Supplementary Material for paper "Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification"

Seokeon Choi Sumin Lee Youngeun Kim Taekyung Kim Changick Kim
Korea Advanced Institute of Science and Technology, Daejeon, Korea
{seokeon, suminlee94, youngeunkim, tkkim93, changick}@kaist.ac.kr

A. Parameter Analysis

We report hyper-parameter analysis on RegDB [6] and SYSU-MM01 [10]. The total loss of our Hi-CMD is defined in Equation (10) in the main manuscript. This total loss can be reformulated as the sum of the ID-PIG and HFL loss terms $\mathcal{L} = (1-\beta)\mathcal{L}^P + \beta\mathcal{L}^H$, where $\mathcal{L}^P$ and $\mathcal{L}^H$ represent the ID-PIG and HFL loss terms respectively, and $\beta \in [0, 1]$ denotes a balancing parameter. The ID-PIG loss term is expressed as $\mathcal{L}^P = \mathcal{L}^{recon} + \lambda_{kl}\mathcal{L}^{kl} + \lambda_{adv}\mathcal{L}^{adv}$, and the HFL loss term is formulated as $\mathcal{L}^H = \lambda_{ce}\mathcal{L}^{ce} + \lambda_{trip}\mathcal{L}^{trip}$. The impacts of different β values on the REID performance are shown as in Fig. 1. We observe that our method achieves the best performance when β is set to 0.5, which indicates that the current weights for ID-PIG and HFL losses are properly balanced.

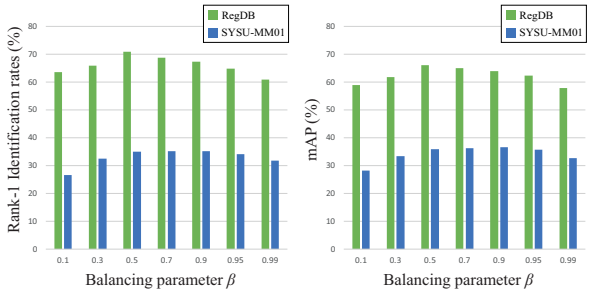

Figure 1. Performance of our Hi-CMD method for different balancing parameters β on RegDB and SYSU-MM01.

B. Results for Additional Evaluation Protocols

Different query settings on RegDB. We evaluate the performance under different query settings on the RegDB dataset, as discussed in [11, 13, 12, 3]. Unlike the original evaluation protocol where the visible image is treated as the query set, infrared images are provided as the query set. Table 1 shows that our method outperforms competing methods regardless of the modalities, which proves that the proposed Hi-CMD is robust to different query settings.

Indoor search on SYSU-MM01. We also evaluate the performance of our method under *single-shot indoor-search* mode on SYSU-MM01. A detailed description of this evaluation protocol is discussed in [10]. This protocol is less challenging than *single-shot all-search* mode since the outdoor scenes are excluded. The performance shown in Table 2 demonstrates that our Hi-CMD still outperforms the state-of-the-art methods under this evaluation protocol.

Dataset	RegDB [6] (*Infrared to visible*)			
Methods	R=1	R=10	R=20	mAP
Zero padding [10]	16.63	34.68	44.25	17.82
TONE [11]	13.86	30.08	40.05	16.98
TONE+HCML [11]	21.70	45.02	55.58	22.24
BDTR [13]	32.72	57.96	68.86	31.10
eBDTR [12]	34.21	58.74	68.64	32.49
HSME [3]	40.67	65.35	75.27	37.50
D-HSME [3]	50.15	72.40	81.07	46.16
Ours (Hi-CMD)	**68.21**	**84.85**	**89.76**	**62.41**

Table 1. Comparison with the state-of-the-arts under different query settings on RegDB. The infrared image is given as the query set as opposed to the original evaluation protocol in the main manuscript. Re-identification rates (%) at rank R and mAP (%).

Dataset	SYSU-MM01 [10] (*Indoor-search*)			
Methods	R=1	R=10	R=20	mAP
SVDNet [7]	20.24	64.32	83.62	28.74
PCB [8]	22.63	65.24	83.92	30.46
Zero padding [10]	20.58	68.38	85.79	26.92
TONE [11]	20.82	68.86	84.46	26.38
cmGAN [1]	31.63	77.23	89.18	42.19
eBDTR (alex) [12]	27.62	74.95	88.12	38.40
eBDTR (resnet) [12]	32.46	77.42	89.62	42.46
Ours (Hi-CMD)	**37.11**	**81.63**	**91.36**	**47.11**

Table 2. Comparison with the state-of-the-art methods under the *single-shot indoor-search* mode on the SYSU-MM01 dataset.

C. Network Architectures

The proposed Hi-CMD method consists of the two prototype encoders E_1^p, E_2^p, the two attribute encoders E_1^a, E_2^a, the two discriminators D_1, D_2, the single decoder G, and the single feature embedding network H. We use ResNet-50 [4] pretrained on ImageNet [2] as the attribute encoder E^a and the feature embedding network H. The attribute encoder has the same structure up to the average pooling layer of ResNet-50, and the feature embedding network borrows the *Conv3* and *Conv4* layers of ResNet-50.

Table 3 shows the remaining three network architectures as follows: (1) The prototype encoder E^p contains several convolutional layers and residual blocks [4] with Instance Normalization (IN) [9]. We also add the Atrous Spatial Pyramid Pooling (ASPP) to exploit multi-scale features. (2) The residual decoder G consists of several convolution layers, upsampling layers, and residual blocks. The Adaptive Instance Normalization (AdaIN) [5] layers in the residual layers help the generator synthesize various attributes in a person image. (3) The discriminator D contains several convolution layers and residual blocks. We use PatchGAN [14] with the three different scales of an input image as 256×128, 128×64, and 64×32.

D. Retrieved Examples

We analyze the top-5 retrieval results of 24 query examples on RegDB and SYSU-MM01 datasets, as illustrated in Fig. 2, 3. For each dataset, two different experiments are performed. In the case of RegDB, the *visible-to-infrared* retrieval is the original setting while the *infrared-to-visible* retrieval is also applied as introduced in Table 1. In the case of SYSU-MM01, the *infrared-to-visible* retrieval is the original setting. Note that we use all the images in the gallery set and query set on the SYSU-MM01 dataset, so the retrieval result is slightly different from the result by the SYSU evaluation protocol, such as *single-shot all-search* and *single-shot indoor-search*.

The images in the first column are the query images, and the retrieved images are sorted from left to right according to the ascending order of the feature distance. The query and retrieved images have different types of poses and illuminations since the cross-modality and intra-modality characteristics are entangled intricately in the image, which represents that this cross-modality retrieval task is very challenging. In this harsh situation, the proposed method achieves successful cross-modality matching results by finding the common ID-discriminative factors, such as the clothing type, pattern, and human body size. Although there are even a few failure cases, we observe that the ID-discriminative characteristics are similar enough to be mistaken by the human eye. In conclusion, our Hi-CMD method is robust to ID-excluded factors and well finds hidden ID-discriminative clues.

References

[1] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 677–683, 2018. 1

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

[3] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8385–8392, 2019. 1

[4] 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 (CVPR)*, pages 770–778, 2016. 2

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 1501–1510, 2017. 2

[6] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 1

[7] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3800–3808, 2017. 1

[8] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 1

[9] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6924–6932, 2017. 2

[10] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5380–5389, 2017. 1

[11] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 1

[12] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 2019. 1

[13] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1092–1099, 2018. 1

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 2223–2232, 2017. 2

Module Name	Layer Name	Input Shape $\rightarrow$ Output Shape	Parameters	Layer Description
Prototype Encoder E_1^p, E_2^p	Conv1	$(H, W, 3) \rightarrow (\frac{H}{2}, \frac{W}{2}, 32)$	$[3 \times 3,\ 32]$	S-2, IN, LReLU
	Conv2	$(\frac{H}{2}, \frac{W}{2}, 32) \rightarrow (\frac{H}{2}, \frac{W}{2}, 64)$	$[3 \times 3,\ 64]$	S-1, IN, LReLU
	Conv3	$(\frac{H}{2}, \frac{W}{2}, 64) \rightarrow (\frac{H}{2}, \frac{W}{2}, 64)$	$[3 \times 3,\ 64]$	S-1, IN, LReLU
	Conv4	$(\frac{H}{2}, \frac{W}{2}, 64) \rightarrow (\frac{H}{4}, \frac{W}{4}, 128)$	$[3 \times 3,\ 128]$	S-2, IN, LReLU
	ResBlocks	$(\frac{H}{4}, \frac{W}{4}, 128) \rightarrow (\frac{H}{4}, \frac{W}{4}, 128)$	$\begin{bmatrix} 3 \times 3,\ 128 \\ 3 \times 3,\ 128 \end{bmatrix} \times 4$	S-1, IN, LReLU
	ASPP	$(\frac{H}{4}, \frac{W}{4}, 128) \rightarrow (\frac{H}{4}, \frac{W}{4}, 256)$	$[1 \times 1,\ 64]$ $\begin{bmatrix} 1 \times 1,\ 64 \\ 3 \times 3,\ 64 \end{bmatrix} \times 3$	S-1, IN, LReLU
	Conv5	$(\frac{H}{4}, \frac{W}{4}, 256) \rightarrow (\frac{H}{4}, \frac{W}{4}, 256)$	$[1 \times 1,\ 256]$	S-1, IN
Residual Decoder G	ResBlocks	$(\frac{H}{4}, \frac{W}{4}, 256) \rightarrow (\frac{H}{4}, \frac{W}{4}, 256)$	$\begin{bmatrix} 3 \times 3,\ 256 \\ 3 \times 3,\ 256 \end{bmatrix} \times 4$	S-1, AdaIN, LReLU
	Upsampling	$(\frac{H}{4}, \frac{W}{4}, 256) \rightarrow (\frac{H}{2}, \frac{W}{2}, 256)$	-	-
	Conv1	$(\frac{H}{2}, \frac{W}{2}, 256) \rightarrow (\frac{H}{2}, \frac{W}{2}, 128)$	$[5 \times 5,\ 128]$	S-1, LN, LReLU
	Upsampling	$(\frac{H}{2}, \frac{W}{2}, 128) \rightarrow (H, W, 128)$	-	-
	Conv2	$(H, W, 128) \rightarrow (H, W, 64)$	$[5 \times 5,\ 64]$	S-1, LN, LReLU
	Conv3	$(H, W, 64) \rightarrow (H, W, 64)$	$[3 \times 3,\ 64]$	LReLU
	Conv4	$(H, W, 64) \rightarrow (H, W, 64)$	$[3 \times 3,\ 64]$	LReLU
	Conv5	$(H, W, 64) \rightarrow (H, W, 3)$	$[1 \times 1,\ 3]$	-
Discriminator D_1, D_2	Conv1	$(H, W, 3) \rightarrow (H, W, 32)$	$[1 \times 1,\ 32]$	S-1, LReLU
	Conv2	$(H, W, 32) \rightarrow (H, W, 32)$	$[3 \times 3,\ 32]$	S-1, LReLU
	Conv3	$(H, W, 32) \rightarrow (\frac{H}{2}, \frac{W}{2}, 32)$	$[3 \times 3,\ 32]$	S-2, LReLU
	Conv4	$(\frac{H}{2}, \frac{W}{2}, 32) \rightarrow (\frac{H}{2}, \frac{W}{2}, 32)$	$[3 \times 3,\ 32]$	S-1, LReLU
	Conv5	$(\frac{H}{2}, \frac{W}{2}, 32) \rightarrow (\frac{H}{4}, \frac{W}{4}, 64)$	$[3 \times 3,\ 64]$	S-2, LReLU
	ResBlocks	$(\frac{H}{4}, \frac{W}{4}, 64) \rightarrow (\frac{H}{4}, \frac{W}{4}, 64)$	$\begin{bmatrix} 3 \times 3,\ 64 \\ 3 \times 3,\ 64 \end{bmatrix} \times 4$	S-1, LReLU
	Conv6	$(\frac{H}{4}, \frac{W}{4}, 64) \rightarrow (\frac{H}{4}, \frac{W}{4}, 1)$	$[1 \times 1,\ 1]$	S-1

Table 3. Network architectures within the proposed overall framework. We express the characteristics of each layer as S : Stride size, IN: Instance Normalization, $AdaIN$: Adaptive Instance Normalization, and LN: Layer Normalization. H and W denote the height and width of the input image, respectively. Both visible and infrared images are resized to $256 \times 128 \times 3$.

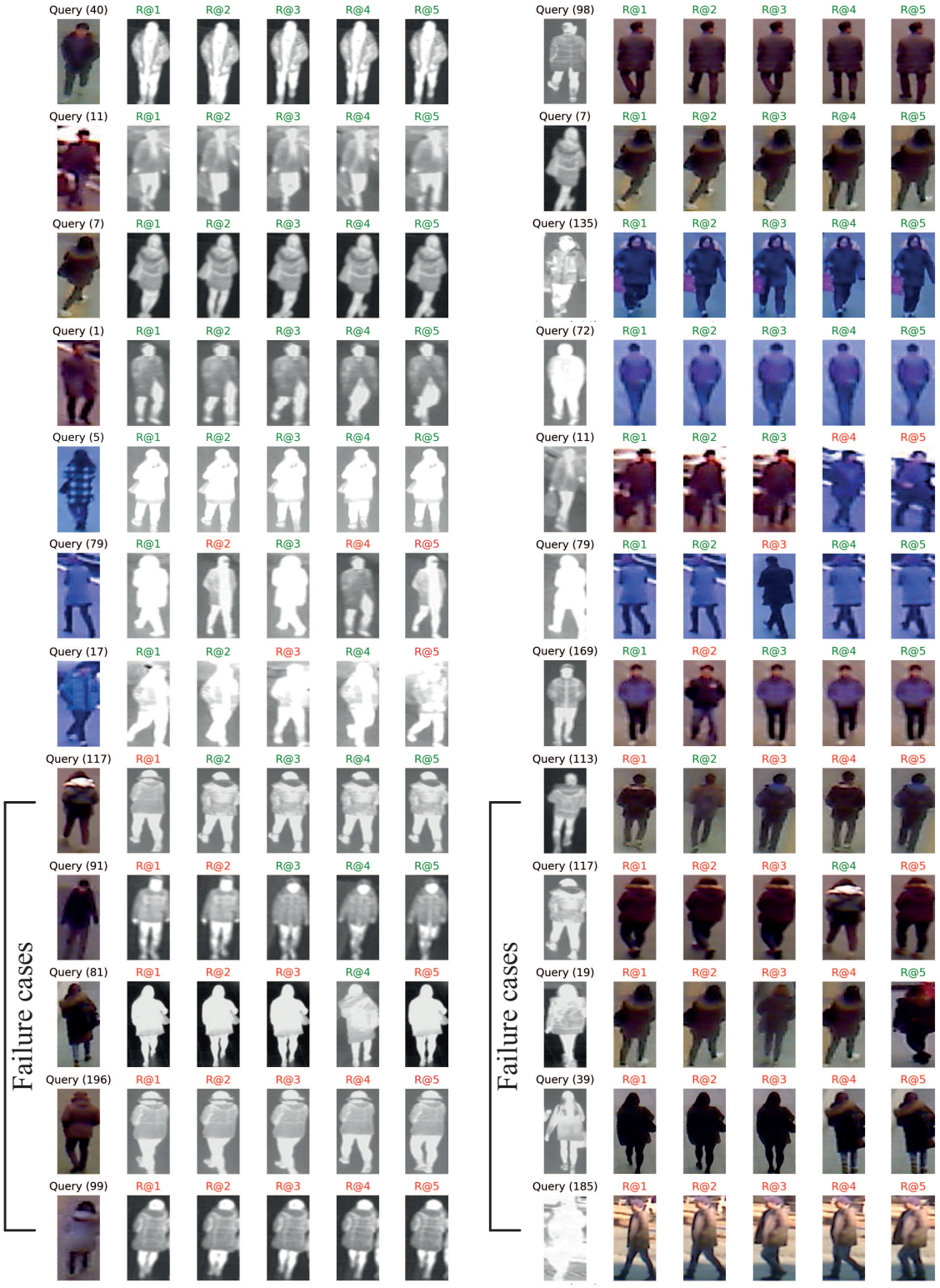

Figure 2. The top-5 retrieval results of our Hi-CMD on the RegDB dataset. The correct matchings and wrong matchings are indicated by **green** and **red** color, respectively. (a) Visible images are given as the query set, which is the original evaluation protocol for RegDB. (b) Infrared images are given as the query set, which is introduced in Table 1. Best viewed in color.

Figure 3. The top-5 retrieval results of our Hi-CMD on the SYSU-MM01 dataset. The SYSU-MM01 dataset is relatively challenging compared to the RegDB dataset. The correct matchings and wrong matchings are indicated by green and red color, respectively. (a) Infrared images are given as the query set, which is the original evaluation protocol for SYSU-MM01. (b) Visible images are given as the query set. Best viewed in color.