

# Supplemental Material

Eugene P.W. Ang<sup>1,2</sup>  
phuaywee001@e.ntu.edu.sg  
Lin Shan<sup>1</sup>  
shan.lin@ntu.edu.sg  
Alex C. Kot<sup>1</sup>  
eackot@ntu.edu.sg

<sup>1</sup> Rapid-Rich Object Search (ROSE) Lab  
Nanyang Technological University  
Singapore  
<sup>2</sup> Defence Science and Technology  
Agency  
Singapore

## 1 Comparison with M<sup>3</sup>L with C-NP as Source Domain

To ensure a comprehensive comparison with [8], we present additional experimental results using the new protocol of CUHK03 [6], C-NP, as a source domain, which were omitted from the main section due to space constraints. Table 1 summarizes each new experiment, detailing the source domains used, target domain and ID/image counts for each.

Sources	Target	Combined-Src-IDs	Combined-Src-Images	Target-Images
C-NP+D+MS	M	2,510	56,508	19,281
C-NP+M+MS	D	2,559	52,922	19,889
C-NP+D+M	MS	2,220	36,823	93,820

Table 1: Details for experiments involving C-NP in Sources.

Table 2 compares our DEX with [8]. Under this different experimental setting, our method continues to surpass leading state-of-the-art methods by significant margins.

## 2 Detailed Differences between DEX and ISDA

Our method is inspired by ISDA [4], a recent work on deep feature augmentation. Still, it is significantly different as we found it impractical to apply it, originally proposed for image classification tasks, directly to multi-source domain generalization for Person ReID (DG-ReID). In the original formulation, a different covariance matrix is stored and updated for each class corresponding to person identities in ReID. However, in DG-ReID the number of classes grows large when multiple source datasets are merged for training, imposing a strong memory overhead as each matrix takes up  $O(n^2)$  space with feature dimension  $n = 2048$  in our case. In our case, the experiment with the lowest number of training classes is already close to 2,500, making direct application of [4] infeasible.

For problems with a large number of classes such as ImageNet [5], ISDA approximated the class-conditional covariance by storing just the matrix diagonals, reducing the per-class memory overhead to  $O(n)$  while still reaping the benefits. However, even with this approximation, using a class-conditional approach could not consistently outperform our baseline

Sources	Method	Market-1501	
		Rank-1	mAP
C-NP+D+MS	DualNorm <sub>50</sub>	76.5	48.5
	QAConv <sub>50</sub>	68.6	39.5
	M <sup>3</sup> L (ResNet-50)	76.5	51.1
	M <sup>3</sup> L (IBN-Net50)	78.3	52.5
	DEX (Ours)	<b>79.7</b>	<b>53.3</b>
Sources	Method	DukeMTMC-reID	
		Rank-1	mAP
C-NP+M+MS	DualNorm <sub>50</sub>	66.1	48.8
	QAConv <sub>50</sub>	64.9	43.4
	M <sup>3</sup> L (ResNet-50)	67.1	48.2
	M <sup>3</sup> L (IBN-Net50)	67.2	48.8
	DEX (Ours)	<b>72.1</b>	<b>53.5</b>
Sources	Method	MSMT17_V2	
		Rank-1	mAP
C-NP+D+M	DualNorm <sub>50</sub>	34.4	13.5
	QAConv <sub>50</sub>	29.9	10.0
	M <sup>3</sup> L (ResNet-50)	32.0	13.1
	M <sup>3</sup> L (IBN-Net50)	37.1	15.4
	DEX (Ours)	<b>42.7</b>	<b>17.9</b>

Table 2: Comparison of our DEX against the most recent state-of-the-art DG Person ReID, for the experiments that use CUHK-NP in Sources.

model. Looking deeper, we believe this is because DG-ReID datasets have far fewer samples per class compared to those used in the original work, making estimation of per-class covariance matrices unstable. Datasets used in [14] have 500 samples per class in the case of CIFAR-100 and 5000 samples per class in CIFAR-10 [14]. In ImageNet [14], a majority of the classes have 1300 samples each, with the lowest being 732. In stark contrast, the median samples-per-identity is between 10 and 25 for ReID, *over an order of magnitude smaller*. Table 3 shows a detailed comparison between the dataset statistics of their datasets and our ReID datasets. Such drastically different distributions indicate that class-conditional covariances may not be the most stable solution.

## 3 Model Training Details

In this section we describe the full details of our training method.

### 3.1 Backbone model

We use an ImageNet [14] pre-trained ResNet-50 [14] with instance normalization applied on the first three out of four Bottleneck blocks, as prescribed in DualNorm [14]. The final layer stride is reduced to 1 (originally 2). A batch normalization layer with no bias is applied to the final layer features [14].

Dataset(s)	Mean	Median	Min	Max
CIFAR-10	5000	5000	5000	5000
CIFAR-100	500	500	500	500
ImageNet [14]	1281.2	1300	732	1300
Market-1501 (M)	17.2	15	2	72
DukeMTMC-reID (D)	23.5	20	6	426
CUHK03 (C)	9.6	10	6	10
MSMT17_V2 (MS)	30.8	25	6	392
C+D+MS	20.0	13	6	426
C+D+M	15.1	10	2	426
C+M+MS	18.6	10	2	392
D+M+MS	24.9	20	2	426

Table 3: Comparison of samples per class between the datasets used in previous work on deep feature augmentation [14] and those from Person ReID. ReID datasets have far fewer samples per class, making estimation of class-conditional covariance matrices unstable.

### 3.2 Loss functions

For regularization, the softmax loss  $\mathcal{L}_{soft}$  applies label smoothing with  $\varepsilon = 0.1$ . The triplet loss is defined as such:

$$\mathcal{L}_{tri} = \sum_{a,p,n} [\delta(\mathbf{x}_a, \mathbf{x}_n) - \delta(\mathbf{x}_a, \mathbf{x}_p) + \gamma]_+$$

where  $[\cdot]_+ = \max(\cdot, 0)$ ,  $(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n)$  refer to the triplets (anchor, positive, negative) that are found in the batch,  $\gamma = 0.3$  is the triplet margin, and  $\delta$  is a metric; in our case, we use Euclidean distance. Our batch sampling scheme follows [14], selecting  $k = 4$  samples per PID in the batch to ensure sufficient triplets for training. The center loss is defined as such:

$$\mathcal{L}_{cen} = \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{c}_{y_i}\|^2$$

where  $i$  indexes a batch size of  $m$ ,  $y_i$  is the class label for  $x_i$  and  $c$  are the class centroids, updated dynamically during training. To recap, the overall loss function for training a baseline model without DEX is:

$$\mathcal{L}_{base} = \beta_{soft} \mathcal{L}_{soft} + \beta_{tri} \mathcal{L}_{tri} + \beta_{cen} \mathcal{L}_{cen}$$

with  $\beta_{soft} = 1, \beta_{tri} = 1, \beta_{cen} = 5 \times 10^{-4}$ . Applying DEX or DEXLite, the overall loss then changes to:

$$\mathcal{L}_{overall}(t) = \mathcal{L}_{DEX/DEXLite}(\lambda_t) + \beta_{tri} \mathcal{L}_{tri} + \beta_{cen} \mathcal{L}_{cen}$$

with the strength of augmentation tempered by epoch  $t$  as described in the main paper.

### 3.3 Input pre-processing

We resized the input to  $[384, 128]$ , pad it 10 pixels around with zeros and crop out  $[384, 128]$ . We then apply a horizontal flip to the image with probability  $p = 0.5$  and random-erasing (RE) augmentation with a probability  $p = 0.1$ .

### 3.4 Other training details

We train for 60 epochs ( $T = 60$ ) with a batch size of 32 using the Adam optimizer, linearly warming up the learning rate  $\eta$  from 0 to  $1.75 \times 10^{-4}$  in 10 epochs. Afterwards,  $\eta$  is reduced by a factor of 0.1 at epochs 30 and 55.

## References

- [1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [3] Jieru Jia, Qiuqi Ruan, and Timothy M. Hospedales. Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice. In *Proc. British Machine Vision Conference (BMVC)*, 2019.
- [4] Alex Krizhevsky and Geoffrey Hinton. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, University of Toronto, 2009.
- [5] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019.
- [6] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1288–1296, 2016.
- [7] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Cheng Wu, and Gao Huang. Implicit Semantic Data Augmentation for Deep Networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [8] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.