Generalizable Person Re-identification Without Demographics – Appendix –

A Extended Related Work

Fully supervised Person ReID. Supervised person ReID depends on the assumption that training and testing data are independent and identically distributed. These methods are usually design to learn discriminative features (Matsukawa et al., 2016; Zhang et al., 2021b) or develop efficient metrics (Koestinger et al., 2012). Single-domain person ReID has achieved great progress with the rapid development of deep Convolutional Neural Networks (CNNs). Despite the encouraging performance under the single-domain setup, current fully-supervised ReID models degrade significantly when deployed to an unseen domain.

Unsupervised-domain adaptation Person ReID. Unsupervised Domain Adaptation (UDA) technologies have great progress (Peng et al., 2020) and have been widely adopted for cross-domain person ReID. The UDA-based ReID methods usually attempt to transfer the knowledge learned from the labeled source domains to target domains, depending on target-domain images (Luo et al., 2020; Huang et al., 2020), features (Wang et al., 2018) or metrics (Peng et al., 2016). Another group of UDA-based methods (Ge et al., 2020; Zhai et al., 2020) 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 model retraining.

Distributionally Robust optimization. Distributionally Robust optimization (Ben-Tal et al., 2009) solve robust versions of ERM, which replace the expected risk under the training data distribution with the worst expected risk over a pre-defined uncertainty set $\mathcal{Q}$ (refer to (Rahimian & Mehrotra, 2019) for a review). Recent studies consititute $\mathcal{Q}$ analytically, such as using moment constraint (Delage & Ye, 2010; Nguyen et al., 2020), f-divergence (Hu & Hong, 2013; Michel et al., 2021), Wasserstein/MMD ball (Sinha et al., 2017; Staib & Jegelka, 2019) or coarse-grained mixture models (Oren et al., 2019; Duchi et al., 2019). We reformulate KL-constraint DRO to an important sampling problem (Unit DRO) and propose an efficient implementation, which scales to large dataset and overparameterized neural network.

B Detailed dataset setting

B.1 Dataset Details

Details of the training datasets are summarized in Table 7 and the test datasets are summarized in Table 8. All the assets (*i.e.,* datasets and the codes for baselines) we use include a MIT license containing a copyright notice and this permission notice shall be included in all copies or substantial portions of the software.

B.2 Evaluation protocols

GRID (Liu et al., 2012) 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 $1025(775 + 250)$ images in the gallery are used for testing.

i-LIDS (Wei-Shi et al., 2009) 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 (Hirzer et al., 2011) 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

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

Table 7: 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 8: Testing Datasets statistics.

	Source Domains	Target Domains	Backbone	Additional Augmentation
Protocol (i)	M/D	D/M+V+P+G+I	Resnet50	Color-Jittering
Protocol (ii)	Training set of MS+D+M	C3	Resnet50	None
Protocol (iii)	M+D+MT	C3	Resnet50	Color-Jittering
Protocol (iv)	M+D+C3+MT	V+P+G+I	Resnet50	Color-Jittering

Table 9: Difference of four DG evaluation protocols. (M:market1501, C2: Cuhk02, C3: Cuhk03, D: DukeMTMC, MT: MSMT17, CS: CUHK-SYSU, V: ViPeR, P: PRID, G: GRID, I: i-LIDS)

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 (Gray et al., 2007) contains 632 pedestrian image pairs. Each pair contains two images of the same individual seen from different camera views 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 view 1 as probe images and their matched images in view 2 as gallery images.

We follow the single-shot setting. The average rank-k (R-k) accuracy and mean Average Precision (mAP) over 10 random splits are reported based on the evaluation protocol

C Extend Experimental Results

C.1 Additional numerical results

In addition to the results in the main manuscript, we also evaluate the performance of Unit DRO in various known experimental settings. We detail the different evaluation protocols settings as follows. (i) One-to-multiple setting mentioned in (Jin et al., 2020). (ii) Multiple-to-one setting mentioned in (Dai et al., 2021b). (iii) Multiple-to-one setting mentioned in Zhao et al. (2021). (iv) Multiple-to-multiple settings are mentioned in (Jin et al., 2020), which is similar to ours while using different source domains. The detailed difference of these protocols are included in Tab. 9. Experimental results for these four protocols are in Tab. 10, Tab. 11, and Tab. 12, all of these results verify the efficiency of Unit DRO.

C.2 Distribution diagrams of step τ^*

Compared to a constant τ^*, weights with step τ^* always have low δ and are more stable. Surprisingly, the weights assigned by Unit DRO are not so far from 1. Just such a small perturbation on sample weights boosts the generalization performance much.

C.3 Additional t-SNE Visualization Results

Figure 6 shows the t-SNE results on four unseen datasets. Figure 7 shows the t-SNE results on five training datasets and Figure 9 shows the t-SNE results on the Market-Duke benchmark. All of these

Source	Methods	Avg		Target:Market1501		Target:Duke		Target:PRID		Target:GRID		Target:VIPeR		Target:iLIDs	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Market1501	A-IN	45.2	44.1	75.3	89.8	24.1	42.7	33.9	21	35.6	27.2	38.1	29.1	64.2	55
	IBN	39.9	39.1	81.1	92.2	21.5	39.2	19.1	12	27.5	19.2	32.1	23.4	58.3	48.3
	A-SN	42.2	40.9	83.2	93.9	20.1	38	35.4	25	29	22	32.2	23.4	53.4	43.3
	IN	45.7	45.1	79.5	90.9	25.1	44.9	35	25	35.7	27.8	35.1	27.5	64	54.2
	SNR	50.9	49.6	84.7	94.4	33.6	55.1	42.2	30	36.7	29	42.3	32.3	65.6	56.7
	Ours	**54.7**	**53.2**	83.5	92.2	**33.8**	**55.5**	**56.7**	**44.5**	**40**	**31**	**44.7**	**35.3**	**69.3**	**60.7**
Duke-MTMC	A-IN	41.2	43.6	21.8	56	64.5	78.9	38.6	29	19.6	13.6	35.1	27.2	67.4	56.7
	IBN	39.9	41.7	26.5	52.5	69.5	81.4	27.4	19	19.9	12	32.8	23.4	63.5	61.7
	A-SN	42.3	45.5	24.6	55	73	85.9	41.4	32	18.8	12.8	31.3	24.1	64.8	63.3
	IN	43.7	45.1	27.2	58.5	68.9	80.4	40.5	27	20.3	13.2	34.6	26.3	70.6	65
	SNR	51.3	52.2	33.9	66.7	72.9	84.4	45.4	35	35.3	26	41.2	32.6	79.3	68.7
	Ours	**55.6**	**56.2**	**36.4**	**69.2**	72.8	81.7	**63.2**	**53.23**	**39.9**	**30.4**	**44.5**	**34.8**	76.7	**68**

Table 10: Comparisons against state-of-the-art DG methods for person ReID on evaluation protocol (i). Unit DRO outperforms SNR by a large margin in average mAP and Rank-1 accuracy. Especially on the PRID dataset, Unit DRO achieves more than 10% points improvement on both mAP and Rank-1 accuracy.

Protocol (ii)					Protocol (iii)		
Method	mAP	Rank-1	Rank-5	Rank-10	Method	mAP	Rank-1
RaMoE	35.5	36.6	54.3	64.6	M3L	29.9	30.7
Ours	**43.8**	**43.6**	**65.3**	**74.5**	Ours	**30.9**	**31.1**

Table 11: Comparisons against state-of-the-art DG methods for person ReID on evaluation protocol (ii) and (iii). Protocol (ii) and (iii) are both multiple-to-one setting which used in RaMoE (Dai et al., 2021b) and M3L Zhao et al. (2021) respectively. Unit DRO beats them in both these two settings.

Method	Avg		Target:PRID		Target:GRID		Target:VIPeR		Target:iLIDs	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SNR	64.6	55.4	60.0	49.0	41.3	30.4	65.0	55.1	91.9	87.0
RaMoE	71.3	63.0	66.8	56.9	53.9	43.4	72.2	63.4	92.3	88.4
Ours	**76.1**	**68.0**	**79.4**	**71.3**	**59.8**	**50.2**	**77.1**	**68.9**	88.2	81.7

Table 12: Comparisons against state-of-the-art DG methods for person ReID on evaluation protocol (iv). Unit DRO outperforms RaMoE (Dai et al., 2021b) in protocols (iv) by a large margin.

Methods	Average		VIPeR (V)				PRID (P)				GRID (G)				i-LIDS (I)			
	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
Baseline(BN)	56.7	65.8	49.9	69.8	75.1	59.0	54.1	78.3	85.5	64.9	46.2	67.5	75.3	55.6	76.7	93.3	96.3	83.8
Baseline(IN)	58.8	67.6	53.6	74.3	81.7	63.1	63.8	82.5	87.7	72.8	40.7	64.5	73.0	51.0	77.0	91.3	94.8	83.3
Baseline(BIN-half)	56.9	65.1	51.5	73.3	78.8	61.2	67.4	84.3	89.9	74.9	35.8	51.0	61.0	43.9	72.7	90.2	95.0	80.5
Baseline(BIN (Nam & Kim, 2018))	62.2	70.4	55.7	74.2	80.2	64.6	67.8	85.5	89.3	76.1	46.4	64.6	74.9	55.7	78.5	94.3	97.5	85.2
Baseline(MetaBIN (Choi et al., 2021))	63.7	72.0	58.3	77.8	82.9	67.5	67.8	86.5	92.1	76.3	48.9	69.7	78.2	58.6	79.8	93.8	97.3	85.7

Table 13: Comparisons over baselines integrated with different batch normalization and instance normalization methods. 'BIN-half' is a channel-wise combination of BN and IN.

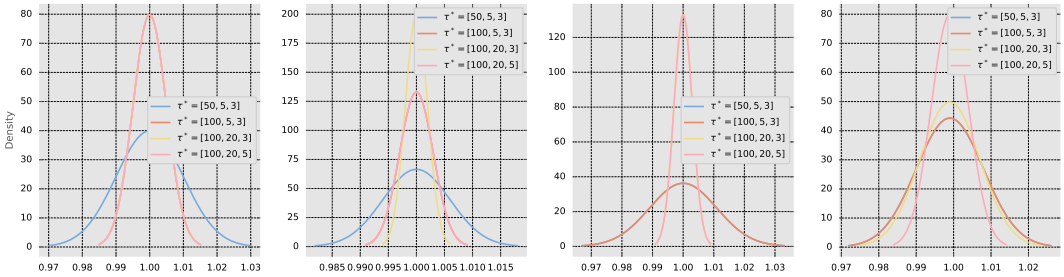
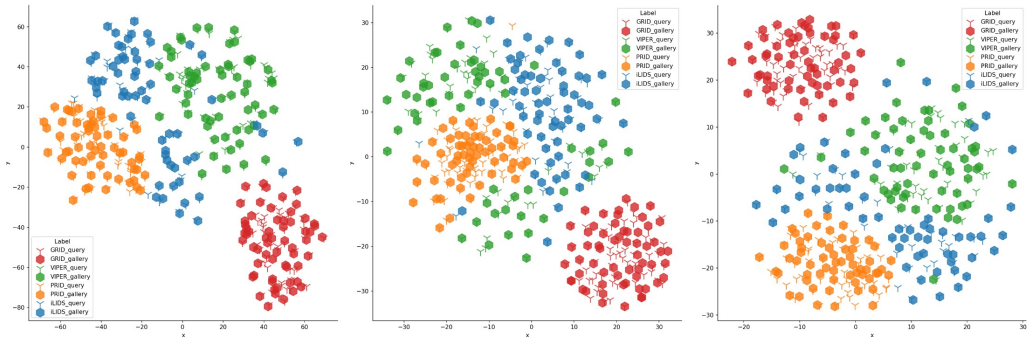


Figure 5: Distribution visualization of sample weights ($|\mathcal{M}| = 800$ by default) of steps $[1000, 50000, 100000, 150000]$ (from left to right). The horizontal axis represents the weight, and the vertical axis represents the density. $\tau^* = [\tau_1, \tau_2, \tau_3]$ means $\tau^* = \tau_1$ initially and decayed to τ_2 and τ_3 at 40 and 70 epochs.

results demonstrate a common pattern, DualNorm (Jia et al., 2019) retain large domain divergences and its embedding vector is far from "domain invariant". MetaBIN (Choi et al., 2021) utilizes a complex framework and expensive demographics, which is able to reduce domain divergences. Unit DRO achieves a comparable or even better result than MetaBIN (Choi et al., 2021) in a simpler and cheaper paradigm. **Consider discriminative capability.** Figure 8 visualizes the probe and gallery samples on four test datasets individually. The utopian discrimination result is that every query-galley pair has the closest intra-identity distance and a relatively large inter-identity distance. Figure 8d and Figure 8b shows that Unit DRO performs well matching on the i-LIDS and the PRID dataset. However, we observe an interesting phenomenon, termed "Inter-Identity Cluster". Specifically, probes and galleries of different identities came together in some clusters. These clusters are always seen on the VIPeR and the GRID datasets (Figure 8a and Figure 8b), which reveals why Unit DRO performs much poorly on these two datasets.



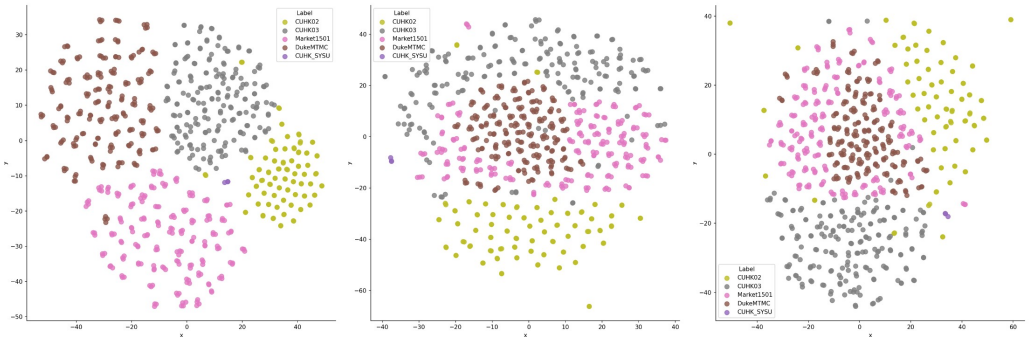
(a) (b) (c)

Figure 6: The t-SNE visualization of embedding vectors on four unseen target datasets. Query and gallery samples are expressed in different shapes. Best viewed in color.

C.4 Implementation of Domain divergence measurement

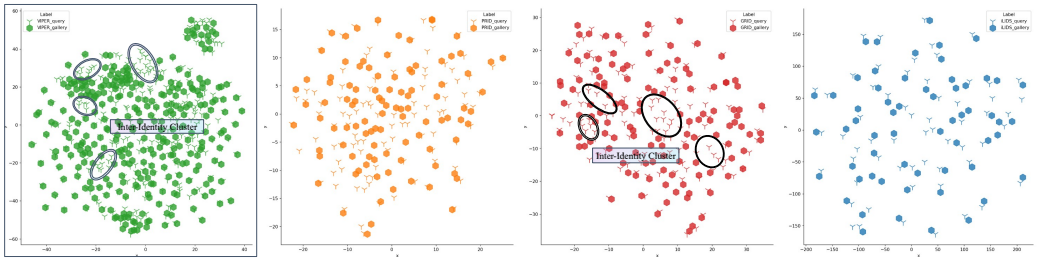
In general, MMD distance (Tolstikhin et al., 2016) is defined by the idea of representing distances between distributions as distances between mean embeddings of features. Following MMFA model (Lin et al., 2018), we use the RBF characteristic kernel with bandwidth $\alpha_2 = 1 : 5 : 10$ to compute the MMD distance. $\mathcal{A}$-distance (Long et al., 2015) can be approximated as $d_{\mathcal{A}}(d_i, d_j) = 2(1 - 2\sigma)$, where σ is the error of a two-sample classifier distinguishing features of samples from two distinct domains d_i, d_j. Note that we have not only two domains. To measure the $\mathcal{A}$-distance or MMD-distance on four unseen datasets, we calculate the average mean distance of each domain pair, namely

$$\mathcal{A}(U) = \frac{1}{6} \sum_{i=1}^{4} \sum_{j=i+1}^{4} \mathcal{A}(d_i, d_j). \tag{11}$$



(a) (b) (c)

Figure 7: The t-SNE visualization of embedding vectors on five training datasets. Best viewed in color.



(a) VIPeR (b) PRID (c) GRID (d) i-LIDS

Figure 8: The t-SNE visualization of embedding vectors on four test datasets individually. Best viewed in color.

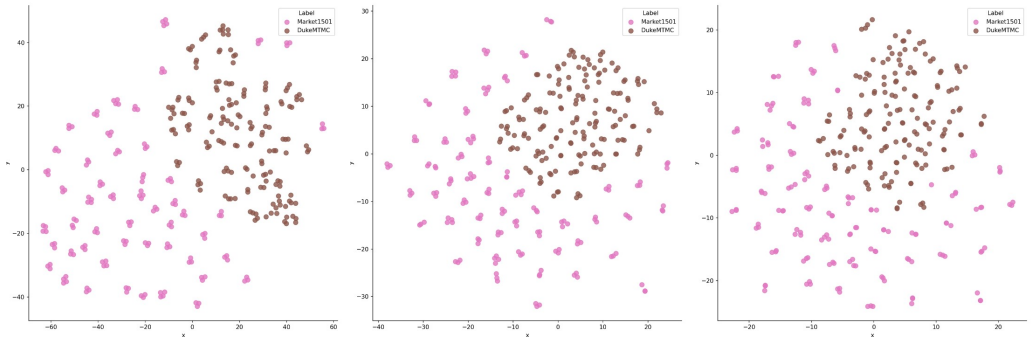
C.5 Additional domain divergence measurement results

The MMD-distance between every dataset pair of all the datasets is plotted in Figure 10a. The MMD-distance between every dataset pair of five training datasets is shown in Figure 10b and that of four test datasets is shown in Figure 10c. For the training dataset, we find that the CUHK02 dataset remains large divergences with almost all the other domains. Namely, the CUHK02 dataset is more likely to be an out-of-distribution dataset and is more important to generalization capability. Hence, Unit DRO assigns relatively higher weights for samples in the CUHK02 dataset. In terms of test datasets, the GRID dataset maintains the largest MMD distance among these datasets. It is also the reason why Unit DRO performs badly on the GRID dataset. However, domain divergence is not the only factor that affects generalization performance. Figure 10c shows that the PRID dataset has a larger domain divergence than VIPeR. However, Unit DRO performs better on the PRID dataset than on the VIPeR dataset. We exploit the underlying reasons in Section C.6.

C.6 Error set analysis

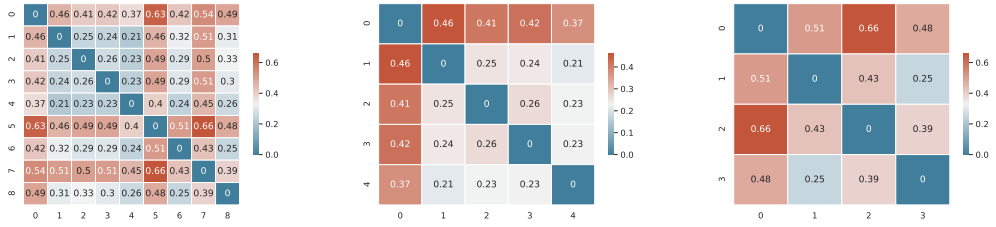
We select some successfully retrieved pairs[6] and failure cases in Figure 11. We plot query images and corresponding gallery images in the top and bottom in these figures respectively. Figure 11a shows that query and gallery images in the failure case have a relatively large view change (front and back shooting). In contrast, successfully retrieved query-gallery pairs in Figure 11b have almost the same camera view. This result shows that Unit DROcannot well overcome the challenges brought by changes in the camera view. Namely, we can leverage advanced structure in supervised ReID methods to eliminate the sensitivity of Unit DRO to camera perspective. Figure 11c and Figure 11b show that the camera perspective changes between query and gallery set in the PRID dataset are small, which is

[6]We name a query-gallery images pair "successfully retrieved pair" such that the distance between the query image and its corresponding gallery image is the closest in all of the gallery images. Other pairs are named failure cases.



(a) (b) (c)

Figure 9: The t-SNE visualization of embedding vectors on Market1501 (Zheng et al., 2015) and DukeMTMC-ReID (Zheng et al., 2017). Model are trained on Market-Duke benchmark. Best viewed in color.



(a) (b) (c)

Figure 10: The heatmaps of MMD distance on training and test dataset pairs. (a, b): 0: CUHK02, 1: CUHK03, 2: Market1501, 3: DukeMTMC, 4: CUHK-SYSU, 5: GRID, 6: VIPeR, 7: PRID, 8: i-LIDS. (c): 0: GRID, 1: VIPeR, 2: PRID, 3: i-LIDS.

one of the reasons why Unit DRO performs much better on the PRID dataset than the GRID dataset[7]. According to error set analysis, we can explain the phenomenon mentioned in Section C.5 that Unit DRO performs superior on the dataset with a relatively high domain divergence (the PRID dataset) than the dataset with low domain divergence (the VIPeR dataset). Figure 11e shows that query-gallery pairs in the VIPeR dataset always maintain camera view changes that more than $90°$, which is more hard to identify compared to the PRID dataset. Finally, the i-LIDS dataset has the lowest MMD distances with other datasets and the camera perspective changes between its query-gallery pairs are always small. These good properties enable Unit DRO to achieve a rank-1 accuracy of 80.7 on the i-LIDS dataset. So far, we can conclude that all of the domain style divergence, intrinsic characteristics of datasets (camera perspective changes), and model capacity [8] affect the performance of DG ReID and DGWD-ReID methods.

[7] Another reason is the domain divergence, as we discussed in Section C.5.

[8] larger backbones and advanced learning paradigm always attains better generalization capability.

Figure 11: Error set analysis. (a): Failture cases in the GRID datasets. (b) Successful retrieved pairs in the GRID datasets. (c) Failture cases in the PRID datasets. (d) Successful retrieved pairs in the PRID datasets. (e): Failture cases in the VIPeR datasets. (f) Successful retrieved pairs in the VIPeR datasets. (g) Failture cases in the i-LIDS datasets. (h) Successful retrieved pairs in the i-LIDS datasets.