Dual Distribution Alignment Network for Generalizable Person Re-Identification

Peixian Chen[1], Pingyang Dai[1]*, Jianzhuang Liu[3], Feng Zheng[2], Mingliang Xu[4], Qi Tian[5], Rongrong Ji[1,6]

[1] Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University
[2] Department of Computer Science and Engineering, Southern University of Science and Technology
[3] Noah's Ark Lab, Huawei Tech
[4] School of Information Engineering, Zhengzhou University
[5] Cloud & AI, Huawei Tech.
[6] Institute of Artificial Intelligence, Xiamen University
pxchen@stu.xmu.edu.cn, {pydai, rrji}@xmu.edu.cn, {liu.jianzhuang, tian.qi1}@huawei.com,
zhengf@sustech.edu.cn, iexumingliang@zzu.edu.cn

Abstract

Domain generalization (DG) offers a preferable real-world setting for Person Re-Identification (Re-ID), which trains a model using multiple source domain datasets and expects it to perform well in an unseen target domain without any model updating. Unfortunately, most DG approaches are designed explicitly for classification tasks, which fundamentally differs from the retrieval task Re-ID. Moreover, existing applications of DG in Re-ID cannot correctly handle the massive variation among Re-ID datasets. In this paper, we identify two fundamental challenges in DG for Person Re-ID: domain-wise variations and identity-wise similarities. To this end, we propose an end-to-end Dual Distribution Alignment Network (DDAN) to learn domain-invariant features with dual-level constraints: the domain-wise adversarial feature learning and the identity-wise similarity enhancement. These constraints effectively reduce the domain-shift among multiple source domains further while agreeing to real-world scenarios. We evaluate our method in a large-scale DG Re-ID benchmark and compare it with various cutting-edge DG approaches. Quantitative results show that DDAN achieves state-of-the-art performance.

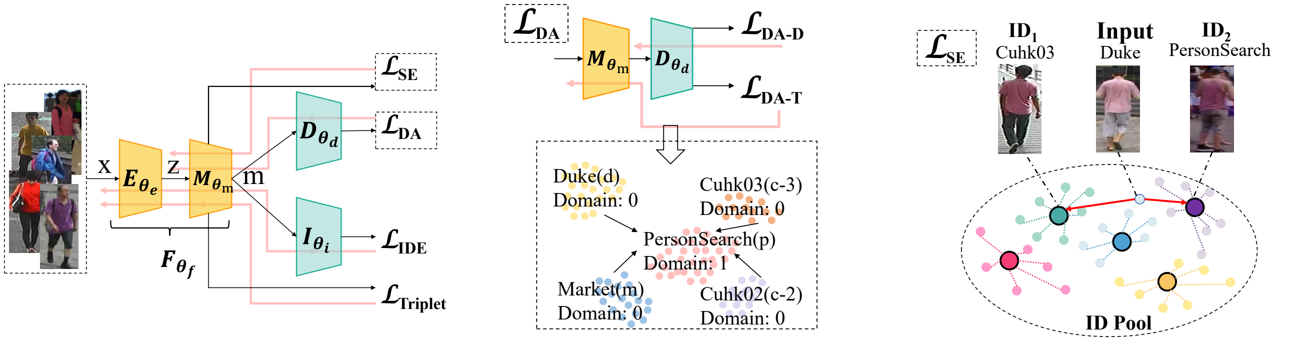
1 Introduction

Person Re-Identification (Re-ID) aims to identify the same pedestrian captured by different cameras under variant viewpoints, lighting, and locations. To date, approaches based on deep Convolution Neural Networks (CNNs) have achieved remarkable performance in this topic (Bak and Carr 2017; Bai, Bai, and Tian 2017; Li, Zhu, and Gong 2017). Unfortunately, these approaches' success heavily relies on the *i.i.d.* assumption between labeled training and test data. In real-world applications, meeting this supervised *i.i.d.* assumption is prohibitively expensive, as it requires low-variation datasets and extensive manual labeling. To this end, Unsupervised Domain Adaptation (UDA) was introduced in person Re-ID (Wang et al. 2018; Lin et al. 2018; Bak et al. 2018). These approaches typically learn a model by mixing data from both the labeled source domain and the unlabeled target domain, hence fitting the target domain's distribution without the cost of labeling. Though being more practical than supervised ones, existing endeavors have shown that UDA approaches require training with many target data to achieve good results (Peng et al. 2016; Deng et al. 2018; Zhong et al. 2019; Fu et al. 2019).

Compared with UDA, Domain Generalization (DG) offers a preferable real-world setting, which learns a model using multiple source domain datasets and expects this model to perform well in an unseen target domain without any model updating, *i.e.*, adaptation or retraining. An optimal DG should learn a feature representation that is discriminative for the task and insensitive to the variation of domain distributions. Therefore, DG is arguably preferable as it does not need any data from the target domain. Unfortunately, existing DG methods are designed explicitly for classification tasks (Li et al. 2018a,b), while Re-ID is a retrieval task, which fundamentally differs. To explain, DG in classification tackles the same set of labels in all source domains. In contrast, DG in retrieval should compare the feature similarity between different identities (corresponding to different labels between source and target domains).

To this end, some works attempted to apply DG in person Re-ID by employing existing techniques, such as meta-learning (Song et al. 2019) and normalization layers (Jia et al. 2019), to learn discriminative features that are insensitive to domain variations. However, these approaches are still not specific to person Re-ID. The variation among Re-ID datasets is usually massive, so directly applying classification DG methods with pair-wise domain alignment can hurt the performance, as there is a statistical trade-off between domain-invariance and classification accuracy (Akuzawa, Iwasawa, and Matsuo 2019). Moreover, they naively mix all source data without associating similar samples from different domains: two pedestrian features maybe even closer than visually-similar ones solely because the former ones are from the same domain. It contradicts real-world scenarios, as visually-similar pedestrians are more

*Corresponding author



(a) The DDAN Architecture (b) Domain-wise Adversarial Learning (c) Identity-wise Similarity Enhancement

Figure 1: An Overview of DDAN. Left: Network Architecture (red arrows denote gradients). Middle: Domain-wise Adversarial Feature Learning, where peripheral domains align to the central domain (detailed in Sec. 3.2). Right: Identity-wise Similarity Enhancement, where an ID pool is used to enhance the inter-domain similarity of feature distributions (detailed in Sec. 3.3). Big dots denote the summarized representation of each ID, to which the similarity of new inputs should be selectively enhanced.

likely to be (incorrectly) identified as the same person.

We have identified two fundamental challenges in DG for person Re-ID from the above observations: *domain-wise variations* and *identity-wise similarities*. To address these two challenges, we propose an end-to-end Dual Distribution Alignment Network (DDAN) with dual-level constraints, which reduces the domain-shift among multiple source domains further while agreeing to real-world scenarios.

As shown in Fig. 1(a), DDAN consists of an encoder E, a feature mapping network M, and four objective modules. In addition to regular Re-ID constraints, such as Identity-Discriminative Embedding (IDE) (Zheng, Yang, and Hauptmann 2016) and Triplet losses, our main contribution originates from the dual-level constraints. At the domain level, we propose a novel adversarial feature learning scheme to align feature distributions of different domains. In particular, we selectively reduce the discrepancy between a *central* domain and each of the *peripheral* domains as shown in Fig. 1(b). We formalize these two kinds of domains using Wasserstein distance, thus minimizing the required distributional shift for alignment (detailed in Sec. 3.2). At the identity level, we enhance the *inter-domain* similarity of visually-similar samples from different domains. In Fig. 1(c), we associate each sample with its top-k similar samples from other domains and reduce their distances to enforce the domain-invariance.

To summarize, the proposed Dual Distribution Alignment Network (DDAN) innovates in the following three aspects:

1. We propose a novel domain-wise adversarial feature learning scheme. It aligns domains with minimal distributional shifts to mitigate the loss of accuracy.
2. We introduce an identity-wise similarity enhancement. It encourages visually-similar features from different domains to be closer than those from the same domain but with fewer visual similarities.
3. We evaluate our method in a large-scale DG Re-ID benchmark (Song et al. 2019) and compare it with various alternative, cutting-edge DG approaches (Song et al. 2019; Jia et al. 2019). Quantitative results show that DDAN achieves state-of-the-art performance, where the rank-1 accuracy on VIPeR, PRID, GRID, and i-LIDS is 56.5%, 62.9%, 50.6%, and 78.5%, respectively.

2 Related Work

2.1 Person Re-Identification (Re-ID)

Existing supervised person Re-ID methods typically learn a distance metric (Köstinger et al. 2012; Xiong et al. 2014; Zheng, Gong, and Xiang 2013), a subspace (Wang et al. 2016; Chen et al. 2018), or view-invariant discriminative features (Zheng et al. 2015; Liao et al. 2015; Gray and Tao 2008). Along with the success of deep CNNs, person Re-ID methods have achieved remarkable performance under the *i.i.d.* assumption between training and test data (Cheng et al. 2016; Paisitkriangkrai et al. 2015; Matsukawa et al. 2016; Chen et al. 2017; Bak and Carr 2017; Bai, Bai, and Tian 2017; Li, Zhu, and Gong 2017). However, the learned models generally overfit to the training data and perform poorly on new unseen datasets.

To this end, UDA was introduced in person Re-ID (Peng et al. 2016; Lin et al. 2018; Wang et al. 2018; Bak et al. 2018). Such a setting typically involves a labeled source domain to help the model fit an unlabeled target domain's distribution. For instance, an asymmetric multi-task dictionary learning scheme was proposed to learn discriminative representations for the target domain (Peng et al. 2016). Another group of UDA-based person Re-ID methods exploited generative adversarial networks (GANs). They employed CycleGAN (Zhu et al. 2017) to translate images from the source domain to the target one (Deng et al. 2018) or generated images with different camera styles in the target domain to enforce invariance (Zhong et al. 2018). Recent works attempted to improve domain alignment performance. For instance, a transferable model was proposed to learn attribute-identity discriminative representation for the target domain jointly (Wang et al. 2018). Some works also investigated the clustering of samples in the target domain for similarity measurement (Zhong et al. 2019), which was later extended to finer-grained with pedestrian's part-level features (Fu et al. 2019). However, all the above UDA approaches generally require a large amount of target data to avoid overfitting and achieve satisfactory results.

2.2 Domain Generalization (DG)

DG aims to learn a generalizable model, which tries to remove the domain-shift without needing the target domain's data during training. In this regard, many DG methods have been proposed for classification. For example, learning a model for each source domain and select the best one for each target domain in the test phase (Xu et al. 2014).

There exist more efficient choices, which learn a model to extract task-specific and domain-invariant features. These methods typically involved advanced ML techniques, such as kennel-based optimization (Muandet, Balduzzi, and Schölkopf 2013), multi-task auto-encoder (Ghifary et al. 2015), domain-distance regularization based on canonical correlation analysis (Yang and Gao 2013), and model-agnostic meta-learning (Li et al. 2019; Dou et al. 2019).

Another group of works adopted adversarial feature alignment to learn a domain-invariant model by reducing pairwise domain discrepancy (Li et al. 2018a,b). However, there is a statistical trade-off between domain-invariance and classification accuracy (Akuzawa, Iwasawa, and Matsuo 2019). Since the variation among Re-ID datasets is usually massive, the pairwise domain alignment can cause extensive reductions in feature discriminability.

To handle DG in person Re-ID, DIMN (Song et al. 2019) adopted meta-learning to learn a model from gallery images, which outputs matching scores between gallery and probe image features. But this meta-learning scheme can increase optimization complexity and significantly decrease the test speed correspondingly. As a more efficient solution, Dual-Norm (Jia et al. 2019) employs both instance and batch normalization to improve the feature extractor's performance. Unfortunately, these methods naively mix all source domains without associating similar samples from different domains. As such, two pedestrian features maybe even closer than visually-similar ones solely because the former ones are from the same domain.

In this paper, we have identified two fundamental challenges in DG for person Re-ID: *domain-wise variations* and *identity-wise similarities*. Therefore, our DDAN is distinguished from the above methods in two folds. At the domain level, instead of adopting the pairwise alignment, we selectively reduce the discrepancy between a *central* domain and other *peripheral* domains, thus minimizing the loss of distinguishability. At the identity level, instead of mixing all source domain data directly, we also recognize the inter-domain similarities between identities in different domains. We have quantitatively proven its effectiveness in improving the feature discriminability.

3 The Proposed Method

For DG in person Re-ID, we have access to M datasets (source domains) with non-overlapping labels in the training phase, as we assume disjoint IDs among different datasets. In the test phase, we "freeze" and apply the trained model directly to a new unseen dataset (target domain) without further model updating or retraining. Let $X_s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ denote inputs from the source domain s, where N_s is the number of labeled data in domain s. As shown in Fig. 1, the encoder $E(\boldsymbol{x};\theta_e)$ parameterized by θ_e maps an image $\boldsymbol{x}$ to a feature map $\boldsymbol{z}$. $M(\boldsymbol{z};\theta_m)$ is a mapping network parameterized by θ_m, which maps $\boldsymbol{z}$ from different distributions to the one in a shared feature space (denoted by $\boldsymbol{m}$). We denote the feature extractor as the composition of the encoder and mapping network, that is, $F = M \circ E$, which is parameterized by $\theta_f = \{\theta_m, \theta_e\}$. We then use a domain discriminator $D(\boldsymbol{m};\theta_d)$ parameterized by θ_d to distinguish the domain to which the inputs belong, and an identity discriminator $I(\boldsymbol{m};\theta_i)$ parameterized by θ_i to increase the effectiveness of learned features. For simplicity, we only write parameters that are updated through back-propagation on the left-hand side of equations in the following of this section.

3.1 Baseline Configuration

We choose the standard aggregation DG method as our baseline, which learns a model with all source domains. In person Re-ID with labeled training data, an effective strategy is to learn identity-discriminative embeddings (IDE) (Zheng, Yang, and Hauptmann 2016). It converts Re-ID into classification by cross-entropy loss J, as shown below.

$$\mathcal{L}_{\text{ide}}(X;\theta_f,\theta_i) = \frac{1}{n_{bs}}\sum_{n=1}^{n_{bs}} J\Big(I\big(F(\boldsymbol{x}_n;\theta_f);\theta_i\big), y_n\Big), \quad (1)$$

where n_{bs} denotes the number of samples in a mini-batch $X = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{n_{bs}}$. After training, the feature extractor F_{θ_f} is used to extract features from input images.

As another common and useful criterion for similarity learning in person Re-ID, the triplet loss $\mathcal{L}_{\text{tri}}$ can reduce intra-class distance while extending inter-class distances (Hermans, Beyer, and Leibe 2017), as shown below.

$$\mathcal{L}_{\text{tri}}(X;\theta_f) = \sum_{\boldsymbol{x}_a \in F(X)} (d(\boldsymbol{x}_a,\boldsymbol{x}_p) - d(\boldsymbol{x}_a,\boldsymbol{x}_n) + m), \quad (2)$$

where d denotes the Euclidean distance, m is the margin, $\boldsymbol{x}_a$ denotes the anchor point of its hardest positive $\boldsymbol{x}_p$ and negative $\boldsymbol{x}_n$ samples within a mini-batch. In other words, $\boldsymbol{x}_p$ is the farthest sample with the same label as $\boldsymbol{x}_a$, and $\boldsymbol{x}_n$ is the nearest sample with a different label as $\boldsymbol{x}_a$.

3.2 Domain-wise Adversarial Feature Learning

We adopt a novel adversarial learning scheme to tackle the first challenge we have observed in DG for person Re-ID: *domain-wise variations*. Specifically, we reduce the overall discrepancy of all source domains to learn a mapping from inputs to a domain-invariant feature space. Notably, we shift various domains' distributions towards a uniform one. Towards an important early exploration, most methods reduce the overall discrepancy by aligning each pair of source domains (Li et al. 2018a,b). However, there is a statistical trade-off between domain-invariance and classification accuracy (Akuzawa, Iwasawa, and Matsuo 2019). Since the variation among Re-ID datasets is usually massive, the pair-wise alignment can spoil learned features, as they have to be heavily shifted to align outlying domains. To explain, some domains in person Re-ID may exceptionally differ from the other domains. In such a case, the pair-wise alignment can introduce unnecessary distributional shifts towards the outlying domain and decrease accuracy.

To overcome this problem, we choose the most "generalizable" domain as the *central* domain and refer the remaining as the *peripheral* domains. Here, the central domain should have a similar distribution to most peripheral domains. In other words, a central domain is the one that minimizes the distributional shift needed for aligning the other peripheral domains to it. We then determine an efficient way to align source domains with minimum negative impacts: Instead of pair-wisely aligning every two domains, we only align the peripheral domains to the central one. Such a setting generalizes better while minimizing the negative impact, as discussed above. We now formally explain our definition of the central and peripheral domains as follows.

We employ Wasserstein distance d_{WS} (Arjovsky, Chintala, and Bottou 2017) and Sinkhorn's approximation (Cuturi 2013) to quantify the needed distributional shift for aligning domains, as d_{WS} is symmetric and supports non-overlapping distributions. The central domain is defined as

$$c^* = \operatorname*{arg\,min}_{c \in \mathcal{S}} \sum_{i \in \mathcal{S} \setminus \{c\}} d_{\mathrm{WS}}\big(X_c, X_i\big), \tag{3}$$

where $\mathcal{S}$ is the set of all source domains, X_c and X_i are the samples from domains c and i, respectively. In this way, the remaining peripheral domains are determined accordingly and aligned with minimized overall distributional shifts. Additionally, we define the "domain label" of each domain by checking if it is central (1) or peripheral (0). The chosen central domain will be specified in Sec.4.3.

Once determining the central and peripheral domains, our adversarial feature learning scheme involves a generator and discriminator pair. We regard the mapping network M_{θ_m} as the generator, which maps the peripheral domains' distributions to a uniform one similar to the central domain. We learn the discriminator D_{θ_d} by minimizing the following cross-entropy loss to correctly distinguish whether an example belongs to the central or peripheral domains.

$$\mathcal{L}_{\mathrm{DA-D}}(X; \theta_d) = \frac{1}{n_{bs}} \sum_{n=1}^{n_{bs}} J(D(F(\boldsymbol{x}_n; \theta_f); \theta_d), c_n), \tag{4}$$

where $X = \{(\boldsymbol{x}_n, c_n)\}_{n=1}^{n_{bs}}$ is the input mini-batch with domain labels. The mapping network M_{θ_m} is trained to fool the discriminator D_{θ_d} by generating domain-invariant features. It can be achieved by minimizing the negative entropy of the predicted domain distributions *w.r.t.* θ_m as

$$\mathcal{L}_{\mathrm{DA-T}}(X; \theta_m) = -\frac{1}{n_{bs}} \sum_{n=1}^{n_{bs}} \log D\Big(F\big(\boldsymbol{x}_n; \theta_f\big); \theta_d\Big). \tag{5}$$

The above learning process adversarially aligns the peripheral domains' feature distributions to a uniform one similar to the central domain. Unlike previous methods that pair-wisely align domains, our alignment conducts in a specific way from peripheral domains to the central domain. Thus, our method minimizes negative impacts on learned features, especially for commonly outlied person Re-ID domains.

3.3 Identity-wise Similarity Enhancement

The above domain-wise adversarial learning scheme effectively aligns domains with minimal distributional shifts, but cannot tackle the second challenge we have observed: *identity-wise similarities*. Therefore, we derive another constraint from a real-world scenario: Two visually-similar pedestrians are more likely to be (incorrectly) identified as the same person. In other words, features of these two pedestrians should be closer than less similar ones, even if these two features are from different domains, thus forcing the identity discriminator I to distinguish them.

To this end, we accumulate the learned knowledge with an ID pool. We store all IDs' representations and enhance the distributional similarity between newly incoming examples and their visually-similar IDs in other domains. We summarize the representation of each ID μ by computing its running mean representation $\bar{\boldsymbol{r}}_\mu$ in an iterative fashion as

$$\bar{\boldsymbol{r}}_\mu^{(e,t+1)} = \frac{1}{t+1}\Big(t \cdot \bar{\boldsymbol{r}}_\mu^{(e,t)} + F(\boldsymbol{x}; \theta_f)\Big), \tag{6}$$

where $\boldsymbol{x}$ is an input image of ID μ, and the superscript (e, t) denotes the t-th update of $\bar{\boldsymbol{r}}_\mu$ in the e-th epoch. We further accumulate this mean representation over epochs, leading to the final effective representation $\hat{\boldsymbol{r}}_\mu$ as:

$$\hat{\boldsymbol{r}}_\mu^{(e+1,t)} = \alpha \cdot \hat{\boldsymbol{r}}_\mu^{(e,\backslash)} + (1-\alpha) \cdot \bar{\boldsymbol{r}}_\mu^{(e+1,t)}, \tag{7}$$

where $\hat{\boldsymbol{r}}_\mu^{(e,\backslash)}$ denotes the final mean representation obtained in epoch e and $\alpha \in [0,1]$ controls the updating rate. All the variables in Eqs. (6) and (7) are initialized to zero.

The mean representation $\hat{\boldsymbol{r}}_\mu$ conveys how samples of a particular ID are generally represented. As such, we associate the representation of each incoming instance with its similar IDs (from different domains) in the ID pool. Since these paired representations are from different domains with exceptionally unmatched entries due to domain variations, we cannot directly make them close in some distance metrics (*e.g.*, ℓ_1 and ℓ_2). Therefore, we normalize these features with softmax and minimize their symmetric KL-divergence.

Specifically, given an image $\boldsymbol{x}_n$ from the peripheral domain, we search for its top-k similar IDs in other domains with cosine similarity. For $\boldsymbol{x}_n$ from the central domain, we search in the same domain instead to stabilize the central domain's distribution. In short, we align the domains in a particular way from peripheral ones to the central one, instead of pair-wisely or the other way around. Then, we minimize

$$\begin{aligned}\mathcal{L}_{\mathrm{SE}}(X; \theta_m) = \sum_{n=1}^{n_{bs}} \Bigg[\frac{1}{k}\sum_{i=1}^{k}\Big[\ell_{\mathrm{KL}}\Big(\mathrm{sm}\big(F(\boldsymbol{x}_n; \theta_f)\big)\Big\|\mathrm{sm}\big(\hat{\boldsymbol{r}}_i\big)\Big) \\ +\ell_{\mathrm{KL}}\Big(\mathrm{sm}\big(\hat{\boldsymbol{r}}_i\big)\Big\|\mathrm{sm}\big(F(\boldsymbol{x}_n; \theta_f)\big)\Big)\Big]\Bigg],\end{aligned} \tag{8}$$

where $\ell_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_r p_r \log(p_r/q_r)$ is the KL-divergence and $\mathrm{sm}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{x}/\tau)$ is the softmax function at temperature τ. This setting eliminates identity-wise domain-shifts further, thus forcing the network to learn discriminative domain-invariant features.

3.4 Overall Objective Function

To tackle the two fundamental challenges we have observed in DG for person Re-ID, our DDAN consists of three components: a baseline model (Sec. 3.1), a novel domain-wise adversarial learning scheme (Sec. 3.2), and an identity-wise similarity enhancement (Sec. 3.3). The overall loss function in a training mini-batch X is thus defined as the sum of:

$$\begin{aligned}\mathcal{L}_1(X;\theta_f,\theta_i) &= \mathcal{L}_{\text{ide}}(X;\theta_f,\theta_i) + \lambda_1 \cdot \mathcal{L}_{\text{tri}}(X;\theta_f),\\ \mathcal{L}_2(X;\theta_m) &= \lambda_2 \cdot \mathcal{L}_{\text{DA}-\text{T}}(X;\theta_m) + \lambda_3 \cdot \mathcal{L}_{\text{SE}}(X;\theta_m),\\ \mathcal{L}_3(X;\theta_d) &= \mathcal{L}_{\text{DA}-\text{D}}(X;\theta_d),\end{aligned} \tag{9}$$

where λ_1, λ_2 and λ_3 are the trade-off parameters. For simplicity, we only write parameters that will be updated through back-propagation on both sides of Eq. (9). To summarize, we set the above loss functions to embed input images into a discriminative domain-invariant feature space, in which our model generalizes better to new unseen domains.

4 Experiments

4.1 Datasets and Settings

Datasets. We conduct experiments on the large-scale DG Re-ID benchmark (Song et al. 2019) to evaluate our DG model for person Re-ID. Specifically, CUHK02 (Li and Wang 2013), CUHK03 (Li et al. 2014), Market-1501 (Zheng et al. 2015), DukeMTMC-ReID (Zheng, Zheng, and Yang 2017) and CUHK-SYSU PersonSearch (Xiao et al. 2016) are taken as the source datasets. All images in these source datasets, regardless of their train/test splits, are used for training, in total 121,765 images of 18,530 identities. The datasets VIPeR (Gray and Tao 2008), PRID (Hirzer et al. 2011), GRID (Loy, Xiang, and Gong 2010), and i-LIDS (Zheng, Gong, and Xiang 2009) are used as the target datasets for testing, in which we follow the single-shot setting with the numbers of probe/gallery images set to: $316/316$ on VIPeR, $100/649$ on PRID, $125/900$ on GRID, and $60/60$ on i-LIDS.

Settings. We implement our model with PyTorch and train it on a single 1080-Ti GPU. The MobileNetV2 (Sandler et al. 2018) with a width multiplier of 1.0 is used as the backbone network for feature extractor F_{θ_f} with ImageNet-pretrained weights. Note that the mapping network M_{θ_m} is actually the last convolution layer of MobileNetV2. The learning rate is initially set to 0.1 and multiplied by 0.1 per 40 epochs. Our domain discriminator D_{θ_d} consists of a 128-D and a 2-D fully connected (FC) layers with batch normalization (BN), while the identity discriminator I_{θ_i} is a 18,530-D (*i.e.*, the number of identities) FC layer with BN. The updating rate α in Eq. (7) is set to 0.05. The triplet loss margin in Eq. (2) is 0.3. The τ of softmax in Eq. (8) is 2×10^{-3}. The weights of the losses in Eq. (9) are set to $\lambda_1 = 1.0$, $\lambda_2 = 0.18$ and $\lambda_3 = 0.05$. The model is trained for 100 epochs with a batch size of 64 (each identity comes with 4 images). We enable $\mathcal{L}_{\text{SE}}$ after the 4th epoch to stabilize learned representations. Test results are averaged over 10 random probe/gallery splits.

4.2 Comparison with State-of-the-Arts

As shown in Tab. 1, we compare DDAN with other methods on VIPeR, PRID, GRID and i-LIDS. The compared methods include 7 supervised training (S), 3 unsupervised domain adpatation (U), and 2 domain generalization (DG) ones.

Although many supervised methods have achieved high performance on large-scale datasets, like CUHK03, Market-1501 or DukeMTMC-ReID, their performance is unfortunately not satisfied on small-scale ones. Many methods were proposed to deal with this problem, among which SSM (Bai, Bai, and Tian 2017) and JLML (Li, Zhu, and Gong 2017) achieve satisfactory results. Nevertheless, given the limited target data, our DDAN achieves better performance since the compared methods suffer from severe over-fitting problems.

Some UDA approaches have shown good results for person Re-ID. However, UDA requires unlabeled data from the target domain and cannot adapt when given insufficient training data. In contrast, DDAN fully utilizes source datasets and outperforms all UDA methods in Tab. 1.

Quantitatively, with MobileNet as the backbone network, the Rank-1 accuracy of DDAN is 52.3%, 54.5%, 50.6% and 78.5% for VIPeR, PRID, GRID, and i-LIDs, respectively. It outperforms the DG method DIMN (Song et al. 2019) in all test datasets, and DualNorm (Jia et al. 2019) in GRID and i-LIDs. Since DualNorm adopts a stronger backbone with MobileNetV2 and normalization layers, we conduct another experiment with the same settings as DualNorm. Results show that DDAN outperforms DualNorm in all test datasets.

4.3 Analysis

Domain-wise Adversarial Feature Learning. We use the baseline model to extract features from each domain, and compute the distance d_{WS} in Eq. (3) between every two domains in Tab.2. As for the pairwise alignment (All domains), since there is no fixed central domain, we instead compute the distance between features extracted by the baseline model and the model trained with pairwise alignment. Notice that the shortest distance appears between CUHK02 and CUHK03, which is consistent with the fact that these two datasets are collected from the same location (CUHK) and thus share some sort of similarities. The PersonSearch dataset also shows a relatively small distance to the CUHK datasets, since a part of this dataset is also collected in the same location as CUHK02 and CUHK03.

In contrast, we find that a significant discrepancy appears between the Duke and Market datasets. This is consistent with our visualization results in Fig.2(a). Since the minimum cost is achieved by PersonSearch, we set it as the central domain in all the experiments.

In addition, Tab. 3 shows the test performance when taking each dataset as the central domain. These results are generally consistent with the above observations in Tab. 2. For example, when setting Duke as the central domain, the cost of aligning the peripheral domains is large. Therefore, the effectiveness of the learned the features could be hurt due to the unsuitable alignment of the peripheral domains to the central domain. Indeed, the resulting model cannot perform well on any of the four test datasets. In contrast, the

Method	Type	VIPeR				PRID				GRID				i-LIDS			
		R-1	R-5	R-10	*m*AP	R-1	R-5	R-10	*m*AP	R-1	R-5	R-10	*m*AP	R-1	R-5	R-10	*m*AP
ImpTrpLoss (Cheng et al. 2016)	S	42.3	71.5	82.9	-	29.8	52.9	66.0	-	-	-	-	-	-	-	-	-
GOG (Matsukawa et al. 2016)	S	49.7	**79.7**	88.7	-	-	-	-	-	24.7	47.0	58.4	-	-	-	-	-
MTDnet (Chen et al. 2017)	S	47.5	73.1	82.6	-	32.0	51.0	62.0	-	-	-	-	-	58.4	80.4	87.3	-
OneShot (Bak and Carr 2017)	S	34.3	-	-	-	41.4	-	-	-	-	-	-	-	51.2	-	-	-
SSM (Bai, Bai, and Tian 2017)	S	53.7	-	**91.5**	-	-	-	-	-	27.2	-	61.2	-	-	-	-	-
JLML (Li, Zhu, and Gong 2017)	S	50.2	74.2	84.3	-	-	-	-	-	37.5	61.4	69.4	-	-	-	-	-
TJAIDL (Wang et al. 2018)	UDA	38.5	-	-	-	34.8	-	-	-	-	-	-	-	-	-	-	-
MMFAN (Lin et al. 2018)	UDA	39.1	-	-	-	35.1	-	-	-	-	-	-	-	-	-	-	-
Synthesis (Bak et al. 2018)	UDA	43.0	-	-	-	43.0	-	-	-	-	-	-	-	56.5	-	-	-
DIMN (Song et al. 2019)	DG	51.2	**70.2**	**76.0**	**60.1**	39.2	**67.0**	**76.7**	52.0	29.3	53.3	65.8	41.1	70.2	**89.7**	**94.5**	78.4
DDAN (Ours)	DG	**52.3**	60.6	71.8	56.4	**54.5**	62.7	74.9	**58.9**	**50.6**	**62.1**	**73.8**	**55.7**	**78.5**	85.3	92.5	**81.5**
DualNorm (Jia et al. 2019)	DG	53.9	62.5	75.3	58.0	60.4	73.6	84.8	64.9	41.4	47.4	64.7	45.7	74.8	82.0	91.5	78.5
DDAN+DualNorm (Ours)	DG	**56.5**	**65.6**	**76.3**	**60.8**	**62.9**	**74.2**	**85.3**	**67.5**	**46.2**	**55.4**	**68.0**	**50.9**	**78.0**	**85.7**	**93.2**	**81.2**

Table 1: Comparing Accuracy (%) with Baselines. "R": rank; "S": supervised training with the target dataset; "-": no report.

Center	Peripheral domains					Sum
	Cuhk02	Cuhk03	Duke	Market	Person	
Cuhk02	0	0.69	1.61	1.37	0.87	4.54
Cuhk03	0.69	0	1.58	1.44	0.72	4.43
Duke	1.61	1.58	0	1.69	1.20	6.08
Market	1.37	1.44	1.69	0	1.10	5.60
Person	0.87	0.72	1.20	1.10	0	3.89
All domains	1.81	1.91	1.92	1.78	1.93	9.35

Table 2: Wasserstein distance d_{WS} of different domains.

Center	VIPeR	PRID	GRID	i-LIDs
Cuhk02	48.4	**48.5**	46.6	**74.8**
Cuhk03	49.0	45.2	**48.4**	**75.0**
Duke	49.2	47.3	45.1	72.5
Market	**49.5**	48.2	46.5	74.1
Person	**50.6**	**50.0**	**47.6**	74.6
All domains	48.7	45.4	44.2	74.5

Table 3: Rank-1 Accuracy with Different Central Domains for "Baseline + $\mathcal{L}_{DA}$" (top-2 accuracies are bolded).

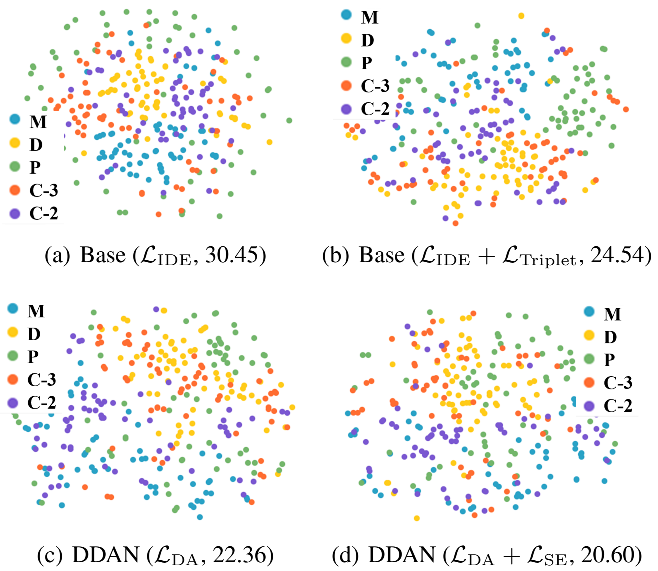
Image A	Image B	Image C	$\mathcal{L}_{SE}$?	cs(A,B)	cs(A,C)
Duke	Duke	Others			
			✗	0.67	0.59
			✓	0.65	0.83
			✗	0.78	0.75
			✓	0.86	0.91
			✗	0.76	0.71
			✓	0.77	0.85

Table 4: Cosine Similarity (cs) of Representative Images.

central domain PersonSearch demonstrates a better performance. In particular, the resulting rank-1 accuracy on PRID is 4.8% higher than the lowest one. Furthermore, we also evaluate the multi-domain approach (All domains) by pairwisely aligning the domains, leading to unsatisfactory performance in all the test datasets.

Identity-wise Similarity Enhancement. We demonstrate the effectiveness of this component in Tab. 4 by comparing the cosine similarity between three representative images. In a general scenario without $\mathcal{L}_{\mathrm{SE}}$, this similarity is dominated by the domain variations; non-similar images (A and B) turn out to be closer than the similar ones (A and C) in the feature space, only because they are in the same domain, which may have a similar hue and lighting. In contrast, the similarities with $\mathcal{L}_{\mathrm{SE}}$ can correctly reflect the relationships among A, B and C, even if the (incorrectly) large similarity between non-similar same-domain images is not penalized. It successfully presents a real-world scenario, where the pedestrians in A are more likely to be the same identity as those in C than B, even if A and B are from the same domain, as long as A and C share more similar appearances than A and B. Therefore, it captures the local similarity between A and C and effectively reduces the domain-shift. Notice that we do not compare by the rows, because the features learned with and without identity-wise enhancement are in different subspaces since the two different models are trained separately. Still, the improvement of row-wise similarities between A and C is larger than A and B, demonstrating that our enhancement is as ideal as described from another perspective.

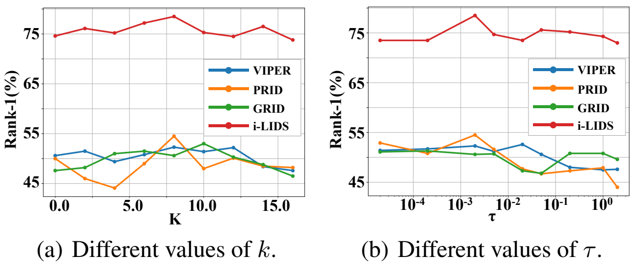
Visualization. We use t-SNE to visualize the distribution of the features obtained by the networks with different loss functions. In Fig. 2, we sample 64 examples from different datasets (denoted by different colors). To better demonstrate the distributional change, we list the sum of Wasserstein distances between every two domains in each caption. For the baseline network with only $\mathcal{L}_{\mathrm{IDE}}$, the distribution in Fig. 2(a) shows clear discrepancy among all domains with few overlaps. Particularly, the features of Market, Duke, and PersonSearch are clearly distinguished from each other. The triplet loss in Fig. 2(b) shortens the intra-class distance and widens the inter-class one, thus the model learns discriminative features while also relatively increasing the distance between each domain, as the labels of each domain are different. In particular, despite the properly aligned CUHK02 and CUHK03, the PersonSearch dataset can be seen as two parts: one well aligned with CUHK02 and CUHK03 that are



(a) Base ($\mathcal{L}_{\text{IDE}}$, 30.45) (b) Base ($\mathcal{L}_{\text{IDE}} + \mathcal{L}_{\text{Triplet}}$, 24.54)

(c) DDAN ($\mathcal{L}_{\text{DA}}$, 22.36) (d) DDAN ($\mathcal{L}_{\text{DA}} + \mathcal{L}_{\text{SE}}$, 20.60)

Figure 2: Feature distribution in source domains. Market (m), Duke (d), Person (p), CUHK03/02 (c-3/2). Features are extracted by models trained with specified losses. Numbers are the sum of pair-wise Wasserstein distances.



(a) Different values of k. (b) Different values of τ.

Figure 3: Evaluation with different k and τ in Eq. (8).

collected from the same location, and the other relatively more independent one collected from movie snippets. Also, the features from Duke are distinguished from the others.

With also the domain-wise adversarial feature learning loss $\mathcal{L}_{\text{DA}}$ in Fig. 2(c), the distributions of different domains are better aligned and more instances tend to be consistent with each other. However, the local distribution of Duke and part of PersonSearch are still distinguished. Lastly, the identity-wise similarity enhancement loss in Fig. 2(d) achieves the ideal scenario expected by DDAN, in which the distributions of similar IDs from different domains are closer. Moreover, the domain shift is greatly reduced to improve the generalization as compared against the baseline.

4.4 Important Parameters

We study the impact of two important hyper-parameters: the temperature τ and the number k of similar IDs in Eq. (8). We change one parameter while fixing the other.

Temperature τ of softmax. In Eq. (8), we use softmax to reduce the influence of exceptionally unmatched entries appearing in the paired representations. We also add a temperature parameter to preserve the distinguishability of features, otherwise the softmax outputs could be small due to the large number of dimensions. We empirically observe that $\tau < 1$ leads to better results in Fig. 3(b). However, the network is hard to converge with a small τ like $2 * 10^{-5}$. We obtain the best results with $\tau \approx 2 * 10^{-3}$.

Loss functions	VIPeR	PRID	GRID	i-LIDs
$\mathcal{L}_{\text{IDE}}$	41.4	30.8	38.1	66.2
$\mathcal{L}_{\text{IDE}} + \mathcal{L}_{\text{Triplet}}$	47.2	46.4	45.3	72.3
$\mathcal{L}_{\text{IDE}} + \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{DA}}$	50.6	50.0	47.6	74.6
$\mathcal{L}_{\text{IDE}} + \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{SE}}$	50.8	48.8	49.6	73.7
$\mathcal{L}_{\text{IDE}} + \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{DA}} + \mathcal{L}_{\text{SE}}$	52.3	54.5	50.6	78.5

Table 5: Ablation Study (Rank-1 Accuracy (%)).

Number k of similar IDs. We study the number k of similar IDs for identity-wise similarity enhancement in Fig. 3(a). $k = 0$ disables this mechanism. For $k \geq 1$, the enhancement generally improves the performance with relatively small values of k. However, setting a large k may incorrectly capture non-similar examples, which could have deleterious effects on the performance. Overall, $k = 8$ achieves the best rank-1 accuracy and mAP in most datasets.

4.5 Ablation Study

We study each component's effectiveness on the full test set, as shown in Tab. 5. $\mathcal{L}_{\text{Triplet}}$ with BNNeck (Luo et al. 2019) greatly improves the effectiveness of learned representations. $\mathcal{L}_{\text{DA}}$ aligns all source domains to learn a domain-invariant feature space, whose improved performance in unseen datasets indicates a better generalization of learned features. $\mathcal{L}_{\text{SE}}$ enforces a real-world distribution, capturing identity-wise similarity to reduce the local domain shift. We also study the effectiveness of $\mathcal{L}_{\text{DA}}$ by disabling it. We observe that all components are effective and jointly contribute to domain-invariant features that are discriminative, and insensitive to domain- and identity-wise variations.

5 Conclusion

This paper identifies two fundamental challenges in DG for person Re-ID: *domain-wise variations* and *identity-wise similarities*. We propose an end-to-end Dual Distribution Alignment Network (DDAN) to learn domain-invariant features with two constraints: the domain-wise adversarial feature learning and the identity-wise similarity enhancement. At the domain level, we align peripheral domains towards the central domain to reduce the domain discrepancy with minimum distributional shifts. At the identity level, we reduce the domain shift by capturing identity-wise similarity with an ID pool across domains. It realizes an ideal scenario, where any group of visually-similar IDs, though from different domains, are closer than non-similar ones from the same domain. Quantitative results on a large-scale DG Re-ID benchmark demonstrate the superior performance of DDAN against other recent methods.

Acknowledgements

This work is supported by the National Science Fund for Distinguished Young (No.62025603), the National Natural Science Foundation of China (No.U1705262, No. 62072386, No. 62072387, No. 62072389, No. 62002305, No.61772443, No.61802324 and No.61702136) and Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049).

References

Akuzawa, K.; Iwasawa, Y.; and Matsuo, Y. 2019. Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization. *CoRR* .

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *CoRR* .

Bai, S.; Bai, X.; and Tian, Q. 2017. Scalable person re-id on supervised smoothed manifold. In *CVPR*.

Bak, S.; and Carr, P. 2017. One-Shot Metric Learning for Person Re-id. In *CVPR*.

Bak, S.; et al. 2018. Domain Adaptation Through Synthesis for Unsupervised Person Re-id. In *ECCV*.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. A Multi-Task Deep Network for Person Re-Id. In *AAAI*.

Chen, Y.; Zhu, X.; Zheng, W.; and Lai, J. 2018. Person Re-Identification by Camera Correlation Aware Feature Augmentation. *TPAMI* .

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person Re-id by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In *CVPR*.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Burges, C. J. C.; Bottou, L.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2292–2300. URL http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-Image Domain Adaptation With Preserved Self-Similarity and Domain-Dissimilarity for Person Re-Identification. In *CVPR*.

Dou, Q.; Castro, D. C.; Kamnitsas, K.; and Glocker, B. 2019. Domain Generalization via Model-Agnostic Learning of Semantic Features. *CoRR* .

Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; and Huang, T. S. 2019. Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification. In *ICCV*.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; and Balduzzi, D. 2015. Domain Generalization for Object Recognition with Multi-task Autoencoders. In *ICCV*.

Gray, D.; and Tao, H. 2008. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *ECCV*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In Defense of the Triplet Loss for Person Re-id. *CoRR* .

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person Re-id by Descriptive and Discriminative Classification. In *SCIA*.

Jia, J.; et al. 2019. Frustratingly Easy Person Re-id: Generalizing Person Re-ID in Practice. *CoRR* .

Köstinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018a. Domain Generalization With Adversarial Feature Learning. In *CVPR*.

Li, W.; and Wang, X. 2013. Locally Aligned Feature Transforms across Views. In *CVPR*.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-id. In *CVPR*.

Li, W.; Zhu, X.; and Gong, S. 2017. Person Re-id by Deep Joint Learning of Multi-Loss Classification. In *IJCAI*.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018b. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *ECCV*.

Li, Y.; Yang, Y.; Zhou, W.; and Hospedales, T. M. 2019. Feature-Critic Networks for Heterogeneous Domain Generalization. In *ICML*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. In *CVPR*.

Lin, S.; Li, H.; Li, C.; and Kot, A. C. 2018. Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Id. In *BMVC*.

Loy, C. C.; Xiang, T.; and Gong, S. 2010. Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding. *IJCV* .

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-id. In *CVPR Workshops*.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical Gaussian Descriptor for Person Re-id. In *CVPR*.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain Generalization via Invariant Feature Representation. In *ICML*.

Paisitkriangkrai, S.; et al. 2015. Learning to rank in person re-id with metric ensembles. In *CVPR*.

Peng, P.; Xiang, T.; Wang, Y.; Pontil, M.; Gong, S.; Huang, T.; and Tian, Y. 2016. Unsupervised Cross-Dataset Transfer Learning for Person Re-identification. In *CVPR*.

Sandler, M.; Howard, A. G.; Zhu, M.; Zhmoginov, A.; and Chen, L. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*.

Song, J.; Yang, Y.; Song, Y.; Xiang, T.; and Hospedales, T. M. 2019. Generalizable Person Re-id by Domain-Invariant Mapping Network. In *CVPR*.

Wang, H.; Gong, S.; Zhu, X.; and Xiang, T. 2016. Human-in-the-Loop Person Re-identification. In *ECCV*.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Id. In *CVPR*.

Xiao, T.; ands Bochao Wang, S. L.; Lin, L.; and Wang, X. 2016. End-to-End Deep Learning for Person Search. *CoRR* .

Xiong, F.; Gou, M.; Camps, O. I.; and Sznaier, M. 2014. Person Re-Identification Using Kernel-Based Metric Learning Methods. In *ECCV*.

Xu, Z.; Li, W.; Niu, L.; and Xu, D. 2014. Exploiting Low-Rank Structure from Latent Domains for Domain Generalization. In *ECCV*.

Yang, P.; and Gao, W. 2013. Multi-View Discriminant Transfer Learning. In *IJCAI*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-id: A Benchmark. In *ICCV*.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person Re-id: Past, Present and Future. *CoRR* .

Zheng, W.; Gong, S.; and Xiang, T. 2009. Associating Groups of People. In *BMVC*.

Zheng, W.; Gong, S.; and Xiang, T. 2013. Reidentification by Relative Distance Comparison. *TPAMI* .

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *ICCV*.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018. Generalizing a Person Retrieval Model Hetero- and Homogeneously. In *ECCV*.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification. In *CVPR*.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*.