DomainMix: Learning Generalizable Person Re-Identification Without Human Annotations

Wenhao Wang*[1]
wangwenhao0716@gmail.com
Shengcai Liao†[2]
scliao@ieee.org
Fang Zhao[2]
fang.zhao@inceptioniai.org
Cuicui Kang[3]
cuicui.kang@mbzuai.ac.ae
Ling Shao[2,3]
ling.shao@ieee.org

[1] Baidu Research
Beijing, China

[2] Inception Institute of Artificial Intelligence (IIAI)
Masdar City, Abu Dhabi, UAE

[3] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Masdar City, Abu Dhabi, UAE

Abstract

Existing person re-identification models often have low generalizability, which is mostly due to limited availability of large-scale ***labeled*** data in training. However, labeling large-scale training data is very expensive and time-consuming, while large-scale synthetic dataset shows promising value in learning generalizable person re-identification models. Therefore, in this paper a novel and practical person re-identification task is proposed, i.e. how to use ***labeled*** synthetic dataset and ***unlabeled*** real-world dataset to train a universal model. In this way, human annotations are no longer required, and it is scalable to large and diverse real-world datasets. To address the task, we introduce a framework with high generalizability, namely DomainMix. Specifically, the proposed method firstly clusters the unlabeled real-world images and selects the reliable clusters. During training, to address the large domain gap between two domains, a domain-invariant feature learning method is proposed, which introduces a new loss, i.e. domain balance loss, to conduct an adversarial learning between domain-invariant feature learning and domain discrimination, and meanwhile learns a discriminative 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 person re-identification datasets under direct cross-dataset evaluation.

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 [21, 22, 23, 24, 27, 28, 42] and gained impressive progress. However, significant performance degradation can be observed when a trained model is tested on a previously unseen dataset. The generalizability

*Wenhao Wang finished his part of work during his internship in IIAI.
†Shengcai Liao is the corresponding author.

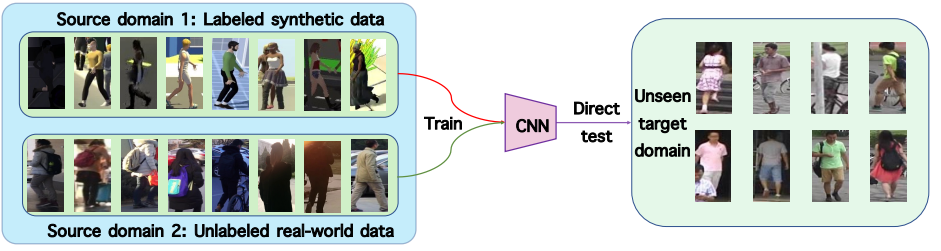

Figure 1: The illustration of the proposed task A+B→C, i.e. how to use labeled synthetic data A and unlabeled real-world data B to train a model that can generalize well to an unseen target domain C.

of known algorithms is hindered by two main aspects. First, the generalizability 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 [32] requires three labelers to work for two months. To address this, RandPerson [31] 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 data alone, the generalizability 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 [31] 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 sub-optimal 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 generalizability. 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 generalizability on C. A task which may seem similar to the proposed one is the semi-supervised learning (SSL). However, for the SSL, both the labeled and unlabeled images are usually from the same domain while in the proposed setting, the images for training are from quite different domains. Besides, that is why we design special method to reduce the domain gap to improve generalizability.

To address this problem, a solution called DomainMix is proposed, for discriminative, domain-invariant, and generalizable person re-identification feature learning. Specifically, to

better utilize unlabeled real-world images, in each given epoch, they are clustered by DBSCAN algorithm. However, because unlike most UDA algorithms, i.e. there is a pre-training on a labeled source dataset, the clustering results may be unreliable and noisy. Therefore, three criteria, i.e. independence, compactness and quantity, are used to select reliable clusters. After clustering in each epoch, the number of identities for training is various. Therefore, it is impossible to use the same classification layer all the time. To address the problem, an adaptive initialization method is utilized: The classification layer can be divided into two parts: one for the synthetic data and the other for the real-world data. The number of identities for the synthetic data part never changes, therefore, it is initialized as the result of the last epoch. However, for the real-world data part, the number of identities changes all the time. As a result, it is initialized as the average of the features of corresponding identity. This initialization method accelerates and guarantees the convergence of training. To deal with the huge domain gap between synthetic and real-world data, a domain-invariant feature learning method is designed. Through alternate training between backbone and discriminator, and with the help of the proposed domain balance loss and other person re-ID metrics, the network can learn discriminative, domain-invariant and generalizable features from two domains jointly. With this framework, the need of human annotations is completely eliminated, and the domain gap between the synthesized and real-world data is reduced, so that the generalizability is improved thanks to the large-scale and diverse training data.

The contributions are summarized as three-fold.

- The paper proposes a novel and practical person re-identification task, i.e. how to combine labeled synthetic dataset with unlabeled real-world dataset to train a model with high generalizability.
- A novel and unified DomainMix framework is proposed to learn discriminative, domain-invariant, and generalizable person re-identification feature from two domains jointly. For the first time, domain generalizable person re-identification can be learned without human annotations completely.
- Experimental results show the proposed annotation-free framework achieves comparable performance with its counterparts trained with full human annotations.

2 Related Work

2.1 Unsupervised Domain Adaptation for Person Re-ID

The goal of Unsupervised Domain Adaptation (UDA) for person re-ID 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 [3, 17, 32, 44], which use a generative adversarial network (GAN) [9] to translate the image style. The second class is feature-level methods [1, 16, 19], which aim to find domain-invariant features between different domains. The last category is cluster-based algorithms [4, 6, 7, 8, 20, 33, 36, 37, 38], which generate pseudo labels to help fine-tune on the target domain.

Although the UDA task and the proposed task both have the source and target domain, they are totally different. The goal of UDA is to use labeled source domain and unlabeled target domain to train a model which can perform well on the known target domain, while the proposed task aims to learn a model from labeled synthetic dataset and unlabeled real-world dataset to generalize well to an unseen domain.

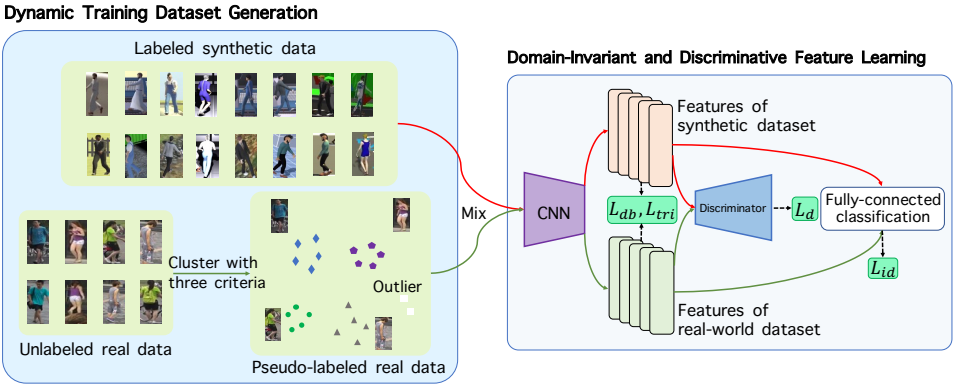

Figure 2: The design of the DomainMix framework. During training, the backbone is trained to extract discriminative, domain-invariant, and generalizable features from two domains jointly with the help of the domain balance loss and other person re-identification metrics.

2.2 Domain Generalization for Person Re-ID

Domain Generalization (DG) for person re-ID was first studied in [34], 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 [12] uses instance normalization to filter out variations in style statistic in earlier layers to increase the generalizability. SNR [13] 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. M^{3}T [39] introduces meta-learning strategy and proposes a memory-based identification loss to enhance the generalization ability of the model. Other works, such as RandPerson [31], focus on using synthetic data to enlarge the diversity and scale of person re-ID datasets.

2.3 Methods for Reducing Domain Gap

Domain gap hinders one trained model performs well on an unseen dataset [14]. In the task of UDA for person re-ID, some methods, such as PTGAN [32], utilize GAN [9] 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) [29]. 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.

3 Proposed Task and 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 $D_{s_1} = \left\{ \left(x_i^{s_1}, y_i^{s_1}\right) \middle|_{i=1}^{N_{s_1}} \right\}$, 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 $D_{s_2} = \left\{ x_i^{s_2} \middle|_{i=1}^{N_{s_2}} \right\}$. Besides, a target domain T, which is a real-world dataset different from D_{s_2}, is given. It is denoted as $D_t = \left\{ x_i^t \middle|_{i=1}^{N_t} \right\}$, 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.

3.2 DomainMix Framework

To tackle the problem mentioned above, we propose the DomainMix framework. In this framework, reliable training dataset is generated dynamically according to three criteria, and before training, the classification layer is initialized adaptively to accelerate the convergence of identity classifier training. When training, together with discriminative metrics, a domain balance loss is proposed to help learning domain-invariant feature. As a result, the proposed DomainMix framework can generalize well to unseen target domains. The framework is shown in Fig. 2.

3.2.1 Two Domains Mixing

Dynamic Training Dataset Generation

The training dataset for DomainMix framework is generated dynamically in each epoch. Given D_{s_2}, the reliable images are selected according to three criteria, i.e. independence, compactness, and quantity.

For independence and compactness, they are proposed in SpCL [8] to judge whether a cluster is far away from others and whether the samples within the same cluster have small inter-samples distances. Together with the *eps* parameter in DBSCAN [5], independence is realized by increasing *eps* to figure out whether more examples are included into original cluster while compactness is realized by decreasing *eps* to find whether a cluster can be split. Please refer to DBSCAN [5] and SpCL [8] to have a deeper understanding about the independence and compactness criteria.

For quantity, we argue that a reliable cluster should contain enough number of images which brings diversity. Further, if clusters with small number of images are selected, there will be too many classes to train an identity classifier well. We denote the pseudo-labels set generated in one epoch as $L_1 = \left\{ l_i |_{i=1}^{M} \right\}$, where l_i is the i-th pseudo label, and M is the total number of pseudo labels. Given the bound b, labels with a total number of images below b are discarded. Thus, the refined pseudo-labels set is obtained as

$$L_2 = \{ l_i \mid l_i \in L_1, S(l_i) \geq b \}, \tag{1}$$

where $S(l_i)$ denotes the number of images belonging to the i-th pseudo label. Note that the quantity criteria is different from the *min_samples* parameter in DBSCAN [5]: the quantity criteria handles the outliers and clusters with few images while *min_samples* parameter controls the core points selection in the process of clustering. Simply adjusting the *min_samples* parameter cannot bring similar improvement with quantity criteria.

After images from D_{s_2} are encoded to features, and features are clustered by a certain algorithm (*e.g.* DBSCAN [5]), the generated clusters are selected by the three criteria. The ablation study part will show the proposed quantity criterion is the key to the outstanding performance while the criteria from [8] only bring slight improvement. In conclusion, the images in reliable clusters are kept, pseudo labeled, and trained with ones from labeled synethetic dataset.

Adaptive Classifier Initialization

Because the training dataset is generated dynamically in each epoch, the number of classes is variant. It is impossible to use the same classification layer in each epoch and random initialization may bring non-convergence problems. As a result, an adaptive classifier initialization method is utilized to accelerate the training of identity classification.

A classification layer can be formed as

$$y = W^T x + \boldsymbol{b}, \tag{2}$$

where x is a batch of features, W is a matrix, and $\boldsymbol{b}$ is a bias which is set as $\mathbf{0}$ for convenience. Given the number of classes M in the generated training dataset and the dim of features d, the shape of matrix W is $d \times M$. Because of the linear properties of matrix, W can be written as (W_1, W_2) in blocks. W_1 is a matrix of shape $d \times N$ and W_2 is a matrix of shape $d \times (M-N)$, where N is the number of classes in synthetic domain.

For W_1, because the classes of synthetic domain never changes during the different epochs, in a new epoch, it is initialized as the final result of the last epoch. For W_2, because clustering and selecting are performed in each epoch, M changes all the time. Denote W_2 as $(w_1, w_2, ..., w_{M-N})$, and w_i is initialized as

$$w_i = \frac{1}{K_i} \sum_{j=1}^{K_i} f_{j_i} \, (i = 1, 2, ..., M-N), \tag{3}$$

where K_i is the number of images belonging to the i-th cluster under the current epoch, and f_{j_i} is the feature of the j-th image in the cluster.

The advantage of this adaptive initialization method lies in two aspects. For the synthetic part, the initialization method enjoys the convenience and stability of fully-supervised learning. For the real-world part, after initialization, the probability of a given feature belongs to its class is much larger than other classes, therefore training the classifier is much easier.

3.2.2 Domain-Invariant and Discriminative Feature Learning

Given the generated training dataset and a well-initialized network, this section focuses on how to learn discriminative, domain-invariant, and generalizable features from two domains. It is realized by training a discriminator and backbone alternately. The discriminator is used to classify a given feature into its domain. Specifically, features of the images from the synthetic and real-world domains are extracted by the backbone. Then a discriminator is trained to judge which domain 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

$$\mathcal{L}_d^s(\theta) = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}_{ce} \left(C_d \left(F \left(x_i^s \mid \theta \right) \right), d_i^s \right), \tag{4}$$

where $F(\cdot|\theta)$ is a feature encoder function, N_s is the sum of the number of images in the current generated dataset, 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 domain, $d_i^s = 0$, otherwise if it belongs to the real-world domain, $d_i^s = 1$.

To encourage the backbone to extract domain-invariant features, it is trained to confuse the domain discriminator. Therefore, a domain balance loss is proposed, which is defined as

$$\mathcal{L}_{db} = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\sum_{j=1}^{n}\left(x_j^i \log\left(x_j^i\right)+a\right)\right), \tag{5}$$

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

$$f(x) = x\log(x)+a, x \in (0,1), \tag{6}$$

the second derivative of f is

$$f''(x) = \frac{1}{x} > 0. \tag{7}$$

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 distance between x_j^i and $1/n$ is shortened. Thus, the probability of a given feature belonging to two domains tends to be the same, i.e. the backbone can extract domain-invariant features by confusing the discriminator.

Beyond learning domain-invariant features, the network is also trained by discriminative metrics in re-ID, therefore an identity classification loss $L_{id}^s(\theta)$ and a triplet loss $L_{tri}^s(\theta)$ [11] are adopted. They are defined as

$$\mathcal{L}_{id}^s(\theta) = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}_{ce}\left(C_s\left(F\left(x_i^s \mid \theta\right)\right), y_i^s\right), \tag{8}$$

and

$$\begin{aligned}\mathcal{L}_{tri}^s(\theta) = \tfrac{1}{N_s}\textstyle\sum_{i=1}^{N_s}\max\Big(0, m + \Big\|F\left(x_i^s \mid \theta\right) - F\left(x_{i,p}^s \mid \theta\right)\Big\| \\ - \Big\|F\left(x_i^s \mid \theta\right) - F\left(x_{i,n}^s \mid \theta\right)\Big\|\Big),\end{aligned} \tag{9}$$

where C_s is an identity classifier, $\|\cdot\|$ 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 generated 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.

Therefore, the final loss is calculated as

$$\mathcal{L}^s(\theta) = \lambda^m \mathcal{L}_{db}(\theta) + \lambda^s \mathcal{L}_{id}^s(\theta) + \mathcal{L}_{tri}^s(\theta), \tag{10}$$

where λ^m and λ^s are the balance parameters. Through alternate training with $\mathcal{L}_d^s(\theta)$ and $\mathcal{L}^s(\theta)$, 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 supplemental material.

Table 1: Ablation studies for each component in the DomainMix framework on the two tasks. '+I/C/Q' denotes the independence/compactness/quantity criteria is used. With or without ACI/DB denotes whether using adaptive classifier initialization/domain balance loss or not. 'Labeled' or 'unlabeled' denotes whether real-world source training data is labeled or not.

RP+MSMT → Market	mAP	rank-1	RP+CUHK → Market	mAP	rank-1
DBSCAN	37.5	64.6	DBSCAN	34.5	61.3
DBSCAN + I + C	37.0	64.2	DBSCAN + I + C	35.5	62.8
DBSCAN + Q	42.4	69.4	DBSCAN + Q	39.5	66.2
DBSCAN + I + C + Q	43.5	70.2	DBSCAN + I + C + Q	39.8	67.5
Without ACI	29.5	56.9	Without ACI	33.8	60.3
With ACI	43.5	70.2	With ACI	39.8	67.5
Without DB	40.1	68.1	Without DB	37.3	66.0
With DB	43.5	70.2	With DB	39.8	67.5
Only RandPerson	36.5	63.6	Only RandPerson	36.5	63.6
Only MSMT (labeled)	32.7	62.0	Only CUHK (labeled)	25.1	50.3
DomainMix (labeled)	45.2	70.5	DomainMix (labeled)	42.7	69.7
DomainMix (unlabeled)	43.5	70.2	DomainMix (unlabeled)	39.8	67.5

4 Experiments

4.1 Datasets and Evaluation Metrics

To evaluate the generalizability of the proposed DomainMix framework, extensive experiments are conducted on four widely used public person re-ID datasets. Among them, RandPerson (RP) [31] 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 [40], CUHK03-NP [15, 43], and MSMT17 [32]. Note that DukeMTMC [41] dataset is not used due to the invasion of privacy. The details of real-world dataset are illustrated in the supplemental material. Evaluation metrics are mean average precision (mAP) and cumulative matching characteristic (CMC) at rank-1.

4.2 Implementation Details

DomainMix is trained on four Tesla-V100 GPUs. The ImageNet-pre-trained [2] ResNet-50 [10] and IBN-ResNet-50 [25] are adopted as the backbone. Adam optimizer is used to optimize the networks with a weight decay of 5×10^{-4}. For more details, please refer to supplemental results.

4.3 Ablation Study

Comprehensive ablation studies are performed to prove the effectiveness of each component in the DomainMix framework. Two different DG tasks are selected: labeled RandPerson [31] with unlabeled MSMT17 [32] to Market-1501 [40] and labeled RandPerson [31] with unlabeled CUHK03-NP [15, 43] to Market-1501 [40]. The experimental results on ResNet-50 [10] are reported below, and the results on IBN-ResNet-50 [25] are shown in supplemental results.

Effectiveness of Dynamic Training Dataset Generation. To investigate the necessity of generating training dataset dynamically and the importance of each component, we com-

pare the domain generalizability of a model trained on two different real-world datasets, i.e. MSMT17 [32] and CUHK03-NP [15, 43]. The baseline model performances are shown in Table 1 as "DBSCAN". If the independence and compactness criteria are used, the performances are denoted as "DBSCAN + I + C", while if the quantity criterion is used, they are denoted as "DBSCAN + Q". "DBSCAN + I + C + Q" denotes all the three criteria are adopted. The quantity criterion brings 4.9% in mAP improvement for the "RP+MSMT → Market" task. For the "RP+CUHK → Market" task, the mAP increases by 5.0%. However, if the independence and compactness criterion are used alone, no stable performance improvement can be observed. It is because, although the two criteria remove the unreliable clusters, there are still many classes including few images to participate in the training, which disturbs the training of the identity classifier and leads to the failure to improve the performance stably. Together with the proposed quantity criterion, the above problem is solved, and the two criteria in [8] can further improve the performance.

Effectiveness of Adaptive Classifier Initialization. To prove the effectiveness of the adaptive classifier initialization method, the experimental results without and with this method are shown in Table 1 and denote as "Without ACI" and "With ACI", respectively. The initialization method brings significant improvement of 14.0% and 6.0% in mAP on the "RP+MSMT → Market" and "RP+CUHK → Market" tasks. The significant improvement comes from the guarantee and acceleration of the convergence.

Influence of Domain Balance Loss. To verify the necessity of using the domain balance loss to learn domain-invariant features, results obtained with and without this loss are compared and shown in Table 1 as "Without DB" and "With DB", respectively. All experiments with the use of domain balance loss show distinct improvement on both the "RP+MSMT → Market" and "RP+CUHK → Market" tasks. Specifically, the mAP increases by 3.4% when the real-world source domain is MSMT17 [32]. As for the 'RP+ CUHK → Market' task, similar mAP improvement of 2.5% can be observed. The improvement brought by domain balance loss on CUHK is displayed in the supplemental material.

We further discuss the **importance of introducing unlabeled real-world dataset**, **whether human annotations are essential for generalizable person re-ID**, and the **comparison with UDA algorithms** in the supplemental material.

4.4 Comparison with the State-of-the-arts

The proposed DomainMix framework is compared with state of the art methods on three DG tasks, i.e. directly testing on Market1501 [40], CUHK03-NP [15, 43], and MSMT17 [32]. The experimental results are shown in Table 2. Note that a fair comparison in anyway is not very feasible, because we only used unlabeled real-world data, although with additional synthesized data, while others used labeled one. So existing results in Table 2 are only provided as a reference to see what we can achieve with a fully annotation-free setting. Secondly, the proposed method is orthogonal to network architecture designs such as IBN-Net [25] and OSNet-IBN [45]. Thus they can also be applied into the framework. The related experimental results are shown in the supplemental material. 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. SNR [13] uses attention mechanism to solve the drawback of instance normalization and improve the performance of IBN-Net [25], and the DomainMix may achieve further performance improvement with the help of this plug-and-play module.

From the comparison in Table 2, the DomainMix framework improves up to 7.0% mAP.

Table 2: Comparison with state-of-the-arts on Market1501 [40], CUHK03-NP [15, 43], and MSMT17 [15, 43]. '§' denotes the results are from github of the original paper, '*' denotes our implementation, and '†' indicates that the results are reproduced based on the authors' codes. 'L' or 'U' denotes whether the real-world source training data is labeled or not.

Method	Source data	Market1501	
		mAP	rank-1
MGN [30, 35]	MSMT (L)	25.1	48.7
ADIN [35]	MSMT (L)	22.5	50.1
ADIN-Dual [35]	MSMT (L)	30.3	59.1
SNR [13]	MSMT (L)	41.4	70.1
QAConv† [18]	MSMT (L)	35.8	66.9
MGN† [30]	RandPerson	17.7	37.4
OSNet-IBN† [45]	RandPerson	39.0	67.0
QAConv§ [18]	RandPerson	34.8	65.6
Baseline*	RandPerson	36.5	63.6
DomainMix	RP+MSMT (U)	43.5	70.2
DomainMix-OSNet-IBN	RP+MSMT (U)	**44.6**	**72.9**
Method	Source data	CUHK03-NP	
		mAP	rank-1
MGN [26, 30]	Market (L)	7.4	8.5
MuDeep [26]	Market (L)	9.1	10.3
QAConv† [18]	MSMT (L)	15.2	16.8
MGN† [30]	RandPerson	7.7	7.4
OSNet-IBN† [45]	RandPerson	12.9	13.6
QAConv§ [18]	RandPerson	11.0	14.3
Baseline*	RandPerson	13.0	14.6
DomainMix	RP+MSMT (U)	16.7	**18.0**
DomainMix-OSNet-IBN	RP+MSMT (U)	**16.9**	17.5
Method	Source data	MSMT17	
		mAP	rank-1
QAConv† [18]	Market (L)	8.3	26.4
MGN† [30]	RandPerson	3.0	10.1
OSNet-IBN† [45]	RandPerson	12.4	34.3
QAConv§ [18]	RandPerson	10.7	34.3
Baseline*	RandPerson	7.9	23.0
DomainMix	RP+Market (U)	9.3	25.3
DomainMix-OSNet-IBN	RP+Market (U)	**13.6**	**36.2**

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 domain balance 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 framework 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-ID.

Acknowledgements

The authors would like to thank Anna Hennig who helped proofreading the paper.

References

[1] 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.

[2] 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.

[3] 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.

[4] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1):1–19, 2020.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[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.

[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.

[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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9757–9766, June 2021.

[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.

[16] 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.

[17] 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.

[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.

[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.

[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.

[21] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020.

[22] Jialun Liu, Jingwei Zhang, Wenhui Li, Chi Zhang, and Yifan Sun. Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory. *arXiv preprint arXiv:2008.09809*, 2020.

[23] 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.

[24] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 542–551, 2019.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6277–6286, 2021.

[40] 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.

[41] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[42] 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.

[43] 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.

[44] 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.

[45] 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.