This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Abstract

Person re-identification (re-id) models are vital in security surveillance systems, requiring transferable adversarial attacks to explore the vulnerabilities of them. Recently, vision-language models (VLM) based attacks have shown superior transferability by attacking generalized image and textual features of VLM, but they lack comprehensive feature disruption due to the overemphasis on discriminative semantics in integral representation. In this paper, we introduce the Attribute-aware Prompt Attack (AP-Attack), a novel method that leverages VLM's image-text alignment capability to explicitly disrupt fine-grained semantic features of pedestrian images by destroying attribute-specific textual embeddings. To obtain personalized textual descriptions for individual attributes, textual inversion networks are designed to map pedestrian images to pseudo tokens that represent semantic embeddings, trained in the contrastive learning manner with images and a predefined prompt template that explicitly describes the pedestrian attributes. Inverted benign and adversarial fine-grained textual semantics facilitate attacker in effectively conducting thorough disruptions, enhancing the transferability of adversarial examples. Extensive experiments show that AP-Attack achieves state-of-the-art transferability, significantly outperforming previous methods by 22.9% on mean Drop Rate in cross-model&dataset attack scenarios. The codes are at https://github.com/yuanbianGit/AP-Attack.

1. Introduction

Person re-identification models are widely employed in security-critical surveillance systems, aiming to retrieve the target person [60, 66]. Despite the significant progress made by deep learning-based re-id methods [1, 5, 17, 27, 28, 47], they also inherit the vulnerability of deep neural networks, *i.e.*, the addition of imperceptible perturbations on benign images can destroy model performance [15, 33]. The intriguing transferability of adversarial examples (AEs) across different models [43] further exposes real-world surveillance systems to safety threats. To obtain reliable re-id models, it is paramount to test the robustness of re-id models by generating highly transferable AEs.

Cross-model and cross-dataset transferability are crucial for adversarial attacks on black-box re-id models due to the uncertainty architecture of the target model and the signif-

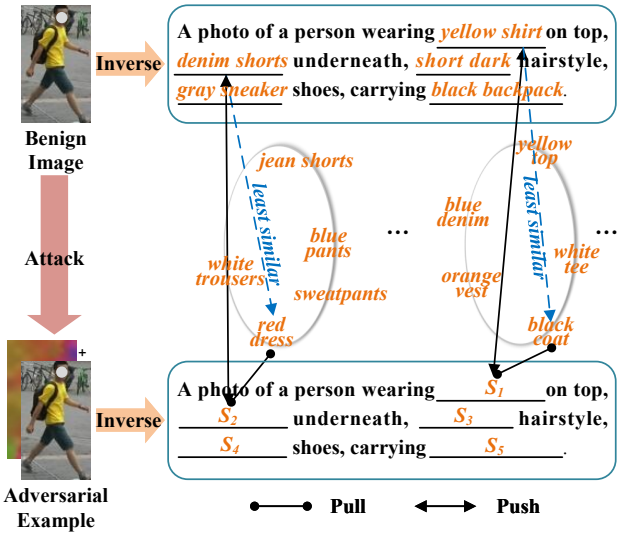

Figure 1. The core idea of our Attribute-aware Prompt Attack. We leverage the image-text alignment capability of vision-language model to invert pedestrian image into pseudo-word tokens S_* that represent attribute-specific semantics, guided by a predefined prompt template explicitly describing person attributes. With inverted semantics, our method enables fine-grained attack by pushing adversarial semantics away from benign ones and pulling them toward the least similar semantics, achieving thorough disruption across all attribute semantic spaces.

Corresponding author: Min Liu (liu_min@hnu.edu.cn)

icant domain gap between training data and unseen query images [40, 54]. While extensive researches have been conducted to enhance cross-model transferability via input transformation [10, 51], gradient modification [9, 14], model ensembling [25, 29], and intermediate feature attacks [19, 48], few studies have attempted to attack generalized features [26, 62] or maximize the fooling gap [34] to improve cross-dataset transferability. These methods primarily depend on the choice of surrogate models, focusing on their architectural similarity to the target model and the generality of the features they extract [59].

Lately, vision-language models [20, 38, 56, 58], such as CLIP [38], have demonstrated excellence in learning generic representations by training on large-scale Internet image-text pair data and the joint vision-language space of VLM enables zero-shot transfers across downstream tasks with natural language prompts [32, 41, 69]. Appreciating the generalization advantage and image-text alignment capability of VLM, some most recent studies [2, 12, 59, 61] have introduced VLM to facilitate the cross-model and cross-dataset transferability of AEs. These methods not only delivered optimal loss gradients by incorporating generic image features, but also introduced predefined prompts like 'A photo of [CLASS]' [2, 12] or learnable prompts such as '$[V_1]$ $[V_2]\ldots[V_M]$ [CLASS]' [54, 59], to guide the implicit semantic destroy in textual cues, thus improving the transferability of AEs.

However, the above VLM-based attack methods solely leveraged global image features or integral textual semantic representation to steer the attacker's learning, which may hinder transferability by overemphasizing discriminative local features and resulting in less comprehensive disruption. To ensure broader and thorough damage to underlying representation, explicitly disrupting fine-grained semantic features is crucial. Nevertheless, destroying fine-grained semantic features in existing prompt-driven attack methods on the classification task is challenging because the attributes of each category differ, making it impractical to create fine-grained attribute guidance through text prompts. But it is worth noting that the re-id task is a retrieval task focusing on distinguishing different identities within the pedestrian category, where all images share the same semantic attributes (*e.g.*, gender, clothing, hairstyle). Therefore, we can leverage VLM's powerful cross-modal comprehension capability to guide the fine-grained image feature disruptions by perturbing semantic text prompts that describe specific attributes of a person, like Fig. 1 shows, leading to more thorough disruption and consequently enhancing the transferability of adversarial person images.

Based on the above analysis, we propose a novel Attribute-aware Prompt Attack (AP-Attack) method to achieve transferable fine-grained semantic perturbations on person re-identification. Specifically, vision-language model CLIP [38] is adopted in our method, and the adversarial generator is trained to produce delta perturbations. In pursuit of explicit person attributes information, we construct a personalized prompt template for individual images: *'A photo of a person wearing* $\underline{S_1}$ *on top,* $\underline{S_2}$ *underneath,* $\underline{S_3}$ *hairstyle,* $\underline{S_4}$ *shoes, carrying* $\underline{S_5}$*.'*, in which pseudo-tokens S_* denotes semantic language description related to each attribute. To obtain these S_*, textual inversion technique [13], which learns to capture unique and varied image concepts to a single word embedding, is introduced. Multiple inversion networks, each corresponding to a specific pedestrian attribute, are designed to generate attribute-aware semantic pseudo-tokens S_*, which are then integrated into the predefined template. With composed text prompts and corresponding images, inversion networks are trained in contrastive learning way and subsequently inverse benign and adversarial semantic representations. In this context, prompt-driven semantic attack loss is devised to push the adversarial semantics away from the original ones while pulling them closer to the least similar semantics, guiding the learning of the adversarial generator to destroy fine-grained semantics. By applying this attribute-aware attack across all attribute semantic spaces, our AP-Attack can thoroughly destroy features of pedestrian images, resulting more transferable adversarial examples.

In summary, our main contributions are as follows:

- We propose a novel Attribute-aware Prompt Attack method that leverages vision-language model's cross-modal comprehension to perturb fine-grained semantic features of pedestrian images.
- To our best knowledge, our method is the first attempt that introduces textual inversion technique to explicitly extract attribute-aware semantic representation to boost the transferability of adversarial examples.
- Our AP-Attack achieves state-of-the-art attack transferability across various domains and model architectures, especially surpassing previous approaches by 22.9% on mean Drop Rate in cross-model&dataset attack scenarios.

2. Related Works

2.1. Adversarial Attack against Re-id

Re-id models are widely deployed in surveillance systems with stringent security requirements, making their robustness against malicious attacks a critical concern. Unlike classification tasks, re-id is an image retrieval task, and numerous white-box attack methods leveraging adversarial feature similarity metrics have been proposed [3, 6, 68]. Given that attackers often need to target unknown models and unseen queries in real-world scenarios, several studies [8, 42, 46, 52, 53] have explored cross-model and cross-dataset transferable attacks for black-box re-id systems. Yang *et al.* [52] and Subramanyam [42] improved cross-

dataset transferability by utilizing multi-source datasets in meta-learning framework for additive and generative attacks, respectively. Wang *et al*. [46] presented a Mis-Ranking formulation and multi-stage discriminator network to extract general and transferable features to boost cross-dataset general attack learning. Ding *et al*. [8] introduced a model-insensitive regularization technique designed to facilitate universal attacks across diverse CNN architectures. Meanwhile, Yang *et al*. [53] proposed a combinatorial attack strategy that integrates functional color manipulation and universal additive perturbations to boost the transferability of attacks across both models and datasets.

2.2. VLM-guided Adversarial Attack

Vision-language models have garnered significant attention for their ability to learn highly generalizable representations through contrastive pretraining on large-scale image-text pairs [20, 38, 56, 58]. Given their broad generalization capacity and alignment of visual and language spaces, VLM have become an appealing target for adversarial attacks. Abhishek *et al*. [2] introduced GAMA, the first VLM-based attack targeting multi-object scenes, using text prompts to force adversarial images to align with the least similar text embeddings. Fang *et al*. [12] enhanced the transferability of multi-target adversarial attacks by incorporating VLM textual knowledge to exploit the rich semantic information of target categories. Ye *et al*. [59] devised an optimization strategy to enhance transferability through iterative attacks on visual inputs while defending text embeddings. Yang *et al*. [54] proposed PDCL-Attack to facilitate the generalization of classes text feature by prompt learning and formulated a prompt-driven contrastive loss to guide the attack training. Notably, these VLM-based attack methods are tailored for classification task, leveraging class-specific text labels for guidance. However, re-id aims to distinguish individual identities within the single 'person' class, making these approaches unsuitable. The PDCL-Attack method is an exception, as it applies prompt learning to generate prompts for each class. However, its reliance on global semantic features may lead to excessive optimization of highly discriminative features, limiting its effectiveness for thorough destroy for pedestrian images.

In contrast to current person re-id attacks and other VLM-based attacks, our approach seeks to thoroughly undermine fine-grained semantic features by leveraging attribute-aware textual inversion networks, utilizing the image-text comprehension capabilities of VLM.

3. Method

In this section, we introduce our AP-Attack method, with an overview provided in Fig. 2 and algorithm summary in Algorithm 1. The preliminaries of the Contrastive Language-Image Pre-training (CLIP) model and generative adversarial attack definition are presented in Sec. 3.1. Details of the attribute-aware textual inversion networks learning are covered in Sec. 3.2, followed by the prompt-driven semantic attack process in Sec. 3.3.

Algorithm 1 Attribute-aware Prompt Attack algorithm

Input: Batch images x, visual encoder $\mathcal{V}$ and textual encoder $\mathcal{T}$, prompt template p, surrogate model $\mathcal{M}$.
Output: Inversion networks f_I, adversarial generator $\mathcal{G}$

Initialize $\mathcal{T}$, $\mathcal{V}$ from pretrained CLIP model and freeze them. Initialize $\mathcal{G}$, f_I randomly. Load $\mathcal{M}$ parameters and freeze.
while in *Textual Inversion Learning* process **do**
  Extract image features v by $\mathcal{V}$ and inverse v to pseudo-tokens S_* by Eq. (4)
  Integrate S_* into p to form $\hat{p}$ and get text embedding $\hat{t}$
  Optimize f_I in contrastive learning manner by Eq. (7)
end while
Freeze the parameter of f_I.
while in *Prompt-driven Semantic Attack* process **do**
  Generate adversarial image x' by $\mathcal{G}$
  Extract image features m, m' by $\mathcal{M}$
  Extract image features v, v' by $\mathcal{V}$ and inverse them to pseudo-tokens S_*, S'_* by Eq. (4)
  Evaluate semantic attack loss by Eq. (11) and optimize $\mathcal{G}$
end while

3.1. Preliminaries

Generative Adversarial Attack. The objective of the proposed AP-Attack is to train the adversarial generator $\mathcal{G}$ to craft perturbations $\mathcal{G}(x)$ for each clean images x. The generated perturbations are applied to produce adversarial examples x' by adding perturbations on input images, aiming to deceive the re-id models into retrieving incorrect results. To ensure the perturbations remain subtle and hard to detect, the maximum perturbation magnitude is constrained by a threshold ϵ.

$$x' = \mathcal{G}(x) + x, \quad \text{s.t.} \|x' - x\|_\infty \leq \epsilon. \tag{1}$$

The adversarial generator is initially trained in a white-box setting, where both the data and the surrogate model are known. Once trained, the generator is kept unchanged and employed to generate perturbations for unseen data to attack black-box models.

Contrastive Language-Image Pre-training. CLIP [38] aims to learn highly generalizable representations by utilizing a dataset of 400 million image-text pairs sourced from the internet for language supervision. CLIP is composed of two primary components: a visual encoder $\mathcal{V}(\cdot)$ that processes images x to image features v by $\mathcal{V}(x)$, and a text encoder $\mathcal{T}(\cdot)$ that transforms tokenized text descriptions p to text representation t by $\mathcal{T}(p)$. With image-text batch $\mathcal{S} = \{(x_n, p_n)\}_{n=1}^{N}$, the core objective is to align images with their corresponding captions while distinguishing mis-

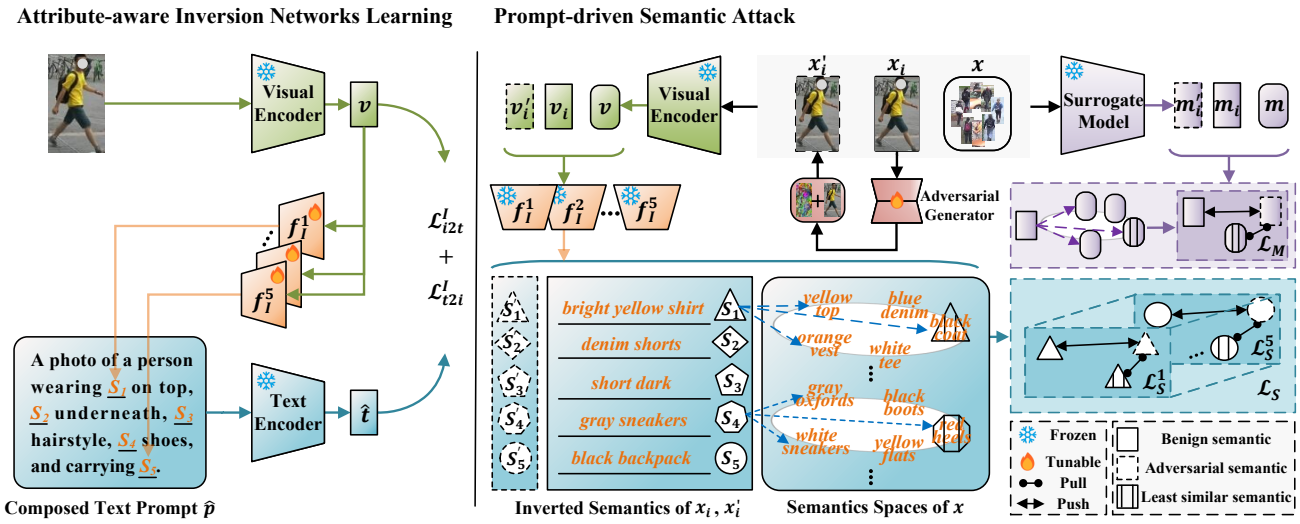

Figure 2. The overview of the proposed Attribute-aware Prompt Attack (AP-Attack) method for person re-id. Our AP-Attack follows two stages. First, attribute-aware inversion networks are trained in the contrastive learning manner with benign pedestrian images and composed text prompts. Then, the trained inversion networks are used to guide the prompt-driven semantic attack. The generated adversarial examples x', benign images x and batch images x_b are fed into surrogate model and VLM visual encoder to produce inverted semantics and surrogate features of them. The adversarial generator is optimized by pushing the adversarial semantics away from the benign ones and pulling them towards the least similar semantics in semantic spaces of each pedestrian attribute and surrogate feature spaces.

matched pairs through contrastive learning by

$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp(sim(v_n, t_n)/\tau)}{\sum_{i=1}^{N}\exp(sim(v_n, t_i)/\tau)}, \quad (2)$$

$$\mathcal{L}_{t2i} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp(sim(t_n, v_n)/\tau)}{\sum_{i=1}^{N}\exp(sim(t_n, v_i)/\tau)}, \quad (3)$$

where τ is the temperature scaling factor, and sim represents cosine similarity.

Utilizing the learned joint vision-language feature space, CLIP facilitates zero-shot transfer to a range of downstream tasks by employing natural language prompts to reference the learned visual concepts, *e.g.*, 'A photo of a [CLASS]' for classification task. However, for re-id task, where labels are typically index-based, there are no accompanying text labels or descriptions for each identity, making it challenging for CLIP to directly transfer its capabilities to re-id. To tackle this challenge, CLIP-ReID [22] and PromptSG [57] introduced the ID-specific learnable prompts 'A photo of a $[X]_1$ $[X]_2$ ... $[X]_M$ person' and 'A photo of a $\underline{S_*}$ person', using automated prompt engineering and textual inversion to craft identity representations. Nevertheless, since these prompts lack explicit descriptions of pedestrian attributes, they are not effective for our method, which requires detailed attribute representation for fine-grained attack.

3.2. Learning Attribute-aware Textual Inversion

Textual inversion [13] is devised in text-to-image generative task to discover pseudo-words within the text encoder's embedding space that encapsulate both high-level semantics and fine visual details, enabling the generation of new scenes based on user-provided natural language instructions. Textual inversion networks have expanded to composed image retrieval [4] and person re-id tasks [57] to retrieve the target object. These textual inversion methods inverse images into coarse-grained textual semantic representations, where a single word embedding is used to represent the visual information of the entire image. Distinct from them, our methods need to map the fine-grained attribute semantics of pedestrian images.

To explicitly inverse the pedestrian images to attribute-aware semantic representations, we first construct a predefined prompt template *'A photo of a person wearing* $\underline{S_1}$ *on top,* $\underline{S_2}$ *underneath,* $\underline{S_3}$ *hairstyle,* $\underline{S_4}$ *shoes, carrying* $\underline{S_5}$*.'*, in which five attributes of pedestrian are described and $\underline{S_*}$ denotes semantic language descriptions related to each attribute. Next, several inversion networks that with the same number of preset attributes are designed to map images to pseudo-tokens that represent each attribute semantics. The inverted pseudo-tokens are then composed to predefined text template. Utilizing composed prompts and corresponding images, inversion networks are trained in the contrastive way in the joint vision-language space of CLIP model.

Specifically, five three-layer fully-connected inversion networks, denoted as f_I, are constructed. During the training of these inversion networks, both the visual encoder $\mathcal{V}$ and the text encoder $\mathcal{T}$ of pretrained CLIP model are kept frozen to provide the joint vision-language space. The pro-

cess begins by passing pedestrian images x through the visual encoder $\mathcal{V}$ to extract global visual features v. These features are then input into the i-th inversion network f_I^i, which maps the visual context to attribute-specific semantic pseudo-tokens S_i.

$$S_i = f_I^i(v). \tag{4}$$

All inverted pseudo-tokens S_* are combined to predefined template p to generate a composed language description $\hat{p}$, which is then undergoes a tokenization process and be fed into the text encoder $\mathcal{T}$ to obtain text embedding $\hat{t}$. Using the original image features v and inverted text embedding $\hat{t}$ pairs, inversion networks are trained by the cycle-consistency contrastive loss to ensure learned pseudo-token effectively align with the semantic information of distinct pedestrian attribute. To handle the cases where images are with the same identity that share the same appearance, we follow [22, 57] to exploit contrastive loss for re-id as

$$\mathcal{L}_{i2t}^I = \frac{1}{N}\sum_{n=1}^{N}\sum_{c^+\in C(n)} \log \frac{\exp(sim(v_n, \hat{t}_{c^+})/\tau)}{\sum_{i=1}^{N}\exp(sim(v_n, \hat{t}_i)/\tau)}, \tag{5}$$

$$\mathcal{L}_{t2i}^I = \frac{1}{N}\sum_{n=1}^{N}\sum_{c^+\in C(n)} \log \frac{\exp(sim(\hat{t}_n, v_{c^+})/\tau)}{\sum_{i=1}^{N}\exp(sim(\hat{t}_n, v_i)/\tau)}, \tag{6}$$

to ensure learned pseudo-tokens S_* are consistent for the same person, where $C(n)$ represents the corresponding samples sharing the same identity as v_n and $\hat{t}_n$. The total contrastive loss for inversion networks is formulated by

$$\mathcal{L}_I = \mathcal{L}_{t2i}^I + \mathcal{L}_{i2t}^I. \tag{7}$$

3.3. Prompt-driven Semantic Attack

Pedestrians are generally recognized as distinct individuals if they differ by even a single semantic information, such as clothing, shoes, or hairstyle, with re-id models relying heavily on these subtle distinctions to accurately identify and differentiate them. In this condition, our method aims to deliberately alter the benign semantic features into other meaningful semantics, thereby misleading the re-id models' recognition. For achieving this, we leverage pre-trained inversion networks, which are capable of generating pseudo-tokens that effectively represent visual attribute features within the joint vision-language space, allowing us to produce both adversarial and clean attribute-aware semantic pseudo-tokens to guide the fine-grained semantic attack.

More formally, adversarial images x' are firstly generated by the adversarial generator $\mathcal{G}$ as defined in Eq. (1). Then, the benign pseudo-tokens S_* for clean images x and the adversarial pseudo-tokens S'_* for perturbed images x' are obtained through the inversion networks f_I. These pseudo-tokens S_* of original batch images form the attribute-specific semantic spaces. In each semantic space, we aim to push the adversarial semantic away from its original images while pulling it closed to its furthest negative semantic by prompt-driven semantic attack loss, which is formulated by

$$\mathcal{L}_S^i = \max\left(0, \|S'_i - S_i^n\|_2 - \|S'_i - S_i\|_2 + \alpha\right), \tag{8}$$

where S_i^n represents the least similar negative semantic in the semantic spaces, α denotes the margin. To boost comprehensive disruption of image features, we apply these constraints across all attribute semantic spaces. The overall prompt-driven fine-grained semantic attack loss is defined as

$$\mathcal{L}_S = \sum_{i=1}^{I}\mathcal{L}_S^i, \tag{9}$$

where I is the number of attribute number crafted in prompt template.

Meanwhile, the generated adversarial and clean images are also input into surrogate models $\mathcal{M}$ to get perturbed features m' and clean features m. The adversarial attack loss that similar to $\mathcal{L}_S$ is conducted by

$$\mathcal{L}_M = \max\left(0, \|m' - m^n\|_2 - \|m' - m\|_2 + \alpha\right), \tag{10}$$

to guide the feature destroy in the surrogate model feature space. Finally, our AP-Attack method optimize the adversarial generator $\mathcal{G}$ by

$$\mathcal{L} = \mathcal{L}_M + \mathcal{L}_S. \tag{11}$$

4. Experiments

4.1. Experimental Setup

Evaluation settings. To assess the effectiveness of our methods, we comprehensively set cross-model, cross-dataset and cross-model&cross-dataset black-box attack scenarios to examine the transferability of generated adversarial examples. The cross-model attack setting involves a black-box target model with different architectures from the surrogate model, while sharing the same training dataset. The cross-dataset attack setting, on the other hand, refers to cases where the victim re-id models trained with different dataset but obtain the same network architecture with surrogate models. **Surrogate and victim models.** For the cross-model attack, we choose classical IDE [66] model as surrogate model and take BOT [30], LSRO [67], MuDeep [37], Aligned [64], MGN [45], HACNN [24], Transreid [17], PAT [35] as the victim re-id models. Significantly, these models are built on diverse backbones, including ResNet [16] (e.g., BOT [30]), ViT [11] (e.g., Transreid [17], PAT [35]), DenseNet [18] (e.g., LSRO [67]), and Inception-v3 [44] (e.g., MuDeep [37]). Additionally, these models represent different architecture types, including global-based (e.g., BOT [30]), part-based (e.g., MGN

Table 1. Results of **cross-dataset** attack: trained on agent model (DukeMTMC) and tested on agent model (MSMT17, Market, CUHK03).

Methods	IDE			aAP↓	mDR↑
	MSMT17	Market	CUHK03		
None	41.9	75.5	52.3	56.6	-
MetaAttack	**3.0**	**4.2**	**3.8**	**3.7**	**93.5**
Mis-Ranking	15.2	26.9	11.1	17.7	68.7
MUAP	3.9	19.3	7.6	10.3	81.9
GAP	5.9	10.4	5.0	7.1	87.5
CDA	7.2	13.3	6.3	8.9	84.2
LTP	5.4	9.1	6.4	7.0	87.7
BIA	3.5	14.8	7.0	8.4	85.1
PDCL-Attack	4.8	7.4	7.7	6.6	88.3
AP-Attack(Ours)	4.2	7.6	5.3	5.7	89.9

Table 2. Results of **cross-model** attack: trained on surrogate model (DukeMTMC) and tested on victim models (DukeMTMC).

Methods	Global-based			Part-based		Attention-based			aAP↓	mDR↑
	BOT	LSRO	MuDeep	Aligned	MGN	HACNN	Transreid	PAT		
None	76.2	55.0	43.0	69.7	66.2	60.2	79.6	70.6	65.0	-
MetaAttack	14.9	44.0	31.8	49.5	57.4	54.6	75.3	64.5	49.0	24.6
Mis-Ranking	14.4	6.8	8.0	16.5	8.4	8.8	34.5	42.9	17.5	73.1
MUAP	16.3	9.2	11.1	23.1	11.4	13.8	34.2	40.4	19.9	69.4
GAP	12.9	14.6	13.7	24.5	16.4	16.5	46.7	45.8	23.9	63.3
CDA	9.6	12.5	12.7	20.8	14.7	15.0	42.3	40.8	21.1	67.6
BIA	14.3	33.1	24.5	44.9	58.0	41.9	71.3	60.8	43.6	32.9
LTP	12.3	22.3	23.3	30.9	37.8	22.5	49.6	45.5	30.5	53.0
PDCL-Attack	11.8	11.1	10.5	22.3	12.6	14.2	37.5	32.0	19.0	70.8
AP-Attack(Ours)	**6.1**	**2.2**	**6.7**	**6.4**	**3.7**	**4.7**	**10.4**	**15.0**	**6.9**	**89.4**

[45]), and attention-based (e.g., HACNN [24]). **Training dataset and test dataset.** For the cross-dataset attack, we train our attacker on the surrogate model that pretrained on DukeMTMC [39] dataset and test it on Market [65], MSMT [49] and CUHK03 [23] pretrained models.

Evaluation metrics. We assess the adversarial performance of generated samples against various re-id models using three metrics: mean Average Precision (mAP) [65], average mAP (aAP), and mean mAP Drop Rate (mDR) [8]. The aAP is defined as

$$aAP = \frac{\sum_{i=0}^{N} mAP_i}{N}, \tag{12}$$

where mAP_i denotes mAP of the i-th re-id model. The mDR metric, indicating the success rate of adversarial attacks across multiple models, is calculated as

$$mDR = \frac{aAP - aAP_{adv}}{aAP}, \tag{13}$$

where aAP represents the average mAP of the re-id models on the benign images and aAP_{adv} on adversarial examples.

Implementation Details. We adopt the ViT-based CLIP-Reid model [22] trained on DukeMTMC [39] as the visual and text encoder for CLIP. The adversarial generator follows the Mis-Ranking approach [46]. Optimization is conducted using the Adam optimizer [21] with a learning rate of 2e-4 for both the adversarial generator and the inversion network parameters. All experiments employ $\mathcal{L}_\infty$-bounded attacks with $\epsilon = 8/255$, setting ϵ as the maximum change per pixel. The training process of our AP-Attack is implemented in PyTorch and runs on one RTX3090 GPU.

4.2. Comparison with State-of-the-art Methods

We evaluate our AP-attack method against state-of-the-art (SOTA) transferable black-box re-id attacks, specifically MUAP [8], Mis-Ranking [46], and MetaAttack [53]. Notably, MetaAttack also includes color-based perturbations, but for consistency, only its additive perturbation performance is compared. Meanwhile, the state-of-the-art transferable generative attack methods GAP [36], CDA [34], LTP [40], BIA [63] and PDCL-Attack [55] are incorporated for comprehensive comparisons. It is worth noting that PDCL-Attack [54] is the latest prompt-driven attack method in literature. All these methods are re-trained with surrogate model IDE [66] on DukeMTMC [39] for fair comparisons.

Comparisons on cross-dataset attack. The results of cross-dataset attack are shown in Tab. 1, from which can be seen that our method gets 5.7% aAP and 89.9% mDR. Our AP-Attack surpasses the SOTA generative attack method PDCL-Attack by 0.9% and 1.6% on aAP and mDR, respec-

Table 3. Results of **cross-model&dataset** attack: trained on surrogate model (DukeMTMC) and tested on victim models (Market).

Methods	Global-based			Part-based		Attention-based			aAP↓	mDR↑
	BOT	LSRO	MuDeep	Aligned	MGN	HACNN	Transreid	PAT		
None	85.4	77.2	49.9	79.1	82.1	75.2	86.6	78.4	76.7	-
MetaAttack	26.3	68.6	37.8	59.4	73.0	63.9	80.0	67.7	59.6	22.3
Mis-Ranking	46.3	36.7	11.9	47.5	46.7	27.0	65.2	63.4	43.1	43.8
MUAP	42.9	35.7	9.7	48.0	40.6	23.8	58.3	59.7	39.8	48.1
GAP	46.1	53.9	19.2	57.7	60.6	41.8	66.5	67.1	51.6	32.7
CDA	46.8	55.9	20.3	58.5	62.3	46.5	69.0	70.1	53.7	30.0
BIA	49.9	60.3	33.9	61.9	69.8	59.0	78.5	66.1	59.9	21.9
LTP	45.3	61.3	32.7	60.7	67.1	52.6	69.8	68.7	57.3	25.3
PDCL-Attack	28.7	36.0	14.4	40.8	49.7	28.1	61.4	50.8	38.7	49.5
AP-Attack(Ours)	**22.0**	**11.1**	**6.1**	**24.1**	**22.3**	**10.3**	**38.0**	**35.4**	**21.2**	**72.4**

tively. Comparing to SOTA re-id attack method MetaAttack that incorporates multi-datasets in meta-learning scheme, our method get comparable performances with only one dataset for training.

Comparisons on cross-model attack. Experimental results in Tab. 2 show that our method achieves the best performances of 6.9% aAP and 89.4% mDR score on cross-model scenarios, significantly outperforming the SOTA methods by 10.6% and 16.3% in terms of aAP and mDR.

Comparisons on cross-model&dataset attack. For the majority of realistic and complex cross-model&dataset attack results in Tab. 3, our AP-Attack method exceeds the SOTA method PDCL-Attack by 17.5% on aAP accuracy and 22.9% on mDR, which further highlights the superiority and effectiveness of our method.

Notably, our method outperforms SOTA prompt-driven attack method PDCL-Attack in all attack settings. The advantage of our method on re-id can be attributed to two main factors. First, our method achieves fine-grained, thorough feature disruption, while PDCL-Attack lacks of comprehensive destroy by perturbing only on global features. Second, the learned prompts in PDCL-Attack are specifically designed for different IDs within the same class 'person' in re-id attack, which differs from prompt learning in classification task where prompts describe different categories, likely leads to prompts that are more stylized rather than universal. In contrast, our approach utilizes the predefined prompt template, guiding the inversion network to convert more specific and generalizable semantic information, resulting in more broadly applicable and transferable results.

4.3. Ablation Studies

The effectiveness of textual inversion. In order to verify the effectiveness of our textual inversion networks, we attempt to interpret the learned pseudo-tokens to meaningful word. We first established an attribute-specific semantic vocabulary by chatGPT, with each attribute corresponding to a distinct, meaningful set of semantic words. These attribute

Details of produced semantic vocabulary shows in supplemental files.

Figure 3. Word cloud visualization of the learned attribute-specific pseudo-tokens, where semantic words for benign images are in black boxes and AEs' are in red boxes. Distinct colors of the word represent different attributes and attributes highlighted in red illustrate the disrupted semantics of AEs. Font size indicates the similarity to tokens, with larger fonts represent greater similarity.

vocabularies capture both color and descriptive features. Then, we calculate the similarity between each word in the vocabulary and the pseudo-tokens, selecting the two words with the highest similarity scores for display, as shown in Fig. 3. From the figure, it can be seen that the semantic words similar to the learned pseudo-tokens can correctly describe the pedestrian image information, indicating the effectiveness of our inversion networks.

The effectiveness of fine-grained attack. To illustrate the superiority of our AP-Attack for guiding fine-grained semantic attack, we compare the results of cross-model&dataset attack using different adversarial triplet losses that incorporate various features. Specifically, we use the loss constraint based solely on the image features from the surrogate model as the baseline, and compare with the results obtained by adding different CLIP feature losses, including global visual features, integral text embeddings, and fine-grained semantic embeddings. As shown in Tab. 5, the results incorporating CLIP feature constraints significantly outperform the baseline. Moreover, the inclusion of text

Table 4. Attack effectiveness against defense methods.

Method	Adv.ResNet	Randomization	JPEG(60%)	aAP ↓	mDR ↑
None	69.6	84.6	83.8	80.0	-
MetaAttack	67.1	67.8	57.9	64.3	19.7
Mis-Ranking	56.1	43.3	51.2	50.2	37.3
MUAP	53.6	48.5	57.4	53.2	33.5
Ours	39.0	26.6	27.8	**31.1**	**61.1**

Table 5. Results of cross-model&dataset attack using different adversarial triplet losses that incorporate various features.

	aAP↓	mDR↑
baseline	45.8	40.3
baseline+global visual feature	27.9	63.6
baseline+integral text embedding	25.9	66.2
baseline+fine-grained semantic embedding	**21.2**	**72.4**

embeddings yields better results than using image features, suggesting that text-based features offer greater universality. Most importantly, our fine-grained semantic embedding constraints achieve the best performance, demonstrating the effectiveness of our method.

To visually demonstrate that our method performs fine-grained attacks on each attribute, we compared the perturbation images under different feature constraints. As shown in Fig. 4, compared to the incorporating global CLIP feature constraints, the perturbations generated by our method cover a larger area, closely resembling the full pedestrian posture in the image. This indicates that our method produces perturbations that attempt to disrupt all semantic features of the pedestrian in a fine-grained manner. Meanwhile, as can be seen from the attribute word cloud of AEs in Fig. 3, our approach is destructive to fine-grained semantics, and can destroy most of the attribute semantics.

4.4. Attack Effectiveness against Defense Method

We conduct evaluations against three defense strategies, including adversarially trained models (Adv. Res [6]), input preprocessing techniques (JPEG compression [7]), and denoising-based methods (Randomization [50]). For JPEG, a compression rate of 60% is applied, and the victim model is BOT(Market). Tab. 4 shows that our method consistently achieves superior attack effectiveness across these defenses, achieves an mDR of 61.1%.

4.5. Transferability to Diverse Types Models

To further evaluate our AP-Attack's generalizability across diverse types of re-id models, we test it against three distinct re-id models: the self-supervised PASS (Market) [70], auxiliary-feature-enhanced PGFA (Occluded-Duke) [31], and CLIP-ReID (Market) [22] based on CLIP. Tab. 6 reveals substantial performance degradation across all tested

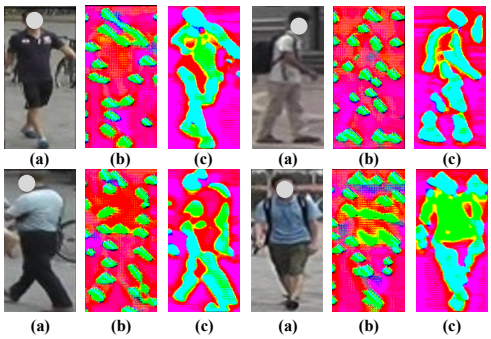

Figure 4. Visualization of perturbations under different feature constraints: (a) shows the original image, (b) depicts the perturbation when incorporating global image features from the CLIP model, and (c) presents the perturbation under our AP-attack with fine-grained semantic feature constraints.

Table 6. Comparisons on self-supervised, auxiliary feature and CLIP-based re-id models.

Method	PASS		PGFA		CLIP-ReID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
None	92.2	96.3	37.3	51.4	89.6	95.5
Ours	12.1	14.3	4.2	5.7	32.7	40.8

architectures, confirming the efficacy of our method across diverse model paradigms.

5. Conclusion

In this paper, we propose a novel Attribute-aware Prompt Attack methods to enhance the transferability of adversarial attacks on person re-id task. Our AP-Attack method leverages the image-text alignment capability of VLM and introduces the attribute-specific inversion networks to map the image feature to attribute semantic textual embeddings. And it attempts to thoroughly destroy the pedestrian features by perturbing fine-grained attribute semantics across all attribute feature spaces. Extensive experimental results validate the superiority of our approach in cross-dataset and cross-model black-box attack scenarios, achieving substantial performance gains over the latest SOTA methods. We believe that our work offers a meaningful contribution to adversarial attack research and holds promise for strengthening the security of machine learning systems in real-world.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62425305, U22B2050 and 62221002, in part by the Science and Technology Innovation Program of Hunan Province under Grant 2023RC1048, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2024JJ3013.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3908–3916, 2015. 1

[2] Abhishek Aich, Calvin-Khang Ta, Akash Gupta, Chengyu Song, Srikanth Krishnamurthy, Salman Asif, and Amit Roy-Chowdhury. Gama: Generative adversarial multi-object scene attacks. *Adv. Neural Inform. Process. Syst.*, 35:36914–36930, 2022. 2, 3

[3] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip HS Torr. Adversarial metric attack and defense for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43 (6):2119–2126, 2020. 2

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Int. Conf. Comput. Vis.*, pages 15338–15347, 2023. 4

[5] Yuan Bian, Min Liu, Xueping Wang, Yi Tang, and Yaonan Wang. Occlusion-aware feature recover model for occluded person re-identification. *IEEE Trans. Multimedia*, 26:5284–5295, 2024. 1

[6] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 794–795, 2020. 2, 8

[7] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017. 8

[8] Wenjie Ding, Xing Wei, Rongrong Ji, Xiaopeng Hong, Qi Tian, and Yihong Gong. Beyond universal person re-identification attack. 16:3442–3455, 2021. 2, 3, 6

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9185–9193, 2018. 2

[10] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4312–4321, 2019. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2020. 5

[12] Hao Fang, Jiawei Kong, Bin Chen, Tao Dai, Hao Wu, and Shu-Tao Xia. Clip-guided networks for transferable targeted attacks. *arXiv preprint arXiv:2407.10179*, 2024. 2, 3

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Int. Conf. Learn. Represent.*, 2023. 2, 4

[14] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Eur. Conf. Comput. Vis.*, pages 307–322. Springer, 2020. 2

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 5

[17] 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. 1, 5

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 5

[19] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Int. Conf. Comput. Vis.*, pages 4733–4742, 2019. 2

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, pages 4904–4916, 2021. 2, 3

[21] DP Kingma. Adam: a method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2014. 6

[22] 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. 4, 5, 6, 8

[23] 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. 6

[24] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2285–2294, 2018. 5, 6

[25] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *AAAI*, pages 11458–11465, 2020. 2

[26] Zihan Li, Weibin Wu, Yuxin Su, Zibin Zheng, and Michael R Lyu. Cdta: a cross-domain transfer-based attack with contrastive learning. In *AAAI*, pages 1530–1538, 2023. 2

[27] Min Liu, Yuan Bian, Qing Liu, Xueping Wang, and Yaonan Wang. Weakly supervised tracklet association learning with

video labels for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5):3595–3607, 2024. 1

[28] Min Liu, Fei Wang, Xueping Wang, Yaonan Wang, and Amit K. Roy-Chowdhury. A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(7):4944–4956, 2024. 1

[29] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Int. Conf. Learn. Represent.*, 2016. 2

[30] 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. 5

[31] 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. 8

[32] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1765–1773, 2017. 1

[34] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 2, 6

[35] Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Part-aware transformer for generalizable person re-identification. In *Int. Conf. Comput. Vis.*, pages 11280–11289, 2023. 5

[36] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4422–4431, 2018. 6

[37] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Int. Conf. Comput. Vis.*, pages 5399–5408, 2017. 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 2, 3

[39] 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. 6

[40] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Adv. Neural Inform. Process. Syst.*, 34:13950–13962, 2021. 2, 6

[41] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18603–18613, 2022. 2

[42] AV Subramanyam. Meta generative attack on person reidentification. *IEEE Trans. Circuit Syst. Video Technol.*, 33(8):4429–4434, 2023. 2

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Int. Conf. Learn. Represent.*, 2014. 1

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2818–2826, 2016. 5

[45] 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. 5, 6

[46] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep misranking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 342–351, 2020. 2, 3, 6

[47] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh 0001. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, page 4, 2018. 1

[48] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Int. Conf. Comput. Vis.*, pages 7639–7648, 2021. 2

[49] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 6

[50] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *Int. Conf. Learn. Represent.*, 2018. 8

[51] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2730–2739, 2019. 2

[52] Fengxiang Yang, Zhun Zhong, Hong Liu, Zheng Wang, Zhiming Luo, Shaozi Li, Nicu Sebe, and Shin'ichi Satoh. Learning to attack real-world models for person re-identification via virtual-guided meta-learning. In *AAAI*, pages 3128–3135, 2021. 2

[53] Fengxiang Yang, Juanjuan Weng, Zhun Zhong, Hong Liu, Zheng Wang, Zhiming Luo, Donglin Cao, Shaozi Li, Shin'ichi Satoh, and Nicu Sebe. Towards robust person re-identification by defending against universal attackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):5218–5235, 2022. 2, 3, 6

[54] 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, 3, 6

[55] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial at-

tacks. In *Eur. Conf. Comput. Vis.*, pages 36–53. Springer, 2025. 6

[56] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19163–19173, 2022. 2, 3

[57] Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17343–17353, 2024. 4, 5

[58] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *Int. Conf. Learn. Represent.*, 2022. 2, 3

[59] Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Mutual-modality adversarial attack with semantic perturbation. In *AAAI*, pages 6657–6665, 2024. 2, 3

[60] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):2872–2893, 2021. 1

[61] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *ACM Int. Conf. Multimedia*, pages 5005–5013, 2022. 2

[62] 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.*, 2021. 2

[63] 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. 6

[64] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 5

[65] 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. 6

[66] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1, 5, 6

[67] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Int. Conf. Comput. Vis.*, pages 3754–3762, 2017. 5

[68] Zhedong Zheng, Liang Zheng, Yi Yang, and Fei Wu. U-turn: Crafting adversarial queries with opposite-direction features. *Int. J. Comput. Vis.*, 131(4):835–854, 2023. 2

[69] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 2

[70] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *Eur. Conf. Comput. Vis.*, pages 198–214, 2022. 8