This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3199013

Towards Robust Person Re-identification by Defending Against Universal Attackers

Fengxiang Yang, Juanjuan Weng, Zhun Zhong, Hong Liu, Zheng Wang,
Zhiming Luo, Donglin Cao, Shaozi Li, Shin'ichi Satoh, Nicu Sebe

Abstract—Recent studies show that deep person re-identification (re-ID) models are vulnerable to adversarial examples, so it is critical to improving the robustness of re-ID models against attacks. To achieve this goal, we explore the strengths and weaknesses of existing re-ID models, *i.e.*, designing learning-based attacks and training robust models by defending against the learned attacks. The contributions of this paper are three-fold: First, we build a holistic attack-defense framework to study the relationship between the attack and defense for person re-ID. Second, we introduce a combinatorial adversarial attack that is adaptive to unseen domains and unseen model types. It consists of distortions in pixel and color space (*i.e.*, mimicking camera shifts). Third, we propose a novel virtual-guided meta-learning algorithm for our attack-defense system. We leverage a virtual dataset to conduct experiments under our meta-learning framework, which can explore the cross-domain constraints for enhancing the generalization of the attack and the robustness of the re-ID model. Comprehensive experiments on three large-scale re-ID benchmarks demonstrate that: 1) Our combinatorial attack is effective and highly universal in cross-model and cross-dataset scenarios; 2) Our meta-learning algorithm can be readily applied to different attack and defense approaches, which can reach consistent improvement; 3) The defense model trained on the learning-to-learn framework is robust to recent SOTA attacks that are not even used during training. Code is available at: https://github.com/WJJLL/Meta-Attack-Defense

1 Introduction

THE goal of person re-identification (re-ID) is to find pedestrians of interest in a surveillance system with non-overlapping cameras. It plays a crucial part in the construction of a smart and safe city, such as assisting law enforcement in the search for fugitive criminals or lost children. Thanks to the rapid development of deep neural networks [1], [2], [3], the leading person re-ID methods [4], [5], [6], [7] have achieved very high accuracies on various datasets. However, recent studies on adversarial attacks [8], [9], [10], [11] reveal the vulnerability of deep models, *i.e.*, a model may suffer catastrophic performance degradation by adding quasi-imperceptible perturbations to test images. This phenomenon also appears in the community of person re-ID [12], [13], [14], arousing the demand of improving the robustness of deep re-ID models.

> "What does not kill me makes me stronger."
>
> —*Friedrich Nietzsche, 1888*

Inspired by this aphorism, we aim to improve the robustness of re-ID models by defending against strong attacks. Specifically, we propose a holistic attack-defense framework, which involves two goals: (1) *learning a strong attack* to mislead re-ID models, and (2) *learning a robust re-ID model* through an effective defense algorithm with the assistance of the attack.

Goal 1: Learning Strong Attacker. Person re-ID is an open-set problem that is usually challenged with a cross-domain scenario. In this case, the testing samples have identities that are not included in the training set and may belong to a domain outside the training set. Therefore, a re-ID attack should have the ability to perturb samples from unseen identities and domains, as well as mislead various types of re-ID models. In this paper, we refer to this ability as *universality*. While recent works show the efficacy of attacking re-ID models, they still have limitations that restrict their universality. These shortcomings are generally manifested in three aspects. First, the works in [13], [14] learn perturbations using the test data (*i.e.*, query and gallery), which severely limits their universality, since the test data is seldom completely accessible in practice. Second, Wang *et al.* propose

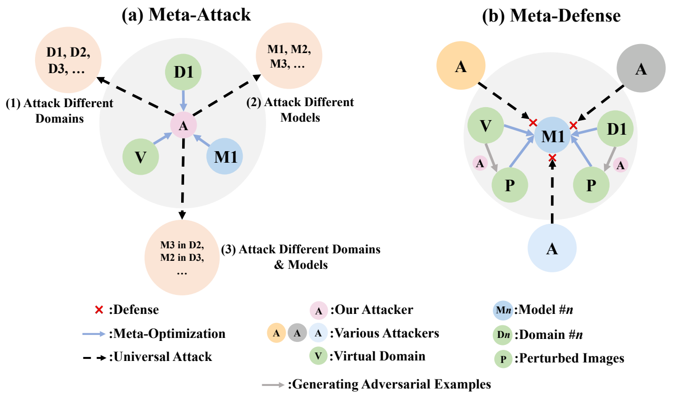


Fig. 1: Schematic illustration of our MetaAttack and MetaDefense. (a) MetaAttack has a good universality, which can mislead models trained on different domains with the same model, the same domain with different models, and different domains with different models. (b) MetaDefense is to learn a robust re-ID model that can defense various attacks achieved by different types of algorithms.

- * *The first two authors contributed equally.*
- *F. Yang, J. Weng, Z. Luo (corresponding author), D. Cao (corresponding author), and S. Li are with the Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China.*
- *Z. Zhong and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Trento 38100, Italy.*
- *H. Liu and S. Satoh are with the National Institute of Informatics, Tokyo 011455, Japan.*
- *Z. Wang is with the School of Computer Science, Wuhan University, Wuhan 430072, China.*

MisRank [12], which learns a generator from the training data and disturbs test samples without additional training. However, the generator is trained on one dataset and may be sensitive to unseen datasets with variation factors. Third, the re-ID has the domain-shift or camera-shift issue [15], [16], which severely impacts the model performance. Therefore, color-distribution variation across cameras is a common and important environmental factor.

In our attack-defense framework, we first propose a combinatorial attack method equipped with a meta-learning algorithm for training a powerful re-ID attack. This combinatorial attack consists of a functional color attack [17] and an additive attack [18], [19]. The former generates adversarial perturbation in a color-space by simulating camera-shift, which can adapt to various models. The latter disturbs several samples with a single universal perturbation, which can generalize well across datasets. Therefore, we can build a powerful combinatorial attack with good universality by using their mutual strengths.

Specifically, learning the attack that includes different variation factors, such as domain variation and pose variation, is an intuitive method to simulate the change of realistic scenarios during the training phase. However, obtaining such rich labeled data is difficult due to data collection and labeling costs, as well as data privacy concerns. The works in [20], [21], [22] have introduced several virtual datasets that automatically simulate many kinds of real-world scenarios and protect data privacy. However, directly using these virtual datasets during training may hamper the performance because of the large domain gap between real and virtual data. To solve this drawback, we propose a novel virtual-guided meta-learning scheme (MetaAttack), in which real data is served as meta-train and virtual data is served as meta-test. We encourage our MetaAttack method to capture more underlying variations by mimicking the cross-domain constraint through meta-learning, therefore enhancing attack generalization on unknown domains. As shown in Fig. 1 (a), our MetaAttack is trained on a real domain and a virtual domain with a specific model, which is flexible and suitable for different attack contexts including various datasets and models.

Goal 2: Learning Robust Re-ID Model. After studying the vulnerability of re-ID models, we further explore the defense system of learning a robust metric-preserving model that can resist those attacks. Recently, few studies [13], [14] have focused on learning defense models based on the adversarial training pipeline [8], [23]. However, they use the same source data for both attack and defense learning, which may cause a biased issue and can only deal with a specific attack. In practice, it is preferable to have a defense model that can resist various kinds of attacks. Interestingly, this generalization ability is similar to that of attacks, where we expect that our attack-defense system can generalize well to different domains, attacks, and environments.

To this end, we further present a virtual-guided meta-defense algorithm (MetaDefense) for training the defense model, which can improve the robustness of re-ID by using our combinatorial attack. As shown in Fig. 1 (b), MetaDefense can effectively defend against different re-ID attacks, in which additional attacks under cross-domain restrictions are simulated with the help of a virtual dataset. This encourages the model to have a good generalization capacity for resisting previously unknown adversarial data.

The main contributions of our work can be summarized as:

- We introduce a holistic attack-defense framework for person re-ID, which facilitates the investigation of the vulnerability of deep models and improves the robustness of deep models, with respect to adversarial samples.
- We design a combinatorial attack for person re-ID. The proposed method takes advantage of a functional color attack and the universal additive attack, which can effectively promote the universality of the attack.
- We propose a unified virtual-guided meta-learning algorithm, which can be applied to the training phases of both the attack and defense. With our method, the cross-domain constraint is explicitly injected into the optimization, effectively improving the universality of the attack and the robustness of the defense model.
- Extensive experimental results on three large-scale re-ID benchmarks demonstrate the effectiveness and strong universality of the proposed MetaAttack in both cross-model and cross-dataset scenarios. Meanwhile, our MetaAttack achieves state-of-the-art attack performance.
- The experiments also demonstrate that MetaDefense can resist various kinds of attacks by learning with a universal re-ID attack. The model trained by our MetaDefense achieves higher accuracies than the one learned by vanilla adversarial training.

Different from our previous work [24], this work further makes four significant contributions. (1) Our previous work only studied the problem of attack. Instead, this paper studies both attack and defense challenges by introducing a holistic attack-defense framework that helps us better understand the relationship between attack and defense. (2) We introduce a combinatorial attack by combining the functional color attack and additive attack. (3) We extend the virtual-guided meta-learning scheme to adversarial learning, which can effectively improve the robustness of the re-ID model. (4) We clarify the definition of universality for attacks and conduct comprehensive experiments to investigate the universality of different attacks. More importantly, we verify our motivation that a re-ID model could be robust to different types of attacks by defending against adversarial samples produced by a strong attack with good universality. We believe this could be a valuable direction for designing robust re-ID models in future studies.

2 Related Work

2.1 Adversarial Attacks

Szegedy *et al.* [9] demonstrated how adversarial examples may easily fool deep models by manipulating images with quasi-imperceptible perturbations. Subsequently, many white-box (*i.e.,* models and data are available) attack methods [19], [25], [26], [27], [28] have been proposed to generate more disruptive adversarial examples. [29] found that adversarial examples are transferable among different models, which then inspired a series of studies on black-box attacks [30], [31], [32]. However, these methods usually need to generate perturbations for each example individually, which is inefficient. Then, Moosavi-Dezfooli *et al.* [19] introduced a Universal Adversarial Perturbation (UAP) method, which uses a single image-agnostic perturbation to attack deep models. Besides, the UAP can be applied to black-box attacks. Wang *et al.* [10] design a new generative attack method using MCMC to generate adversarial examples more efficiently. Different from previous attacks on classification tasks, attacking re-ID [13], [14] or image retrieval systems [18], [33], [34] focuses on corrupting the pair-wise similarities in the training set. Bai *et al.* [13] proposed an attack scheme for re-ID models by extending various classification attacks to metric-attacks. This attack is achieved by pushing the feature of an input away from its intra-class features *(*i.e., *non-targeted attack)* or pulling towards a target from inter-class *(*i.e., *targeted attack)*. Bouniot *et al.* [14] propose a self-supervised attack method SMA to corrupt the accuracies of re-ID models without using any labels. However, these two methods have to optimize their attack during test time on one or more images from test set, which is not practical in real-world re-ID attacks. Li *et al.* [18] tried to attack the image retrieval systems, including re-ID, with a UAP by optimizing with label-wise, pair-wise and list-wise constraints. Wang *et al.* [12] proposed to attack re-ID models by learning a perturbation generator based on adversarial training. Although this generator is learned on a specific model and dataset, it can also be applied to attack other re-ID systems. In sum, existing attack methods can be mainly divided into two types: instance-level attacks [13], [14] and universal attacks [12], [18]. The former requires additional computation power during the test time to optimize the perturbation for each test image. In contrast, the latter directly applies the unique perturbation, which is learned during an offline training stage on training data, to attack all test images. Therefore, considering the advantages of low-cost, universality and ease of use, this paper focuses on the universal attack methods and does not compare with the instance-level attack methods when evaluating attack performance. In this paper, we design a combinatorial universal attack with the UAP and the color-space perturbation to achieve a more universal attack. In addition, we propose a meta-learning scheme with the assistance of a virtual data for the training of attack, which can further improve the universality of the learned attack.

2.2 Adversarial Defense

Adversarial defense aims to improve the robustness of models to resist adversarial samples. Existing adversarial defense methods can be mainly categorized into three types: (1) adversarial training [8], [23], [35], [36], [37], (2) defensive distillation [38], [39], [40], and (3) detection-based methods [41], [42]. Adversarial training is one of the popular defense schemes, which is formulated as a min-max game that injects adversarial examples into the training set and jointly trained with original samples [8], [23].

Defensive distillation learns a smoother model by model distillation, which introduces a smaller amplitude of gradients around input points [43]. Detection-based methods try to distinguish adversarial examples from benign ones by using the region information of images [41] or additional language descriptions [42]. Although most defense methods achieve great success, they can not be directly applied to re-ID due to the property of open-set [44] problems. Bai *et al.* [13] proposed the first defense model, which merges the original examples and adversarial examples to train a new defense reID model. Subsequently, Bouniot *et al.* [14] attempted to improve the defense accuracy with guide-sampling adversarial training. These works extend defense algorithms to the re-ID problem and achieve satisfactory performance. Nevertheless, they overlook the generalization ability for defending different attacks, which may limit their practicality in a real-world attack scenario. Different from previous works, we adopt additional virtual data and meta-learning algorithm to build underlying cross-domain constraints during training, which can effectively improve the robustness of the re-ID models.

2.3 Meta Learning

Meta-learning, also known as “learning to learn”, is initially introduced to train models that can learn new tasks rapidly with a few training examples. The popular meta-learning algorithms include optimizing-based [45], [46], [47], model-based [48], [49], and metric-based [50], [51] methods. The optimizing-based methods are to obtain an ideal model that can be fast adapted to new tasks with just a few steps of fine-tuning. Finn *et al.* [45] proposed model-agnostic meta-learning (MAML) to acquire the ideal weight by stimulating the learning process of new tasks with meta-test set. Subsequently, Reptile [46] sped up the learning process of MAML with a first-order approximation. For applications in attack and defense, Yin *et al.* [52] designed a defense algorithm with meta-learning to help model resist FGSM [8] attack. This work adopts training samples as meta-train and perturbed images as meta-test, which may suffer from the domain-shift issue brought by different attacks. However, [52] is designed for image classification and can not be directly employed in our re-ID task. Different from their work, we propose a virtual-guided meta-learning algorithm based on optimization, which can be used in both attack and defense schemes for re-ID. Our proposed meta-learning approach mainly focuses on improving the generalization ability of attacks and defense models, in which an attack is learned to mislead different models while a defense model is optimized to resist different types of attacks.

3 Attacking Re-ID Models

In the context of universal re-ReID attack, we are given a source dataset $\mathcal{S}$ and a model ϕ trained on $\mathcal{S}$. We aim to learn an attacker that can mislead different re-ID models on different domains. In general, these re-ID models are (1) the model ϕ (source-attack), (2) the same model trained on unseen datasets $\mathcal{T}$ (cross-dataset attack), (3) different model trained on $\mathcal{S}$ (cross-model attack), and (4) different model trained on unseen datasets $\mathcal{T}$ (cross-model & dataset attack). Note that, the unseen datasets and new models are not available during the learning of attack, making it hard to achieve a universal attack. In this section, we propose a novel combinatorial universal attack for person re-ID, which has a good universality under different attack scenarios. In addition, we introduce a virtual-guided meta-learning algorithm to further improve the universality of the proposed attack, which is achieved by simulating cross-domain constraints with an extra virtual dataset.

3.1 Universal Adversarial Attack

For the universal attack, the attacker should be learned on the training data and can be directly applied to unseen samples without further optimization. Moreover, this attack cannot access query images to optimize the perturbation. In this paper, we consider two types of attack that satisfy the requirements of universal attack, and integrate them into a combinatorial universal attack.

Additive Delta Attack. The additive attack is the most widely used adversarial attack method, which perturbs the input image by adding a small perturbation to each pixel. To learn a universal perturbation, we adopt universal perturbation attack (UAP) [18] to threaten re-ID systems, formulated by:

$$\widetilde{\boldsymbol{x}} = (\boldsymbol{x}_1 + \delta_1, \cdots, \boldsymbol{x}_n + \delta_n), \tag{1}$$

where $\boldsymbol{x}_i$ is the i-th pixel ($\boldsymbol{x}_i \in \mathbb{R}^3$ for RGB images) in $\boldsymbol{x}$ and δ_i is the i-th perturbation of $\boldsymbol{\delta}$. n denotes the number of pixels in

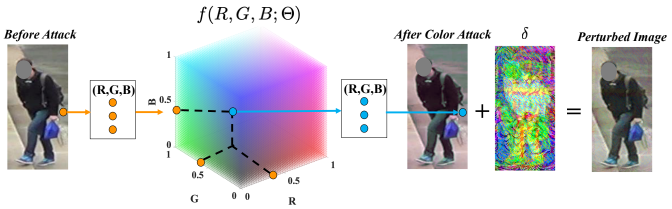


Fig. 2: The process of the proposed combinatorial universal attack. Given an image, we first perturb it by the functional color attack. Then, we produce the final adversarial example by adding the learned universal perturbation δ to the color-attacked image.

an image $\boldsymbol{x}$, and $\widetilde{\boldsymbol{x}}$ is the adversarial counterpart. Generally, the maximum change caused by $\boldsymbol{\delta}$ can not exceed ϵ-ball of the original image. The UAP can produce adversarial examples with only one single perturbation for all samples. This property enables UAP to naturally be applied to unseen samples that are from different identities. Since we use symbol delta to represent the perturbation, we call this attack method as an additive delta attack to distinguish it from the following functional color attack.

Functional Color Attack. Camera-shift or domain-shift is an important issue that can largely influence the retrieval performance of re-ID. In other words, re-ID models are sensitive to color changes caused by camera-shift or domain-shift [15], [16], [53]. To mimic the camera-shift or domain-shift for a better attack, we leverage the ReColorAdv [17] as the adversarial color attack. The ReColorAdv learns a color-transform function that disturbs the deep re-ID models by uniformly adjusting the color-space with a small perturbation. The color-transform function can be formulated as:

$$f : C \rightarrow C \quad s.t. \, |c_j - \widetilde{c}_j| < \epsilon, \;\; j = \{1, 2, 3\}, \tag{2}$$

where C is the color-space (RGB in our case) and c_j denotes the j-th channel of a pixel $\boldsymbol{x}_i \in \mathbb{R}^3$. In practice, $f(\cdot)$ is defined on a mesh grid in the color space and parameterized by a series of trainable parameters $\Theta \in R^{x \times y \times z}$. The transformation of other points in the color space is computed by trilinear interpolation from those mesh points.

For a sample $\boldsymbol{x}$, we obtain the corresponding color-transformed adversarial counterpart for each pixel in $\boldsymbol{x}$ by:

$$\widetilde{\boldsymbol{x}}_i = (\widetilde{c}_{i,1}, \widetilde{c}_{i,2}, \widetilde{c}_{i,3}) = f(c_{i,1}, c_{i,2}, c_{i,3}), \tag{3}$$

where $\widetilde{\boldsymbol{x}}_i$ is the i-th pixel for the perturbed image $\widetilde{\boldsymbol{x}}$, and $\{\widetilde{c}_{i,1}, \widetilde{c}_{i,2}, \widetilde{c}_{i,3}\}$ are corresponding RGB channels for the given perturbed pixel $\widetilde{\boldsymbol{x}_i}$. The functional color attack has two advantages: (1) it can change the overall color-distribution of samples with one color-transformer; (2) it is effective in attacking different models, since re-ID models commonly are sensitive to color changes (*i.e.*, domain/camera shift).

Combinatorial Attack. By combining the above two attacks, we then have the final universal adversarial attack to fool re-ID models, which is formulated as

$$\widetilde{\boldsymbol{x}} = f(\boldsymbol{x}) + \boldsymbol{\delta}, \tag{4}$$

where $f(\cdot)$ is the learned color-transformer and $\boldsymbol{\delta}$ is the learned universal perturbation.

Similar to previous methods [12], [18], [54], we perform an L_∞-bounded attack to generate perturbed images that is limited in the ϵ-ball of the original image. The overall attacking process is shown in Fig. 2. Given an input image, we first use the functional color attack $f(\cdot)$ to uniformly change its color-space. Then, we generate the final adversarial example by adding the learned perturbation $\boldsymbol{\delta}$ to the image. The proposed combinatorial attack can exploit the merits of both additive attack and functional color attack to achieve better universality. Note that, although we use two attack processes in our method, the overall ϵ is the same as using each individual attack.

3.2 Objective Function for Attack

To attack a re-ID model, it is crucial to disturb the pair-wise similarities of training samples. Therefore, different from attack methods in classification, we adopt the pair-wise constraint, which is implemented based on the triplet loss, to optimize the attack proposed in Sec. 3.1. In contrast to the triplet loss function for the normal training of re-ID model, we use a reverse version to mislead the original pairwise relations. Specifically, in a training batch, we aim at pushing an adversarial sample away from its hard-positive sample while pulling it close to its furthest negative example. The loss function for misleading the pair-wise relations is defined as:

$$L_{pair}(\widetilde{\boldsymbol{x}}, \boldsymbol{x}; \Theta, \delta) = \Big[||\phi(\boldsymbol{x}_n) - \phi(\widetilde{\boldsymbol{x}})||_2 - ||\phi(\boldsymbol{x}_p) - \phi(\widetilde{\boldsymbol{x}})||_2 + m \Big]_+, \tag{5}$$

where $[x]_+$ is the $\max(0, x)$ function, $\boldsymbol{x}_p$ and $\boldsymbol{x}_n$ are the hard-positive and furthest negative examples of $\boldsymbol{x}$ within the training batch, respectively. $\widetilde{\boldsymbol{x}}$ is the adversarial example of $\boldsymbol{x}$. $\phi(\cdot)$ extracts the feature for a given sample. $|| \cdot ||_2$ is the L2-norm function. m is the margin.

3.3 MetaAttack for Universal Re-ID Attacker

As mentioned in the previous sections, an effective way to improve the universality of attack is training with larger, more diverse data, so that the attack can capture more underlying variation factors. However, the abundant training data is hard to obtain in practice due to data privacy and labeling costs. Therefore, in this work, we propose to use a virtual dataset to enlarge the training data and help us learn a more universal attack. Furthermore, to overcome the potential performance degradation caused by the large domain shift between virtual and real images, we propose to optimize our attack under the meta-learning framework. The proposed algorithm is called **virtual-guided meta-learning**, which simulates the cross-domain constraint during training and encourages the attack to capture more underlying variations that are important for universality. Specifically, we first regard the real dataset as meta-train set $\mathcal{D}_{tr}$ and the virtual set as meta-test set $\mathcal{D}_{te}$, and then train the attack with the following three steps: *i.e.*, *Meta-Train*, *Meta-Test* and *Meta-Update*. Next, we will introduce these three processes in detail.

Step 1: Meta-train. In the first step, we aim to optimize the attack with meta-train set. We sample a mini-batch $\boldsymbol{x}_{tr}$ with N_b images from $\mathcal{D}_{tr}$ and generate their perturbed counterparts $\widetilde{\boldsymbol{x}}_{tr}$ to update our attack. The loss function utilized in the meta-train step can be formulated as:

$$L_{mtr}(\widetilde{\boldsymbol{x}}_{tr}, \boldsymbol{x}_{tr}; \Theta, \delta) = \frac{1}{N_b} \sum_{i=1}^{N_b} \Big[L_{pair}(\widetilde{\boldsymbol{x}}_{tr_i}, \boldsymbol{x}_{tr_i}; \Theta, \delta) \Big], \tag{6}$$

where Θ and δ are trainable parameters of our combinatorial attack. By optimizing the meta-train loss through SGD with momentum, we can obtain a new attack. This new attack is termed

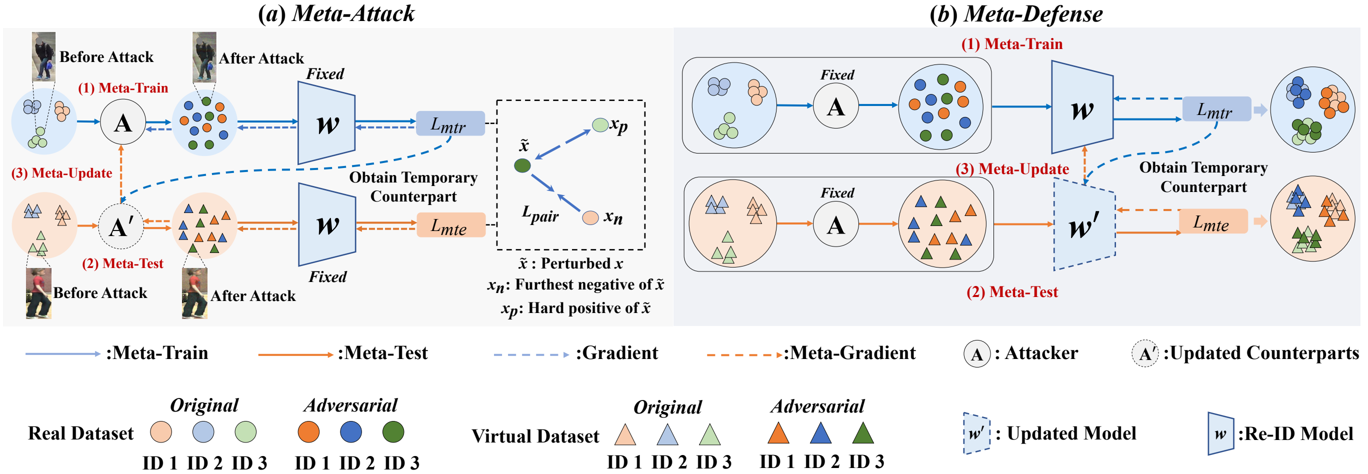


Fig. 3: The overall framework of the proposed "MetaAttack" and "MetaDefense". **Overall**: in both "MetaAttack" and "MetaDefense", we regard the real dataset as meta-train and the virtual dataset as meta-test. A meta-learning strategy is applied to learn the universal attack (MetaAttack) and the robust re-ID model (MetaDefense). **(a) MetaAttack**: (1) We first generate adversarial samples for the meta-train with the original attack and obtain a temporary attack with the meta-train loss L_{mtr}. (2) We then generate adversarial samples for the meta-test with the obtained temporary attack and calculate the meta-test loss L_{mte}. (3) The original attack is updated with the combination of L_{mtr} and L_{mte}. In this way, the proposed MetaAttack encourages the attack to capture more underlying variations across domains, and thus to have better universality. **(b) MetaDefense**: Similar to MetaAttack, we train the re-ID model with three steps. The difference is that we fix the attack and update the re-ID model with corresponding defense losses. With our MetaDefense, the re-ID model is enforced to be robust to unseen adversarial examples.

as temporary attack, which is parameterized by Θ' and δ' and is utilized in the next step.

Step 2: Meta-Test. In the meta-test step, we attempt to evaluate the temporary attack with meta-test images. Specifically, we sample a mini-batch $\boldsymbol{x}_{te}$ from $\mathcal{D}_{te}$ with N_b images and generate its perturbed counterparts $\widetilde{\boldsymbol{x}}_{te}$ by the temporary attack. Then, the meta-test loss is obtained by:

$$L_{mte}(\widetilde{\boldsymbol{x}}_{te}, \boldsymbol{x}_{te}; \Theta', \delta') = \frac{1}{N_b}\sum_{i=1}^{N_b}\Big[L_{pair}(\widetilde{\boldsymbol{x}}_{te_i}, \boldsymbol{x}_{te_i}; \Theta', \delta')\Big]. \tag{7}$$

Step 3: Meta-update. In the final step, the gradient from meta-train loss and meta-gradient from meta-test loss are aggregated to optimize parameters of the original attack. The loss function in this step can be formulated as:

$$L_{meta} = L_{mtr}(\widetilde{\boldsymbol{x}}_{tr}, \boldsymbol{x}_{tr}; \Theta, \delta) + L_{mte}(\widetilde{\boldsymbol{x}}_{tr}, \boldsymbol{x}_{tr}; \Theta', \delta'). \tag{8}$$

The former item aims to learn basic knowledge with meta-train, while the latter item aims to capture common factors across domains that are helpful in improving universality. It should be noted that although the meta-test loss is obtained from the temporary parameters Θ' and δ', the gradient *w.r.t.* original Θ and δ can also be obtained through chain rule. The gradient for optimizing Θ can be formulated as:

$$\frac{\partial L_{meta}}{\partial \Theta} = \frac{\partial L_{mtr}}{\partial \Theta} + \frac{\partial L_{mte}}{\partial \Theta'}\frac{\partial \Theta'}{\partial \Theta}. \tag{9}$$

The gradient of δ shares a similar formulation to Θ. The overall training process is listed in Alg. 1. In this work, we call the proposed virtual-guided meta-learning for combinatorial attack as "MetaAttack". The framework of the proposed MetaAttack is illustrated in Fig. 3-(a).

Algorithm 1 The Procedure of MetaAttack.

Inputs: Meta-train $\mathcal{D}_{tr}$ (source dataset $\mathcal{S}$), meta-test $\mathcal{D}_{te}$ (virtual dataset $\mathcal{P}$), batch size N_b, re-ID model trained on source domain $\mathcal{S}$, number of training epochs T, learning rate α

Outputs: Final attack F parameterized by Θ^* and δ^*.

```
Initialize $\Theta$ and $\delta$ with $\mathbf{0}$;
for $t$ in $T$ do
  repeat
    Sample mini-batches $\boldsymbol{x}_{tr}$ and $\boldsymbol{x}_{te}$ with $N_b$ images from $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$, respectively;
    Disturb $\boldsymbol{x}_{tr}$ and $\boldsymbol{x}_{te}$ to generate $\widetilde{\boldsymbol{x}}_{tr}$ and $\widetilde{\boldsymbol{x}}_{te}$;
    // Meta-train
    Compute meta-train loss $L_{mtr}$ with Eq. 6;
    Update $\Theta$ and $\delta$ via momentum SGD to obtain temporary $\Theta'$ and $\delta'$;
    // Meta-Test
    Compute meta-test loss $L_{mte}$ through Eq. 7;
    // Meta-update
    Compute final loss with Eq. 8;
    Update the original $\Theta$ and $\delta$ via momentum SGD;
  until $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ are enumerated;
end for
$\Theta^* \leftarrow \Theta, \delta^* \leftarrow \delta$;
Return the attack $F$ parameterized $\Theta^*$ and $\delta^*$;
```

4 Defending Re-ID Models

As a counter-strike to adversarial attacks, this section aims to learn a robust metric-preserving re-ID model that can resist these attacks. Previous re-ID defense methods [13], [14] follow the traditional adversarial training pipeline to improve robustness of the re-ID model, which jointly optimize the re-ID model with the original training images and their adversarial counterparts. This training strategy can lead the model to resist a specific attack that is used during defense learning. However, the learned model may fail to defend against other types of attacks that can not be seen during training. In practice, a robust model should have the ability to resist different types of attacks. This generalization ability is similar to that of the attack task. In attack and defense tasks, we hope that (1) the attack can mislead different models, and that (2) the defense model can be robust to different types of attacks. Taking the above considerations, we draw inspiration from the proposed "MetaAttack" and propose a virtual-guided meta-defense algorithm for defending re-ID model, which is called as "MetaDefense" (illustrated in Fig. 3-(b)). Next, we will introduce the basic loss functions for defense learning and the proposed "MetaDefense".

4.1 Basic Losses for Defense

We adopt triplet and cross-entropy losses to improve the robustness of the original re-ID model ϕ. The triplet loss used in our method is formulated as:

$$L_{tri}(\boldsymbol{x};\boldsymbol{w}) = \Big[||\phi(\boldsymbol{x}_p) - \phi(\boldsymbol{x})||_2 - ||\phi(\boldsymbol{x}_n) - \phi(\boldsymbol{x})||_2 + m\Big]_+ , \tag{10}$$

where $\boldsymbol{x}$ are images or their adversarial counterparts. $\boldsymbol{x}_p$ and $\boldsymbol{x}_n$ are corresponding hard-positive and hard-negative examples of $\boldsymbol{x}$. $\phi(\cdot)$ computes the feature for an input image.

The cross-entropy loss for the defense re-ID model is as follows:

$$L_{ce}(\boldsymbol{x};\boldsymbol{w}) = -\sum_{i=1}^{N_I} q_i \log p(\boldsymbol{x_i}), \tag{11}$$

where N_I is the number of identities in the training dataset. $p(\boldsymbol{x}_i)$ is the predicted probability of $\boldsymbol{x}_i$ belonging to identity i, and q_i is the i-th element of $\boldsymbol{x}_i$'s one-hot vector.

4.2 MetaDefense for Robust Re-ID Model

Similar to "MetaAttack", the proposed "MetaDefense" leverages an extra virtual dataset and meta-learning strategy to improve the robustness of the re-ID model. We regard the real dataset as the meta-train set $\mathcal{D}_{tr}$ and the virtual dataset $\mathcal{D}_{te}$ as the meta-test set. The overall defense scheme can be summarized into *Meta-Train*, *Meta-Test* and *Meta-Update* steps.

Meta-Train. Based on loss functions for adversarial training, we sample N_b images from the meta-train set $\mathcal{D}_{tr}$ and generate N_{adv} perturbed images with an attacker. Then we adopt the triplet and cross-entropy losses for meta-train:

$$\begin{aligned} L_{mtr}(\boldsymbol{x}_{tr}, \widetilde{\boldsymbol{x}}_{tr}; \boldsymbol{w}) = &\sum_{i=1}^{N_b} \Big[L_{tri}(\boldsymbol{x}_i;\boldsymbol{w}) + L_{ce}(\boldsymbol{x}_i;\boldsymbol{w})\Big] \\ &+ \sum_{j=1}^{N_{adv}} \Big[L_{tri}(\widetilde{\boldsymbol{x}}_j;\boldsymbol{w}) + L_{ce}(\widetilde{\boldsymbol{x}}_j;\boldsymbol{w})\Big], \end{aligned} \tag{12}$$

where $\boldsymbol{x}_{tr}$ and $\widetilde{\boldsymbol{x}}_{tr}$ are original and the perturbed images of meta-train set $\mathcal{D}_{tr}$. $\boldsymbol{w}$ are parameters of the current re-ID model. We update $\boldsymbol{w}$ with SGD optimizer to generate its updated version $\boldsymbol{w}'$ for further optimization. For the triplet loss, we select hard-positive and hard-negative samples of $\boldsymbol{x}_{tr}$ or $\widetilde{\boldsymbol{x}}_{tr}$ in the mini-batch that includes both original and perturbed images (*i.e.*, size $=N_b+N_{adv}$).

Algorithm 2 The Procedure of MetaDefense.

```
Inputs: Meta-train $\mathcal{D}_{tr}$ (source dataset $\mathcal{S}$), meta-test $\mathcal{D}_{te}$ (virtual dataset $\mathcal{P}$), number of training epochs $T$, batch size for original images $N_b$, batch size for adversarial examples $N_{adv}$, learning rate $\alpha$, attack $F$, pre-trained re-ID model $\phi$ parameterized by $\boldsymbol{w}$.
Outputs: metric-preserving re-ID model $\phi^*$ parameterized by $\boldsymbol{w}^*$.
for $t$ in $T$ do
    repeat
        Sample mini-batches $\boldsymbol{x}_{tr}$ and $\boldsymbol{x}_{te}$ with $N_b$ images from $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$, respectively;
        // Meta-Train
        Randomly select $N_{adv}$ examples from $\boldsymbol{x}_{tr}$ to generate adversarial examples $\widetilde{\boldsymbol{x}}_{tr}$ by attack $F$;
        Compute meta-train loss $L_{mtr}$ through Eq. 12;
        Obtain temporary model by $\boldsymbol{w}' = \boldsymbol{w} - \alpha\nabla_{\boldsymbol{w}} L_{mtr}$;
        // Meta-Test
        Randomly select $N_{adv}$ examples from $\boldsymbol{x}_{te}$ to generate adversarial examples $\widetilde{\boldsymbol{x}}_{te}$ by attack $F$;
        Compute meta-test loss $L_{mte}$ through Eq. 13;
        // Meta-Update
        Compute the final loss $L_{meta}$ with Eq. 14;
        Update the original model by $\boldsymbol{w} = \boldsymbol{w} - \alpha\nabla_{\boldsymbol{w}} L_{meta}$;
    until $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ are enumerated.
end for
$\boldsymbol{w}^* \leftarrow \boldsymbol{w}$
Return the robust re-ID model $\phi^*$ parameterized by $\boldsymbol{w}^*$;
```

Meta-Test. In this step, we use the meta-test set $\mathcal{D}_{te}$ to validate the robustness of the updated model, and to further improve the performance of the model with guidance from $\mathcal{D}_{te}$. Similar to meta-train, we sample N_b images $\boldsymbol{x}_{te}$ from $\mathcal{D}_{te}$ and randomly select N_{adv} examples to generate adversarial examples $\widetilde{\boldsymbol{x}}_{te}$. Then the loss is defined with the updated parameters $\boldsymbol{w}'$ to compute meta-test loss:

$$\begin{aligned} L_{mte}(\boldsymbol{x}_{te}, \widetilde{\boldsymbol{x}}_{te}; \boldsymbol{w}') = &\sum_{i=1}^{N_b} \Big[L_{tri}(\boldsymbol{x}_i;\boldsymbol{w}') + L_{ce}(\boldsymbol{x}_i;\boldsymbol{w}')\Big] \\ &+ \sum_{j=1}^{N_{adv}} \Big[L_{tri}(\widetilde{\boldsymbol{x}}_j;\boldsymbol{w}') + L_{ce}(\widetilde{\boldsymbol{x}}_j;\boldsymbol{w}')\Big], \end{aligned} \tag{13}$$

where $\boldsymbol{x}_{te}$ and $\widetilde{\boldsymbol{x}}_{te}$ are original and their perturbed counterparts, respectively.

Meta-Update. Based on the aforementioned meta-train and meta-test losses, the final loss for updating the original $\boldsymbol{w}$ is defined as follows:

$$L_{meta} = L_{mtr}(\boldsymbol{x}_{tr}, \widetilde{\boldsymbol{x}}_{tr}; \boldsymbol{w}) + L_{mte}(\boldsymbol{x}_{te}, \widetilde{\boldsymbol{x}}_{te}; \boldsymbol{w}'). \tag{14}$$

The overall MetaDefense training scheme is illustrated in Alg. 2.

5 Experiments on Attack

5.1 Experimental Setup

Evaluation Settings. To verify the effectiveness of the attack, we conduct experiments under different scenarios. Specifically, we first train the attack on the source training set (or combined with a virtual dataset) with a re-ID model, and then test the learned attack on four settings: 1) source attack, 2) cross-model attack, 3) cross-dataset attack and 4) cross-dataset&model attack, depending on the testing samples and models. Note that, in our attack settings,

TABLE 1: Performance of source attack with three methods (color attack, delta attack and combinatorial attack). "None": before attack, "Color": color attack. "Delta": delta attack, "Col.+Del.": combinatorial attack

Model	Attack	Meta	Duke		Market	
			mAP	rank-1	mAP	rank-1
IDE	None	×	66.7	80.9	78.2	88.7
	Color	×	18.5	26.1	28.5	36.7
	Delta	×	2.5	3.0	2.6	2.0
	Col.+Del.	×	1.6	1.9	2.3	2.3
	Color	✓	12.1	16.3	18.1	24.2
	Delta	✓	1.1	1.2	0.6	0.1
	Col.+Del.	✓	**0.6**	**0.6**	**0.4**	**0.1**
PCB	None	×	68.0	84.1	76.7	91.3
	Color	×	19.8	30.0	26.2	36.3
	Delta	×	13.5	19.6	27.5	36.5
	Col.+Del.	×	7.4	9.9	13.1	16.5
	Color	✓	13.5	20.7	23.9	34.2
	Delta	✓	8.2	12.7	14.7	21.0
	Col.+Del.	✓	**3.7**	**4.7**	**4.6**	**4.9**

TABLE 2: Performance of cross-model attack with three methods. "Color": color attack. "Delta": delta attack. "Col.+Del.": our combinatorial attack. "IDE → PCB": attack PCB model with the attacker trained with IDE model. "PCB → IDE": attack IDE model with the attacker trained with PCB model.

Training Data	Methods	Meta	IDE → PCB		PCB→IDE	
			mAP	rank-1	mAP	rank-1
Market	None	×	76.7	91.3	78.2	88.7
	Color	×	29.1	41.3	26.4	33.4
	Delta	×	55.6	77.0	33.9	46.1
	Col.+Del.	×	22.9	32.9	15.0	20.0
	Color	✓	**22.4**	**32.7**	29.9	40.4
	Delta	✓	60.4	80.9	21.2	28.9
	Col.+Del.	✓	26.6	38.2	**4.4**	**4.6**
Duke	None	×	68.0	84.1	66.7	80.9
	Color	×	21.2	30.8	18.4	25.3
	Delta	×	45.2	64.5	20.7	27.9
	Col.+Del.	×	21.0	32.3	11.2	14.1
	Color	✓	**10.7**	**15.8**	20.0	29.1
	Delta	✓	49.9	70.3	16.8	22.7
	Col.+Del.	✓	15.8	24.3	**7.0**	**9.6**

the query and gallery sets are not allowed to train the perturbations or attack models, which is an important constraint in the universal attack. The source attack belongs to white-box setting, since the re-ID model is entirely available. The cross-dataset & cross-model attacks belong to black-box setting, since we do not know the structure and parameters of the targeted re-ID model.

- **Source attack**: We perturb the query set of the source domain and use the obtained adversarial examples to fool the source model that is used to train the attack.
- **Cross-model attack**: We perturb the query set of the source domain and use the obtained adversarial examples to fool another model optimized with source training data.
- **Cross-dataset attack**: We perturb the query set of target domains and use the obtained adversarial examples to fool the re-ID models that are trained on these target domains.
- **Cross-dataset&model attack**: We perturb the query set of the target domains and use the obtained adversarial examples to fool another re-ID model optimized on other domains.

Evaluation Protocol. The overall performance is evaluated with the mean average precision (mAP) and rank-1 accuracy, **lower** mAP and rank-1 denote better attack performance. We use Euclidean distance for the evaluations of all re-ID attacks, unless otherwise specified.

Datasets. In our experiments, we use three real-world datasets as the source data or the test data, including Market-1501 (Market) [55], DukeMTMC-reID (Duke) [56], [57] and MSMT-17 (MSMT) [58]. *Market* contains $1,501$ identities ($32,668$ images) taken by six cameras, of which 750 identities ($12,936$ images) are used for training and the other 751 identities for evaluation. *Duke* consists of $36,411$ images of $1,404$ identities obtained from eight cameras. *MSMT* covers $126,441$ images from $4,101$ identities captured by fifteen cameras. For the virtual-guided meta-learning algorithm, we use PersonX-456, a subset of PersonX [20] dataset with rich background variations, as the virtual data in our experiments. PersonX-456 is composed of $39,852$ images, which are obtained from 410 persons and is initially designed to explore the impact of viewpoints on re-ID systems. The viewpoints of one identity in PersonX are sampled at an interval of $10°$.

Implementation Details. We use IDE [59], PCB [5], and AGW [60] as our re-ID models, of which the feature extractors are constructed based on ResNet-50 [1], and train universal adversarial attacks separately. We control the mesh-grid resolution of the parameters Θ over the learned color-transformer $f(\cdot)$ and set its dimension $x = y = z = 25$. The other hyper-parameters in our experiments are as follows: the batch size $N_b = 64$, the number of training epoch $T = 10$, margin $m = 0.5$, and the learning rate $\alpha = \epsilon/10$. We apply a stochastic gradient descent with momentum [18] to update the universal attack. We use the model of the last epoch for evaluation. By default, all experiments based on attacks are performed by L_∞-bounded attacks with $\epsilon = 8/255$, where ϵ is the upper bound for the change of each pixel ($\| \widetilde{\boldsymbol{x}} - \boldsymbol{x} \|_\infty \leq \epsilon$). To satisfy this constraint, we adopt the same strategy as [18], *i.e.*, clipping the generated perturbation after the optimization. For simplicity, we regard "Color" as the functional color attack, "Delta" as the additive delta attack and "Col.+Del." as the proposed combinatorial attack. We utilize PersonX as the meta-test set for our meta-learning approach, unless otherwise noted.

5.2 Evaluations on Source Attack

In this section, we assess the performance of the proposed attack method in the source attack setting. We mainly evaluate two important factors in our method, *i.e.*, (1) the types of attacks (color attack, delta attack, and combinatorial attack), (2) the effectiveness of meta-learning. For experiments without meta-learning, both meta-optimization and an additional virtual dataset were not used in the training. From the results reported in Tab. 1, we have the following three observations.

(1) Both delta attack and color attack can substantially mislead the accuracies of re-ID models. Firstly, we notice that the most widely used additive attack can successfully corrupt the deep Re-ID models. For example, the additive delta attack (w/o meta-learning) successfully reduces the mAP of the IDE model on Duke from 66.7% to 2.5%. On the other hand, the color attack can also significantly reduce the performance of re-ID models, such as the mAP drops to 18.5% and 19.8% for the IDE and

TABLE 3: Performance of cross-dataset attack with three methods (color attack, delta attack and combinatorial attack). "None": before attack. "Color": color attack. "Delta": additive delta attack. "Col.+Del.": the proposed combinatorial attack. "Dataset A → Dataset B": using the attack trained on Dataset A to mislead the inferring on Dataset B.

Model	Methods	Meta	Duke → Market		Duke → MSMT		Market → Duke		Market → MSMT	
			mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
IDE	None	×	78.2	88.7	42.3	69.8	66.7	80.9	42.3	69.8
	Color	×	37.3	50.0	18.0	49.4	31.2	42.8	22.7	57.8
	Delta	×	3.1	2.6	4.4	10.8	13.9	18.8	4.1	10.2
	Col.+Del.	×	2.4	2.0	2.4	5.5	11.4	16.2	3.0	7.1
	Color	✓	23.6	31.5	18.9	50.7	15.4	21.0	21.1	56.0
	Delta	✓	**1.8**	**1.2**	3.1	7.3	5.5	7.2	2.5	5.7
	Col.+Del.	✓	2.2	1.4	**1.5**	**3.0**	**2.4**	**2.6**	**1.2**	**2.4**
PCB	None	×	76.7	91.3	50.8	88.9	68.0	84.1	50.8	88.9
	Color	×	26.3	38.5	8.8	21.2	23.0	33.7	15.9	38.7
	Delta	×	29.4	41.7	4.0	8.3	22.2	34.2	5.3	11.8
	Col.+Del.	×	16.6	24.5	2.4	4.8	14.0	20.9	3.9	8.7
	Color	✓	29.5	41.6	13.4	34.7	17.3	26.5	11.5	29.3
	Delta	✓	18.3	25.2	2.4	4.8	16.2	24.2	3.0	6.4
	Col.+Del.	✓	**11.5**	**16.1**	**2.1**	**4.4**	**4.9**	**7.2**	**2.1**	**4.7**

TABLE 4: The results of cross-model & dataset attack. For experiments with "IDE → PCB" and "Market → Duke", we optimize the attack with the IDE model trained on Market and then corrupt another PCB model trained on Duke.

Model	Methods	Meta	Market → Duke		Duke → Market	
			mAP	rank-1	mAP	rank-1
IDE ↓ PCB	None	×	68.0	84.1	76.7	91.3
	MisRank	×	57.6	73.3	39.7	53.4
	Color	×	23.5	34.9	33.8	49.4
	Delta	×	50.0	69.0	57.5	79.2
	Col.+Del.	×	20.6	30.8	32.0	48.0
	Color	✓	**13.6**	**19.5**	22.6	32.7
	Delta	✓	51.7	72.3	59.2	80.2
	Col.+Del.	✓	17.0	27.3	**22.1**	**31.3**
PCB ↓ IDE	None	×	66.7	80.9	78.2	88.7
	MisRank	×	45.9	60.8	49.5	63.9
	Color	×	30.7	41.4	37.4	49.5
	Delta	×	31.1	44.2	33.4	45.4
	Col.+Del.	×	21.3	29.1	17.9	22.7
	Color	✓	16.9	23.9	36.3	47.4
	Delta	✓	21.2	28.9	27.1	37.2
	Col.+Del.	✓	**10.6**	**14.2**	**13.3**	**18.8**

PCB on Duke (w/o meta-learning). These results prove that our motivation of using color transformation to mimic camera shift for attacking re-IDs is feasible.

(2) Combining the color attack and delta attack can further improve the attack performance. The results reveal that deep re-ID models are vulnerable to the proposed combinatorial attack on different models and datasets, regardless of using or not using meta-learning. For example, the combinatorial attack (w/o meta-learning) successfully fools the IDE model trained on Duke from 66.7% to 1.6%, which is better than the individual color attack or delta attack. After using the meta-learning, the combinatorial attack further reduces the mAP to 0.6%, which also outperforms individual color or delta attack with meta-learning.

(3) Meta-learning plays an essential role in improving attacking performance. By comparing the results of each attacking method trained w/ and w/o meta-learning, we find that attacks endowed with meta-learning achieve better performance. For example, in the scenario of IDE trained on the Duke, the mAP of the color attack w/ and w/o meta-learning is 12.1% vs. 18.5%, the delta attack is 1.1% vs. 2.5%, and the "Col. + Del." attack is 0.6% vs. 1.6%. Accordingly, these results indicate the effectiveness of the proposed meta-learning strategy that consistently improves the performance of different types of attacks.

5.3 Evaluations on Cross-model Attack

Different from the source attack in Tab. 1, we further evaluate the universality of the proposed attack method under the cross-model attack setting (the target model is different), and report the results in Tab. 2. From the Tab., we have the following observations.

(1) The color attack shows better consistency than the delta attack in this setting. In Tab. 2, we find that the color attack learned by one model can consistently decrease the accuracy of the other different model for all the cases. For the case of "IDE→PCB" in the delta attack, we observe the learned universal delta perturbations (on Market and Duke) struggle to corrupt the PCB model. For example, the rank-1 only drops from 91.3% to 77.0% (w/o meta) and 80.9% (with meta) on the Market, respectively.

(2) The combination of color and delta is beneficial to improving the performance. For the case of "PCB→IDE" on both Market and Duke, the combinatorial attack can achieve better performance than the individual color and delta attack. Although for the setting of "IDE→PCB", the combinatorial attack fails to outperform the color attack. It still can largely alleviate the significant negative influence brought by the delta attack.

(3) The meta-learning is generally effective for improving cross-model attack. For most of the cases, we find a substantial improvement of attacking performance after using the proposed meta-optimization procedure. However, we also notice that the attacking of color attack learned by the PCB model is decreased on both Market and Duke datasets.

5.4 Evaluations on Cross-dataset Attack

Concurrent with the cross-model attack, we also evaluate the universality of our attack methods under the setting of cross-dataset attack. In this section, we conduct extensive experiments on three large-scale benchmarks and list the results in Tab. 3. Similar to the cross-model attack, we can find that the combinatorial attack

TABLE 5: Comparison between the proposed attack (MetaAttack with combinatorial attack) and the state-of-the-arts (MisRank [12] and UAP-Retrieval [18]) under the settings of source attack and cross-dataset attack. All experiments are conducted with $\epsilon = 8$ in default. "Market → MSMT": cross-dataset attack that uses perturbation trained on Market to attack MSMT model. "Before Attack": re-ID accuracies without attack. For cross-dataset attack, we report the re-ID accuracies of target model without attack in "Before Attack".

Model	Methods	Market		Duke		Market → MSMT		Duke → MSMT	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
IDE [59]	Before Attack	78.2	88.7	66.7	80.9	42.3	69.8	42.3	69.8
	MisRank	19.2	25.9	19.9	26.9	11.1	28.5	11.7	30.0
	MisRank+PersonX	24.4	32.7	29.1	39.0	12.4	31.0	20.9	55.8
	UAP-Retrieval	3.6	4.5	4.2	9.9	5.3	13.9	5.5	15.4
	UAP w/ SMA	1.5	3.4	2.0	2.4	3.2	8.2	2.6	5.7
	Ours	**0.4**	**0.1**	**0.6**	**0.6**	**1.2**	**2.4**	**1.5**	**3.0**
PCB [5]	Before Attack	76.7	91.3	68.0	84.1	50.8	88.9	50.8	88.9
	MisRank	36.3	49.2	39.8	56.1	14.4	28.5	21.1	47.7
	MisRank+PersonX	35.3	47.8	40.5	56.2	18.8	39.6	18.8	39.6
	UAP-Retrieval	10.7	15.1	14.3	20.3	4.3	8.9	4.4	9.1
	UAP w/ SMA	8.3	12.2	6.7	10.4	4.6	9.5	5.8	10.7
	Ours	**4.6**	**4.9**	**3.7**	**4.7**	**2.1**	**4.7**	**2.1**	**4.4**
AGW [60]	Before Attack	88.2	95.3	77.0	87.2	49.3	68.3	49.3	68.3
	MisRank	57.1	72.5	31.6	42.5	17.5	33.7	24.9	46.1
	MisRank+PersonX	61.3	73.2	34.2	44.1	19.2	34.2	27.0	47.8
	MisRank (ϵ=16)	13.2	18.1	8.9	11.2	4.7	8.5	9.8	18.4
	UAP-Retrieval	3.7	3.9	2.0	2.7	9.4	16.7	6.4	10.4
	UAP w/ SMA	2.2	4.1	3.1	4.4	6.2	9.7	5.2	8.7
	Ours	**0.5**	**0.3**	**0.6**	**0.4**	**0.4**	**2.5**	**0.5**	**2.5**

and the meta-learning are both effective in this Cross-dataset attack setting. On the other hand, we notice that the delta attack has better performance than the color attack in this setting. For example, the learned delta perturbation on the Duke with IDE (w/o meta) successfully deteriorates the mAP on the Market and MSMT to 3.1% and 4.4%, respectively. These results suggest that the learned delta perturbations are less sensitive to the domain shift of different re-ID datasets.

5.5 Evaluations on Cross-model & Dataset Attack

We also conduct experiments to evaluate cross-model & dataset attack performance, which is the most challenging and practical form of real-world re-ID attack. The results are shown in Tab. 4. Based on the results, we have similar conclusions to previous sections. *i.e.*, meta-learning and the color information play essential roles in the cross-model & dataset attack. Besides, the proposed combinatorial attack takes advantage of the individual color and delta attacks. Moreover, we also evaluate the performance of the MisRank [12] in this setting and find that the MisRank only slightly decreases the accuracy of different models. These results demonstrate the universality of the proposed methods for attacking unknown re-ID models.

5.6 Comparison with State-of-the-Arts

We compare the proposed MetaAttack with two state-of-the-art approaches, *i.e.*, MisRank [12] and UAP-Retrieval [18]. It should be noted that we do not take [13] and [14] into our comparison. In detail, both [13] and [14] will use images from testing set to optimize their perturbations, which is different from our experimental setting and is not a practical scenario in real world re-ID attack. Although SMA [14] can not be directly applied to our attack scheme, we try our best to make it usable in our experiments by applying the self-supervised loss in SMA to UAP. Concretely, we reduce the pair-wise cosine similarities between the features of the original training images and their perturbed counterparts to attack re-ID systems. This variant is called "UAP w/ SMA". Results are evaluated on the settings of source attack and cross-dataset attack. All experiments are conducted with $\epsilon = 8$. For source attack, we evaluate the attacks on Market and Duke. For cross-dataset attack, we train the attacks on Market (or Duke) and test them on MSMT without any modification. Apart from IDE and PCB, we also evaluate the recently published re-ID model (AGW [60]), which achieves state-of-the-art performance on most benchmarks. We show the comparison results in Tab. 5 and make the following conclusions. (1) Our attack largely outperforms the MisRank and UAP-Retrieval in all settings, showing that we achieve new state-of-the-art attack performance for re-ID. For example, when attacking the IDE model with $\epsilon = 8$, our attack achieves mAP = 0.4% on Market and mAP = 0.6% on Duke, respectively. These results are 3.2% and 3.6% lower (better) than the best competitor (UAP-Retrieval). Similar superiority of our attack can be found in other settings (source attack and cross-dataset attack on IDE, PCB and AGW). (2) "UAP w/ SMA" achieves better attack performance than UAP-Retrieval but is still inferior to our method. (3) Directly using the virtual dataset (PersonX) as the training data fails to bring consistent improvement for MisRank. For example, when additionally using the virtual data, the mAP on Market (source attack, IDE model) is increased from 19.2% (MisRank) to 24.4% (MisRank+PersonX), leading to a worse attack performance. This indicates that using a virtual dataset is not trivial in attack re-ID.

5.7 Evaluations on Using Different Meta-Test Sets

In this section, we use four different datasets to evaluate the characters of choosing additional data for meta-learning. They are (1) Real Dataset (MSMT [58] in our experiments) (2) PersonX [20] (3) UnrealPerson [21] (4) RandPerson [22]. The last three virtual

TABLE 6: Ablation study on meta-learning and meta-test sets. We adopt MSMT-17 (Real) [58], RandPerson (RP) [22], PersonX (PX) [20], and UnrealPerson (UP) [21] as meta-test set to study the characters of choosing additional data for meta-learning. "Meta": meta-learning, "Extra": extra data, "Color": color attack, "Delta": additive delta attack, "Col.+Del.": the proposed combinatorial attack.

No.	Extra	Meta	Delta				Color				Col.+Del.			
			Market → Duke		Market		Market → Duke		Market		Market → Duke		Market	
			mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
0	×	×	13.9	18.8	2.6	2.0	31.2	42.8	28.5	36.7	11.4	16.2	2.3	2.3
1	Real	×	12.7	17.0	2.8	2.3	32.0	43.1	27.5	36.2	10.8	15.2	2.3	2.0
2	PX	×	12.5	17.3	1.9	1.1	33.8	45.9	33.9	44.6	10.7	14.7	1.6	1.0
3	RP	×	13.9	19.5	2.2	1.5	25.8	34.6	29.9	40.7	9.8	13.8	2.2	2.3
4	UP	×	11.9	16.3	1.7	0.9	23.5	31.9	27.6	36.3	9.1	12.3	1.7	1.4
5	Real	✓	6.8	8.7	**0.5**	**0.1**	19.2	26.4	20.9	28.7	3.6	4.0	0.4	0.1
6	PX	✓	**5.5**	**7.2**	0.6	**0.1**	15.4	21.0	18.1	24.2	**2.4**	**2.6**	0.4	0.1
7	RP	✓	8.2	10.3	**0.5**	0.2	14.7	19.6	16.8	21.4	5.7	7.5	**0.3**	**0.0**
8	UP	✓	7.1	9.0	**0.5**	**0.1**	**13.6**	**18.4**	**16.7**	**20.4**	4.6	5.5	**0.3**	0.1

TABLE 7: Comparison of using different training datasets. The combinatorial attack (Color+Delta) is applied.

Methods	Market		Duke	
	mAP	rank-1	mAP	rank-1
Real Only (R)	2.3	2.3	1.6	1.9
PersonX Only (PX)	10.8	13.2	11.1	14.8
MetaAttack (R+PX)	**0.4**	**0.1**	**0.6**	**0.6**

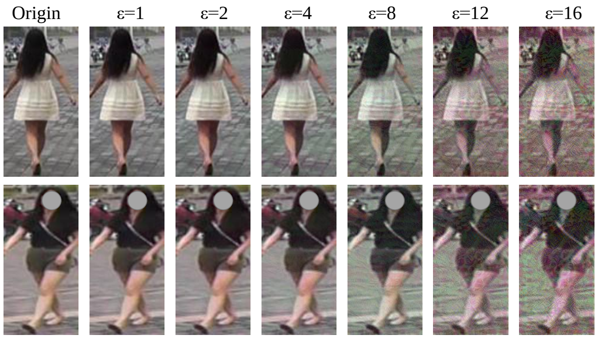


Fig. 4: Examples of corrupted queries with different ϵ.

datasets are designed for different purposes and have different properties. As mentioned in Sec. 5.1, *PersonX* has balanced viewpoints and can be utilized to explore the impact of viewpoints on re-ID systems. *RandPerson* is composed of $8,000$ identities captured by 19 cameras, which is designed to improve the quality of synthesized data with diversified backgrounds, style of clothes, and occlusion. *UnrealPerson* is another large-scale virtual dataset, which contains $120,000$ images of $1,200$ IDs captured by 34 virtual cameras and is proposed to enrich the details of virtual pedestrians. The experimental results are listed in Tab. 6. We can make the following observations.

(1) Directly combining source data with the additional dataset for joint-optimization achieves limited improvements. By comparing *No. 0* and *No. 1 - No. 4* in Tab. 6, we notice that three attacks yield limited improvements on corrupting re-ID models. The most significant improvement occurs in the cross-dataset color attack (Market→Duke), which reduces the mAP score from 31.2% (w/o extra data) to 23.5% (w/ extra data) with the assistance of UnrealPerson. However, this improvement is not substantial when compared with those trained by meta-optimization (*No. 5 - No. 8* in Tab. 6).

(2) Meta-learning is essential for improving the performance of attack. By comparing the four rows w/o meta-learning (*No. 1 - No. 4*) and the last four rows w/ meta-learning (*No. 5 - No. 8*), we observe that attacks optimized with meta-learning always have better performance than those without meta-learning for both extra virtual and extra real data. These results demonstrate the importance of using meta-optimization for learning more powerful universal attacks.

(3) Viewpoint is an important factor for training powerful attacks. The RandPerson, UnrealPerson and MSMT have much more training samples than PersonX. However, as shown in Tab. 6, the attacks optimized with PersonX achieve competitive results with other attacks, regardless of the amount of training data. Therefore, attacks may be more powerful when optimized with dataset that have balanced viewpoints.

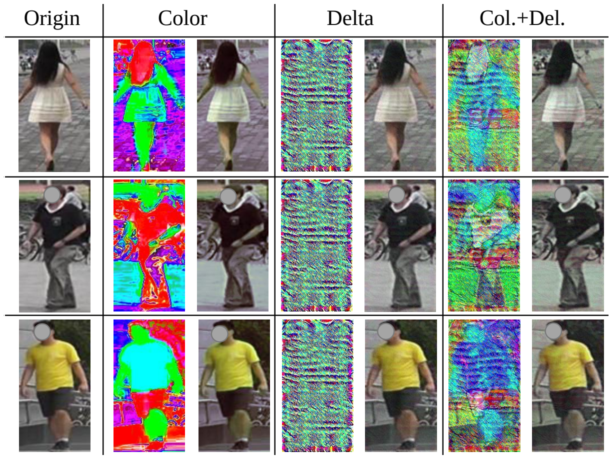


Fig. 5: Examples of the perturbations and corrupted queries for three attacks trained with our method. We use IDE as the training method and Market as the source data.

In Tab. 7, we also show the results of only using the virtual PersonX as the training data. We can observe that only using the PersonX achieves largely worse attack performance than the method of using the real source data and our MetaAttack method. This indicates the importance of the real data and further verifies the advantage of the proposed meta-learning.

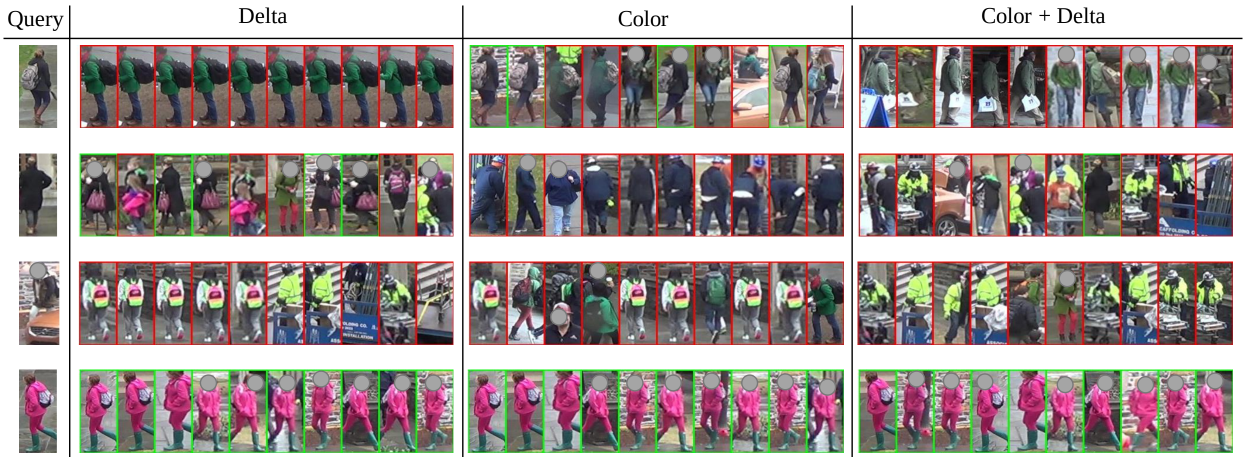


Fig. 6: Visualization of ranking lists under cross-dataset attack. We train the adversarial perturbation with PCB model trained on Market and attack the PCB model trained on Duke. The queries are evaluated on the Duke dataset.

TABLE 8: Sensitivity analysis of ϵ.

Col.+Del.	Market		Duke	
	mAP	rank-1	mAP	rank-1
ϵ=1	70.5	86.1	60.6	75.8
ϵ=2	47.3	62.0	43.8	57.8
ϵ=4	7.5	8.3	13.8	19.0
ϵ=8	0.4	0.1	0.6	0.6
ϵ=12	0.2	0.0	0.2	0.2
ϵ=16	0.2	0.0	0.2	0.0

5.8 Sensitivity Analysis

In Tab. 8, we analyze the impact of ϵ by changing the value of it between 1 and 16. Experiments are evaluated on our method (combinatorial attack with meta-learning) for the setting of source attack. Clearly, a higher value of ϵ can easily damage re-ID accuracies. However, in practice, we should ensure that the added perturbation is quasi-imperceptible and thus should not assign a value that is too large to ϵ. In Fig. 4, we visualize some perturbed queries with different values of ϵ. We can perceive changes when ϵ is larger than 8, e.g. ϵ=12. Based on this observation, we suggest setting ϵ to 8 for the attacking re-ID systems, which can produce acceptable attack results with quasi-imperceptible changes.

5.9 Image Quality of Adversarial Examples

In general, the generated adversarial images should be indistinguishable from real images. We then estimate the SSIM [61] scores between the real images and their perturbed counterparts. SSIM is a popular metric for evaluating image quality, a large SSIM score indicates better quality and less distortion. In Tab. 9, we compare the SSIM scores of different attacks, including functional color attack, additive delta attack, combinatorial universal attack, and MisRank [12]. The first three attacks are learned with meta-learning. For fair comparison, we set $\epsilon = 8$ for all attacks. We can find that all of our attacks achieve higher SSIM scores than MisRank, showing the advantage of our method in terms of the image quality. On the other hand, by comparing our methods, the additive delta attack produces a higher SSIM score than the functional color attack and the combinatorial attack.

TABLE 9: Comparison of SSIM scores between different attack methods. We set ϵ=8 for all attacks. The “Color”, “Delta” and “Col.+Del.” attacks are trained with our MetaAttack.

Dataset	Color	Delta	Col.+Del.	MisRank
Duke	0.2116	**0.2141**	0.2126	0.1985
Market	0.1904	**0.1997**	0.1921	0.1889

TABLE 10: Attack results of using two virtual datasets. ϵ=8 is used. UP:UnrealPerson, PX:PersonX.

Method	Meta Train	Meta Test	Duke		Market	
			mAP	rank-1	mAP	rank-1
Before Attack	-	-	66.7	80.9	78.2	88.7
Ours	UP	PX	61.4	68.9	59.6	62.5
Ours	Market	PX	2.4	2.6	0.4	0.1
Ours	Duke	PX	0.6	0.6	2.2	1.4

5.10 Attack without Real Data

The utilization of the virtual dataset can partially address the privacy issue in re-ID. To explore the possibility of fully addressing the privacy concern, we conduct an experiment by training without real images. Specifically, we use UnrealPerson [21] as the meta-train and PersonX [20] as meta-test. Results in Tab. 10 show that only using virtual data achieves largely worse attack results than using both real and virtual data. This demonstrates the importance of real data in achieving good attack performance. Although we do not completely address the data privacy issue, this paper has made a non-trivial step to solve this problem. With our method, we can use publicly available datasets and virtual datasets to train universal re-ID attacks and learn robust re-ID models. This avoids the usage of data in unseen scenarios and thus largely prevents data privacy in real-world applications where the testing process commonly requires higher privacy.

5.11 Visualization

Perturbation and Corrupted Samples. In Fig. 5, we visualize the perturbations and adversarial examples of different attacks (color, delta, and color+delta). All attacks are trained with our MetaAttack. We can find that: (1) the functional color attack generates different perturbations for images and changes the overall color-distributions of the images; (2) the additive delta attack generates adversarial examples with the same perturbation;

TABLE 11: Cross-dataset attack on DeepChange ($\epsilon = 8$). Before Attack: Performance of target model before being attacked.

	Market→DeepChange		Duke→DeepChange	
	mAP	rank-1	mAP	rank-1
Before Attack	9.2	28.4	9.2	28.4
UAP	0.8	1.1	1.0	2.4
Ours	1.3	2.6	1.2	3.2

TABLE 12: Results of using DeepChange as meta-test ($\epsilon = 8$).

Meta-Test	Market→Duke		Duke→Market	
	mAP	rank-1	mAP	rank-1
PersonX	2.4	2.6	2.2	1.4
DeepChange	13.7	18.6	14.7	19.2

and (3) the combinatorial attack produces adversarial examples that have the properties of both color and delta attacks while does not bring perceptible visual changes.

Ranking Lists under Cross-dataset Attack. We further visualize the query results (ranking lists) under the setting of cross-dataset attack. Specifically, the three attacks ("Color", "Delta" and "Col.+Del.") are trained with the PCB model on Market. We visualize the query results on the Duke dataset in Fig. 6. The first two rows visualize samples with correct retrieval results under either the delta or color attacks. In the third row, the query image with partial occlusion is successfully corrupted by all three attacks. Moreover, we provide a sample that can have correct retrievals for three attacks in the last row of Fig. 6. We notice that pink clothes can effectively neglect the influence of perturbations.

5.12 Evaluation on Long-Term Person re-ID

Currently, the long-term re-id with clothes changes becomes a very hot topic in the re-ID community [62], [63]. To explore the potential of our MetaAttack in long-term re-ID, we conduct several experiments on the DeepChange dataset [63].

First, we use the proposed MetaAttack, which adopts the Market dataset or Duke dataset as the source domain, to attack the re-ID model trained on DeepChange. The results are shown in Tab. 11. We can find that our method successfully reduces mAP and rank-1 accuracy but fails to outperform the UAP [18] attack. The main reason may be that the simulation on meta-test (PersonX) does not contain instances of changing clothes, hindering our method achieves further improvement under clothes changing context. On the other hand, clothes changing indeed can be regarded as a type of physical attack [64], [65], which however is very different from traditional imperceptible re-ID attack. Therefore, some conclusions in the traditional re-ID attack may not be applicable for the scenario of clothes changing.

Second, we train our MetaAttack by using DeepChange as the meta-test set. The results are reported in Tab. 12. We can find that, using DeepChange as the meta-test achieves worse attack performance compared to using the virtual PersonX. The main reason is similar to that of cross-dataset attack (in Tab. 11). That is, the DeepChange dataset includes conspicuous changes of clothes for the same person, which may not bring the same benefit as the traditional re-ID dataset during our meta-learning process. Since the model trained on long-term re-ID may be better robust to clothes changing, one possible attack solution for long-term re-ID could be enforcing the attack model to add perturbations to clothes-unrelated regions, such as face and background. We would like to study this interesting but challenging problem in the future.

6 Experiments on Defense

6.1 Experimental Setup

In the defense experiments, we utilize IDE to train the metric-preserving model "M". During optimization, the adversarial counterparts of the training data are generated by the MetaAttack as described in Sec. 3, where the used re-ID model for the MetaAttack is IDE (default) or PCB. Moreover, we use PersonX-456 [20] and its perturbed data as the meta-test. During the evaluation, we apply seen and unseen attacks to disturb the query of the source dataset to test the defense performance of the metric-preserving model "M". For other parameters in defense, we set the sample rate $N_b = 64$ and $N_{adv} = 0.5 \times N_b$ in Eq. 12. The learning rate α of Adam optimizer is initialized to 0.0003, and the metric-preserving model "M" is trained for 85 epochs. We use the model of the last epoch for evaluation. The defense experiments are conducted on Market and Duke datasets and evaluated by the mAP and rank-1, where higher numbers indicate better defense performance.

6.2 Robustness to Different Types of Attacks

To evaluate the metric defense performance, we adopt different universal adversarial attacks to disrupt the metric-preserving model "M". In Tab. 13, we compare the results of the normally trained model and defending model "M" against different attacks from the IDE model on both the Market and Duke. The included attacks are MisRank [12], UAP-Retrieve [18], SMA [14] and our proposed attacks (Col., Del., and Col.+Del.). We adopt Euclidean distance for most of the evaluations of attack methods, except SMA that uses cosine similarity. This is because SMA achieves better performance when using cosine similarity.. **Note that, the adversarial samples used for optimizing the "M" are with the same value of** ϵ, and the re-ID model used for optimization is the one used for disturbing the query images. From Tab. 13, we have the following conclusions.

(1) When testing the original clean query set, the metric-preserving models have a similar performance compared with normally trained models, as shown in the first row of Market and Duke.

(2) When purely testing the adversarial version of query images, the performance of metric-preserving models guarding against the attacks in our paper (delta attack, color attack, and combinatorial attack) is significantly increased. For instance, on the Market, the delta attack "Delta" fools normally trained models, which sharply reduces the rank-1 from 88.7% to 0.1%. While the metric-preserving models "M(Delta)" can increase rank-1 from 0.1% to 61.6% by improving the robustness of the model.

(3) Our defense method is also capable of helping the re-ID model survive from the state-of-the-art instance-level attacks like SMA [14], MisRank [12] and UAP-Retrieve [18], which indicates the effectiveness of our method. In detail, SMA can easily reduce the rank-1 of a Market re-ID model from 88.7 to 0.0. After utilizing our method to train the metric-preserving model "M(Col.+Del.)", the rank-1 has been recovered to 65.4% and other metric-preserving models can also improve the robustness against SMA (rank-1=34.6% for "M(Color)" and rank-1=60.4% for "M(Delta)").

(4) Our defense models learned by strong attacks can resist unseen universal attacks, in which the defense model does not know the type of attacks in advance. For example, the metric-preserving models "M(Col.+Del.)" can achieve a distinct effect

TABLE 13: Performance of metric-preserving models trained with our virtual-guided meta-learning. “Normal” denotes normally trained models, “Color”: color attack. “Delta”: additive delta attack. “Col.+Del.”: the proposed combinatorial attack. “M(Color)” means a metric-preserving model trained by source training data and their adversarial counterparts under color attack with $\epsilon = 8$. In all experiments, the adversaries of training data under “M(·)” are generated by MetaAttack and IDE model.

Datasets	Attack Methods	Normal		M(Color)		M(Delta)		M(Col.+Del.)	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
Market	Original	78.2	88.7	74.5	88.7	70.0	86.8	75.5	88.9
	Color	18.1	24.2	40.0	54.3	43.4	59.5	41.6	55.1
	Delta	0.6	0.1	19.1	24.1	46.5	61.6	46.7	62.4
	Col.+Del.	0.4	0.1	7.6	8.7	25.9	35.6	31.7	43.1
	MisRank [12]	19.2	25.9	71.7	86.7	64.5	83.3	72.1	86.3
	SMA (Cosine) [14]	0.2	0.0	25.4	34.6	45.7	60.4	50.2	65.4
	UAP-Retrieval [18]	3.6	4.5	25.3	33.8	43.6	58.3	49.3	64.5
Duke	Original	66.7	80.9	62.0	77.8	57.0	74.6	56.2	73.2
	Color	12.1	16.3	37.5	52.2	34.7	51.2	36.7	53.8
	Delta	1.1	1.2	19.4	27.5	33.1	47.1	39.6	56.1
	Col.+Del.	0.6	0.6	8.2	11.3	14.4	19.9	24.9	37.8
	MisRank [12]	19.9	26.9	59.4	76.3	52.6	71.1	53.9	72.6
	SMA (Cosine) [14]	0.2	0.2	23.6	34.7	30.2	44.6	35.3	49.2
	UAP-Retrieval [18]	4.2	9.9	21.6	30.6	33.5	48.8	39.0	55.9

TABLE 14: Comparison with state-of-the-arts. We use different attack methods (MisRank [12], SMA [14], and UAP-Retrieve [18]) to attack metric-preserving models. We use $\epsilon = 8$ for each attack method and the optimization of each metric-preserving model trained by adversarial examples with $\epsilon = 8$.

Datasets	Attacks	M(SMA)		M(Col.+Del.)	
		mAP	rank-1	mAP	rank-1
Market	MisRank	69.4	83.8	72.1	86.3
	SMA (Cosine)	10.4	15.3	50.2	]65.4
	UAP-Retrieval	42.6	62.3	49.3	64.5
Duke	MisRank	46.7	66.9	53.9	72.6
	SMA (Cosine)	8.2	14.6	35.3	49.2
	UAP-Retrieval	24.9	39.0	39.0	55.9

of resisting the other types of attacks (color, delta, MisRank [12] and UAP [18]). On Market, the rank-1 increases from 0.1% to 62.4% against “Delta” and from 24.2% to 55.1% against “Color” for the “M(Col.+Del.)”, respectively. Notably, we find that all our defense models (M(Color), M(Delta), and M(Col.+Del.) can resist the MisRank and UAP attack because an obvious improvement can be seen after re-training.

6.3 Comparison with State-of-the-Arts

We also compare our method with state-of-the-art re-ID defense works with $\epsilon = 8$ and IDE model in Tab. 14. Based on the results, we find that our method outperforms state-of-the-art defense method SMA [14]. For example, the metric-preserving model “M(Col.+Del.)” can recover the mAP score to 50.2% when attacked by SMA on Market. However, the metric-preserving IDE model optimized with SMA can only recover the mAP to 10.4%. These results fully demonstrate the effectiveness of our method.

6.4 Further Experiments

We further explore our defense model of resisting unseen attacks from other three aspects, *i.e.*, 1) defense from the attacks of different models, 2) defense from the non-meta attacks, and 3) defense from the attacks of different values of ϵ.

Defense from the Attack of Different Models. In Tab. 15, we evaluate the performance of using our metric-preserving model “M(Col.+Del.)” to resist perturbations learned by different attack methods and re-ID models. The re-ID models include IDE, PCB and AGW. The attack methods include MisRank [12] and our meta-attacks. When using IDE or PCB as the model for generating perturbations, our defense model “M(Col.+Del.)” can clearly improve the mAP and rank-1 accuracy for all settings. For example, on Market, the defense model “M(Col.+Del.)” can increase the mAP from 36.3% to 74.0% for resisting MisRank based on the PCB model. When generating perturbations by AGW, our defense model “M(Col.+Del.)” can consistently improve the performance under all settings except resisting MisRank on Duke. These results suggest that our adversarial training paradigm with virtual-guided meta-learning can improve the robustness for dealing attacks that are not seen during training.

Defense from the Non-Meta Attacks. We also explore our defense model for resisting non-meta attacks. In general, by comparing the performance of resisting the non-meta attacks and meta attacks in Tab. 15, we can find that the defense model commonly has a slightly higher performance in defending against non-meta attacks. Besides, as demonstrated in the attack experiments, the meta attacks usually have a strong attack ability than those non-meta attacks. Therefore, these results mean that the defending model learned with a strong attack is robust to weak attacks.

Defense from the Attacks of Different values of ϵ. To fully evaluate the robustness of defense models, we employ attack models with different values of ϵ to attack the metric-preserving model “M(Col.+Del.)”. From Tab. 16, we observe following phenomena. **(1)** The defense model can achieve dramatic improvement in resisting attacks trained by a lower epsilon. For example, the mAP can increase from 8.3% to 68.6%, when “M(Col.+Del.)” faces adversarial examples generated by the combinatorial attack with $\epsilon = 4$ on Market. **(2)** When further escalating the perturbation of adversarial samples to $\epsilon = 12$, our MetaDefense only can slightly improve the accuracy in resisting the combinatorial attack, such as the mAP only reaches 4.3%. **(3)** The defense model is robust for resisting MisRank attack even with $\epsilon = 16$. Compared to the clean case, the mAP only decreases from 78.2% to 67.1%, and 66.7% to 55.7% on Market and Duke, respectively.

These experimental results further confirm that our defense method can enhance the intrinsic robustness of models, especially

TABLE 15: Performance of metric-preserving model against attacks generated by different models. “M(Col.+Del.)” means a metric-preserving model trained by the original clean training data and the adversarial examples generated by Alg. 1 based on IDE model.

Attack Models	Meta	Attack Methods	Market				Duke			
			Normal		M(Col.+Del.)		Normal		M(Col.+Del.)	
			mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
None	×	None	78.2	88.7	75.5	88.9	66.7	80.9	56.2	73.2
IDE [59]	×	MisRank	19.2	25.9	72.1	86.3	19.9	26.9	54.7	73.4
		Color	28.5	36.7	48.9	62.4	18.5	26.1	36.5	53.7
		Delta	2.6	2.0	46.4	61.5	2.5	3.0	39.4	55.8
		Col.+Del.	2.3	2.3	22.0	28.0	1.6	1.9	23.5	35.3
	✓	Color	18.1	24.2	41.6	55.1	12.1	16.3	36.7	53.8
		Delta	0.6	0.1	46.7	62.4	1.1	1.2	39.6	56.1
		Col.+Del.	0.4	0.1	31.7	43.1	0.6	0.6	24.9	37.8
PCB [5]	×	MisRank	36.3	49.2	74.0	88.1	39.8	56.1	53.9	72.6
		Color	26.4	33.4	52.1	67.0	18.4	25.3	34.1	50.7
		Delta	33.9	46.1	53.0	67.4	20.7	27.9	28.9	41.8
		Col.+Del.	15.9	20.0	33.9	45.4	11.2	14.1	18.5	27.3
	✓	Color	29.9	40.4	44.7	58.1	20.2	29.1	34.1	50.8
		Delta	21.2	28.9	47.9	61.5	16.8	22.7	28.8	41.6
		Col.+Del.	4.4	4.6	16.9	21.2	7.0	9.6	18.5	27.4
AGW [60]	×	MisRank	61.4	74.4	70.4	85.5	63.6	77.8	57.2	74.5
		Color	37.6	47.7	47.1	61.1	28.3	38.0	32.2	46.3
		Delta	8.4	10.2	39.1	52.1	2.5	3.2	35.0	49.3
		Col.+Del.	5.8	7.1	30.3	40.5	0.8	0.9	25.6	36.8
	✓	Color	27.6	35.7	35.4	45.8	17.6	23.9	31.3	45.5
		Delta	1.6	1.6	33.8	44.8	1.9	2.4	37.7	53.3
		Col.+Del.	0.5	0.3	19.6	25.5	0.6	0.4	27.3	40.2

TABLE 16: Performance of metric-preserving model against MetaAttacks of different values of ϵ. “M(Col.+Del.)” denotes a metric-preserving model trained by original clean training data and the adversarial version of training data under Alg. 1 with $\epsilon = 8$.

Attack Methods	ϵ	Market				Duke			
		Normal		M(Col.+Del.)		Normal		M(Col.+Del.)	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
None	\	78.2	88.7	75.5	88.9	66.7	80.9	56.2	73.2
Col.+Del.	ϵ=4	7.5	8.3	53.1	68.6	13.8	19.0	40.1	56.7
	ϵ=8	0.4	0.1	31.7	43.1	1.1	1.0	24.9	37.8
	ϵ=12	0.2	0.0	4.3	4.5	0.2	0.0	9.2	13.4
MisRank	ϵ=8	19.2	25.9	73.1	86.3	19.9	26.9	54.7	73.4
	ϵ=16	4.4	3.7	67.1	82.9	2.7	3.3	55.7	73.2

TABLE 17: Ablation study on the proposed MetaDefense algorithm. We use our universal attack to train the metric-preserving model.

No.	Duke		Market		Virtual data	Meta
	mAP	rank-1	mAP	rank-1	PersonX	Learning
1	5.5	7.4	10.2	16.8	×	×
2	16.2	22.7	16.4	19.8	×	✓
3	16.7	21.5	21.3	30.3	✓	×
4	24.9	37.8	31.7	43.1	✓	✓

in resisting different types of adversarial attacks.

6.5 Ablation Study

Evaluation on Virtual-Guided Meta-Learning. To evaluate the contributions of different components, we gradually add the virtual dataset and meta-learning into our metric-preserving model. The training protocols for four settings in Tab. 17 are as follows: the *No.1* is the baseline which is trained by using only source data and its adversarial counterpart. In *No.2*, we leverage another real dataset (MSMT) and its adversarial counterpart as the meta-test. In *No.3*, the model is trained with mixed data (source and PersonX) and their adversarial counterparts. The *No.4* is our full defense training protocol with both virtual data and meta-learning.

From Tab. 17, we can have the following findings. **(1)** The comparison between *No.1* and *No.2* suggests that using the real data as the meta-test has a positive effect on improving the robustness of the metric preserving model. **(2)** From the comparison of *No.1* vs. *No.3*, we can find that introducing virtual data into defense training can significantly improve the metric defense performance. **(3)** Comparing the *No.3* vs. *No.4*, we can observe a further improvement of mAP and rank-1 on both Duke and Market. These results indicate the mutual benefit of leveraging the virtual data in meta-learning to increase the robustness of the metric-preserving model.

Evaluation on the Losses for MetaDefense. In Tab. 18, we investigate the impact of different loss functions for our MetaDefense, including triplet loss and cross-entropy loss. We can find that the triplet loss achieves higher defending performance than the cross-entropy loss on both datasets. In addition, by combining these two loss functions, the defending performance can be further improved.

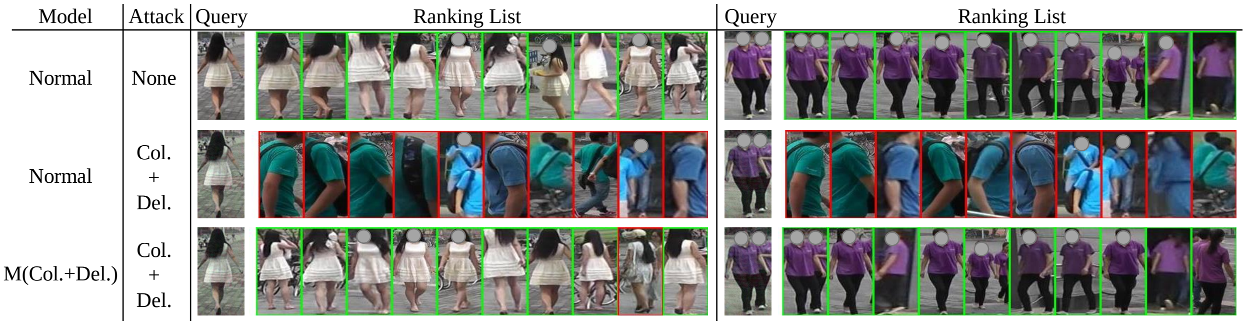


Fig. 7: Visualization of ranking lists for two query images under attack "Col.+Del." disturbing a normally trained model "Normal" and the proposed metric-preserving model "M(Col.+Del.)" .

TABLE 18: Effect of loss functions for the proposed MetaDefense. Our MetaAttack is used to train the metric-preserving model.

Loss	Market		Duke	
	mAP	rank-1	mAP	rank-1
L_{ce}	15.4	26.4	3.1	5.3
L_{tri}	26.1	32.7	15.3	22.9
$L_{tri} + L_{ce}$	31.7	43.1	24.9	37.8

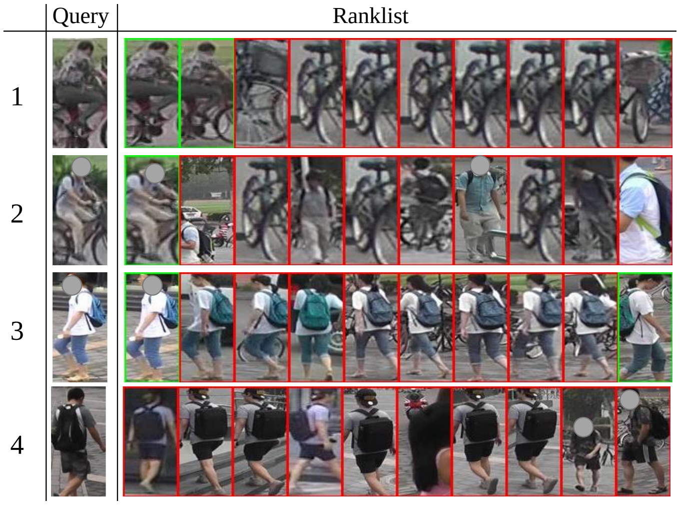


Fig. 8: Visualization of incorrect ranking lists for query images under "M(Col.+Del.)" against "Col.+Del.".

6.6 Visualization

Ranking Lists in MetaDefense. We visualize the ranking list of two query images in Fig. 7. The first row is the ranking list of a normally trained model for retrieving non-attacked samples. The second and third rows represent the combinatorial attack "Col.+Del." that misleads the normal-trained model "Normal" and the proposed metric-preserving model "M(Col.+Del.)", respectively. From the second row, we can find that the proposed attack method can disturb the whole ranking list of the queries. In the third row, our metric-preserving model can correct the ranking list for the attacked query.

Destructive Queries in MetaDefense. Our MetaDefense can largely resist other attack methods, while it has certain limitations in combating the combinatorial attack. In Fig 8, we further visualize some destructive query images and their ranking lists under "M(Col.+Del.)" against the MetaAttack of "Col.+Del.". From the results, we may conjecture that the query image with extra objects, such as a bag or bicycle, is not easy to defend.

TABLE 19: Analysis of computational cost.

Method		Training Time (s / epoch)
Our Attack	w/ meta	63.0
	w/o meta	22.0
Our Defense	w/ meta	271.0
	w/o meta	31.0

6.7 Computational Cost

In Table 19, we compare the computational cost of training models with or without the proposed meta-learning. When using meta-learning, the training time will be increased by 3 times and 9 times for the attack and defense model, respectively. Since we need to update the whole re-ID model during meta-defense, the increased training time is higher than that of meta-attack. Note that our meta-learning approach does not increase the testing time.

7 Conclusion

Deep person re-identification (re-ID) systems are vulnerable to adversarial attacks, raising the security risk of real applications. In this study, we present a holistic attack-defense framework to investigate the relationship between attack and defense in person re-ID. Specifically, we propose a universal adversarial attack to fool the re-ID, which combines the functional color attack and additive adversarial attack. Experiments show the universality of the proposed universal attack. Moreover, we design a novel virtual-guided meta defense algorithm, which can be applied to the learning of both attack and defense. Experiments verify that our meta-learning can effectively improve the universality of the attack and the robustness of the re-ID model. Importantly, we find that re-ID models learned by defending strong adversarial examples that have a good universality could be robust to different attacks. This observation may inspire us to develop robust re-ID models by learning with stronger (universality) attacks in the future.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017.

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

[4] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM MM*, 2018.

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

[6] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proc. AAAI*, 2019.

[7] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, X. Sun, and W. Zheng, "Aware loss with angular regularization for person re-identification," in *Proc. AAAI*, 2020.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. ICLR*, 2014.

[10] H. Wang, G. Li, X. Liu, and L. Lin, "A hamiltonian monte carlo method for probabilistic adversarial attack and learning," *IEEE TPAMI*, 2020.

[11] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal adversarial attack on attention and the resulting dataset damagenet," *IEEE TPAMI*, 2020.

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

[13] S. Bai, Y. Li, Y. Zhou, Q. Li, and P. H. Torr, "Adversarial metric attack and defense for person re-identification," *IEEE TPAMI*, 2020.

[14] Q. Bouniot, R. Audigier, and A. Loesch, "Vulnerability of person re-identification models to metric adversarial attacks," in *Proc. CVPRW*, 2020.

[15] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. CVPR*, 2018.

[16] F. Yang, Z. Zhong, Z. Luo, Y. Cai, S. Li, and S. Nicu, "Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification," in *Proc. CVPR*, 2021.

[17] C. Laidlaw and S. Feizi, "Functional adversarial attacks," in *Proc. NeurIPS*, 2019.

[18] J. Li, R. Ji, H. Liu, X. Hong, Y. Gao, and Q. Tian, "Universal perturbation attack against image retrieval," in *Proc. ICCV*, 2019.

[19] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. CVPR*, 2017.

[20] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. CVPR*, 2019.

[21] T. Zhang, L. Xie, L. Wei, Z. Zhuang, Y. Zhang, B. Li, and Q. Tian, "Unrealperson: An adaptive pipeline towards costless person re-identification," in *Proc. CVPR*, 2021.

[22] Y. Wang, S. Liao, and L. Shao, "Surpassing real-world source training data: Random 3d characters for generalizable person re-identification," in *Proc. ACM MM*, 2020.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018.

[24] F. Yang, Z. Zhong, H. Liu, Z. Wang, Z. Luo, S. Li, N. Sebe, and S. Satoh, "Learning to attack real-world models for person re-identification via virtual-guided meta-learning." in *Proc. AAAI*, 2021.

[25] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proc. CVPR*, 2016.

[26] Y. Fan, B. Wu, T. Li, Y. Zhang, M. Li, Z. Li, and Y. Yang, "Sparse adversarial attack via perturbation factorization," in *Proc. ECCV*, 2020.

[27] N. Akhtar, M. Jalwana, M. Bennamoun, and A. S. Mian, "Attack to fool and explain deep networks," *IEEE TPAMI*, 2021.

[28] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. ICML*, 2020.

[29] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM ASIAACS*, 2017.

[30] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proc. CVPR*, 2018.

[31] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu, "Dast: Data-free substitute training for adversarial attacks," in *Proc. CVPR*, 2020.

[32] Z. Zhao, Z. Liu, and M. Larson, "Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter," in *Proc. BMVC*, 2020.

[33] X. Li, J. Li, Y. Chen, S. Ye, Y. He, S. Wang, H. Su, and H. Xue, "Qair: Practical query-efficient black-box attacks for image retrieval," in *Proc. CVPR*, 2021.

[34] M. Zhou, Z. Niu, L. Wang, Q. Zhang, and G. Hua, "Adversarial ranking attack and defense," in *Proc. ECCV*, 2020.

[35] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "Stylized adversarial defense," *IEEE TPAMI*, 2020.

[36] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proc. ICML*, 2020.

[37] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in *Proc. ICLR*, 2021.

[38] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE S&P*, 2016.

[39] M. Soll, T. Hinz, S. Magg, and S. Wermter, "Evaluating defensive distillation for defending text processing neural networks against adversarial examples," in *Proc. ICANN*, 2019.

[40] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein, "Adversarially robust distillation," in *Proc. AAAI*, 2020.

[41] S. Li, S. Zhu, S. Paul, A. Roy-Chowdhury, C. Song, S. Krishnamurthy, A. Swami, and K. S. Chan, "Connecting the dots: Detecting adversarial perturbations using context inconsistency," in *Proc. ECCV*, 2020.

[42] M. Yin, S. Li, Z. Cai, C. Song, M. S. Asif, A. K. Roy-Chowdhury, and S. V. Krishnamurthy, "Exploiting multi-object relationships for detecting adversarial attacks in complex scenes," in *Proc. ICCV*, 2021.

[43] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NeurIPSW*, 2015.

[44] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *Proc. ICCV*, 2017.

[45] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017.

[46] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, 2018.

[47] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, "Meta-learning with implicit gradients," in *Proc. NeurIPS*, 2019.

[48] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. ICML*, 2016.

[49] T. Munkhdalai and H. Yu, "Meta networks," *Proc. ICML*, 2017.

[50] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. NeurIPS*, 2017.

[51] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. CVPR*, 2018.

[52] C. Yin, J. Tang, Z. Xu, and Y. Wang, "Adversarial meta-learning," *arXiv preprint arXiv:1806.03316*, 2018.

[53] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE TPAMI*, 2020.

[54] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE TPAMI*, 2018.

[55] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015.

[56] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, 2016.

[57] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proc. ICCV*, 2017.

[58] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018.

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

[60] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE TPAMI*, 2021.

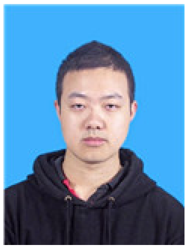
[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004.

[62] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Long-term cloth-changing person re-identification," in *Proc. ACCV*, 2020.

[63] P. Xu and X. Zhu, "Deepchange: A long-term person re-identification benchmark," *arXiv preprint arXiv:2105.14685*, 2021.

[64] Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi, "advpattern: physical-world attacks on deep person re-identification via adversarially transformable patterns," in *Proc. ICCV*, 2019.

[65] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *Proc. ECCV*, 2020.

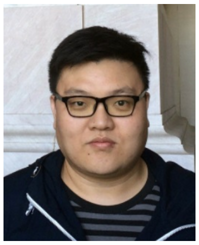


Fengxiang Yang is currently a Ph.D. candidate at Xiamen University, China. He received his master degree in Pattern Recognition & Intelligent System from Xiamen University, China, in 2020. His research interests include person re-identification and domain adaptation.

Juanjuan Weng is currently a Ph.D. candidate at Xiamen University, China. She received her bachelor's degree in Computer Science and Technology from Henan University of Science and Technology, China, in 2019. Her research interests include person re-identification and deep adversarial learning.

Zhun Zhong received the Ph.D. Degree in Computer Science and Technology from Xiamen University, China, in 2019. He was also a joint Ph.D. student at University of Technology Sydney. He is currently a Fellowship Researcher in University of Trento. His research interests include person re-identification and domain adaptation.

Hong Liu obtained his Ph.D. degree from Xiamen University, China, in 2020. He is now a JSPS Fellowship researcher at the National Institute of Informatics, Japan. His research interests include large-scale image retrieval and deep adversarial learning. He has published about 20+ papers in top journals and conferences. He was awarded the JSPS International Fellowship, Outstanding Doctoral Dissertation Award of the China Society of Image and Graphics, and Top-100 Chinese New Stars in Artificial Intelligence by Baidu Scholar.



Zheng Wang received the B.S., M.S., and Ph.D. degrees from Wuhan University in 2006, 2008, 2017, respectively. He was a JSPS Fellowship Researcher at the National Institute of Informatics, Japan, and a Project Assistant Professor at The University of Tokyo, Japan. He is currently a Professor at Wuhan University, China. His research interests include person re-identification and instance search.

Zhiming Luo received the B.S. degree from the Cognitive Science Department, Xiamen University, Xiamen, China, in 2011; the Ph.D. degree in computer science with Xiamen University and University of Sherbrooke, Sherbrooke, QC, Canada, in 2017. His research interests include surveillance video analytic, computer vision, and machine learning.

Donglin Cao received the B.S. degree and M.S. degree from Xiamen University, and the Ph.D. degree from Chinese Academy of Sciences. He is currently an Assistant Professor in the Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen, China. His research interest is cross-media information retrieval.

Shaozi Li received the B.S. degree from Hunan University, and the M.S. degree from Xi'an Jiaotong University, and the Ph.D. degree from National University of Defense Technology. He is currently a full professor in the Department of Artificial Intelligence, Xiamen University, the Vice Director of Technical Committee on Collaborative Computing of CCF, the Vice Director of the Fujian Association of Artificial Intelligence. He is also the senior Member of IEEE, ACM and China Computer Federation (CCF). His research interests cover Artificial Intelligence and Its Applications, Moving Objects Detection and Recognition, Machine Learning, Computer Vision, Multimedia Information Retrieval, etc.

Shin'ichi Satoh received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. He has been a Full Professor with the National Institute of Informatics, Tokyo, Japan, since 2004. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997. His current research interests include image processing, video content analysis, and multimedia databases.

Nicu Sebe is Professor in the University of Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding. He was the General Co-Chair of the IEEE FG 2008 and ACM Multimedia 2013. He was a program chair of ACM Multimedia 2011 and 2007, ECCV 2016, ICCV 2017 and ICPR 2020. He is a general chair of ACM Multimedia 2022 and a program chair of ECCV 2024. He is a fellow of IAPR.