Learning Domain Invariant Representations for Generalizable Person Re-Identification

Yi-Fan Zhang[1], Zhang Zhang[1], Da Li[1], Zhen Jia[1], Liang Wang[1], Tieniu Tan[1]

[1]Institute of Automation, Chinese Academy of Sciences

Abstract

Generalizable person Re-Identification (ReID) has attracted growing attention in recent computer vision community. In this work, we construct a structural causal model (SCM) among identity labels, identity-specific factors (clothing/shoes color etc.), and domain-specific factors (background, viewpoints etc.). According to the causal analysis, we propose a novel Domain Invariant Representation Learning for generalizable person Re-Identification (DIR-ReID) framework. Specifically, we first propose to disentangle the identity-specific and domain-specific feature spaces, based on which we propose an effective algorithmic implementation for backdoor adjustment, essentially serving as a causal intervention towards the SCM. Extensive experiments have been conducted, showing that DIR-ReID outperforms state-of-the-art (SOTA) methods on large-scale domain generalization (DG) ReID benchmarks.
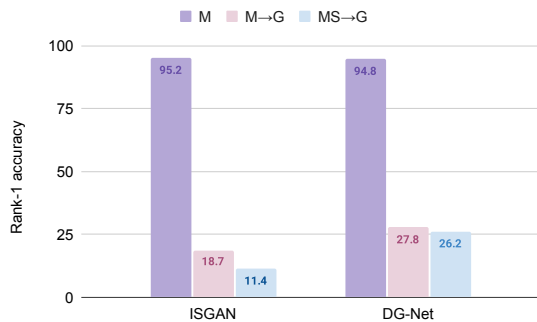
1. Introduction

Figure 1. Performance of two representative models. M: Train and test on Market1501. M→G: Trained on Market1501 and tested on GRID . MS→G: Trained on multi-source datasets and tested on GRID.

Person Re-IDentification (ReID) aims at matching person images of the same identity across multiple camera views. In previous works, most ReID methods are trained and tested on the same dataset, termed fully-supervised methods [69, 72], or tested on new target domains different from the training one, termed domain adaptation (DA) methods [32, 45]. Although recent fully-supervised methods have achieved remarkable performance, they tend to fail catastrophically when tested in out-of-distribution (OOD) settings. Figure 1 illustrates the fragility of two representative fully-supervised works, *i.e.*, the DG-Net [72] and IS-GAN [12], which get very high rank-1 accuracies above $94\%, 95\%$ respectively when tested on Market1501 dataset. However, the rank-1 accuracies drop to 18.7% and 27.8% respectively when directly tested on the dataset of GRID, which suggests the weak extrapolation capability and poor robustness of fully-supervised methods. We further train these two models over multiple source domains (the details of sources are in Section 4.1). However, the even worse performance are obtained, which indicate the challenge of cross-domain ReID. To tackle these problems, a number of DA methods have been proposed to mitigate the domain gap without the need for extensive manual annotations for new camera settings. However, as the deployment information of new cameras is unknown, large amounts of unlabeled data are required for DA retraining. These problems severely hinder real-world applications of current person ReID systems.

To tackle the above challenges, we focus on a more realistic and practical setting: generalizable person ReID, where the model is trained on multiple large-scale public datasets and tested on unseen domains. The generalizable person ReID is originally formulated as a problem of domain generalization (DG) [52], which is more practicable than the traditional ReID paradigm since the ready-made models can work on any new settings without repeating the whole process of data collection, annotation, and model updating.

Inspired by the observation that human beings can easily distinguish the identity-specific factors (e.g. appearances, body shapes) from domain-specific factors (e.g. imaging conditions such as backgrounds, viewing angles, illuminations *etc*.) under various visual contexts, we first present a

causal view on the generalizable person ReID and use the SCM to provide insights for the poor generalization of traditional ReID models when applied to unseen domains. The assumption behind is that person images are affected by two sets of latent random variables, *i.e.* identity-specific factors S, and domain-specific factors V. In this paper, we highlight the potential reason for poor generalization ability: domain-specific factors can confound generalizable ReID model and the spurious correlations between V and S hinders the model to make robust prediction on identity label Y based on S. Thus a novel DIR-ReID paradigm is proposed to disentangle and combat combating spurious features.

Specially, a Multi-Domain Disentangled Adversarial Neural Network (MDDAN) is first proposed to jointly learn two encoders for embedding identity-specific and domain-specific factors from multiple public ReID datasets, where the adversarial learning principle is adopted to exclude the domain (identity) related information from the embedded identity (domain) specific representations. Then a backdoor adjustment block (BA) is proposed to approximate the interventional distribution [36], which can pursue the true causality between identity-specific factors and identity label. These two components (MDDAN and BA) are integrated as a whole system, termed DIR-ReID.

To sum up, the contributions of our work can be summarized as:

- For the first time, a causal perspective on the analysis of generalizable person ReID is introduced. Our Structural Causal Model (SCM) assumption shows that the domain-specific factor is essentially a confounder, which has spurious correlations with identity-specific factors.

- A causal approximation to the backdoor adjustment named DIR-ReID is proposed, where an MDDAN block is adopted to identify identity-specific and domain-specific factors from multiple data sources. Then a BA block is adopted as causal interventions. Mathematical analysis proves the characteristics of our method;

- Comprehensive experiments are conducted to demonstrate the effectiveness of the learned representations. Our method achieves superior performance in comparison with state-of-the-art methods on large-scale generalizable person ReID benchmarks.

2. Related Work

Single-domain Person ReID Existing works of single-domain person ReID (*i.e.* supervised person ReID) usually depend on the assumption that in training and testing data are independent and identically distributed (i.i.d.). They usually design or learn discriminative features [34] and develop efficient metrics [23]. With the rapid development of deep Convolutional Neural Networks (CNNs), single-domain person ReID has achieved great progress. Some of the CNN-based methods introduce human parts [24, 54], poses [68], and masks [51] to improve the robustness of extracted features. And some other methods use deep metric learning to learn appropriate similarity measurement [9]. Despite the encouraging performance under the single-domain setup, current fully-supervised models for person ReID degrade significantly when deployed to an unseen domain.

Cross-domain Person ReID Unsupervised Domain Adaptation (UDA) has achieved great progress [44] and is commonly adopted for cross-domain person ReID to achieve good performance on unlabeled target domain with both the labeled source data and unlabeled target data. The UDA-based ReID methods usually attempt to transfer the knowledge from source domains to the target one depending on target-domain images [32, 20], features [57] or metrics [43]. Another group of UDA-based methods [14, 65] propose to explore hard or soft pseudo labels in unlabeled target domain using its data distribution geometry. Though UDA-based methods improve the performance of cross-domain ReID to a certain extent, most of them require a large amount of unlabeled target data for adaptation.

Generalizable Person ReID Recently, Generalizable person ReID methods [52] are proposed to learn a model that can generalize to unseen domains without target data and the adaptation. Existing methods mainly follow a meta-learning pipeline or utilize domain-specific heuristics. Jia *et al.* [21] learn the domain-invariant features by integrating the Instance Normalization (IN) into the network to filter out style factors. Jin *et al.* [22] extend the work [21] by restituting the identity-relevant information to network to ensure the model discrimination. Lin *et al.* [28] propose a feature generalization mechanism by integrating the adversarial auto-encoder and Maximum Mean Discrepancy (MMD) alignment. Song *et al.* [52] propose a Domain-Invariant Mapping Network (DIMN) following the meta-learning pipeline and maintaining a shared memory bank for training datasets. There also have some works try to learn domain-invariant features, *e.g.* DANN [13], DDAN [7] and CaNE [63]. The differene between DIR-ReID with these methods are detailed in Appendix F.

Domain Generalization Domain/Out-of-distribution generalization [18] aims to learn representations $\Phi(X)$ that is invariant across environments $\mathcal{E}$ so that model can well extrapolate in unseen environments. The problem can be formulated as $\min_{\Phi} \max_{e \in \mathcal{E}} \mathbb{E}[l(y, \Phi(x)) \mid E = e]$. Representative approaches like IRM [1] are proposed to tackle this challenge. However, IRM would fail catastrophically unless the test data are sufficiently similar to the training

distribution [46]. To address those challenges, we proposed DIR-ReID to explicitly remove the confounding effects of spurious features via backdoor adjustment.

Causality for CV Causal Representation Learning [49] combines machine learning and causal inference, has attracted increasing attention on a learning paradigm with improved generalization and trustworthiness. Simultaneously, there is a growing number of computer vision tasks that benefit from causality [2, 33]. Most of them focus on causal effects: disentangling the desired model effects [5], and modularizing reusable features that generalize well [39]. Recently, causal intervention is also introduced into computer vision [58, 61, 56]. Specifically, CONTA [11] removes the confounding bias in image-level classification by backdoor adjustment and thus provides better pseudo-masks as ground-truth for the subsequent segmentation model. IFSL [64] believes that pre-training is a confounder hurting the few-shot learning performance. Specifically, they propose a SCM in the process of FSL and then develop three practical implementations based on the backdoor adjustment. We also adopt the SCM of Pearl [41] to model the causal effects of person ReID. We use causality to provide the explanations why traditional methods work poorly on unseen domains and further propose methods to tackle the problem based on the backdoor adjustment.

3. Learning Disentangled and Invariant Representations

In this section, we first introduce the proposed SCM to illustrate the spurious correlations between S and V. Then, a DIR-ReID model is proposed to approximate the interventional distribution, which captures the true causality between S and Y. Finally, a theoretical analysis is taken for a better understanding of our method.

3.1. SCM for Generalizable Person ReID

Inspired by current research of harnessing causality in machine learning [2], we propose a SCM to analyze the disentanglement and generalization in person ReID models.

Following the causal models in [6, 53], we use the SCM (in Figure 2(a)) to describe the causal relationships between person images and person identities, where Y denotes the observable variables of identity labels, S and V are the latent variables indicating identity-specific and domain-specific factors respectively. We write $x_i = G(s_i, v_i)$, where G is the map from the latent space to the pixel space. As shown in the model, there are three kinds of causal relationships as follows:

$S \rightarrow Y$. Identity-specific factors S directly cause Y, which means the person identities are determined by their identity-specific information, such as clothing styles, body shapes, *etc.*, instead of the backgrounds, illuminations and camera viewpoints.
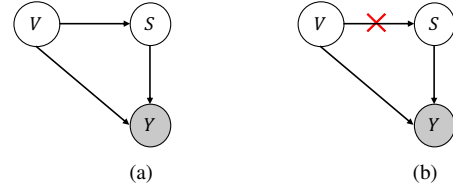

Figure 2. Graphical representation of person ReID methods. X: pedestrian images, S, V: identity-specific and domain-specific factors, Y, D: pedestrian and domain labels. Gray circles denote observable variables. (a) Traditional ReID model where $P(Y|S) \neq P(Y|do(S))$. (b) Interventional ReID model where $P(Y|S) \approx P(Y|do(S))$.

$V \rightarrow S, V \rightarrow Y$. There exist many confounders $v_i \in V$ (*e.g.*, backgrounds, illuminations, and viewpoints) that affect both identity-factors S and person identity Y. The $V \rightarrow S$ edge indicates the spurious correlations between S and V in the real world, $V \rightarrow Y$ denotes the influences of V on Y. For example, pedestrians mostly appear on campus in the CUHK03 dataset while in a busy underground station in the GRID dataset and in sidewalks in the PRID dataset. Due to different illumination conditions and backgrounds, even for pedestrians wearing the same color clothes, their semantic factors vary greatly. Especially, Pedestrians in column 1 of Figure 3 are all wearing white while their semantic factors (especially clothing colors) will be influenced by confounders[1]. So the pedestrian identities and their variation factors are spuriously correlated.

An ideal ReID model should capture the true causality between S and Y to generalize to unseen samples. However, from the SCM in Figure 2(a), the conventional correlation $P(Y|S)$ fails to do so, because the likelihood of Y given S is not only due to "S causes Y" via $S \rightarrow Y$, but also the spurious correlations via $V \rightarrow Y$. Therefore, to pursue the true causality between S and Y, we need to use the causal intervention [36] $P(Y|do(S))$ instead of the likelihood $P(Y|S)$ for the ReID model.

3.2. Causal Intervention via Backdoor Adjustment

The above formulation only gives a causal quantity $P(Y|do(S))$ without further *identification* and grounded method for computing it from a purely statistical quantity. Therefore, we propose to use the backdoor adjustment [42] to identify and compute $P(Y|do(S))$ without the need for an ideal dataset[2]. The back-door adjustment assumes that we can observe and stratify the confounders,

---

[1]The white cloth of P_1 in GRID is more yellowish and that of P_1 in PRID is more greenish

[2]An ideal dataset has images of every pedestrian in all backgrounds and illumination conditions, which is unbiased and has few spurious correlations. Namely there isn't any edge $V \rightarrow S$ and then $P(Y|S) = P(Y|do(S))$.
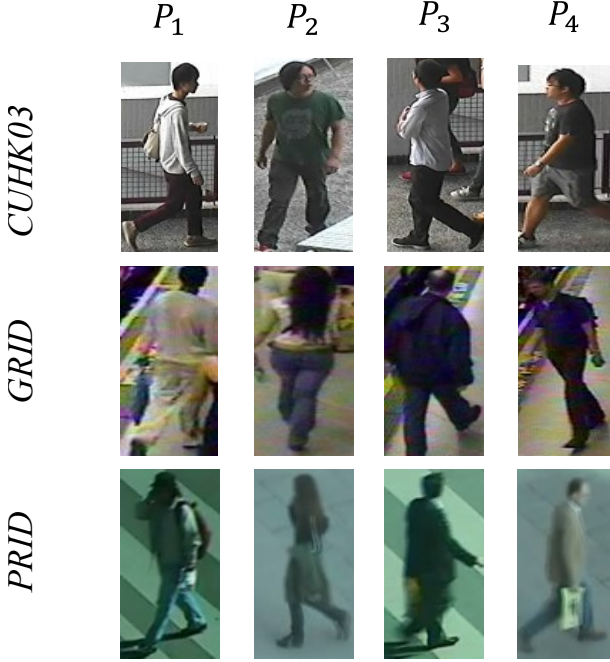
Figure 3. Spurious correlations between domain-specific factors and identity-specific factors.

Figure 4. **Schematic description of the proposed approach.** x_0, x_1, x_2: Three images from one mini-batch. f_S, f_V: encoders of identity-specific and domain-specific factors. s, v: identity-specific and domain-specific factors represented in latent space. $\mathcal{C}_S, \mathcal{C}_V, \mathcal{C}_C$: classifiers for identity-specific factors, domain-specific factors, and concatenated vectors. $\oplus$: concatenation of two latent vectors.

i.e. $V = \{v_i\}_{i=1}^{|V|}$, where each v_i is the domain-specific factor corresponding to special camera view, illumination condition *etc*. Formally, the backdoor adjustment for the graph in Figure 2(b) is (the detailed proof is shown in the Section A of the Appendix):

$$P(Y|do(S = s_k)) = \sum_{i=1}^{|V|} P(Y|S = s_k, V = v_i)P(V = v_i) \tag{1}$$

There are two challenges: (i) it is hard to instantiate v and s, namely learning the inverse map $v_i = G_V^{-1}(x_i)$ and $s_i = G_S^{-1}(x_i)$. (ii) it is almost impossible to enumerate all domain-specific factors v. Next, we will offer a practical implementation of Eq. 1 for interventional ReID: DIR-ReID.

3.3. The DIR-ReID Framework

Notations and Problem Formulation. For generalized person ReID, we have access to G labeled datasets $\mathcal{D} = \{\mathcal{D}_g\}_{g=1}^G$. Each dataset $\mathcal{D}_g = \{(x_i, y_i, d_i)\}_{i=1}^{N_g}$, where N_g is the number of images in $\mathcal{D}_g$ and (x_i, y_i, d_i) is the image, identity label and domain label for the image respectively. In the training phase, we train a DG model using $N = \sum_{g=1}^G N_g$ aggregated image-label pairs with $M = \sum_{g=1}^G M_g$ i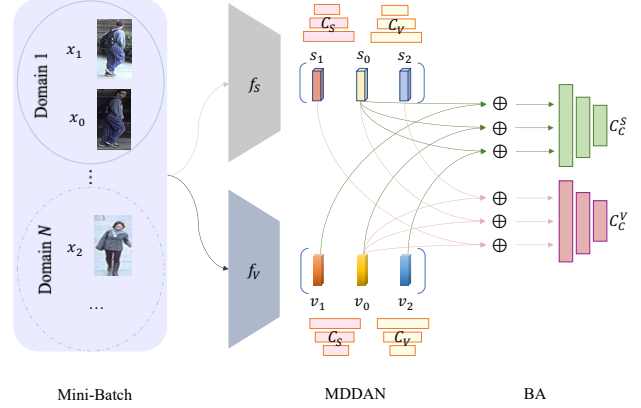dentities from all source domains. In the testing phase, we perform a retrieval task on unseen target domains without additional model updates.

As analyzed in Section 3.2, the challenges are (i) how to get the representations of S and V from observable variable X. (ii) how to approximately marginalize over the feature space V. Here, we propose DIR-ReID to tackle these challenges. DIR-ReID consists of three blocks, as shown in Figure 4.

(i) Identity adversarial learning block, which is a domain-agnostic, identity-aware encoder f_S (namely G_S^{-1} mentioned above) to obtain identity-specific factors.

(ii) Domain adversarial learning block, which is a domain-aware, identity-agnostic encoder f_V (namely G_V^{-1} mentioned above) to identify domain-specific factors.

(iii) Backdoor adjustment block, which approximates the backdoor adjustment in Eq. 1 based on the disentangled representation space V and S.

Components (i) and (ii) are named MDDAN. As shown in figure 4, the overall process includes feeding images x_0, x_1, x_2 randomly selected from N source domain. into identity-specific and domain-specific encoders to get disentangled representations s_0, s_1, s_2 and v_0, v_1, v_2 via adversarial learning. Then backdoor adjustment is performed based on these representations to train the encoders f_S and f_V for learning invariant representations.

Identity Adversarial Learning Block. An identity-aware encoder f_S is adopted to extract identity-specific factors. Then, a classifier $\mathcal{C}_s$ is used to identify the ID label for a given person image x_i. The cross-entropy loss with label smoothing [35] is calculated for training the encoder

4

f_S, which is defined as:

$$\mathcal{L}_{id}^s = -\sum_{i=1}^{N} \log P\left(Y = y_i | \mathcal{C}_S\left(f_S\left(x_i\right)\right)\right) \qquad (2)$$

where θ_S is the parameters of f_S, $\mathcal{C}_S$ is the identity classifier. y_i is the labeled person identity of x_i.

To exclude all the domain information from identity-specific factors, a domain classifier $\mathcal{C}_V$ is adopted for adversarial learning. One promising way is using gradient reversal layer (GRL) technique in DANN [13] to train the encoder f_S and classifier $\mathcal{C}_v$ simultaneously, where a misclassification loss is adopted to enforce an image not to be classified into its true domain class. However, "misclassification" does not mean "indistinguishability". Intuitively, "indistinguishability" means classifying one image into all the domains equiprobably. Hence, we adopt a loss of maximum entropy (minimization of negative entropy [6]) termed the domain-indistinguishability loss as follows:

$$\mathcal{L}_{indis}^{dom} = -\mathcal{H}(D|S), \qquad (3)$$

where $\mathcal{H}$ is the entropy to measure the uncertainty of predicted domain class given identity-specific factors, *i.e.*,

$$\begin{aligned} \mathcal{H}(D|S)) = &-\sum_{i=1}^{N}\sum_{g=1}^{G} P\left(D = g | \mathcal{C}_V\left(f_S\left(x_i\right)\right)\right) \\ &\cdot \log P\left(D = g | \mathcal{C}_V\left(f_S\left(x_i\right)\right)\right), \end{aligned} \qquad (4)$$

It encourages the identity-specific factors of all images to contain less domain information about their datasets.

As a result, the parameters θ_S of the identity-aware encoder f_S are optimized by jointly minimizing the identity-classification loss and the domain-indistinguishability loss. The overall objective function of identity adversarial learning is:

$$\min_{\theta_S} \mathcal{L}_{id}^s + \lambda_1 \mathcal{L}_{indis}^{dom}. \qquad (5)$$

Domain Adversarial Learning Block. This block aims to extract the domain-specific factors from person images x_i and exclude identity-specific factors. As it is a dual task to the identity adversarial learning, we can use a similar training schema to learn the parameters of f_V. Here, the identity-classification loss and the domain-indistinguishability loss in the above section are replaced by the domain-classification loss $\mathcal{L}_{dom}^v$ and the identity-indistinguishability loss $\mathcal{L}_{indis}^{id}$ respectively. The objective of the domain adversarial learning block is as follows:

$$\min_{\theta_V} \mathcal{L}_{dom}^v + \lambda_2 \mathcal{L}_{indis}^{id}. \qquad (6)$$

The domain-classification loss $\mathcal{L}_{dom}^v$ is defined as,

$$\mathcal{L}_{dom}^v = -\sum_{i=1}^{N} \log P(D = d_i | \mathcal{C}_V(f_V(x_i))), \qquad (7)$$

where d_i is the domain label of image x_i. The identity-indistinguishability loss $\mathcal{L}_{indis}^{id}$ is similar to Equation 4 but has different number of classes (number of classes for $\mathcal{L}_{indis}^{id}$ is M and G for $\mathcal{L}_{indis}^{dom}$).

Backdoor Adjustment Block. With the disentangled representations learnt from MDDAN, we can implement backdoor adjustment now. Given samples $\{x_i\}_{i=1}^N$. We first project them into f_S and f_V and get $\{v_i\}_{i=1}^N$ and $\{s_i\}_{i=1}^N$. Then we follow two similar assumptions as [64].
(i) $P(v_i) = 1/|V|$, where we assume a uniform prior for the adjusted domain-specific factors.
(ii) $P(Y|S = s_k, V = v_i) = P(Y|\mathbf{s_k} \oplus \mathbf{v_i})$, where $\oplus$ denotes vector concatenation. Then the overall backdoor adjustment is:

$$P(Y|do(S = s_k)) = \frac{1}{|V|}\sum_{i=1}^{|V|} P(Y|\mathbf{s} \oplus \mathbf{v})) \qquad (8)$$

However, we still cannot enumerate the feature space V. Here we propose 4 different approximate implementation to select K domain-specific factors.

- K-Random. For each x_i, we randomly select K different domain-specific factors $\{v_j\}_{j=1}^K$ where $j \neq i$.

- K-Hardest. For each x_i, we select K domain-specific factors $\{v_j\}_{j=1}^K$ which are the most dissimilar to v_i.

- K-Mixup. For each x_i, we can create more domain-specific factors by mixup [66, 62]. We generate K mixed sample feature $\tilde{v}$ by interpolation of two randomly selected features (a pair $\langle v_j, v_j' \rangle$), denoted by

$$\tilde{v} = \alpha v_j + (1 - \alpha)v_j', \qquad (9)$$

where $\alpha \in [0, 1]$ controls the interpolation degree and we empirically set $\alpha = 0.5$ in our experiments.

- K-MixHard. For each x_i, we firstly select K domain-specific factors $\{v_j\}^K$ least like v_i and generate K mixed features by randomly interpolating these factors.

Then we can approximate the backdoor adjustment by

$$P(Y|do(S = s_k)) = \frac{1}{K}\sum_{i=1}^{K} P(Y|s_k \oplus v_i)) \qquad (10)$$

In the experiments we just ignore the constant $1/K$ and the final loss function is

$$\mathcal{L}_{invar}^{id} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \log P(Y = y_i | \mathcal{C}_C^S(s_i \oplus v_k)), \qquad (11)$$

3.4. Model Summary

Recalling the adversarial learning of identity-aware encoder f_S and domain-aware encoder f_V (Eq. (2) and Eq. (7)), the identity and domain classifiers need to be trained with the inputs of domain-specific and identity-specific factors. The classification losses are calculated in terms of the cross entropy function.

$$\mathcal{L}_{id}^{v} = -\sum_{i=1}^{N} \log P\left(Y = y_i | \mathcal{C}_S\left(f_V\left(x_i\right)\right)\right). \quad (12)$$

$$\mathcal{L}_{dom}^{s} = -\sum_{i=1}^{N} \log P\left(D = d_i | \mathcal{C}_V(f_S(x_i))\right). \quad (13)$$

Finally, given the parameters ϕ_S, ϕ_V of classifiers $\mathcal{C}_S, \mathcal{C}_V$, the total loss function is summarized as follows.

$$\min_{\phi_S, \phi_V} \mathcal{L}_{id}^{s} + \mathcal{L}_{dom}^{s} + \mathcal{L}_{dom}^{v} + \mathcal{L}_{id}^{v}. \quad (14)$$

And the overall loss functions for the other components are defined as

$$\begin{aligned}
&\min_{\theta_S, \phi_C^S} \mathcal{L}_{id}^{s} + \lambda_1 \mathcal{L}_{indis}^{dom} + \lambda_3 \mathcal{L}_{invar}^{id}, \\
&\min_{\theta_V, \phi_C^V} \mathcal{L}_{dom}^{v} + \lambda_2 \mathcal{L}_{indis}^{id} + \lambda_4 \mathcal{L}_{invar}^{id},
\end{aligned} \quad (15)$$

where ϕ_C^S, ϕ_C^V are parameters of classifiers $\mathcal{C}_C^S$ and $\mathcal{C}_C^V$ respectively. With the above components, each mini-batch training process is divided into two phases. In the Phase I, the encoders f_S and f_V as well as the augmented data classifiers C_C^S and C_C^V are trained by Eq. (15), while the identity and domain classifiers are fixed. Then in the Phase II, identity and domain classifiers are trained by Eq. (14), while other components are fixed.

3.5. Analysis

Lemma 3.1 *Let $\mathcal{D}_i$ denote one of the source domains, s^τ are identity-specific factors of images from the τ-th domain. Let $p(s^\tau | \mathcal{D}_i)$ be the class-conditional density function of the τ-th domain given domain information $\mathcal{D}_i$. It can be proved that, carrying out the MDDAN will lead to*

$$p(s^\tau | \mathcal{D}_i) = p_\tau(s^\tau), \forall s^\tau \in \mathcal{D}_\tau; i = 1, ..., K. \quad (16)$$

It indicates that in the latent space of identity-specific representation, any image will be invariant to K different domains. Its class-conditional density function value for these domains (e.g., $p(s^\tau | \mathcal{D}_i)$) is just equal to its prior density function value in its own domain (e.g., $p_\tau(s^\tau)$), but not dependents on the domain index i. The proof is shown in the Section B of the Appendix.

Lemma 3.2 *From the view of information theory, MDDAN is minimizing the mutual information between identity-specific factors S and domain information D, namely $\min \mathcal{I}(S, D)$. The proof is in the Section C of the Appendix.*

Collection	Dataset	IDs	Images
	CUHK02	1,816	7,264
	CUHK03	1,467	14,097
MS	DukeMTMC-Re-Id	1,812	36,411
	Market-1501	1,501	29,419
	CUHK-SYSU	11,934	34,547

Table 1. Training Datasets Statistics. All the images in these datasets, regardless of their original train/test splits, are used for model training.

Dataset	Probe		Gallery	
	Pr. IDs	Pr. Imgs	Ga. IDs	Ga. imgs
PRID	100	100	649	649
GRID	125	125	1025	1,025
VIPeR	316	316	316	316
i-LIDS	60	60	60	60

Table 2. Testing Datasets statistics.

4. Experiments

4.1. Datasets and Settings

Following [52, 21], we evaluate the DIR-ReID with multiple sources (MS), where source domains cover five large-scale ReID datasets, including CUHK02 [25], CUHK03 [26], Market1501 [71], DukeMTMC-ReID [73], CUHK-SYSU PersonSearch [60]. Details of MS are summarized in Table 1. The unseen test domains are VIPeR [16], PRID [19], QMUL GRID [30] and i-LIDS [59]. We follow the single-shot setting, where the number of probe/gallery images are summarized in Table 2. The evaluation protocols can be found in Section D of the Appendix. The average rank-k (R-k) accuracy and mean Average Precision (mAP) over 10 random splits are reported based on the evaluation protocol. In this way, we simulate the real-world setting that a ReID model is trained with all the public datasets to evaluate the generalization capability to unseen domains.

4.2. Implementation Details

Following previous generalizable person ReID methods, we use MobileNetV2 [48] as the domain-specific encoder f_V and use MobileNetV2 with IN layer [29] as identity-specific encoder f_S. Our classifiers $\mathcal{C}_S, \mathcal{C}_V, \mathcal{C}_C^S, \mathcal{C}_C^V$ are simply composed of a single fully-connected layer. Images are resized to 256×128 and the training batch size is set to 128. Random cropping, random flipping, and color jitter are applied as data augmentations. The label smoothing parameter is 0.1. SGD is used to train all the components from scratch with a learning rate 0.02 and momentum 0.9. The training process includes 150 epochs and the learning rate is divided by 10 after 100 epochs. At test time, DIR-ReID only involves identity-specific encoder f_S, which is of a comparable network size to most ReID methods. The

tradeoff weights are set to $\lambda_2 = 0.1$ and $\lambda_1 = \lambda_3 = \lambda_4 = 1$ empirically.

4.3. Comparisons Against State-of-the-art

Comparison with single domain methods. Many supervised methods report high performance on large-scale benchmarks, but their performance is still poor on small-scale ones. We select 6 representative models (labeled as 'S' in Table 3) trained with data and labels the same as the training splits of target datasets. Although data from the target domain are inaccessible for DIR-ReID, it achieves competitive or better performance on all four benchmarks, which indicates that sufficient source data and our model based on domain invariance learning can alleviate the need for data from the target domain.

Comparison with cross-domain Re-ID methods. We also select representative cross-domain Re-ID methods for the comparison, which is shown in Table. 4. ("Market1501→DukeMTMC" indicates that Market-1501 is a labeled source domain and DukeMTMC-ReID is an unseen target domain). The setting of cross-domain Re-ID is challenging for us because there has only one source domain, the MDDAN block cannot work directly. Even though we see different cameras as different domains, the style variation within a single dataset is still small. Despite this, Table. 4 shows that our DIR-ReID framework achieved comparable performance on both settings. It indicates that, even with small domain discrepancy, DIR-ReID can improve the generalization capability.

Unfortunately, DIR-ReID still has a large margin between with current state-of-the-art domain adaptation methods (*e.g.* MMT [14] achieves more than 75% rank 1 accuracy in "Market-to-Duke" dataset setting, which is much superior than current DG methods). However, as discussed earlier, supervised methods require labeled target datasets and unsupervised domain adaptation methods require unlabeled images from the target domain, which both need additional training data collections. The expensive manual cost of data collections hinders the ReID models from real-world deployments. DIR-ReID attains superior performance without using any images or labels from the target domain and additional adaptation, which is crucial for practical ReID deployment.

Comparison with DG methods. Finally, we compare DIR-ReID with existing methods about generalizable person ReID. As far as we know, there are a few publications focusing on person ReID generalization problem, including DIMN [52], DualNorm [21], [55] and DDAN [7]. From the third row in Table 3, the DIR-ReID has achieved the best performance in terms of mAP against other SOTA DG-ReID methods. Although our method falls behind others on the i-LIDS and the GRID datasets in terms of Rank-5 and Rank-10, the DIR-ReID obtains the best Rank-1 per-

formance on three of four datasets. Interestingly, methods such as SNR [22] and AugMining [55] perform very well in i-LIDS, while having low performance in other datasets. It is not an exemplification of high generalization ability. Following [47], we consider the worst-domain accuracy (WDA), which is 32.3% rank 1 accuracy for AugMining, 29.3% rank 1 accuracy for DIMN, 40.2% rank 1 accuracy for SNR, 41.4% rank 1 accuracy for DualNorm, and 46.2% rank 1 accuracy for DDAN. Compared to them, the DIR-ReID has a superior WDA value, which is 47.8% rank 1 accuracy when tested on the GRID dataset.

4.4. Ablation Studies

There are two main components in the proposed DIR-ReID: the MDDAN block and the BA block. Here, we firstly analyze the effectiveness of the blocks respectively, then demonstrate their contributions to the final performance of the whole DIR-ReID model.

Analysis of MDDAN. The superiority of MDDAN is verified by comparison with the dual DANN [13] block. The latter means inserting the GRL layers between f_S, C_V and f_V, C_S. Simultaneously, the dual DANN block replaces the maximum entropy loss in MDDAN with the maximum misclassification loss. Results are shown in Table 5. The dual DANN method only has a slight improvement in comparison with the baseline, while our MDDAN block leads to significant improvements.
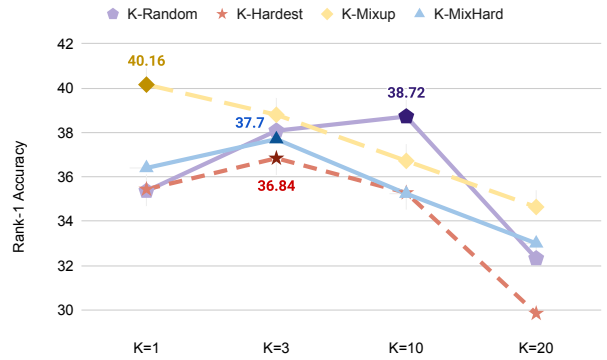

Figure 5. Ablation study on BA block. The metric is the rank-1 accuracy on the GRID dataset. Considering the expensive cost of training with five datasets, all the models here are trained on three datasets, *i.e.* Market-1501, CUHK02 and CUHK03.

Analysis of methods for backdoor adjustment. We compare four different implementations of BA and the results are shown in Figure 7. K-Random is the simplest method while it works well, attaining 38.72% rank 1 accuracy when $K = 10$. Surprisingly, K-Hardest is no better than K-Random. It may because that these selected hardest vectors may hinder the convergence speed of our

Methods	Type	Source	VIPeR (V)				PRID (P)				GRID (G)				i-LIDS (I)			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
DeepRank [8]	S	Target	38.4	69.2	81.3	-	-	-	-	-	-	-	-	-	-	-	-	-
DNS [67]	S	Target	42.3	71.5	82.9	-	29.8	52.9	66.0	-	-	-	-	-	-	-	-	-
MTDnet [10]	S	Target	47.5	73.1	82.6	-	32.0	51.0	62.0	-	-	-	-	-	58.4	80.4	87.3	-
JLML [67]	S	Target	50.2	74.2	84.3	-	-	-	-	-	37.5	61.4	69.4	-	-	-	-	-
SSM [3]	S	Target	53.7	-	91.5	-	-	-	-	-	27.2	-	61.2	-	-	-	-	-
SpindleNet [70]	S	Target	58.3	74.1	83.2	-	67.0	89.0	89.0	-	-	-	-	-	66.3	86.6	91.8	-
AugMining [55]	DG	MS	49.8	70.8	77.0	-	34.3	56.2	65.7	-	46.6	67.5	76.1	-	76.3	93.0	95.3	-
DIMN [52]	DG	MS	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
DualNorm [21]	DG	MS	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
SNR [22]	DG	MS	52.9	-	-	61.3	52.1	-	-	66.5	40.2	-	-	47.7	84.1	-	-	89.9
DDAN [7]	DG	MS	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
DIR-ReID	DG	MS	58.3	66.9	77.3	62.9	71.1	82.4	88.6	75.6	47.8	51.1	70.5	52.1	74.4	83.1	90.2	78.6

Table 3. Comparisons against state-of-the-art methods. 'S': single domain, 'DA': domain adaptation (cross-domain), 'DG': domain generalization, 'M': Market1501, 'D': DukeMTMC-ReID, 'comb': the combination of VIPeR, PRID, CUHK01, i-LIDS, and CAVIAR datasets. 'C3': CUHK03, '-': no report. 1st and 2^{ed} highest accuracy are indicated by **blue** and red color.

Method	Cross-domain Re-ID (single-source DG)							
	Market-Duke				Duke-Market			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
IBN-Net [38]	43.7	59.1	65.2	24.3	50.7	69.1	76.3	23.5
OSNet [74]	44.7	59.6	65.4	25,9	52.2	67.5	74.7	24.0
OSNet-IBN [74]	47.9	62.7	68.2	27.6	57.8	74.0	79.5	27.4
CrossGrad [50]	48.5	63.5	69.5	27.1	56.7	73.5	79.5	26.3
QAConv [27]	48.8	-	-	28.7	58.6	-	-	27.6
L2A-OT [75]	50.1	64.5	70.1	29.2	63.8	80.2	84.6	30.2
OSNet-AIN [74]	52.4	66.1	71.2	30.5	61.0	77.0	82.5	30.6
SNR [22]	55.1	-	-	33.6	66.7	-	-	33.9
DIR-ReID	54.5	66.8	72.5	33.0	68.2	80.7	86.0	35.2

Table 4. performance (%) comparison with the state-of-the-arts on the cross-domain Re-ID problem. 1st and 2^{ed} highest accuracy are indicated by **blue** and red color.

	PRID	VIPeR	i-LIDS	GRID
Baseline (DualNorm [21])	38.9	54.0	64.3	34.4
w/ Dual DANN [13]	41.0	53.3	66.2	35.4
Improvements↑	2.1	-0.7	1.9	1.0
w/ MDDAN	42.7	56.5	65.7	36.6
Improvements↑	3.8	2.5	1.4	2.2
w/ BA	43.8	61.4	64.2	40.2
Improvements↑	4.9	7.4	-0.1	5.8

Table 5. Ablation studies on three different blocks: dual DANN block, our MDDAN block and the BA block. The metric is the rank-1 accuracy. Models here are all trained on three datasets, including Market-1501, CUHK02 and CUHK03. The improvements denote the difference from the baseline.

optimization process. K-Mixup outperforms other methods and attains 40.17% rank 1 accuracy, which verified the importance of mixup for data augmentation. However, as we increase the value of K, the performance will not be improved. When we combine K-Hardest and mixup, K-MixHard has a slightly higher performance than K-Hardest, while it still has a large margin compared to K-MixUp.

Ablation study of different blocks. To evaluate the contribution of each component, we gradually add the MDDAN and the BA to the baseline, and the overall ablation studies are reported in Table 5. The MDDAN improves the rank-1 accuracy from 34.4% to 36.6% in the GRID dataset. The

results in PRID, VIPeR, and i-LIDS datasets are consistently improved, which validate that the MDDAN removes some of the domain-specific information from our identity-specific representations and yields consistent generalization performance improvements. The BA provides greater improvement gains in the GRID dataset (5.8% higher rank-1 accuracy). It validates that BA can exclude domain-specific information efficiently. Unfortunately, the DIR-ReID seems not to work well on i-LIDS dataset, we will explore the reason and make improvements in future work.

We also conduct additional ablation studies and disentanglement experiments on rotated MNIST. The dataset and results in Section E of the Appendix.

5. Conclusions and Future Work

We propose a novel generalizable person ReID framework based on disentanglement and backdoor adjustment from a causal invariant learning perspective. Specifically, a MDDAN block is proposed to disentangle identity-specific and domain-specific factors from multi-source ReID training data. We then propose BA block to learn interventional distribution and reduce the confounding effects via backdoor adjustment. The comprehensive experimental results show that DIR-ReID achieves state-of-the-art performance.

To upgrade DIR-ReID, we can improve the model performance with other tricks *i.e.* triplet loss [31], Maximum Mean Discrepancy (MMD) [17] regularization. One promising way is to generate realistic images from concatenated vectors. The augmented feature vectors are guided by a reconstruction loss, which will further improve the disentanglement of identity-specific and domain-specific factors. These generated images can also be used for interpretability and augmenting the training set. Besides, we will seek other methods for better disentanglement performance such as replacing the multi-domain adversarial learning with mutual information minimization [4] or f-divergence [37] maximization. We believe DIR-ReID may promote the development of new paradigms for DG beyond the person ReID.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 2

[2] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 3

[3] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 8

[5] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[6] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, page 2060. NIH Public Access, 2019. 3, 5

[7] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. 2, 7, 8, 14

[8] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016. 8

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 2

[10] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3988–3994, 2017. 8

[11] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *NeurIPS*, 2020. 3

[12] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 5298–5309, 2019. 1

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 5, 7, 8, 13, 14

[14] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 2, 7

[15] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 13

[16] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007. 6, 13

[17] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS*, volume 23, pages 673–681, 2009. 8

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 2

[19] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 6, 13

[20] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14094, 2020. 2

[21] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422*, 2019. 2, 6, 7, 8

[22] 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. 2, 7, 8

[23] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012. 2

[24] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017. 2

[25] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 6

[26] 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 (CVPR)*, June 2014. 6

[27] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. *arXiv preprint arXiv:1904.10424*, 2019. 8

[28] Shan Lin, Chang-Tsun Li, and Alex C Kot. Multi-domain adversarial feature generalization for person re-identification. *arXiv preprint arXiv:2011.12563*, 2020. 2

[29] Shan Lin, Haoliang Li, Chang Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *29th British Machine Vision Conference, BMVC 2018*, 2018. 6

[30] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision*, pages 391–401. Springer, 2012. 6, 13

[31] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing (TIP)*, 26(7):3492–3506, 2017. 8

[32] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[33] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10846–10856, 2018. 3

[34] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372, 2016. 2

[35] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019. 4

[36] Leland Gerson Neuberg. Causality: Models, reasoning, and inference, 2003. 2, 3

[37] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016. 8

[38] 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. 8

[39] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *ICML*, 2018. 3

[40] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. 12

[41] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 2000. 3

[42] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 3

[43] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[44] Xingchao Peng, Yichen Li, and Saenko Kate. Domain2vec: Domain embedding for unsupervised domain adaptation. *European conference on computer vision (ECCV 2020)*, 2020. 2

[45] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 12

[46] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization, 2020. 3

[47] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 7

[48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[49] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*, 2021. 3

[50] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 8

[51] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 2

[52] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 6, 7, 8

[53] Xinwei Sun, Botong Wu, Chang Liu, Xiangyu Zheng, Wei Chen, Tao Qin, and Tie-yan Liu. Latent causal invariant model. *arXiv preprint arXiv:2011.02203*, 2020. 3

[54] 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)*, September 2018. 2

[55] Masato Tamura and Tomokazu Murakami. Augmented hard example mining for generalizable person re-identification. *arXiv preprint arXiv:1910.05280*, 2019. 7, 8

[56] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 3

[57] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 2

[58] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[59] Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. Associating groups of people. In *Proceedings of the British Machine Vision Conference*, pages 23–1, 2009. 6, 13

[60] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2(2), 2016. 6

[61] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 3

[62] Mang Ye, Jianbing Shen, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 5

[63] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3589–3598, 2020. 2, 14

[64] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020. 3, 5, 12

[65] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. *arXiv preprint arXiv:2007.01546*, 2020. 2

[66] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[67] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 8

[68] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019. 2

[69] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[70] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

[71] 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 (ICCV)*, December 2015. 6

[72] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[73] 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 (ICCV)*, Oct 2017. 6

[74] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 8

[75] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, pages 561–578. Springer, 2020. 8

A. Derivation of the Interventional Distribution $P(Y|do(S))$ for the Proposed Causal Graph

The following proof is similar to [64] with three rules of *do*-calculus [40]: Insertion/deletion of observations, Action/observation exchange and Insertion/deletion of actions.

In our causal formulation, the desired interventional distribution $P(Y|do(S = s))$ can be derived by:

$$
\begin{aligned}
P(Y|do(S)) &= \sum_v P(Y|do(S = s), V = v)P(V = v|do(S = s)) \\
&= \sum_v P(Y|do(S = s), V = v)P(V = v) \\
&= \sum_v P(Y|S = s, V = v)P(V = v),
\end{aligned}
$$

(17)

where line 1 follow the law of total probability; line 2 uses Rule 3 given $S \perp\!\!\!\perp V$ in $\mathcal{G}_{\overline{X}}$; line 3 uses Rule 2 to change the intervention term to observation as $(Y \perp\!\!\!\perp S|V)$ in $\mathcal{G}_{\underline{X}}$.

B. Analysis of the Multi-Domain Disentangled Adversarial Neural Network (MDDAN)

Here we adapt from [45] and prove the Lemma 1 in the original manuscript. All the analyses are conducted in the shared ID-specific representation space. We first prove the following lemma.

Lemma B.1 *When MDDAN is carried out and the uncertainty is maximized, given any ID-specific representation $\mathbf{s}^\tau$ of an image from domain $\mathcal{D}_\tau$, the conditional probability of any $\mathcal{D}_i$ given $\mathbf{s}^\tau$ will equal $1/K$, namely for any $i = 1, ..., K$, we have $p(\mathcal{D}_i|\mathbf{s}^\tau) = 1/K$.*

Proof B.1 *Here we slightly abuse the notation: omitting the domain classifier $\mathcal{C}_V$ and $p(k|\mathbf{s}^\tau)$ is equivalent to $p(\mathcal{D}_k|\mathbf{s}^\tau)$. MDDAN is defined as follows.*

$$
\begin{aligned}
&\max - \sum_{k=1}^K p(k|\mathbf{s}^\tau) \log p(k|\mathbf{s}^\tau), \\
&s.t. \sum_{k=1}^K p(k|\mathbf{s}^\tau) = 1; p(k|\mathbf{s}^\tau) \geq 0, \forall k.
\end{aligned}
$$

(18)

Let $\mathcal{H}(p_k) = p_k \log p_k$, where $p_k = p(k|\mathbf{s}^\tau)$. We simplify the above optimization problem as

$$
\min \sum_{k=1}^K \mathcal{H}(p_k), \ s.t. \sum_{k=1}^K p_k = 1; p_k \geq 0, \forall k,
$$

(19)

where $\mathcal{H}(p_k)$ is convex as we have $\mathcal{I}''(p_k) = \frac{1}{p_k \ln 2} > 0$. The sum of convex functions $\sum_{k=1}^K \mathcal{H}(p_k)$ is also a convex

function. Namely, this problem is a convex optimization problem. To prove the lemma, it is now equivalent to show that the minimum value of this convex optimization problem is obtained when $p_1 = p_2 = ... = p_K = 1/K$.

We use the augmented Lagrangian method to solve the problem, where the Lagrange function is defined as

$$
L(p, \lambda) = \sum_{k=1}^K p_k \log p_k + \lambda(\sum_{k=1}^K p_k - 1)).
$$

(20)

We take partial derivatives for each p_k and get

$$
\frac{\partial L(p, \lambda)}{\partial p_k} = \log p_k + \frac{1}{\ln 2} + \lambda = 0.
$$

(21)

Then we have $p_k = 2^{-\lambda - 1/\ln 2}$. As $\sum_{k=1}^K p_k = 1$, we have $K * 2^{-\lambda - 1/\ln 2} = 1$, and thus $p_k = 1/K, i = 1, 2, ..., K$. Since the local minimum of the convex function is the global minimum, when $p_1 = p_2 = ... = p_K = 1/K$, $\sum_{k=1}^K \mathcal{H}(p_k)$ achieves the minimum value, $\log \frac{1}{K}$. In other words, when MDDAN is carried out and achieves the maximum uncertainty, for all $i = 1, ..., K$, we have $p(\mathcal{D}_i|\mathbf{s}^\tau) = 1/K$.

Now we are ready to prove Lemma 1. we can calculate any domain $\mathcal{D}_i$'s conditional probability given $\mathbf{s}^\tau$, which is

$$
p(\mathcal{D}_i|\mathbf{s}^\tau) = \frac{p(\mathbf{s}^\tau|\mathcal{D}_i)p(\mathcal{D}_i)}{p_\tau(\mathbf{s}^\tau)}, \forall \mathbf{s}^\tau \in \mathcal{D}_\tau; i = 1, ..., K,
$$

(22)

where $p(\mathbf{s}^\tau|\mathcal{D}_i)$ denotes the conditional probability of $\mathbf{s}^\tau$ given domain information $\mathcal{D}_i$, $p_\tau(\mathbf{s}^\tau)$ denotes the probability function of the ID-specific representation in its domain $\mathcal{D}_\tau$, and $p(\mathcal{D}_i)$ is the priori probability of domain classes. Without generality, we set equal priori probability for each domain, namely $p(\mathcal{D}_i) = 1/K$. Further, from the Lemma B we know that, optimizing MDDAN leads to $p(\mathcal{D}_i|\mathbf{s}^\tau) = 1/K$ for all $i = 1, ..., K$. Hence, Eq.(22) becomes

$$
\begin{aligned}
&1/K = \frac{p(\mathbf{s}^\tau|\mathcal{D}_i)1/K}{p_\tau(\mathbf{s}^\tau)}, \forall \mathbf{s}^\tau \in \mathcal{D}_\tau; i = 1, ..., K, \\
&\Rightarrow p(\mathbf{s}^\tau|\mathcal{D}_i) = p_\tau(\mathbf{s}^\tau), \forall \mathbf{s}^\tau \in \mathcal{D}_\tau; i = 1, ..., K,
\end{aligned}
$$

(23)

thus completes the proof.

C. Understanding MDDAN by Information Theory

Minimizing the mutual information between the ID-specific factors S and the domain information D is defined as

$$
\begin{aligned}
\min \mathcal{I}(S, D) &= \min \mathcal{H}(D) - \mathcal{H}(D|S) \\
&= \min -\mathcal{H}(D|S) \\
&= \max \mathcal{H}(D|S)
\end{aligned}
$$

(24)

The second line is derived since the entropy of domain distribution $\mathcal{H}(D)$ is not related to our optimization. Namely, our MDDAN is essentially minimizing the mutual information between the ID-specific factors and the domain information.

D. Evaluation Protocols.

Evaluation protocols are as follows.

GRID [30] contains 250 probe images and 250 true match images of the probes in the gallery. Besides, there are a total of 775 additional images that do not belong to any of the probes. We randomly take out 125 probe images. The remaining 125 probe images and $775 + 250$ images in the gallery are used for testing.

i-LIDS [59] has two versions, images and sequences. The former is used in our experiments. It involves 300 different pedestrian pairs observed across two disjoint camera views 1 and 2 in public open space. We randomly select 60 pedestrian pairs, two images per pair are randomly selected as probe image and gallery image respectively.

PRID2011 [19] has single-shot and multi-shot versions. We use the former in our experiments. The single-shot version has two camera views A and B, which capture 385 and 749 pedestrians respectively. Only 200 pedestrians appear in both views. During the evaluation, 100 randomly identities presented in both views are selected, the remaining 100 identities in view A constitute probe set and the remaining 649 identities in view B constitute gallery set.

VIPeR [16] contains 632 pedestrian image pairs. Each pair contains two images of the same individual seen from different cameras 1 and 2. Each image pair was taken from an arbitrary viewpoint under varying illumination conditions. To compare to other methods, we randomly select half of these identities from camera 1 as probe images and their matched images in 2 as gallery images.

E. Ablation studies on Rotated MNIST.

E.1. Dataset and Setting

To verify whether the capability of DIR-ReID model to disentangling s and v, we first construct rotated MNIST datasets following [15]. 100 images per class (10 classes totally) are randomly sampled from the MNIST training dataset, which is denoted by $\mathcal{M}_{0°}$. We then rotated the images in $\mathcal{M}_{0°}$ by $15, 30, 45$, and 60 degrees, creating four additional domains. To plot the latent subspaces directly without applying dimensionality reduction, we restrict the size of each latent space to 2 dimensions. In experiments, we train the MDDAN block with an additional reconstruction loss, which is enough to attain a encouraging result.

block	details
1	Conv2d(32, 5), BatchNorm2d, ReLU
2	MaxPool2d(2, 2)
3	Conv2d(64, 5), BatchNorm2d, ReLU
4	MaxPool2d(2, 2)
5	Linear(2)

Table 6. Architecture for encoders f_S, f_V. The parameters for Conv2d are output channels and kernel size. Theparameters for MaxPool2d are kernel size and stride. The parameter for Linear is output features.

block	details
For $\mathcal{C}_V$	ReLU,Linear(5)
For $\mathcal{C}_S$	ReLU,Linear(10)
For $\mathcal{C}_C^S$	ReLU,Linear(5)
For $\mathcal{C}_C^v$	ReLU,Linear(5)

Table 7. Architecture of classifiers for ID-specific factors, Domain-Specific factors, and concatenated vectors. The parameter for Linear is output features.

	Test Accuracy on $\mathcal{M}_{75°}$
Baseline	46.4
w/ Dual DANN [13]	53.5
Improvements↑	7.1
w/ MDDAN	58.2
Improvements↑	11.8
w/ BA	61.9
Improvements↑	15.5

Table 8. Ablation study of three different blocks: dual GRL block, our MDDAN block and the BA block. The reported validation metric is the accuracy of the $\mathcal{M}_{75°}$ dataset. The baseline is only using the encoder f_S and classifier $\mathcal{C}_S$.

E.2. Architecture and Implementation Details

The architectures of the encoders, classifiers are displayed in Table 6, Table 7 respectively. All methods are trained for 500 epochs and the batch size is set to 100. Adam is used to train all the components from scratch with learning rate as 0.001. We also use warmup to linearly increase the learning rate from 0 to 0.001 during the first 100 epochs of training.

E.3. Additional Experimental Results

Analysis of MDDAN. As shown in Table 5, directly applying the dual GRL block benefits the generalization ability, while the proposed MDDAN improve the test accuracy on $\mathcal{M}_{75°}$ dataset by a even more large margin, which is 11.8% points.

Analysis of methods for backdoor adjustment. The comparison results are shown in Figure 7. Different from BA for Re-ID, here K-MixHard attains the most superior performance, which is 61.9% test accuracy. In general,
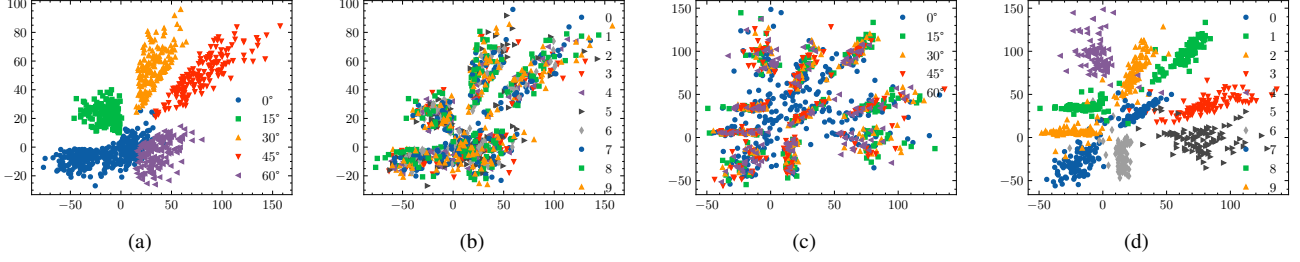
Figure 6. 2D embeddings of all two latent subspaces. (a,b): v encoded by f_V. (c,d): s encoded by f_S. (a,c) are colored according to their domains and (b,d) are colored according to their classes.
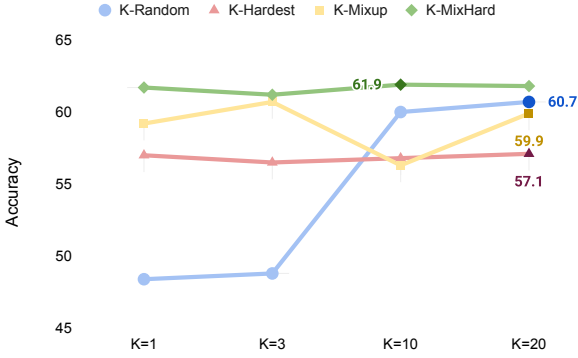

Figure 7. Ablation study of backdoor adjustment methods. The reported validation metric is the test accuracy of the $\mathcal{M}_{70\circ}$ dataset.

$K-$Random can attain an encouraging performance gain and we can choose other kinds of methods and parameters based on the difficulty of the dataset.

Ablation study of different blocks. By adding the multi-domain disentangled block and the backdoor adjustment block successively, we improve the generalization accuracy from 46.4% to 58.2% and 61.9% respectively (Table 8), showing the effectiveness of the proposed model again.

Disentanglement on rotated MNIST. The disentanglement results are demonstrated in Figure 6. We can find a correlation between the rotation angle (domain labels) and v in Figure 6(a), five domains are clustered into five distinct clusters, while in Figure 6(b) no clustering is visible, namely there is very weak correlation between v and class labels. By contrast, in Figure 6(c) no clustering is visible according to the rotation angle. But Figure 6(d) shows ten distinct clusters, where each cluster corresponds to a class. From these initial qualitative results, we conclude that the dual MDDAN is disentangling the information contained in x as intended. v only contains domain information and s only contains identity information. The great disentangle results ensure the backdoor adjustment can be carried out effectively.

F. Comparision with Similar Adversarial Learning-Based Methods

Here we discuss similar works DANN [13], DDAN [7] and CaNE [63], which use adversarial training to reduce domain divergence or nuisance divergence. There are three differences: (i) The applications are different. Given n domains, DANN needs n binary classifiers to check which domain one images belongs to. DDAN selects one central domain, using one binary classifier to check if one image belongs to the central (1) or peripheral (0) domain. CaNE implements adversarial training in nuisance attributes (camera ID and video timestamps). We directly use entropy maximization with multi-domain classification results, which extends the assumption of binary classification in DANN. (ii) The goals are different. Our DIR-ReID is an algorithmic implementation of backdoor adjustment and adversarial training block (MDDAN) is just one step of our method. We use adversarial training for simultaneously reducing domain divergence and disentangling semantic/variation factors. (iii) We show the underlying optimization objective of entropy maximization is highly correlated with fusing domain information (Appendix. B). Secondly, we observe a correlation between MDDAN optimization with mutual information minimization (Appendix. C).