GENERALIZABLE PERSON RE-IDENTIFICATION WITHOUT DEMOGRAPHICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable Person Re-Identification (DG ReID) aims to learn ready-to-use cross-domain representations for direct cross-data evaluation. It typically fully *exploit demographics information*, *e.g.,* the domain information and camera IDs to learn features that are domain-invariant. However, the protected demographic features are not often accessible due to privacy and regulation issues. Under this more realistic setting, distributionally robust optimization (DRO) provides a promising way for learning robust models that are able to perform well on a collection of possible data distributions (the "uncertainty set") without demographics. However, the convex condition of KL DRO may not hold for overparameterized neural networks and applying KL DRO fails to generalize under distribution shifts in real scenarios. Instead, by applying the change-of-measure technique and the analytical solution of KL DRO, we propose a simple yet efficient approach, **Unit DRO**. Unit DRO minimizes the loss over a reweighted dataset where important samples (*i.e.,* samples on which models perform poorly) will be upweighted and others will be downweighted. Empirical results show that Unit DRO achieves superior performance on large-scale DG ReID and cross-domain ReID benchmarks compared to standard baselines.

1 INTRODUCTION

Person Re-IDentification (ReID) aims at matching person images of the same identity across multiple camera views. In previous work, ReID models mainly follow three settings: (i) As shown in Figure 1a, most ReID models are trained and tested on *i.i.d* datasets, termed **fully-supervised methods** (Zhang et al., 2020). Although recent fully-supervised methods have achieved remarkable performance, they are non-robust when tested in out-of-distribution (OOD) settings. (ii) Figure 1b illustrates the settings of **unsupervised domain adaptation (UDA) methods and cross-domain (CD) person ReID** methods (Luo et al., 2020). However, UDA ReID relies on large amounts of unlabeled data for retraining and CD ReID cannot exploit the benefits brought by multi source domains. These problems severely hinder real-world applications of current person ReID techniques. Recently, (iii) **generalizable person ReID methods (DG)** (Dai et al., 2021a) are proposed (Figure 1c) in a more realistic setting, where the model is trained on multiple large-scale datasets. The trained model is tested on unseen domains directly without any data collection, annotation, and model updating.

However, generalizable person ReID methods come at a serious disadvantage: they require demographics (*e.g.,* domain labels (Choi et al., 2021; Zhao et al., 2021), camera ID (Zhang et al., 2021a; Dai et al., 2021a) and video timestamps (Yuan et al., 2020)) as extra supervision. Such demographics implicitly define variations in training data that the learned models should be invariant or robust to[1]. However, such demographics usually are not available to use for the following reasons: (i) The collection of demographics inevitably creates privacy risks (Veale & Binns, 2017), *e.g.,* exposing the geographical location and environment information. ReID is often used for high-privacy tasks such as security, on which the privacy disclosure is unacceptable. (ii) domain labels collection are expensive and ethically fraught endeavours (Michel et al., 2021), and (iii) manually collected domain labels may be noisy or suboptimal (Creager et al., 2021) and such coarsed grained labels may exacerbate *hidden stratification*, which hinders safe-critical applications (Oakden-Rayner et al., 2020). We aim to overcome the difficulty of manual demographics collection by *developing a new setting without*

[1]The commonly used paradigm is to enforce representations to be invariant to domain labels.

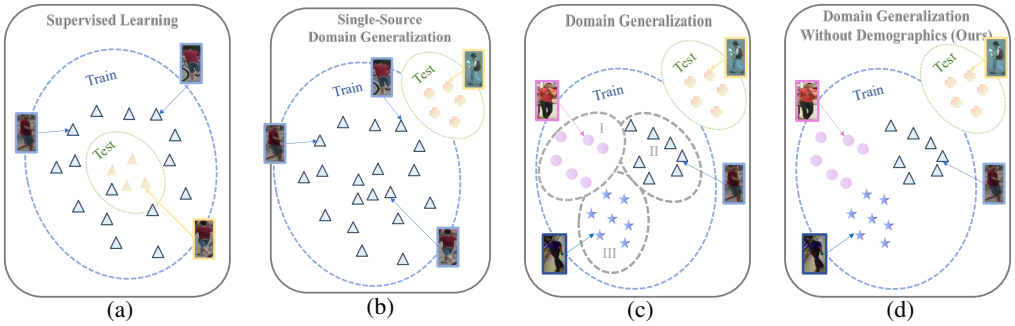


Figure 1: Universal person ReID settings. (a) Supervised person ReID. (b) CD ReID and UDA ReID. (c)DG ReID. (d) Illustration of our setting for generalizable person ReID without demographics.

the need for demographics. Figure 1d depicts the **D**omain **G**eneralizable Person Re-identification **W**ithout **D**emographics (DGWD) setting, where models are also trained on multiple large-scale datasets while the demographics are unavailable.

DRO is a promising paradigm to tackle the problem above by explicitly obtaining prediction functions robust to distribution shifts (Hu et al., 2018). Specifically, DRO considers a minimax game: the inner optimization objective is to shift the training distribution within a pre-specified uncertainty set so as to maximize the expected loss on the test distribution. The outer optimization minimizes the adversarial expected loss (see Section 3.1 for details). DRO with f-divergences has been well studied, which defines the uncertainty set by an f-divergence ball from the training distribution Hu & Hong (2013). However, the convex assumption usually does not hold in real scenarios, which leads to inferior performance in the context of overparameterized neural networks.

In this paper, we first solve the inner step optimization problem and obtain a closed-form expression of the optimal objective. Different from previous work that converts the minimax DRO problem into a single minimization problem by the closed-form expression (Hu & Hong, 2013), we implement a change-of-measure technique and reformulate the minimax optimization as an importance sampling problem, termed **Unit DRO**[2]. Unit DRO avoids the troublesome bi-level optimization in traditional DRO problems and scales well to over-parameterized regimes. Specifically, Unit DRO upweights samples which are prone to be misclassified and downweigts others. It assigns a normalized weight $e^{\ell/\tau^*}/\mathbb{E}[e^{\ell/\tau^*}]$ to each data and label pair (x, y), where ℓ is the error incurred by (x, y) and τ^* is a hyperparameter. There are two main challenges here, (i) The optimization parameter τ^* is hard to determine and we observe that a constant τ^* always achieves inferior performance; (ii) The normalization factor $\mathbb{E}[e^{\ell/\tau^*}]$ requires taking an expectation over the training distribution. To tackle the first problem, we propose **step** τ^* to determine the value of τ^* by the training step. We then maintain a **weights queue** which stores historical sample weights to better estimate $\mathbb{E}[e^{\ell/\tau^*}]$ over the training distribution. Compared to DG ReID methods, the implementation of Unit DRO is simple yet effective, avoiding the need for meta-learning pipelines or complicated model structure engineering.

We empirically evaluate and analyze the proposed implementation. First, we compare Unit DRO with both CD and DG methods. Unit DRO achieves improved performance by a large margin on DG and CD benchmarks even compared to these methods that rely on demographics. Second, we take comprehensive ablation studies of the step τ^* and the weights queue, providing justification for these two blocks. Finally, we visualize the learned weight distributions, t-SNE embeddings, and measure the domain divergence and error set to show the invariant learning capability of Unit DRO. Empirical results show that Unit DRO can retrieve valuable samples or subgroups without demographics.

2 Related Work

Domain generalization. Domain/Out-of-distribution generalization (Muandet et al., 2013) aims to learn a model that can extrapolate well in unseen environments. Representative methods like Invariant

[2]The name "Unit" is a contrast to "Group". Group DRO (Sagawa et al., 2019) assigns weights for every domain and our proposed Unit DRO assigns weights for every sample.

Risk Minimization (IRM) (Arjovsky et al., 2020) and its variant (Ahuja et al., 2020) are recently proposed to tackle this challenge. IRM center on the objective of extracting data representations that lead to invariant prediction across environments under a multi-environment setting. The main difference here is that we propose to learn invariant representations without demographics.

Generalizable Person ReID. Generalizable person ReID methods (Song et al., 2019; Choi et al., 2021) are recently proposed to learn invariant representations that can generalize to unseen domains. Existing methods mainly utilize domain divergence minimization strategies or a meta-learning pipeline. DualNorm(Jia et al., 2019) integrate the Instance Normalization (IN) into the network to filter out style factors, boosting generalization capability. Other works aim to learn domain-invariant features, *e.g.,* (Chen et al., 2021) and (Zhang et al., 2021a). However, neither meta learning-based methods nor domain divergence minimization strategies work properly without demographics.

Fairness without demographics. Methods in Fairness (Dwork et al., 2012) aim to develop a model that performs well for worst-case group assignments according to some fairness criteria for addressing the underperformance in minority subgroups. Although there are some works consider *fariness without demographics* (Liu et al., 2021; Creager et al., 2021), they mostly evaluate their algorithms in datasets with predefined distribution shifts. Note that DGWD-ReID problem is more challenging than the category-level recognition problem considered in the existing *fariness w or w/o demographics* study. In DGWD-ReID, the target identities are different from source ones and we need to tackle both domain gap and disjoint label space problems at the same time.

3 METHODOLOGIES

Notations and Problem Formulation. Consider current DG setting, where we have access to one labeled dataset which consists of several distinct training[3] distributions (domains): $\mathcal{P} = \{P_k\}_{k=1}^{|\mathcal{P}|} = \{\{x_i, y_i\}\}_{i=1}^{|P_k|}\}_{k=1}^{|\mathcal{P}|}$, where $|\mathcal{P}|$ is the number of domains, $|P_k|$ is the number of images in domain P_k, $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ is the image and the corresponding label. In the training phase, we train a DG model using all the aggregated image-label pairs. In the testing phase, we perform a retrieval task on the unseen target domain G without additional model updates. Our goal is to learn a model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, parameterized by $\theta \in \Theta$, that minimizes the error in G:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\in G} \left[\ell(x, y; \theta)\right]. \tag{1}$$

This objective encodes the goal of learning a model that does not depend on spurious correlations (*e.g.,* domain-specific information). If a model makes decisions according to domain-specific information, it is natural to be brittle in an entirely distinct domain. Previous studies mostly leverage demographics (*e.g.,* domain IDs, camera IDs, video timestamps) to clip the spurious correlations. In this paper, we consider a novel setting where all of these demographics are not known during training, which makes empirical sense that annotating demographics is expensive and likely to expose private information.

Baseline Algorithms. Here we describe the learning objectives used in the baseline model. The first is the cross-entropy loss $\mathcal{L}_{ce}$, which seeks to minimize the average ID-classification loss over all the training samples. Given n training points $\{(x_1, y_1), ..., (x_n, y_n)\}$, $\mathcal{L}_{ce}$ is defined as follows:

$$\mathcal{L}_{ce} = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta) \tag{2}$$

The label-smoothing method is applied to prevent our model from overfitting to the training IDs.

Besides, following most of ReID methods, we introduce triplet loss to enhance the intra-class compactness and inter-class separability in the Euclidean space. Following (Hermans et al., 2017), given Euclidean distance $d(\cdot, \cdot)$ and an anchor sample x_i^a, we select the hardest positive sample x_i^p and the hardest negative sample x_i^n within a mini-batch. The triplet loss then can be defined as:

$$\mathcal{L}_{tr}(x_i^a) = \max\left\{d(f_\theta(x_i^a), f_\theta(x_i^p) - d(f_\theta(x_i^a), f_\theta(x_i^n)) + m, 0\right\}, \tag{3}$$

where m is the margin parameter. The BNNeck structure (Luo et al., 2019) is used to maximize the synergy between $\mathcal{L}_{ce}$ and $\mathcal{L}_{tr}$. Meanwhile, we integrate the mixture of Batch Normalization and Instance Normalization with learnable parameters (Choi et al., 2021) in the baseline, which has proved very useful for the DG problem.

[3]We use training distributions, empirical distributions interchangeably.

3.1 UNIT DRO

We now propose Unit DRO, a novel generalization framework that does not require priors about demographics. We begin from an effective algorithmic framework: distributionally robust optimization (DRO) (Ben-Tal et al., 2009; Rahimian & Mehrotra, 2019). In DRO, we use the worst-case expected risk over a predefined family of distributions $\mathcal{Q}$ (termed *uncertainty set*) to replace the expected risk under an unseen target distribution G in equation 1. Hence, the target is as follows,

$$\min_{\theta\in\Theta}\max_{q\in\mathcal{Q}}\mathbb{E}_{(x,y)\in q}[\ell(x,y;\theta)]. \tag{4}$$

The uncertainty set $\mathcal{Q}$ encodes the possible test distributions that we want our model to perform well on. If $\mathcal{Q}$ contains G, the DRO object can upper bound the expected risk under G. An important question about DRO modeling is how to choose the uncertainty set (see Appendix.A for details). Note that in many practical situations, we can obtain only the empirical (training) data distribution. Then, a reasonable approach is to construct the uncertainty set by requiring the distribution within a certain distance from the training distribution. Previous work chooses a KL-divergence ball (Hu & Hong, 2013)/MMD ball (Sinha et al., 2017) around the training distribution, which confers robustness to a wide set of distributional shifts. However, it can also lead to overly pessimistic models which optimize for implausible worst-case distributions (Duchi et al., 2019). In other words, $\mathcal{Q}$ should be sufficiently large to contain G, but if it is too large it may contain *noisy* distributions where no model can perform well (Michel et al., 2021). Group DRO (Sagawa et al., 2019) leverages demographics to define the uncertainty set $\mathcal{Q}$ and attains superior OOD performance. Here we consider a natural extension to improve OOD generalization in the DRO framework without demographics.

Construction of the uncertainty set based on the KL-divergence ball. In this paper, we construct $\mathcal{Q}$ as a KL-divergence ball around the empirical distribution $\mathcal{P}$. Given a KL upper bound (radius) η, we can formulate the uncertainty set as $\mathcal{Q}=\{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta\}$[4]. Then the min-max problem in equation 4 can be reformulated as

$$\min_{\theta\in\Theta}\max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta}\mathbb{E}_{(x,y)\in Q}\left[\ell(x,y;\theta)\right]. \tag{5}$$

Lemma 1 *(Modified from (Hu & Hong, 2013), Section 2) Assume the model family $\theta\in\Theta$ and $\mathcal{Q}$ to be convex and compact. The loss ℓ is continuous and convex for all $x\in\mathcal{X}, y\in\mathcal{Y}$. Suppose empirical distribution $\mathcal{P}$ has density $p(x,y)$. Then the inner maximum of equation 5 has a closed-form solution*

$$q^*(x,y)=\frac{p(x,y)e^{\ell(x,y;\theta)/\tau^*}}{\mathbb{E}_{\mathcal{P}}\left[e^{\ell(x,y;\theta)/\tau^*}\right]}, \tag{6}$$

where τ^* satisfies $\mathbb{E}_{\mathcal{P}}\left[\frac{e^{\ell(x,y;\theta)/\tau^*}}{\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]}\left(\frac{\ell(x,y;\theta)}{\tau^*}-\log\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]\right)\right]=\eta$ and $q^*(x,y)$ is the optimal density of Q. Then the min-max problem in equation 5 is equivalent to

$$\min_{\theta\in\Theta,\tau>0}\tau\log\mathbb{E}_{\mathcal{P}}\left[e^{\ell(x,y;\theta)/\tau}\right]+\eta\tau. \tag{7}$$

Reformulate KL DRO to Unit DRO. We name equation 7 **KL DRO**. Unfortunately, the convex condition of KL DRO is not held for over-parameterized neural networks, such that applying KL DRO often fails to generalize under distribution shifts in real scenarios. Therefore, we do not follow KL DRO that uses the inner maximum directly. Instead, we reformulate equation 5 as follows.

$$\begin{aligned}
\min_{\theta\in\Theta}\max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta}\mathbb{E}_{(x,y)\in Q}[\ell(x,y;\theta)] &= \min_{\theta\in\Theta}\max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta}\int\ell(x,y;\theta)q(x,y)d_xd_y\\
&= \min_{\theta\in\Theta}\max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta}\int\ell(x,y;\theta)\frac{q(x,y)}{p(x,y)}p(x,y)d_xd_y\\
&= \min_{\theta\in\Theta}\max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta}\mathbb{E}_{(x,y)\in\mathcal{P}}\left[\frac{q(x,y)}{p(x,y)}\ell(x,y;\theta)\right]\\
&= \min_{\theta\in\Theta}\mathbb{E}_{(x,y)\in\mathcal{P}}\left[\frac{e^{\ell(x,y;\theta)/\tau^*}}{\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]}\ell(x,y;\theta)\right]
\end{aligned} \tag{8}$$

[4]The $\mathcal{Q}$ mentioned below is the Q with KL constraint by default.

To get the third lines, we apply the change-of-measure technique. The fourth line replaces the inner maximum by its closed-form solution $q^*(x, y)$ in equation 6. Note that both the value of τ^* and the normalizer $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$ depend on the expectation of losses over the entire training data, which is untrackable at each optimization step. For simplicity, we can serve τ^* as a hyper-parameter and take the average over each mini-batch as a preliminary estimator of the normalizer. We term the resulting formulation **Unit DRO v1**.

$$\mathcal{L}_{\textbf{Unit DRO v1}}(\theta, \tau^*) = \min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{e^{\ell(x,y;\theta)/\tau^*}}{\frac{1}{N}\sum_{i=1}^{N}\left(e^{\ell(x,y;\theta)/\tau^*}\right)} \ell(x, y; \theta) \right), \tag{9}$$

where N is the batch size. In practice, Unit DRO v1 does not perform very well. The following problems and solutions are depicted point by point.

Step τ^*. The first problem is that a constant hyper-parameter τ^* is sub-optimal for the learning process. We visualize the densities of the weights $e^{\ell(x,y;\theta)/\tau^*}/\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$ over different learning steps with constant τ^* values in Figure 2 (The detailed experimental setting is in Section 4.3). A small τ^* leads to weights distribution with high variance and is sensitive to outliers. So, models cannot converge to a well optimal point. A large τ^* is so conservative that the weights for all samples are almost similar. So, the performance is similar to the baselines. To tackle this problem, we propose step τ^*, which declines the value of τ^* during training. The intuition behind step τ^* is that: at the beginning of the training, the model assigns almost similar losses to all samples and cannot identify which sample is more important or not. For this reason, we can allocate a large τ^* which hardly affects the training process. After some steps, we decrease the value of τ^* and improve the weights for more important (hard-to-distinguish) samples.

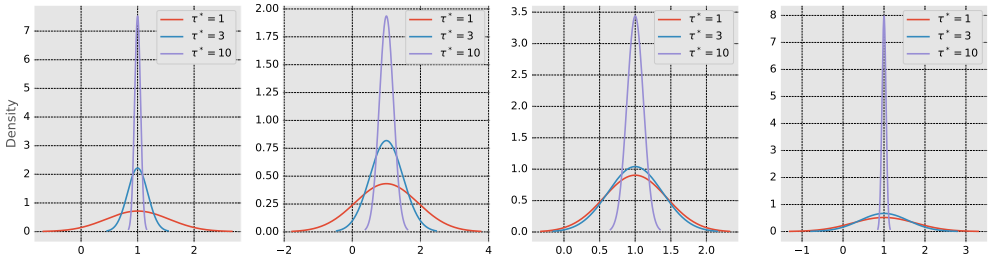


Figure 2: Distribution visualization of sample weights at steps $[1000, 5000, 10000, 20000]$ (from left to right). The horizontal axis represents the weight, and the vertical axis represents the density.

Weights queue $\mathcal{M}$. The second problem is that the expectation over each mini-batch is not a good estimator of the normalizer $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$. We preserve a queue $\mathcal{M} = \{w_i := e^{\ell(x_i,y_i;\theta)/\tau^*}\}_{i=1}^{|\mathcal{M}|}$ that stores historical weights and serve $|\mathcal{M}|$ a hyper-parameter. $|\mathcal{M}|$ is an integer multiple of batch size N and determines how well $\mathcal{M}$ can estimate $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$. The detailed analysis is in Section 4.3.

The resulting target combining step τ^* and weights queue $\mathcal{M}$ is termed **Unit DRO v2**.

$$\mathcal{L}_{\textbf{Unit DRO v2}}(\theta, \tau^*(t)) = \min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{e^{\ell(x,y;\theta)/\tau^*(t)}}{\frac{1}{|\mathcal{M}|}\sum_{w_i \in \mathcal{M}}\left(w_i\right)} \ell(x, y; \theta) \right), \tag{10}$$

where t is the training step and $\tau^*(t)$ means τ^* is a function of t. Algorithm 1 depicts the online optimization algorithm. Note that in group DRO (Algorithm 1 in (Sagawa et al., 2019)), samples in one domain share the same weight, which is actually a special case of Unit DRO (Algorithm 1 here). One key improvement from previous group DRO to (Sagawa et al., 2019) is the implementation trick that the group weights is updated using exponentiated gradient ascent instead of picking the group with worst average loss at each step. (Sagawa et al., 2019) shows such an improvement is important for stability and obtaining convergence guarantees but cannot explain why it works. In contrast, the weights in this work are interpretable: the optimal distribution of DRO with KL constraint is proportional to the empirical distribution composite with the exponential term $e^{\ell(x,y;\theta)/\tau^*}$.

Algorithm 1: Online optimization algorithm for **Unit DRO v2**

Input: $\mathcal{P} = \{P_g\}_{g=1}^{|\mathcal{P}|}$, batch size N, learning rate η, SGD hyper-parameters β, training iterations T.
Initial: Parameters θ^0 and $\mathcal{M}^0 = \{1\}_{i=1}^{|\mathcal{M}|}$.
for t = $1, ..., T$ **do**

$(x_i, y_i)_{i=1}^N \sim \mathcal{P}$ *//Data sampling*

$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e^{\ell(x,y;\theta^{t-1})/\tau^*(t)}}{\frac{1}{|\mathcal{M}|}\sum_{w_i\in\mathcal{M}}(w_i)}\ell(x,y;\theta^{t-1})\right)$ *//Calculate the reweighted loss*

$\mathcal{M}^t = \left[\mathcal{M}^{t-1}[N:], \{e^{\ell(x_i,y_i;\theta^{t-1})/\tau^*(t)}\}_{i=1}^N\right]$ *//Update weights queue by current weights*

$\theta^t \leftarrow \mathrm{SGD}\left(\mathcal{L}, \theta^{t-1}, \eta, \beta\right)$ *//Update model parameters*

end

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS.

Here we aim to answer the following questions:

- *Without demographics, how does Unit DRO perform compared to advanced CD and DG methods?*
- *How do hyperparameters in Unit DRO influence the performance?*
- *Why Unit DRO can achieve performance improvements and when Unit DRO will fail?*

To answer the first question, we compare Unit DRO with baselines of both DG ReID and CD ReID on several benchmarks. We perform detailed ablation studies to answer the second question. Comprehensive analysis are conducted for the third question, *e.g.,* error set analysis, feature visualization and domain divergence measure, *etc*. The main setups of the experiments are as follows.

Dataset and Setting. Following (Song et al., 2019; Jia et al., 2019; Zhang et al., 2021a), we evaluate the DIR-ReID with multiple data sources (MS), where source domains cover five large-scale ReID datasets, including CUHK02 (Li & Wang, 2013), CUHK03 (Li et al., 2014), Market1501 (Zheng et al., 2015), DukeMTMC-ReID (Zheng et al., 2017), and CUHK-SYSU PersonSearch (Xiao et al., 2016). The unseen test domains are VIPeR (Gray et al., 2007), PRID (Hirzer et al., 2011), QMUL GRID (Liu et al., 2012), and i-LIDS (Wei-Shi et al., 2009). We include the detailed illustration of datasets and evaluation protocols in Appendix B.1. In the CD domain setting, we employ Market1501 and DukeMTMC-ReID. We alternately construct the two datasets into source and target domains.

Baselines. We compare our model with (i) **DG ReID methods**, *e.g.,* AugMining (Tamura & Murakami, 2019), DIMN (Song et al., 2019), DualNorm (Jia et al., 2019), SNR (Jin et al., 2020), DDAN (Chen et al., 2021), DIR-ReID (Zhang et al., 2021a), and MetaBIN (Choi et al., 2021). (ii) **CD ReID methods**, *e.g.,* CrossGrad (Shankar et al., 2018), QAConv (Liao & Shao, 2019), L2A-OT (Zhou et al., 2020), OSNet-AIN (Zhou et al., 2021), SNR (Jin et al., 2020), DIR-ReID (Zhang et al., 2021a), and MetaBIN (Choi et al., 2021).

Implementation. Following previous generalizable person ReID methods, we use MobileNetV2 (Sandler et al., 2018) with a multiplier of 1.4 as the backbone network, which is pretrained on ImageNet (Deng et al., 2009). Images are resized to 256×128 and the training batch size N is set to 80. The SGD optimizer is used to train all the components with a learning rate of 0.01, a momentum of 0.9 and a weight decay of 5×10^{-4}. The learning rate is warmed up in the first 10 epochs and decayed to its $0.1\times$ and $0.01\times$ at 40 and 70 epochs. The step τ^* is initialized with $\tau^* = 100$ and decayed to $20, 5$ at 40 and 70 epochs. The default size for $\mathcal{M}$ is 800. The training process includes 100 epochs and we use the automatic mixed-precision training to increase memory efficiency in the entire process. For the hyperparameters of losses: the label-smoothing parameter is 0.1 and the margin in the triplet loss is 0.3. We serve $\mathcal{L}_{ce}$ as the $\ell(x, y; \theta)$ in Unit DRO. We compare different normalization methods in Table. 13 of the Appendix and integrate the mixture of BN and IN with learnable balancing parameters (Choi et al., 2021) to the proposed Unit DRO. We conduct all the experiments on a machine with i7-8700K, 32G RAM and four GTX2080ti.

Methods	Average		VIPeR				PRID				GRID				i-LIDS			
	R-1	mAP	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
AugMining	51.8	-	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	47.5	57.9	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	57.6	61.8	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	59.0	63.1	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
DDAN w/DualNorm	60.9	65.1	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	63.8	71.2	58.5	76.9	**83.3**	67.0	69.7	85.8	91.0	77.1	48.2	67.1	76.3	57.6	79.0	**94.8**	97.2	83.4
DIR-ReID†	62.3	70.8	57.2	74.1	80.2	64.9	67.6	87.1	91.6	76.6	47.2	66.1	75.4	57.0	77.3	93.3	97.2	84.5
MetaBIN	64.7	72.3	56.9	76.7	82.0	66.9	72.5	**88.2**	91.3	**79.8**	49.7	67.5	76.8	**58.1**	79.7	93.3	97.0	85.5
MetaBIN†	64.2	71.9	59.3	76.8	81.9	67.6	70.6	86.5	91.5	78.2	47.3	66.0	74.0	56.4	79.5	93.0	**97.5**	85.5
Group DRO	57.1	65.9	48.5	68.4	77.2	57.8	66.1	86.5	90.6	74.8	38.7	58.8	66.6	48.6	74.8	90.8	96.8	81.9
Group DRO†	56.7	65.6	48.5	68.9	76.6	58.1	65.4	85.4	89.8	74.1	38.4	58.6	66.1	48.4	74.5	91.0	96.0	81.7
Unit DRO†	**65.4**	**72.8**	**60.0**	**78.2**	82.8	**68.4**	**73.5**	85.3	**91.7**	79.4	47.5	**69.3**	**77.4**	57.2	**80.7**	94.0	97.0	**86.2**

Table 1: Comparisons against state-of-the-art DG methods for person ReID, where '†' indicates the reported result is simply from the last checkpoint. The 1^{st} highest accuracy is indicated by **red bold**.

Method	Market-Duke				Duke-Market			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
CrossGrad	48.5	63.5	69.5	27.1	56.7	73.5	79.5	26.3
QAConv	48.8	-	-	28.7	58.6	-	-	27.6
L2A-OT	50.1	64.5	70.1	29.2	63.8	80.2	84.6	30.2
OSNet-AIN	52.4	66.1	71.2	30.5	61.0	77.0	82.5	30.6
SNR	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
MetaBIN	55.2	69.0	74.4	33.1	69.2	83.1	87.8	35.9
Unit DRO	**55.5**	**70.3**	**74.9**	**33.8**	**69.2**	**83.7**	**88.0**	**36.4**

Table 2: Performance (%) comparison with the state-of-the-arts on the CD ReID problem. All of these methods adopt ResNet50 as the backbone.

τ^*	$\lvert\mathcal{M}\rvert$	R-1	mAP
$[50, 5, 3]$	800	64.2	72.1
$[100, 5, 3]$	0	63.4	71.6
$[100, 5, 3]$	800	65.4	72.8
$[100, 5, 3]$	4000	63.9	71.9
$[100, 10, 3]$	800	64.2	71.8
$[100, 10, 3]$	1600	64.2	72.0
$[100, 20, 3]$	800	64.8	72.3
$[100, 20, 5]$	800	65.4	72.8

Table 3: Ablation studies of step τ^*. $\tau^* = [\tau_1, \tau_2, \tau_3]$ means $\tau^* = \tau_1$ initially and decayed to τ_2 and τ_3 at 40 and 70 epochs.

4.2 Numerical Results

To the best of our knowledge, there is no work focusing on DGWD-ReID setting. We first evaluate one representative method under the fairness/OOD setting termed Group DRO (Sagawa et al., 2019) on the DG ReID benchmark. We find η for Group DRO within $[10^{-3}, 10^{-1}]$ and the optimal value is 0.01. Table. 1 shows that it doesn't work well. Then we then compare the proposed Unit DRO with the existing methods about DG ReID and CD ReID.

Comparison with DG ReID. We observe that current DG ReID methods (Choi et al., 2021; Zhang et al., 2021a; Chen et al., 2021) all apply an utopian model selection method, they all choose the checkpoint with the best performance on the test datasets and report their results. We argue that such a model selection method is inadvisable. Under the DG setting, we should restrict access to the test domain data (Gulrajani & Lopez-Paz, 2020). Instead, we simply use the last checkpoint and report its results as the final performance over all test datasets. Among the competitors, although some methods have achieved advantages sporadically on one or two datasets, the proposed Unit DRO attains the best performance in the average R-1 accuracy and average mAP over most of the test sets. Note that such comparison is unfair because DG methods can utilize demographics. We also report the last checkpoints' results of other methods. Table 1 shows that, without the utopian model selection method, there is a certain degree of performance decline of these methods, which indicates the effectiveness of Unit DRO again.

Comparison with CD ReID. Table 2 shows the comparison under the CD setting, where 'Market-Duke' indicates that Market1501 is the labeled source domain and DukeMTMC-ReID is an unseen target domain. Because the style variation within a single dataset is relatively small, previous DG ReID methods must utilize fine-grained demographics,*e.g.,* camera IDs (Zhang et al., 2021a), or tune the hyper-parameters carefully (Choi et al., 2021). Instead, Unit DRO does not require additional data augmentation or any changes to model structures and hyper-parameters. For a fair comparison, we employ the Resnet50 backbone with color jittering and Table 2 shows that Unit DRO outperforms current CD methods. So far, we have verified that the proposed Unit DRO has the potential to improve the generalization ability on both multi-source datasets and single-source dataset settings.

Ablation Studies: impact of τ^* and $|\mathcal{M}|$ in Unit DRO. We conduct ablation studies on various components. Table 4 shows that the results with a constant τ^* are not better than the baseline (the

results on the first row). Cooperated with a weights queue with a size of $|\mathcal{M}| = 800$ boosts the performance slightly. However, maintaining a large $\mathcal{M}$ with a size of 5000 is harmful to Unit DRO. Step τ^* with no $\mathcal{M}$ or a large $\mathcal{M}$ all behave not well. We perform a careful search for the most suitable $|\mathcal{M}|$ and step τ^* in Table 3, which brings great performance gain. These empirical results show that both the weights queue $\mathcal{M}$ and the step τ^* play an important role in Unit DRO.

4.3 Analysis

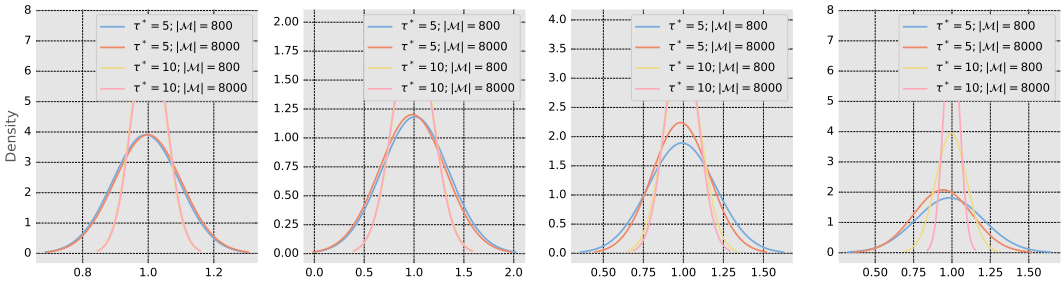


Figure 3: Distribution visualization of sample weights of steps $[1000, 5000, 10000, 20000]$ (from left to right). The horizontal axis represents the weight, and the vertical axis represents the density.

Distributions of sample weights on different parameters. We train models in DG benchmarks 100 epochs and each epoch consists of 1850 steps. Per 1000 steps, all the sample weights[5] will be saved and the mean and variance of these weights will be calculated. We assume these weights obey Gaussian distribution $\mathcal{N}(\mu, \delta)$ and plot diagrams based on the mean μ and variance δ. The x-coordinate of these diagrams is just the value between $[\mu - 3 * \delta, \mu + 3 * \delta]$, not the real values of weights. Based on the loss values of each sample, we calculate the weights with the following three methods. (i) *Sample weights for Unit DRO.* In this case, these weights are normalized in their batches, so the mean of all distributions here is 1. Figure 2 shows the results and we had discussed them in Section 3.1. (ii) *Sample weights for Unit DRO with different* $|\mathcal{M}|$. We plot the sample distribution of steps $[1000, 5000, 10000, 20000]$ in this case. With an additional queue, Figure 3 indicates that weight distributions have different means during training. Theoretically we need a large $|\mathcal{M}|$ to estimate $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$. However, as $|\mathcal{M}|$ becomes larger, the estimation will be inaccurate. Consider an extreme case: $|\mathcal{M}| = T - 1$ and then the queue absolutely contains all the training data. It is catastrophic to estimate $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$ in step T by such a queue. The large queue contains very old sample weights which is unsuitable for the current model. Figure 3 depicts the phenomenon, where the distribution with a larger $|\mathcal{M}|$ always has smaller μ. We plot and discuss the distribution diagrams of step τ^* in Appendix C.2.

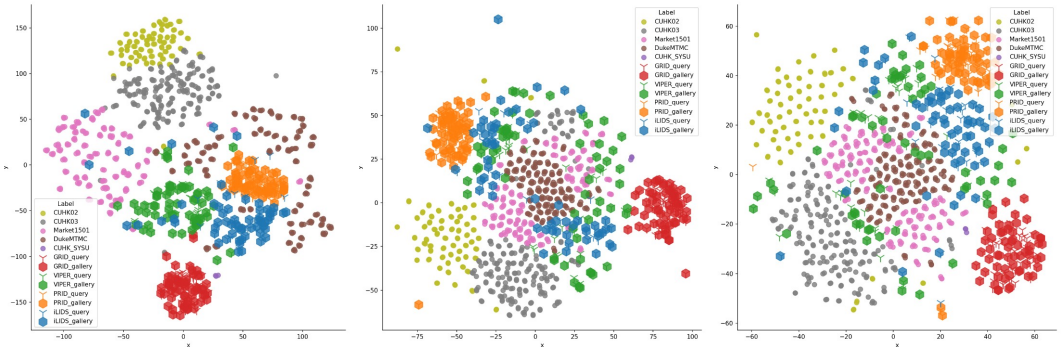
Feature visualization using t-SNE. We compare the proposed Unit DRO to MetaBIN and DualNorm through t-SNE visualization. We observe a distinct division of different domains in Figure 4a, which denotes that a domain-specific feature space is learned by the DualNorm. MetaBIN and the proposed Unit DRO tackle this problem well and the overlaps in Figure 4b and Figure 4c between different domains are more prominent. The t-SNE visualization shows that Unit DRO can learn domain-invariant representations while keeping discriminative capability for ReID tasks. However, MetaBIN follows a meta-learning pipeline and needs expensive demographics. In contrast, no demographics is required by Unit DRO and the framework is simpler. We supply more visualization results and further analysis about the discriminative capability in Section C.3 of Appendix.

Domain divergence measure using $\mathcal{A}$-distance and MMD-distance. We study the MMD distance (Tolstikhin et al., 2016) and $\mathcal{A}$-distance (Long et al., 2015) as measures of domain discrepancy (Ben-David et al., 2010). Detailed implementation is depicted in Section C.4 of Appendix. Table 5 shows that Unit DRO can learn comparable or even more invariant representations compared to MetaBIN, which outperforms DualNorm by a large margin. We also study the correlation between the weights for each dataset and the MMD distance. For each dataset, we calculate the sum of MMD distance between it to all other datasets. Besides, we calculate the average weights of the final model

[5]The distribution of step 5000", which is equal to the distribution of sample weights in steps $[4000, ..., 4999]$

$\tau^* = 1$	$\tau^* = 10$	$\tau^* = 20$	$\lvert\mathcal{M}\rvert = 800$	$\lvert\mathcal{M}\rvert = 5000$	Step τ^*	R-1	mAP
						63.7	72.0
✓						42.0	52.4
	✓					63.5	71.8
		✓				63.6	71.5
		✓	✓			64.1	72.0
	✓			✓		63.5	71.2
					✓	63.8	72.2
				✓	✓	63.9	71.8
			✓		✓	65.4	72.8

Table 4: Ablation study of Unit DRO.



(a) DualNorm (Jia et al., 2019) (b) MetaBIN (Choi et al., 2021) (c) Unit DRO

Figure 4: The t-SNE visualization of the embedding vectors of training and test datasets. Query and gallery samples of these unseen datasets are expressed in different shapes. Best viewed in color.

for each dataset. Table 6 shows that for a tough dataset (*e.g.,* , CUHK02) that has a large divergence to other datasets, Unit DRO assigns a relatively higher average weight. *This phenomenon depicts that even without demographics, Unit DRO can also find meaningful subgroups and upweight them.* We can also see that Unit DRO upweights samples in CUHK-SYSU which has a relatively small MMD distance with other datasets. It is because the generalization ability is not only dependent on domain divergence, but also some other factors. We discuss these influence factors and perform error set analysis in Section C.6 of Appendix. We also plot the MMD-distance of every dataset pair and give further analysis in Section C.5 of Appendix.

5 Conclusion

It is common that traditional DG ReID methods fail to work in cases where domain information, camera ID, or other demographics are difficult to obtain due to privacy issues. To this end, We introduce DGWD-ReID, a new setting that needs to learn domain-invariant representation without demographics. Under DGWD-ReID, we further propose Unit DRO, a new method reformulated from KL constraint DRO. Unit DRO learns domain-invariant features and outperforms previous DG ReID methods that even require demographics. Empirical results and detailed analysis have verified that Unit DRO can find semantically meaningful samples and subgroups without demographics.

Method	MMD↓ (U)	MMD↓ (T)	MMD↓ (A)	$\mathcal{A}\downarrow$ (U)	$\mathcal{A}\downarrow$ (T)	$\mathcal{A}\downarrow$ (A)
DualNorm	0.52	0.21	0.41	1.96	1.91	1.88
MetaBIN	0.41	0.19	0.36	1.96	1.89	1.86
Unit DRO	0.41	0.19	0.35	1.95	1.89	1.85

Table 5: Divergence measurement on four unseen datasets (U), five training datasets (T) and all of these datasets (A).

	Cuhk02	Cuhk03	Market	Duke	SYSU
Weight	1.02	0.99	0.99	1.00	1.01
MMD	1.66	1.17	1.15	1.16	1.04

Table 6: Average weight and one-to-all MMD distance for training datasets.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning (ICML)*, 2020.

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

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 2010.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton university press, 2009.

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

Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning (ICML)*, 2021.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, 2021b.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 2010.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, 2012.

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.

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.

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

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

Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

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.

Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition (CVPR)*, June 2013.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

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

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 *British Machine Vision Conference (BMVC)*, 2018.

Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision (ECCV)*. Springer, 2012.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015.

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.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Modeling the second player in distributionally robust optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, 2013.

Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. *arXiv preprint arXiv:1805.07925*, 2018.

Viet Anh Nguyen, Nian Si, and Jose Blanchet. Robust bayesian classification using an optimistic score ratio. In *International Conference on Machine Learning (ICML)*, 2020.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, 2020.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

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 *Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Xingchao Peng, Yichen Li, and Saenko Kate. Domain2vec: Domain embedding for unsupervised domain adaptation. *European Conference on Computer Vision (ECCV)*, 2020.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

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.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

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.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 2017.

Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. Associating groups of people. In *British Machine Vision Conference (BMVC)*, 2009.

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.

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*, 2020.

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.

Yi-Fan Zhang, Hanlin Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. Learning domain invariant representations for generalizable person re-identification. *arXiv preprint arXiv:2103.15890*, 2021a.

Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhong Zhang, Haijia Zhang, and Shuang Liu. Person re-identification using heterogeneous local graph attention networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Sebe Nicu. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *International Conference on Computer Vision (ICCV)*, December 2015.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *International Conference on Computer Vision (ICCV)*, Oct 2017.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision (ECCV)*, 2020.

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.