Look Closer to Your Enemy: Learning to Attack via Teacher-Student Mimicking

Mingjie Wang, Jianxiong Guo, *Member, IEEE*, Sirui Li, Dingwen Xiao, and Zhiqing Tang, *Member, IEEE*

***Abstract*—Deep neural networks have significantly advanced person re-identification (ReID) applications in the realm of the industrial internet, yet they remain vulnerable. Thus, it is crucial to study the robustness of ReID systems, as there are risks of adversaries using these vulnerabilities to compromise industrial surveillance systems. Current adversarial methods focus on generating attack samples using misclassification feedback from victim models (VMs), neglecting VM's cognitive processes. We seek to address this by producing authentic ReID attack instances through VM cognition decryption. This approach boasts advantages like better transferability to open-set ReID tests, easier VM misdirection, and enhanced creation of realistic and undetectable assault images. However, the task of deciphering the cognitive mechanism in VM is widely considered to be a formidable challenge. In this paper, we propose a novel inconspicuous and controllable ReID attack baseline, LCYE (*Look Closer to Your Enemy*), to generate adversarial query images. Specifically, LCYE first distills VM's knowledge via teacher-student memory mimicking the proxy task. This knowledge prior serves as an unambiguous cryptographic token, encapsulating elements deemed indispensable and plausible by the VM, with the intent of facilitating precise adversarial misdirection. Further, benefiting from the multiple opposing task framework of LCYE, we investigate the interpretability and generalization of ReID models from the view of the adversarial attack, including cross-domain adaption, cross-model consensus, and online learning process. Extensive experiments on four ReID benchmarks show that our method outperforms other state-of-the-art attackers with a large margin in white-box, black-box, and target attacks. The source code can be found at https://github.com/MingjieWang0606/LCYE-attack_reid.**

***Index Terms*—Adversarial Attack, GAN, ReID, Memory Module, Image Classification.**

I. Introduction

PERSON re-identification (ReID) [1]–[3] aims to associate images of a person across disjoint cameras. In recent years, ReID methods predicated on deep learning [2] have not only achieved accuracy rates exceeding 90% but also transcended human-level competence. Nonetheless, it has been discovered that deep ReID models display a high susceptibility to adversarial instances, implying that even slight alterations to the initial instances could lead the model to erroneous conclusions with a high degree of certainty [4]. Within the context of practical applications, such a susceptibility could be exploited either malevolently (for example, enabling criminals to evade ReID-based surveillance systems) or benevolently (for instance, permitting individuals with sensitive identities to remain concealed within a database).

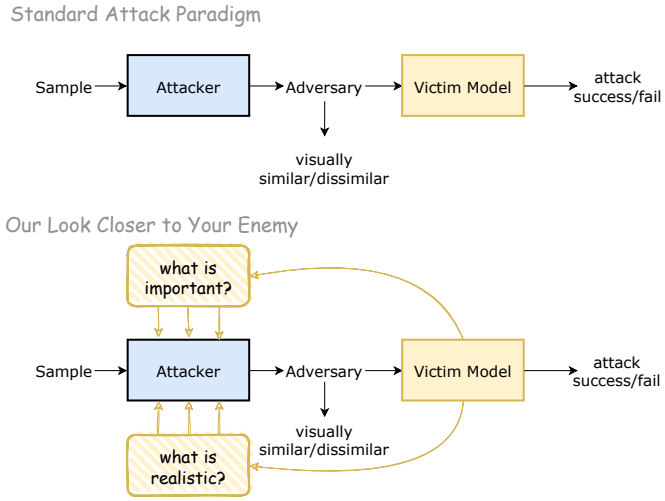


Fig. 1. Illustration of standard attack paradigm and the idea of our *Look Closer to Your Enemy*.

Current research endeavors attacking image retrieval models, as reflected in studies [4], [5], predominantly adopt paradigms borrowed from classification attacks. The standard attack paradigm as shown in Figure 1, under the supervision for attack and imperceptible purposes, the attacker gradually optimizes toward a stable status between the attack performance and imperceptibility. However, we believe this paradigm is suboptimal. Firstly, it lacks pixel-level guidance since attack supervision gives global feedback for attack and imperception. Secondly, there exists a discrepancy in the training and attacking mechanisms for victim models (VMs). Specifically, ReID VMs are trained using varied structures and losses, such as PCB [6] and circle loss [7], yet they are attacked using a similar global mis-ranking loss [4]. Lastly, the adversarial identity consistency for personalized and realistic generation remains unverified, both in RGB and latent space, particularly in Generative Adversarial Network (GAN)-based attackers [4]. In such attackers, the discriminator distinguishes between real and fake, without indicating an association with the intended identity.

To handle the problems above, we propose to manipulate

Mingjie Wang, Sirui Li, and Dingwen Xiao are with the Guangdong Key Lab of AI and Multi-Modal Data Processing, Department of Computer Science, BNU-HKBU United International College, Zhuhai 519087, China. (e-mail: mjwang0606@gmail.com; sil089@ucsd.edu; dxiaoaf@connect.ust.hk)

Jianxiong Guo is with the Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai 519087, China, and also with the Guangdong Key Lab of AI and Multi-Modal Data Processing, BNU-HKBU United International College, Zhuhai 519087, China. (e-mail: jianxiongguo@bnu.edu.cn)

Zhiqing Tang is with the Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai 519087, China. (e-mail: zhiqingtang@bnu.edu.cn)

(Corresponding author: Jianxiong Guo; Zhiqing Tang.)

Manuscript received April xxxx; revised August xxxx.

the perceptual framework of the VM to mount the attack. Specifically, the attacker may eliminate regions that the VM identifies as reflective of the genuine identity, while concurrently introducing elements associated with a different identity. A salient example of this can be seen in the attribute-level descriptions provided by datasets such as Market1501 [8] and DukeMTMC [9], which might state *"a young man wears a yellow shirt, black pants, and a hat, and carries no bag"*. These attributes provide concrete instantiations of an individual. We can execute an attack that alters specific attributes, such as replacing a yellow shirt with green short sleeves. Simultaneously, it is imperative to uphold the logical coherence of the attack, for instance, by avoiding contradictory modifications like forcing an individual to appear as if they are wearing both trousers and shorts concurrently. This particular form of attack possesses three inherent advantages:

- The attacker could be transferable to a cross-domain test set due to the transferability of well-posed VM knowledge.
- It provides pixel-level guidance taking into account both realistic and harmful generation for the attack.
- By manipulating the VM's perceptual framework, it mitigates the need for subverting the VM during training to garner adequate negative feedback.

Driven by this insight, we propose not to seek an instance-level accurate image, *i.e.,* coercing one image to morph into another, but rather to search for a generalized identity-level structural prototype. In this paper, we introduce a novel, subtle, and controllable attack baseline harnessing the identity-level understanding of VMs. As depicted in Figure 2, we first capture the particular attribute of the VM using our proposed attention-like memory module via a teacher-student mimicry approach. Subsequently, we enable the attacker to retrieve the potential target prototype from the memory module. In specific terms, the knowledge of what the VM interprets as an identity is learned online and stored within a parameterized memory module. Both the generator and the discriminator then retrieve relevant prototypes from this memory, but with differing objectives. The generator obtains the prototype of potential targets for the attack, while the identity-wise discriminator verifies whether the generated image maintains the prototype of the targeted identity. We integrate the pixel-level attack and perceptibility measurement within a preservation-consistent Generative Adversarial Network framework.

An additional contribution of this paper is the exploration of the interpretability and generalization capacities of Recognition using IDentity (ReID) models from an adversarial attack perspective, encompassing cross-domain adaptation, cross-model consensus, and online learning processes. The discrete teacher and student models within the mimicking branch facilitate the analysis of explainable ReID across various hybrid combinations. In summary, our contributions can be outlined as follows:

- We propose a novel ReID attack baseline named *Look Closer to Your Enemy* (LCYE) to solve the issues from joint attack and imperceptibility optimization.
- We delve deep into interpretable and generalizable ReID from the aspect of adversarial attack by evaluating model consensus, covering cross-model and cross-domain covariance, and the online learning process.
- Our method obtains a promising attack success rate with inconspicuous noise. Experimental validations on four of the most extensive ReID benchmarks with both CNN and Transformer-based ReID SOTAs [2], [6] validate our method's superior efficiency and transferability in white-box, black-box, and target attacks.

Organization: In Section II, we discuss the related work of Person Reid and Adversarial Attack. In Section III, we establish the basic settings of our attack algorithm, introduce our novel ReID attack baseline, LCYE, and its method of generating adversarial query images, and discuss the advantages of our approach and the challenges in deciphering VM's cognitive mechanism. Section V details the experiments carried out on four ReID benchmarks, comparing our method against other SODA attackers. Lastly, Section VI wraps up the study and mentions potential directions for future research.

II. Related Work

In this section, we discuss the related work of the realm of person re-identification (ReID) and Adversarial Attacks. ReID refers to the task of matching images of the same person across different camera views or time instances, typically used in surveillance and security applications. With the rise of deep learning techniques, ReID systems have shown significant improvements. However, they remain vulnerable to Adversarial Attacks. An Adversarial Attack introduces small, often imperceptible perturbations to the input data, designed to mislead a trained model into making a false classification or decision. By integrating these attacks into ReID, it becomes crucial to understand their implications and devise potential defense mechanisms to ensure the robustness of ReID systems.

Person Re-identification aims to spot the appearance of the same person in different observations [7]. Deep feature-based methods [10], [11] and metric learning-based methods [12], [13] have achieved significant progress in supervised ReID. For deep feature-based methods, a cascade structure has already been studied in the neural network literature in the 1980s [14]. Although fully connected cascade networks trained with batch gradient descent [15] are effective on small datasets. This method only applies to networks with a few hundred parameters. In [16], it has been found to be effective for various vision tasks to utilize multi-level features in CNNs through skip-connections. Then, the pure theoretical framework of networks with cross-layer connections is derived [17]. Highway Networks was the first batch of architectures that provided effective training methods for more than 100 layers of end-to-end networks [18].

Additionally, ResNets have achieved impressive performance like ImageNet and COCO object detection [19]. ResNets with pre-activation can also help train state-of-the-art networks with more than 1000 layers [20]. For metric learning-based methods, TriNet [21] samples the most negative samples in the batch to achieve rapid convergence. The negative samples were found by Harwood et al. [22] from the increasing

TABLE I
KEY TERMINOLOGIES AND CONCEPTS FOR LCYE

Symbol	Statement
x	original image
x'	adversarial counterpart of the original image
N	the identity number of the training set
H, W	height and width of the image
$f_i \in f$	the feature from the input
$k_j \in K$	the slices from prototype the matrix
w_{ij}	the normalized weight
$\mathcal{I}$	the clean image
$\hat{\mathcal{I}}$	the adversary of the clean image
C_n	the number of samples drawn from the n-th person ID
$\mathcal{I}_c^n$	the c-th images of the n ID in a mini-batch
c_s	the samples from the same ID
c_d	the samples from the different ID
$\mathcal{L}_*$	the loss of $*$
λ, α_*, β_*	the tradeoff factors
$\mathcal{G}$	the generator of the attacker
$\mathcal{P}$	memory module
$\mathcal{M}$	the misclassification model
$\mathcal{H}$	the pixel-level perception loss
$\mathcal{A}$	the global-level attack loss
∂L	the joint gradient
$\mathcal{M}'$	the subnet of VM
n	noise
$\mathcal{O}$	mask predictor
m	mask
h_{attack}	the attack cues
$\mathcal{S}$	log-softmax function
$\mathbb{H}$	indicator function

search space defined by the nearest neighbor distance. For example, TransReID [2] achieves state-of-the-art performance, with the participation of a local-aware Transformer [23]. Moreover, other innovative works [24] already focus on the interpretable and generalizable ReID which is verified by improved results. However, these methods may not be suitable and scalable for different tasks and models. In this paper, we aim to jointly analyze different ReID models in the point of supervised and unsupervised adversarial attack, keeping the fairness and flexibility of assessing the robustness and generalization. This attempt is not limited to cross-validation via simple white-box, black-box attacks but covers implicit decision-making.

Adversarial Attack is to extract samples from real data to fool the learning model and help evaluate the robustness of the target models [25]–[27]. The security problems of the current most advanced model [28], [29] and more insights into the CNN mechanism [30] was raised by Szegedy et al. The fast-gradient sign method [30], which generates adversarial examples in one step, is one of the earliest works of gradient-based attack. The primary iterative method [31], deep fool [32] and iterative momentum method [11] extend the fast-gradient sign method [30] to update the adversarial images with small step sizes iteratively. Score-based attacks rely on searching input space. Single-pixel perturbation out of the valid image range can successfully lead to misclassification on small-scale images [33], which can be extended to large-scale images by local greedy searching. In addition to pixel modification, the

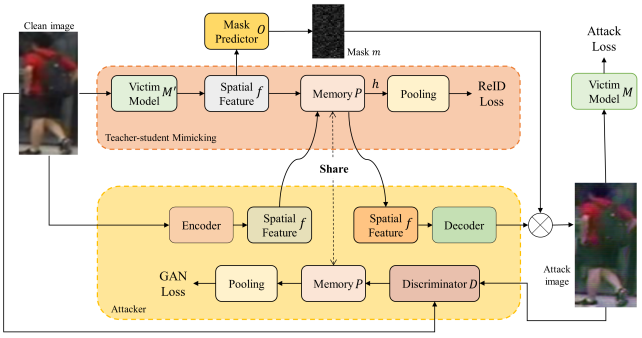


Fig. 2. The unrolled framework of our *Look Closer to Your Enemy* (LCYE). The generator of the attacker has an encoder and a decoder which retrieves relevant prototypes from the memory module at the middle of the encoder-decoder.

adversarial examples can also result in sample generation via spatial transform [34]. The iterative least-likely class method [31] increases the prediction probability of the least possible class through constraints, so the classification model outputs interesting errors.

Adversarial Attack on ReID needs to generalize to unseen query images using misranking loss, which differs from adopting misclassification loss in close-set recognition tasks [35]. For optimization-based methods, [36] proposes an ODFA that exploits feature-level adversarial gradients. [5] proposes a metric attack to distort the distance between the attacked image and other similar images. Besides, [4] introduces a GAN-based misranking attacker whose transferability is proved by its post-hoc results. However, these methods still ignore pixel-level supervision considering both attack and perception, which is essential to tackle their tradeoff. In this paper, we propose our LCYE to solve this issue by cheating VM's mind to generate harmful but realistic images.

III. METHODOLOGY

In this section, we elucidate the challenges and methods associated with adversarial attacks. After detailing the mechanisms behind clean image misclassification, we discuss the constraints of close-set recognition tasks and the intricacies of GAN-based attackers. Notably, the conventional paradigm of adversarial attacks is contrasted with the adversarial reinforcement approach from VM. To address these complexities, we introduce a novel method named LCYE, designed to deceive VM's sample-driven decision-making process. The frequently used notations are summarized in Table I.

A. Overall Framework

The unrolled framework of our method is illustrated in Figure 2. We facilitate the learning of adversarial the attacker by jointly solving two opposing tasks: (1) memorizing the underlying recognition cues for ReID of each identity via teacher-student memory mimicking and (2) interpolating this prior knowledge of VM to the attacker to guide the efficient adversarial generation.

The objective of an adversarial attack, given a clean image and its ground truth label $(\boldsymbol{x}, y)$, is to lead the model $\mathcal{M}$ into making a misclassification on the input $\boldsymbol{x}$. The adversary $\boldsymbol{x}'$ is determined based on:

$$\min_{\boldsymbol{x}'} \mathcal{H}\left(\boldsymbol{x}, \boldsymbol{x}'\right) \text{and} \min_{\boldsymbol{x}'} \mathcal{A}(\mathcal{M}\left(\boldsymbol{x}'\right), y), \tag{1}$$

where $\mathcal{H}$ represents the pixel-level perception loss and $\mathcal{A}$ signifies the global-level attack loss. Specifically, Pixel-level perception loss, denoted by $\mathcal{H}$, measures the perceptual difference between the original image $\boldsymbol{x}$ and its adversarial counterpart $\boldsymbol{x}'$. The main purpose of this loss is to ensure that the adversarial image remains visually similar to the original image. Mathematically, this can be represented as:

$$\mathcal{H}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \frac{1}{H \times W} \sum\nolimits_{i=1}^{H} \sum\nolimits_{j=1}^{W} \|x_{i,j} - x'_{i,j}\|^2, \tag{2}$$

where H and W are the height and width of the image, respectively, and $\|\cdot\|$ denotes the L2 norm. By minimizing this loss, the perturbations added to the original image are kept subtle and almost imperceptible to human eyes. On the other hand, the global-level attack loss, denoted by $\mathcal{A}$, focuses on the task of misleading the model $\mathcal{M}$ into making incorrect predictions. Specifically, the loss ensures that the adversarial example $\boldsymbol{x}'$ is classified differently than the original label y of image $\boldsymbol{x}$. Formally, it can be expressed as:

$$\mathcal{A}(\mathcal{M}\left(\boldsymbol{x}'\right), y) = -\log(\mathbb{P}_{\mathcal{M}}(y|\boldsymbol{x}')), \tag{3}$$

where $\mathbb{P}_{\mathcal{M}}(y|\boldsymbol{x}')$ indicates the model's predicted probability of the adversarial image $\boldsymbol{x}'$ belonging to the correct class y. By maximizing this loss, we can ensure that the adversarial image is classified incorrectly by the model.

In essence, the balance between these two losses is crucial. While $\mathcal{H}$ ensures that the adversarial image remains visually close to the original, $\mathcal{A}$ guarantees that the model is effectively fooled by the perturbed image. Therefore, to avoid optimization collapse in solving such multiple opposing tasks, we deliberately separate mimicking and attack branches as much as possible. In particular, the clean images $\boldsymbol{x}$ are sent to the mimicking branch composed by the subnet $\mathcal{M}'$ of VM, memory module $\mathcal{P}$, to learn identity-wise structural prototypes, which will be described in Section III-B.

Afterward, the generator $\mathcal{G}$ of the attacker, composed of an encoder and a decoder, retrieves a similar but different identity prototype to generate, which determines to mislead the VM in prior. Meanwhile, according to the searchable realistic memory, the discriminator $\mathcal{D}$ distinguishes the adversarial generation belonging to the identity it claimed, following consistent identity preservation. This parts will be described in Section III-C.

B. Teacher-student Memory Mimicking

Inspired by [1], we insert an external-attention memory module $\mathcal{P}$ into the VM to remember such prototypes in a learnable tensor cache. Specifically, we keep and freeze the subnet $\mathcal{M}'$ of VM before pooling, local split, or equivalents to obtain informative spatial features where we can access the sampled-driven decision-making process. Then, for the following evaluation fairness, we apply the identical memory module $\mathcal{P}$, MaxPooling, Batch Normalization (BN), and classification head sequentially after $\mathcal{M}'$ for different VMs. During fine-tuning for ReID, $\mathcal{P}$ could dynamically record what VM recognizes images for each identity and what are real images embedded in latent space, *i.e.,* identity-wise structural prototype.

The prototype memory module $\mathcal{P}$ contains N learnable identity prototypes, which are recorded by a matrix $\boldsymbol{K} \in \mathbb{R}^{N \times C}$ with fixed feature dimension C. Identity prototype number N equals the identity number of the training set where each identity contains one prototype in memory. Then an attention-based addressing operator for accessing the memory, *i.e.*, memory reader and writer, is used to assign each image into spare prototypes. The role of the memory reader and writer depends on whether $\boldsymbol{K}$ updates in the current step. Given the output $\boldsymbol{f} \in \mathbb{R}^{H \times W \times C}$ of $\mathcal{M}'$ in which H and W denote the height and width resolution, our memory module could be denoted as:

$$w_{ij} = \frac{\exp\left(d\left(\boldsymbol{f}_i, \boldsymbol{k}_j\right)\right)}{\sum_{j=1}^{N} \exp\left(d\left(\boldsymbol{f}_i, \boldsymbol{k}_j\right)\right)} \tag{4}$$

where $\boldsymbol{f}_i \in \mathbb{R}^C$ and $\boldsymbol{k}_j \in \mathbb{R}^C$ are feature and prototype slices from input $\boldsymbol{f}$ and prototype matrix K. w_{ij} is the normalized weight measuring the cosine similarity $d(\cdot, \cdot)$ between $\boldsymbol{f}_i$ and $\boldsymbol{k}_j$. Thus, the assigned prototype $\boldsymbol{h} \in \mathbb{R}^{H \times W \times C}$ from feature $\boldsymbol{f}$ could be calculated by using the Memory Regulation and Alignment (MRA) module proposed in [1]:

$$\boldsymbol{h} = MRA(\boldsymbol{f}, K) = \oplus_{i=1}^{H \times W} \sum\nolimits_{j=1}^{N} w_{ij} \boldsymbol{k}_j, \tag{5}$$

where the $\oplus$ indicates that there are $H \times W$ vectors concatenating together. Thus, the dimension of $\boldsymbol{h}$ is $\mathbb{R}^{H \times W \times C}$, which is aligned with $\boldsymbol{x}$.

C. Memory-guided Adversarial Attacker

The motivation of our LCYE is to build a prototype-consistent generation for an adversarial attack: The generator accords to one potential identity prototype to generate the adversary, where the selected prototype claims the identity belonging to the adversary. Then the discriminator could also tell whether the adversary belongs to the claimed identity by checking the prototype. Thus, our memory-guided attacker obtains three main components: adversarial generator $\mathcal{G}$ to generate noise $\boldsymbol{n}$, identity-wise multi-stage discriminator $\mathcal{D}$ to ensure identity preservation consistency, and a mask predictor $\mathcal{O}$ to estimate the effective noise mask $\boldsymbol{m}$. The final attack image is composed of $\boldsymbol{x}' = \boldsymbol{m} \odot \boldsymbol{n} + \boldsymbol{x}$. Besides, VM's knowledge learned and recorded in $\boldsymbol{K}$, is interpolated into the generator and discriminator to facilitate the inconspicuous adversarial generation. Meanwhile, the mask predictor selects the location of recognition cues from VM for the accurate hit. Given the real image $\boldsymbol{x}$, our generator first encode it to $\boldsymbol{f}$ and let it retrieve the target prototype from $\mathcal{P}$ as attack cues $\boldsymbol{h}_{attack}$ to generate noise $\boldsymbol{n}$:

$$\boldsymbol{h}_{attack} = MRA(\boldsymbol{f}, \boldsymbol{K}), \tag{6}$$

$$n = \mathcal{G}(x, h_{attack}), \tag{7}$$

Hence, mask predictor $\mathcal{O}$ maps the location of discriminative region to image resolution as mask $\boldsymbol{m}$ according to the clues of spatial features $\boldsymbol{f}$ from VM:

$$\boldsymbol{f} = \mathcal{M}'(\boldsymbol{x}), \boldsymbol{m} = \mathcal{O}(\boldsymbol{f}). \tag{8}$$

Finally, the multi-stage identity-aware discriminator classifies whether attack images $\boldsymbol{x}'$ and original images $\boldsymbol{x}$ meet the prototype belonging to the claimed identity. In particular, three subnetworks, receiving $\{1, 1/4, 1/16\}$ areas of the original images as the input, are introduced in $\mathcal{D}$ to obtain a multi-scale response. Then, by pyramiding the features of different discriminator levels as [4], a series of downsampled results with a ratio of $\{1/32, 1/16, 1/8, 1/4\}$ of the image is thus formulated for final prediction. We empirically let features $s_{1/4}$ with a 1/4 resolution ratio to retrieve the relevant prototype from $\boldsymbol{K}$ to check whether the image meets the imagination of VM. Solely interpolating VM knowledge to the discriminator brings implicit external-internal semantic consistency but still lacks explicit identity consistency for personalized supervision. Therefore, $\mathcal{D}$ is designed to estimate multi-identity probability $\boldsymbol{p} \in \mathbb{R}^{N+1}$, where N is the total number of identities in the training set, and the additional dimension denotes the fake class. Thus, we have

$$\boldsymbol{p} = \mathcal{D}(\boldsymbol{x}, \boldsymbol{h}_d), \boldsymbol{h}_d = MRA(s_{1/4}, \boldsymbol{K}). \tag{9}$$

Note that this prototype interpolation in the generator and discriminator seems to have no detailed map with one-hot ground truth since the learnable prototype matrix $\boldsymbol{K}$ records the general representation of each identity but does not know the correspondence. It means LCYE lacks explicit and strict prototype supervision for each identity and we can only hope the model learns such correspondence. However, this risk is mitigated by the adversarial generative mechanism and static memory reading. The competitive generator and discriminator can not lasso the memory module since they can only read it. In particular, when the memory is meaningful for proxy recognition, the focus of competition thus moves to how to use the memory for better generation, rather than ignoring it.

D. Objectives

The mis-ranking loss, *i.e.,* attack loss, includes four variants: (1) standard misclassification loss **cent**; (2) misclassification proposed by [4] **xent**; (3) adversarial triplet loss **etri**; and (4) misclassification and adversarial triplet losses **xcent+etri**. We omit the description of standard misclassification loss **cent** since it is a simple adversarial cross-entropy loss. Given a clean image $\mathcal{I}$ and its adversary $\hat{\mathcal{I}}$, the misclassification loss **xent** proposed by [4] is: $\mathcal{L}_{xent} =$

$$-\sum_{n=1}^{N} \mathcal{S}(\mathcal{T}(\hat{\mathcal{I}}))_n \left((1-\delta)\mathbb{H}_{argmin\ \mathcal{T}(\mathcal{I})_n} + \delta v_n\right), \tag{10}$$

where $\mathcal{S}$ is the log-softmax function, N is the total identity number and $v = \left[\frac{1}{N-1}, \cdots, 0, \cdots, \frac{1}{N-1}\right]$ is smoothing regularization in which v_k equals to $\frac{1}{N-1}$ everywhere except when n is the ground-truth ID. $\mathbb{H}$ is the indicator function. The adversarial triplet loss **etri** is:

$$\mathcal{L}_{etri} = \sum_{n=1}^{N} \sum_{c=1}^{C_n} \Bigg[\max_{j \neq n; c_d = 1 \ldots C_j} \left\| \mathcal{T}\left(\hat{\mathcal{I}}_c^n\right) - \mathcal{T}\left(\hat{\mathcal{I}}_{c_d}^j\right) \right\|_2^2 - \min_{c_s = 1 \ldots C_n} \left\| \mathcal{T}\left(\hat{\mathcal{I}}_c^n\right) - \mathcal{T}\left(\hat{\mathcal{I}}_{c_s}^n\right) \right\|_2^2 + \Delta \Bigg]_+, \tag{11}$$

where C_n is the number of samples drawn from the n-th person ID, $\mathcal{I}_c^n$ is the c-th images of the n ID in a mini-batch, c_s and c_d are the samples from the same ID and the different IDs, $||\cdot||_2$ is the square of L2 norm used as the distance metric, and Δ is a margin threshold.

For visual perception loss, we borrow two choices of **SSIM** and **MS-SSIM** from [4]. The difference between them is interpolating multi-scale (**MS**) measurement. Therefore, we mainly represent **MS-SSIM** here: $\mathcal{L}_{MS-SSIM}(\mathcal{I}, \hat{\mathcal{I}}) =$

$$\left[l_L(\mathcal{I}, \hat{\mathcal{I}})\right]^{\alpha_L} \cdot \prod_{j=1}^{L} \left[c_j(\mathcal{I}, \hat{\mathcal{I}})\right]^{\beta_j} \left[s_j(\mathcal{I}, \hat{\mathcal{I}})\right]^{\gamma_j}, \tag{12}$$

where c_j and s_j are the measures of the contrast comparison and the structure comparison at the j-th scale respectively, which are calculated by $c_j(\mathcal{I}, \hat{\mathcal{I}}) = \frac{2\sigma_{\mathcal{I}}\sigma_{\hat{\mathcal{I}}} + C_2}{\sigma_{\mathcal{I}}^2 + \sigma_{\hat{\mathcal{I}}}^2 + C_2}$ and $s_j(\mathcal{I}, \hat{\mathcal{I}}) = \frac{\sigma_{\mathcal{I}\hat{\mathcal{I}}} + C_3}{\sigma_{\mathcal{I}}\sigma_{\hat{\mathcal{I}}} + C_3}$ where σ is the variance/covariance.

For GAN loss, we adopt the multi-discriminator and multi-label GAN loss as:

$$\mathcal{L}_{GAN} = \mathbb{E}_{(I_{cd}, I_{cs})}\left[\log \mathcal{D}_{1,2,3}\left(I_{cd}, I_{cs}\right)\right] + \mathbb{E}_{\mathcal{I}}\left[\log\left(1 - \mathcal{D}_{1,2,3}(\mathcal{I}, \hat{\mathcal{I}})\right)\right], \tag{13}$$

where the subscript $1, 2, 3$ denotes our multi-stage discriminator. The identity-aware supervision could be deemed as the multi-label binary version of standard real/fake loss.

E. Objective Function

The total objective includes (1) mis-ranking loss $\mathcal{L}_{mr}$ for attack; (2) GAN loss $\mathcal{L}_{GAN}$ for personalized realistic generation; (3) visual perception loss $\mathcal{L}_{VP}$ for inconspicuous change; and (4) cross-entropy loss $\mathcal{L}_{ce}$ and triplet loss $\mathcal{L}_{tri}$ for teacher-student mimicking:

$$\mathcal{L}_{attack} = \alpha_1 \cdot \mathcal{L}_{mr} + \alpha_2 \cdot \mathcal{L}_{GAN} + \alpha_3 \cdot \mathcal{L}_{VP}, \tag{14}$$

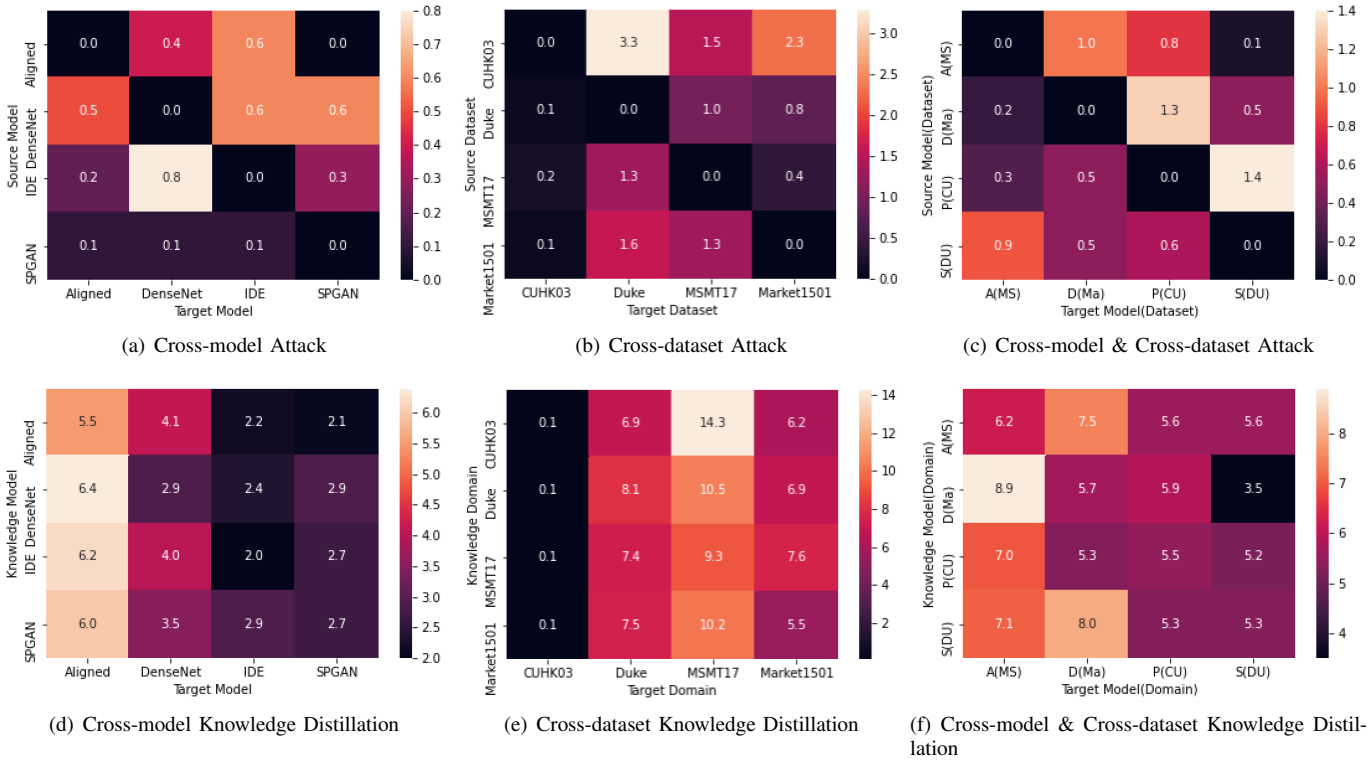
$$\mathcal{L}_{mimic} = \beta_1 \cdot \mathcal{L}_{ce} + \beta_2 \cdot \mathcal{L}_{tri}, \tag{15}$$

where α_* and β_* are tradeoff factors. $\mathcal{L}_{attack}$ is the loss for the attack branch, while $\mathcal{L}_{mimic}$ is for mimicking the branch. Since the objectives are already proposed or largely identical as Mis-ranking [4] that we modify from, we prefer to only discuss the capability of inserting VM's knowledge in the attacker, which eases the dependence on both mis-ranking loss $\mathcal{L}_{mr}$ and visual perception loss $\mathcal{L}_{VP}$ in Section V-C. Empirically, we set $\alpha_1 = 1, \alpha_2 = 1,\ \alpha_3 = 1,\ \beta_1 = 1$, and $\beta_2 = 1$.

IV. How Memory Help Attack?

To comprehend the underlying mechanics, we ought to revisit the foundational framework delineated in Sec. III-A. Systematically, this can be partitioned into three components:

Firstly, the marriage of pixel-level perception loss and global-level attack loss appears incomplete, given that it does



(a) Cross-model Attack (b) Cross-dataset Attack (c) Cross-model & Cross-dataset Attack

(d) Cross-model Knowledge Distillation (e) Cross-dataset Knowledge Distillation (f) Cross-model & Cross-dataset Knowledge Distillation

Fig. 3. Rank1 results (%) of *black-box attack* and cross-model/domain knowledge distillation. In cross-modal&dataset(domain) evaluation, we abbreviate AlignedReID(MSMT17)→ A(MS), DenseNet(Market1501)→D(Ma), PCB(CUHK03)→P(CU), SPGAN(DukeMSMT)→S(DU).

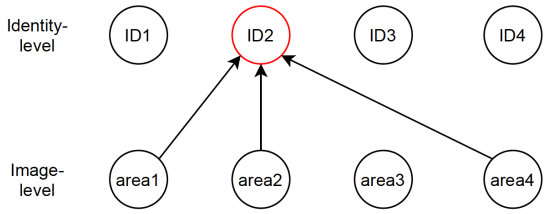
not offer pixel-level supervision that takes into account their combined objectives. The perception loss $\mathcal{H}$ and adversarial loss $\mathcal{A}$ operate in separate domains, probing the potential adversary $\boldsymbol{x}'$ in both RGB and latent spaces. The composite gradient $\partial L = \frac{\partial \mathcal{H}}{\partial \boldsymbol{x}} + \lambda \cdot \frac{\partial \mathcal{A}}{\partial \boldsymbol{x}}$ is guided by the predetermined trade-off factor λ, which tends to confine the optimization process to a localized, biased optimum.

Secondly, in the realm of close-set recognition tasks—where training and test datasets have identical categories—Virtual Models (VM) $\mathcal{M}$s are educated using a cross-entropy loss and subsequently assaulted with an adversarial variant of the same. This seems intuitive. Nonetheless, when it comes to the open-set ReID models $\mathcal{M}$, their training procedures exhibit a wide variation in terms of structure and loss functions. As an illustration, while PCB aims to mine local representation, $\mathcal{M}$ finds itself under the assault of an adversarial triplet loss, which targets global features. This approach seems to neglect the unique attributes of individual models. Although tailoring attack supervision for each model could be a potential remedy, its implementation remains challenging.

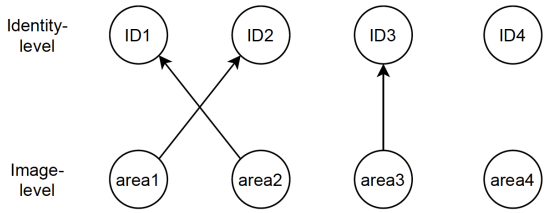
Thirdly, the commitment to adversarial category consistency within GAN-based attackers [4], [37] appears somewhat diluted. These methodologies typically employ a fixed VM as an outboard category discriminator while concurrently training an online real/fake discriminator to satiate the need for authentic generation. However, such a static external VM can be effortlessly deceived by the generator. Concurrently, the rudimentary real/fake discriminator struggles to discern genuinely individualized images from merely authentic ones. A deeper dive into this concern can be found in Section V-B, wherein we highlight that the images attacked via the GAN-based approach tend to manifest more apparent noise.

At its core, we posit that the prevailing adversarial attack framework resembles an adversarial reinforcement exercise steered by VM. The challenges we've pinpointed stem from a tendency to overlook learning from samples and understanding the VM's perspective. To address these intertwined issues, this paper introduces a groundbreaking method, dubbed LCYE, designed to deceive the sample-driven decision-making machinery of the VM. Specifically,

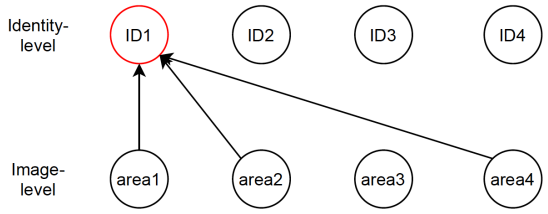
1) The pixel-level adversarial direction is derived from the collaboration of the mask predictor and the memory reading component within the generator. These mechanisms not only determine the spatial regions to target but also decipher the identity transformation in the latent space. In contrast, existing techniques [4] either fall short in terms of precise control or predominantly rely on misclassification-based adversarial feedback.
2) Addressing the unique characteristics of the victim model, our LCYE introduces a universal teacher-student mimicry approach. Instead of seeking an adversarial analogue, such as the hard adversarial triplet loss or the simple triplet loss, LCYE preserves the original architecture and protocols of the victim model, while embracing a multi-faceted adversarial strategy.
3) Through prototype interpolation combined with identity-aware generation—termed as category cycle consistency—we ensure that every modified region directly



(a) Clean images



(b) Adversary of current methods with standard paradigm



(c) Adversary of our LCYE

Fig. 4. Illustration of decision making of victim model using clean images, adversarial images generated by current methods, and our LCYE. We omit the insignificant links from area to identity (ID).

contributes to the recognition of the desired identity. Figure 4 illustrates this distinction: whereas conventional paradigms distribute the significance of different regions across multiple identities, our approach coherently aligns these influential areas to a singular, erroneous higher-level concept by consistently associating the same identity prototype with the image.

Concluding, based on the aforementioned tripartite decomposition, we can ascertain that the conventional methodologies embody certain limitations and biases when juxtaposed against the demands of real-world adversarial tasks. Our LCYE, conceived in response to these shortcomings, not only provides a holistic approach but also ensures precise and controllable adversarial perturbations. By profoundly understanding the VM's standpoint and meticulously addressing each of its intricacies, LCYE offers a robust mechanism to effectively operate under diverse adversarial conditions.

V. Experiments

In this section, we compare our proposed CLYE with other commonly used baselines.

A. Datasets and Experimental Setup

Datasets. Four ReID benchmarks are used for the evaluation of our method, including Market1501 [8] (1,501 identities with 32,688 images), CUHK03 [38] (1,467 identities and 28,192 images), DukeMTMC [9] (1,404 identities with 34,183 images), and MSMT17 [39] (4,101 identities and 126,441 images). For attack evaluation metrics, we use Rank1, 5, and 10, and mAP for ReID attack where the lower numerical value means a better success attack rate in an attack problem.

Implementation Details. We use batch size 32, learning rate 0.0002 for GAN, and 0.0003 for mimicking on a single GTX P40 GPU. We use **xent+etri** for misranking loss $\mathcal{L}_{mr}$ and **MS-SSIM** for visual perception loss $\mathcal{L}_{VP}$. The triplet margin is set to 0.3 for mimicking and attacking branches. Compared with other attackers, we use full-size images as possible masks to attack. Our generator uses ResNet Block with $4\times$ downsampling and $4\times$ upsampling. The sub-discriminator adopts the basic structure of Mis-ranking. For target attacks, since ReID is an open-set task where training and test sets have non-overlapping identities, it is unfeasible to follow a close-set target attack. We separate it into two evaluations: target consistency on adversarial query images and standard attack. For each identity $n \in N$, we randomly select γ query images for each identity to generate adversarial query images with a total number $\gamma \cdot N$. Then we sent these images to the victim model and calculated the Euclidean distance of their embeddings. With pseudo labels, we calculate the Rank1 accuracy as their target consistency.

Implementation Details and Protocols. Our victim models include CNN-based methods (AlignedReID [40], DenseNet [10], etc [41]–[43]), GAN-based methods (CamStyle [44], SPGAN [45], LSRO [46], HHL [47]) and Transformer-based methods (TransReID [2]). The basic framework of our LCYE largely benefited from Mis-ranking [4], including the hyper-parameters, basic model structure etc. Therefore, we compare it in ablation to verify the capability brought from VM knowledge. For a fair comparison, we adopt the same protocols as [4] by L_∞-bounded attacks with $\varepsilon = 16$. The *black-box attack* includes a cross-model attack, cross-dataset attack, and cross-model-dataset attack as the standard setting. For the *target attack*, we achieve it by modifying Eqn. (5) to $\oplus_{i=1}^{H\times W}\sum_{j=1}^{N} w_{ij}\boldsymbol{q}_j\boldsymbol{k}_j$ where $\boldsymbol{q}_j \in Q \in \mathbb{R}^{N\times 1}$ is the indicator of selected identity and adding corresponding identity consistency supervision in GAN. Due to the open-set essence of the ReID task, we evaluate the target attack performance by (1) the similarity matrix of embeddings of adversarial query images which is simplified as Rank1↑ as target consistency and (2) the attack success rate, *i.e.,* Rank1↓ and mAP↓.

For the ablation study, we clarify that (1) *baseline mimicking* is to let $\mathcal{M}'$ fixed and learn the memory; (2) *online mimicking* is to allow $\mathcal{M}'$ pre-trained on ImageNet to update with mimicking branch (learn from scratch as a ReID task); and (3) *offline mimicking* is to first train to mimic branch with fixed $\mathcal{M}'$ and then train the pending attack branch. If without specific instruction, all ablation experiments are done on AlignedReID with Market1501.

B. Attacking State-of-the-Art ReID Models

White-box Attack. As shown in Table II, Table III and Table IV, we demonstrate the superior performance of our method against other attackers. Results on multiple datasets and models show that our LCYE essentially improves the

TABLE II
ATTACKING THE STATE-OF-THE-ART REID SYSTEMS ON MARKET1501. BLUE AND RED DENOTE PREVIOUS BEST AND CURRENT BEST RESULTS. ↓ MEANS THE LOWER NUMERICAL VALUE IS BETTER FOR ATTACK.

Methods with Market1501		Rank1↓				Rank5↓				Rank10↓				mAP↓			
		Before	PGD	MR	Ours	Before	PGD	MR	Ours	Before	PGD	MR	Ours	Before	PGD	MR	Ours
Backbone	IDE(ResNet50)	83.1	4.5	3.7	0.3	91.7	8.7	8.3	1.2	94.6	12.1	11.5	2.4	63.3	4.6	4.4	0.3
	DenseNet121	89.9	1.2	1.2	0	96.0	1.0	1.3	0.2	97.3	1.5	2.1	0.4	73.7	1.3	1.3	0.2
	Mudeep(Inceptionv3)	73.0	2.6	1.7	0.0	90.1	5.5	1.7	0.2	93.1	6.9	5.0	0.5	49.9	2.0	1.8	0.2
Part-Aligned	AlignedReID	91.8	10.2	1.4	0.1	97.0	15.8	3.7	0.9	98.1	19.1	5.4	1.8	79.1	8.9	2.3	0.3
	PCB	88.6	6.1	5.0	0.0	95.5	12.7	10.7	0.1	97.3	15.8	14.3	0.2	70.7	4.8	4.3	0.2
	HACNN	90.6	6.1	0.9	2.8	95.9	8.8	1.4	8.1	97.4	10.6	2.3	12.9	75.3	5.3	1.5	1.2
GAN	CamStyle+Era(IDE)	86.6	15.4	3.9	0.1	95.0	23.9	7.5	0.7	96.6	29.1	10.0	1.4	70.8	12.6	4.2	0.2
	LSRO(DenseNet121)	89.9	7.2	0.9	0.8	96.1	13.1	2.2	2.2	97.4	15.2	3.1	3.5	77.2	8.1	1.3	1.7
	HHL(IDE)	82.3	5.7	3.6	0.1	92.6	9.8	7.3	0.7	95.4	12.2	9.7	1.4	64.3	5.5	4.1	0.2
	SPGAN(IDE)	84.3	10.1	1.5	0.0	94.1	16.7	3.1	0.6	96.4	20.9	4.3	1.6	66.6	8.6	1.6	0.2
Transformer	TransReID(ViT+baseline)	94.6	-	6.2	0.9	98.2	-	10.0	1.5	99.2	-	12.1	2.7	87.1	-	6.3	0.8
	TransReID(ViT)	95.1	-	5.2	0.9	98.4	-	10.1	1.7	99.1	-	12.0	2.6	89.0	-	6.5	0.9

TABLE III
ATTACKING THE STATE-OF-THE-ART REID SYSTEMS ON CUHK03.

Methods with CUHK03		Rank1↓				Rank5↓				Rank10↓				mAP↓			
		Before	PGD	MS	Ours	Before	PGD	MS	Ours	Before	PGD	MS	Ours	Before	PGD	MS	Ours
Backbone	IDE(ResNet50)	24.9	0.8	0.4	0.0	43.3	1.2	0.7	0.4	51.8	2.1	1.5	0.4	24.5	0.8	0.9	0.2
	DenseNet121	48.4	0.1	0.0	0.0	50.1	0.1	0.2	0.6	70.1	0.3	0.6	1.2	84.0	0.2	0.3	0.4
	Mudeep(Inceptionv3)	32.1	0.4	0.1	0.0	53.3	1.0	0.5	0.2	64.1	1.5	0.8	0.4	30.1	0.8	0.3	0.1
Part-Aligned	AlignedReID	61.5	1.4	1.4	0.0	79.4	2.2	3.7	0.6	85.5	4.1	5.4	1.1	59.6	2.1	2.1	0.3
	PCB	50.6	0.5	0.2	0.0	71.4	2.1	1.3	0.2	78.7	4.5	1.8	0.8	48.6	1.2	0.8	0.3
	HACNN	48.0	0.4	0.1	0.0	69.0	0.9	0.3	0.4	78.1	1.3	0.4	1.1	47.6	0.8	0.4	0.3

TABLE IV
ATTACKING THE STATE-OF-THE-ART REID SYSTEMS ON DUKEMTMC.

Methods with DukeMTMC		Rank1↓				Rank5↓				Rank10↓				mAP↓			
		Before	PGD	MR	Ours	Before	PGD	MR	Ours	Before	PGD	MR	Ours	Before	PGD	MR	Ours
GAN-based	CamStyle+Era(IDE)	76.5	22.9	1.2	0.6	86.8	34.1	2.6	1.5	90.0	39.9	3.4	2.6	58.1	16.8	1.5	0.3
	LSRO(DenseNet121)	72.0	7.2	0.7	0.5	85.7	12.5	1.6	1.4	89.5	18.4	2.2	2.2	55.2	8.1	0.9	0.8
	HHL(IDE)	71.4	9.5	1.0	0.1	83.5	15.6	2.0	0.8	87.7	19.0	2.5	1.7	51.8	7.4	1.3	0.2
	SPGAN(IDE)	73.6	12.4	0.1	0.4	85.2	21.1	0.5	1.2	88.9	26.3	0.6	2.5	54.6	10.2	0.3	0.3

attack success scope by 1% for Rank1, 3% for Rank5, 4% for Rank10, and 2% for mAP, respectively, in most cases. Moreover, our method performs favorably against Mis-ranking (MR) and PGD [48] with a large margin on different lines of ReID models.

Black-box Attack. Figure 3 shows the cross-model & cross-dataset & cross-model-dataset evaluation. Benefiting from cheating VM's knowledge, our method achieves a similar attack performance as a white-box attack. We find that training on domain adaption SPGAN seems more versatile than others in the cross-model case, and results on CUHK03 in the cross-dataset evaluation show the domain-specific vulnerability.

Target Attack. To the best of our knowledge, our LCYE is the first method of target attack on ReID. In Table V, our method could successfully transfer the prototype of the desired identity to random images with over 60% consistency accuracy and high attack performance, except on TransReID. One possible reason is that, compared to CNN, Transformers always focus on a larger area of recognition cues, making personalized target attacks harder. Moreover, compared with *non-target attack*, *i.e.,* white-box attack, the performance only drops 2%, 4.1%, 1.2%, and 1.2% for four models, respectively. This indicates that our LCYE is not sensitive to personalized assignments.

TABLE V
Target attack RESULTS (%) ON MARKET1501.

Method	Target Consistency	Attack	
	Rank1 ↑	Rank1 ↓	mAP ↓
DenseNet121	71.2	2.0	1.3
AlignedReID	78.6	4.2	1.7
SPGAN	63.6	1.2	1.4
TransReID	45.0	2.1	2.0

C. Ablation Study

Cross-model & Cross-domain Knowledge Distillation. For cross-domain adaption and cross-model consensus, we use different $\mathcal{M}'$ to attack $\mathcal{M}$ with attack loss and find their attack performance is unexpectedly good without obvious distinction

(a) Original images (b) SSIM, Mis-ranking (c) SSIM, LCYE

(d) Original images (Same as (a)) (e) MS-SSIM, Mis-ranking (f) MS-SSIM, LCYE

Fig. 5. Visualization of using different visual perception losses for Mis-ranking and our method.

TABLE VI
ABLATION ON '*does the memory learn the belief of victim model?*' IN ONLINE MIMICKING AND OFFLINE MIMICKING WITH THREE MIMICKING VARIANTS. REID RESULTS ARE EVALUATED BY MIMICKING THE BRANCH.

Mimicking	ReID		Attack	
	Rank1↑	mAP↑	Rank1↓	mAP ↓
baseline	61.3	54.2	5.5	2.1
online	85.2	74.3	5.6	2.1
offline	61.3	54.2	5.5	2.1

with Table II. We believe that strong attack supervision may aggressively direct the whole optimization. Thus, we choose to remove the mis-ranking loss $\mathcal{L}_{mr}$ to check the model consensus for better interpretability. This experiment is similar to a black-box attack in meaning but free from attack guidance. As shown in Figure 3 (d)-(e), the diversity of knowledge consensus from models is more significant than from domains. Expect CUHK03, which also achieves the lowest value in cross-dataset attack; all domains' result seems uniform. This finding is also consistent with the cross-model dataset, indicating domain distribution makes fewer senses than a model structure for knowledge commonsense of robustness. As shown in Table VI, our LCYE achieves a similar performance as Table 1 in the original paper. For cross-domain knowledge distillation, it shows a similarly good performance. Thus, we believe the attack supervision may make whole optimization aggressive without showing the knowledge property of target model $\mathcal{M}$ and knowledge model $\mathcal{M}'$.

Sensitivity of Visual Perception Loss. As shown in Figure 5, we provide more visualization to analyze the insensitivity of our method to different objectives. We can find that our method poses less dark green or purple background on original images and generates clearer images without heavy blur. We contribute this benefit to the interpolation of VM knowledge to both generator and discriminator since it also conveys the configuration of realistic images.

Online Mimicking and Offline Mimicking. One common concern is: Does the memory learn the belief of the victim model? One explanation is the cross-model/domain knowledge distillation and target attack, which show the specific knowledge of each model and the possibility of attack. Besides, to determine the influence of mimicking the manner in the interaction with the attacker, we conduct experiments with three variants, i.e., *baseline mimicking, online mimicking, offline mimicking*, without attack loss. As shown in Table VI, ReID results evaluate the leftover property of the mimicking

(a) Original images

(b) Original images (Same as (a))

(c) Mis-ranking, $\varepsilon = 3$

(d) LCYE, $\varepsilon = 3$

(e) Mis-ranking, $\varepsilon = 10$

(f) LCYE, $\varepsilon = 10$

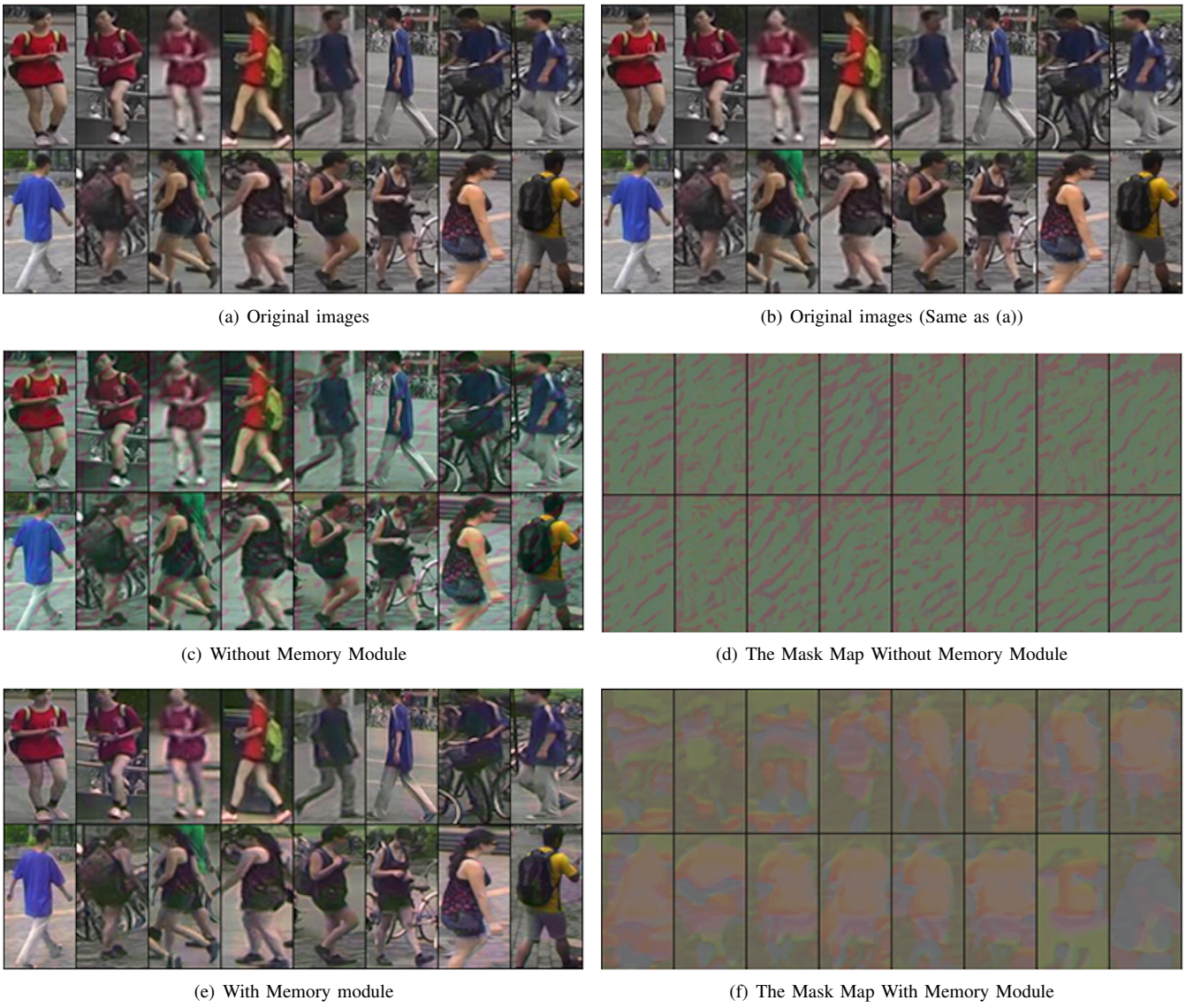
(g) Mis-ranking, $\varepsilon = 16$

(h) LCYE, $\varepsilon = 16$

Fig. 6. Visualization of using different visual perception losses for Mis-ranking and our method.

branch after our simple mimicking manner. The identical results of *baseline* and *offline* mean the memory module is not affected by the attacked branch during training and the quality of memory does not influence the performance of the attacker (joint training and two-stage training are different from memory retrieved by the attacker in each iteration). Furthermore, the obtained memory of *online* is different from that of the victim model since the model structure is partly different. But the performance does not drop for this knowledge difference. We owe this phenomenon to the effectiveness of our LCYE paradigm, which is not sensitive to knowledge.

Number of Pixels to be Attacked. We further ablate the



(a) Original images (b) Original images (Same as (a))

(c) Without Memory Module (d) The Mask Map Without Memory Module

(e) With Memory module (f) The Mask Map With Memory Module

Fig. 7. Visualization of our LCYE with/without memory module. And LCYE can clearly capture the contours of the human body from the mask map.

TABLE VII
PROPORTION OF ADVERSARIAL POINTS († IS THE RESULTS WITH APPROPRIATE RELAXATION AND RATIO IS THE ADVERSARIAL POINTS/TOTAL POINTS).

Ratio	Mis-ranking		ours	
	Rank1	mAP	Rank1	mAP
full size	1.4	2.3	0.1	0.3
1/2	39.3	31.5	11.1	3.6
1/4	72.7	85.9	11.1	3.5
1/8	91.8	79.1	11.1	3.5
1/16	91.8	79.1	11.0	3.5
1/16†	8.2	14.7	1.3	0.6
1/32†	59.4	47.3	1.3	0.7
1/64†	75.5	61.5	1.6	1.0

pixel demand to attack in Table VII. Our LCYE achieves promising performance even with a small ratio. Note that it keeps 11.1% Rank1 and 3.6% mAP from 1/2 to 1/16. We believe LCYE needs much fewer pixels than other attackers. The relaxation only brings about less than 10% improvement, compared to Mis-ranking, which heavily depends on it. This benefit may come from the accurate hit of our mask predictor, and salient region believed by the victim model.

Comparisons of Different ε. Larger ε could effectively boost attack performance but sacrifice the visual quality. We manually control the magnitude of ε to verify the effectiveness of our LCYE. As shown in Table VIII, smaller ε would not limit our attack performance. It meets the lower bound ($\backsim 0.4\%$) of Rank1 at $\varepsilon = 10$ or even much earlier. Especially when using $\varepsilon = 10$ or 5, our LCYE also shows promising superiority over Mis-ranking with 20%-60% gains. We further provide a visualization comparison in Figure 6. With small magnitudes of ε, Mis-ranking not only has strong blue atmospheres and obvious color blocks over original images but also poses striped Gaussian blur. However, our method generates a

TABLE VIII
ABLATION ON DIFFERENT ε. THE RESULTS IS REPORTED ON ALIGNEDREID WITH MARKET1501.

	Mis-ranking				ours			
	R1	R5	R10	mAP	R1	R5	R10	mAP
40	0.0	0.2	0.6	0.2	0.0	0.1	0.1	0.0
20	0.1	0.4	0.8	0.4	0.0	0.1	0.1	0.0
16	1.4	3.7	5.4	2.3	0.1	0.9	1.8	0.3
10	24.4	38.5	46.6	21.0	0.4	1.6	3.2	0.3
5	69.2	82.6	87.0	56.4	4.2	10.2	15.1	1.6
3	83.9	92.5	95.1	70.2	8.1	18.1	23.8	2.7

TABLE IX
ABLATION ON DIFFERENT VALUES OF DIMENSION C. THE RESULTS ARE REPORTED ON ALIGNEDREID WITH MARKET1501.

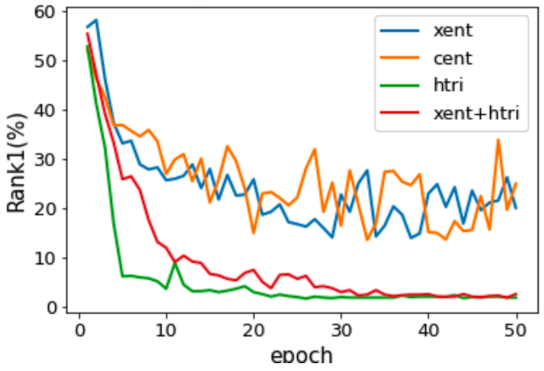
C	1	1/2	1/4	2	4
Rank1	1.2	0.9	1.2	1.1	1.6
mAP	1.3	0.9	1.6	1.4	1.7

much more reasonable adversary compared to Mis-ranking.

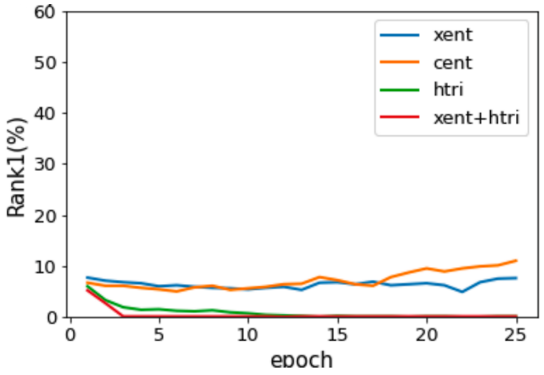
Cycle Consistency for Generation. Our LCYE keeps cycle consistency via identity-aware adversarial learning and interpolating VM knowledge to generator and discriminator, which are termed as explicit and implicit guidance. Figure 7 shows the noise and adversary visualization where the cycle consistency facilitates the inconspicuous perception.

***Discussion*.** The transferability of ReID models are always evaluated by applying it to another dataset where the uninspiring results usually show the poor generalization ability of models. We instead analyze the commonsense of the vulnerability of different models/domains and find that even without attack guidance, domain-specific and model-specific decision-making is fragile, with a minor distribution shift coming from the knowledge belonging to others or even themselves. It means the improvement brought from either data augmentation (GAN) or local awareness (Transformer) can not facilitate the robustness across models and domains in robustness. We reconsider the capability of our LCYE, including lower pixel demand and lower objective dependency, from the aspect of inverse reinforcement learning. Namely, the attacker estimates a possible distribution of adversaries by the attack award from victim models. Our LCYE benefits from (1) a more direct way to observe and memorize the environment (VM) and (2) providing dense awards for each pixel. In particular, the memory module is a shortcut interface accessing the decision-making process of the victim model rather than indirectly predicting it by attack award. As shown in Figure 4, we illustrate the decision-making from the image level in neurons. Our method redirects the map from each salient locality to the same wrong higher semantics since it assigns the same identity prototype to the image, while current methods always scatter them without pixel-level supervision.

Sensitivity of Different Losses. Mis-ranking heavily relies on attack loss $\mathcal{L}_{mr} \in \{$**cent, xent, etri, xcent+etri**$\}$ and $\mathcal{L}_{VP} \in \{$**SSIM, MS-SSIM**$\}$, since it needs enough adversarial feedback via the overturning victim model. In Figure 8, our LCYE has sharper and more stable curves



(a) Mis-ranking with different attack losses



(b) LCYE with different attack losses

Fig. 8. Ablation on the sensitivity of different losses. (a) and (b) record the test evaluation results after each epoch.

TABLE X
COMPLEXITY COMPARISON WITH BASELINE MODEL IN TRAINING AND TESTING.

	Traning		Testing	
Model	baseline	LCYE	baseline	LCYE
Paras	2.5616×10^7	3.0021×10^7	1.9508×10^7	2.4733×10^7
FLOPs	1.9899×10^9	3.6822×10^9	1.3614×10^9	2.8385×10^9

than Mis-ranking using different attack losses, indicating that memory reduces the necessity of well-designed objectives. This observation is also consistent with Figure 3, where our LCYE achieves promising attack performance without attack supervision. Moreover, we also ablate the dependence on visual perception loss. As shown in Figure 3, our method has a more inconspicuous visualization than the counterpart without dark purple or green background. Thus, exquisite losses are not a necessity of our LCYE.

Sensitivity of Different parameters of prototype matrix. We investigate the impact of prototype matrix scaling on attack effectiveness through comprehensive ablation studies. As Table IX reveals: (1) LCYE maintains stable performance across different scale factors ($0.25\times - 4\times$), with Rank-1 variance within $\pm 0.7\%$; (2) The $4\times$ scaling achieves optimal attack success (1.6% Rank-1), suggesting moderate dimension expansion enhances attack potency while preserving stealthiness. This dimensional robustness confirms our method's adaptability to different memory configurations.

Complexity Comparison. As shown in Table X, our LCYE could boost the performance by a large margin with minor complexity addition. The additional parameter complexity mainly comes from the prototype memory module in both training and testing. Besides, the FLOPs complexity is primarily increased in training since the victim model needs to process clean images for mimicking branches, while in testing, we directly use the memory module for attack without processing images in the victim model again.

VI. Conclusion and Future Work

In this paper, we propose the Look Closer to Your Enemy, LCYE, a novel method for attacking victim models by exploiting their decision-making processes through teacher-student mimicking. Unlike traditional attack strategies, LCYE enhances the accessibility of the victim model's internal workings, enabling a more nuanced and powerful adversarial strategy. Extensive experiments have demonstrated LCYE's strong performance in white-box, black-box, and targeted attack scenarios, making it a highly transferable attack method across various models and datasets. Besides, the computational cost of our method remains within acceptable limits for real-world applications, ensuring its feasibility for deployment in large-scale, production environments.

For future work, we plan to explore the integration of LCYE into existing person ReID systems, as well as other machine learning tasks. Although our current experiments primarily focus on controlled environments, understanding how LCYE can be effectively incorporated into real-world surveillance workflows is essential. This involves addressing challenges such as compatibility with existing ReID architectures and evaluating the impact of LCYE on system performance in production environments, where practical considerations such as computational overhead and response times become critical. Importantly, while this study focuses on attack methodology, we recognize the critical need to investigate defense mechanisms tailored to LCYE's unique attack paradigm. Future efforts will systematically evaluate existing defense strategies (e.g., input preprocessing, adversarial training) and develop task-specific countermeasures for retrieval-based adversarial scenarios, ultimately contributing to a more comprehensive security framework for industrial vision systems. Beyond ReID, we recognize the potential of LCYE in other settings, including action recognition and cross-domain image matching. Extending LCYE to these domains could offer valuable insights into its generalizability across different tasks. Lastly, we believe that the domain adaptation and domain generalization capabilities of LCYE warrant further investigation. Currently, these aspects are explored independently, but they could be integrated into a unified approach for more robust attacks. This nuanced, directional nature of the attack could lead to more sophisticated adversarial strategies that combine model interpretability with adversarial manipulation. We plan to explore these areas in future work, aiming to enhance the robustness of victim models against both general and domain-specific attacks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62202055, the Start-up Fund from Beijing Normal University under Grant No. 310432104/312200502510, the Internal Fund from BNU-HKBU United International College under Grant No. UICR0400003-24, the Project of Young Innovative Talents of Guangdong Education Department under Grant No. 2022KQNCX102, and the Interdisciplinary Intelligence Supercomputer Center of Beijing Normal University (Zhuhai).

References

[1] F. Chen, F. Wu, Q. Wu, and Z. Wan, "Memory regulation and alignment toward generalizer rgb-infrared person re-identification," *arXiv preprint arXiv:2109.08843*, 2021.

[2] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," *arXiv preprint arXiv:2102.04378*, 2021.

[3] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *AAAI*, vol. 34, no. 04, 2020, pp. 4610–4617.

[4] H. Wang, G. Wang, Y. Li, D. Zhang, and L. Lin, "Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking," in *IEEE CVPR*, 2020, pp. 342–351.

[5] S. Bai, Y. Li, Y. Zhou, Q. Li, and P. H. Torr, "Metric attack and defense for person re-identification," *arXiv e-prints*, pp. arXiv–1901, 2019.

[6] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018, pp. 480–496.

[7] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *IEEE CVPR*, 2020, pp. 6398–6407.

[8] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[9] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*. Springer, 2016, pp. 17–35.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 4700–4708.

[11] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *IEEE CVPR*, 2018, pp. 9185–9193.

[12] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 1318–1327.

[13] G. Chen, Y. Lu, J. Lu, and J. Zhou, "Deep credible metric learning for unsupervised domain adaptation person re-identification," in *ECCV*. Springer, 2020, pp. 643–659.

[14] S.E.Fahlman and C.Lebiere, "The cascade-correlation learning architecture," in *NeurIPS*, 1989, pp. 524–532.

[15] B. M. Wilamowski and H. Yu, "Neural network learning without backpropagation," *IEEE Transactions on Neural Networks*, vol. 21, no. 11, pp. 1793–1803, 2010.

[16] S. Yang and D. Ramanan, "Multi-scale recognition with dag-cnns," in *IEEE CVPR*, 2015, pp. 1215–1223.

[17] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang, "Adanet: Adaptive structural learning of artificial neural networks," in *ICML*. PMLR, 2017, pp. 874–883.

[18] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *NeurIPS*, vol. 28, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[20] ——, "Identity mappings in deep residual networks," in *ECCV*. Springer, 2016, pp. 630–645.

[21] B. Harwood, V. Kumar BG, G. Carneiro, I. Reid, and T. Drummond, "Smart mining for deep metric learning," in *IEEE CVPR*, 2017, pp. 2821–2829.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] S. Liao and L. Shao, "Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting," in *ECCV*, 2020, pp. 456–474.

[25] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1419–1428, 2018.

[26] K. Zhang, Y. Yao, R. Xie, X. Han, Z. Liu, F. Lin, L. Lin, and M. Sun, "Open hierarchical relation extraction," in *NAACL*, 2021, pp. 5682–5693.

[27] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su, "Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts," in *ICLR*, 2023.

[28] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *ACM CCS*, 2016, pp. 1528–1540.

[29] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *IEEE CVPR*, 2018, pp. 1625–1634.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[31] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[32] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *IEEE CVPR*, 2016, pp. 2574–2582.

[33] N. Narodytska and S. P. Kasiviswanathan, "Simple black-box adversarial perturbations for deep networks," *arXiv preprint arXiv:1612.06299*, 2016.

[34] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv preprint arXiv:1801.02612*, 2018.

[35] G. Zhang, H. Zhang, Y. Chen, and Y. Zheng, "Close-set camera style distribution alignment for single camera person re-identification," *Neurocomputing*, vol. 486, pp. 93–103, 2022.

[36] Z. Zheng, L. Zheng, Y. Yang, and F. Wu, "Query attack via opposite-direction feature: Towards robust image retrieval," *arXiv preprint arXiv:1809.02681*, 2018.

[37] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on gan," *arXiv preprint arXiv:1702.05983*, 2017.

[38] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE CVPR*, 2014, pp. 152–159.

[39] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE CVPR*, 2018, pp. 79–88.

[40] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[41] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[42] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *ICCV*, 2017, pp. 5399–5408.

[43] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *IEEE CVPR*, 2018, pp. 2285–2294.

[44] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE CVPR*, 2018, pp. 5157–5166.

[45] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE CVPR*, 2018, pp. 994–1003.

[46] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE CVPR*, 2017, pp. 3754–3762.

[47] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *ECCV*, 2018, pp. 172–188.

[48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

Mingjie Wang received his M.Phil. degree from the Department of Computer Science, BNU-HKBU United International College (UIC), in 2023, and his B.E. degree from the Department of Computer Science and Technology, Longdong University, China, in 2021. He is currently a research assistant at UIC, where his research interests focus on financial time series analysis and federated learning.

Jianxiong Guo (Member, IEEE) received his Ph.D. degree from the Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA, in 2021, and his B.E. degree from the School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou, China, in 2015. He is currently an Associate Professor with the Advanced Institute of Natural Sciences, Beijing Normal University, and also with the Guangdong Key Lab of AI and Multi-Modal Data Processing, BNU-HKBU United International College, Zhuhai, China. He is a member of IEEE/ACM/CCF. He has published more than 90 peer-reviewed papers and been the reviewer for many famous international journals/conferences. His research interests include social networks, wireless sensor networks, combinatorial optimization, and machine learning.

Sirui Li received the B.S. degree from the Department of Statistics and Data Science, BNU-HKBU United International College (UIC), in 2023 and is currently an M.Sc. candidate in Halıcıoğlu Data Science Institute, University of California San Diego. Her main interests are Data Mining, Image Analysis, and Interpretable Deep Learning.

Dingwen Xiao received the B.S. degree from the Department of Statistics and Data Science, BNU-HKBU United International College (UIC), in 2023 and is currently an M.Sc. candidate in Data Driven Modeling, Hong Kong University of Science and Technology. His main interests are Data Mining, Image Semantic Analysis, and Interpretable Deep Learning.

Zhiqing Tang (Member, IEEE) received the B.S. degree from School of Communication and Information Engineering, University of Electronic Science and Technology of China, China, in 2015 and the Ph.D. degree from Department of Computer Science and Engineering, Shanghai Jiao Tong University, China, in 2022. He is currently an assistant professor with the Advanced Institute of Natural Sciences, Beijing Normal University, China. His current research interests include edge computing, resource scheduling, and reinforcement learning.