Cross-Category Subjectivity Generalization for Style-Adaptive Sketch Re-ID

Supplementary Material

1. More Ablation Study Results

We also evaluate the impact of non-pedestrian data inclusion on the MaSk1k dataset. To ensure a zero-shot evaluation, we train the model exclusively on pedestrian samples in style A from MaSk1k and test it on the remaining unseen styles. As shown in Table 1, the results reinforce findings in the main manuscript: naively integrating non-person data provides minimal benefits to the sketch re-ID task, highlighting the effectiveness of the proposed AIP method in leveraging non-person data for person sketch re-ID.

train config	mAP	R1	mINP
AIP	**18.6**	**14.7**	**12.8**
MaSk1k only (CLIP-Base)	14.0	11.9	9.2
Sketchy $\rightarrow$ MaSk1k	15.3	12.8	10.1
Sketchy + MaSk1k	13.7	12.4	8.9

Table 1. Retrieval results on MaSk1k dataset, considering different training configurations. Pedestrian sketches in style A are used for training and the rest are for testing.

2. Fully Supervised Results

Fully Supervised (FS) Configuration Results. To comprehensively evaluate the proposed AIP method, we also compare it with the FS configuration. Table 2 presents re-ID results on the PKU dataset with FS configuration (all methods are trained and tested on the PKU dataset). As observed, our method still outperforms the latest DALNet method, which requires extra pre-learned edge detector U^2-Net [8] for auxiliary generation and alignment.

In Table 3, the comparison between the AIP method and existing works on the MaSk1k dataset with FS configuration is presented. As observed, our method also outperforms all existing works, demonstrating the proposed AIP method's superiority in terms of familiar styles.

3. More Retrieval Results Visualization

In Figure 1, we present additional top-5 retrieval visualizations. The middle section further illustrates the performance of AIP without the style prompt generator ϕ.

Consistent with our main findings, AIP not only improves overall retrieval accuracy but also demonstrates robustness to sketch style variations. Moreover, while AIP without ϕ still outperforms the CLIP-base model, it remains susceptible to style-induced discrepancies, as evidenced by the rank gaps between sketches of the same ID but different styles. In contrast, incorporating ϕ significantly mitigates this effect, reinforcing its effectiveness in addressing subjective style variations.

method	mAP	R1	R5	R10
Triplet SN [15]	-	9.0	26.8	42.2
GN Siamese [10]	-	28.9	54.0	62.4
CD-AFL [7]	-	34.0	56.3	72.5
LMDI [4]	-	49.0	70.4	80.2
CDAC [20]	-	60.8	80.6	88.8
SketchTrans [1]	-	84.6	94.8	98.2
CCSC [19]	83.7	86.0	98.0	100.0
DALNet [6]	86.2	90.0	98.6	100.0
CLIP-Base	88.2	90.0	96.0	98.0
AIP (ours)	**92.4**	**98.0**	98.0	100.0

Table 2. Sketch-based re-ID results on PKU-Sketch dataset with Fully Supervised configuration.

method	mAP	R1	R5	R10
DDAG [13]	12.13	11.22	25.40	35.02
CM-NAS [3]	0.82	0.70	2.00	3.90
CAJ [14]	2.38	1.48	3.97	7.34
MMN [17]	10.41	9.32	21.98	29.58
DART [12]	7.77	6.58	16.75	23.42
DCLNet [11]	13.45	12.24	29.20	39.58
DSCNet [18]	14.73	13.84	30.55	40.34
DEEN [16]	12.62	12.11	25.44	30.94
AttrAlign [5]	19.61	18.10	38.95	50.75
CLIP-Base	31.51	29.21	55.54	67.14
AIP (ours)	**38.04**	**36.46**	**61.81**	**74.30**

Table 3. Retrieval results on MaSk1k with Fully-Supervised configuration and single-query setting.

4. Extend to Traditional RGB re-ID

One interesting extension of our AIP method could be: could the traditional RGB re-ID task also benefit from inclusion of extra non-pedestrian RGB data?

We first test **prompt-tuning CLIP for traditional RGB re-ID** on Market-1501. As shown in Table 4, **Prompt-Tuning (PT) alone achieves strong performance**, with additional **LayerNorm (LN)** tuning as in [9], the result is further improved.

Intuitively, **extending our method to RGB re-ID is**

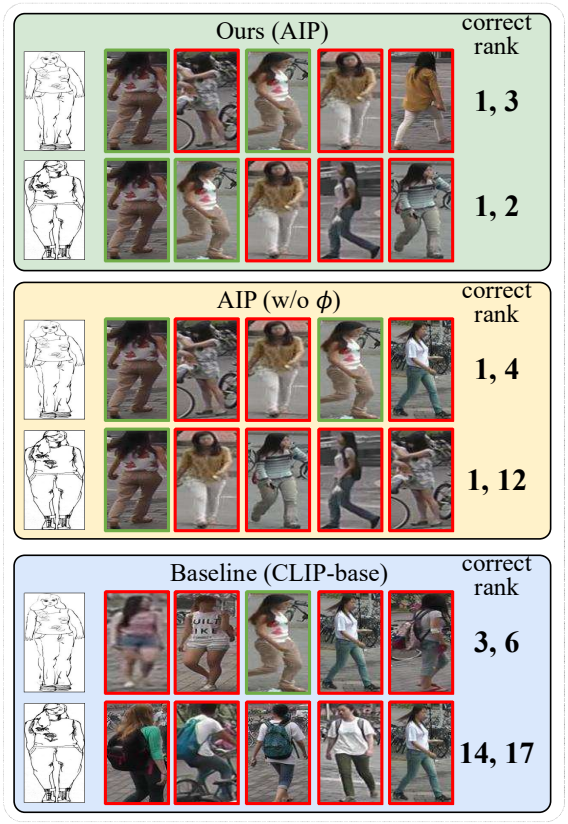

Figure 1. More Qualitative results of the proposed AIP and the CLIP-base. Top-5 retrieval results are shown, with correct retrievals outlined in green and incorrect ones in red. Each query has two corresponding RGB images in total. "AIP (w/o ϕ) means without the style prompt generator applied.

method	mAP	R1
PT	71.0	98.4
PT + LN	86.6	99.4
PT + Sketchy	72.2	99.0
PT + LN + Sketchy	87.2	99.7

Table 4. Results of our method on traditional RGB re-ID task dataset Market-1501.

Stage	Time cost (s)	
	Single step	Whole stage
1) CIA	1.00	1000
2) ISA	1.38	1380

Table 5. Training time cost (in seconds) for each stage of the proposed AIP method. "Single step" indicates the time cost for processing one step. "Whole stage" means the total time cost for completing the entire stage.

Model	Batch Size	FLOPs(G)	Param (M)	Time (ms)
ViT-B/16	32	562.6	85.8	34.9
AIP (ours)		791.5	102.4	56.2

Table 6. Floating Point Operations (FLOPs in Giga), Parameter count (in Millions), and forward time cost (in Million Seconds) of the proposed AIP framework. By default, a batch (batch size is 32) image of size 224×224 serves as input. "ViT" refers to a standard base version of the ViT model (ViT-B/16) [2] for comparison.

promising, as style variation in sketch-RGB re-ID resembles intra-ID variation in RGB re-ID. However, while RGB variations (e.g., illumination, angle, occlusion) are predictable, style variation is subjective and unpredictable. If external sketch data enhances robustness to style shifts, extra RGB samples should similarly improve alignment under varying conditions. **However, it requires fine-grained multi-category RGB data with rich intra-ID variation.**

To explore this, we use photos from Sketchy as non-person data, simulating intra-ID variations via random crops, illumination changes, etc. As shown in the bottom table (rows with "+Sketchy"), adding non-person data still improves performance, though gains are less pronounced than in sketch-RGB re-ID. This is expected as: **1)** the augmented Sketchy data provides limited variation, as it is manually adapted for this experiment. More diverse non-person data could potentially yield greater benefits. **2)** the RGB person data (more than 10k train photos with shared cameras for test) is already more diverse than the sketch person data (only 2k train sketches with non-overlap style for test)

It should be noted that we perform this simple toy experiment only to show whether the idea of utilizing non-pedestrian data can also benefit traditional RGB re-ID. We do not tend to compare or push forward the performance in traditional RGB re-ID as it is a different task to our research work in the main manuscript.

5. Model efficiency

In this section, we evaluate the computational cost of the proposed AIP method during both the training and testing. All experiments were conducted on an NVIDIA RTX 4090 GPU. The ViT-B/16 [2] model, which shares the same architecture as the backbone of the image encoder in CLIP, is used as a baseline for comparison.

Training Phase We assess the time cost for each stage of the proposed AIP method during training. As shown in Table 5, the first CIA stage (left side of Figure 2 of the main manuscript) requires approximately 1 second per forward and backward pass, amounting to around 1000 seconds for 1000 steps of training. The second ISA stage (right side of Figure 2 of the main manuscript) takes about 1.38 seconds

per step, resulting in approximately 1400 seconds for 1000 steps.

Testing Phase. We evaluate both the FLOPs (Floating Point Operations) and inference time for a batch of images as input during testing. The batch size is set to 32 and each image has a resolution of 224×224, consistent with the training stage. As shown in Table 6, for a batch of images, the FLOPs of the proposed AIP method is 791.5G, moderately higher than the base ViT-B/16 model's 562.6G. This increase in computational cost is attributed to the prompt generator ϕ and the additional prompts, which enhance the model's adaptability and performance. Consequently, the inference time for batch feature extraction increases from 34.9ms (ViT-B/16) to 56.2ms, representing a reasonable trade-off for improved generalization and robustness.

References

[1] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis. In *ACM Int. Conf. Multimedia*, pages 4012–4020, 2022. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021. 2

[3] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Int. Conf. Comput. Vis.*, pages 11823–11832, 2021. 1

[4] Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing*, 403:294–303, 2020. 1

[5] Kejun Lin, Z. Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In *ACM Int. Conf. Multimedia*, pages 2078–2089, 2023. 1

[6] Xingyu Liu, Xu Cheng, Haoyu Chen, Hao Yu, and Guoying Zhao. Differentiable auxiliary learning for sketch re-identification. In *AAAI*, pages 3747–3755, 2024. 1

[7] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *ACM Int. Conf. Multimedia*, pages 609–617, 2018. 1

[8] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 1

[9] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2765–2775, 2023. 1

[10] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35 (4):1–12, 2016. 1

[11] Hanzhe Sun, J. Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In *ACM Int. Conf. Multimedia*, pages 5333–5341, 2022. 1

[12] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xiaocui Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14308–14317, 2022. 1

[13] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Eur. Conf. Comput. Vis.*, pages 229–247. Springer, 2020. 1

[14] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Int. Conf. Comput. Vis.*, pages 13567–13576, 2021. 1

[15] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 799–807, 2016. 1

[16] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2153–2162, 2023. 1

[17] Yukang Zhang, Y. Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *ACM Int. Conf. Multimedia*, pages 788–796, 2021. 1

[18] Yiyuan Zhang, Yuhao Kang, Sanyuan Zhao, and Jianbing Shen. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 18:1554–1565, 2022. 1

[19] Yafei Zhang, Yongzeng Wang, Huafeng Li, and Shuang Li. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In *ACM Int. Conf. Multimedia*, pages 3347–3355, 2022. 1

[20] Fengyao Zhu, Yu Zhu, Xiaoben Jiang, and Jiongyao Ye. Cross-domain attention and center loss for sketch re-identification. *IEEE Transactions on Information Forensics and Security*, 17:3421–3432, 2022. 1