Camera-Driven Representation Learning for Unsupervised Domain Adaptive Person Re-identification - Supplementary materials

Geon Lee[1] Sanghoon Lee[1] Dohyung Kim[1] Younghoon Shin[2] Yongsang Yoon [2] Bumsub Ham[1*]

[1]Yonsei University [2]Robotics Lab, Hyundai Motor Company

In the supplementary material, we present training cost comparisons (Sec. S1), training epochs (Sec. S2), results for IBN-ResNet50 (Sec. S3), qualitative results (Sec. S4), and detailed descriptions for setting hyperparameters (Sec. S5).

S1. Training cost comparisons

To demonstrate the efficiency of our approach, we compare the training cost with the state of the art [1, 3]. We measure GPU memory consumption during training and total training hours on MSMT17 [2]-to-Market1501 [4], and present the results in Table S1. For a fair comparison, we average the numbers over 10 executions using the official implementations provided by the authors[1], on the same machine with 4 Geforce RTX 2080Ti GPUs. Our approach offers faster training while using less GPU memory, because it does not employ multiple reID models [3], or explicitly generate more features [1], while outperforming them for all cases (Tables 1 and 2). Note that IDM even updates pseudo labels more frequently to maximize the performance.

Table S1: Quantitative comparisons of ours, UNRN [3], and IDM [1] in terms of GPU memory consumption and training hours.

Methods	GPU memory	Training hours
Ours	**27GB**	**3.2 hours**
UNRN [3]	29GB	4.3 hours
IDM [1]	54GB	6.8 hours

S2. Training epochs

Our model is trained with 4 epochs for each curriculum stage, except for the final one with 30 epochs. Namely, the total number of epochs is $4(C-1)+30$, where C is the number of cameras in a target domain. We use a half of target images on average for each stage, except for the final one with all images. The number of iterations is thus represented as $(2C-2)\lceil\frac{N}{B}\rceil + 30\lceil\frac{N}{B}\rceil$, where N and B are the number of target images and the batch size, respectively. The number of iterations for IDM [1] and UNRN [3] is $50\lceil\frac{N}{B}\rceil$, as they use 50 epochs with all target images for entire stages. If C is smaller than 11, *e.g.*, Market1501 [4] captured by 6 cameras, ours uses fewer iterations than IDM and UNRN. It requires more iterations for MSMT17 [2] with 15 cameras.

S3. Results for IBN-ResNet50

We have adopted IBN-ResNet50 as a backbone network, following IDM [1], and obtain the result on Market1501 [4]-to-MSMT17 [2]. Our model outperforms the IDM counterpart by 2.9% and 2.1% in terms of mAP and rank-1, respectively, demonstrating the effectiveness of our approach for UDA reID.

S4. Qualitative results

We provide in Fig. S1 additional visual comparisons of retrieval results among variants of our model on MSMT17 [2]-to-Market1501 [4], and Market1501-to-MSMT17. We can see that the baseline model is distracted by different persons with, *e.g.*, illumination (top-left), and similar pose (bottom-left). On the contrary, our model is able to offer person representations that are robust to various intra-class variations, and successfully retrieves person images with the same ID as a query person.

S5. Hyperparameters

We mainly adopt hyperparameter settings from recent works [1, 3] (*e.g.*, batch size, learning rate, and momen-

*Corresponding author

[1]We directly adopt official codes from https://github.com/zkcys001/UDAStrongBaseline and https://github.com/SikaStar/IDM for UNRN [3] and IDM [1], respectively.

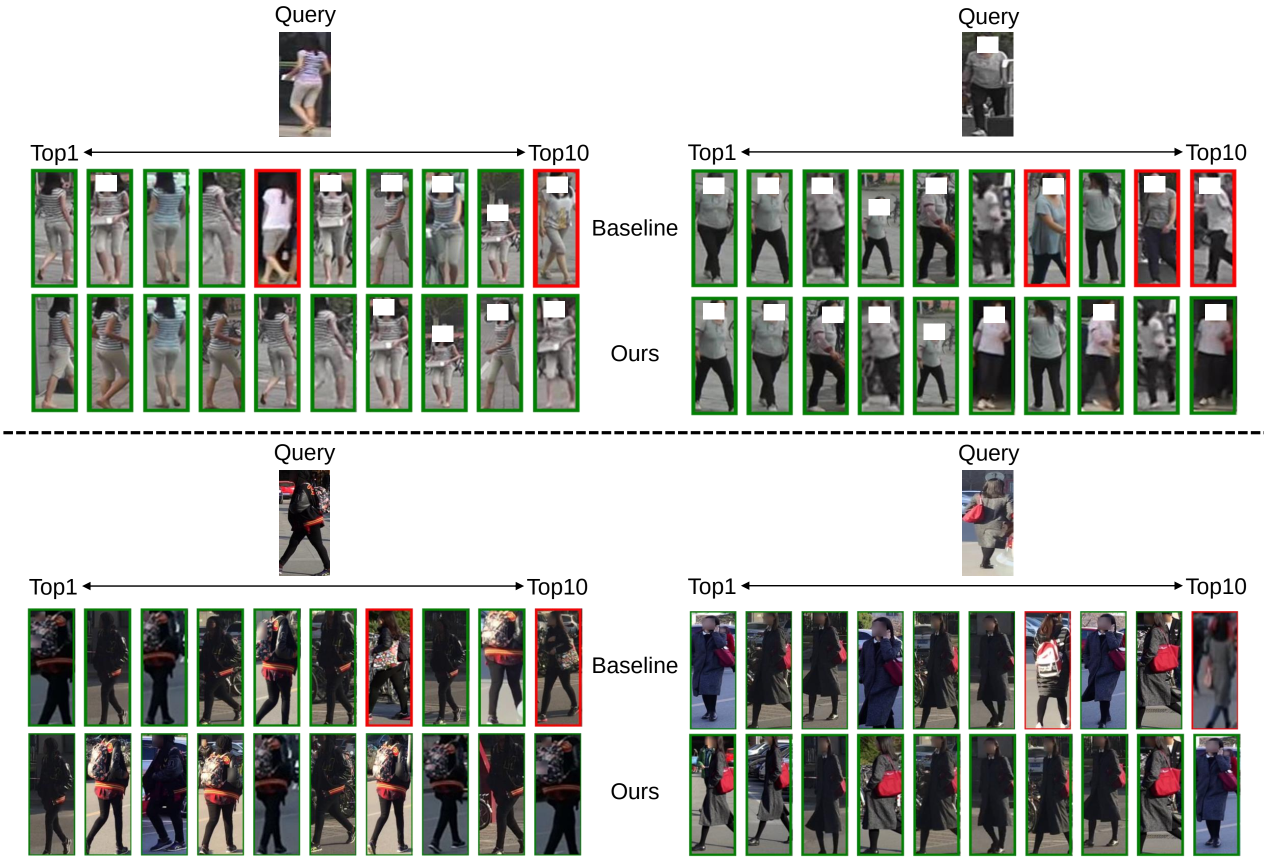

Figure S1: Visual comparisons of retrieval results on MSMT17 [2]-to-Market1501 [4] (top) and Market1501-to-MSMT17 (bottom). Results with green boxes have the same identity as the query, while those with red boxes do not. (Best viewed in color.)

Table S2: Quantitative comparisons for the hyperparameter ϵ_{init} and M_{int}. Numbers in bold indicate the best performance and underscored ones indicate the second best.

ϵ_{init}	rank-1 (%)	mAP (%)
0.3	70.2 ± 1.5	45.1 ± 1.0
0.4	$\mathbf{71.1 \pm 1.0}$	$\mathbf{46.3 \pm 0.5}$
0.5	70.1 ± 0.8	44.9 ± 1.1
M_{int}	rank-1 (%)	mAP (%)
1	71.0 ± 0.4	46.1 ± 0.4
3	$\mathbf{71.1 \pm 0.5}$	$\mathbf{46.3 \pm 0.3}$
5	69.2 ± 1.0	44.5 ± 0.2

tum value for an EMA update), except for cluster density threshold ϵ for the DBSCAN algorithm and the frequency of updating pseudo labels M_{int}. We randomly split 1041 training IDs in MSMT17 [2] into 841 and 200 IDs for training and validation, respectively, and perform cross-validation on Market1501 [4]-to-MSMT17 [2]. We set the threshold value to ϵ_{init} during the initial curriculum stage, and linearly increase the value at each stage, up to 0.6 at the final stage. This is consistent with recent works [1, 3] that set ϵ to 0.6 for training with all target images. We perform a grid search for the initial density threshold ϵ_{init} over values in $\{0.2, 0.3, 0.4, 0.5\}$. We perform a grid search for M_{int} over $\{1, 3, 5\}$. We adopt the corresponding ϵ_{init} and M_{int} values across all settings and scenarios. We provide in Table S2 the results for various ϵ_{init} and M_{int} values.

References

[1] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. IDM: An intermediate domain module for domain adaptive person re-id. In *ICCV*, 2021.

[2] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018.

[3] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha. Exploiting sample uncertainty for domain adaptive person re-identification. In *AAAI*, 2021.

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