BIT: Matching-based Bi-directional Interaction Transformation Network for Visible-Infrared Person Re-Identification

Supplementary Material

1. Dataset

We evaluate the BIT on three widely-used VI-ReID benchmarks: SYSU-MM01, LLCM and RegDB.

SYSU-MM01 [8] is the largest and most widely used benchmark for VI-ReID. It contains 287,628 RGB images and 15,792 IR images from 491 identities. The dataset is split into a training set with 395 identities (22,258 RGB and 11,909 IR images) and a testing set with the remaining 96 identities. The testing set includes 3,803 infrared query images and 301 visible gallery images. Evaluation is conducted under two standard protocols: *all-search mode*, which considers both indoor and outdoor gallery images, and *indoor-search mode*, which includes only indoor gallery images. All evaluations are performed under both *single-shot* and *multi-shot* setting, and results are averaged over 10 random trials.

RegDB [6] comprises 8,240 images of 412 identities, captured by paired visible and infrared cameras. Each identity has 10 visible and 10 infrared images. The dataset is divided equally for training and testing. Two cross-modality matching settings are evaluated: visible-to-infrared (V2I) and infrared-to-visible (I2V), where queries and galleries belong to different modalities.

LLCM [10] is a large scale dataset designed for VI-ReID. It is split into training and testing sets at a 2:1 ratio. Similar to RegDB, LLCM supports both V2I and I2V matching scenarios, providing a more realistic and challenging environment for evaluating cross-modality robustness under adverse illumination.

Table 1. Dataset statistics of RegDB, SYSU-MM01 and LLCM.

Datasets	IDs	Images	VIS / IR cam
RegDB	412	8,240	1 / 1
SYSU-MM01	491	287,628	4 / 2
LLCM	1,064	46,767	9 / 9

2. Implementation Details

All experiments are conducted on a single NVIDIA L20 GPU using PyTorch. We follow the same backbone configuration as PMT [5], employing a ViT-B/16 [2] model pre-trained on ImageNet [1]. The number of stacked BCI blocks T is set to 3. To balance computational efficiency and representation quality, the overlap stride is set to 12. All person images are resized to 256×128 before being fed into the network. For data augmentation, we apply random horizontal flipping and random erasing to both modalities. Additionally, color jitter and Gaussian blur are applied exclusively to infrared images to enhance modality robustness.

Stage 1: Backbone Training. In the first training stage, we train the backbone to obtain strong modality-invariant representations. Each mini-batch contains 64 images sampled from 8 identities, with 4 visible and 4 infrared images per identity. The model is optimized using AdamW with a cosine annealing learning rate scheduler. The initial learning rate is set to 3×10^{-4}, and the weight decay is set to 1×10^{-4}. This stage is trained for 40 epochs on all datasets.

Stage 2: BIT Training. In the second stage, the backbone network is frozen and only the proposed BIT modules are optimized. Each mini-batch is constructed from 4 randomly sampled identities, each contributing 4 visible and 4 infrared images, resulting in 32 images per batch. Unlike traditional methods that operate on individual images, BIT takes visible-infrared feature pairs as input. This design yields 256 cross-modal pairs per batch, effectively transforming a batch of B images into $B^2/4$ training samples. We continue to use AdamW with the same learning rate and scheduler settings. This stage is trained for 60 epochs.

3. Controlled Data-Imbalance Experiments

To further validate BIT's effectiveness toward data imbalance, we simulated controlled imbalance experiments by reducing training samples per identity on the SYSU-MM01 dataset, whose original visible/infrared ratio in the training set is 1.9. We consider four settings: randomly removing 10% and 20% of infrared images (which exacerbate the imbalance), and removing 10% and 20% of visible images (which alleviate it). Note that reducing training data inevitably causes some performance drop for all methods. The results are shown in Tab. 2.

Across all four settings, BIT consistently achieves the best performance on all metrics compared with current SOTA methods. We also observe that existing rigid feature learning approaches suffer significantly larger degradation when infrared samples are reduced than when an equal proportion of visible samples is removed (*e.g.*, 10% inf reduce setting v.s. 10% vis reduce setting), even though the absolute number of removed visible images is higher. This indicates that rigid feature learning methods are inherently sensitive to data imbalance.

In contrast, BIT remains markedly stable under all imbalance settings. Moreover, the performance gain of BIT over the baseline PMT becomes even larger in the more severely imbalanced infrared reduction settings, further

Table 2. Comparison with SOTA methods on the SYSU-MM01 dataset under controlled Data-Imbalance settings. The best are highlighted in bold. † denotes our reproduced results.

Settings	Model	All-Search		Indoor-Search	
		Rank-1	mAP	Rank-1	mAP
Original	MCLNet [4]	65.40	61.98	72.56	76.58
	MPANet [9]	70.58	68.24	76.74	80.95
	SAAI [3]	75.90	77.03	83.20	88.01
	DEEN [10]	74.70	71.80	80.30	83.30
	HOS-Net [7]	75.60	74.20	84.20	86.70
	PMT (baseline)† [5]	69.23	66.02	73.24	77.91
	BIT (ours)	**80.53** (+11.3)	**79.76** (+13.74)	**87.42**(+14.18)	**89.25** (+11.28)
10% Inf Reduce	MCLNet	60.31	57.46	67.31	73.42
	MPANet	66.34	64.21	72.67	76.47
	SAAI	71.43	72.97	78.43	83.42
	DEEN	70.21	65.32	76.74	78.95
	HOS-Net	70.45	69.98	80.26	81.78
	PMT(baseline)	65.20	62.43	70.45	72.54
	BIT (ours)	**79.67**(+14.47)	**79.39**(+16.96)	**86.17**(+15.72)	**87.96**(+15.42)
20% Inf Reduce	MCLNet	56.42	53.86	63.52	69.37
	MPANet	62.67	60.93	69.42	72.81
	SAAI	66.57	68.91	73.78	77.92
	DEEN	65.76	61.34	72.89	72.79
	HOS-Net	64.70	64.71	75.54	76.47
	PMT(baseline)	60.40	58.37	66.87	69.65
	BIT (ours)	**77.80**(+17.40)	**78.65**(+20.28)	**85.45**(+18.58)	**86.47**(+16.82)
10% Vis Reduce	MCLNet	62.79	59.86	70.47	74.89
	MPANet	68.92	67.31	73.79	78.06
	SAAI	74.32	74.97	80.43	85.76
	DEEN	73.07	68.92	77.87	80.25
	HOS-Net	72.76	72.10	82.36	82.95
	PMT(baseline)	67.16	63.92	71.64	73.17
	BIT (ours)	**79.51**(+12.35)	**79.42**(+15.5)	**86.65**(+15.01)	**88.21**(+15.04)
20% Vis Reduce	MCLNet	59.82	57.93	66.82	70.60
	MPANet	65.78	64.42	69.75	75.32
	SAAI	70.24	71.66	77.81	81.62
	DEEN	69.65	66.42	75.31	76.62
	HOS-Net	67.51	68.92	78.62	77.41
	PMT(baseline)	63.84	60.31	68.02	70.57
	BIT (ours)	**77.51**(+13.67)	**78.32**(+18.01)	**85.71**(+17.69)	**86.43**(+15.86)

demonstrating BIT's strong robustness to uneven modality distributions and its clear advantage over prior SOTA methods. In addition, we also observe that BIT yields a larger performance gain over the baseline even when the visible samples are reduced. This indicates that BIT is not only robust to data imbalance, but also resilient to reductions in the overall training data scale.

4. More Visualization Analysis

4.1. More Retrieval Results

To qualitatively evaluate the effectiveness of our proposed BIT framework, we present additional retrieval results on the SYSU-MM01 dataset in Fig. 1, comparing BIT with the baseline method. For each query image, we visualize the top-10 retrieved gallery images, where green boxes denote correctly matched identities and red boxes indicate incorrect matches. As shown in the figure, BIT consistently yields more accurate retrievals, with a greater number of correct matches appearing in the top ranks compared to the baseline. This improvement highlights BIT's enhanced ability to bridge the modality gap and capture identity-discriminative features, even under challenging cross-modality conditions. In particular, BIT demonstrates superior robustness in complex scenes where the baseline tends to misidentify hard negatives or be distracted by modality-specific cues.

Notably, in the third row, we observe that the baseline method retrieves gallery images that are visually similar to the correct match but belong to incorrect identities. This behavior indicates that the baseline relies on a rigid fea-

(a) Baseline (b) BIT

Figure 1. Illustration of Rank-10 retrieval results on SYSU- MM01. The left is baseline, the right is our BIT.

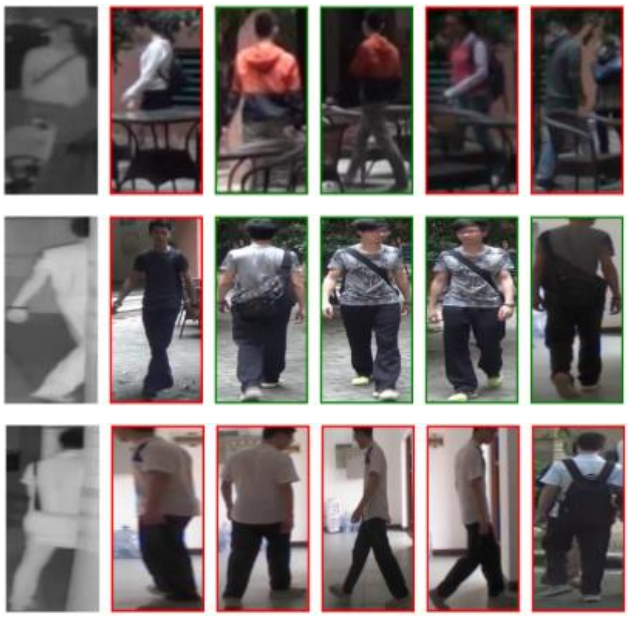

Figure 2. Illustration of some unsuccessful retrieval results.

ture extraction paradigm, which struggles to model the complex and nonlinear relationships inherent in cross-modality scenarios, as discussed in the main manuscript. In contrast, BIT successfully avoids such false positives by directly computing the similarity between query and gallery instances in a more adaptive and identity-aware manner, retrieving the correct matches despite appearance-level distractions.

These qualitative results further corroborate the quantitative improvements reported in the main paper, demonstrating that BIT not only improves performance metrics but also enhances the visual reliability and interpretability of retrieval outcomes in practical applications.

4.2. Some Unsuccessful Retrieval Results

Fig. 2 presents several representative failure cases of our BIT framework. As shown, most of these unsuccessful retrievals occur in scenarios involving severe occlusion, where critical identity cues are partially blocked. These cases reveal a compounded challenge: not only must the model bridge the modality gap between visible and infrared images, but it must also cope with significant spatial and semantic information loss caused by occlusion.

The current design of BIT primarily focuses on query aware matching. However, these results suggest that under heavy occlusion, even matching-based method is insufficient when essential visual cues are missing or distorted.

Addressing such challenging cases remains an open problem. Future work could explore integrating occlusion-aware modeling mechanisms, such as part-based feature reasoning, visibility prediction, or generative completion, which can enhance the model's robustness under partial observation. In particular, designing solutions that can

jointly reason about modality heterogeneity and occlusion patterns may offer a promising direction for advancing visible-infrared person re-identification in real-world environments.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[3] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *ICCV*, pages 11270–11279, 2023. 2

[4] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *ICCV*, pages 16403–16412, 2021. 2

[5] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *AAAI*, pages 1835–1843, 2023. 1, 2

[6] 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*, 17(3):605, 2017. 1

[7] Liuxiang Qiu, Si Chen, Yan Yan, Jing-Hao Xue, Da-Han Wang, and Shunzhi Zhu. High-order structure based middle-feature learning for visible-infrared person re-identification. In *AAAI*, pages 4596–4604, 2024. 2

[8] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 1

[9] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *CVPR*, pages 4330–4339, 2021. 2

[10] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, pages 2153–2162, 2023. 1, 2