Supplementary Material for "Prompt-driven Transferable Adversarial Attack on Person Re-Identification with Attribute-aware Textual Inversion"

Yuan Bian[1,2], Min Liu[1,2], Yunqi Yi[1,2], Xueping Wang[3], Shuai Jiang[1,2], Yaonan Wang[1,2]
[1]School of Artificial Intelligence and Robotics, Hunan University
[2]National Engineering Research Center of Robot Visual Perception and Control Technology
[3]College of Information Science and Engineering, Hunan Normal University
{yuanbian, liu_min, y0512321, wang_xueping, svyj, yaonan}@hnu.edu.cn

In the supplementary material, we would like to show more experiments, visualization results and broader application discussion of our methods.

1. Experiments

1.1. Real-world Implications

Due to privacy and legal concerns in real surveillance systems, we simulate real-world implications by testing AP-Attack on real-world pedestrian data and low-resource scenarios. Specifically, we collected images of 50 pedestrians from practical scenarios, integrated them into the Market dataset for retrieval experiments and evaluated attack time on GPU(1050Ti)/CPU(i5-7300HQ) platforms using PyTorch/C++. Results in Fig. 1 show our method generates effective adversarial perturbations, reducing mAP from 74.7% to 32.4% with an inference time of 8.9/90.3ms(enables real-time attacks on 10-30 fps surveillance systems), showing practical for real-world attack.

1.2. Prompt Template Selection

We employed ChatGPT to generate several templates that accurately describe pedestrian attributes. As demonstrated in Tab. 2, these fine-grained templates consistently deliver satisfactory results, with the template we selected achieving the best overall performance. A comparison of the 5-th and 6-th prompts reveals that the inclusion of word 'person' in the prompt significantly improves the inversion network's ability to learn human attributes, enhances the attacker's capacity to disrupt specific attributes through perturbations.

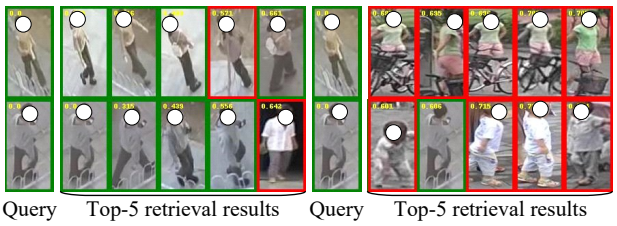

Figure 1. Retrieval results of BOT(Market) before(Left) and after(Right) our attack on real-world data.

Table 1. Computation cost and cross-model&dataset performance.

Methods	Parameters	FLOPs	Training Time	aAP ↓	mDR ↑
Baseline	8.419 K	180.355 M	4.0 h	45.8	40.3
AP-Attack	8.419 K	180.355 M	5.2 h	**21.2**	**72.4**

1.3. Computation Complexity and Training Time

Tab. 3 shows that our method has the same computational complexity during inference as the baseline, as both use a simple generator after training. For training time, our two-stage method only increases by 30%, taking 5.2 GPU hours.

1.4. Semantic Vocabulary Details

In order to verify the effectiveness of our textual inversion networks, we aim to interpret the learned pseudo-tokens as meaningful words. To achieve this, we first established an attribute-specific semantic vocabulary using ChatGPT, where each attribute corresponds to a distinct set of meaningful words. These vocabularies encapsulate both color and descriptive features associated with each attribute. The specific semantic vocabularies for various attributes are as follows:

- **Upper Body** : red, yellow, blue, green, black, white, pink, purple, orange, gray, brown, jacket, shirt, suit, checkered, solid, turtleneck, hoodie, sweater, coat, tracksuit, waistcoat, dress.
- **Lower Body**: red, yellow, blue, green, black, white, pink, purple, orange, gray, brown, shorts, trousers, tights, flares, sweatpants, jeans.
- **Hair**: black, yellow, brown, long, short, wavy, straight.
- **Shoes**: red, yellow, blue, black, white, pink, purple, orange, gray, brown, boots, sneakers, sandals, leather.

Table 2. Results of cross-model&dataset attack using different prompts.

No	Prompt	aAP↓	mDR↑
1	*'A photo capturing someone dressed in S_1 on the upper body, with S_2 on the lower body, sporting a S_3 hairstyle and S_4 shoes, and handling S_5.'*	21.9	71.4
2	*'A close-up photo of a person styled in S_1 as their upper garment, layered with S_2 as the lower piece. With a S_3 hairstyle and S_4 shoes, they are equipped with S_5, either held in hand or worn.'*	22.8	70.3
3	*'A photo of a person wearing S_1 on top, S_2 underneath, S_3 hairstyle, S_4 shoes, and holding S_5 in their hand or over their shoulder.'*	23.3	69.6
4	*'An image showing a person in S_1 as their top layer, matched with S_2 on the bottom, styled with a S_3 hairstyle, wearing S_4 shoes, and keeping S_5 close by.'*	22.4	70.8
5	*'A photo of a S_1 wearing S_2 on top, S_3 underneath, S_4 hairstyle, S_5 shoes, carrying S_6.'*	23.6	69.2
6	*'A photo of a person wearing S_1 on top, S_2 underneath, S_3 hairstyle, S_4 shoes, carrying S_5.'**(Ours)***	**21.2**	**72.4**

- **Bag**: red, yellow, blue, green, black, white, pink, purple, gray, brown, suitcase, backpack, bag, nobag.

We compute the similarity between the pseudo-token and the semantic vocabulary. The word with the highest similarity is assigned a font size of 14pt, while the word with the lowest similarity is set to 4pt. The font sizes of the remaining words are determined by linearly interpolating the similarity scores, where the font size is adjusted based on the distance from the maximum similarity, creating a proportional representation of each word's relevance to the pseudo-token.

1.5. SOTA Methods Reproduction Details

Firstly, we note that all the methods compared are trained using the IDE [12] model, which was trained on the DukeMTMC [6] dataset as the surrogate model. Specifically, for the LTP [7] and BIA [10] methods, adversarial perturbations are generated by targeting intermediate feature layers. In our experiments, we tested different feature layers to identify the one that optimizes adversarial transferability. Our results indicate that the last feature layer provides the best transferability for adversarial attacks on the re-id task. As for PDCL-Attack [9], this method employs prompt learning to guide the semantic learning of class features. For our experiments, we directly utilized the prompts learned from the CLIP-ReID [2] method for pedestrian images, which were then used to reproduce the PDCL-Attack methodology.

1.6. Failure Cases Analysis

Fig. 3 shows that attacking targets with distinct-textured clothing remains challenging. The second row of Fig. 3 reveals that mis-retrieved targets retain query-like textures but differ in color, indicating our method disrupts semantics rather than texture patterns. We will further explore the potential methods for texture pattern disruption with the guidance of prompt.

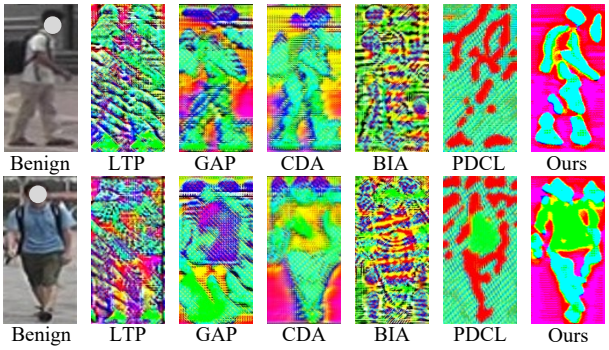

Figure 2. The perturbations generated by different attack methods.

Figure 3. Adversarial results of failure cases.

Table 3. Cross-model&dataset results with dynamic attributes.

Attributes	$S_{1,2}$	$S_{1,2,3}$	$S_{1,2,4}$	$S_{1,2,5}$	$S_{1,2,3,4}$	$S_{1,2,3,5}$	$S_{1,2,3,4,5}$
aAP ↓	30.1	29.2	28.7	28.4	27.3	25.2	21.2
mDR ↑	60.8	62.0	62.6	63.1	64.4	67.2	72.4

1.7. Attributes Analysis

The selection of the five specific attributes was manually determined, as these attributes represent the most salient features of pedestrian appearance. Additionally, we conducted a redundancy analysis by initially considering the most prominent upper and lower clothing attributes as the baseline and incrementally incorporating other attributes as loss constraints. As demonstrated in Tab. 3, the optimal performance was achieved when all five attributes were included, with incremental gains observed as more attributes were added. This validates that the inclusion of all five at-

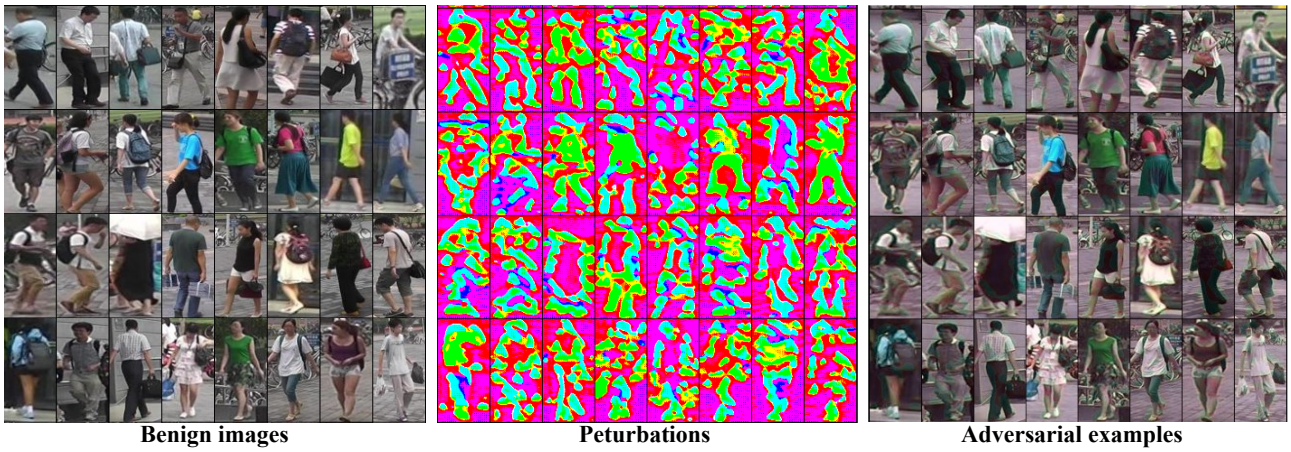

Figure 4. The visualization of benign images, perturbations and adversarial examples on Market dataset.

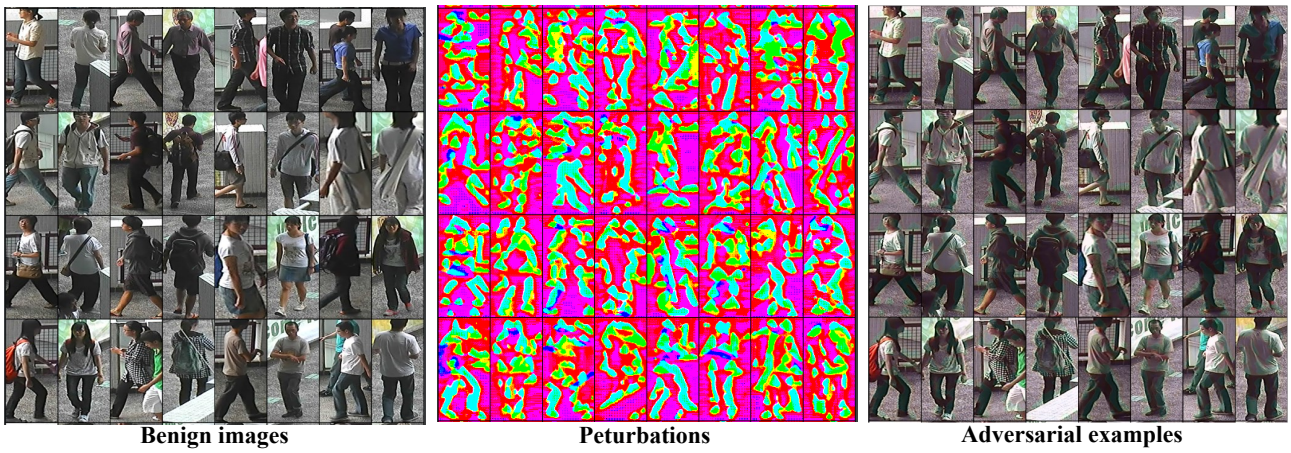

Figure 5. The visualization of benign images, perturbations and adversarial examples on CUHK03 dataset.

Table 4. Results of cross-model attack on Occlued-DukeMTMC.

Methods	None	MetaAttack	Mis-Ranking	MUAP	PDCL-Attack	Ours
aAP ↓	40.8	31.3	10.5	12.0	15.9	**7.8**
mDR ↑	0.0	23.4	72.2	70.5	61.0	**81.1**

tributes is non-redundant and contributes to the overall effectiveness.

1.8. Missing Attributes Scenarios

We conduct experiments on occluded dataset Occluded-DukeMTMC [5], where queries are all occluded by objects. The cross-model results in Tab. 4 shows that ours also get SOTA performance in this missing attributes scenarios. This is because only training stage needs attribute information, once adversarial generator is trained, it strives to disrupt perceptible attributes.

2. Visualization

2.1. Adversarial Perturbations

In order to better compare the superiority of our methods, we visualize in Fig. 2 the perturbations generated by different methods, showing our method generates more comprehensive and semantically meaningful perturbations than others.

2.2. Adversarial Examples

We showcase additional adversarial examples generated on the Market [11] and CUHK03 [3] datasets, as illustrated in Fig. 4 and Fig. 5. As depicted, the adversarial perturbations are visually indistinguishable yet effectively disrupt all semantic information related to human attributes, demonstrating the strength of our approach in attacking comprehensive human semantics.

Retrieval results on benign queries.

Retrieval results on attacked queries.

Figure 6. The rank-10 retrieval results of BOT(Market)[4] before and after our attack on Market-1501.

2.3. Destruction of re-id models

We illustrate the destruction effects on the Market dataset [11] by presenting Rank-10 matches from different re-id models (i.e., BOT(Market) [4], MGN(Market) [8], TransReID(Market) [1]) and IDE(Market)[12] before and after adversarial attacks, as shown in Fig. 3, Fig. 4, and Fig. 5. In these visualizations, green boxes represent correct matches, red boxes indicate mismatches, and query images appear in the first column. These results effectively showcase how our approach disrupts the matching accuracy of various re-id models. These results effectively showcase how our approach disrupts the matching accuracy of various re-id models.

3. Broader Application Discussion

Our method can be migrated to tasks where target objects belong to one general category, such as face recognition, vehicle re-id, and fine-grained image classification, by setting specialized prompts for each task. For other tasks with diverse object categories, our method can be explored and extended to automatically generate object text attribute descriptions for different targets using multimodal foundation models to enable fine-grained feature disruption.

References

[1] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Int. Conf. Comput. Vis.*, pages 15013–15022, 2021. 4, 5

[2] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, pages 1405–1413, 2023. 2

[3] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 152–159, 2014. 3

[4] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. 4

[5] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Int. Conf. Comput. Vis.*, pages 542–551, 2019. 3

[6] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Eur. Conf. Comput. Vis.*, pages 17–35. Springer, 2016. 2

[7] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Adv. Neural Inform. Process. Syst.*, 34:13950–13962, 2021. 2

[8] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Int. Conf. Multimedia*, pages 274–282, 2018. 4, 5

[9] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial attacks. In *Eur. Conf. Comput. Vis.*, pages 36–53, 2024. 2

[10] Qilong Zhang, Xiaodan Li, YueFeng Chen, Jingkuan Song, Lianli Gao, Yuan He, et al. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In *Int. Conf. Learn. Represent.*, 2022. 2

[11] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, pages 1116–1124, 2015. 3, 4

[12] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2, 4, 5

Retrieval results on benign queries.

Retrieval results on attacked queries.

Figure 7. The rank-10 retrieval results of MGN(Market)[8] before and after our attack on Market-1501.

Retrieval results on benign queries.

Retrieval results on attacked queries.

Figure 8. The rank-10 retrieval results of Transreid(Market)[1] before and after our attack on Market-1501.

Retrieval results on benign queries.

Retrieval results on attacked queries.

Figure 9. The rank-10 retrieval results of IDE(Market)[12] before and after our attack on Market-1501.