Supplementary Material
Learning by Aligning:
Visible-Infrared Person Re-identification using Cross-Modal Correspondences

Hyunjong Park* Sanghoon Lee* Junghyup Lee Bumsub Ham†
School of Electrical and Electronic Engineering, Yonsei University
https://cvlab.yonsei.ac.kr/projects/LbA

In this supplementary material, we present additional performances under different quantitative metrics (Sec. S1) and results on applying our method to single-modality person re-identification (reID) (Sec. S2). We also present qualitative comparisons with the state-of-the-art methods for visible-infrared person reID (VI-reID) (Sec. S3), along with additional visualizations of co-attention maps (Sec. S4), and cross-modal correspondences (Sec. S5).

S1. Quantitative comparisons

We report in Table S1 the rank-10 accuracy (%) on RegDB [3] and SYSU-MM01 [6] datasets, as these metrics provide additional insights in evaluating the robustness of a VI-reID model. The results are obtained by taking an average value over 4 training and test runs.

Table S1: Average and standard deviations of rank-10 accuracy (%) on RegDB [3] and SYSU-MM01 [6] datasets. The results are obtained over 4 training and test runs.

RegDB [3]		SYSU-MM01 [6]	
V to I	*I to V*	*All-search*	*Indoor-search*
r10	r10	r10	r10
87.66 ± 0.52	87.37 ± 0.13	91.12 ± 0.44	94.13 ± 0.20

Table S2: Quantitative comparison of reID methods trained with and without our components (*i.e.*, a CMAlign module, and ID, ID consistency and dense triplet terms). We use the ID loss alone for the baseline and PCB [5]. We measure mAP (%) and rank-1 accuracy (%) on DukeMTMC-reID [4].

Method	DukeMTMC-reID [4]	
	mAP	rank-1
Baseline [2]	69.09	83.62
Baseline [2] w/ Ours	**71.26**	**85.37**
PCB [5]	69.76	84.16
PCB [5] w/ Ours	**70.47**	**84.38**

S2. RGB reID

With minor effort, our feature learning framework is applicable to other reID tasks. To verify the generalization ability, we apply our method to single-modality person reID. To this end, we train our baseline using ResNet50 [2] as a backbone and PCB [5], a representative work for person reID, with and without a CMAlign module and our losses. Table S2 shows the results on DukeMTMC-reID [4]. We can see that the CMAlign module and our losses bring performance improvements for both cases in terms of mAP and rank-1 accuracy. This demonstrates that our approach is also effective to the single-modality setting, and it boosts performance for other reID methods. Note that our approach does not require any additional parameters or computational overhead at test time.

S3. Qualitative comparisons

We provide in Figs. S1 and S2 the visual comparisons of retrieval results with the state-of-the-art methods [1, 7] on SYSU-MM01 [6] and RegDB [3], respectively. We sort the retrieval results according to cosine distances w.r.t each query image. In contrast to other methods, our model successfully retrieves person images with the same identity as the input query, even under severe intra-class variations across person images. We also show in Fig. S3 failure cases on SYSU-MM01 [6]. Current VI-reID methods including ours fail to discriminate persons with similar silhouettes. Retrieving person images of different modalities is still extremely challenging. The main reason is that IR sensors do not capture scene details, such as cloth patterns and color. In particular, our method tends to fail when a person image depicts an object of which correspondences are ambiguous, *e.g.*, *a backpack* in the last row of Fig. S3. Training with the CMAlign module in this case would be ineffective in enhancing the discriminative power of local features.

*Equal contribution. †Corresponding author.

S4. Co-attention map

We incorporate co-attention maps to focus on reconstructions between image regions that are visible in both RGB/IR person images for discriminative feature learning with the dense triplet loss $\mathcal{L}_{\mathrm{DT}}$. We show in Fig. S4 examples of co-attention maps on SYSU-MM01 [6]. Our co-attention map ignores disassociated human parts for a given pair of RGB/IR person images. For example, in Fig. S4(a), when a person's feet are not visible in the IR image, the co-attention map ignores these parts, while highlighting regions that are mutually visible.

S5. Cross-modal correspondences

For each pair of RGB/IR person images, we pick top 20 matches according to the matching probabilities on SYSU-MM01 [6], and show the results in Figs. S5 and S6. We categorize image pairs into three groups depending on the type of intra-class variations within each pair: scale (Fig. S5(a)), occlusion (Fig. S5(b)) and viewpoint (Fig. S5(c)). We can see that our model offers local features robust to the cross-modal discrepancies between RGB and IR images, even under severe intra-class variations across person images. Figure S6 compares correspondences obtained from models trained with different configurations of training losses. Our full model (Fig. S6(c)) establishes dense correspondences between regions that are semantically related, while being robust to the cross-modal discrepancies between RGB and IR images.

References

[1] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, 2020. 1, 3, 4, 5

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 2017. 1, 4

[4] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 1

[5] 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*, 2018. 1

[6] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *ICCV*, 2017. 1, 2, 3, 5

[7] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020. 1, 3, 4, 5

Figure S1: Retrieval results of (a) Hi-CMD [1], (b) DDAG [7], and (c) ours on SYSU-MM01 [6]. For each row, the person image in the left is given as a query. The retrieved images in green boxes indicate correct results and the red ones are the opposite. We also show retrieval results, obtained with an RGB image as a query, although we do not evaluate this case quantitatively following the experimental protocol of SYSU-MM01 [6]. (Best viewed in color.)

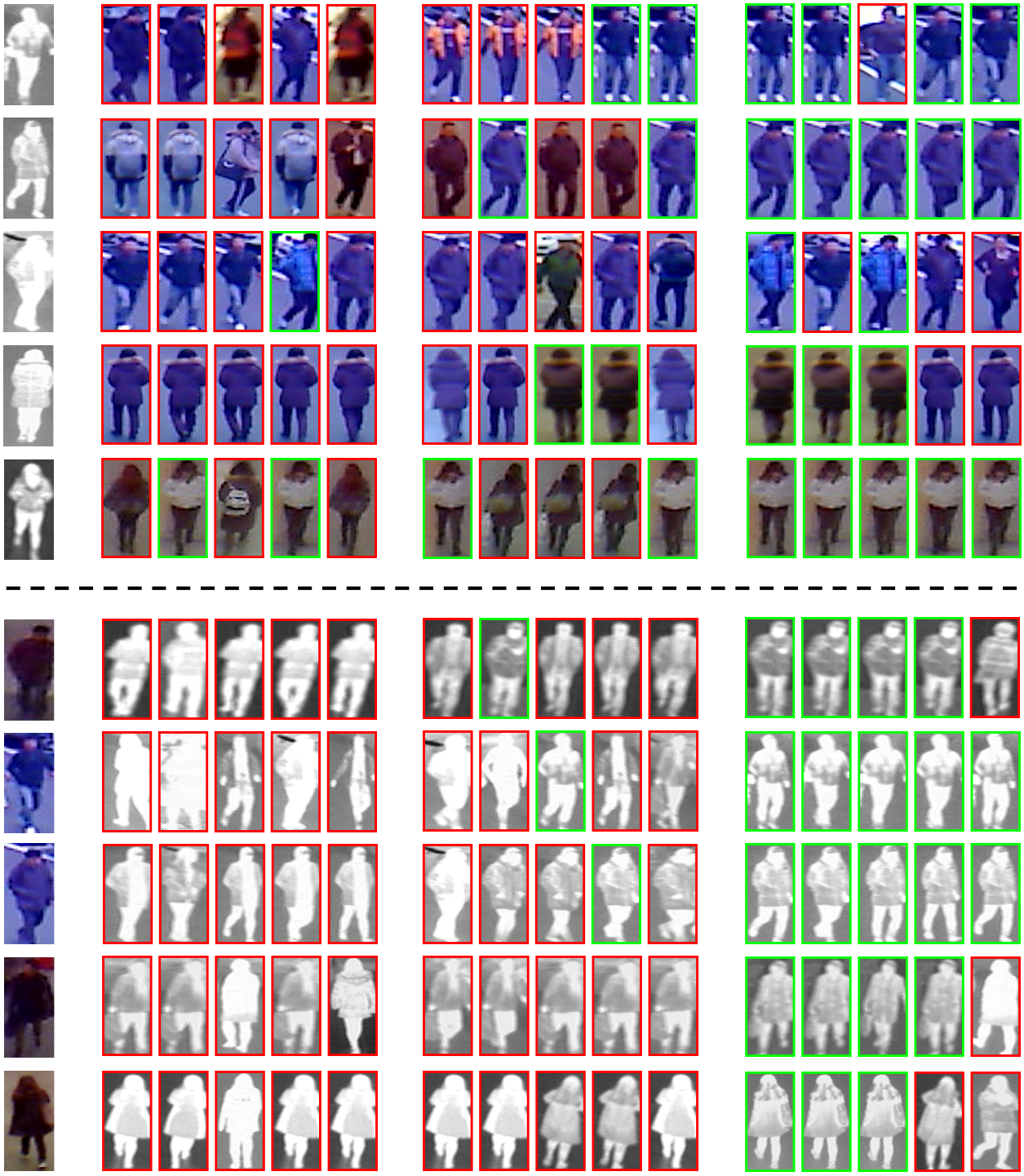


Figure S2: Retrieval results of (a) Hi-CMD [1], (b) DDAG [7], and (c) ours on RegDB [3]. For each row, the person image in the left is given as query. For each row, the person image in the left is given as a query. The retrieved images in green boxes indicate correct results and the red ones are the opposite. (Best viewed in color.)

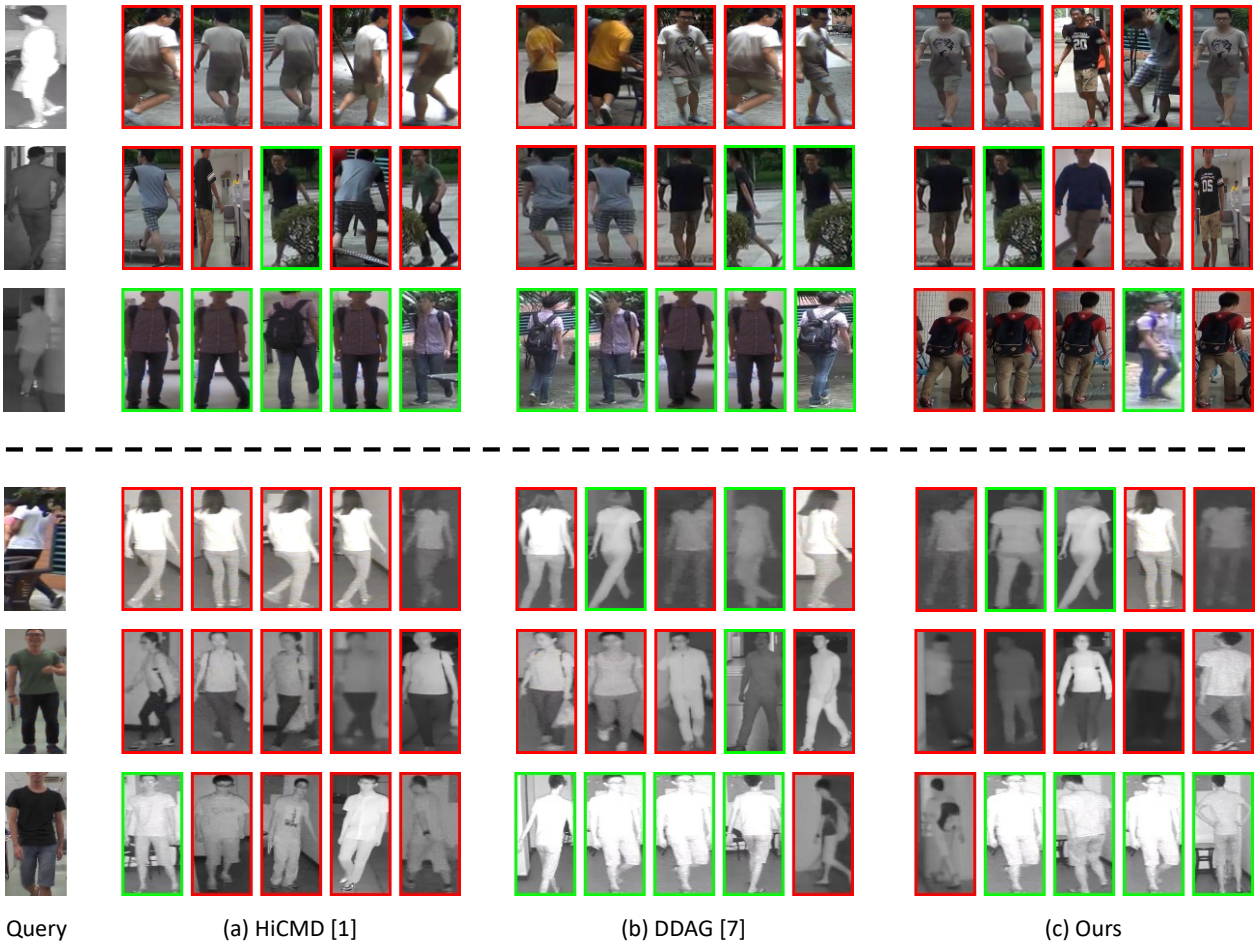


Figure S3: Failure examples of (a) Hi-CMD [1], (b) DDAG [7], and (c) ours on SYSU-MM01 [6]. Current methods including ours typically fail to retrieve, particularly when IR images do not contain enough discriminative cues. Last rows of each group illustrate cases when ours fail while other methods are successful. (Best viewed in color.)

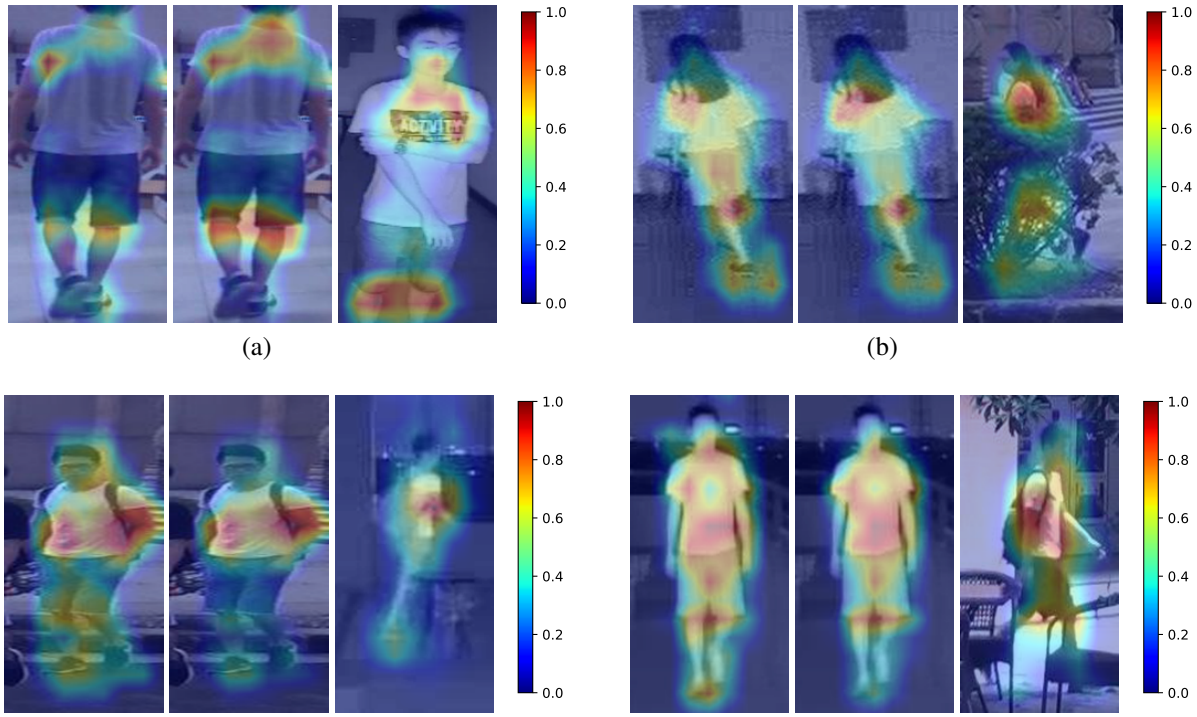
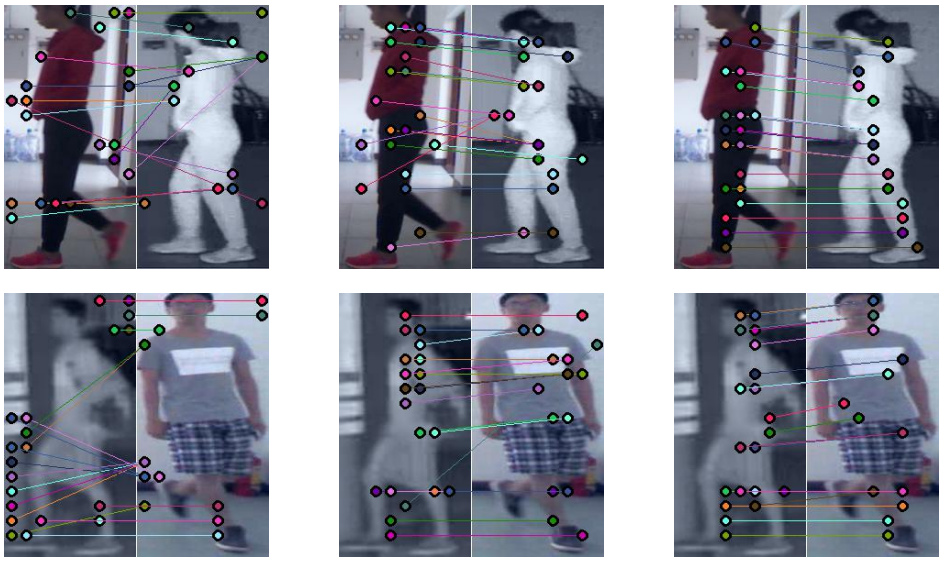


Figure S4: For each RGB/IR pair, we overlay person images and corresponding masks, and visualize the co-attention map in the middle. The co-attention map ignores human parts disassociated within a pair of RGB/IR person images, *e.g.*, (a)(d) feet, (b) abdomen and (c) leg. (Best viewed in color.)

(a) Scale variation (b) Occlusion (c) Viewpoint variation

Figure S5: Visualization of dense correspondences established by our model. The model trained using our framework is able to offer local features that are robust to the cross-modal discrepancies and large intra-class variations. (Best viewed in color.)



(a) $\mathcal{L}_{\mathrm{ID}}$ (b) $\mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{IC}}$ (c) $\mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{IC}} + \mathcal{L}_{\mathrm{DT}}$

Figure S6: Dense correspondences obtained from models trained with: (a) $\mathcal{L}_{\mathrm{ID}}$, (b) $\mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{IC}}$ and (c) $\mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{IC}} + \mathcal{L}_{\mathrm{DT}}$. The model trained with $\mathcal{L}_{\mathrm{ID}}$ alone tends to establish incorrect correspondences, *e.g.*, matches between background and person regions, as this term does not handle cross-modal discrepancies explicitly. Incorporating $\mathcal{L}_{\mathrm{IC}}$ somewhat alleviates this problem. Our full model in (c) offers more discriminative features, providing more correct matches between semantically related regions, particularly for person regions. (Best viewed in color.)