Generalizable Person Re-identification Without Demographics

Yi-Fan Zhang[1], Zhang Zhang[1], Baosheng Yu[2], Liang Wang[1], Dacheng Tao[2], Tieniu Tan[1]
[1]Institute of Automation, Chinese Academy of Sciences, [2]The University of Sydney.

Abstract

Domain generalizable person re-identification (DG-ReID) aims to learn a ready-to-use domain-agnostic model directly for cross domain evaluation, while current methods mainly explore the demographic information such as domain and/or camera labels for domain-invariant representation learning. However, the above-mentioned demographic information is not always accessible in practice due to the privacy and security issues. In this paper, we consider the problem of person reid in a more general setting, i.e. *domain generalizable person re-id without demographics (**DGWD-ReID**). To address the underlying uncertainty of domain distribution, we introduce distributionally robust optimization (DRO) to learn robust person reid models without demographics. However, directly applying the popular Kullback-Leibler divergence constrained DRO (or KL-DRO) fails to generalize well, since the convex condition may not hold for overparameterized neural networks. Inspired by this, we reformulate the popular KL-DRO, and then propose a simple yet efficient approach, **Unit-DRO**, which minimizes the loss over a new dataset with hard samples upweighted and other samples downweighted. We perform extensive experiments on both DG reid tasks, and the empirical results on several large-scale benchmarks show that Unit-DRO achieves superior performance without using demographics.*

1. Introduction

Person re-identification (ReID) aims to find the correspondences between person images from the same identity across multiple camera views. **Domain generalizable person ReID (DG-ReID)** models are trained on multiple large-scale datasets and tested on unseen domains directly without extra data collection/annotation and model updating on new domains. Therefore, DG-ReID is receiving increasing attention from the community due to its great value in real-world person retrieval applications.

However, current DG-ReID research usually comes at a serious disadvantage: it requires the demographic information (*e.g*. domain labels [7, 62], camera IDs [61], and video timestamps [57]) as the extra supervision for model training. Such demographics implicitly define the variations in training data that the learned model should be invariant or robust to. Unfortunately, the demographic information is usually not available in practice due to the following reasons: 1) the collection of demographics inevitably leads to privacy problems [52], *e.g*. the risks of exposing the geographical location and/or the environment information; 2) the collection/annotation of domain labels is very expensive and ethically fraught endeavours [35]; and 3) such coarse-grained labels and the noise of manual annotation collected domain labels may exacerbate the *hidden stratification issue*, which hinders a variety of safety-critical applications [8, 24, 38] (refer to Appendix A for more discussions). Therefore, we consider a more general setting for ReID, *i.e*. DG Person Re-identification Without Demographics (DGWD-ReID), where the model is trained without demographics.

To address the underlying uncertainty of domain distribution without using demographics, distributionally robust optimization (DRO) is a promising paradigm [19]. 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. The uncertainty set defined by an f-divergence ball (such as Kullback-Leibler divergence) from the training distribution has been very popular, which is also known as KL-DRO [20]. However, the convex assumption in KL-DRO usually does not hold in real-world scenarios, thus leading to the inferior performance in the context of overparameterized neural networks. To this end, we address the above-mentioned issue and reformulate KL-DRO to first solve the inner step optimization problem and then obtain a closed-form expression of the optimal objective. By doing this, the proposed Unit-DRO avoids the troublesome bi-level optimization in traditional DRO problems and scales well to overparameterized regimes. Compared to previous DG-ReID methods, the proposed Unit-DRO is simple yet effective, which also avoids the need for either meta-learning pipelines or complicated model structure.

In this paper, we evaluate the proposed Unit-DRO for person ReID by comparing it with existing DG-ReID methods. Unit-DRO outperforms recent methods, even including those methods using demographics. To better understand the proposed Unit-DRO, we perform comprehensive ablation studies on several important components. We also visualize t-SNE embeddings, and measure the domain divergence and error set to show the good invariant learning capability of Unit-DRO. Empirical results show that the proposed Unit-DRO can effectively retrieve valuable samples or subgroups without demographics.

2. Method

2.1. Overview

Problem Formulation. Given current DG-ReID setting, there is a labeled set of training data from several different domains: $\mathcal{P} = \cup_{k=1}^{N} P_k$ and $P_k = \{(x_i, y_i)\}_{i=1}^{N_k}$, where N is the number of domains, N_k is the number of images in domain P_k, and $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ indicate an image and its corresponding label, respectively. During training, we use all aggregated image-label pairs from $\mathcal{P}$. During testing, we evaluate the person retrieval performance on the unseen target domain G without any additional model updating. Therefore, the goal of DG-ReID is to learn a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ that minimizes the error on the target domain G:

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

where ℓ is the predefined loss function. Previous studies mostly leverage demographics (*e.g.* domain/camera labels and video timestamps) to clip the spurious correlations for more robust models, which is not always available in real-world applications. Therefore, we consider a more general setting where the above-mentioned demographic information is unknown during training, *i.e.* DG-ReID without demographics or DGWD-ReID.

Baseline Algorithm. We introduce the objectives used in our baseline as follows. The first one is the cross-entropy loss. Given n training points $\{(x_1, y_1), ..., (x_n, y_n)\}$, we then have the loss for person identity classification: $\mathcal{L}_{ce} = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$, where ℓ indicates the cross-entropy loss function. Label-smoothing is also applied to prevent the model from overfitting to the identity labels. Inspired by recent ReID methods, we further introduce triplet loss to enhance the intra-class compactness and inter-class separability in the embedding space. Given an anchor sample x_i^a, we then evaluate triplet loss using the hardest positive and negative samples, x_i^p and x_i^n:

$$\mathcal{L}_{tr}(x_i^a, x_i^p, x_i^n; \theta) = \max\left\{d(x_i^a, x_i^p; \theta) - d(x_i^a, x_i^n; \theta) + m,\ 0\right\}, \tag{2}$$

where $d(\cdot, \cdot)$ indicates a pairwise distance such as the normalized Euclidean distance, and m is the margin between positive and negative pairs. Similar to [34], we use a BN-Neck structure to maximize the synergy between $\mathcal{L}_{ce}$ and $\mathcal{L}_{tr}$. We also integrate a mixture of batch normalization and instance normalization with learnable parameters [7], which has proved to be very useful for DG-ReID.

2.2. Unit-DRO

To address the underlying uncertainty of domain distribution without demographics, we introduce Unit-DRO, a novel generalization framework that does not require priors about demographics. We first introduce the basic DRO framework [5, 42] as follows, where the worst-case expected risk over a predefined family of distributions $\mathcal{Q}$ (termed *uncertainty set*) is used to replace the expected risk on the unseen target distribution G in Equ.(1) and the objective is,

$$\min_{\theta \in \Theta} \max_{q \in \mathcal{Q}} \mathbb{E}_{(x,y)\in q}[\ell(x, y; \theta)]. \tag{3}$$

Specifically, 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 bound the risk under G.

An important question for using DRO is how to choose the uncertainty set. Note that in real-world applications, we can obtain only the empirical (training) data distribution. The uncertainty set can thus be constructed by collecting the distributions within a certain distance from the training distribution. For example, previous work may choose a KL-divergence ball [20]/MMD ball [46] around the training distribution, which confers robustness to a wide set of distribution shifts. However, it can also lead to overly pessimistic models which optimize for implausible worst-case distributions [12]. In other words, $\mathcal{Q}$ should be sufficiently large to contain G, while it may also contain noisy distributions [35]. Group-DRO [43] thus leverages demographics to define the uncertainty set $\mathcal{Q}$ and attains superior OOD performance. Here we consider a new extension of DRO to improve OOD generalization *without demographics*.

KL-DRO. We first introduce the construction of $\mathcal{Q}$ based on the KL-divergence ball around the empirical distribution $\mathcal{P}$ as follows. Given the KL upper bound (radius) η, we have $\mathcal{Q} = \{Q : \mathrm{KL}(Q||\mathcal{P}) \le \eta\}$. The min-max problem in Equ.(3) can then be reformulated as

$$\min_{\theta \in \Theta} \max_{Q:\mathrm{KL}(Q||\mathcal{P})\le\eta} \mathbb{E}_{(x,y)\in Q}\left[\ell(x, y; \theta)\right]. \tag{4}$$

Lemma 1 *(Modified from Section 2 in [20]) 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 Equ.(4) has a closed-form solution*

$$q^*(x, y) = \frac{p(x, y)e^{\ell(x,y;\theta)/\tau^*}}{\mathbb{E}_{\mathcal{P}}\left[e^{\ell(x,y;\theta)/\tau^*}\right]}, \tag{5}$$

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. The min-max problem in Equ.(4) is then equivalent to

$$\min_{\theta\in\Theta, \tau>0} \tau \log \mathbb{E}_{\mathcal{P}}\left[e^{\ell(x,y;\theta)/\tau}\right] + \eta\tau. \tag{6}$$

We refer to Equ.(6) as **KL-DRO**. Unfortunately, the convex condition of KL-DRO is not held for overparameterized neural networks, such that applying it may fail to generalize under the distribution shifts in real-world scenarios. As illustrated in Figure 1, we compare the training statistics with the baseline, where KL-DRO is highly unstable and attains inferior results. Therefore, instead of following KL-DRO to directly use the inner maximum, we reformulate Equ.(4) as follows.

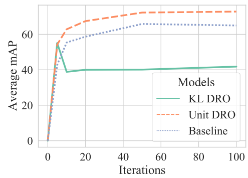


Figure 1. Training statistics.

$$
\begin{aligned}
&\min_{\theta\in\Theta} \max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta} \mathbb{E}_{(x,y)\in Q}[\ell(x,y;\theta)] \\
&= \min_{\theta\in\Theta} \max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta} \int \ell(x,y;\theta) q(x,y) d_x d_y \\
&= \min_{\theta\in\Theta} \max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta} \int \ell(x,y;\theta) \frac{q(x,y)}{p(x,y)} p(x,y) d_x d_y \\
&= \min_{\theta\in\Theta} \max_{Q:\mathrm{KL}(Q||\mathcal{P})\leq\eta} \mathbb{E}_{(x,y)\in\mathcal{P}}\left[\frac{q(x,y)}{p(x,y)}\ell(x,y;\theta)\right] \\
&= \min_{\theta\in\Theta} \mathbb{E}_{(x,y)\in\mathcal{P}}\left[\frac{e^{\ell(x,y;\theta)/\tau^*}}{\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]}\ell(x,y;\theta)\right].
\end{aligned} \tag{7}
$$

Specifically, to obtain the third line, we apply the change-of-measure technique. The fourth line replaces the inner maximum by its closed-form solution $q^*(x,y)$ in Equ.(5). 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 all training data, which is untrackable. For simplicity, we can serve τ^* as a parameter and take the average over mini-batch as an estimator of the normalizer. Therefore, we have the formulation of vanilla Unit-DRO:

$$
\mathcal{L}_{\text{Unit-DRO}}(\theta,\tau^*) = \min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e^{\ell(x,y;\theta)/\tau^*}}{\frac{1}{N}\sum_{i=1}^{N}\left(e^{\ell(x,y;\theta)/\tau^*}\right)}\ell(x,y;\theta)\right), \tag{8}
$$

where N is the batch size. However, vanilla Unit-DRO does not work well in practice, and we address the following two problems to form a robust Unit-DRO solution.

Multi-Step τ^*. The first problem is that a constant hyperparameter τ^* is usually suboptimal for the whole learning process. As shown in Figure 2, we visualize the densities of the weight $e^{\ell(x,y;\theta)/\tau^*}/\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$ at different optimization steps when using a constant τ^* (please refer to Section C.3 for the details). Specifically, we find that: 1) a small τ^* leads to the high variance on the weight distribution and is also sensitive to outliers; 2) a large τ^* is so conservative that the weights for all samples are almost similar. To this end, we propose a multi-step solution for the hyparameter τ^*, which declines with the training/optimization steps. The intuition behind the multi-step τ^* is that: at the beginning, we use a large τ^*, and the model thus assigns almost similar weights to all samples and cannot identify which sample is more important or not. With the increase of training steps, we decrease the value of τ^* and improve the weights for important (*i.e.* hard-to-distinguish) samples.

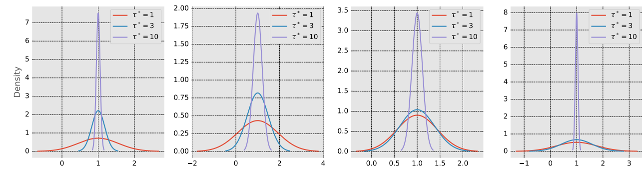


Figure 2. Visualizing the distribution of sample weight at $1k, 5k, 10k, 20k$ steps, respectively (from left to right). The horizontal axis represents the weight.

Weight Queue $\mathcal{M}$. The second problem is that the expectation over each mini-batch may not be a good estimator of the normalizer $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$. To address this problem, we introduce a queue $\mathcal{M} = \{w_i := e^{\ell(x_i,y_i;\theta)/\tau^*}\}_{i=1}^{M}$ to maintain the historical weights, where M depends on the batch size N and determines how well $\mathcal{M}$ can estimate $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$. (see more analysis in Section C.3). Lastly, we have the objective function of **Unit-DRO**:

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

where t is the index of training step and τ^* is a piecewise function of t. As shown in Figure 1, the training statistics of Unit-DRO is more stable than KL-DRO, and it's performance also outperforms baseline methods by a large margin. Note that in Algorithm 1 of Group-DRO [43], all samples in the same domain share the same weight, which can be seen as a special case of the proposed Unit-DRO. Comparing with Group-DRO, one of the key improvements is the implementation trick that the group weights are updated using exponential gradient ascent instead of picking the group with the worst average loss at each step. Specifically, Group-DRO shows that such an improvement is useful for training stability and model convergence but cannot explain why it works. In contrast, the adaptive weights used in this paper are interpretable: the optimal distribution of DRO with KL constraint is proportional to the empirical distribution composite with the exponential term $e^{\ell(x,y;\theta)/\tau^*}$.

3. Experiments

In this section, we evaluate the proposed Unit-DRO and try to answer the following questions: "without demographics, how does Unit-DRO perform compared to other CD-ReID and DG-ReID methods? what is the influence of different hyperparameters in Unit-DRO? why Unit-DRO improves the baseline?". To answer the first question, we compare Unit-DRO with baseline methods on both DG-ReID and CD-ReID benchmarks. We then perform detailed ablation studies to answer the second question. Comprehensive analyses are conducted for the third question, *e.g.* error set

Protocol (i)	mAP	Rank-1	Protocol (ii)	mAP	Rank-1
SNR	54.3	48.48	RaMoE	31.2	32.4
Ours	**58.84**	**52.7**	Ours	**41.7**	**40.4**
Protocol (iii)	mAP	Rank-1	Protocol (iv)	mAP	Rank-1
M3L	26.7	27.9	RaMoE	68.8	58.9
Ours	**27.8**	**29.1**	Ours	**73.2**	**65.4**

Table 1. Comparison with SOTA DG-ReID methods under different evaluation protocols, where the Duke is removed from source and target domains. The best accuracy is highlighted by **bold**.

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

analysis, feature visualization, and domain divergence measure. For space limit, we place the full experimental results in the appendix C and leave part of generalization results and analysis in the main manuscript.

Experimental Setup. Following previous DG-ReID methods, we use MobileNetV2 [44] with the width multiplier of 1.4 as the backbone network, which is initialized using the weights pretrained on ImageNet [11]. All training images are resized to 256×128 pixels and the batch size is $N = 80$. We use the SGD optimizer with a momentum 0.9 and the weight decay $5e-4$. The learning rate starts from 0.01 and then decays to its $0.1\times$ at 40 and 70 epochs. We also use a warmup learning rate schedule at the first 10 epochs. We initialize the multi-step τ^* with $\tau^* = 100$, which is then decayed to 20 and 5 at 40 and 70 epochs, respectively. The default size of the weight queue is $M = 800$.

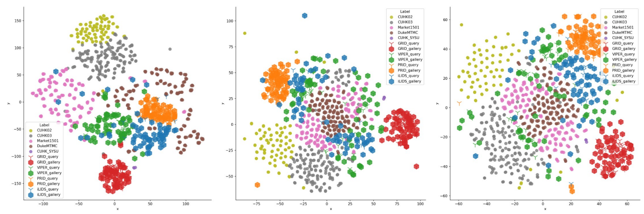
DG-ReID Protocols. We compare Unit-DRO with other methods using the following protocols: (1) one-to-multiple setting [23]; (2) multiple-to-one setting [9]; (3) multiple-to-one setting [62]; and (4) multiple-to-multiple setting [23]. Due to the page limit, please see the results in Appendix. Besides, due to privacy issue, the Duke dataset is not appropriate for using. We thus conduct experiments under different evaluation protocols but remove the Duke from source/target domains, and Table 1 shows that the performance margin between Unit-DRO and other baselines becomes larger. Because these protocols are used in different DG-ReID papers, we choose the SOTA method under every protocol for comparison.

Ablation Studies. We conduct ablation studies on different Unit-DRO components, including the multi-step τ^* and the weight queue. Results in Tab. 3 verify the importance of both of these two components.

Domain Divergence Analysis. We explore MMD distance [51] and $\mathcal{A}$-distance [32] as the measure of domain discrepancy [4]. Table 2 shows that Unit-DRO can learn comparable or even more invariant representations compared to MetaBIN, which outperforms DualNorm by a large margin. With t**-SNE visualization**, the superior of Unit-DROis verified again and representations inFigure 3c are more invariant.

$\tau^* = 10$	$\tau^* = 20$	$\lvert\mathcal{M}\rvert = 800$	$\lvert\mathcal{M}\rvert = 5000$	Multi-Step τ^*	R-1	mAP
✓					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 3. Ablation studies on different Unit-DRO components.



(a) DualNorm [22] (b) MetaBIN [7] (c) Unit-DRO

Figure 3. Visualization of the embeddings on training and test datasets. Query and gallery samples of these unseen datasets are shown using different types of mark. Best viewed in color.

Refer to Appendix B for discussions about DG, DG ReID, DRO and related works.

4. Conclusion

Traditional DG-ReID methods fail to work in the cases where domain information are not available due to the security and privacy issues. To this end, we introduce DGWD-ReID, a more general setting that requires the model to learn domain-invariant representations without demographics. To address this problem, we propose Unit-DRO, which is a simple yet effective algorithm that substantially improves the model generalization performance without requiring expensive demographics during training. Extensive experimental results demonstrate that the proposed Unit-DRO not only achieves comparable or better performance comparing with other DG-ReID methods using the demographic information, but also can be used for other DG applications.

Different from typical image classification tasks, where domains are partitioned by image styles, person ReID datasets have more fine-grained variation factors, *e.g.* the image styles of different datasets, camera perspective changes within one dataset, and the shooting conditions at different times on the same camera. We believe that simply specifying each dataset as a separate domain is suboptimal and a better domain inference method that considers the above variation factors will be the subject of future study.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning (ICML)*, 2020. 8

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 8, 10

[3] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2123–2132, 2021. 9

[4] 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. 4

[5] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton university press, 2009. 2, 9

[6] 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. 9

[7] 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. 1, 2, 4, 8, 9, 10

[8] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning (ICML)*, 2021. 1, 8

[9] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, 2021. 4, 9, 11

[10] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 2010. 9

[11] 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. 4

[12] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019. 2, 9

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, 2012. 8

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 10

[15] 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. 9

[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. 9

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 10

[18] 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. 9

[19] 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. 1

[20] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013. 1, 2, 9

[21] 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. 9

[22] 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. 4, 9, 10

[23] 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. 4, 8, 9

[24] Young Eun Kim and Seong-Whan Lee. Domain generalization with pseudo-domain label for face anti-spoofing. *arXiv preprint arXiv:2107.06552*, 2021. 1

[25] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition (CVPR)*, June 2013. 9

[26] 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. 9

[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. 9

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

[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 *British Machine Vision Conference (BMVC)*, 2018. 13

[30] 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. 9

[31] 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. 8

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015. 4, 13

[33] 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. 9

[34] 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. 2

[35] Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Modeling the second player in distributionally robust optimization. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 9

[36] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, 2013. 8

[37] Viet Anh Nguyen, Nian Si, and Jose Blanchet. Robust bayesian classification using an optimistic score ratio. In *International Conference on Machine Learning (ICML)*, 2020. 9

[38] 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. 1

[39] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019. 9

[40] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person reidentification. In *Computer Vision and Pattern Recognition (CVPR)*, June 2016. 9

[41] Xingchao Peng, Yichen Li, and Saenko Kate. Domain2vec: Domain embedding for unsupervised domain adaptation. *European Conference on Computer Vision (ECCV)*, 2020. 9

[42] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. 2, 9

[43] 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. 2, 3, 10

[44] 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. 4

[45] 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. 9

[46] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017. 2, 9

[47] 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. 8, 9

[48] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR, 2020. 8

[49] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 9

[50] Masato Tamura and Tomokazu Murakami. Augmented hard example mining for generalizable person re-identification. *arXiv preprint arXiv:1910.05280*, 2019. 9

[51] 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. 4, 13

[52] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 2017. 1

[53] 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. 9

[54] Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. Associating groups of people. In *British Machine Vision Conference (BMVC)*, 2009. 9

[55] 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. 9

[56] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 8, 10

[57] 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. 1

[58] 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. 9

[59] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. *arXiv preprint arXiv:2111.13839*, 2021. 8

[60] Yi-Fan Zhang, Hanlin Zhang, Jindong Wang, Zhang Zhang, Baosheng Yu, Liang Wang, Dacheng Tao, and Xing Xie. Domain-specific risk minimization, 2022. 8

[61] 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*, 2021. 1, 8, 9

[62] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021. 1, 4, 9, 11

[63] 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. 9, 13

[64] 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. 9, 13

[65] 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. 9

[66] 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. 9

[67] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *European Conference on Computer Vision*, pages 140–157. Springer, 2020. 9