Generalizing Person Re-Identification by Camera-Aware Invariance Learning and Cross-Domain Mixup
— *Supplementary Material* —

Chuanchen Luo[1,2], Chunfeng Song[1,2], and Zhaoxiang Zhang[1,2,3]

[1] University of Chinese Academy of Sciences
[2] Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS
{luochuanchen2017,chunfeng.song,zhaoxiang.zhang}@ia.ac.cn

In the supplementary material, we provide additional discussions and experimental results to validate the superiority of our method further.

1 Discussions

1.1 Sensitivity to Neighborhood Range ϵ

We adopt the relative similarity ratio ϵ rather than the number of neighbors k to define the range of neighborhood. The feasible scope shrinks from $(0, N)$ for k to $(0, 1)$ for ϵ, where N denotes the size of the dataset. In this case, a small change of ϵ corresponds to a drastic change of k. Thus, sensitivity is magnified. However, the optimal setting for ϵ is generalizable, which frees us from tuning separate hyper-parameters for intra-camera matching and inter-camera matching. In preliminary experiments, we find that the optimal setting of k is quite different between intra-camera matching ($k = 5$) and inter-camera matching ($k = 15$). While the performance reaches the optimum around $\epsilon = 0.8$ for both matching cases.

1.2 Insights of the Virtual Classifier in Cross-Domain Mixup

The proposed design of classifier in cross-domain mixup is based on the following insights: 1. According to the setup of cross-domain re-ID, the label space of the target domain is disjoint with that of the source domain. For this reason, we should differentiate target instances from source identities. Otherwise, the learned identity embedding could be corrupted irrationally. 2. Without access to identity annotations, it is infeasible to define a complete classifier for the target domain. By introducing a dynamic virtual prototype, we can bypass this issue and meanwhile ensure the separation between source and target instances. As for the constraint between target instances, we leave it to Neighborhood Invariance component to impose.

2 Quantitative Results

2.1 Ablation studies on MSMT17 and CUHK03

We have reported ablation studies on Market-1501 [6] and DukeMTMC-reID [7,4] in the paper. Here, we further validate the effectiveness of each proposed component by ablation studies on MSMT17 [5] and CUHK03 [3]. In terms of CUHK03, we use labeled annotations and follow the protocol proposed in [8]. From Tab. A, we observe a similar phenomenon on MSMT17 and CUHK03. Both camera-aware invariance learning and cross-domain mixup improve the transfer performance significantly. Interestingly, our final variant even outperforms the fully-supervised counterpart on CUHK03. This indicates that our method can take full advantage of the knowledge learned in the source domain to benefit the learning in the target domain.

Methods	MSMT17				CUHK03			
	Src.	R-1	R-5	mAP	Src.	R-1	R-5	mAP
Supervised Learning	$\mathcal{MS}$	72.2	84.4	43.7	$\mathcal{C}$	44.0	64.7	39.7
Direct Transfer	$\mathcal{M}$	10.2	17.6	3.08	$\mathcal{M}$	6.36	12.4	5.4
$\mathcal{L}_s + \mathcal{L}_{ag}$	$\mathcal{M}$	16.2	23.7	6.8	$\mathcal{M}$	10.7	19.6	11.2
$\mathcal{L}_s + \mathcal{L}_{intra}$	$\mathcal{M}$	28.3	41.4	11.3	$\mathcal{M}$	22.6	37.1	21.5
$\mathcal{L}_m + \mathcal{L}_{intra}$	$\mathcal{M}$	36.0	49.9	16.4	$\mathcal{M}$	26.2	44.0	26.0
$\mathcal{L}_s + \mathcal{L}_{intra} + \mathcal{L}_{inter}$	$\mathcal{M}$	33.3	46.5	13.2	$\mathcal{M}$	42.7	61.7	39.1
$\mathcal{L}_m + \mathcal{L}_{intra} + \mathcal{L}_{inter}$	$\mathcal{M}$	43.7	56.1	20.4	$\mathcal{M}$	55.2	73.6	51.9
Direct Transfer	$\mathcal{D}$	18.2	28.3	5.7	$\mathcal{D}$	7.5	17.4	7.2
$\mathcal{L}_s + \mathcal{L}_{ag}$	$\mathcal{D}$	21.1	30.1	8.6	$\mathcal{D}$	8.6	15.7	9.4
$\mathcal{L}_s + \mathcal{L}_{intra}$	$\mathcal{D}$	32.4	46.6	13.2	$\mathcal{D}$	20.8	35.0	20.0
$\mathcal{L}_m + \mathcal{L}_{intra}$	$\mathcal{D}$	41.6	55.7	18.9	$\mathcal{D}$	21.6	36.4	21.3
$\mathcal{L}_s + \mathcal{L}_{intra} + \mathcal{L}_{inter}$	$\mathcal{D}$	38.0	51.9	15.2	$\mathcal{D}$	39.0	57.1	35.7
$\mathcal{L}_m + \mathcal{L}_{intra} + \mathcal{L}_{inter}$	$\mathcal{D}$	51.7	64.0	24.3	$\mathcal{D}$	53.3	71.2	49.9

Table A. Ablation studies on MSMT17 and CUHK03. **Supervised Learning**: Model trained with labeled target data. **Direct Transfer**: Model trained with only labeled source data. $\boldsymbol{\mathcal{M}}$: Market-1501. $\boldsymbol{\mathcal{D}}$: DukeMTMC-reID. $\boldsymbol{\mathcal{MS}}$: MSMT17. $\boldsymbol{\mathcal{C}}$: CUHK03.

2.2 Design Choices of the Classification for Target Images

In addition to assigning each target instance a dynamic virtual class, we also explore other design choices. To be more specific, we assign a random source class or the closest source class to each target instance. As shown in Tab. B, the two designs degrade rank-1 accuracy by 5.7% and 13.5% on Market-1501, respectively. The reason behind such a degradation is straightforward. According to the open-set setup of the cross-domain re-ID, the label space of the target domain is disjoint with that of source domain. Direct assigning target instances to one of the source classes confuses the learning of discriminative features.

Methods	Duke→Market				Market→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-1	mAP
Ours	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2
Closest	74.6	85.3	89.0	50.8	75.9	86.7	89.0	60.5
Random	82.4	91.5	94.3	63.1	77.2	87.1	90.2	61.3

Table B. Ablation study of classification supervision for target images. **Closest**: we label the target images with their closest source classes. **Random**: random labels are assigned to the target images.

2.3 Variants with Source Loss

As we have stated in the paper, the proposed cross-domain mixup constraint $\mathcal{L}_m$ includes moderate supervision for the source domain. Thereby, we replace $\mathcal{L}_s$ with $\mathcal{L}_m$ while incorporating cross-domain mixup into the training. Such a practice brings two advantages: First, we omit the feature extraction of source images. In this case, the introduction of $\mathcal{L}_m$ only leads to negligible growth in computational overhead. Second, we prevent the excessive bias towards the source domain. Imposing too strong constraints on the source data is detrimental to the transfer ability. As reported in Tab. C, appending $\mathcal{L}_s$ to the variants with $\mathcal{L}_m$ degrades the performance. On the basis of the variant $\mathcal{L}_t + \mathcal{L}_m$, the involvement of $\mathcal{L}_s$ degrades the rank-1 accuracy by 2.9% and 1.2% on Market-1501 and DukeMTMC-reID, respectively. This confirms our argument about the source constraints.

Methods	Duke→Market				Market→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-1	mAP
$\mathcal{L}_{intra} + \mathcal{L}_m$	76.8	89.0	92.4	54.9	73.7	84.2	88.1	57.3
$\mathcal{L}_{intra} + \mathcal{L}_m + \mathcal{L}_s$	75.1	88.0	91.8	51.1	72.8	83.6	86.7	55.2
$\mathcal{L}_t + \mathcal{L}_m$	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2
$\mathcal{L}_t + \mathcal{L}_m + \mathcal{L}_s$	85.2	92.8	95.1	64.2	78.3	88.0	90.8	62.7

Table C. Ablation study of employing $\mathcal{L}_m$ and $\mathcal{L}_s$ simultaneously. $\mathcal{L}_t = \mathcal{L}_{intra} + \mathcal{L}_{inter}$.

2.4 Derivation of Virtual Prototype

In our final implementation, the virtual prototype is derived from *the feature stored in the memory bank* rather than *the up-to-date feature*. The former can be deemed as a temporal ensemble [2] of the model prediction, which is more stable and representative than the latter. To verify the effectiveness of our practice, we evaluate the variants that use the up-to-date feature to derive the virtual prototype. As shown in Tab. D, such variants lead to inferior performance. To be more specific, the rank-1 accuracy degrades by 6.2% and 2.3% on Market-1501 and DukeMTMC-reID, respectively.

Methods	Duke→Market				Market→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-1	mAP
Memory	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2
Feature	81.9	91.6	94.3	62.0	77.2	86.8	89.8	60.7

Table D. Ablation study of the virtual prototype. **Feature**: the variant that computes the virtual prototype using the up-to-date feature. **Memory**: our final variant that computes the virtual prototype using the stored feature in the memory bank.

2.5 Comparison with Style Transfer

We formulate the proposed cross-domain mixup (CDM) as a transition state between the source domain and the target domain, which borrows the insight from cross-domain style transfer. Different from style transfer, cross-domain mixup interpolates the samples continuously and can benefit from smooth transition across domains. Besides, it does not require extra generative models. To compare the two techniques, we translate the source images to the target style using SPGAN [1]. As shown in the first two rows of Tab. E, the model trained with stylized source data performs favorably against the variant without any cross-domain bridge components. Especially on Market-1501, the rank-1 accuracy improves by 2.0%. Nevertheless, it still lags behind the variant with cross-domain mixup. Combining the two techniques can further improve performance.

Methods	Duke→Market				Market→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-1	mAP
ours w/o CDM	81.2	91.7	94.2	59.2	76.2	87.5	90.4	59.6
ours w/o CDM + ST	83.2	92.9	94.7	60.8	76.7	86.6	89.8	60.0
ours	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2
ours + ST	88.6	94.8	96.6	58.5	79.6	89.4	92.2	65.9

Table E. The effect of style transfer. **ST**: the style of source images is transfered to the target style. **CDM**: cross-domain mixup.

2.6 Necessity of Interpolation

As expressed in Eq. (10), the virtual prototype for the target instance is exclusive from the label space of the source domain. Thus, the proposed cross-domain mixup has an effect to separate source data and target data. Someone may suspect that the improvement is mainly brought by such a separation effect, since prior research [1] has validated the effectiveness of such a dissimilarity constraint in style transfer. To verify the necessity of interpolation, we sample the mixing coefficient λ from $\{0, 1\}$ rather than Beta distribution. In this case, only the dissimilarity constraint is preserved. As reported in Tab. F, the variant without interpolation is far inferior to our method. The rank-1 accuracy degrades from 88.1% to 81.4% and 79.5% to 75.6% on Market-1501 and DukeMTMC-reID, respectively. Therefore, the interpolation is essential for the promising transfer performance.

Methods	Duke→Market				Market→Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-1	mAP
w/ interpolation	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2
w/o interpolation	81.4	91.3	94.0	58.5	75.6	86.3	89.3	59.1

Table F. Ablation study of the interpolation.

3 Qualitative Results

In addition to quantitative results, we also show the ranking lists produced by different variants. As shown in Fig. A, the quality of the ranking list improves step by step.

References

1. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
2. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
3. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
4. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshop (2016)
5. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: CVPR (2018)
6. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
7. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV (2017)
8. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)

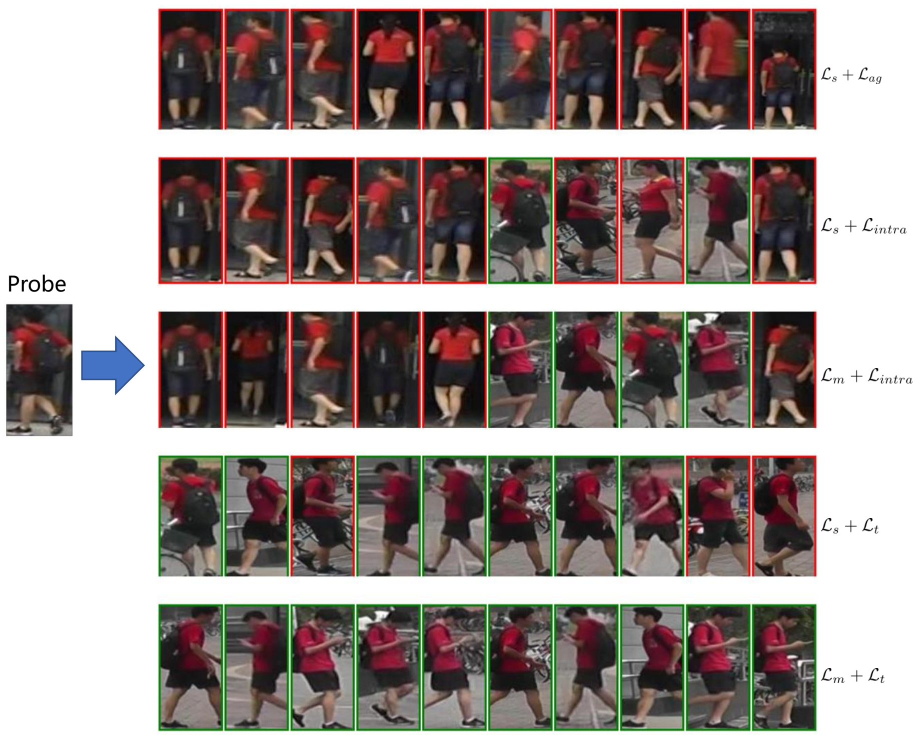

Fig. A. The visualization of the ranking lists on Market-1501. The green frame indicates positive matches, while the red frame indicates negative matches. The text at the end of each line indicates the variant that produces the ranking list.