Supplemental Material for DomainMix: Learning Generalizable Person Re-Identification Without Human Annotations

Wenhao Wang*[1]
wangwenhao0716@gmail.com
Shengcai Liao†[2]
scliao@ieee.org
Fang Zhao[2]
fang.zhao@inceptioniai.org
Cuicui Kang[3]
cuicui.kang@mbzuai.ac.ae
Ling Shao[23]
ling.shao@ieee.org

[1] Baidu Research
Beijing, China
[2] Inception Institute of Artificial Intelligence (IIAI)
Masdar City, Abu Dhabi, UAE
[3] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Masdar City, Abu Dhabi, UAE

1 Further Details of Real-world Datasets and Implementation

1.1 Datasets and Evaluation Metrics

To evaluate the generalizability of the proposed DomainMix framework, extensive experiments are conducted on four widely used public person re-ID datasets. Among them, RandPerson (RP) [16] is selected as the synthetic dataset. Its subset contains 8,000 persons in 132,145 images. Nineteen cameras were used to capture them under eleven scenes. All images in the subset are used as training data, i.e. no gallery or query is available. The real-world datasets used are Market-1501 [19], CUHK03-NP [10, 21], and MSMT17 [17]. Market-1501 [19] includes 1,501 labeled persons in 32,668 images. Note that DukeMTMC [20] dataset is not used due to the invasion of privacy. The training set has 12,936 images of 751 identities. For testing, the query has 3,368 images and the gallery has 19,732 images. CUHK03-NP [10, 21] contains 1,467 persons from six cameras. In this dataset, 7,365 images of 767 identities are used for training. For testing, there are 1,400 queries and 5,332 gallery images. MSMT17 [17] is the most diverse and challenging re-ID dataset, consisting of 126,441 bounding boxes of 4,101 identities taken by 15 cameras. There are 32,621 images for training, while the query has 11,659 images and the gallery has 82,161 images.

Evaluation metrics are mean average precision (mAP) and cumulative matching characteristic (CMC) at rank-1. The models trained on the source domains are directly tested on the target domain without transfer learning. Single-query evaluation protocols without post-processing methods is adopted.

*Wenhao Wang finished his part of work during his internship in IIAI.
†Shengcai Liao is the corresponding author.

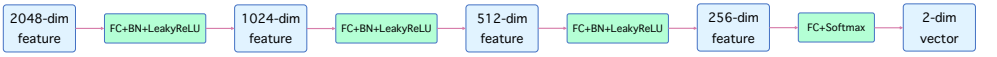

Figure 1: The design of the discriminator. Through multiple fully-connected and non-linear layers, the discriminator can classify the feature of a given image into its domain.

1.2 Implementation Details

DomainMix is trained on four Tesla-V100 GPUs. The ImageNet-pre-trained [2] ResNet-50 [8] and IBN-ResNet-50 [13] are adopted as the backbone. Adam optimizer is used to optimize the networks with a weight decay of 5×10^{-4}. All images are resized to 256×128 before being fed into the networks. Each training batch includes 64 person images of 16 actual or generated identities. The design of the discriminator is shown in Fig. 1. The λ^m and λ^s in equation 1 are both set to 1. The total number of epochs is 60. Due to the difficulty in training the mixed identity classifier, before alternating training, the model is trained for 30 epochs with the metrics in re-ID. The number of iterations in each epoch is $2,000$. The initial learning rate is set to 3.5×10^{-4}, and it is decreased to $1/10$ of its previous value on the 10th, 15th, 30th, 40th, and 50th epoch.

2 Pseudo Codes for the DomainMix Framework

$$\mathcal{L}^s(\theta) = \lambda^m \mathcal{L}_{db}(\theta) + \lambda^s \mathcal{L}^s_{id}(\theta) + \mathcal{L}^s_{tri}(\theta), \tag{1}$$

$$\mathcal{L}^s_d(\theta) = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}_{ce}\left(C_d\left(F\left(x^s_i \mid \theta\right)\right), d^s_i\right). \tag{2}$$

Algorithm 1: DomainMix framework for generalizable person re-ID

Require: Labeled synthetic dataset D_{s_1} and unlabeled real-world dataset D_{s_2};
Require: Weighting factors λ^m and λ^s for Eq. (1);

```
for n ← 1 to num_epochs do
    Generate and select training dataset D_s according to the three criteria;
    Initialize the identity classifier adaptively;
    for each mini-batch {x_i^s, y_i^s} ⊂ D_s do
        if i ≡ 0 (mod iters) then
            Update the discriminator by minimizing the objective function Eq. (2)
              with backbone fixed;
        end
        else
            Update the backbone by minimizing the objective function Eq. (1) with
              discriminator fixed;
        end
    end
end
```

Table 1: Ablation studies for each component in the DomainMix framework on the two tasks. '+I/C/Q' denotes the independence/compactness/quantity criteria is used. With or without ACI/DB denotes whether using adaptive classifier initialization/domain balance loss or not. The used backbone is IBN-ResNet-50 [13].

RP+MSMT → Market	mAP	rank-1	RP+CUHK → Market	mAP	rank-1
DBSCAN	41.5	70.0	DBSCAN	38.0	65.7
DBSCAN + I + C	42.2	70.2	DBSCAN + I + C	37.8	65.3
DBSCAN + Q	45.1	72.5	DBSCAN + Q	44.5	71.2
DBSCAN + I + C + Q	45.7	73.0	DBSCAN + I + C + Q	45.2	71.9
Without ACI	34.1	61.2	Without ACI	34.3	62.0
With ACI	45.7	73.0	With ACI	45.2	71.9
Without DB	42.3	71.0	Without DB	42.8	70.5
With DB	45.7	73.0	With DB	45.2	71.9

Table 2: Ablations studies for domain balance loss on CUHK dataset. The effectiveness of domain balance loss is proved by comparing "Without DB" with "With DB".

Without DB	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
RP+MSMT → CUHK	13.9	14.9	15.3	15.6
RP+Market → CUHK	14.3	15.6	15.7	15.9
With DB	**ResNet-50**		**IBN-ResNet-50**	
	mAP	rank-1	mAP	rank-1
RP+MSMT → CUHK	16.7	18.0	18.3	19.2
RP+Market → CUHK	16.2	17.4	17.3	17.4

3 Further Ablation Study

3.1 Experimental Results on the IBN-ResNet-50 Backbone

To further prove the efficacy of each component in the proposed DomainMix framework, we also repeat the ablation study on the IBN-ResNet-50 [13] backbone. The experimental results are shown in Table 1. Similar improvement brought by dynamic training dataset generation, adaptive classifier initialization, and domain balance loss, can be observed.

3.2 Improvement Brought by Domain Balance Loss on CUHK03-NP

To prove that the pivotal component, i.e. the domain balance loss, can improve the baseline performance significantly, experimental results on CUHK03-NP [10, 21] dataset are displayed in the Table 2. The used backbones are ResNet-50 [8] and IBN-ResNet-50 [13]. The proposed domain balance loss is still effective when the target dataset changed to CUHK03-NP [10, 21].

3.3 Importance of Introducing Unlabeled Real-world Dataset

We also verify the importance of using the unlabeled real-world dataset. Further, whether human annotations are essential for generalizable person re-ID is discussed. The baselines are denoted as "Only RP/MSMT (labeled)/CUHK (labeled)" in Table 3. On the one hand, compared to only training with synthetic data, mixing unlabeled real-world data with synthetic data brings up to 7.0% improvement in mAP. Further, if only labeled real-world data is

Table 3: The experimental results on Market dataset. "Only RP/MSMT (labeled)/CUHK (labeled)" denotes the baseline model is only trained on the RandPerson/MSMT (labeled)/CUHK (labeled) dataset.

Method	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
Only RandPerson	36.5	63.6	40.3	68.6
Only MSMT (labeled)	32.7	62.0	39.3	69.4
DomainMix (labeled)	45.2	70.5	48.7	74.6
DomainMix (unlabeled)	43.5	70.2	45.7	73.0
Method	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
Only RandPerson	36.5	63.6	40.3	68.6
Only CUHK (labeled)	25.1	50.3	36.7	64.8
DomainMix (labeled)	42.7	69.7	47.2	72.9
DomainMix (unlabeled)	39.8	67.5	45.2	71.9

Table 4: The experimental results for real-world datasets to Market or CUHK.

Real-world → Market	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
MSMT (L) + CUHK (U)	35.1	62.6	40.5	67.7
MSMT (U) + CUHK (L)	31.2	58.3	37.8	64.8
MSMT (L) + CUHK (L)	40.4	66.7	47.6	72.4
Real-world → CUHK	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
MSMT (L)+Market (U)	14.7	14.0	20.1	20.5
MSMT (U)+Market (L)	9.7	9.5	16.2	15.3
MSMT (L)+Market (L)	17.4	16.4	22.9	21.2

adopted for training, the mAP drops by up to 14.7%. On the other hand, compared to adding labeled real-world data to synthetic data, though performance decreases can be observed, using unlabeled real-world data still achieves competitive performance. Thus, the real-world data is necessary for learning domain-invariant features and improving performance. Further, the experimental results of three settings, i.e. MSMT (labeled) + CUHK/Market (labeled), MSMT (labeled) + CUHK/Market (unlabeled), and MSMT (unlabeled) + CUHK/Market (labeled) are shown in Table 4. The results show that the setting without human annotations, such as RandPerson + MSMT (unlabeled), achieves quite competitive performance. Therefore, the proposed method is quite promising in achieving competitive performance completely without human annotations.

3.4 Comparison with UDA Algorithms

To show the state of the art UDA algorithms cannot handle the proposed task well, the performance of them is in Table 5. "RP → MSMT/CUHK (SDA/MMT/SpCL)" denotes three state of the art UDA algorithms. SDA [6] uses the GAN to reduce the domain gap between RandPerson and MSMT/CUHK. However, obvious performance degradation on the unseen domain can be observed because of the bias to MSMT/CUHK and the neglection of RandPerson. SpCL [5] is a cluster-based algorithm, which uses domain specific batch normalization (DSBN) [1] and combines the source domain with the target domain for training. However, the DSBN hinders the generalizability because the BN statistics are biased to

Table 5: Comparison between the proposed DomainMix and the state of the art UDA algorithms. SDA [6], MMT [4], and SpCL [5] are three state of the art UDA algorithms. It can be seen that the UDA algorithms cannot handle the proposed task well.

RP+MSMT → Market	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
RP→MSMT (SDA)	26.6	56.3	31.3	60.9
RP→MSMT (MMT)	22.7	46.5	30.0	57.5
RP→MSMT (SpCL)	24.2	49.8	33.5	60.4
DomainMix	43.5	70.2	45.7	73.0
RP+CUHK → Market	ResNet-50		IBN-ResNet-50	
	mAP	rank-1	mAP	rank-1
RP→CUHK (SDA)	26.6	55.1	30.4	58.6
RP→CUHK (MMT)	24.6	51.2	29.9	56.3
RP→CUHK (SpCL)	9.3	24.1	18.3	39.4
DomainMix	39.8	67.5	45.2	71.9

a certain domain. Further, we find the contrastive loss in SpCL [5] is harmful for domain generalization.

4 Further Analysis of GAN-based UDA algorithms

Unsupervised Domain Adaptation (UDA) for person re-identification aims at learning a model on a labeled source domain and adapting it to an unlabeled target domain. Some methods, such as [3, 11, 17, 22], try to reduce the domain gap between two domains using a Generative Adversarial Network (GAN) [7]. Our proposed task aims at combining a labeled synthetic dataset with unlabeled real-world data to learn a ready-to-use model that can generalize well to an unseen target domain. One possible solution to learn domain-invariant feature is reducing the domain gap between synthetic and real-world data. The similar point between two tasks is how to reduce the domain gap between two different datasets. However, the GAN-based UDA algorithms cannot perform well on the proposed task because after the transfer, the model will learn domain-specific features of the real-world data and ignore the diversity of the synthetic data.

To further analyze why GAN-based UDA algorithms cannot work well on the proposed task, we visualize the images transferred from RandPerson [16] to MSMT17 [17] or CUHK03-NP [10, 21] in Fig. 2. First, the environmental lighting is diverse in RandPerson [16], but when the images are transferred to CUHK03-NP [10, 21], the environmental lighting appears to be with a single source. Second, the colors of image backgrounds are similar when RandPerson [16] is transferred to MSMT17 [17] or CUHK03-NP [10, 21]. Finally, the transferring process is imperfect, and it may induce the change of colors, the distortion of people, and so on.

The reduction of environmental lighting diversity hinders the accuracy of person matching under different cameras. Besides, similar backgrounds may prevent the model from learning domain-invariant features. Last, transferring algorithms' imperfection may cause the neural network not to fit the data well.

Therefore, though GAN-based UDA algorithms can reduce the domain gap between two domains from image-level, they cannot generalize well to an unseen target domain, and they are not suitable for the proposed task.

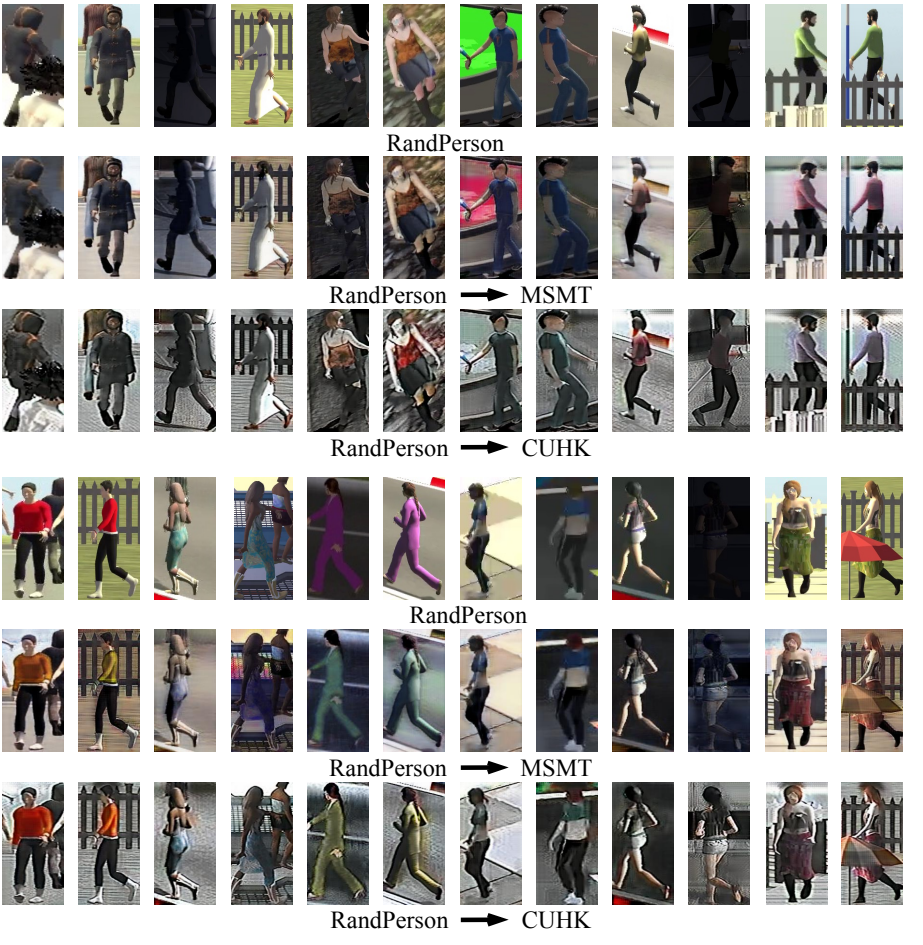

Figure 2: The visualization of the transferred images from RandPerson [16] to MSMT17 [17] or CUHK03-NP [10, 21].

5 Further Comparisons between the DomainMix and the State-of-the-arts

Some experimental results on the other backbones are shown in this section. Further, to conduct somewhat fair comparisons with the state-of-the-art algorithms, we also use authors' codes to evaluate their performances when trained on RandPerson [16]. The experimental results are shown in Table 6. It can be observed that, compared to existing methods trained on either labeled MSMT17 [17] or RandPerson [16], the proposed DomainMix generally performs better, thanks to the ability of additionally using the unlabeled MSMT17 [17]. Note that QAConv [12] achieves the best rank-1 on MSMT17 [17]. However, our method is general and can be built upon other baseline methods like QAConv [12].

Table 6: Comparison with state-of-the-arts on Market1501 [19], CUHK03-NP [10, 21], and MSMT17 [10, 21]. '§' denotes the results are from github of the original paper, '*' denotes our implementation, and '†' indicates that the results are reproduced based on the authors' codes. 'L' or 'U' denotes the used source data is labeled or unlabeled, respectively.

Method	Source data	Market1501	
		mAP	rank-1
MGN [15, 18]	MSMT (L)	25.1	48.7
ADIN [18]	MSMT (L)	22.5	50.1
ADIN-Dual [18]	MSMT (L)	30.3	59.1
OSNet-IBN† [23]	MSMT (L)	35.2	64.9
SNR [9]	MSMT (L)	41.4	70.1
QAConv† [12]	MSMT (L)	35.8	66.9
MGN† [15]	RandPerson	17.7	37.4
MGN-IBN† [15]	RandPerson	20.1	41.4
OSNet-IBN† [23]	RandPerson	39.0	67.0
QAConv§ [12]	RandPerson	34.8	65.6
QAConv-IBN§ [12]	RandPerson	36.8	68.0
Baseline*	RandPerson	36.5	63.6
Baseline-IBN*	RandPerson	40.3	68.6
DomainMix	RP+MSMT (U)	43.5	70.2
DomainMix-OSNet-IBN	RP+MSMT (U)	44.6	72.9
DomainMix-IBN	RP+MSMT (U)	**45.7**	**73.0**
Method	Source data	CUHK03-NP	
		mAP	rank-1
MGN [15, 18]	Market (L)	7.4	8.5
MuDeep [14]	Market (L)	9.1	10.3
QAConv† [12]	MSMT (L)	15.2	16.8
MGN† [15]	RandPerson	7.7	7.4
MGN-IBN† [15]	RandPerson	8.4	9.1
OSNet-IBN† [23]	RandPerson	12.9	13.6
QAConv§ [12]	RandPerson	11.0	14.3
QAConv-IBN§ [12]	RandPerson	10.8	12.9
Baseline*	RandPerson	13.0	14.6
Baseline-IBN*	RandPerson	13.6	14.3
DomainMix	RP+MSMT (U)	16.7	18.0
DomainMix-OSNet-IBN	RP+MSMT (U)	16.9	17.5
DomainMix-IBN	RP+MSMT (U)	**18.3**	**19.2**
Method	Source data	MSMT17	
		mAP	rank-1
QAConv† [12]	Market (L)	8.3	26.4
MGN† [15]	RandPerson	3.0	10.1
MGN-IBN† [15]	RandPerson	4.0	12.5
OSNet-IBN† [23]	RandPerson	12.4	34.3
QAConv§ [12]	RandPerson	10.7	34.3
QAConv-IBN§ [12]	RandPerson	12.1	**36.6**
Baseline*	RandPerson	7.9	23.0
Baseline-IBN*	RandPerson	10.9	30.6
DomainMix	RP+Market (U)	9.3	25.3
DomainMix-OSNet-IBN	RP+Market (U)	**13.6**	36.2
DomainMix-IBN	RP+Market (U)	12.1	33.1

References

[1] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.

[4] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020.

[5] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Advances in Neural Information Processing Systems*, 2020.

[6] Yixiao Ge, Feng Zhu, Rui Zhao, and Hongsheng Li. Structured domain adaptation for unsupervised person re-identification. *arXiv preprint arXiv:2003.06650*, 2020.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020.

[10] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[11] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7919–7929, 2019.

[12] 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.

[13] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018.

[14] Xuelin Qian, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Leader-based multi-scale attention deep architecture for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, pages 371–385, 2019.

[15] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[16] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3422–3430, 2020.

[17] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[18] 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 *The IEEE Winter Conference on Applications of Computer Vision*, pages 3589–3598, 2020.

[19] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[20] 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.

[21] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

[22] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, 2018.

[23] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.