one hand, compared to only training with synthetic data, mixing unlabeled real-world data with synthetic data brings up to 7.7% improvement in mAP. Further, if only labeled real-world data is adopted for training, the mAP drops by up to 12.3%. On the other hand, compared to adding labeled real-world data to synthetic data, though performance decreases can be observed, using unlabeled real-world data still achieves competitive performance. Thus, the real-world data is necessary for learning domain-invariant features and improving performance.

Table 2. Ablations studies for DomainMix loss on several domain generalizable person re-ID tasks. The effectiveness of DomainMix loss is proved by comparing “DirectMix” with “DomainMix”.

DirectMix	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
RP+MSMT → Market	41.0	70.2	47.2	76.0
RP+MSMT → CUHK	14.4	15.1	15.9	16.3
RP+Market → MSMT	11.7	31.6	16.0	40.3
RP+Market → CUHK	14.5	15.9	16.5	16.4
RP+CUHK → Market	39.0	67.6	44.6	72.9
RP+CUHK → MSMT	11.3	30.7	15.2	39.9

DomainMix	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
RP+MSMT → Market	42.9	72.0	48.9	76.8
RP+MSMT → CUHK	15.7	16.1	17.3	18.0
RP+Market → MSMT	11.9	31.9	16.3	40.5
RP+Market → CUHK	15.8	16.9	17.9	18.5
RP+CUHK → Market	40.5	68.5	46.4	74.2
RP+CUHK → MSMT	11.5	31.0	15.6	40.5

Table 3. Comparison with state-of-the-arts on Market1501 [36], MSMT17 [38, 15], and CUHK03-NP [38, 15]. ‘*’ denotes our implementation, and ‘†’ indicates that the results are reproduced based on the authors’ codes. ‘L’ or ‘U’ denotes the used source data is labeled or unlabeled, respectively.

Method	Source data	Market1501	
		mAP	rank-1
MGN [27]	MSMT (L)	25.1	49.7
CaNE [32]	MSMT (L)	22.5	50.1
CaNE-Dual [32]	MSMT (L)	30.3	59.1
OSNet-IBN [40]	MSMT (L)	37.2	66.5
SNR [14]	MSMT (L)	41.4	70.1
QAConv† [18]	MSMT (L)	34.7	65.4
Baseline*	RandPerson	36.6	65.9
OSNet-IBN† [40]	RandPerson	39.0	67.0
Baseline-IBN*	RandPerson	41.2	70.5
DomainMix	RP+MSMT (U)	42.9	72.0
DomainMix-OSNet-IBN	RP+MSMT (U)	45.1	72.4
DomainMix-IBN	RP+MSMT (U)	**48.9**	**76.8**

Method	Source data	MSMT17	
		mAP	rank-1
QAConv† [18]	Market (L)	6.1	20.4
Baseline*	RandPerson	9.6	27.7
OSNet-IBN† [40]	RandPerson	12.4	34.3
Baseline-IBN*	RandPerson	12.3	34.3
DomainMix	RP+Market (U)	11.9	31.9
DomainMix-OSNet-IBN	RP+Market (U)	15.1	38.9
DomainMix-IBN	RP+Market (U)	**16.3**	**40.5**

Method	Source data	CUHK03-NP	
		mAP	rank-1
MGN [27]	MSMT (L)	7.4	8.5
MuDeep [23]	Market (L)	9.1	10.3
QAConv† [18]	MSMT (L)	14.6	16.7
Baseline*	RandPerson	13.2	14.4
OSNet-IBN† [40]	RandPerson	12.9	13.6
Baseline-IBN*	RandPerson	13.3	14.2
DomainMix	RP+Market (U)	15.8	16.9
DomainMix-OSNet-IBN	RP+Market (U)	16.5	17.7
DomainMix-IBN	RP+Market (U)	**17.9**	**18.5**

Comparison with UDA algorithms and improvement of different stages. To show the state-of-the-art UDA algorithms cannot handle the task well and prove effectiveness of each stage in the proposed DomainMix, the performance of them is shown in Table 1. Because SpCL [8] is used to generate pseudo labels, “DomainMix (Stage 1)” is “RP → MSMT/CUHK (SpCL)”. “RP → MSMT/CUHK (SDA/MMT/SpCL)” denotes three state-of-the-art UDA algorithms. SDA [9] uses the GAN to reduce the domain gap between RandPerson and MSMT/CUHK. However, obvious performance degradation on the unseen domain can be observed because of the bias to MSMT/CUHK and the neglection of RandPerson. SpCL [8] is a cluster-based algorithm, which uses domain specific batch normalization (DSBN) [1] and combines the source domain with the target domain for training. However, the DSBN hinders the generalization capability because the BN statistics are biased to a certain domain. When following the pipeline of the DomainMix, performance improvement can be found.

4.4. Comparison with the State-of-the-arts

The proposed DomainMix approach is compared with state-of-the-art methods on three domain generalization tasks, *i.e*. directly testing on Market1501 [36], MSMT17 [29], and CUHK03-NP [38, 15]. The experimental results are shown in Table 3. Note that different source training datasets are used in Table 3, but we used unlabeled real-world data while existing results used labeled one. We obtain significant improvements of 7.5%, 10.2%, and 3.3% in mAP on Market1501 [36], MSMT17 [29], and CUHK03-NP [38, 15] domain generalization tasks, respectively. Although some methods use much stronger backbones, the DomainMix approach still outperforms them. Specifically, OSNet-IBN [40] uses the Omni-Scale Network with IBN layers, CaNE-Dual [32] uses ResNet-50 with a dual-branch, and MGN [27] uses the Multiple Granularity Network. Further, for example, the performance of the proposed DomainMix can be improved by using OSNet-IBN as backbone. For QAConv [18], though its performance is relatively high, because it needs to store feature maps of images rather than features to match, more memory is needed. It also displays instability in performance across different experiments with the same settings. SNR [14] proposes a style normalization and restitution module, and the DomainMix may achieve further performance improvement with the help of this plug-and-play module.

From the comparison in Table 2 and 3, the improvement in performance is attributed to two aspects. First, directly combining the training of the synthetic dataset and unlabeled real-world dataset increases the source domain's diversity and scale. Second, the DomainMix loss further forces the network to learn domain-invariant features and minimizes the domain gap between the synthetic dataset and real-world dataset in the source domain.

5. Conclusion

In this paper, a more practical and generalizable person re-ID task is proposed, *i.e*., how to combine a labeled synthetic dataset with unlabeled real-world data to train a more generalizable model. To deal with it, the DomainMix approach is introduced, with which the requirement of human annotations is completely removed, and the gap between synthesized and real-world data is reduced. Extensive experiments show that the proposed annotation-free method is superior for generalizable person re-identification.

References

[1] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. 8

[2] Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M Hospedales. Disjoint label space transfer learning with common factorised space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3288–3295, 2019. 3

[3] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9637–9646, 2019. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[5] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018. 3

[6] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6112–6121, 2019. 3

[7] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 3, 7

[8] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Advances in Neural Information Processing Systems*, 2020. 3, 6, 7, 8

[9] Yixiao Ge, Feng Zhu, Rui Zhao, and Hongsheng Li. Structured domain adaptation for unsupervised person re-identification. *arXiv preprint arXiv:2003.06650*, 2020. 7, 8

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3, 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[13] Jieru Jia, Qiuqi Ruan, and Timothy M. Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. In *British Machine Vision Conference*, 2019. 3

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020. 3, 8

[15] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 6, 7, 8

[16] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7919–7929, 2019. 3

[17] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–178, 2018. 3

[18] Shengcai Liao and Ling Shao. Interpretable and Generalizable Person Re-Identification with Query-Adaptive Convolution and Temporal Lifting. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 8

[19] Shan Lin, Haoliang Li, Chang-Tsun Li, and A. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *British Machine Vision Conference*, 2018. 3

[20] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8738–8745, 2019. 3

[21] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, pages 2597–2609, 2019. 1

[22] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 6, 7

[23] Xuelin Qian, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Leader-based multi-scale attention deep architecture for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, pages 371–385, 2019. 8

[24] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3750–3759, 2019. 1

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 1

[26] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 4

[27] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 8

[28] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3422–3430, 2020. 1, 2, 3, 6, 7

[29] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 1, 3, 4, 6, 7, 8

[30] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12597–12604, 2020. 3

[31] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition*, pages 34–39, 2014. 3

[32] Ye Yuan, Wuyang Chen, Tianlong Chen, Yang Yang, Zhou Ren, Zhangyang Wang, and Gang Hua. Calibrated domain-invariant learning for highly generalizable large scale re-identification. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3589–3598, 2020. 4, 8

[33] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13657–13665, 2020. 3

[34] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9021–9030, 2020. 3

[35] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 1–18, 2020. 3

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 6, 7, 8

[37] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 1

[38] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 6, 7, 8

[39] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, 2018. 3

[40] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019. 8