Learning to Attack Real-World Models for Person Re-identification via Virtual-Guided Meta-Learning

Fengxiang Yang[1], Zhun Zhong[2], Hong Liu[3], Zheng Wang[4], Zhiming Luo[6*], Shaozi Li[1*], Nicu Sebe[2,5], Shin'ichi Satoh[3,4]

[1]Artificial Intelligence Department, Xiamen University, China
[2]Department of Information Engineering and Computer Science, University of Trento, Italy
[3]National Institute of Informatics, Japan [4]The University of Tokyo, Japan [5] Huawei Research, Ireland
[6]Post Doctoral Mobile Station of Information and Communication Engineering, Xiamen University, China

Abstract

Recent advances in person re-identification (re-ID) have led to impressive retrieval accuracy. However, existing re-ID models are challenged by the adversarial examples crafted by adding quasi-imperceptible perturbations. Moreover, re-ID systems face the domain shift issue that training and testing domains are not consistent. In this study, we argue that learning powerful attackers with high universality that works well on unseen domains is an important step in promoting the robustness of re-ID systems. Therefore, we introduce a novel universal attack algorithm called "MetaAttack" for person re-ID. MetaAttack can mislead re-ID models on unseen domains by a universal adversarial perturbation. Specifically, to capture common patterns across different domains, we propose a meta-learning scheme to seek the universal perturbation via the gradient interaction between meta-train and meta-test formed by two datasets. We also take advantage of a virtual dataset (PersonX), instead of real ones, to conduct meta-test. This scheme not only enables us to learn with more comprehensive variation factors but also mitigates the negative effects caused by biased factors of real datasets. Experiments on three large-scale re-ID datasets demonstrate the effectiveness of our method in attacking re-ID models on unseen domains. Our final visualization results reveal some new properties of existing re-ID systems, which can guide us in designing a more robust re-ID model. Code and supplemental material are available at https://github.com/FlyingRoastDuck/MetaAttack_AAAI21.

1 Introduction

Person re-identification (re-ID) (Sun et al. 2018; Wang et al. 2018) aims to match pedestrians across non-overlapping cameras. Recent advances in person re-ID have witnessed great progress with the developments of deep models (Ye et al. 2020; Wang et al. 2020b). However, the robustness of deep re-ID models is challenged by the adversarial examples (Szegedy et al. 2014; Wang et al. 2020a). By disturbing images with quasi-imperceptible noises, re-ID models will suffer from catastrophic performance degradation. This makes the design of robust re-ID systems that are insensitive to adversarial examples become an urgent issue to be resolved.

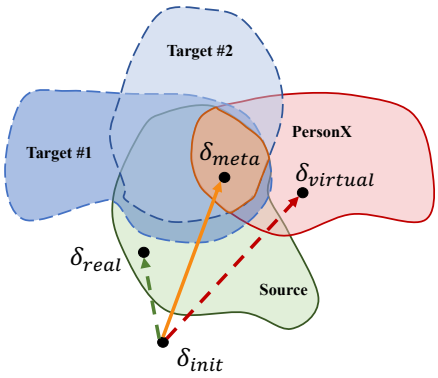

Figure 1: Schematic illustration of attacking in re-ID. Adversarial perturbation sets of real source, real target, and virtual (PersonX) datasets are visualized in different colors. Each perturbation set crushes the re-ID model on its corresponding dataset. The common region represents perturbations that can attack models on all datasets. This common region is hard to reach when directly training with only one dataset, *e.g.,* optimizing with real source ($\delta_{init} \rightarrow \delta_{real}$) or PersonX ($\delta_{init} \rightarrow \delta_{virtual}$). Our MetaAttack leverages the interacted gradients from real source and PersonX to guide the initialized perturbation to the common region ($\delta_{init} \rightarrow \delta_{meta}$).

"Our strength grows out of our weakness."

—Ralph Waldo Emerson

Inspired by this quote, in this work, we argue that *learning powerful attackers helps verifying and improving the robustness of re-ID models, especially ones with high universality*. Then, our goal is designing such an attacker, which will help reveal and understand the weaknesses of re-ID systems.

Most current studies in adversarial attack mainly focus on image classification (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Moosavi-Dezfooli et al. 2017; Goodfellow, Shlens, and Szegedy 2015), while few (Tolias, Radenovic, and Chum 2019; Li et al. 2019; Wang et al. 2020a) have touched upon the attacking scheme of image retrieval, especially person re-ID. Different from image classification, 1)

*Corresponding author: {zhiming.luo, szlig}@xmu.edu.cn.

person re-ID is an open-set (Panareda Busto and Gall 2017) problem in which identities in the training and testing sets are non-overlapped, and 2) person re-ID often encounters a large domain shift issue that training and testing sets are from different domains (Zhong et al. 2018). Hence, it is important to learn an adversarial attacker that is appropriate for different person identities and can generalize to different unseen domains. Recently, MisRank (Wang et al. 2020a) focuses on attacking unseen domains by simulating the impact of image scale with a multi-scale feature extractor. Despite its effectiveness, MisRank has two shortcomings: 1) it requires to generate a unique noise for each query image, limiting the efficiency and flexibility; 2) it ignores various factors (*e.g.*, illumination and viewpoint) that significantly influence the testing results, which should also be considered in generalization. In this paper, we aim to design an attacker that is 1) efficient and flexible, and 2) robust to more variations that may exist in unseen domains.

For the first aspect, we propose to adopt Universal Adversarial Perturbation (UAP) (Moosavi-Dezfooli et al. 2017) for person re-ID. The goal of UAP is to mislead models with a single universal perturbation, which can speed up the attacking process. In addition, the learned UAP can reflect the distribution bias of data (Moosavi-Dezfooli et al. 2017), which benefits the interpretation of the model and the proposal of a defense scheme. However, since the adversarial perturbation is, in fact, a kind of feature (Ilyas et al. 2019), it may excessively focus on the biased factors of training data and as such may fail to achieve good performance on unseen domains. In this study, we assume that *there exists a universal perturbation that captures common factors across domains and can attack most domains*. Taking Fig. 1 as an example, we consider the perturbation sets of different domains have intersections with each other. The "common region" represents the perturbations that can attack most domains and we aim to learn a perturbation belonging to it.

To achieve this, a straightforward way is to train with a much larger and comprehensive dataset. However, in practice, due to the difficulty of labeling, the size of existing datasets are limited. Although we can directly train with the combination of existing datasets, the domain shifts between different datasets will hamper capturing the shared knowledge among datasets. In addition, due to the data privacy, we can actually only access few datasets. To solve this problem, we propose to utilize the meta-learning (Finn, Abbeel, and Levine 2017) to simulate the cross-domain process using two datasets during training. The meta-learning separates the training data into meta-train and meta-test, which enables us to learn the basic knowledge from meta-train as well as generalize the variation factors from meta-test. Intuitively, if a meta-test can cover as many common factors as possible, then the learned perturbations would be potentially located at the "common region" and thus can well attack more unseen domains. However, the existing real-world datasets are seriously limited by the biased environmental factors (Sun and Zheng 2019), hindering the strength of the meta-learning. Therefore, *for the second aspect*, we consider to take a virtual (synthetic) dataset, PersonX (Sun and Zheng 2019), into the meta-training process. PersonX is designed to simulate important variation factors of the re-ID system, such as pose, viewpoint, illumination and background. Moreover, more factors can be involved by simply controlling and generating data with a unity engine. It is easier for virtual dataset to consider more environmental factors. Based on these, PersonX is more appropriate to be the meta-test for capturing common patterns across domains during meta-learning.

This work proposes a novel universal adversarial perturbation method, named "MetaAttack", for person re-ID through virtual-guided meta-learning. In our method, we use two datasets during training: a real dataset regarded as the source domain and a virtual dataset (PersonX) regarded as the extra association domain. MetaAttack is designed to learn a universal perturbation that disturbs queries and causes significant performance drop on different unseen (target) domains without any online modification. Specifically, we take the real dataset as meta-train and PersonX as meta-test. This enables us to simulate the cross-domain constraint and to learn the universal perturbation with a meta-learning. During optimization, the gradient from meta-train and meta-gradient from meta-test are aggregated to obtain the final gradient, which is used to update the perturbation and can improve the universality of the learned perturbation. To sum up, our contributions mainly lie in three aspects:

- We propose a meta-learning scheme to learn universal perturbation for person re-ID. With our method, the cross-domain constraint is explicitly injected into the optimization, improving universality of the learned perturbation.
- We adopt a virtual dataset as meta-test during meta-optimization. The diverse, balanced virtual data enable us to capture more common patterns across domains.
- Extensive experiments on three large-scale benchmarks demonstrate the effectiveness of the proposed MetaAttack. Our method can outperform state-of-the-art approaches when attacking unseen domains, even using a smaller perturbation budget ϵ.

2 Related Work

2.1 Adversarial Attack

Szegedy et al. (2014) reveal the existence of adversarial examples and propose to learn perturbations by one-step gradient optimization. Subsequently, a lot of adversarial attack methods have been explored (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Madry et al. 2018; Poursaeed et al. 2018; Fan et al. 2020), but they need to generate perturbation for each image individually, which are not efficient in practice. Su, Vargas, and Sakurai (2019) propose to attack classification models with only one pixel. However, it dramatically suffers from increased searching time when handling large-scale datasets. To speed up the attacking process, Moosavi-Dezfooli *et al.* propose the Universal Adversarial Perturbation (UAP) (Moosavi-Dezfooli et al. 2017) algorithm that can attack deep models with a single adversarial noise. It can also be used for black-box attack, where the target model is not available during the optimization. Most of existing attack

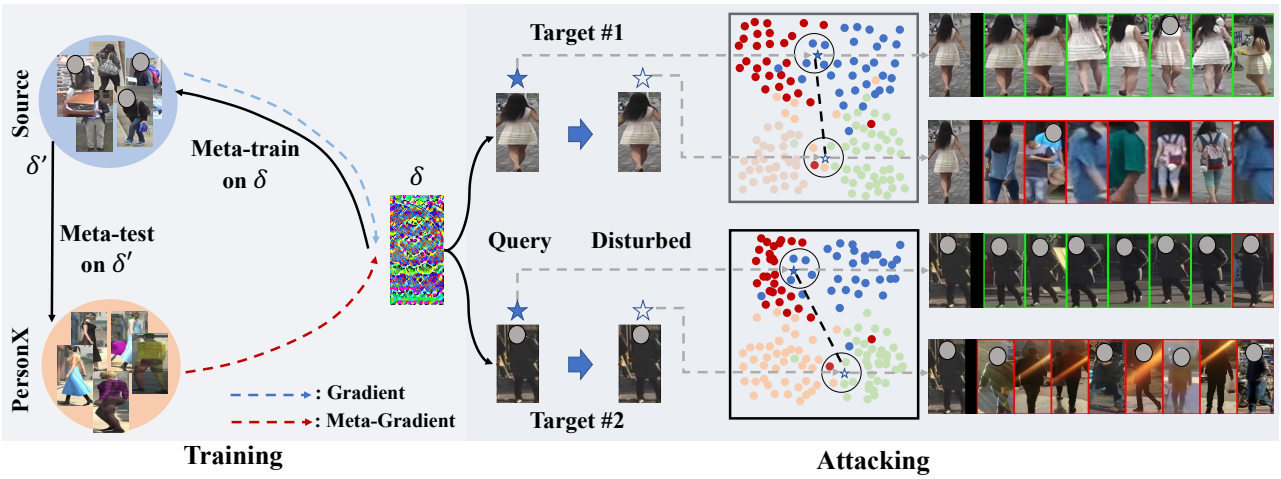

Figure 2: The framework of the proposed MetaAttack. During training, we use the source dataset $\mathcal{S}$ as meta-train $\mathcal{M}_{tr}$ and PersonX as meta-test $\mathcal{M}_{te}$ to simulate cross-domain attack. The aggregation of gradients computed by meta-train and meta-test is used to optimize the perturbation δ. During testing, δ can attack both source domain and unseen target domains.

methods focus on misleading classification models. Different from them, this paper concentrates on fooling person re-ID systems with a universal perturbation.

2.2 Attack Person Re-ID System

There are few existing works that contribute to the attack of image retrieval problem (Li et al. 2019; Wang et al. 2020a), especially person re-ID. Li et al. (2019) design an attack scheme for image retrieval by corrupting label-wise, pair-wise and list-wise relationships in the training set. Zheng et al. (2018) and Bai et al. (2019) study the effectiveness of different attack methods for person re-ID. The above methods do not explicitly consider the context of universal attack, which aims to attack unseen domains during testing and is the goal of this work. Our work is most related to the method in (Wang et al. 2020a) that designs a generator to produce perturbation and verify the universality of the generated perturbation. Different from their work, our method does not require any generators and adopts a virtual-guided meta-learning scheme to learn a UAP.

2.3 Meta Learning

Meta-learning is designed to learn new tasks with limited training samples and improve the generalization to different tasks (Li et al. 2018; Guo et al. 2020). Existing meta-learning methods can be mainly divided into three classes: *metric-based* (Snell, Swersky, and Zemel 2017; Sung et al. 2018), *model-based* (Santoro et al. 2016) and *optimizing-based* methods (Finn, Abbeel, and Levine 2017; Nichol and Schulman 2018). Our algorithm is constructed based on MAML (Finn, Abbeel, and Levine 2017), which is an optimizing-based method. MAML attempts to obtain a good initialized weight that can fast adapt to new tasks by simulating the learning process of new tasks with meta-test. Different from MAML that focuses on few-shot learning problem, this work aims to learn a universal perturbation that can be used for misleading re-ID models on unseen domains.

3 Methodology

Problem Definition. We aim to seek a universal adversarial perturbation δ that can mislead ranking results of re-ID models in both source domain $\mathcal{S}$ and unseen target domains $\mathcal{T}$. The attack operation is achieved by adding δ to a query image I. The perturbed query $I^{'}$ ($I^{'} = I + \delta$) is used to retrieve from the gallery and mislead victim re-ID model $\mathcal{F}$.

3.1 Overall Framework

In Fig. 2, we show the overall framework of the proposed MetaAttack. *In the training stage*, we propose to optimize δ by meta-learning with a source dataset and an extra association dataset. The source data is a real dataset (*e.g.*, Duke), which is adopted as the meta-train for basic optimization. The extra association dataset is a virtual dataset (PersonX) that is utilized as meta-test to mimic possible real-world scenarios and improve the universality of δ. Our method tries to learn a δ locating at the "common region" that can successfully attack different domains. *In the attack stage*, the obtained δ fools re-ID models, resulting in incorrect ranking lists. Next, we will introduce our method in detail.

3.2 Basic Losses for Attacking Re-ID Models

In this work, we aim to cheat re-ID models with a single universal perturbation. We use pair-wise and label-wise relations among training samples for the perturbation learning.
Misleading Pair-wise Relations. We follow (Wang et al. 2020a), which applies triplet loss to pull dissimilar pairs close and push similar pairs away. Different from (Wang et al. 2020a), we do not use the labels of training data to estimate the pair-wise relations between samples. Instead, we apply the centroids generated by clustering, which can better reveal the sample similarities of re-ID models (Li et al. 2019; Radenović, Tolias, and Chum 2018).

Optimizing perturbations with cluster centroids can be regarded as directly corrupting the feature space of the re-ID

model. Therefore, we first quantize the training data into k centroids by the k-means algorithm with the features generated by re-ID models. Then, based on the obtained cluster centroids, we use the triplet loss to mislead pair-wise relations between samples, which is formulated as:

$$L_{tri}(\mathbf{f}^{'};\delta) = \Big[||\mathbf{c_n} - \mathbf{f}^{'}||_2 - ||\mathbf{c_p} - \mathbf{f}^{'}||_2 + m\Big]_+, \quad (1)$$

where $[\cdot]_+$ is the $\max(\cdot, 0)$ function. $\mathbf{f}^{'} \in \mathbb{R}^{d\times 1}$ is the disturbed feature of the disturbed image $I^{'}$ ($\mathbf{f}^{'} = \mathcal{F}(I^{'})$), and d is the dimension of feature. $\mathbf{c_p}$ and $\mathbf{c_n}$ are closest and furthest cluster centroids of original image feature in the training data, respectively. m is the hyper-parameter that controls the margin of positive and negative pairs.

Misleading Label-wise Relations. We also use the loss function proposed in (Li et al. 2019) to mislead label-wise relations among training samples. In this function, a mis-classification loss is used to enforce a sample away from its nearest centroid while pull it close to its second-nearest centroid. The mis-classification loss is formulated as:

$$L_{cls}(\mathbf{f}^{'};\delta) = \Big[\mathbf{f}^{'\mathrm{T}}\mathbf{c_1} - \mathbf{f}^{'\mathrm{T}}\mathbf{c_2}\Big]_+, \quad (2)$$

where $\mathbf{c_1}$ and $\mathbf{c_2}$ represent the nearest and second-nearest centroids of original image feature, respectively.

3.3 Virtual-Guided Meta-Learning

To improve the universality of perturbation δ, an intuitive way is to train with a larger dataset. This is, however, not applicable in the real world because of the difficulty of labeling re-ID data and data privacy. We propose to generalize perturbation with meta-learning. In our method, the training data is formed by two datasets that are regarded as meta-train $\mathcal{M}_{tr}$ and meta-test $\mathcal{M}_{te}$, respectively. As shown in Fig. 3, the final gradient for meta-optimization is obtained by combining gradients from both $\mathcal{M}_{tr}$ and $\mathcal{M}_{te}$, which calibrates the universal perturbation δ to the direction that can perform well on both parts. Intuitively, if a meta-test set includes as many common factors as possible, the learned perturbation will more easily be located at the "common region," as shown in Fig. 1, and thus the perturbation has a better universality. To achieve this, we require a dataset that has more common and balanced factors to form $\mathcal{M}_{te}$. PersonX is a virtual synthetic dataset that contains several important and comprehensive variation factors of re-ID system. Hence PersonX is more appropriate for meta-learning. We present a virtual-guided meta-learning algorithm for attacking re-ID models. The algorithm is summarized in Alg. 1, which contains three steps.

Step 1: Meta-train. We utilize the source dataset $\mathcal{S}$ as the meta-train set and learn the perturbation δ with the loss functions introduced in Sec. 3.2. For the given meta-train batch with N_b samples, we perturb and extract their features with the re-ID model trained on the source data. The loss function for meta-train is formulated as:

$$L_{mtr}(\mathbf{F}^{'}_{mtr};\delta) = \frac{1}{N_b}\sum_{i=1}^{N_b}\Big[L_{cls}(\mathbf{f}^{'}_i;\delta) + \lambda L_{tri}(\mathbf{f}^{'}_i;\delta)\Big], \quad (3)$$

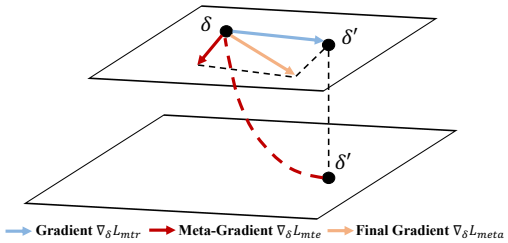

Figure 3: An illustration of our meta-learning process. Given current perturbation δ, we update it with gradient computed on meta-train data $\mathcal{M}_{tr}$ and obtain temporary δ'. Then, we compute meta-gradient on meta-test data $\mathcal{M}_{te}$ with δ'. Note that the meta-gradient is obtained from δ rather than δ'. The final gradient is the combination of gradient and meta-gradient, which is used for updating δ.

where $\mathbf{F}^{'}_{mtr}$ represents N_b features of current disturbed meta-train batch, $\mathbf{f}^{'}_i$ is the i-th feature, and λ is the balancing parameter. We follow (Dong et al. 2018; Li et al. 2019) to obtain a updated temporary δ' through stochastic gradient descent (SGD) with momentum, which is formulated as:

$$\begin{aligned} g' &= \mu g + \frac{\nabla_\delta L_{mtr}}{||\nabla_\delta L_{mtr}||_1}, \\ \delta' &= clip(-\epsilon, \epsilon, \delta - \alpha \cdot sign(g')), \end{aligned} \quad (4)$$

where g' is the momentum for updating δ and g is the momentum. μ is the weight of momentum and α is the learning rate. $clip(\cdot)$ is the function that ensures the constraint $||\delta||_\infty \leq \epsilon$. $\delta^{'}$ is the updated temporary perturbation, which will be used in the following meta-test step.

Step 2: Meta-test. We use PersonX to be the meta-test $\mathcal{M}_{te}$ and compute meta-test loss with temporary δ' obtained in **Step 1**, which is defined as:

$$L_{mte}(\mathbf{F}^{'}_{mte};\delta') = \frac{1}{N_b}\sum_{i=1}^{N_b}\Big[L_{cls}(\mathbf{f}^{'}_i;\delta') + \lambda L_{tri}(\mathbf{f}^{'}_i;\delta')\Big], \quad (5)$$

where $\mathbf{F}^{'}_{mte}$ indicates N_b features of current disturbed meta-test batch, $\mathbf{f}^{'}_i$ is the i-th feature extracted by re-ID model. The meta-test is used to mimic the perturbation attack process on unseen target domains. We use the meta-test loss to calculate the meta-gradient on the original δ, which can be considered as a regularization term to guide the original δ to the "common region" in the final step.

Step 3: Meta-update. The final step is based on the losses in the aforementioned two steps. Specifically, our final meta-loss function for learning δ is:

$$L_{meta}(\mathbf{F}^{'}_{mtr}, \mathbf{F}^{'}_{mte};\delta) = L_{mtr}(\mathbf{F}^{'}_{mtr};\delta) + L_{mte}(\mathbf{F}^{'}_{mte};\delta^{'}). \quad (6)$$

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.

3.4 Universal Attack

The optimized δ is utilized to proceed universal attack, as shown in the right part of Fig. 2. We mainly aim to attack

Algorithm 1 Procedure of MetaAttack.

Inputs: Meta-train $\mathcal{M}_{tr}$ (source dataset $\mathcal{S}$), meta-test $\mathcal{M}_{te}$ (association dataset), number of centroids k, batch size N_b, re-ID model trained on source domain $\mathcal{F}_{\mathcal{S}}$, maximum iterations max_iter, learning rate α.
Outputs: Universal perturbation δ.

Use k-means clustering to obtain k centroids on $\mathcal{M}_{tr}$ and $\mathcal{M}_{te}$, respectively;
Initialize δ with $\mathbf{0}$;
for i in max_iter **do**
  repeat
    Sample mini-batches m_{tr} and m_{te} with N_b images from $\mathcal{M}_{tr}$ and $\mathcal{M}_{te}$, respectively;
    Disturb m_{tr} and m_{te} with δ to obtain $m_{tr}^{'}$ and $m_{te}^{'}$;
    Extract features for disturbed $m_{tr}^{'}$ and $m_{te}^{'}$ with $\mathcal{F}_{\mathcal{S}}$ to obtain $\mathbf{F}_{mtr}^{'} \in \mathbb{R}^{d \times N_b}$ and $\mathbf{F}_{mte}^{'} \in \mathbb{R}^{d \times N_b}$;
    // Step 1: Meta-train
    Compute meta-train loss and obtain temporary δ' with $\mathbf{F}_{mtr}^{'}$ (Eq. 3 and Eq. 4);
    // Step 2: Meta-test
    Compute meta-test loss with δ' and $\mathbf{F}_{mte}^{'}$ (Eq. 5);
    // Step 3: Meta-update
    Compute final loss and update δ through SGD with momentum (Eq. 6);
  until $\mathcal{M}_{tr}$ and $\mathcal{M}_{te}$ are enumerated;
end for
Return δ;

re-ID models of unseen domains. This goal is achieved by directly adding δ to all queries and corrupting their corresponding retrieval ranking lists.

4 Experiments

Datasets. We use three large-scale re-ID benchmarks to verify our algorithm, *i.e.* Market-1501 (Market) (Zheng et al. 2015), DukeMTMC-reID (Duke) (Ristani et al. 2016; Zheng, Zheng, and Yang 2017) and MSMT-17 (MSMT) (Wei et al. 2018). Market contains $32,668$ images of $1,501$ identities obtained from six cameras. Duke consists of $36,411$ labeled images of $1,404$ identities pictured by eight different cameras. MSMT has $126,441$ images from $4,101$ pedestrians captured by fifteen cameras. For each dataset, nearly half of the identities are used for training. We only use PersonX-456 as meta-test, which removes all samples without backgrounds in PersonX (Sun and Zheng 2019) and contains $39,852$ images from 410 identities.

Evaluation Protocol. To show the universality of different attack methods, we learn δ on a source dataset and then adopt δ to corrupt queries of other (target) datasets. In this paper, only the real datasets will be used as source and target datasets. The virtual dataset (PersonX) is an extra association dataset for our MetaAttack. The widely used mAP and rank-1 accuracy are used for evaluation. Lower mAP and rank-1 accuracies indicate better attack performance.

Experimental Settings. We test our method on both global-based and part-based models. For the first, we use IDE (Zheng, Yang, and Hauptmann 2016) to train the re-ID model and extract *pooling-5* feature to compute Eq. 6 for meta-optimization. For the second, we use PCB (Sun et al. 2018) to train the re-ID model. Specifically, PCB considers pedestrians as six parts and extract 256-dim feature for each part.[1] We use the ResNet-50 (He et al. 2016) as the backbone for both models.

All hyper-parameters in our experiments are set as follow: the number of centroids $k = 512$, the batch size $N_b = 50$, the iteration number $max\_iter = 20$, margin $m = 0.5$, and the learning rate $\alpha = \epsilon/10$. We use SGD with momentum (Dong et al. 2018; Li et al. 2019) to update δ, and the weight of momentum $\mu = 1$. The balancing factor λ is set to 10. We perform L_∞-bounded attacks with $\epsilon = 8$ unless otherwise noted. ϵ is the upper bound for each pixel of the generated δ, *i.e.*, $||\delta||_\infty \leq \epsilon$.

4.1 Comparison with State-of-the-Art

We first compare our method with two state-of-the-art algorithms: MisRank[2] (Wang et al. 2020a) and UAP-Retrieval[2] (Li et al. 2019). In most experiments, we set the $\epsilon = 8$ to obtain quasi-imperceptible perturbation. We also report results when $\epsilon = 16$ for fair comparison with MisRank (Wang et al. 2020a). In addition, since our method uses PersonX as the extra association dataset, we report the results of training MisRank with both source data and PersonX ("MisRank+PersonX"). In Tab. 1, the first two columns of results (Duke $\rightarrow$ Market and Duke $\rightarrow$ MSMT) use Duke as the source domain and the other two datasets (Market and MSMT) as target domains. Similar settings are used for the last two columns of results.

From Tab. 1, we have the following conclusions. **(1)** Our method can achieve the best attack results with the same ϵ in all settings. This demonstrates the effectiveness of our method in attacking unseen domains and shows that our method is capable of attacking both global- and part-based models. **(2)** The effect of MisRank largely relies on a larger ϵ. When $\epsilon = 8$, MisRank fails to achieve competitive attacking results while our method obtains reasonable results that clearly outperform MisRank. Importantly, our method with $\epsilon = 8$ can obtain better results than MisRank with $\epsilon = 16$ in some settings. For example, in the setting of Duke $\rightarrow$ Market, our method with $\epsilon = 8$ reduces the mAP to 4.9%. This is 5.4% lower than MisRank with $\epsilon = 16$. **(3)** PersonX can not bring improvement for MisRank. When additionally training with PersonX, the attacking results of MisRank are even worse compared to the one trained with only source data. This suggests that PersonX may be not suitable for generator-based method and that leveraging the extra virtual dataset is not trivial in attack re-ID.

[1] During clustering, part features are aggregated into $1,536$-dim feature to obtain cluster centroids. During optimization, each cluster centroid is divided into six parts for computing attack losses of each corresponding part.

[2] We reproduced the experiments based on the authors' code.

Table 1: Results for attacking re-ID systems. We use our method to attack different backbones (IDE (Zheng, Yang, and Hauptmann 2016) and part-based PCB (Sun et al. 2018)), then compare our method with state-of-the-arts (MisRank (Wang et al. 2020a) and UAP-Retrieval (Li et al. 2019)). "Before Attack": re-ID accuracies of unseen target model on target set.

Backbone	Methods	Duke → Market		Duke → MSMT		Market → Duke		Market → MSMT	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
IDE	Before Attack	78.2	88.7	42.3	69.8	66.7	80.9	42.3	69.8
	MisRank	28.2	38.6	11.7	30.3	36.7	48.8	11.1	28.5
	MisRank + PersonX	38.5	51.5	20.9	55.8	43.4	71.2	12.4	31.0
	MisRank ($\epsilon = 16$)	10.3	13.0	3.0	7.2	13.7	18.3	1.6	4.2
	UAP-Retrieval	8.2	9.7	5.5	15.4	14.8	20.4	5.3	13.9
	MetaAttack (Ours)	**4.9**	**7.0**	**3.5**	**8.3**	**11.2**	**15.2**	**3.4**	**8.3**
	MetaAttack (Ours, $\epsilon = 16$)	**0.7**	**0.9**	**0.3**	**0.7**	**1.0**	**1.3**	**0.5**	**1.1**
PCB	Before Attack	76.7	91.3	50.8	88.9	68.0	84.1	50.8	88.9
	MisRank	48.1	64.2	21.1	47.7	31.2	45.4	14.4	28.5
	MisRank + PersonX	52.4	70.6	18.8	39.6	38.0	51.4	18.8	39.6
	MisRank ($\epsilon = 16$)	11.5	13.8	5.2	9.6	12.4	17.8	8.2	17.0
	UAP-Retrieval	21.6	30.4	4.4	9.1	29.0	41.9	4.3	8.9
	MetaAttack (Ours)	**19.5**	**28.2**	**4.2**	**8.7**	**26.9**	**39.9**	**3.8**	**8.2**
	MetaAttack (Ours, $\epsilon = 16$)	**4.5**	**5.9**	**0.6**	**1.4**	**4.1**	**6.6**	**0.9**	**1.9**

Table 2: Ablation study on the proposed virtual-guided meta-learning algorithm.

No.	Duke → MSMT		Market → MSMT		Extra Data		Meta
	mAP	rank-1	mAP	rank-1	Real	PersonX	Learning
1	5.6	14.3	5.8	14.9	×	×	×
2	5.1	14.5	5.7	14.3	✓	×	×
3	4.8	10.4	5.0	12.6	✓	×	✓
4	4.6	9.9	5.5	14.2	×	✓	×
5	**3.5**	**8.3**	**3.4**	**8.3**	×	✓	✓

4.2 Ablation Study

To show the effectiveness of the proposed method, we conduct experiments by adding extra training data and meta-learning into the baseline. Results with IDE model are reported in Tab. 2. The first row (*No.1*) is the baseline that only uses the basic losses functions (Sec. 3.2) on the training data. For the extra real data, we use Market when using Duke as the source domain, vice versa.

The effectiveness of meta-learning. To verify the significance of the meta-learning strategy, we compare with the variant that directly trained with the source data and the extra data. From the comparison of *No.2* vs *No.3* and *No.4* vs *No.5*, we can observe that 1) directly combing the source and extra data brings limited improvement; and 2) training with the meta-learning strategy can consistently improve the attack results and universality of learned perturbation.

The benefit of virtual data in meta-learning. Another important component of our method is adopting a virtual dataset instead of a real one during meta-learning. The comparison of *No.3* vs *No.5* shows that virtual-guided meta-learning outperforms the real-guided one. For example, when using Duke as the source domain, virtual-guided meta-learning (*No.5*) reduces the mAP to 3.5%, which is lower than real-guided one (*No.3*) by 1.3%. These results indicate that using a dataset with less biased factors can improve the universality of learned perturbation.

Table 3: Results on source domain.

Backbone	Method	Duke		Market	
		mAP	rank-1	mAP	rank-1
IDE	Before Attack	66.7	80.9	78.2	88.7
	UAP-Retrieval	4.2	9.9	3.6	4.5
	Ours	**3.6**	**6.4**	**3.1**	**3.4**
PCB	Before Attack	68.0	84.1	76.7	91.3
	UAP-Retrieval	14.3	20.3	**10.7**	**15.1**
	Ours	**11.2**	**16.5**	10.9	15.4

4.3 Performance on Source Domain

In Tab. 3, we report results on the testing set of source domain and compare our MetaAttack with UAP-Retrieval for both IDE and PCB models. Tab. 3 shows that our model can effectively corrupt the accuracies of the ranking list on the source domain, and can achieve better results than UAP-Retrieval in most settings. Since UAP-Retrieval uses almost the same basic loss functions to our MetaAttack, it can be regarded as the reduction of MetaAttack, which does not use virtual-guided meta-learning. Then, we can conclude that our MetaAttack can also improve the university of perturbation in the source domain.

4.4 Visualization

In this section, we visualize the obtained δ and some perturbed query images to give an intuitive presentation of the proposed MetaAttack algorithm.

Robust Queries in MetaAttack. Our MetaAttack can effectively corrupt re-ID accuracies with slight modifications to query images. However, it remains several robust queries that can defend our attack. To find out their common attributes, we show some examples in Fig. 4(a) and (b). All the experiments in this part use the IDE model.

We observe two kinds of situations that may help defend our attack. The first kind of robust query is caused by occlusion and we visualize its ranking lists before and after attack in the first and second rows of Fig. 4(c). Since there

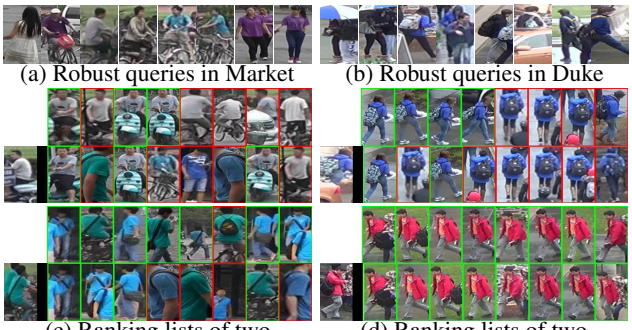

Figure 4: The visualization of some robust queries.

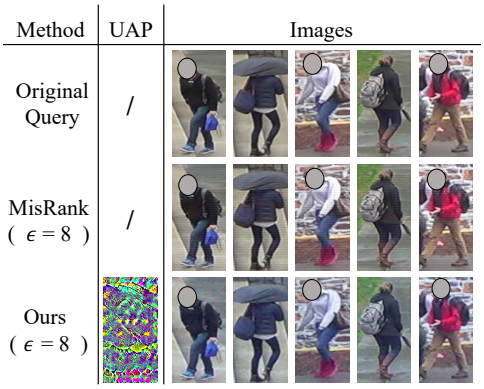

Figure 5: Visualizations of corrupted queries and obtained δ.

are few occluded samples in the training data, the learned perturbation will only capture the overall distribution of non-occluded images but be sensitive to occluded images.

Therefore, during testing, we can try to add an occluding process before forwarding queries to the re-ID model, for defending the adversarial samples, e.g., adding erasing (Carmon et al. 2019; Zhong et al. 2020). Another kind of robust query is caused by the camera shift in the dataset. As shown in the third (before attack) and fourth (after attack) rows of Fig. 4(c), the pedestrian in the query image with a green T-shirt, changes to a blue T-shirt in its correctly matched nearest neighbor. The change of appearance is caused by camera shift (Zhong et al. 2018) and has been a long-standing problem in person re-ID. Both the source data and the PersonX do not contain such kind of cases, which causes our failure. In fact, each domain may contain its own specific camera shift that is very different to other domains. *Therefore, we argue that adding domain-specific camera shift to query images may help defend our attack, which can be used as a reference for designing defense models of re-ID*. Similar results and conclusions can also be found in Duke (Fig. 4(d)).

Visualizations of δ and Perturbed Images. In Fig. 5, we visualize the obtained δ and some perturbed images. Our method is more efficient and flexible than MisRank, because our method only requires a single perturbation for all queries while MisRank needs to generate new perturbations for different queries. We can see that under the same magnitude ϵ ($\epsilon = 8$), our generated adversarial examples are much better than those of MisRank. The quality of generated adversarial examples is further evaluated in Sec. 4.5 with SSIM.

Table 4: SSIM scores of generated adversarial examples between (Wang et al. 2020a) and our method.

	Duke	Market
MisRank	0.1985	0.1889
Ours	**0.2121**	**0.1963**

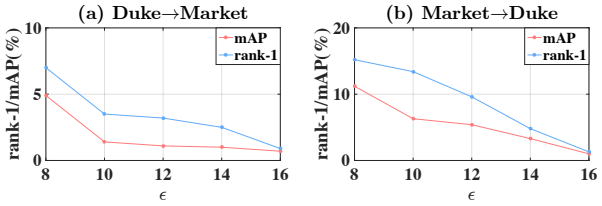

Figure 6: Sensitive analysis of ϵ.

4.5 Further Experiments

Image Quality. SSIM (Wang et al. 2004) is a kind of metric to measure the similarity of two images and has been widely used to evaluate the quality of GAN-made (Goodfellow et al. 2014) virtual images. A larger SSIM score between synthetic and natural images indicates better quality and less distortion. We, therefore, utilize SSIM to evaluate the degree of distortion for adversarial examples. We report SSIM scores in Tab. 4. Compared with MisRank, our method produces higher SSIM scores, indicating that our method can generate higher quality adversarial images and achieve better attack performance.

Sensitive Analysis. We change the value of ϵ from 8 to 16 and study the influence of perturbation budget. In Fig. 6, we plot the curve of mAP and rank-1 scores under two settings (Market→Duke and Duke→Market). The results show that a larger ϵ can easily damage re-ID accuracies. However, to make the obtained perturbation quasi-imperceptible, we suggest using a small ϵ to attack real-world re-ID models if a good attack results can be achieved.

5 Conclusion

In this paper, we propose a novel universal attack algorithm for person re-ID, which is based on the virtual-guided meta-learning. Our method takes the source dataset to be meta-train and the synthetic PersonX dataset as meta-test. By combining the gradients from both meta-train and meta-test sets during meta-optimization, the obtained perturbation can learn to generalize in unseen target domains and achieve satisfactory results. The proposed method performs well on three large-scale datasets with both IDE and PCB models. In our future work, we consider applying our observations and perspectives to design robust re-ID that can defend against adversarial samples.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (No. 61876159, 61806172, 61662024 and U1705286); the China Postdoctoral Science Foundation Grant (No. 2019M652257); the Fundamental Research Funds for the Central Universities (Xiamen University, No. 20720200030); the Italy-China cooperation project TALENT and the PRIN project PREVUE.

References

Bai, S.; Li, Y.; Zhou, Y.; Li, Q.; and Torr, P. H. 2019. Metric Attack and Defense for Person Re-identification. *arXiv preprint arXiv:1901.10650* .

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Duchi, J. C.; and Liang, P. S. 2019. Unlabeled data improves adversarial robustness. In *NeurIPS*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *CVPR*.

Fan, Y.; Wu, B.; Li, T.; Zhang, Y.; Li, M.; Li, Z.; and Yang, Y. 2020. Sparse adversarial attack via perturbation factorization. In *ECCV*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Guo, J.; Zhu, X.; Zhao, C.; Cao, D.; Lei, Z.; and Li, S. Z. 2020. Learning meta face recognition in unseen domains. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. In *NeurIPS*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018. Learning to generalize: Meta-learning for domain generalization. In *AAAI*.

Li, J.; Ji, R.; Liu, H.; Hong, X.; Gao, Y.; and Tian, Q. 2019. Universal perturbation attack against image retrieval. In *ICCV*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *CVPR*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.

Nichol, A.; and Schulman, J. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* .

Panareda Busto, P.; and Gall, J. 2017. Open set domain adaptation. In *ICCV*.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *CVPR*.

Radenović, F.; Tolias, G.; and Chum, O. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE TPAMI* 41(7): 1655–1668.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *ECCV*.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *ICML*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NeurIPS*.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE TEC* 23(5): 828–841.

Sun, X.; and Zheng, L. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

Tolias, G.; Radenovic, F.; and Chum, O. 2019. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In *ICCV*.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*.

Wang, H.; Wang, G.; Li, Y.; Zhang, D.; and Lin, L. 2020a. Transferable, Controllable, and Inconspicuous Adversarial Attacks on Person Re-identification With Deep Mis-Ranking. In *CVPR*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP* 13(4): 600–612.

Wang, Z.; Wang, Z.; Zheng, Y.; Wu, Y.; Zeng, W.; and Satoh, S. 2020b. Beyond intra-modality: A survey of heterogeneous person re-identification. In *IJCAI*.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *CVPR*.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. H. 2020. Deep Learning for Person Re-identification: A Survey and Outlook. *arXiv preprint arXiv:2001.04193* .

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *CVPR*.

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

Zheng, Z.; Zheng, L.; Hu, Z.; and Yang, Y. 2018. Open set adversarial examples. *arXiv preprint arXiv:1809.02681* .

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. In *ICCV*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *AAAI*.

Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; and Yang, Y. 2018. Camera style adaptation for person re-identification. In *CVPR*.