advPattern: Physical-World Attacks on Deep Person Re-Identification via Adversarially Transformable Patterns

Zhibo Wang[†], Siyan Zheng[†], Mengkai Song[†], Qian Wang[†,*], Alireza Rahimpour[‡], Hairong Qi[‡]

[†]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, P. R. China

[‡]Dept. of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, USA

{zbwang, zhengsy, mksong, qianwang}@whu.edu.cn, {arahimpo, hqi}@utk.edu

Abstract

Person re-identification (re-ID) is the task of matching person images across camera views, which plays an important role in surveillance and security applications. Inspired by great progress of deep learning, deep re-ID models began to be popular and gained state-of-the-art performance. However, recent works found that deep neural networks (DNNs) are vulnerable to adversarial examples, posing potential threats to DNNs based applications. This phenomenon throws a serious question about whether deep re-ID based systems are vulnerable to adversarial attacks.

In this paper, we take the first attempt to implement robust physical-world attacks against deep re-ID. We propose a novel attack algorithm, called advPattern, for generating adversarial patterns on clothes, which learns the variations of image pairs across cameras to pull closer the image features from the same camera, while pushing features from different cameras farther. By wearing our crafted "invisible cloak", an adversary can evade person search, or impersonate a target person to fool deep re-ID models in physical world. We evaluate the effectiveness of our transformable patterns on adversaries' clothes with Market1501 and our established PRCS dataset. The experimental results show that the rank-1 accuracy of re-ID models for matching the adversary decreases from 87.9% to 27.1% under Evading Attack. Furthermore, the adversary can impersonate a target person with 47.1% rank-1 accuracy and 67.9% mAP under Impersonation Attack. The results demonstrate that deep re-ID systems are vulnerable to our physical attacks.

1. Introduction

* Person re-identification (re-ID) [9] is an image retrieval problem that aims at matching a person of interest across
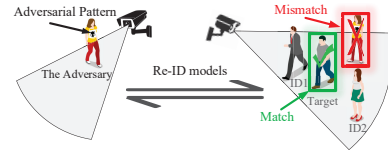

Figure 1. The illustration of Impersonation Attack on re-ID models. The adversary with the adversarial patterns lures re-ID models into mismatching herself as the target person.

multiple non-overlapping camera views. It has been increasingly popular in research area and has broad applications in video surveillance and security, such as searching suspects and missing people [30], cross-camera pedestrian tracking [33], and activity analysis [20]. Recently, inspired by the success of deep learning in various vision tasks [12, 14, 26, 27, 34, 35], deep neural networks (DNNs) based re-ID models [1, 4, 5, 6, 7, 18, 29, 31, 32] started to become a prevailing trend and have achieved state-of-the-art performance. Existing deep re-ID methods usually solve re-ID as a classification task [1, 18, 32], or a ranking task [4, 6, 7], or both [5, 29].

Recent studies found that DNNs are vulnerable to adversarial attack [3, 10, 13, 16, 17, 22, 24, 28]. These carefully modified inputs generated by adding visually imperceptible perturbations, called adversarial examples, can lure DNNs into working in abnormal ways, which pose potential threats to DNNs based applications, e.g., face recognition [25], autonomous driving [8], and malware classification [11]. The broad deployment of deep re-ID in security related systems makes it critical to figure out whether such adversarial examples also exist on deep re-ID models. Serious consequences will be brought if deep re-ID systems are proved to be vulnerable to adversarial attacks, for example, a suspect who utilizes this vulnerability can escape from the person search of re-ID based surveillance systems.

To the best of our knowledge, we are the first to investigate robust physical-world attacks on deep re-ID. In this paper, we propose a novel attack algorithm, called advPat-

---

tern, to generate adversarially transformable patterns across camera views that cause image mismatch in deep re-ID systems. An adversary cannot be correctly matched by deep re-ID models by printing the adversarial pattern on his clothes, like wearing an "invisible cloak". We present two different kinds of attacks in this paper: Evading Attack and Impersonation Attack. The former can be viewed as an untargeted attack that the adversary attempts to fool re-ID systems into matching him as an arbitrary person except himself. The latter is a targeted attack which goes further than Evading Attack: the adversary seeks to lure re-ID systems into mismatching himself as a target person. Figure 1 gives an illustration of Impersonation Attack on deep re-ID models.

The main challenge with generating adversarial patterns is that *how to cause deep re-ID systems to fail to correctly match the adversary's images across camera views with the same pattern on clothes*. Furthermore, the adversary might be captured by re-ID systems in any position, *but the adversarial pattern generated specifically for one shooting position is difficult to remain effective in other varying positions*. In addition, other challenges with physically realizing attacks also exist: (1) *How to allow cameras to perceive the adversarial patterns but avoid arousing suspicion of human supervisors*? (2) *How to make the generated adversarial patterns survive in various physical conditions, such as printing process, dynamic environments and shooting distortion of cameras*?

To address these challenges, we propose advPattern that formulates the problem of generating adversarial patterns against deep re-ID models as an optimization problem of minimizing the similarity scores of the adversary's images across camera views. The key idea behind advPattern is to amplify the difference of person images across camera views in the process of extracting features of images by re-ID models. To achieve the scalability of adversarial patterns, we approximate the distribution of viewing transformation with a multi-position sampling strategy. We further improve adversarial patterns' robustness by modeling physical dynamics (e.g., weather changes, shooting distortion), to ensure them survive in physical-world scenario. Figure 2 shows an example of our physical-world attacks on deep re-ID systems.

To demonstrate the effectiveness of advPattern, we first establish a new dataset, PRCS, which consists of 10,800 cropped images of 30 identities, and then evaluate the attack ability of adversarial patterns on two deep re-ID models using the PRCS dataset and the publicly available Market1501 dataset. We show that our adversarially transformable patterns generated by advPattern achieve high success rates under both Evading Attack and Impersonation Attack: the rank-1 accuracy of re-ID models for matching the adversary decreases from 87.9% to 27.1% under Evading Attack, meanwhile the adversary can impersonate as a target per-
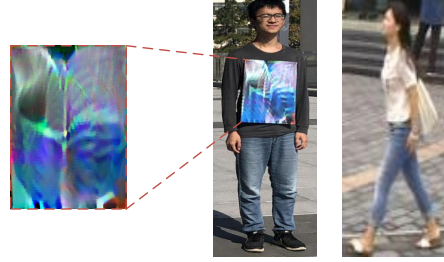

Figure 2. An example of Impersonation Attack in physical world. Left: the digital adversarial pattern; Middle: the adversary wearing a clothes with the physical adversarial pattern; Right: The target person randomly chosen from Market1501 dataset.

son with 47.1% rank-1 accuracy and 67.9% mAP under Impersonation Attack. The results demonstrate that deep re-ID models are indeed vulnerable to our proposed physical-world attacks.

In summary, our main contributions are three-fold:

- To the best of our knowledge, we are the first to implement physical-world attacks on deep re-ID systems, and reveal the vulnerability of deep re-ID modes.

- We design two different attacks, Evading Attack and Impersonation Attack, and propose a novel attack algorithm advPattern for generating adversarially transformable patterns, to realize adversary mismatch and target person impersonation, respectively.

- We evaluate our attacks with two state-of-the-art deep re-ID models and demonstrate the effectiveness of the generated patterns to attack deep re-ID in both digital domain and physical world with high success rate.

The remainder of this paper is organized as follows: we review some related works in Section 2 and introduce the system model in Section 3. In Section 4, we present the attack methods for implementing physical-world attacks on deep re-ID models. We evaluate the proposed attacks and demonstrate the effectiveness of our generated patterns in Section 5 and conclude with Section 6.

2. Related Work

Deep Re-ID Models. With the development of deep learning and increasing volumes of available datasets, deep re-ID models have been adopted to automatically learn better feature representation and similarity metric [1, 4, 5, 6, 7, 18, 29, 31, 32], achieving state-of-the art performance. Some methods treat re-ID as a classification issue: Li et al. [18] proposed a filter pairing neural network to automatically learn feature representation. Yi et al. [32] used a siamese deep neural network to solve the re-ID problem. Ahmed et al. [1] added a different matching layer to improve original deep architectures. Xiao et al. [31] utilized multi-class classification loss to train model with data from

multiple domains. Other approaches solve re-ID as a ranking task: Ding et al. [7] trained the network with the proposed triplet loss. Cheng et al. [6] introduced a new term to the original triplet loss to improve model performance. Besides, two recent works [5, 29] considered two tasks simultaneously and built networks to jointly learn representation from classification loss and ranking loss during training.

Adversarial Examples. Szegedy et al. [28] discovered that neural networks are vulnerable to adversarial examples. Given a DNNs based classifier $f(\cdot)$ and an input x with ground truth label y, an adversarial example x' is generated by adding small perturbations to x such that the classifier makes a wrong prediction, as $f(x') \neq y$, or $f(x') = y^*$ for a specific target $y^* \neq y$. Existing attack methods generate adversarial examples either by one-step methods, like the Fast Gradient Sign Method (FGSM) [10], or by solving optimization problems iteratively, such as L-BFGS [28], Basic Iterative Methods(BIM) [15], DeepFool [22], and Carlini and Wagner Attacks(C&W) [3]. Kurakin et al. [15] explored adversarial attack in physical world by printing adversarial examples on paper to cause misclassification when photographed by cellphone camera. Sharif et al. [25] designed eyeglass frame by printing adversarial perturbations on it to attack face recognition systems. Evtimov et al. [8] created adversarial road sign to attack road sign classifiers under different physical conditions. Athalye et al. [2] constructed physical 3D-printed adversarial objects to fool a classifier when photographed over a variety of viewpoints.

In this paper, to the best of our knowledge, we are the first to investigate physical-world attacks on deep re-ID models, which differs from prior works targeting on classifiers as follows: (1) Existing works on classification task failed to generate transformable patterns across camera views against image retrieval problems. (2) Attacking re-ID systems in physical world faces more complex physical conditions, for instance, adversarial patterns should survive in printing process, dynamic environments and shooting distortion under any camera views. These differences make it impossible to directly apply existing physical realizable methods on classifiers to attack re-ID models.

3. System Model

In this section, we first present the threat model and then introduce the our design objectives.

3.1. Threat Model

Our work focuses on physically realizable attacks against DNNs based re-ID systems, which capture pedestrians in real-time and automatically search a person of interest across non-overlapping cameras. By comparing the extracted features of a probe (the queried image) with features from a set of continuously updated gallery images collected from other cameras in real time, a re-ID system outputs images from the gallery which are considered to be the most similar to the queried image. We choose re-ID system as our target model because of the wild deployment of deep re-ID in security-critical settings, which will throw dangerous threats if successfully implementing physical-world attacks on re-ID models. For instance, a criminal can easily escape from the search of re-ID based surveillance systems by physically deceiving deep re-ID models.

We assume the adversary has *white-box access* to well-trained deep re-ID models, so that he has knowledge of model structure and parameters, and *only* implements attacks on re-ID models in the inference phase. The adversary is not allowed to manipulate either the digital queried image or gallery images gathered from cameras. Moreover, the adversary is not allowed to change his physical appearance during attacking re-ID systems in order to avoid arousing human supervisor's suspicion. These reasonable assumptions make it challenging to realize successfully physical-world attacks on re-ID systems.

Considering that the stored video recorded by cameras will be copied and re-ID models will be applied for person search only when something happens, the adversary has no idea of when he will be treated as the person of interest and which images will be picked for image matching, which means that the queried image and gallery images are completely unknown to the adversary. However, with the white-box access assumption, the adversary is allowed to construct a generating set X by taking images at each different camera view, which can be realized by stealthily placing cameras at the same position of surveillance cameras to capture images before implementing attacks.

3.2. Design Objectives

We propose two attack scenarios, Evading Attack and Impersonation Attack, to deceive deep-ID models.

Evading Attack. An Evading Attack is an *untargeted attack*: *Re-ID models are fooled to match the adversary as an arbitrary person except himself, which looks like that the adversary wears an "invisible cloak"*. Formally, a re-ID model $f_\theta(\cdot, \cdot)$ outputs a similarity score of an image pair, where θ is the model parameter. Given a probe image p_{adv} of an adversary, and an image g_{adv} belonging to the adversary in the gallery G_t at time t, we attempt to find an adversarial pattern δ attached on the adversary's clothes to fail deep re-ID models in person search by solving the following optimization problem:

$$\max D(\delta), \quad s.t. \; Rank(f_\theta(p_{adv+\delta}, g_{adv+\delta})) > K \quad (1)$$

where $D(\cdot)$ is used to measure the reality of the generated pattern. Unlike previous works aiming at generating visually inconspicuous perturbations, we attempt to generate *visible patterns* for camera sensing, while *making generated patterns indistinguishable from naturally decorative pattern*
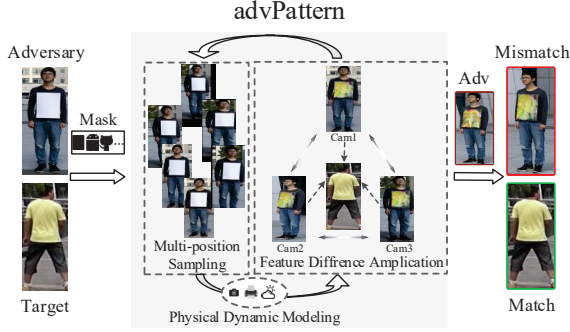
Figure 3. Overview of the attack pipeline.

on clothes. $Rank(\cdot)$ is a sort function which ranks similarity scores of all gallery images with p_{adv} in the decreasing order. An adversarial pattern is successfully crafted only if the image pair $(p_{adv+\delta}, g_{adv+\delta})$ ranks behind the top-K results, which means that the re-ID systems cannot realize cross-camera image match of the adversary.

Impersonation Attack. An Impersonation Attack is a *targeted attack* which can be viewed as an extension of Evading Attack: The adversary attempts to *deceive re-ID models into mismatching himself as a target person*. Given our target's image I_t, we formulate Impersonation Attack as the following optimization problem:

$$\max D(\delta), \quad s.t. \begin{cases} Rank(f_\theta(p_{adv+\delta}, g_{adv+\delta})) > K \\ Rank(f_\theta(p_{adv+\delta}, I_t)) < K \end{cases} \quad (2)$$

we can see that, besides the evading constraint, the optimization problem for an Impersonation Attack includes another constraint that the image pair $(p_{adv+\delta}, I_t)$ should be within the top-K results, which implies that the adversary successfully induces the re-ID systems into matching him to the target person.

Since the adversary has no knowledge of the queried image and the gallery, it is impossible for the adversary to solve the above optimization problems. In the following section, we will present the solution that approximately solve the above optimization problems.

4. Adversarial Pattern Generation

In this section, we present a novel attack algorithm, called advPattern, to generate adversarial patterns for attacking deep re-ID systems in real-world. Figure 3 shows an overview of the pipeline to implement an Impersonation Attack in physical world. Specifically, we first generate transformable patterns across camera views for attacking the image retrieval problem as described in Section 4.1. To implement position-irrelevant and physical-world attacks, we further improve the scalability and robustness of adversarial patterns in Section 4.2 and Section 4.3.

4.1. Transformable Patterns across Camera Views

Existing works [6, 38] found that there exists a common image style within a certain camera view, while dramatic variations across different camera views. To ensure that the same pattern can cause cross-camera image mismatch in deep re-ID models, we propose an adversarial pattern generation algorithm to generate transformable patterns that amplify the distinction of the adversary's images across camera views in the process of extracting features of images by re-ID models.

For the **Evading Attack**, given the generating set $X = (x_1, x_2, ..., x_m)$ constructed by the adversary, which consists of the adversary's images captured from m different camera views. For each image x_i from X, we compute the adversarial image $x_i' = o(x_i, T_i(\delta))$. $o(x_i, T_i(\delta))$ denotes overlaying the corresponding areas of x_i after transformation $T_i(\cdot)$ with the generated pattern δ. Here $T_i(\delta)$ is a perspective transformation operation of the generated pattern δ, which ensures the generated pattern to be in accordance with transformation on person images across camera views. We generate the transformable adversarial pattern δ by solving the following optimization problem:

$$\arg\min_\delta \sum_{i=1}^m \sum_{j=1}^m f_\theta(x_i', x_j'), \quad s.t. \quad i \neq j \quad (3)$$

We iteratively minimize the similarity scores of images of an adversary from different cameras to gradually pull farther extracted features of the adversarys images from different cameras by the generated adversarial pattern.

For the **Impersonation Attack**, given a target person's image I_t, we optimize the following problem:

$$\arg\min_\delta \sum_{i=1}^m \sum_{j=1}^m f_\theta(x_i', x_j') \\ - \alpha(f_\theta(x_i', I_t) + f_\theta(x_j', I_t)), \quad s.t. \quad i \neq j \quad (4)$$

where α controls the strength of different objective terms. By adding the second term in Eq. 4, we additionally maximize similarity scores of the adversary's images with the target person's image to generate a more powerful adversarial pattern to pull closer the extracted features of the adversary's images and the target person's image.

4.2. Scalable Patterns in Varying Positions

The adversarial patterns should be capable of implementing successful attacks at any position, which means our attacks should be position-irrelevant. To realize this objective, we further improve the scalability of the adversarial pattern in terms of varying positions.

Since we cannot capture the exact distribution of viewing transformation, we augment the volume of the generat-
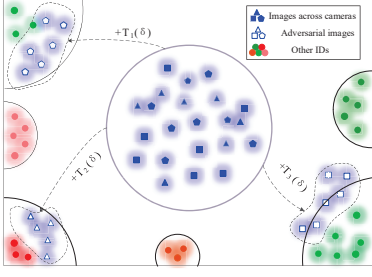
Figure 4. The illustration of how scalable adversarial patterns work. By adding the generated adversarial pattern, the adversarial images from the same camera view are clustered together in the feature space. Meanwhile, the distance of adversarial images from different cameras becomes farther.

ing set with a multi-position sampling strategy to approximate the distribution of images for generating scalable adversarial patterns. The augmented generating set X^C for an adversary is built by collecting the adversary's images with various distances and angles from each camera view, and synthesized instances generated by image transformation such as translation and scaling on original collected images.

For the **Evading Attack**, given a triplet $tri_k = < x_k^o, x_k^+, x_k^- >$ from X^C, where x_k^o and x_k^+ are person images from the same camera, while x_k^- is the person image from a different camera, for each image x_k from tri_k, we compute the adversarial image $x_k{}'$ as $o(x_k, T_k(\delta))$. We randomly chose a triplet at each iteration for solving the following optimization problem:

$$\arg\min_{\delta} \mathbb{E}_{tri_k \sim X^C} f_\theta((x_k^o)', (x_k^-)') - \beta f_\theta((x_k^o)', (x_k^+)')$$
(5)

where β is a hyperparameter that balances different objectives during optimization. The objective of Eq.(5) is to minimize the similarity scores of x_k^o with x_k^- to discriminate person images across camera views, while maximizing similarity scores of x_k^o with x_k^+ to preserve the similarity of person images from the same camera view. During optimization the generated pattern learns the scalability from the augmented generating set X^C to pull closer the extracted features of person images from the same camera, while pushing features from different cameras farther, as shown in Figure 4.

For the **Impersonation Attack**, given an image set I^t of the target person, and a quadruplet $quad_k = < x_k^o, x_k^+, x_k^-, t_k >$ consisting of a triplet tri_k and a person image t_k from I^t, we randomly choose a quadruplet at each iteration, and iteratively solve the following optimization problem:

$$\arg\max_{\delta} \mathbb{E}_{quad_k \sim \{X^C, I^t\}} f_\theta((x_k^o)', t_k)$$
$$+ \lambda_1 f_\theta((x_k^o)', (x_k^+)') - \lambda_2 f_\theta((x_k^o)', (x_k^-)')$$
(6)

where λ_1 and λ_2 are hyperparameters that control the

strength of different objectives. We add an additional objective that maximizes the similarity score of x_k^o with t_k to pull closer the extracted features of the adversary's images to the features of the target person's images.

4.3. Robust Patterns for Physically Realizable Attack

Our goal is to implement physically realizable attacks on deep re-ID systems by generating physically robust patterns on adversaries' clothes. To ensure adversarial patterns to be perceived by cameras, we generate large magnitude of patterns with no constraints over them during optimization. However, introducing conspicuous patterns will in turn make adversaries be attractive and arouse suspicion of human supervisors.

To tackle this problem, we design unobtrusive adversarial patterns which are visible but difficult for humans to distinguish them from the decorative patterns on clothes. To be specific, we choose a mask M_x to project the generated pattern to a shape that looks like a decorative pattern on clothes (e.g., commonplace logos or creative graffiti). In addition, to generate smooth and consistent patches in our pattern, in other words, colors change only gradually within patches, we follow Sharif et al. [25] that adds total variation (TV) [21] into the objective function:

$$TV(\delta) = \sum_{p,q} ((\delta_{p,q} - \delta_{p+1,q})^2 + (\delta_{p,q} - \delta_{p,q+1})^2)^{\frac{1}{2}}$$
(7)

where $\delta_{p,q}$ is a pixel value of the pattern δ at coordinates (p, q), and $TV(\delta)$ is high when there are large variations in the values of adjacent pixels, and low otherwise. By minimizing $TV(\delta)$, the values of adjacent pixels are encouraged to be closer to each other to improve the smoothness of the generated pattern.

Implementing physical-world attacks on deep re-ID systems requires adversarial patterns to survive in various environmental conditions. To deal with this problem, we design a degradation function $\varphi(\cdot)$ that randomly changes the brightness or blurs the adversary's images from the augmented generating set X^C. During optimization we replace x_i with the degraded image $\varphi(x_i)$ to improve the robustness of our generated pattern against physical dynamics and shooting distortion. Recently, the non-printability score (NPS) was utilized in [8, 25] to account for printing error. We introduce NPS into our objective function but find it hard to balance NPS term with other objectives. Alternatively, we constrain the search space of the generated pattern δ in a narrower interval $\mathbb{P}$ to avoid unprintable colors (e.g., high brightness and high saturation). Thus, for each image x_k from tri_k, we use $o(x_k, T_k(M_x \cdot \delta))$ to compute the adversarial images $x_k{}'$, and generate robust adversarial patterns to implement physical-world attack as solving the
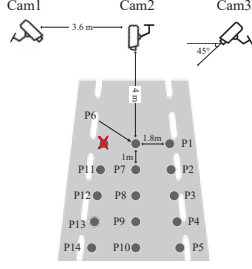
Figure 5. The scene setting of physical-world tests. We choose 14 testing points under each camera which vary in distances and angles.

following optimization problem:

$$\underset{\delta}{\arg\min}\, \mathbb{E}_{tri_k \sim X^C}\, f_\theta(\varphi(x_k^o)', \varphi(x_k^-)') \\ - \beta f_\theta(\varphi(x_k^o)', \varphi(x_k^+)') + \kappa \cdot TV(\delta), \quad s.t. \quad \delta \in \mathbb{P} \tag{8}$$

where λ and κ are hyperparameters that control the strength of different objectives. Similarly, the formulation of the Impersonation Attack is analogous to that of the Evading Attack, which is as follows:

$$\underset{\delta}{\arg\min}\, \mathbb{E}_{quad_k \sim \{X^C, I^t\}}\, f_\theta((x_k^o)', t_k) + \lambda_1 f_\theta((x_k^o)', (x_k^+)') \\ - \lambda_2 f_\theta((x_k^o)', (x_k^-)') + \kappa \cdot TV(\delta), \quad s.t. \quad \delta \in \mathbb{P} \tag{9}$$

Finally, we print the generated pattern over the the adversary's clothes to deceive re-ID into mismatching him as an arbitrary person or a target person.

5. Experiments

In this section, we first introduce the datasets and the target deep re-ID models used for evaluation in Section 5.1. We evaluate the proposed advPattern for attacking deep re-ID tools both under digital environment (Section 5.2) and in physical world (Section 5.3). We finally discuss the implications and limitations of advPattern in Section 5.4.

5.1. Datasets and re-ID Models

Market1501 Dataset. Market1501 contains 32,668 annotated bounding boxes of 1501 identities, which is divided into two non-overlapping subsets: the training dataset contains 12,936 cropped images of 751 identities, while the testing set contains 19,732 cropped images of 750 identities.

PRCS Dataset. We built a Person Re-identification in Campus Streets (PRCS) dataset for evaluating the attack method. PRCS contains 10,800 cropped images of 30 identities. During dataset collection, three cameras were deployed to capture pedestrians in different campus streets. Each identity in PRCS was captured by both three cameras and has at least 100 cropped images per camera. We chose

Table 1. Re-ID performance of model A and model B on Market1501 and PRCS datasets. (ss = Similarity Score)

Model	Dataset	rank-1	rank-5	rank-10	mAP	ss
A	Market-1501	77.8%	90.2%	93.5%	62.7%	0.796
	PRCS	87.9%	93.4%	100.0%	78.6%	0.876
B	Market-1501	74.5%	89.0%	92.8%	57.3%	0.732
	PRCS	84.7%	95.4%	99.0%	77.2%	0.857

Table 2. Digital-environment attack results on model A and model B under Evading Attack (GS = Generating Set, TS = Testing Set).

Model	Dataset	rank-1	rank-5	rank-10	mAP	ss
A	GS	0.0%	0.0%	0.0%	4.4%	0.394
	TS	4.2%	8.3%	16.7%	7.3%	0.479
B	GS	0.0%	0.0%	0.0%	4.5%	0.422
	TS	10.4%	13.3%	16.7%	16.3%	0.508

30 images of each identity per camera to construct the testing dataset for evaluating the performance of the trained re-ID models and our attack method.

Target Re-ID Models. We evaluated the proposed attack method on two different types of deep re-ID models: model A is a siamese network proposed by Zheng et al. [37], which is trained by combining verification loss and identification loss; model B utilizes a classification model to learn the discriminative embeddings of identities as introduced in [36]. The reason why we choose the two models as target models is that classification networks and siamese networks are widely used in the re-ID community. The effectiveness of our attacks on the two models can imply the effectiveness on other models. Both of the two models achieve the state-of-the-art performance (i.e., rank-k accuracy and mAP) on Market1501 dataset, and also work well on PRCS dataset. The results are given in Table 1.

We use the ADAM optimizer to generate adversarial patterns with the following parameters setting: learning rate $= 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We set the maximum number of iterations to 700.

5.2. Digital-Environment Tests

We first evaluate our attack method in digital domain where the adversary's images are directly modified by digital adversarial patterns[†]. It is worth noting that attacking in digital domain is actually not a realistic attack, but a necessary evaluation step before successfully implementing real physical-world attacks.

Experiment Setup. We first craft the adversarial pattern over a generating set for each adversary, which consists of real images from varying positions, viewing angles and synthesized samples. Then we attach generated adversarial pattern to the adversary's images in digital domain to evaluate the attacking performance on the target re-ID models.

[†]The code is avaliable at `https://github.com/whuAdv/AdvPattern`

Table 3. Digital-environment attack results on target models under Impersonation Attack. The performance of matching the adversary as the target person (Target) and himself (Adversary) are both given. (GS = Generating Set, TS = Testing Set).

Model	Matched person	PRCS					Market1501				
		rank-1	rank-5	rank-10	mAP	ss	rank-1	rank-5	rank-10	mAP	ss
A	Target(GS)	86.7%	97.2%	100.0%	89.7%	0.848	68.0%	83.8%	89.8%	70.7%	0.814
	Adversary(GS)	5.50%	5.70%	8.30%	8.10%	0.524	4.10%	4.10%	10.4%	9.20%	0.565
B	Target(GS)	92.8%	97.7%	100.0%	91.8%	0.858	94.0%	98.1%	100.0%	60.7%	0.775
	Adversary(GS)	2.50%	5.80%	8.40%	5.50%	0.486	1.80%	6.90%	8.70%	10.9%	0.606
A	Target(TS)	74.4%	83.3%	91.7%	81.5%	0.824%	55.8%	70.8%	79.2%	45.1%	0.803
	Adversary(TS)	19.4%	22.2%	38.9%	18.4%	0.633	14.1%	17.1%	27.5%	12.9%	0.638
B	Target(TS)	78.4%	83.3%	91.7%	88.2%	0.812	68.7%	79.2%	91.6%	51.9%	0.749
	Adversary(TS)	16.7%	18.9%	41.7%	24.8%	0.652	28.4%	34.7%	50.4%	31.3%	0.659

Table 4. Physical-world attack results on target models at varying distances and angles. Distance&Angle of each points with camera 3 are given. Δrank-1, ΔmAP and Δss indicate the drop of target models' performance due to adversarial patterns.

Distance&Angle	Evading Attack						Impersonation Attack		
	rank-1	Δrank-1	mAP	ΔmAP	ss	Δss	rank-1	mAP	ss
P1 (4.39, 24.2)	0.00%	100%	18.9%	69.2%	0.689	0.146	40.0%	58.4%	0.774
P2 (5.31, 19.8)	20.0%	80.0%	25.6%	66.5%	0.694	0.149	80.0%	83.8%	0.746
P3 (6.26, 16.7)	20.0%	80.0%	24.6%	66.7%	0.687	0.163	0.0%	33.6%	0.728
P4 (7.23, 14.4)	20.0%	80.0%	21.1%	68.9%	0.680	0.180	80.0%	84.9%	0.737
P5 (8.20, 12.7)	20.0%	80.0%	23.4%	64.2%	0.660	0.124	20.0%	58.4%	0.748
P6 (5.38, 42.0)	0.00%	100%	18.6%	71.2%	0.685	0.192	40.0%	68.0%	0.784
P7 (6.16, 35.8)	40.0%	60.0%	27.6%	62.7%	0.709	0.164	40.0%	65.2%	0.761
P8 (7.00, 31.0)	0.00%	100%	15.9%	40.8%	0.663	0.083	80.0%	85.8%	0.762
P9 (7.87, 27.2)	0.0%	100%	18.0%	71.6%	0.669	0.158	40.0%	67.1%	0.767
P10 (8.77, 24.2)	40.0%	60.0%	40.0%	57.5%	0.714	0.167	40.0%	59.5%	0.761
P11 (7.36, 47.2)	20.0%	80.0%	22.6%	65.6%	0.695	0.147	60.0%	72.1%	0.768
P12 (8.07, 42.0)	60.0%	40.0%	27.6%	47.6%	0.688	0.124	80.0%	82.3%	0.771
P13 (8.84, 37.6)	80.0%	20.0%	47.5%	36.8%	0.749	0.071	40.0%	78.3%	0.789
P14 (9.65, 34.0)	40.0%	60.0%	20.9%	66.7%	0.699	0.148	20.0%	52.6%	0.768
Average	**27.1%**	**74.3%**	**25.2%**	**61.1%**	**0.692**	**0.144**	**47.1%**	**67.9%**	**0.762**

We choose every identity from PRCS as an adversary to attack deep re-ID. In each query, we choose an image from adversarial images as the probe image, and construct a gallery by combining 12 adversarial images from other cameras with images from 29 identities in PRCS, and 750 identities in Market1501. For Impersonation Attack, we take two identities as target for each adversary: one is randomly chosen from Market1501, and another one is chosen from PRCS. We ran 100 queries for each attack.

Experiment Results. Table 2 shows the attack results on two re-ID models under Evading Attack in digital environment. We can see that the matching probability and mAP drops significantly for both re-ID models, which demonstrate the high success rate of implementing the Evading Attack. The similarity score of the adversary's images decreases to less than 0.5, making it hard for deep re-ID models to correctly match images of the adversary in the large gallery. Note that the attack performance on testing set is close to generating set, e.g., rank-1 accuracy from 4.2% to 0% and mAP from 7.3% to 4.4%, which demonstrates the scalability of the digital adversarial patterns when implementing attacks with unseen images.

Table 3 shows the attack results to two re-ID models under Impersonation Attack in the digital environment. In PRCS, the average rank-1 accuracy is above 85% when matching adversarial images from the generating set as a target person, which demonstrates the effectiveness of implementing the targeted attack. The patterns are less effective when targeting an identity from Market1501: the rank-1 accuracy of model A decreases to 68.0% for the generating set, and 41.7% for the testing set. We attribute it to the large variations in physical appearance and image styles between two datasets. Though, the high rank-5 accuracy and mAP demonstrate the strong capability of digital patterns to deceive target models. Note that the rank-k accuracy and mAP decrease significantly for matching the adversary's images across cameras, which means that the generated patterns can also cause mismatch across camera views in targeted attacks. Again, that the attack performance on testing set is close to generating set demonstrates the scalability of the adversarial patterns with unseen images.

5.3. Physical-World Evaluation

On the basis of digital-environment tests, we further evaluate our attack method in physical world. We print the adversarial pattern and attach it to the adversary's clothes for implementing physical-world attacks.

Experiment Setup. The scene setting of physical-world tests is shown in Figure 5, where we take images of the adversary with/without the adversarial pattern in 14 testing

Figure 6. Examples of physically realizable attacks. Top row: an Evading Attack (the adversary: ID1 from PRCS). Middle row: an Impersonation Attack targeting identity from PRCS (the adversary: ID2 in PRCS, the target: ID12 from PRCS). Bottom row: an Impersonation Attack targeting identity from Market1501 (the adversary: ID3 in PRCS, the target: ID728 from Market1501)

points with variations in distances and angles from cameras. These 14 points are sampled with a fix interval in the filed of cameras views. We omit the left-top point in our experiment due to the constraint of shooting conditions. The distance between cameras and the adversary is about from $5m$ to $10m$ for better perceiving the adversarial pattern.

We choose 5 identities as adversaries from PRCS to implement physically realizable attacks. In each query, we randomly choose the adversary's image under a testing point as the probe image, while adding 12 adversarial images from other cameras into the gallery. Two identities are randomly chosen from Market1501 and PRCS respectively to serve as target person. 100 queries for each testing point are performed. We evaluate physical-world attacks for Evading Attack with model A, while Impersonation Attack with model B.

Experiment Results. Table 4 shows the physical-world attack results of 14 different testing positions with varying distances and angles from cameras. Note that Δrank-1 denotes the drop of match probability due to adversarial patterns. Similar meanings happen to ΔmAP and Δss. For the Evading Attack, we can see that it significantly decreases the match probability to the adversary with the crafted adversarial pattern. The average Δrank-1 and ΔmAP are 62.2% and 61.1%. The average of the rank-1 accuracy and mAP are 47.1% and 67.9% under Impersonation Attack, respectively. The results demonstrates the effectiveness of adversarial patterns to implement physical-world attacks in varying positions with considerable success rate.

For Evading Attack, the average rank-1 accuracy drops to 11.1% in 9 of 14 positions, which demonstrates that the generated adversarial patterns can physically attack deep re-ID systems with high success rate. Note that adversarial patterns are less effective in some testing points, e.g., P12 and P13. We attribute it to the larger angles and farther

distance between these points and cameras, which makes it more difficult for cameras to perceive the patterns. For Impersonation Attack, The rank-1 accuracy for matching the adversary as the target person is 56.4% in 11 of 14 positions , which is close to the result of digital patterns targeting on Market1501. The high mAP and similarity scores when matching the adversary as the targeted person demonstrate the effectiveness of adversarial patterns to implement targeted attack in physical world. Still, there exists few points (P3, P5, P14) where the adversary has trouble to implement successful attack with adversarial patterns. Figure 6 shows examples of physical-world attacks on deep re-ID systems.

5.4. Discussion

Black Box Attacks. In this paper, we start with the white-box assumption to investigate the vulnerability of deep re-IDs models. Nevertheless, it would be more meaningful if we can realize adversarial patterns with black-box setting. Prior works [19, 23] demonstrated successful attacks without any knowledge of model's internals by utilizing the transferability of adversarial examples. We will leave black-box attacks as our future work.

AdvPattern vs. Other Approaches. AdvPattern allows the adversary to deceive deep re-ID systems without any digital modifications of person images or any physical appearance change. Although there are simpler ways to attack re-ID systems, e.g., directly object removal in digital domain, or changing physical appearance in different camera view, we argue that our adversarial pattern is the most reasonable method because: (1) for object removal methods, it is unrealistic to control the queried image and gallery images; (2) changing physical appearance makes adversaries attractive to human supervisors.

6. Conclusion

This paper designed Evading Attack and Impersonation Attack for deep re-ID systems, and proposed advPattern for generating adversarially transformable patterns to realize adversary mismatch and target person impersonation in physical world. The extensive evaluations demonstrate the vulnerability of deep re-ID systems to our attacks.

Acknowledgments

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proc. of IEEE CVPR*, pages 3908–3916, 2015.

[2] Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *arXiv:1707.07397*, 2017.

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

[4] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proc. of AAAI*, pages 3988–3994, 2017.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. of IEEE CVPR*, pages 1335–1344, 2016.

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

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proc. of IEEE CVPR*, pages 1625–1634, 2018.

[9] Shaogang Gong, Marco Cristani, Chen Change Loy, and Timothy M Hospedales. The re-identification challenge. In *Person re-identification*, pages 1–20. 2014.

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

[11] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial perturbations against deep neural networks for malware classification. *arXiv:1606.04435*, 2016.

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

[13] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *Proc. of IEEE SPW*, pages 36–42, 2018.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*, pages 1097–1105, 2012.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv:1607.02533*, 2016.

[16] Bo Li and Yevgeniy Vorobeychik. Feature cross-substitution in adversarial classification. In *Proc. of NIPS*, pages 2087–2095, 2014.

[17] Bo Li and Yevgeniy Vorobeychik. Scalable optimization of randomized operational decisions in adversarial classification settings. In *Proc. of AISTATS*, pages 599–607, 2015.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *Proc. of IEEE CVPR*, pages 152–159, 2014.

[19] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv:1611.02770*, 2016.

[20] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *Proc. of IEEE CVPR*, pages 1988–1995, 2009.

[21] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proc. of IEEE CVPR*, pages 5188–5196, 2015.

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

[23] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv:1605.07277*, 2016.

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

[25] 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 *Proc. of ACM CCS*, pages 1528–1540, 2016.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. of IEEE CVPR*, pages 1–9, 2015.

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[29] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proc. of IEEE CVPR*, pages 1288–1296, 2016.

[30] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

[31] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proc. of IEEE CVPR*, pages 1249–1258, 2016.

[32] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Proc. of IEEE ICPR*, pages 34–39, 2014.

[33] Shoou-I Yu, Yi Yang, and Alexander Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *Proc. of IEEE CVPR*, pages 3714–3720, 2013.

[34] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proc. of IEEE CVPR*, pages 5810–5818, 2017.

[35] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proc. of IEEE CVPR*, pages 7982–7991, 2019.

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

[37] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2017.

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