Supplementary Material: Deep Frequency Filtering for Domain Generalization

Shiqi Lin[1*] Zhizheng Zhang[2] Zhipeng Huang[1*] Yan Lu[2] Cuiling Lan[2] Peng Chu[2]
Quanzeng You[2] Jiang Wang[2] Zicheng Liu[2] Amey Parulkar[2] Viraj Navkal[2] Zhibo Chen[1]
[1]University of Science and Technology of China [2]Microsoft
{linsq047,hzp1104}@mail.ustc.edu.cn chenzhibo@ustc.edu.cn
{zhizzhang,yanlu,culan,pengchu,quyou,jiangwang,zliu,amey.parulkar,vnavkal}@microsoft.com

1. More Datasets and Implementation Details

We evaluate the effectiveness of our proposed Deep Frequency Filtering (DFF) for Domain Generalization (DG) on **Task-1**: the close-set classification task and **Task-2**: the open-set retrieval task, *i.e.*, person re-identification (ReID). More details about the datasets and our experiment configurations are introduced in this section.

1.1. Datasets for Task-1

We use the most commonly used Office-Home [20] and PACS dataset [20]. Specifically, Office-Home consists of 4 domains (Art (Ar), Clip Art (Cl), Product (Pr), Real-World (Rw)), each consisting of 65 categories, with an average of 70 images per category, for a total of 15,500 images. PACS consists of 9991 samples in total from 4 domains (*i.e.*, Photo (P), Art Painting (A), Cartoon (C) and Sketch (S)). All these 4 domains share 7 object categories. They are commonly used domain generalization (DG) benchmark on the task of classification. We validate the effectiveness of our proposed method for generalization in close-set classification task on Office-Home and PACS. Following the typical setting, we conduct experiments on this dataset under the leave-one-out protocol (see Table 1 Protocol-1 and Protocol-2), where three domains are used for training and the remaining one is considered as the unknown target domain.

1.2. Datasets for Task-2

Person re-identification (ReID) is a representative open-set retrieval task, where different domains and datasets do not share their label space. We employ existing ReID protocols to evaluate the generalization ability of our method. i) For Protocol-3 and Protocol-4, we also follow the leave-one-out protocols as in [27, 48]. Among the four datasets (CUHK-SYSU (CS) [40], MSMT17 (MS) [39], CUHK03 (C3) [25] and Market-1501 (MA) [50]), three are selected as the seen domain for training and the remaining one is selected the unseen domain data for testing. Differently, Protocol-3 only adopts the training set of seen domains for model training while in Protocol-3, the testing set of the seen domains are also included for training model. ii) For Protocol-5 in Table 1, several large-scale ReID datasets *e.g.*, CUHK02 (C2) [24], CUHK03 (C3) [25], Market-1501 (MA) [50] and CUHK-SYSU (CS) [40], are viewed as multiple source domains. Each small-scale ReID dataset including VIPeR [10], PRID [15], GRID [28] and iLIDS [51] is used as an unseen target domain, respectively. To comply with the General Ethical Conduct, we exclude DukeMTMC from the source domains. The final performance is obtained by averaging 10 repeated experiments with random splits of training and testing sets.

Table 1. The evaluation protocols. "Com-" refers to combining the train and test sets of source domains for training. "Pr", "Ar", "Cl", "Rw" are short for the Product, Art, Clip Art and Real-World domains in Office-Home dataset [20], respectively. "P", "A", "C", "S" are short for the Photo, Painting, Cartoon, Sketch domains in PACS dataset [20], respectively. "MA", "CS", "C3", "MS" denote Market-1501 [50], CUHK-SYSU [40], CUHK03 [25], MSMT17 [39], respectively. Note that for person ReID, the commonly used DukeMTMC [53] has been withdrawn by its publisher, is thus no longer used.

Task	Setting	Training Data	Testing Data
Close-set classification	Protocol-1	Cl,Pr,Rw	Ar
		Ar,Pr,Rw	Cl
		Ar,Cl,Rw	Pr
		Ar,Cl,Pr	Rw
	Protocol-2	C,P,S	A
		A,P,S	C
		A,C,P	S
		A,C,S	P
Open-set retrieval	Protocol-3	CS+C3+MS	MA
		MA+CS+MS	C3
		MA+CS+C3	MS
	Protocol-4	Com-(CS+C3+MT)	MA
		Com-(MA+CS+MS)	C3
		Com-(MA+CS+C3)	MS
	Protocol-5	Com-(MA+C2+C3+CS)	PRID
			GRID
			VIPeR
			iLIDs

*This work was done when Shiqi Lin and Zhipeng Huang were interns at Microsoft Research Asia.

1.3. Networks

Following the common practices of domain generalizable classification (Task-1) [2, 22, 33, 55] and person ReID (Task-2) [3, 6, 17, 26], we build the networks equipped with our proposed Deep Frequency Filtering (DFF) for these two tasks on the basis of ResNet-18 and ResNet-50, respectively. As introduced in the Sec. 3.4 of our manuscript, we evaluate the effectiveness of our proposed DFF based on the two-branch architecture of Fast Fourier Convolution (FFC) in [5]. In particular, we adopt our DFF operation to the spectral transformer branch of this architecture. Unless otherwise stated, the ratio r in splitting $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ into $\mathbf{X}^g \in \mathbb{R}^{rC\times H\times W}$ and $\mathbf{X}^l \in \mathbb{R}^{(1-r)C\times H\times W}$ is set to 0.5. We conduct an ablation study on this ratio in this Supplementary as follows.

1.4. Training

Following common practices [2, 12, 22, 29, 30, 44], we adopt ResNet-18 and ResNet-50 [11] as our backbone for Task-1 and Task-2, respectively. Each convolution layer of the backbone is replaced with our DDF module. Unless specially stated, we first pretrain the models on ImageNet [32] then fine-tune them on Task-1 or Task-2, referring to the common practices [2, 3, 6, 33, 55]. We introduce our training configurations with more details in the following.

Pre-training on ImageNet. Following the common practices [11, 16, 41, 43] in this field, we adopt the commonly used data augmentation strategies including color jittering, random flipping and center cropping. The input image size is 224×224. We use the SGD optimizer with the base initial learning rate of 0.4, the momentum of 0.9 and the weight decay of 0.0001, and perform learning rate decay by a factor of 0.1 after 30, 60 and 80 epochs. A linear warm-up strategy is adopted in the first 5 epochs, where the learning rate is increased from 0.0 to 0.4 linearly. All models are trained for 90 epochs with the batchsize of 1024.

Fine-tuning on Task-1. The initial learning rate in this stage is set to 0.001. We train all models on this task using the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0001. The batch size is set to 32. Following prior works [2, 22, 31, 33], we adopt the data augmentation strategies including random cropping, horizontal flipping, and random grayscale. The input images are resized to 224×224. On the PACS dataset [20], we train the models for 3,500 iterations; and on the Office-Home [37] dataset, we train the models for 3,000 iterations. The experiment results on Office-Home have been presented in the main paper while the results on PACS are placed in this Supplementary due to the length limitation.

Table 2. Performance (classification accuracy %) comparison with the state-of-the-art methods under Protocol-2 (*i.e.*, on PACS dataset) on close-set classification task. We use ResNet-18 as backbone. Best in bold.

Method	Source→Target				Avg
	C,P,S→A	A,P,S→C	A,C,P→S	A,C,S→P	
Baseline	77.6	73.9	70.3	94.4	79.1
MMD-AAE [22]	75.2	72.7	64.2	96.0	77.0
CrossGrad [33]	78.7	73.3	65.1	94.0	77.8
MetaReg [1]	79.5	75.4	72.2	94.3	80.4
JiGen [2]	79.4	77.3	71.4	96.0	81.0
MLDG [19]	79.5	75.3	71.5	94.3	80.7
MASF [8]	80.3	77.2	71.7	95.0	81.1
Epi-FCR [21]	82.1	77.0	**73.0**	93.9	81.5
MMLD [30]	81.3	77.2	72.3	**96.1**	81.7
Ours	**82.2**	**78.5**	72.5	95.5	**82.2**

Fine-tuning on Task-2. Following the common practices for domain generalizable person ReID [6, 7, 17, 49], we adopt the widely used data augmentation strategies, including cropping, random flipping, and color jittering. We use Adam [18] optimizer with the momentum of 0.9 and weight decay of 0.0005. The learning rate is initialized by 3.5×10^{-4} and decayed using a cosine annealing schedule. The batch size is set to 128, including 8 identities and 16 images per identity. For the Protocol-3, Protocol-4 and Protocol-5, the models are trained for 60 epochs on their corresponding source datasets. Similar to previous work [29], the last spatial down-sampling in the "conv5_x" block is removed. The input images are resized to 384×128. Following [12], we use task-related loss including cross-entropy loss, arcface loss, circle loss and triplet loss. And we adopt a gradient reversal layer [9] encouraging the learning of domain-invariant features.

2. More Experiment Results

In this section, we present more experiment results to further evaluate the effectiveness of our proposed DFF.

2.1. More Experiments on the Task-1 (PACS)

We further evaluate the effectiveness of our proposed DFF on another commonly used dataset, *i.e.*, Office-Home [37], for investigating the domain generalization on the close-set classification. This dataset contains four domains (Aritistic, Clipart, Product and Real World) with 15,500 images of classes for home and office object recognition. Similar to the Protocol-1 on PACS dataset [20], we adopt a "Leave-One-Out" protocol for the evaluation on Office-Home where three domains are used for training while the remaining one is for testing. The experiment results are shown in Table 2. Our proposed DFF achieves significant improvements relative to the *Baseline* model, and outperforms the state-of-the-art methods on this dataset by a clear margin over all evaluation settings. This further demon-

Table 3. Performance comparisons of different frequency transformations. In *"Baseline"*, we take vanilla ResNet-18/-50 as the backbone models. *"Wavelet (db3)"* and *"Wavelet (Haar)"* denote the wavelet transforms with the Daubechies3 and Haar filters, respectively.

Method	Source→Target					
	MS+CS+C3→MA		MS+MA+CS→C3		MA+CS+C3→MS	
	mAP	R1	mAP	R1	mAP	R1
Base	59.4	83.1	30.3	29.1	18.0	41.9
Wavelet (db3)	61.5	83.7	30.7	29.8	18.3	42.2
Wavelet (Haar)	61.1	83.6	30.5	29.7	18.5	42.3
FFT (Ours)	**71.1**	**87.1**	**41.3**	**41.1**	**25.1**	**50.5**

Table 4. Performance comparisons of different dimensions on which the Fast Fourier Transform (FFT) is performed. *"FFT (CHW)"* refers to the models in which FFT is performed across the height (H), width (W) and channel (C) dimensions. In *"FFT (HW)"*, we just perform FFT across the height and width dimensions, *i.e.*, for each feature map independently, which is the default setting in this paper.

Method	Source→Target					
	MS+CS+C3→MA		MS+MA+CS→C3		MA+CS+C3→MS	
	mAP	R1	mAP	R1	mAP	R1
Base	59.4	83.1	30.3	29.1	18.0	41.9
FFT (CHW)	59.2	83.0	30.0	28.8	17.5	38.5
FFT(HW)	**71.1**	**87.1**	**41.3**	**41.1**	**25.1**	**50.5**

Table 5. Performance comparisons of our proposed DFF with different ratios. All models are built based on ResNet-18 for Task-1 while ResNet-50 for Task-2.

Ratio	Source→Target					
	MS+CS+C3→MA		MS+MA+CS→C3		MA+CS+C3→MS	
	mAP	R1	mAP	R1	mAP	R1
0.0	59.4	83.1	30.3	29.1	18.0	41.9
0.25	67.4	84.1	38.1	38.1	22.9	48.4
0.5(Ours)	**71.1**	**87.1**	**41.3**	**41.1**	**25.1**	**50.5**
0.75	70.8	86.8	40.7	40.6	21.0	44.9
1.0	64.2	83.4	29.3	28.1	17.6	40.4

Table 6. Performance comparisons of our proposed DFF with the corresponding ResNet baselines on ImageNet-1K classification. "*DFF-ResNet-18/-50*" denote the ResNet-18/-50 models equipped with our DFF.

Method	Parameters	GFLOPs	Top-1 Acc.
ResNet-18	11.7M	1.8	69.8
DFF-ResNet-18	12.2M	2.0	72.3
ResNet-50	25.6M	4.1	76.3
DFF-ResNet-50	27.7M	4.5	77.9

strates the effectiveness of DFF.

2.2. More Ablation Studies

Experiments with other frequency transforms. We preliminarily investigate the effectiveness of using other frequency transforms in implementing our conceptualized DFF. In particular, we replace the Fast Fourier Transform (FFT) in our proposed scheme by the wavelet transform with two widely used filters, *i.e.*, db3 and Haar. From the experiment results in Table 3, we observe that adopting the wavelet transform also delivers improvements compared to *Baseline*, but is inferior to adopting FFT. This is because the wavelet transform is a low frequency transformation such that our proposed filtering operation is performed in a local space, thus limiting the benefits of DFF.

Table 7. Performance comparisons of our proposed DFF with the state-of-the-art methods on supervised person ReID. "*Base.*" refers to the baseline model.

Model	Market-1501(MA)		MSMT17(MT)	
	mAP	R1	mAP	R1
PCB [35]	81.60	93.80	-	-
BoT [29]	85.90	94.50	-	-
MGN [38]	86.90	95.70	-	-
JDGL [52]	86.00	94.80	52.30	77.20
GASM [13]	84.70	95.30	52.50	79.50
FPR [14]	86.58	95.42	-	-
HCTL [47]	81.80	93.80	43.60	74.30
OSNet [54]	84.90	94.80	52.90	78.70
RGA-SC [46]	88.40	96.10	57.50	80.30
CDNet [23]	86.00	95.10	54.70	78.90
Circle Loss [34]	87.40	96.10	52.10	76.90
AMD [4]	87.15	94.74	-	-
FIDI [42]	86.80	94.50	-	-
MPN-tuple [45]	88.70	95.30	60.10	82.20
ResNet-50 Base.	81.63	93.89	50.84	76.78
DFF-ResNet-50	**90.21**	**96.17**	**60.21**	**82.95**

Design choices of performing FFT. In our proposed scheme, for the given intermediate feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we perform FFT for each channel independently to obtain the latent frequency representations, as described in the Sec. 3.2 of the main paper. Here, we investigate other design choices of perform FFT. In the Table 4, we find that performing FFT across H, W, C dimensions leads to performance drop compared to *Baseline*. For the intermediate feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, its different channels correspond to the outputs of different convolution kernels, which are independent in fact. Thus, we perform FFT on each channel of $\mathbf{X}$ independently.

Ablation study on the ratio r. We follow the overall architecture design of [5] to split the given intermediate feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ into $\mathbf{X}^g \in \mathbb{R}^{rC\times H\times W}$ and $\mathbf{X}^l \in \mathbb{R}^{(1-r)C\times H\times W}$ along its channel dimension with a ratio $r \in [0,1]$. Our proposed filtering operation is only performed on $\mathbf{X}^g$. When setting $r=0$, the models degenerate to the ResNet-18/-50 baselines. Setting $r=1$ means that we perform DFF on the entire intermediate feature $\mathbf{X}$. As the experiment results in Table 5, we empirically find that the models with $r=0.5$ achieve the best performance, exploiting the complementarity of features in the frequency and original spaces.

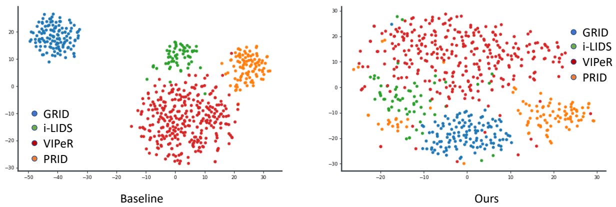

Figure 1. The t-SNE [36] visualization of ReID feature vectors learned by baseline (left) and our DFF (right) on four unseen target datasets (GRID, i-LIDS, VIPeR and GRID). Best viewed in color.

2.3. More Visualization Results

We perform t-SNE visualization for the ReID feature vectors extracted by the baseline model and the model with our proposed DFF on four unseen datasets. As shown in Fig. 1, the four unseen target domains distribute more separately for the baseline model than that of ours. This indicates the domain gaps are effectively mitigated by our proposed Deep Frequency Filtering (DFF).

2.4. Effectiveness on ImageNet-1K Classification

ImageNet-1K [32] classification widely serves as a pre-training task, providing pre-trained weights as the model initialization for various downstream task. We present the effectiveness of our conceptualized DFF on ImageNet-1K classification to showcase its potentials for more general purposes. As the results shown in Table 6, our DFF achieves 2.5%/1.6% improvements on the Top-1 classification accuracy compared to the corresponding baselines ResNet-18 and ResNet-50, respectively. Note that these improvements are achieved with the simple instantiation introduced in the Sec.3.3 of the main body. We believe more effective instantiations of DFF are worth exploring to make DFF contribute more in a wider range of fields.

2.5. Effectiveness on Supervised Person ReID

In the main body, we target domain generalization and present the effectiveness of our proposed DFF on domain generalizable person ReID. In this supplementary material, we also showcase its potential on improving supervised person ReID. Following the previous works [13, 23, 47, 52, 54] in this field, we evaluate our DFF on two most widely used datasets Market-1501 [50] and MSMT17 [39]. Note that another popular dataset DukeMTMC [53] has been taken down by its publisher. As shown in Table 7, the ResNet-50 equipped with DFF significantly outperforms the baseline model and reaches the SOTA performance on this task. This demonstrates the proposed DFF is also potentially beneficial for capturing discriminative features. We expect that it can contribute to more tasks.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *NIPS*, 31, 2018. 2

[2] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019. 2

[3] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. In *AAAI*, volume 6, 2021. 2

[4] Xiaodong Chen, Xinchen Liu, Wu Liu, Xiao-Ping Zhang, Yongdong Zhang, and Tao Mei. Explainable person re-identification with attribute-guided metric distillation. In *ICCV*, pages 11813–11822, 2021. 3

[5] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *NeurIPS*, pages 4479–4488, 2020. 2, 3

[6] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2021. 2

[7] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, pages 16145–16154, 2021. 2

[8] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6447–6458, 2019. 2

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2

[10] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[12] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 2

[13] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *ECCV*, pages 357–373, 2020. 3, 4

[14] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *ICCV*, pages 8450–8459, 2019. 3

[15] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 1

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[17] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32, 2018. 2

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 1, 2

[21] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019. 2

[22] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 2

[23] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *CVPR*, pages 6729–6738, 2021. 3, 4

[24] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 1

[25] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 1

[26] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *ECCV*, pages 456–474. Springer, 2020. 2

[27] Jiawei Liu, Zhipeng Huang, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Debiased batch normalization via gaussian process for generalizable person re-identification. In *AAAI*, 2022. 1

[28] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009. 1

[29] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, pages 0–0, 2019. 2, 3

[30] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, volume 34, pages 11749–11756, 2020. 2

[31] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5715–5725, 2017. 2

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2, 4

[33] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *ICLR*, 2018. 2

[34] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 3

[35] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 3

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4

[37] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 2

[38] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*, pages 274–282, 2018. 3

[39] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 1, 4

[40] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3415–3424, 2017. 1

[41] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[42] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Trans Multimedia*, 2021. 3

[43] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 2

[44] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019. 2

[45] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Shih-Fu Chang. Beyond triplet loss: Meta prototypical n-tuple loss for person re-identification. *IEEE Trans Multimedia*, 2021. 3

[46] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 3

[47] Cairong Zhao, Xinbi Lv, Zhang Zhang, Wangmeng Zuo, Jun Wu, and Duoqian Miao. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *IEEE Trans Multimedia*, pages 3180–3195, 2020. 3, 4

[48] 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

[49] 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*, pages 6277–6286, 2021. 2

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 4

[51] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, volume 2, pages 1–11, 2009. 1

[52] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 3, 4

[53] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 1, 4

[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 3, 4

[55] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *TIP*, 2021. 2