Transferable, Controllable, and Inconspicuous Adversarial Attacks on Person Re-identification With Deep Mis-Ranking

Hongjun Wang[1]* Guangrun Wang[1]* Ya Li[2] Dongyu Zhang[2] Liang Lin[1,3]†
[1]Sun Yat-sen University [2]Guangzhou University [3]DarkMatter AI
[1]{wanghq8,wanggrun,zhangdy27}@mail2.sysu.edu.cn [2]liya@gzhu.edu.cn [3]linliang@ieee.org

Abstract

The success of DNNs has driven the extensive applications of person re-identification (ReID) into a new era. However, whether ReID inherits the vulnerability of DNNs remains unexplored. To examine the robustness of ReID systems is rather important because the insecurity of ReID systems may cause severe losses, e.g., the criminals may use the adversarial perturbations to cheat the CCTV systems.

In this work, we examine the insecurity of current best-performing ReID models by proposing a learning-to-mis-rank formulation to perturb the ranking of the system output. As the cross-dataset transferability is crucial in the ReID domain, we also perform a back-box attack by developing a novel multi-stage network architecture that pyramids the features of different levels to extract general and transferable features for the adversarial perturbations. Our method can control the number of malicious pixels by using differentiable multi-shot sampling. To guarantee the inconspicuousness of the attack, we also propose a new perception loss to achieve better visual quality.

Extensive experiments on four of the largest ReID benchmarks (i.e., Market1501 [45], CUHK03 [17], DukeMTMC [33], and MSMT17 [40]) not only show the effectiveness of our method, but also provides directions of the future improvement in the robustness of ReID systems. For example, the accuracy of one of the best-performing ReID systems drops sharply from 91.8% to 1.4% after being attacked by our method. Some attack results are shown in Fig. 1. The code is available at https://github.com/whj363636/Adversarial-attack-on-Person-ReID-With-Deep-Mis-Ranking.

1. Introduction

The success of deep neural networks (DNNs) has benefited a wide range of computer vision tasks, such as person re-identification (ReID), a crucial task aiming at matching pedestrians across cameras. In particular, DNNs have benefited ReID in learning discriminative features and adaptive distance metrics for visual matching, which drives ReID to a new era [36,44]. Thanks to DNNs, there have been extensive applications of ReID in video surveillance or criminal identification for public safety.

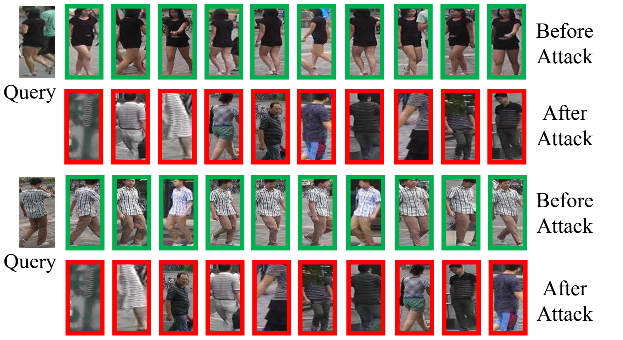


Figure 1. The rank-10 predictions of AlignedReID [36] (one of the state-of-the-art ReID models) before and after our attack on Market-1501. The green boxes represent the correctly matching images, while the red boxes represent the mismatching images.

Despite the impressive gain obtained from DNNs, whether ReID inherits the vulnerability of DNNs remains unexplored. Specifically, recent works found that DNNs are vulnerable to adversarial attacks [23,35] (An adversarial attack is to mislead a system with adversarial examples). In the past two years, the adversarial attack has achieved remarkable success in fooling DNN-based systems, e.g., image classification. Can the recent DNN-based ReID systems survive from an adversarial attack? The answer seems not promising. Empirically, evidence has shown that a person wearing bags, hats, or glasses can mislead a ReID system to output a wrong prediction [7,11,16,22,43]. These examples may be regarded as natural adversarial examples.

To examine the robustness of ReID systems against adversarial attacks is of significant importance. Because the insecurity of ReID systems may cause severe losses, for example, in criminal tracking, the criminal may disguise themselves by placing adversarial perturbations (e.g., bags, hats, and glasses) on the most appropriate position of the

*Equal contribution
†Corresponding author

body to cheat the video surveillance systems. By investigating the adversarial examples for the ReID systems, we can identify the vulnerability of these systems and help improve the robustness. For instance, we can identify which parts of a body are most vulnerable to the adversarial attack and require future ReID systems to pay attention to these parts. We can also improve ReID systems by using adversarial training in the future. In summary, developing adversarial attackers to attack ReID is desirable, although no work has been done before.

As the real-world person identities are endless, and the queried person usually does not belong to any category in the database, ReID is defined as a ranking problem rather than a classification problem. But existing attack methods for image classification, segmentation, detection, and face recognition do not fit a ranking problem. **Moreover**, since the image domains vary at different times and in different cameras, examining the robustness of ReID models by employing a cross-dataset black-box attack should also be taken into consideration. However, existing adversarial attack methods often have poor transferability, i.e., they are often designed for a sole domain of task (e.g., Dataset A) and can not be reused to another domain (e.g., Dataset B) due to their incapacity to find general representations for attacking. **Furthermore**, we focus on attacks that are inconspicuous to examine the insecurity of ReID models. Existing adversarial attack methods usually have a defective visual quality that can be perceived by humans.

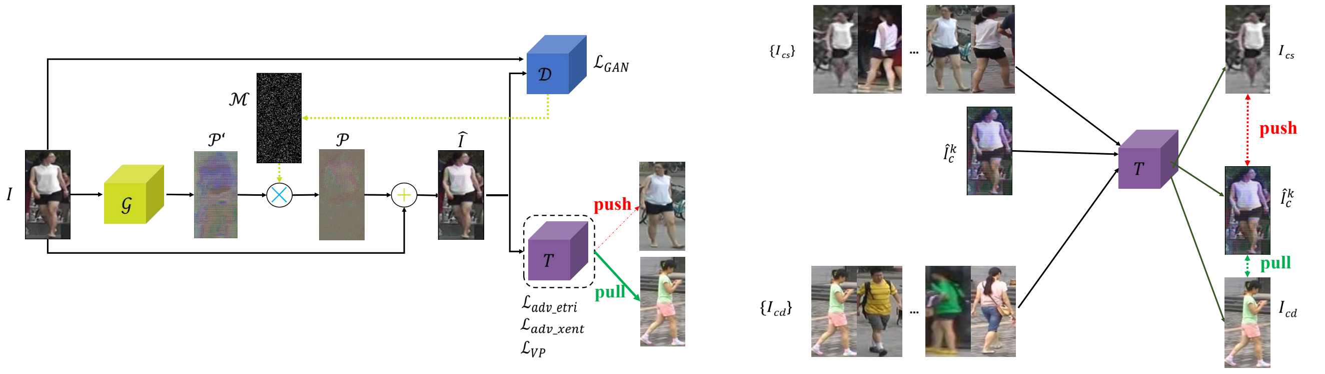
To address the aforementioned issues, we design a transferable, controllable, and inconspicuous attacker to examine the insecurity of current best-performing ReID systems. We propose a learning-to-mis-rank formulation to perturb the ranking prediction of ReID models. A new mis-ranking loss function is designed to attack the ranking of the potential matches, which fits the ReID problem perfectly. Our mis-ranking based attacker is complementary to existing misclassification based attackers. Besides, as is suggested by [12], adversarial examples are features rather than bugs. Hence, to enhance the transferability of the attacker, one needs to improve the representation learning ability of the attacker to extract the general features for the adversarial perturbations. To this end, we develop a novel multi-stage network architecture for representation learning by pyramiding the features of different levels of the discriminator. This architecture shows impressive transferability in black-box attack for the complicated ReID tasks. The transferability leads to our joint solution of both white- and black-box attack. To make our attack inconspicuous, we improve the existing adversarial attackers in two aspects. **First**, the number of target pixels to be attacked is controllable in our method, due to the use of a differentiable multi-shot sampling. Generally, the adversarial attack can be considered as searching for a set of target pixels to be contaminated by noise. To make the search space continuous, we relax the choice of a pixel as a Gumbel softmax over all possible pixels. The number of target pixels is determined by the dynamic threshold of the softmax output and thus can be controllable. **Second**, a new perception loss is designed by us to improve the visual quality of the attacked images, which guarantees the inconspicuousness.

Experiments were performed on four of the largest ReID benchmarks, i.e., Market1501 [45], CUHK03 [17], DukeMTMC [33], and MSMT17 [40]. The results show the effectiveness of our method. For example, the performance of one of the best-performing systems [44] drops sharply from 91.8% to 1.4% after attacked by our method. Except for showing a higher success attack rate, our method also provides interpretable attack analysis, which provides direction for improving the robustness and security of the ReID system. Some attack results are shown in Fig. 1. To summarize, our contribution is four-fold:

- To attack ReID, we propose a learning-to-mis-rank formulation to perturb the ranking of the system output. A new mis-ranking loss function is designed to attack the ranking of the predictions, which fits the ReID problem perfectly. Our mis-ranking based adversarial attacker is complementary to the existing misclassification based attackers.
- To enhance the transferability of our attacker and perform a black-box attack, we improve the representation capacity of the attacker to extract general and transferable features for the adversarial perturbations.
- To guarantee the inconspicuousness of the attack, we propose a differentiable multi-shot sampling to control the number of malicious pixels and a new perception loss to achieve better visual quality.
- By using the above techniques, we examine the insecurity of existing ReID systems against adversarial attacks. Experimental validations on four of the largest ReID benchmarks show not only the successful attack and the visual quality but also the interpretability of our attack, which provides directions for the future improvement in the robustness of ReID systems.

2. Related Work

Person Re-identification. ReID is different from image classification tasks in the setup of training and testing data. In an image classification task, the training and test set share the same categories, while in ReID, there is no category overlap between them. Therefore, deep ranking [4] is usually in desire for ReID. However, deep ranking is sensitive to alignment. To address the (dis)alignment problem, several methods have been proposed by using structural messages [18, 36]. Recently, Zhang *et al.* [44] introduce the shortest path loss to supervise local parts align-



(a) **(b)**

Figure 2. (a) The framework of our method. Our goal is to generate some noise $\mathcal{P}$ to disturb the input images $\mathcal{I}$. The disturbed images $\hat{\mathcal{I}}$ is able to cheat the ReID system $\mathcal{T}$ by attacking the visual similarities. (b) Specifically, the distance of each pair of samples from different categories (e.g., ($\hat{I}_c^k$, I), $\forall I \in \{I_{cd}\}$) is minimized, while the distance of each pair of the samples from the same category (e.g., ($\hat{I}_c^k$, I), $\forall I \in \{I_{cs}\}$) is maximized. The overall framework is trained by a generative adversarial network (GAN).

ing and adopt a mutual learning approach in the metric learning setting, which has obtained the surpassing human-level performance. Besides the supervised learning mentioned above, recent advance GANs have been introduced to ReID to boost performance in some unsupervised manner [3, 47, 49, 50]. Despite their success, the security and robustness of the existing ReID system have not yet been examined. Analyzing the robustness of a ReID system to resist attacks should be raised on the agenda.

Adversarial Attacks. Since the discovery of adversarial examples for DNNs [38], several adversarial attacks have been proposed in recent years. Goodfellow *et al.* [6] proposes to generate adversarial examples by using a single step based on the sign of the gradient for each pixel, which often leads to sub-optimal results and the lack of generalization capacity. Although DeepFool [28] is capable of fooling deep classifiers, it also lacks generalization capacity. Both methods fail to control the number of pixels to be attacked. To address this problem, [30] utilize the Jacobian matrix to implicitly conduct a fixed length of noise through the direction of each axis. Unfortunately, it cannot arbitrarily decide the number of target pixels to be attacked. [35] proposes to modify the single-pixel adversarial attack. However, the searching space and time grow dramatically with the increment of target pixels to be attacked. Besides the image classification, the adversarial attack is also introduced to face recognition [5, 34]. As discussed Section 1, all of the above methods do not fit the deep ranking problem. Also, their transferability is poor. Furthermore, many of them do not focus on the inconspicuousness of the visual quality. These drawbacks limit their applications in open-set tasks, e.g., ReID, which is our focus in this work. Although [1] has studied in metric analysis in person ReID, it does not provide a new adversarial attack method for ReID. It just uses the off-the-shelf methods for misclassification to examine very few ReID methods.

3. Methodology

3.1. Overall Framework

The overall framework of our method is presented in Fig. 2 (a). Our goal is to use the generator $\mathcal{G}$ to produce deceptive noises $\mathcal{P}$ for each input image $\mathcal{I}$. By adding the noises $\mathcal{P}$ to the image $\mathcal{I}$, we obtain the adversarial example $\hat{\mathcal{I}}$, using which we are able to cheat the ReID system $\mathcal{T}$ to output the wrong results. Specifically, the ReID system $\mathcal{T}$ may consider the matched pair of images dissimilar, while considering the mismatched pair of images similar, as shown in Fig.2 (b). The overall framework is trained by a generative adversarial network (GAN) with a generator $\mathcal{G}$ and a novel discriminator $\mathcal{D}$, which will be described in Section 3.3.

3.2. Learning-to-Mis-Rank Formulation For ReID

We propose a learning-to-mis-rank formulation to perturb the ranking of system output. A new mis-ranking loss function is designed to attack the ranking of the predictions, which fits the ReID problem perfectly. Our method tends to minimize the distance of the mismatched pair and maximize the distance of the matched pair simultaneously. We have:

$$\mathcal{L}_{adv\_etri} = \sum_{k=1}^{K}\sum_{c=1}^{C_k}\Big[\max_{\substack{j\neq k\\ j=1\ldots K\\ c_d=1\ldots C_j}} \left\|\mathcal{T}(\hat{\mathcal{I}}_c^k)-\mathcal{T}(\hat{\mathcal{I}}_{c_d}^j)\right\|_2^2 - \min_{c_s=1\ldots C_k}\left\|\mathcal{T}(\hat{\mathcal{I}}_c^k)-\mathcal{T}(\hat{\mathcal{I}}_{c_s}^k)\right\|_2^2+\Delta\Big]_+, \quad (1)$$

where C_k is the number of samples drawn from the k-th person ID, $\mathcal{I}_c^k$ is the c-th images of the k ID in a mini-batch, c_s and c_d are the samples from the same ID and the different IDs, $\|\cdot\|_2^2$ is the square of L2 norm used as the distance metric, and Δ is a margin threshold. Eqn.1 attacks the deep ranking in the form of triplet loss [4], where the distance of the *easiest* distinguished pairs of inter-ID images are encouraged to small, while the distance of the *easiest* distinguished pairs of intra-ID images are encouraged to large.

Remarkably, using the mis-ranking loss has a couple of advantages. **First**, the mis-ranking loss fits the ReID problem perfectly. As is mentioned above, ReID is different from image classification tasks in the setup of training and testing data. In an image classification task, the training and test set share the same categories, while in ReID, there is no category overlap between them. Therefore, the mis-ranking loss is suitable for attacking ReID. **Second**, the mis-ranking loss not only fits the ReID problem; it may fit all the open-set problems. Therefore, the use of mis-ranking loss may also benefit the learning of general and transferable features for the attackers. In summary, our mis-ranking based adversarial attacker is perfectly complementary to the existing misclassification based attackers.

3.3. Learning Transferable Features for Attacking

As is suggested by [12], adversarial examples are features rather than bugs. Hence, to enhance the transferability of an attacker, one needs to improve the representation learning ability of the attacker to extract the general features for the adversarial perturbations. In our case, the representation learners are the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$ (see Fig. 2 (a)). For the generator $\mathcal{G}$, we use the ResNet50. For the discriminator $\mathcal{D}$, recent adversarial defenders have utilized cross-layer information to identify adversarial examples [2, 19, 20, 26, 42]. As their rival, we develop a novel multi-stage network architecture for representation learning by pyramiding the features of different levels of the discriminator. Specifically, as shown in Fig. 3, our discriminator $\mathcal{D}$ consists of three fully convolutional sub-networks, each of which includes five convolutional, three downsampling, and several normalization layers [13, 27]. The three sub-networks receives $\{1, 1/2^2, 1/4^2\}$ areas of the original images as the input, respectively. Next, the feature maps from these sub-networks with the same size are combined into the same *stage* following [21]. A *stage pyramid* with series of downsampled results with a ratio of $\{1/32, 1/16, 1/8, 1/4\}$ of the image is thus formulated. With the feature maps from the previous stage, we upsample the spatial resolution by a factor of 2 using bilinear upsampling and attach a 1×1 convolutional layer to reduce channel dimensions. After an element-wise addition and a 3×3 convolutions, the fused maps are fed into the next stage. Lastly, the network ends with two atrous convolution layers and a 1×1 convolution to perform feature re-weighting, whose final response map λ is then fed into downstream sampler $\mathcal{M}$ discussed in Section 3.4. Remarkably, all these three sub-networks are optimized by standard loss following [25].

3.4. Controlling the Number of the Attacked Pixels

To make our attack inconspicuous, we improve the existing attackers in two aspects. The first aspect is to control the number of the target pixels to be attacked. Generally, an adversarial attack is to introduce a set of noise to a set of target pixels for a given image to form an adversarial example. Both the noise and the target pixels are unknown, which will be searched by the attacker. Here, we present the formulation of our attacker in searching for the target pixels. To make the search space continuous, we relax the choice of a pixel as a Gumbel softmax over all possible pixels:

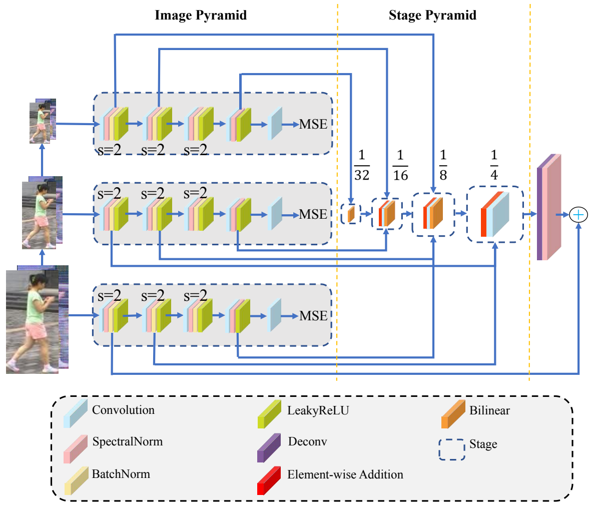


Figure 3. Detail of our multi-stage discriminator.

$$p_{i,j} = \frac{\exp((log(\lambda_{i,j} + \mathcal{N}_{i,j}))/\tau)}{\sum_{i,j=1}^{H,W} \exp(log(\lambda_{i,j} + \mathcal{N}_{i,j})/\tau)}, \tag{2}$$

where $i \in (0, H), j \in (0, W)$ denote the index of pixel in a feature map of size $H \times W$, where H/W are the height/width of the input images. The probability $p_{i,j}$ of a pixel to be chosen is parameterized by a softmax output vector $\lambda_{i,j}$ of dimension $H \times W$. $\mathcal{N}_{i,j} = -log(-log(U))$ is random variable at position (i, j), which is sampled from Gumbel distribution [8] with $U \sim Uniform(0, 1)$. Note that τ is a temperature parameter to soften transition from uniform distribution to categorical distribution when τ gradually reduces to zero. Thus, the number of the target pixels to be attacked is determined by the mask $\mathcal{M}$:

$$\mathcal{M}_{ij} = \begin{cases} \mathcal{K}eep\mathcal{T}opk(p_{i,j}), & \text{in forward propagation} \\ p_{i,j}, & \text{in backward propagation} \end{cases} \tag{3}$$

where $\mathcal{K}eep\mathcal{T}opk$ is a function by which the top-k pixels with the highest probability $p_{i,j}$ are retained in $\mathcal{M}$ while the other pixels are dropped during the forward propagation. Moreover, the difference between the forward and backward propagation ensures the differentiability. By multiplying the mask $\mathcal{M}$ and the preliminary noise $\mathcal{P}'$, we obtain the final noise $\mathcal{P}$ with controllable number of activated pixels. The usage of $\mathcal{M}$ is detailed in Fig. 2 (a).

3.5. Perception Loss for Visual Quality

In addition to controlling the number of the attacked pixels, we also focus on the visual quality to ensure the inconspicuousness of our attackers. Existing works introduce noises to images to cheat the machines without considering the visual quality of the images, which is inconsistent with human cognition. Motivated by MS-SSIM [39] that is able to provide a good approximation to perceive image quality for visual perception, we include an perception loss $\mathcal{L}_{VP}$ in our formulation to improve the visual quality:

$$\mathcal{L}_{VP}(\mathcal{I},\hat{\mathcal{I}}) = [l_L(\mathcal{I},\hat{\mathcal{I}})]^{\alpha_L} \cdot \prod_{j=1}^{L}[c_j(\mathcal{I},\hat{\mathcal{I}})]^{\beta_j}[s_j(\mathcal{I},\hat{\mathcal{I}})]^{\gamma_j}, \quad (4)$$

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. L is the level of scales, α_L, β_j, and γ_j are the factors to re-weight the contribution of each component. Thanks to $\mathcal{L}_{VP}$, the attack with high magnitude is available without being noticed by humans.

3.6. Objective Function

Besides the mis-ranking loss $\mathcal{L}_{adv\_etri}$, the perception loss $\mathcal{L}_{VP}$, we have two additional losses, i.e., a misclassification loss $\mathcal{L}_{adv\_xent}$, and a GAN loss $\mathcal{L}_{GAN}$.

Misclassification Loss. Existing works usually consider the least likely class as the target to optimize the cross-entropy between the output probabilities and its least likely class. However, the model may misclassify the inputs as any class except for the correct one. Inspired by [37], we propose a mechanism for relaxing the model for non-targeted attack by:

$$\mathcal{L}_{adv\_xent} = -\sum_{k=1}^{K}\mathcal{S}(\mathcal{T}(\hat{\mathcal{I}}))_k((1-\delta)\mathbb{1}_{\arg\min\mathcal{T}(\mathcal{I})_k}+\delta v_k), \quad (5)$$

where $\mathcal{S}$ denotes the log-softmax function, K is the total number of person IDs and $v = [\frac{1}{K-1},\ldots,0,\ldots,\frac{1}{K-1}]$ is smoothing regularization in which v_k equals to $\frac{1}{K-1}$ everywhere except when k is the ground-truth ID. The term $\arg\min$ in Eqn. 5 is similar to numpy.argmin which returns the indices of the minimum values of an output probability vector, indicating the least likely class. In practice, this smoothing regularization improves the training stability and the success attack rate.

GAN Loss. For our task, the generator $\mathcal{G}$ attempts to produce deceptive noises from input images, while the discriminator D distinguishes real images from adversarial examples as much as possible. Hence, the GAN loss $\mathcal{L}_{GAN}$ is given as:

$$\mathcal{L}_{GAN} = \mathbb{E}_{(I_{cd},I_{cs})}[\log\mathcal{D}_{1,2,3}(I_{cd},I_{cs})]+\mathbb{E}_{\mathcal{I}}[\log(1-\mathcal{D}_{1,2,3}(\mathcal{I},\hat{\mathcal{I}}))], \quad (6)$$

where $\mathcal{D}_{1,2,3}$ is our multi-stage discriminator shown in Fig. 3. We access to the final loss function:

$$\mathcal{L} = \mathcal{L}_{GAN} + \mathcal{L}_{adv\_xent} + \zeta\mathcal{L}_{adv\_etri} + \eta(1-\mathcal{L}_{VP}), \quad (7)$$

where ζ and η are loss weights for balance.

4. Experiment

We first present the results of attacking state-of-the-art ReID systems and then perform ablation studies on our method. Then, the generalization ability and interpretability of our method are examined by exploring black-box attacks.

Datasets. Our method is evaluated on four of the largest ReID datasets: Market1501 [45], CUHK03 [17] DukeMTMC [33] and MSMT17 [40]. Market1501 is a fully studied dataset containing 1,501 identities and 32,688 bounding boxes. CUHK03 includes 1,467 identities and 28,192 bounding boxes. To be consistent with recent works, we follow the **new** training/testing protocol to perform our experiments [48]. DukeMTMC provides 16,522 bounding boxes of 702 identities for training and 17,661 for testing. MSMT17 covers 4,101 identities and 126,441 bounding boxes taken by 15 cameras in both indoor and outdoor scenes. We adopt the standard metric of mAP and rank-$\{1,5,10,20\}$ accuracy for evaluation. *Note that in contrast to a ReID problem, lower rank accuracy and mAP indicate better success attack rate in a attack problem.*

Protocols. The details about training protocols and hyper-parameters can be seen in Appendix C. The first two subsections validate a white-box attack, i.e., the attacker has full access to training data and target models. In the third subsection, we explore a black-box attack to examine the transferability and interpretability of our method, i.e., the attacker has no access to the training data and target models. Following the standard protocols of the literature, all experiments below are performed by L_∞-bounded attacks with $\varepsilon = 16$ without special instruction, where ε is an upper bound imposed on the amplitude of the generated noise ($\{\|\hat{\mathcal{I}} - \mathcal{I}\|_{1,2,\mathrm{or}\,\infty} \leq \epsilon\}$) that determines the attack intensity and the visual quality.

4.1. Attacking State-of-the-Art ReID Systems

To demonstrate the generality of our method, we divide the state-of-the-art ReID systems into three groups as follows.

Attacking Different Backbones. We first examine the effectiveness of our method in attacking different best-performing network backbones, including ResNet-50 [9] (i.e., IDE [46]), DenseNet-121 [10], and Inception-v3 [37] (i.e., Mudeep [32]). The results are shown in Table 1 (a) & (b). We can see that the rank-1 accuracy of all backbones drop sharply approaching zero (e.g, from 89.9% to 1.2% for DenseNet) after it has been attacked by our method, suggesting that changing backbones cannot defend our attack.

Attacking Part-based ReID Systems. Many best-performing ReID systems learn both local and global similarity by considering part alignment. However, they still fail to defend our attack (Table 1 (a)(b)). For example, the accuracy of one of the best-performing ReID systems (AlignedReID [44]) drops sharply from 91.8% to 1.4% after it has been attacked by our method. This comparison

Table 1. Attacking the state-of-the-art ReID systems. *IDE:* [46]; *DenseNet-121:* [10]; *Mudeep:* [32]; *AlignedReid:* [44]; *PCB:* [36]; *HACNN:* [18]; *LSRO:* [47]; *HHL:* [49]; *SPGAN:* [3]; *CamStyle+Era:* [50]. We select GAP [31] and PGD [24] as the baseline attackers.

(a) Market1501

Methods		Rank-1				Rank-5				Rank-10				mAP			
		Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours
Backbone	IDE (ResNet-50)	83.1	5.0	4.5	**3.7**	91.7	10.0	8.7	**8.3**	94.6	13.9	12.1	**11.5**	63.3	5.0	4.6	**4.4**
	DenseNet-121	89.9	2.7	**1.2**	**1.2**	96.0	6.7	**1.0**	1.3	97.3	8.5	**1.5**	2.1	73.7	3.7	**1.3**	**1.3**
	Mudeep (Inception-V3)	73.0	3.5	2.6	**1.7**	90.1	5.3	5.5	**1.7**	93.1	7.6	6.9	**5.0**	49.9	2.8	2.0	**1.8**
Part-Aligned	AlignedReid	91.8	10.1	10.2	**1.4**	97.0	18.7	15.8	**3.7**	98.1	23.2	19.1	**5.4**	79.1	9.7	8.9	**2.3**
	PCB	88.6	6.8	6.1	**5.0**	95.5	14.0	12.7	**10.7**	97.3	19.2	15.8	**14.3**	70.7	5.6	4.8	**4.3**
	HACNN	90.6	2.3	6.1	**0.9**	95.9	5.2	8.8	**1.4**	97.4	6.9	10.6	**2.3**	75.3	3.0	5.3	**1.5**
Data Augmentation	CamStyle+Era (IDE)	86.6	6.9	15.4	**3.9**	95.0	14.1	23.9	**7.5**	96.6	18.0	29.1	**10.0**	70.8	6.3	12.6	**4.2**
	LSRO (DenseNet-121)	89.9	5.0	7.2	**0.9**	96.1	10.2	13.1	**2.2**	97.4	12.6	15.2	**3.1**	77.2	5.0	8.1	**1.3**
	HHL (IDE)	82.3	5.0	5.7	**3.6**	92.6	9.8	9.8	**7.3**	95.4	13.5	12.2	**9.7**	64.3	5.4	5.5	**4.1**
	SPGAN (IDE)	84.3	8.8	10.1	**1.5**	94.1	18.6	16.7	**3.1**	96.4	24.5	20.9	**4.3**	66.6	8.0	8.6	**1.6**

(b) CUHK03

Methods		Rank-1				Rank-5				Rank-10				mAP			
		Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours
Backbone	IDE (ResNet-50)	24.9	0.9	0.8	**0.4**	43.3	2.0	1.2	**0.7**	51.8	2.9	2.1	**1.5**	24.5	1.3	**0.8**	**0.9**
	DenseNet-121	48.4	2.4	0.1	**0.0**	50.1	4.4	**0.1**	**0.2**	70.1	5.9	**0.3**	0.6	84.0	1.6	**0.2**	**0.3**
	Mudeep (Inception-V3)	32.1	1.1	0.4	**0.1**	53.3	3.7	1.0	**0.5**	64.1	5.6	1.5	**0.8**	30.1	2.0	0.8	**0.3**
Part-Aligned	AlignedReid	61.5	2.1	**1.4**	**1.4**	79.4	4.6	**2.2**	3.7	85.5	6.2	**4.1**	5.4	59.6	3.4	**2.1**	**2.1**
	PCB	50.6	0.9	0.5	**0.2**	71.4	4.5	2.1	**1.3**	78.7	5.8	4.5	**1.8**	48.6	1.4	1.2	**0.8**
	HACNN	48.0	0.9	0.4	**0.1**	69.0	2.4	0.9	**0.3**	78.1	3.4	1.3	**0.4**	47.6	1.8	0.8	**0.4**

(c) DukeMTMC

Methods		Rank-1				Rank-5				Rank-10				mAP			
		Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours
Data augmentation	CamStyle+Era (IDE)	76.5	3.3	22.9	**1.2**	86.8	7.0	34.1	**2.6**	90.0	9.6	39.9	**3.4**	58.1	3.5	16.8	**1.5**
	LSRO (DenseNet-121)	72.0	1.3	7.2	**0.7**	85.7	2.9	12.5	**1.6**	89.5	4.0	18.4	**2.2**	55.2	1.4	8.1	**0.9**
	HHL (IDE)	71.4	1.8	9.5	**1.0**	83.5	3.4	15.6	**2.0**	87.7	4.2	19.0	**2.5**	51.8	1.9	7.4	**1.3**
	SPGAN (IDE)	73.6	5.3	12.4	**0.1**	85.2	10.3	21.1	**0.5**	88.9	13.4	26.3	**0.6**	54.6	4.7	10.2	**0.3**

proves that the testing tricks, e.g., extra local features ensemble in AlignedReID [44] and flipped image ensembling in PCB [36], are unable to resist our attack.

Attacking Augmented ReID Systems. Many state-of-the-art ReID systems use the trick of data augmentation. Next, we examine the effectiveness of our model in attacking these augmentation-based systems. Rather than conventional data augmentation trick (e.g., random cropping, flipping, and 2D-translation), we examine four new augmentation tricks using GAN to increase the training data. The evaluation is conducted on Market1501 and DukeMTMC. The results in Table 1 (a)(c) show that although GAN data augmentations improve the ReID accuracy, they cannot defend our attack. In contrast, we have even observed that better ReID accuracy may lead to worse robustness.

Discussion. *We have three remarks for rethinking the robustness of ReID systems for future improvement.* ***First****, there is no effective way so far to defend against our attacks, e.g., after our attack, all rank-1 accuracies drop below 3.9%.* ***Second****, the robustness of Mudeep [32] and PCB [36] are strongest. Intuitively, Mudeep may benefit from its nonlinear and large receptive field. For PCB, reprocessing the query images and hiding the network architecture during evaluation may improve the robustness.* ***Third****, HACNN [18] has the lowest rank-1 to rank-20 accuracy after the attack, suggesting that attention mechanism may be harmful to the defensibility. The returns from the target ReID system before and after the adversarial attack are provided in Appendix A.*

4.2. Ablation Study

We conduct comprehensive studies to validate the effectiveness of each component of our method. AlignedReID [44] is used as our target model in the rest of the paper for its remarkable results in ReID domain.

Different Losses. We report the rank-1 accuracy of four different losses to validate the effectiveness of our loss. The results are shown in Table 2 (a), where the four rows represent **(A)** the conventional misclassification loss, **(B)** our misclassification, **(C)** our mis-ranking loss, and **(D)** our misclassification + our mis-ranking loss, respectively. Actually, we observe that conventional misclassification loss A is incompatible with the perception loss, leading to poor attack performance (28.5%). In contrast, our visual mis-ranking loss D achieves very appealing attack performance (1.4%). We also observe that our misclassification loss B and our visual mis-ranking loss C benefit each other. Specifically, by combining these two losses, we obtain Loss D, which outperforms all the other losses.

Multi-stage vs. Common Discriminator. To validate the effectiveness of our multi-stage discriminator, we compare the following settings: **(A)** using our multi-stage discriminator and **(B)** using a commonly used discriminator. Specifically, we replace our multi-stage discriminator with PatchGAN [14]. Table 2 (c) shows a significant degradation of attack performance after changing the discriminator, demonstrating the superiority of our multi-stage discriminator to capture more details for a better attack.

Using MS-SSIM. To demonstrate the superiority of MS-SSIM, we visualize the adversarial examples under different perception supervisions in Fig. 4. We can see that at the same magnitude, the adversarial example generated under the supervision of MS-SSIM are much better than those generated under the supervision of SSIM and without any supervision. This comparison verifies that MS-SSIM is critical to reserve the raw appearance.

Comparisons of Different ϵ. Although using perception loss has great improvement for visual quality with large ϵ,

Table 2. **Ablations.** We present six major ablation experiments in this table. **R-1,R-5,& R-10:** Rank-1, Rank-5, & Rank-10.

	R-1	R-5	R-10	mAP
(A) cent	28.5	43.9	51.4	23.8
(B) xent	13.7	22.5	28.7	12.5
(C) etri	**4.5**	9.1	12.5	**5.1**
(D) xent+etri	**1.4**	**3.7**	**5.4**	**2.3**

(a) **Different Objectives**: The modified xent loss outperforms the cent loss, but both of them are unstable. Our loss brings more stable and higher fooling rate than misclassification.

	R-1	R-5	R-10	mAP
ϵ=40	0.0	0.2	0.6	0.2
ϵ=20	0.1	0.4	0.8	0.4
ϵ=16	1.4	3.7	5.4	2.3
ϵ=10	24.4	38.5	46.6	21.0

(b) **Comparisons of different** ϵ: Results on the variants of our model using different ϵ. Our proposed method achieves good results even when $\epsilon = 10$.

	R-1	R-5	R-10	mAP
PatchGAN (ϵ=40)	48.3	65.8	73.1	37.7
Ours (ϵ=40)	**0.0**	**0.2**	**0.6**	**0.2**
PatchGAN (ϵ=10)	53.3	69.2	75.6	43.2
Ours (ϵ=10)	**24.4**	**38.5**	**46.6**	**21.0**

(c) **Multi-stage vs. Common discriminator**: Multi-stage technique improves results under both large and small ϵ for utilizing the information from previous layers.

	R-1	R-5	R-10	mAP
Market→CUHK	4.9	9.2	12.1	6.0
CUHK→Market	34.3	51.6	58.6	28.2
Market→Duke	17.7	26.7	32.6	14.2
Market→MSMT	35.1	49.4	55.8	27.0

(d) Crossing Dataset. **Market→CUHK**: noises learned from Market1501 mislead inferring on CUHK03. All experiments are based on AlignedReID model.

	R-1	R-5	R-10	mAP
→PCB	31.7	46.1	53.2	22.9
→HACNN	14.8	24.4	29.8	13.4
→LSRO	17.0	28.9	35.1	14.8

(e) Crossing Model. →**PCB**: noises learned from **AlignedReID** attack pretrained PCB model. All experiments are performed on Market1501.

	R-1	R-5	R-10	mAP
→PCB(C)	6.9	12.9	18.9	8.2
→HACNN(C)	3.6	7.1	9.2	4.6
→LSRO(D)	19.4	30.2	34.7	15.2
→Mudeep(C)*	19.4	27.7	34.9	16.2

(f) Crossing Dataset & Model. → **PCB(C)**: noises learned from **AlignedReID** pretrained on Market-1501 are borrowed to attack PCB model inferred on CUHK03. * denotes **4k**-pixel attack.

Table 3. Proportion of adversarial points. † denotes the results with appropriate relaxation.

	R-1	R-5	R-10	mAP
Full size	1.4	3.7	5.4	2.3
Ratio=1/2	39.3	55.0	62.4	31.5
Ratio=1/4	72.7	85.9	89.7	58.3
Ratio=1/4†	0.3	1.5	2.7	0.7
Ratio=1/8†	0.6	1.8	3.0	1.1
Ratio=1/16†	8.2	14.7	17.8	6.9
Ratio=1/32†	59.4	76.5	82.2	47.3
Ratio=1/64†	75.5	87.6	91.6	61.5

Table 4. Effectiveness of our multi-shot sampling.

	(A) random location		**(B) our learned location**	
	R-1	mAP	R-1	mAP
Gaussian noise	81.9	68.1	79.4	65.3
Uniform noise	51.1	40.1	50.7	39.2
Ours	-	-	**39.7**	**30.7**

we also provide baseline models with smaller ϵ for full studies. We manually control ϵ by considering it as a hyperparameter. The comparisons of different ϵ are reported in Table 2 (b). Our method has achieved good results, even when $\epsilon = 15$. The visualization of several adversarial examples with different ϵ can be seen in Appendix D.

Number of the Pixels to be Attacked. Let H and W denote the height and the width of the image. We control the number of the pixels to be attacked in the range of $\{1,$ $1/2,$ $1/4,$ $1/8,$ $1/16,$ $1/32,$ $1/64\}$ $\times HW$ respectively by using Eqn. 3. We have two major observations from Table 3. **First**, the attack is definitely successful when the number of the pixels to be attacked $> \frac{HW}{2}$. This indicates that we can fully attack the ReID system by using a noise number of only $\frac{HW}{2}$. **Second**, when the number of pixels to be attacked $< \frac{HW}{2}$, the success attack rate drops significantly. To compensate for the decrease in noise number, we propose to enhance the noise magnitude without significantly affecting the perception. In this way, the least number of pixels to be attacked is reduced to $\frac{HW}{32}$, indicating that the number and the magnitude of the noise are both important.

Effectiveness of Our Multi-shot Sampling. To justify the effectiveness of our learned noise in attacking ReID, we compare them with random noise under the restriction of $\varepsilon = 40$ in two aspects in Table 4. **(A)** Random noise is imposed on random locations of an image. The results suggest that rand noise is inferior to our learned noise. **(B)** Random noise is imposed on our learned location of an image. Interestingly, although (B) has worse attack performance than our learned noise, (B) outperforms (A). This indicates our method successfully finds the sensitive location to attack.

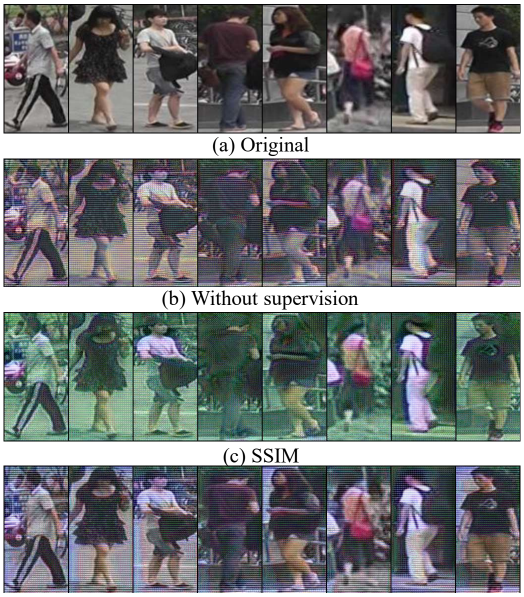
Interpretability of Our Attack. *After the analysis of the superiority of our learned noise, we further visualize the noise layout to explore the interpretability of our attack in ReID. Unfortunately, a single image cannot provide intuitive information (see Appendix B). We statistically display query images and masks when noise number equals to* $\frac{HW}{8}$ *in Fig. 5 for further analysis. We can observe from Fig. 5 (b) that the network has a tendency to attack the top half of the average image, which corresponds to the upper body of a person in Fig. 5(a). This implies that the network is able to sketch out the dominant region of the image for ReID. For future improvement of the robustness of ReID systems, attention should be paid to this dominant region.*

4.3. Black-Box Attack

Different from the above white-box attack, a black-box attack denotes that the attacker has no access to the training data and target models, which is very challenging.

Cross-dataset attack. Cross-dataset denotes that the attacker is learned on a known dataset, but is reused to attack a model that is trained on an unknown dataset. Table 2 (d) shows the success of our cross-dataset attack in AlignedReID [44]. We also observe that the success rate of the cross-dataset attack is almost as good as the naive white-box attack. Moreover, MSMT17 is a dataset that simulates the real scenarios by covering multi-scene and multi-time. Therefore, the successful attack on MSMT17 proves that our method is able to attack ReID systems in the real scene without knowing the information of real-scene data.

Cross-model attack. Cross-model attack denotes that



(a) Original

(b) Without supervision

(c) SSIM

(d) MS-SSIM

Figure 4. Visual comparison of using different supervisions.

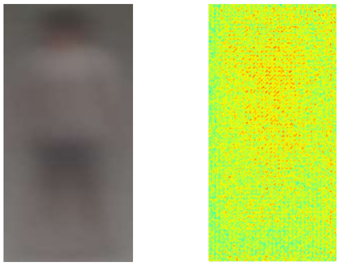
the attacker is learned by attacking a known model, but is reused to attack an unknown model. Experiments on Market1501 show that existing ReID systems are also fooled by our cross-model attacked (Table 2 (e)). It is worth to mention that PCB seems to be more robust than others, indicating that hiding the testing protocol benefits the robustness.

Cross-dataset-cross-model attack. We further examine the most challenging setting, i.e., our attacker has no access to both the training data and the model. The datasets and models are randomly chosen in Table 2 (f). Surprisingly, we have observed that our method has successfully fooled all the ReID systems, even in such an extreme condition. Note that Mudeep has been attacked by only 4,000 pixels.

Discussion. *We have the following remarks for future improvement in ReID.* ***First****, although the bias of data distributions in different ReID datasets reduces the accuracy of a ReID system, it is not the cause of security vulnerability, as is proved by the success of cross-dataset attack above.* ***Second****, the success of cross-model attack implies that the flaws of networks should be the cause of security vulnerability.* ***Third****, the success of a cross-dataset-cross-model attack drives us to rethink the vulnerability of existing ReID systems. Even we have no prior knowledge of a target system; we can use the public available ReID model and datasets to learn an attacker, using which we can perform the cross-dataset-cross-model attack in the target systems. Actually, we have fooled a real-world system (see Appendix D).*

4.4. Comparison with Existing Attackers

To show the generalization capability of our method, we perform an additional experiment on image classification using CIFAR10. We compare our method with four advanced white-box attack methods in adversarial examples community, including DeepFool [28], NewtonFool [15], CW [2], and GAP [41]. We employ adversarially trained ResNet32 as our target model and fix $\varepsilon = 8$. Other hyper-parameters are configured using default settings the same as [29]. For each attack method, we list the accuracy of the resulting network on the full CIFAR10 *val* set. The results in Table 5 imply that our proposed algorithm is also effective in obfuscating the classification system. Note that changing ε to other numbers (e.g., $\varepsilon = 2$) does not reduce the superiority of our method over the competitors.



(a) Average image (b) Position statistics

Figure 5. **Left**: The average image of all queries on Market1501. **Right**: The frequency of adversarial points appears at different positions among Market1501 when ratio=1/8. The higher the color temperature is, the frequently the position tends to be selected.

Table 5. Accuracy after non-targeted white-box attacks on CIFAR10. *Original:* the accuracy on clean images. *DeepFool:* [28]; *NewtonFool:* [15]; *NewtonFool:* [15]; *CW:* [2]; *GAP:* [41];

Method	Accuracy (%)			
Original	90.55			
DeepFool	$\varepsilon = 8$	58.22	$\varepsilon = 2$	58.59
NewtonFool		69.79		69.32
CW		52.27		53.44
GAP		51.26		51.8
Ours		**47.31**		50.3

5. Conclusion

We examine the insecurity of current ReID systems by proposing a learning-to-mis-rank formulation to perturb the ranking of the system output. Our mis-ranking based attacker is complementary to the existing misclassification based attackers. We also develop a multi-stage network architecture to extract general and transferable features for the adversarial perturbations, allowing our attacker to perform a black-box attack. We focus on the inconspicuousness of our attacker by controlling the number of attacked pixels and keeping the visual quality. The experiments not only show the effectiveness of our method but also provides directions for the future improvement in the robustness of ReID.

Acknowledgement

This work was supported in part by the State Key Development Program (No. 2018YFC0830103), in part by NSFC (No.61876224,U1811463,61622214,61836012,61906049), and by GD-NSF (No.2017A030312006,2020A1515010423).

References

[1] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip H. S. Torr. Metric attack and defense for person re-identification, 2019. 3

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *S&P*, pages 39–57. IEEE, 2017. 4, 8

[3] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018. 3, 6

[4] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 48(10):2993–3003, 2015. 2, 3

[5] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7714–7722, 2019. 3

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. 3

[7] Mengran Gou, Xikang Zhang, Angels Rates-Borras, Sadjad Asghari-Esfeden, Octavia I. Camps, and Mario Sznaier. Person re-identification in appearance impaired scenarios. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016. 1

[8] Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*. US Govt. Print. Office, 1954. 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 5, 6

[11] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pages 1–8, 2019. 1

[12] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *NeurIPS*, 2019. 2, 4

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 4

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 6

[15] Uyeong Jang, Xi Wu, and Somesh Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *ACSAC*, pages 262–277. ACM, 2017. 8

[16] Annan Li, Luoqi Liu, Kang Wang, Si Liu, and Shuicheng Yan. Clothing attributes assisted person reidentification. *IEEE Trans. Circuits Syst. Video Techn.*, 25(5):869–878, 2015. 1

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 1, 2, 5

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 2, 6

[19] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *ICCV*, pages 5764–5772, 2017. 4

[20] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, pages 1778–1787, 2018. 4

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4

[22] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognit.*, 95:151–161, 2019. 1

[23] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017. 1

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 6

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 4

[26] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *ICLR*, 2017. 4

[27] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 4

[28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016. 3, 8

[29] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v0.6.0. *CoRR*, 2018. 8

[30] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, pages 372–387. IEEE, 2016. 3

[31] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[32] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5409–5418, 2017. 5, 6

[33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35. Springer, 2016. 1, 2, 5

[34] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540, 2016. 3

[35] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *TEVC*, 2019. 1, 3

[36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 1, 2, 6

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 5

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 3

[39] Zhou Wang, Eero P Simoncelli, Alan C Bovik, et al. Multiscale structural similarity for image quality assessment. In *ACSSC*, volume 2, pages 1398–1402. Ieee, 2003. 5

[40] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 1, 2, 5

[41] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *IJCAI*, pages 3905–3911, 2018. 8

[42] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, pages 501–509, 2019. 4

[43] Jia Xue, Zibo Meng, Karthik Katipally, Haibo Wang, and Kees van Zon. Clothing change aware person identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2112–2120, 2018. 1

[44] 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. *CoRR*, 2017. 1, 2, 5, 6, 7

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 2, 5

[46] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *CoRR*, 2016. 5, 6

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 3, 6

[48] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 5

[49] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018. 3, 6

[50] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, pages 5157–5166, 2018. 3, 6