Supplementary Materials: Dual-Granularity Cross-Modal Identity Association for Weakly-Supervised Text-to-Person Image Matching

Yafei Zhang
Faculty of Information Engineering and Automation, Kunming University of Science and Technology
Kunming, China
zyfeimail@163.com

Yongle Shang
Faculty of Information Engineering and Automation, Kunming University of Science and Technology
Kunming, China
sylmail99@163.com

Huafeng Li*
Faculty of Information Engineering and Automation, Kunming University of Science and Technology
Kunming, China
hfchina99@163.com

1 Visualization Analysis

To further illustrate the effectiveness of our method, we present the t-SNE visualization of pedestrian features in 2D space, as shown in Figure 1. Dots represent pedestrian image features, triangles represent text features, and different colors denote different pedestrian identities. Compared to the baseline method, our approach clusters features of the same identity more effectively and better separates different identities, indicating that our model performs better in pedestrian identity matching. To illustrate the model's cross-modal alignment, we perform attention visualization. As shown in Figure 2, under text guidance, the model effectively attends to prominent visual cues such as clothing and accessories, particularly in vibrant color regions. Furthermore, as seen in Figure 2 (a), words related to these visual cues are assigned high attention by the model, further demonstrating its cross-modal alignment capability.

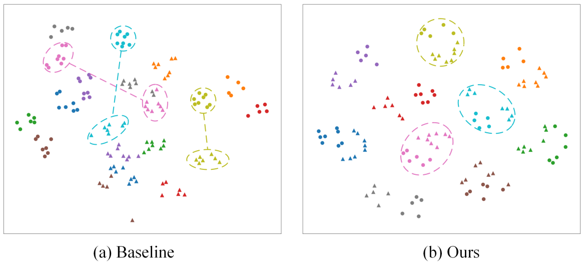
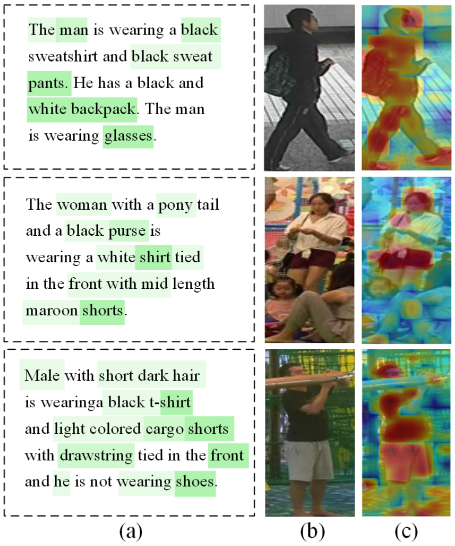

(a) Baseline
(b) Ours

Figure 1: t-SNE visualization of features on the RSTPReid dataset. (a) Features extracted by baseline method. (b) Features extracted by the proposed method. Dots represent pedestrian image features, triangles represent text features, and different colors denote different pedestrian identities.

2 Further Discussion

Analysis of Cross-Dataset Generalization Ability. To evaluate the generalization capability of our method, we conduct cross-dataset experiments among CUHK-PEDES [3], ICFG-PEDES [1], and RSTPReid [7]. As shown in Table 1, although cross-domain performance inevitably drops due to domain shift, our method consistently demonstrates reasonable transferability across all three datas ets. Notably, when trained on CUHK-PEDES and evaluated on RSTPReid, our approach achieves a Rank-1 accuracy of 53.45 %, reflecting a certain degree of robustness in cross-domain scenarios.


The man is wearing a black sweatshirt and black sweat pants. He has a black and white backpack. The man is wearing glasses.

The woman with a pony tail and a black purse is wearing a white shirt tied in the front with mid length maroon shorts.

Male with short dark hair is wearinga black t-shirt and light colored cargo shorts with drawstring tied in the front and he is not wearing shoes.

Figure 2: Visualization of cross-modal attention maps on the CUHK-PEDES dataset. (a) Attention strength map for each word in the Text, (b) Original Input Image, (c) Attention strength map for the pedestrian image. Darker green in the text indicates higher model attention, while warmer colors in the heatmap in (c) indicate higher attention strength from the model.

Table 1: Cross-dataset generalization performance (%) across CUHK-PEDES, ICFG-PEDES, and RSTPReid. The * indicates that the model is trained and evaluated on the same dataset.

Train Datasets	Test on CUHK-PEDES		Test on ICFG-PEDES		Test on RSTPReid	
	R1	mAP	R1	mAP	R1	mAP
CUHK-PEDES	73.06*	64.88*	42.43	21.51	53.45	39.45
ICFG-PEDES	34.62	30.99	63.71*	36.87*	47.05	37.79
RSTPReid	32.26	29.25	31.73	20.06	61.30*	47.20*

Impact of text quality on retrieval performance. To further investigate the limitations of our method under challenging scenarios, we analyze failure cases in which retrieval performance deteriorates due to incomplete or ambiguous textual descriptions. These cases typically arise in complex real-world settings, where text inputs may lack critical semantic cues or exhibit noise. To

[*] Corresponding author.

assess the robustness of our model under such conditions, we simulate degraded text inputs by applying random masking at varying ratios to the query texts. This controlled perturbation introduces different levels of semantic degradation, allowing us to evaluate the method's sensitivity to text quality. Specifically, we compare our approach with IRRA [2], a supervised method using the same CLIP backbone, with performance comparable to ours. As shown in Table 2, retrieval performance declines for both methods as the masking ratio increases. However, our method exhibits progressively larger advantages over IRRA as the masking ratio increases, indicating a superior capacity to cope with textual degradation. In particular, when the mask ratio reaches 0.5, our method achieves notable improvements over IRRA, with Rank-1 accuracy gains of 3.62%, 3.17%, and 2.30%, and mAP gains of 2.44%, 1.17%, and 1.67% on CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively. These results clearly demonstrate the effectiveness of our cross-modal structure modeling and inconsistent information pairing strategy in mitigating the adverse impact of incomplete or noisy text inputs, thereby enhancing retrieval robustness under weakly supervised settings.

Table 2: Impact of Text Modality Quality on Text-to-Person Image Matching Performance (%). Results are reported on CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets.

Dataset	Mask Ratio	IRRA		Ours		Δ = Ours − IRRA	
		R1	mAP	R1	mAP	R1	mAP
CUHK-PEDES	0.1	67.12	60.87	67.21	60.15	+0.09	-0.72
	0.2	59.73	54.56	60.59	54.72	+0.86	+0.16
	0.3	51.24	47.43	52.89	48.15	+1.65	+0.72
	0.4	41.24	39.02	44.30	40.92	+3.06	+1.90
	0.5	31.79	31.07	35.41	33.51	+3.62	+2.44
ICFG-PEDES	0.1	58.38	34.59	58.69	33.79	+0.31	-0.80
	0.2	52.65	30.82	53.18	30.27	+0.53	-0.55
	0.3	46.07	26.76	47.65	26.65	+1.58	-0.11
	0.4	37.62	22.05	40.12	22.73	+2.50	+0.68
	0.5	29.24	17.26	32.41	18.43	+3.17	+1.17
RSTPReid	0.1	53.55	42.87	55.40	43.66	+1.85	+0.79
	0.2	47.15	38.37	48.85	38.87	+1.70	+0.50
	0.3	39.91	32.93	41.75	33.99	+1.84	+1.06
	0.4	33.75	28.58	35.75	29.78	+2.00	+1.20
	0.5	25.55	22.47	27.85	24.14	+2.30	+1.67

Evaluation of Model Complexity and Performance. To comprehensively evaluate the complexity of our proposed method, we report the number of parameters, FLOPs, and inference time on the CUHK-PEDES dataset, as summarized in Table 3. Inference time is measured by the total time required to evaluate the entire CUHK-PEDES test set (6156 texts, 3074 images) with all images resized to 384 × 128. Due to the lack of publicly available code and weights for CPCL [6], we select CMMT [5] as a representative weakly-supervised method. For comparison, we also include IRRA [2] and LuP-MLLM [4], both of which are supervised approaches utilizing the same CLIP backbone. As shown, our model contains 44.9M fewer parameters than IRRA and LuP-MLLM. Although our FLOPs are slightly higher (by approximately 0.72G), resulting in longer inference time, the overall complexity remains within a reasonable range.

More importantly, unlike supervised methods that rely on extensive identity annotations, our method achieves comparable performance without such annotations, demonstrating a favorable balance between efficiency and effectiveness. We will further explore more lightweight architecture designs in future work to enhance real-world applicability.

Table 3: Comprehensive evaluation of model performance and complexity on the CUHK-PEDES dataset. WS: weakly supervised; S: supervised.

Method	Parameters (M)	FLOPs (G)	Inference Time (s)	R (%)	mAP (%)
CMMT (WS)	41.8	4.02	22.5	57.10	-
IRRA (S)	194.5	19.53	14.5	73.38	66.13
LuP-MLLM (S)	194.5	19.53	13.6	78.13	68.75
Ours (WS)	149.6	20.25	34.9	73.06	64.88

References

[1] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666* (2021).

[2] Ding Jiang and Mang Ye. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2787–2797.

[3] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5187–5196.

[4] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 17127–17137.

[5] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. 2021. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 11395–11404.

[6] Yanwei Zheng, Xinpeng Zhao, Chuanlin Lan, Xiaowei Zhang, Bowen Huang, Jibin Yang, and Dongxiao Yu. 2024. CPCL: Cross-Modal Prototypical Contrastive Learning for Weakly Supervised Text-based Person Re-Identification. *arXiv preprint arXiv:2401.10011* (2024).

[7] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia.* 209–217.