Generalizable Person Re-identification Without Demographics – Appendix –

A. Supplementary disadvantages of DG ReID setting

Privacy Risks. 1) The specialized domain information of mostly common used ReID datasets, *e.g*. Market1501 and CUHK is described by the author in the article, such as the campus of Tsinghua University and the Chinese University of Hong Kong. In practice, different from these public datasets for research purposes, ReID domain/cameras information acquisition is often completed automatically, based on some inherent physical properties of cameras (such as network cameras MAC address). With the increasing development of the Internet of Things, the utilization of cameras' physical properties will definitely increase the leakage risk of location information, as each device will be online in the future. 2) Even if the location information is eliminated, the relationships between different people are still exposed by the annotated domain information. If two-person IDs are marked with the same camera, their social relationships can be simply measured by counting their occurrence frequency in the same camera. So the private social relationships among persons may be leaked, which may be more sensitive than location information. In summary, removing these domain/camera information in data collection is definitely a safer and efficient usage mode of the person images. **Manually collected domain labels may be noisy or suboptimal.** Previous studies in this area have defined each domain or each camera simply as a dataset and focused on developing learning techniques. However, how best to partition domains that can benefit the learning process most is still unclear [48]. Defining each domain or each camera as a dataset is too simple and sometimes unreasonable. As illustrated in Figure 4, images in each column is from the same datasets and different camera, however, we can see that some images in different datasets and also different camera are very similar (images in each row). In contrast, some images in one dataset are very dissimilar (images in each column). Namely, the artificially defined domain ID and camera ID are not optimal for downstream learning, and finding more rational domain partitions for DGWD-ReIDtasks is an open and important problem.

B. Related Works

DG-ReID. Generalizable methods are recently proposed to learn invariant representations that can generalize to unseen domains [7,47,59–61]. Existing methods mainly utilize domain divergence minimization strategies or a meta-learning pipeline. In view of the current research trend (Table 4), most methods rely on demographics to learn invariant features. Though existing strong baseline [28], normalization [23], and augmentation methods [56] require no demographics, they are plug-and-play modules and thus orthogonal to the proposed Unit-DRO. Different from existing studies, DGWD-ReID adds a strict restriction on demographics and has ambitious targets that **"can we learn invariant features even without demographics? can we partition domains better?"**.

Fairness without Demographics. Methods in Fairness [13] 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. Though there are several works considering *fariness without demographics* [8, 31], they mostly evaluate their algorithms in datasets with predefined distribution shifts. Note that DGWD-ReID 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 simultaneously. **Domain generalization.** Domain/Out-of-distribution generalization [36, 59] aims to learn a model that can extrapolate well in unseen environments. Representative methods like Invariant Risk Minimization (IRM) [2] and its variant [1] are recently proposed to tackle this

Figure 4. Samples on ReID datasets.

Method	Source	Domain	Camera
DIR-ReID [61]	Arxiv 21	✓	✓
MetaBIN [7]	CVPR 21	✓	✓
M3L [62]	CVPR 21	✓	
DMG-Net [3]	CVPR 21	✓	✓
RaMoE [9]	CVPR 21	✓	
CBN [67]	ECCV 20		✓
CAIL [33]	ECCV 20	✓	✓
QAConv [27]	ECCV 20	Backbone	
SNR [23]	CVPR 20	Normalization	

Table 4. The current research trend of DG-ReID.

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.

Unsupervised-domain adaptation Person ReID. Unsupervised Domain Adaptation (UDA) technologies have great progress [41] 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 [21,33], features [53] or metrics [40]. Another group of UDA-based methods [15,58] 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 [5] 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 [42] for a review). Recent studies consititute $\mathcal{Q}$ analytically, such as using moment constraint [10, 37], f-divergence [20, 35], Wasserstein/MMD ball [46, 49] or coarse-grained mixture models [12, 39]. 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.

C. Experiments

C.1. Experimental Setup

Datasets. Following [22, 47, 61], we evaluate the Unit-DRO with multiple data sources (MS), where source domains cover five large-scale ReID datasets, including CUHK02 [25], CUHK03 [26], Market1501 [63], DukeMTMC-ReID [64], and CUHK-SYSU PersonSearch [55]. The unseen test domains are VIPeR [16], PRID [18], QMUL GRID [30], and i-LIDS [54]. Details of the training datasets are summarized in Table 7 and the test datasets are summarized in Table 8. The average rank-k (R-k) accuracy and mean Average Precision (mAP) over 10 random splits are reported based on the evaluation protocol

Baselines We compare our model with 1) **DG-ReID** methods, including AugMining [50], DIMN [47], DualNorm [22], SNR [23], DDAN [6], DIR-ReID [61], and MetaBIN [7]; and 2) **CD-ReID** methods, including CrossGrad [45], QAConv [27], L2A-OT [66], OSNet-AIN [65], SNR [23], DIR-ReID [61], and MetaBIN [7].

C.2. Results

DG-ReID Protocols. We summarize the detailed difference of different protocols in Tab. 5. As shown in Tab. 11, 9, 10 , Unit-DRO outperforms other methods with a clear margin in both average mAP and Rank-1 accuracy, which demonstrate the robustness of the proposed Unit-DRO across different evaluation protocols.

Protocol	Source	Target	Backbone	Augmentation
(1)	M/D	D/M+V+P+G+I	ResNet-50	Color-Jittering
(2)	MS+D+M (train)	C3	ResNet-50	None
(3)	M+D+MT	C3	ResNet-50	Color-Jittering
(4)	M+D+C3+MT	V+P+G+I	ResNet-50	Color-Jittering

Table 5. Summary of different DG-ReID protocols. (M:market1501, C2: Cuhk02, C3: Cuhk03, D: DukeMTMC, MT: MSMT17, CS: CUHK-SYSU, V: ViPeR, P: PRID, G: GRID, I: i-LIDS)

Domain	PACS					VLCS				
	A	C	P	S	Avg	C	L	S	V	Avg
IRM	85.7 ± 1.0	79.3 ± 1.1	97.6 ± 0.4	75.9 ± 1.0	84.6	97.6 ± 0.5	64.7 ± 1.1	69.7 ± 0.5	76.6 ± 0.7	77.2
Group-DRO	88.2 ± 0.7	82.4 ± 0.8	97.7 ± 0.2	80.6 ± 0.9	87.2	97.8 ± 0.0	66.4 ± 0.5	68.7 ± 1.2	76.8 ± 1.0	77.4
MIXUP	87.4 ± 1.0	80.7 ± 1.0	97.9 ± 0.2	79.7 ± 1.0	86.4	98.3 ± 0.3	66.7 ± 0.5	73.3 ± 1.1	76.3 ± 0.8	78.7
DANN	86.4 ± 1.4	80.6 ± 1.0	97.7 ± 0.2	77.1 ± 1.3	85.5	95.3 ± 1.8	61.3 ± 1.8	74.3 ± 1.0	79.7 ± 0.9	77.7
Unit-DRO	**88.3 ± 0.1**	**84.8 ± 0.1**	96.4 ± 0.1	**82.1 ± 0.1**	**87.9**	**98.1 ± 0.1**	**68.0 ± 0.0**	**71.3 ± 0.1**	**78.8 ± 0.0**	**79.1**

Table 6. Results for general DG tasks.

General Domain Generalization. Apart from person ReID, we also compare Unit-DRO with other general domain generalization methods, including IRM [2], Group-DRO [43], DANN [14], and Mixup [56]. For fair comparison, we use test-domain validation, which is one of the most important methods in [17]. Specifically, this strategy is an oracle selection one since we choose the model maximizing the accuracy on a validation set that follows the distribution of the test domain. As shown in Table. 6, Unit-DRO consistently outperforms all baseline methods for general domain generalization tasks with a clear margin without using demographics.

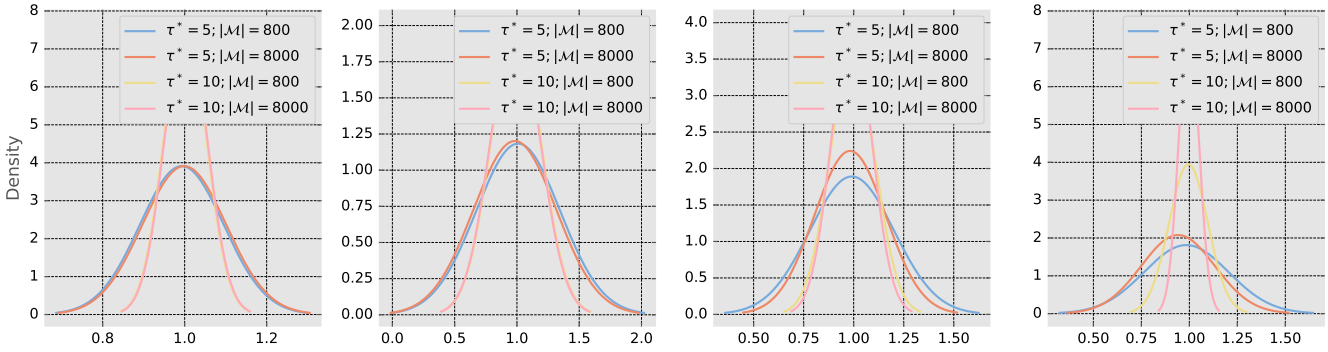


Figure 5. Visualizing the distribution of the sample weight at $1k, 5k, 10k, 20k$ steps, respectively (from left to right).

C.3. Analysis

Sample Weights. Considering that the proposed Unit-DRO will upweight and downweight different samples, we thus visualize the distribution of sample weight to better understand the influences of different components. Specifically, during training, we save the mean and variance of sample weights for every 1k iterations/steps. We assume these weights follow the 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 under the following two settings: 1) **sample weights without the weight queue.** In this case, these weights are normalized in their batches, so the mean of all distributions here is 1. As shown in Figure 2, we have already discussed this setting in the former sections; 2) **sample weights with different length of weight queue $|\mathcal{M}|$.** In Figure 5, we show the distribution of sample weight at $1k, 5k, 10k, 20k$ training steps, which indicates how the weight distribution changes during training. Intuitively, we may need a large $|\mathcal{M}|$ to better estimate $\mathbb{E}_{\mathcal{P}}[e^{\ell(x,y;\theta)/\tau^*}]$. However, as $|\mathcal{M}|$ becomes larger, the estimation will become inaccurate. For example, we consider an extreme case: $|\mathcal{M}| = T - 1$ and then the queue absolutely contains all training data. Therefore, 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 too much old weights which is unsuitable for the current model. Figure 5 depicts the phenomenon, where the distribution with a larger $|\mathcal{M}|$ always has smaller μ. For more visualization results and discussions about the distribution diagrams of the multi-step τ^*, please see Appendix C.3.

Distribution diagrams of step τ^* Compared to a constant τ^*, weights with step τ^* always have low δ and are more stable.

Additional t-SNE Visualization Results Figure 7 shows the t-SNE results on four unseen datasets. Figure 8 shows the t-SNE results on five training datasets and Figure 10 shows the t-SNE results on the Market-Duke benchmark. All of these results demonstrate a common pattern, DualNorm [22] retain large domain divergences and its embedding vector is far from "domain invariant". MetaBIN [7] 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 [7] in a simpler and cheaper paradigm.

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	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 9. 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 10. 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 [9] and M3L [62] 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 11. Comparisons against state-of-the-art DG methods for person ReID on evaluation protocol (iv). Unit DRO outperforms RaMoE [9] in protocols (iv) by a large margin.

Consider discriminative capability. Figure 9 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 9d and Figure 9b 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

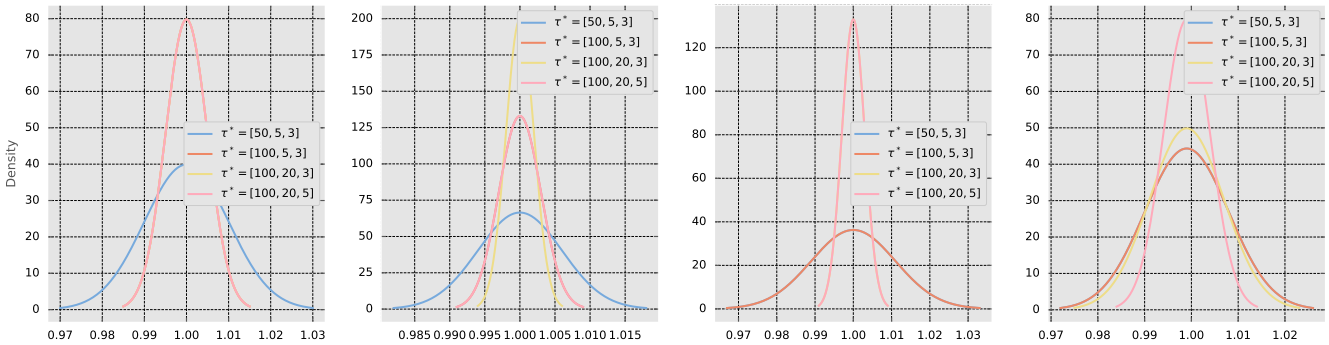
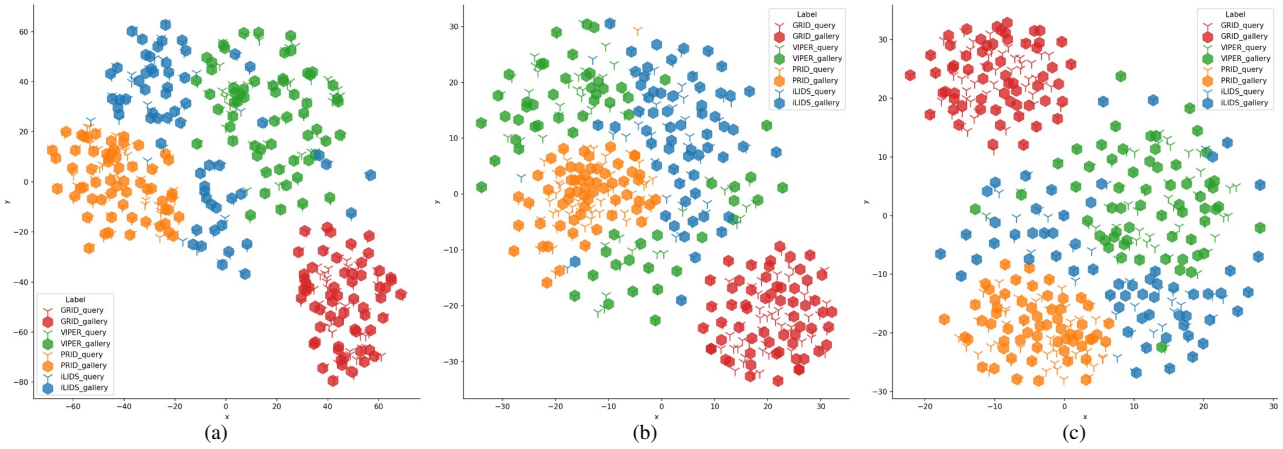


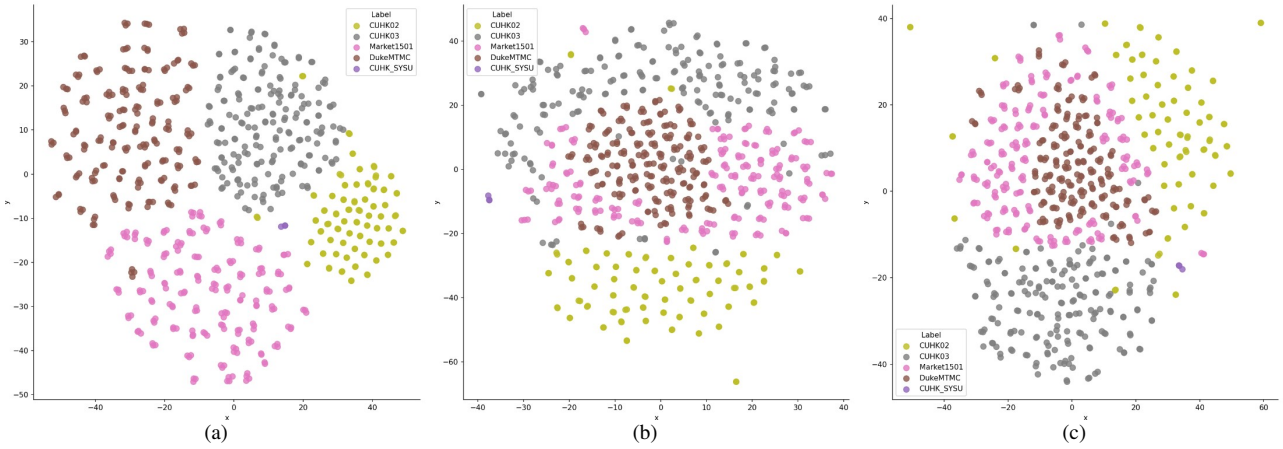
Figure 6. Distribution visualization of sample weights ($|\mathcal{M}| = 800$ by default) of steps $[1000, 50000, 100000, 150000]$ (from left to right). The horizontal axis represents the weight, and the vertical axis represents the density. $\tau^* = [\tau_1, \tau_2, \tau_3]$ means $\tau^* = \tau_1$ initially and decayed to τ_2 and τ_3 at 40 and 70 epochs.

of different identities came together in some clusters. These clusters are always seen on the VIPeR and the GRID datasets (Figure 9a and Figure 9b), which reveals why Unit-DRO performs much poorly on these two datasets.



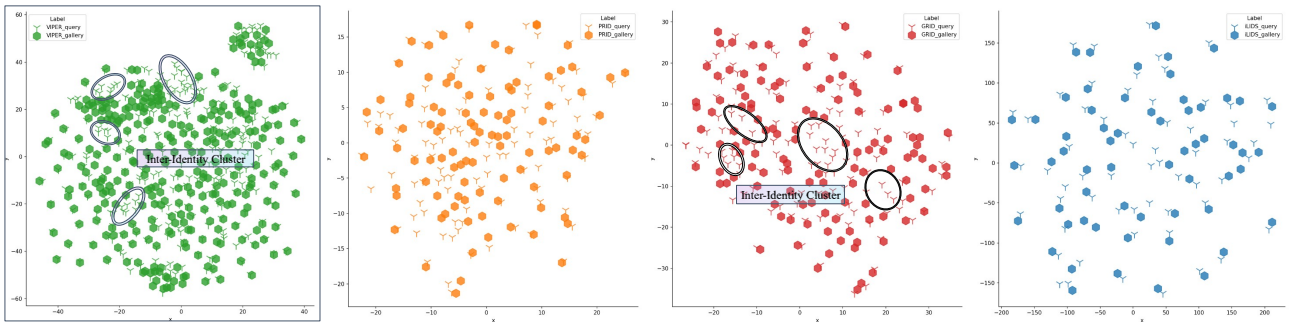
(a) (b) (c)

Figure 7. 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.



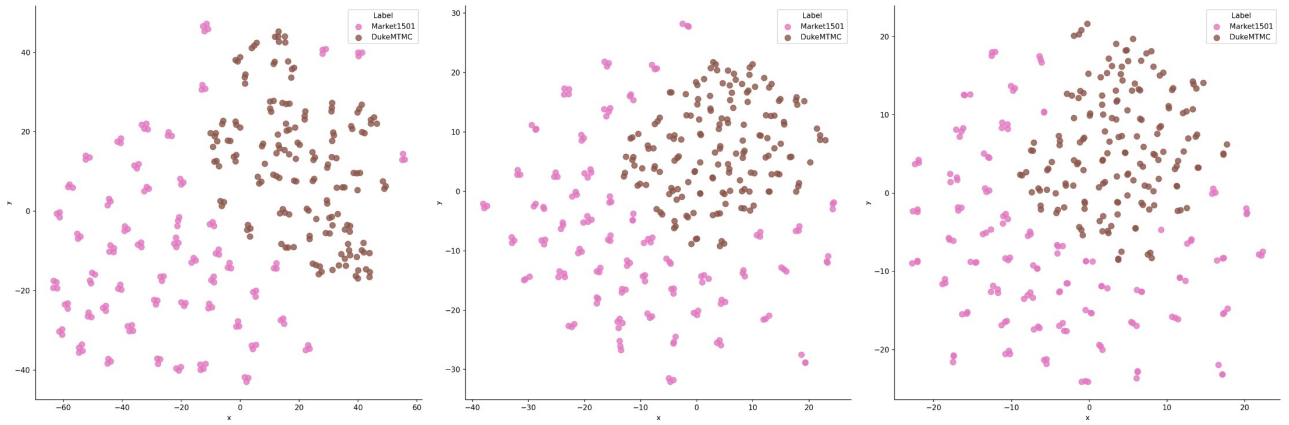
(a) (b) (c)

Figure 8. 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 9. The t-SNE visualization of embedding vectors on four test datasets individually. Best viewed in color.



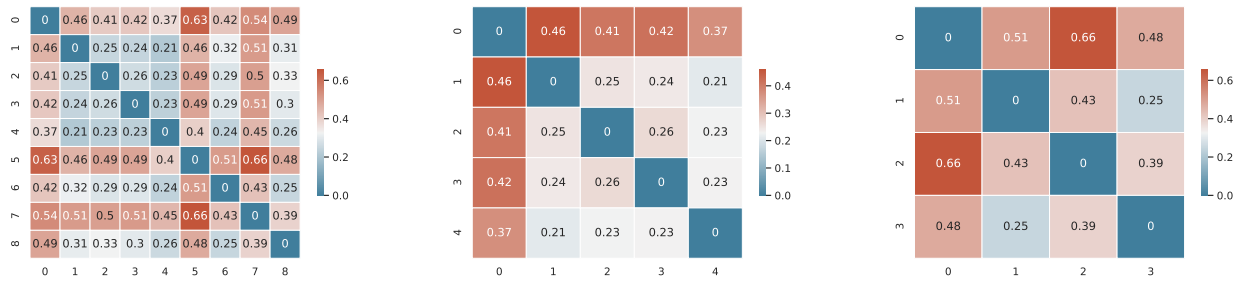
(a) (b) (c)

Figure 10. The t-SNE visualization of embedding vectors on Market1501 [63] and DukeMTMC-ReID [64]. Model are trained on Market-Duke benchmark. Best viewed in color.

Implementation of Domain divergence measurement In general, MMD distance [51] is defined by the idea of representing distances between distributions as distances between mean embeddings of features. Following MMFA model [29], we use the RBF characteristic kernel with bandwidth $\alpha_2 = 1 : 5 : 10$ to compute the MMD distance. $\mathcal{A}$-distance [32] 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

$$\mathcal{A}(U) = \frac{1}{6} \sum_{i=1}^{4} \sum_{j=i+1}^{4} \mathcal{A}(d_i, d_j). \tag{10}$$

Additional domain divergence measurement results The MMD-distance between every dataset pair of all the datasets is plotted in Figure 11a. The MMD-distance between every dataset pair of five training datasets is shown in Figure 11b and that of four test datasets is shown in Figure 11c. 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 11c 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.



(a) (b) (c)

Figure 11. 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.