Universal Adversarial Perturbations Against Person Re-Identification

Wenjie Ding\textsuperscript{1}  Xing Wei\textsuperscript{1}  Xiaopeng Hong\textsuperscript{1}  Rongrong Ji\textsuperscript{2}  Yihong Gong\textsuperscript{1}
\textsuperscript{1}Xi’an Jiaotong University  \textsuperscript{2}Xiamen University

Abstract

Person re-identification (re-ID) has made great progress and achieved high performance in recent years with the development of deep learning. However, as an application related to security issues, there are few researches considering the safety of person re-ID systems. In this paper, we attempt to explore the robustness of current person re-ID models against adversarial samples. Specifically, we attack the re-ID models using universal adversarial perturbations (UAPs), which are especially dangerous to the surveillance systems because it could fool most pedestrian images with a little overhead. Existing methods for UAPs mainly consider classification models, while the tasks in open set scenarios like re-ID are rarely explored. Re-ID attack is different from classification ones in the sense that the former discards decision boundary during test and cares more about the ranking list. Therefore, we propose an effective method to train UAPs against person re-ID models from the global list-wise perspective. Furthermore, to increase the impact of attack to different models and datasets, we propose a novel UAPs learning method based on total variation minimization. Extensive experiments validate the effectiveness of our proposed method.

Introduction

In the past few years, deep learning based person re-identification (re-ID) has made remarkable progress and achieved high performance. Despite the great success, the vulnerability of neural network gradually increases people’s vigilance in more recent years. It is shown that the output of convolution neural networks (CNNs) can be easily changed by a small perturbation on the input (Szegedy et al. 2013). Existing researches in person re-ID mainly focus on learning image features with the goal to be invariant to the large appearance variations caused by illuminations, human poses (Liu et al. 2018), occlusions and camera viewpoints (Saquib Sarfraz et al. 2018). However, the robustness of re-ID models against adversarial attack, which is a significant issue for surveillance systems, is hardly ever explored. The work (Bai et al. 2019) attacks re-ID models by generating a new gallery composed of adversarial samples. In (Zhedong Zheng 2019), adversarial query images are generated by pushing the features of perturbed images in the opposite direction of the raw features. UAA-GAN (Zhao et al. 2019) trains a generative adversarial network to generate perturbations. Existing studies on re-ID models against adversarial samples are usually image-specific and barely touches the impact of image-agnostic adversarial samples. Unlike image-specific attacking methods which have to generate tailored adversarial samples for every image, the Universal Adversarial Perturbation (UAP) is image-agnostic and can fool a given CNN model with high probability, independent of the input applied to the model (Moosavi-Dezfooli et al. 2017). This kind of attack is efficient and requires little extra computation, which makes it more dangerous to the surveillance system. However, existing works on UAP mainly consider classification analysis, such as image classification (Moosavi-Dezfooli et al. 2017) and semantic image segmentation (Hendrik Metzen et al. 2017) and few works have been devoted to the task of person re-ID.

In this paper, we make the first attempt to attack person re-ID models using universal adversarial perturbations, as illustrated in Fig. 1. This research is necessary and significant, since previous UAP attack methods are inappro-
prietate on this problem due to the reasons explained below. Most attack methods for image classification are on the basis of pushing features of adversarial samples out of decision boundaries (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Mei and Zhu 2015; Goodfellow, Shlens, and Szegedy 2014). Nonetheless, on the person re-ID task, the concept of decision boundaries is discarded during test, since person re-ID is an open set problem where training and testing images belong to different identities. As a result, those attack methods which stress the decision boundaries are insufficient for attacking person re-ID which aims at disordering the global ranking list.

To this end, we propose to use a list-wise attack objective for corrupting the global ranking of person re-ID systems. We perform extensive experiments using models with five different CNN architectures including ResNet-50, DenseNet-121, VGG-16, SENet-154 and ShuffleNet, and trained on two widely used re-ID datasets. The experimental results show that the accuracies of person re-ID models fall considerably under attack using our proposed UAP trained on two widely used re-ID datasets. The experimental results show that the accuracies of person re-ID models fall considerably under attack using our proposed UAP method. For instance, the mean Average Precision (mAP) of the mean Average Precision of different identities. As a result, those attack methods which stress the decision boundaries are insufficient for attacking person re-ID which aims at disordering the global ranking list.

To this end, we propose to use a list-wise attack objective for corrupting the global ranking of person re-ID systems. We perform extensive experiments using models with five different CNN architectures including ResNet-50, DenseNet-121, VGG-16, SENet-154 and ShuffleNet, and trained on two widely used re-ID datasets. The experimental results show that the accuracies of person re-ID models fall considerably under attack using our proposed UAP method. For instance, the mean Average Precision (mAP) of a state-of-the-art re-ID model is decreased from 85.32% to 62.02% with a mean Drop Rate (mDR) of 99.38% defined in Eq. (12) on the Market-1501 dataset after attack.

Moreover, we further observe that the attack performance may fall down noticeably when the source model used for training UAPs and the target model to attack use different CNN structures, or employ different training data. For example, the mean Drop Rate of ResNet-50 trained on DukeMTMC-reID is 98.15% under the attack of UAP trained on the CNN architecture itself. However, the mDR is only 62.02% using UAP trained on VGG-16 on the same dataset. Similar observations have been demonstrated in many existing UAP works (Mopuri, Ganeshan, and Radhakrishnan 2018; Moosavi-Dezfooli et al. 2017). Therefore, we explore this phenomenon and further propose a method to improve the impact of UAP attack across models and datasets by reducing the biasing of perturbations on the training model. Experiment results show that such an approach can significantly improves the generalization ability of UAP attack.

To be summarized, this work makes the following contributions:

- We make the first attempt to ‘universally’ attack state-of-the-art re-ID models.
- We propose a novel method to improve the effectiveness of attack to different CNN architectures and datasets.
- Our proposed method achieves high attack performance in extensive experiments performed with five different CNN architectures on two widely used re-ID datasets.

Related Work

In this section, we briefly introduce the prior works on person re-identification and adversarial attack.

Person re-identification. As an important part of surveillance system, person re-ID is widely concerned in recent years. Most of existing works focus on learning discriminative local part information and global information. (Sun et al. 2018) propose PCB to learn part-level features and align them using refined part pooling(RPP). In contrast to PCB where the part region is predefined, MGN (Wang et al. 2018a) propose to learn discriminative information with various granularities. In (Zheng et al. 2019a), a coarse-to-fine pyramidal model is proposed to incorporates local and global information. DG-Net (Zheng et al. 2019c) learn robust appearance features by separating appearance and structure information using generative adversarial networks (GAN). There are also some works trying to solve the misalignment problem caused by various pose and camera views. The work (Zhang et al. 2019) makes use of fine grained semantics to address the misalignment problems to learn better global features. CASN (Zheng et al. 2019b) enforce the attention consistency between images of the same person by proposing an attention-driven siamese network. More recently, BoT (Luo et al. 2019) propose an effective training method that can achieve state-of-the-art performance without complex network structure.

Adversarial attack. Adversarial attack has gained much attention since it is shown that an image can be misclassified by neural networks through applying a tiny perturbation that is imperceptible to human eye (Szegedy et al. 2013). Lots of methods have been proposed to effectively attack the neural network based models (Ho et al. 2019; Dong et al. 2018a; Goodfellow, Shlens, and Szegedy 2014). One step gradient-based method (Goodfellow, Shlens, and Szegedy 2014) was proposed to generate adversarial samples by maximizing the networks prediction error for one single step. I-FGSM (Kurakin, Goodfellow, and Bengio 2016) further improves the attack performance by applying FGSM iteratively. MI-FGSM (Dong et al. 2018a) considers the direction of previous optimization step and achieves better performance. Gradient-based method generate perturbations based on back-propagation gradients, but there are cases where the model parameters could not be accessed. Since it is observed that adversarial perturbations trained on one model can usually fool other models (Szegedy et al. 2013), a typical way to perform attack in this scenario is to train adversarial samples on a source model and use it to attack the target model (Liu et al. 2017). In contrast to most of the attack methods which generate perturbations specific to image, (Moosavi-Dezfooli et al. 2017) propose an image-agnostic adversarial perturbation that can fool the CNN models with high probability.

Methodology

UAPs Against Person Re-ID

Problem definition. Let \( \mathcal{X} \subset \mathbb{R}^m \) be the space of pedestrian images, and given a model \( f_\theta \), where \( \theta \) is its parameter set, e.g., a CNN which embeds an image \( x \in \mathcal{X} \) to a feature vector space \( f_\theta(x) \in \mathbb{R}^n \). Typically, the model has been trained with the objective that the distances of features with the same person IDs (positive pairs) are smaller than those with different IDs (negative pairs). The goal of UAP attack is to seek the perturbation vectors \( u \) with small magnitudes that fool the model \( f_\theta \) where the negatives rank above almost all the positives on almost all samples in \( \mathcal{X} \).
Algorithm 1 Learning Universal Adversarial Perturbations
Against Person Re-Identification

**Input:** Database \( D \), source model \( f_\theta \), stop criterion \( \xi \), number of epochs \( T \), \( \ell_p \) norm of perturbation bound \( \mu \), momentum \( \beta \), learning rate \( \eta \).

**Output:** The universal perturbation vector \( u \).

1. Initialize \( u \leftarrow 0 \).
2. **while** \( mAP > \xi \) and epochs \( \leq T \) **do**
3. **for** each sample \( q \in D \) **do**
4. Update perturbation vector \( u \) using Eq. (8).
5. **if** \( u \) get saturated **then**
6. Constrain perturbation vector \( u \) using the projection defined in Eq. (9).
7. **end if**
8. **end for**
9. **end while**

Formally, let \( D \subseteq \mathcal{X} \) be a dataset. For a query person image \( q \in D \), let \( D^+_q \) denotes the positive set, and \( D^-_q \) denotes the negative set in \( D \), respectively. Given a distance metric \( d(\cdot, \cdot) \) and for any query image \( q \in D \), let \( (x_1, x_2, \ldots, x_t) \) be a ranking list with respect to \( q \). Let \( D \) denote the positive set, and \( \neg \) denote the negative set in \( D \). For any query image \( q \), we consider a more practically situation where only the positive set is available.

We assume the CNN embedding \( f_\theta(\cdot) \) is \( \ell_2 \) normalized, and let \( d(f_\theta(q'), f_\theta(x_i)) \) be the Euclidean distance of CNN embeddings of the attacked query and gallery image which lies in \([0, 2]\). Given a histogram bin number \( b \) and equally divide the distance interval into \( b \)-1 parts with length \( \Delta = \frac{2}{b-1} \). Similar to (Revaud et al. 2019; He et al. 2018), we define a soft indicator function \( \delta : \mathbb{R} \times \{1, 2, \ldots, b\} \rightarrow [0, 1] \), thus the contribution of the each gallery instance \( x_i \) to the \( k \)-th bin \((k \in [1, b]) \) is calculated by:

\[
\delta(x_i, k) = \max \left( 1 - \frac{\|d(f_\theta(q'), f_\theta(x_i)) - (k-1)\Delta\|_1}{\Delta} \right),
\]

where \( \| \cdot \| \) is the \( \ell_1 \) norm. In this way, we can calculate precision at each bin instead of at each rank position to avoid the sorting operation. The precision at the \( k \)-th bin is:

\[
\hat{P}_k = \sum_{k'=1}^{k} \sum_{i=1}^{l} \delta(x_i, k') \cdot \mathbb{1}[x_i \in D^+_q],
\]

Finally the approximated average precision is:

\[
AP = \frac{1}{|D^+_q|} \sum_{k=1}^{b} \hat{P}_k \cdot \left( \sum_{i=1}^{l} \delta(x_i, k) \cdot \mathbb{1}[x_i \in D^+_q] \right).
\]

Learning method. Due to the sorting operation, \( AP \) is non-smooth and non-differential, thus it is difficult to optimize \( AP \) in neural networks using the standard backpropagation training method. Here we adopt an approximation method proposed in (Revaud et al. 2019), where the non-differential sorting operation is replaced by histogram binning and soft indication.

The generalizability of UAPs across different models and different environments is important for person re-ID attack in practice. However, we observe that our UAP attack
learned using the above method works well in the same model and in the same dataset, but the performance of attack falls significantly when applying across different ones. The reasons account for the generalization problem is that the only training objective is to minimize \( AP \), which depends on the CNN architecture \( f \) and its parameter set \( \theta \), thus the learned UAP vector \( u \) can be biased by \( f_\theta \). In order to improve the attack generalizability, we propose an additional regularization to train the UAP vector. The method is explained as follows.

**Total variation prior.**  Total variation (TV) is a measure of the complexity of an image with respect to its spatial variation. It was first introduced by Rudin, Osher, and Fatemi (ROF) in their pioneering work on image denoising (Rudin, Osher; and Fatemi 1992). As a statistical character (Huang and Mumford 1999) that the corrupted noisy images often have much larger TV values than natural clean images, TV has become a successful prior in many image restoration tasks (Chan and Wong 1998; Beck and Teboulle 2009). In our attack problem, we observed that 1) UAPs trained using the \( AP \) loss function can remarkably increase TVs of the attacked images, which makes them statistically different from uncontaminated images; 2) TV is an general image prior works on the inputs which is independent to the CNN model \( f_\theta \), thus it is helpful to increase the generalizability of \( u \) by avoiding falling into local minima on a specific model \( f_\theta \). The TV regularizer on the attacked image \( q' = q + u \) can be formulated as follows:

\[
TV = \sum_{i=1}^{m} \| (\nabla q')_i \|_1 = \sum_{i=1}^{m} (\| (\partial_x q')_i \|_1 + \| (\partial_y q')_i \|_1),
\]

(10) where \( q' \) is the input, \( \nabla \) denotes the gradient operation which is expressed in an anisotropic manner in 2D with two directions \( \partial_x \) and \( \partial_y \).

When considering TV regularizer, the training loss function becomes:

\[
\mathcal{L} = AP + \lambda \cdot TV,
\]

(11) where \( \lambda \) is a parameter balancing the \( AP \) and TV terms defined in Eqs. (7) and (10), respectively. Thus the objective is to train a UAP vector \( u \) to minimize average precision as much as possible, while regularizes TV values of attacked images at the same time. An example is shown in Fig. 2, where (a)-(b) display two UAPs \( u_1 \) and \( u_2 \) trained without and with TV regularizer, (c)-(e) display the input \( q \) and attacked \( q + u_1 \) and \( q + u_2 \) images, and (f)-(g) display pixel values \( q \) of a scan line, their gradients \( \partial_x q \), attacked pixel values \( q + u_1 \) and \( q + u_2 \), and their gradients \( \partial_x (q + u_1) \) and \( \partial_x (q + u_2) \), respectively. As can be seen that, \( q + u_2 \) varies smoothly as well as \( q \), thus \( \partial_x (q + u_2) \) is close to \( \partial_x q \). On the other hand, there are large variations in \( q + u_1 \) and thus significantly increase the magnitude of \( \partial_x (q + u_1) \). We calculate average TVs of the original and attacked images using \( u_1 \) and \( u_2 \) on two datasets, and present the results in Fig. 3. As it can be seen, \( q + u_1 \) can increase TV remarkably, while \( q + u_2 \) maintains the same level as \( q \) with an additional perturbation vector.

In contrast to \( AP \), TV regularization is applied on the input image, thus it is independent to both the CNN architecture \( f \) and the distribution of its parameter \( \theta \). Consequently, TV regularizer rectifies \( u \) as a prior of natural image to avoid biasing on the model \( f_\theta \) exclusively. We show that such a regularization can significantly improve the performance of attack across different models in the experiment. TV is computational efficient, and the learning method can be upgraded with small overheads. The only difference to the vanilla learning method is computing the composite loss function for error back-propagation in step-4 of Algorithm 1 and other steps remain the same.


## Experiments

### Experimental Settings

In this section, we perform extensive experiments to validate the effectiveness of our proposed methods. The experimental results and visualization will also be discussed in this section.

### Evaluation metric

In order to evaluate the attack performance quantitatively, inspired by the fooling rate metric in classification tasks, we propose a new metric named mean drop rate (mDR) defined as follows:

$$mDR(q, u) = \frac{mAP(q) - mAP(q + u)}{mAP(q)}, \quad (12)$$

where $mAP(q)$ denotes the mean average precision of $q$, and $u$ refers to our UAP vector. A positive $mDR$ means the attack is successful, and higher the value, better the attack.

### Datasets

We perform experiments on three publicly available and widely used datasets DukeMTMC-reID, Market-1501, and MSMT17.

- **DukeMTMC-reID** (Ristani et al. 2016) contains 36411 images of 1404 individuals captured by 8 different cameras in campus. 16522 images of 702 identities are used for training and 19889 images of 702 other identities for testing. During test, 2228 images are used as query images and others used as gallery images.

- **Market-1501** (Zheng et al. 2015) contains 32668 pictures corresponding to 1501 identities taken by 6 cameras. 12936 images of 751 identities are used for training and 19732 images of the rest 750 identities used for testing. During test, 3368 images are used for query and others used as gallery images.

- **MSMT17** (Wei et al. 2018) contains 126441 images of 4101 identities taken by 15 cameras in various scenes with 12 time slots. 32621 images of 1041 identities are used for training and 93820 images of the rest 3060 identities are used for testing.

### CNN Models

BoT (Luo et al. 2019) propose an effective training method for person re-ID task which achieves state-of-the-art performance on DukeMTMC-reID and Market-1501. We follow the training method proposed in (Luo et al. 2019) and train models with five different backbones.

We achieve state-of-the-art performance with ResNet-50 and DenseNet-121. The results with different backbones are reported in Table 1.

### Implementation details

We perform extensive experiments to validate the effectiveness of our proposed methods. We first perform UAP attack within the same dataset. In this setting, we train our UAP and test the attack performance on all the models trained on this dataset. To show the generality of our UAP to different datasets, we further test the attack performance of our UAP across datasets. Specifically, in this setting, we train UAP using images and models trained on MSMT17. In both settings, we train UAP with 800 images randomly selected from the corresponding training set. The input images are resized to 256 x 128, and we set $\lambda = 10$, the stop criterion $\xi = 0.01$, momentum $\beta = 0.4$, the learning rate $\eta = 0.25$, and the maximum number of epochs $T = 100$.

### Experimental Results

#### Cross CNN architecture attack within the same dataset.

We report the mDR on DukeMTMC-reID and Market-1501 using $\{p = 2, \mu = 2000, \lambda = 10\}$ and $\{p = \infty, \mu = 10, \lambda = 10\}$ in Table 2 and Table 4, respectively. For every dataset, each element in the table represents the mDR achieved by UAP when using the model of the row as source model and the model of the column as target model. The diagonal of the table marked in gray is the result where we train and test UAP on the same CNN architecture. As is seen from the table, mDR of our UAP attack is 97%+ using AP loss, which means that the ranking list is almost entirely disrupted under UAP attack. However, its generalization ability to other CNN architectures is unsatisfying. For instance, the mDR is only 10.24% when using UAP trained on ShuffleNet to attack SENet-154 on Market-1501. On the other hand, our method with TV regularization significantly improves the generalization ability to other CNN architectures with minor performance loss under the same backbone. Specifically, we improve the attack performance from 23.57% to 92.42% in terms of mDR on DukeMTMC when using UAP trained on ResNet-50 to attack DenseNet-121 as shown in Table 2.

#### Cross CNN architecture and dataset attack.

In this subsection, we evaluate UAP attack across CNN model and datasets. We use the UAPs trained on the MSMT17 dataset to attack the models trained on Market-1501 and DukeMTMC-reID. The attack performance in this scenario is reported in Table 3 and Table 5 using $\{p = 2, \mu = 2000, \lambda = 10\}$ and $\{p = \infty, \mu = 10, \lambda = 10\}$, respectively. The diagonal of the table marked in gray represents the results in the special case where the source and target models share the same CNN architecture but with different parameters. It can be seen from the table that UAP attack can still remarkably disrupt the ranking list when the same CNN architecture is employed. Similar to previous experiments, the attack performance falls considerably when testing on a different CNN model. For example, when VGG-16 is used as backbone, mDR falls from 71.58% when attacking the model with VGG-16 architecture to 9.57% when at-

### Table 1: The performances before (mAP) and after attack (mAP†) of the re-ID method (Luo et al. 2019) with different backbones. UAPs are trained using the same backbones on the training subset of the same data.

| Models   | DukeMTMC-reID | Market-1501 |
|----------|---------------|-------------|
|          | mAP | mAP† | mAP | mAP† |
| ResNet-50| 75.90% | 1.40% | 85.32% | 0.53% |
| DenseNet-121 | 73.47% | 0.77% | 81.49% | 0.59% |
| VGG-16   | 66.39% | 0.56% | 76.74% | 0.36% |
| SENet-154| 66.93% | 2.84% | 74.16% | 1.10% |
| ShuffleNet | 66.92% | 0.28% | 76.02% | 0.24% |
tackling the model with SENet-154 as backbone in Table 3. An interesting phenomenon is found in ShuffleNet where it generalizes well to different datasets but the performance is unsatisfactory when attacking a different CNN model. The reason may reside in that ShuffleNet is a lightweight CNN with much fewer parameters, which is helpful to avoid overfitting on datasets. However, its special operations such as channel shuffle are not used in other CNNs. On the other hand, the further proposed method with additional TV regularization increases the attack performance significantly in most cases. Specifically, the attack performance is improved from 8.89\% m\textit{DR} to 88.19\% in terms of m\textit{DR} when using the UAP trained on MSMT17 with ResNet-50 to attack the SeNet-154 architecture trained on Market-1501.

We also evaluate the performance of UAP attack against several other re-ID methods including PCB (Sun et al. 2018), SA-reID (Wang et al. 2018b) and MGN (Wang et al. 2018a), which all use ResNet-50 as their backbone. Here we use our UAPs trained on DenseNet-121 for attacking. As can be seen from Table 6, these approach are also vulnerable to UAPs, especially to the UAP trained with TV regularizer.

**Visualization** In this subsection, we visualize the perturbations trained on different networks without and with TV regularization in Fig. 4. Perturbations in the first and second row are trained on DukeMTMC-reID and Market-1501, respectively. As it can be seen, the differences between perturbations trained on different CNN architectures are larger than those between perturbations trained on different datasets. This is consistent with our experimental results that the attack performances of UAPs across datasets is better than those across models. On the other hand, the perturbations trained with TV regularization are apparently more smooth with less fine-grained textures than those trained without TV regularization, and share some similar patterns such as dots among different models and datasets. We also find it interesting that there is a blurred silhouette in the center of each perturbation trained with TV, which indicates that the trained UAPs learn some inherent structures of the pedestrian datasets.

---

### Table 2: UAP attack performance evaluation within the same dataset using $\|u\|_2 \leq 2000$.

| Mean Drop Rate (mDR) | Test: DukeMTMC-reID | Test: Market-1501 |
|----------------------|---------------------|-------------------|
|                      | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet |
| ResNet-50 + TV        | 98.15%     | 23.57%       | 37.21%   | 11.08%    | 18.05%     | 99.38%     | 25.93%       | 38.12%   | 14.45%    | 30.07%     |
| DenseNet-121 + TV     | 96.64%     | 92.42%       | 91.36%   | 88.18%    | 87.82%     | 96.22%     | 81.27%       | 82.19%   | 76.25%    | 66.29%     |
| VGG-16 + TV           | 89.35%     | 98.95%       | 75.96%   | 48.14%    | 75.37%     | 55.59%     | 99.27%       | 60.49%   | 18.05%    | 58.47%     |
| SENet-154 + TV        | 92.20%     | 98.45%       | 91.56%   | 86.36%    | 91.04%     | 74.66%     | 98.41%       | 86.20%   | 68.52%    | 68.12%     |
| ShuffleNet + TV       | 92.02%     | 63.88%       | 99.15%   | 24.23%    | 32.32%     | 24.76%     | 29.09%       | 99.53%   | 12.49%    | 31.75%     |

---

Please note that the above table and text are the natural representation of the document content.
### Table 3: UAP attack performance evaluation across different datasets using $\|u\|_2 \leq 2000$.  

| Mean Droop Rate (mDR) | Test: DukeMTMC-reID | Test: Market-1501 |
|-----------------------|----------------------|------------------|
|                       | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet |
| ResNet-50             | 58.09%   | 83.56%   | 79.82% | 80.53% | 70.94% | 71.76% | 67.59% | 79.82% | 86.77% | 88.19% | 83.08% |
| DenseNet-121          | 30.82%   | 85.19%   | 90.35% | 88.20% | 77.91% | 72.00% | 29.04% | 55.05% | 35.17% | 15.55% | 47.36% |
| VGG-16                | 39.59%   | 75.06%   | 81.40% | 74.08% | 78.43% | 67.04% | 22.67% | 19.35% | 83.04% | 9.64%  | 42.97% |
| SENet-154             | 58.95%   | 70.67%   | 71.42% | 78.58% | 87.83% | 73.80% | 54.47% | 56.38% | 93.61% | 65.87% | 53.18% |
| ShuffleNet            | 6.74%    | 73.46%   | 74.04% | 85.47% | 88.63% | 91.10% | 7.88%  | 7.25%  | 9.45%  | 8.93%  | 98.83% |

### Table 4: UAP attack performance evaluation within the same dataset using $\|u\|_\infty \leq 10$.  

| Mean Droop Rate (mDR) | Test: DukeMTMC-reID | Test: Market-1501 |
|-----------------------|----------------------|------------------|
|                       | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet |
| ResNet-50             | 98.49%   | 92.14%   | 79.56% | 81.51% | 71.28% | 71.68% | 99.39% | 93.16% | 92.17% | 97.10% | 87.04% |
| DenseNet-121          | 89.89%   | 84.55%   | 98.87% | 73.15% | 54.83% | 71.12% | 66.11% | 99.36% | 73.00% | 32.44% | 79.68% |
| VGG-16                | 65.91%   | 81.27%   | 98.44% | 80.00% | 74.58% | 84.26% | 84.69% | 98.48% | 89.37% | 85.86% | 86.25% |
| SENet-154             | 76.18%   | 79.27%   | 72.49% | 83.32% | 97.27% | 74.75% | 24.44% | 23.17% | 99.42% | 14.45% | 42.75% |
| ShuffleNet            | 78.30%   | 73.78%   | 11.95% | 84.18% | 84.93% | 93.44% | 59.82% | 51.33% | 58.05% | 98.95% | 74.18% |

### Table 5: UAP attack performance evaluation across different datasets using $\|u\|_\infty \leq 10$.  

| Mean Droop Rate (mDR) | Test: DukeMTMC-reID | Test: Market-1501 |
|-----------------------|----------------------|------------------|
|                       | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet | ResNet-50 | DenseNet-121 | VGG-16 | SENet-154 | ShuffleNet |
| ResNet-50             | 70.22%   | 83.36%   | 83.01% | 88.12% | 79.34% | 67.45% | 83.70% | 82.93% | 89.38% | 85.5%  | 68.38% |
| DenseNet-121          | 33.44%   | 82.37%   | 91.24% | 89.94% | 87.55% | 81.67% | 34.76% | 54.37% | 35.65% | 22.4%  | 49.97% |
| VGG-16                | 78.50%   | 75.80%   | 20.71% | 47.46% | 11.01% | 38.89% | 75.99% | 92.64% | 92.43% | 91.21% | 85.66% |
| SENet-154             | 70.06%   | 76.61%   | 71.75% | 89.02% | 85.12% | 72.32% | 72.25% | 81.51% | 93.34% | 93.06% | 73.4%  |
| ShuffleNet            | 11.76%   | 80.35%   | 6.97%  | 10.44% | 7.41%  | 89.65% | 12.26% | 12.47% | 16.33% | 17.98% | 97.73% |

### Table 6: Attacking other re-ID methods (which employ ResNet-50 as their backbone) using UAPs trained on DenseNet-121.  

In this paper, we attempt to attack state-of-the-art re-ID models using universal adversarial perturbations. To effectively disorder the ranking list of re-ID models, we propose to minimize the average precision ($AP$) from the global ranking perspective. Extensive experiments show that our UAP attack method can significantly decrease the performance of re-ID models. Specifically, we achieve $97\% + mDR$ on all the five models trained on two widely used public datasets. In addition, we further observe that the attack performance of UAP falls down noticeably when applied to models with different CNN architectures or trained on different datasets. Thereafter we propose a $TV$ minimization based approach to enhance the generalization ability of UAP attack. Extensive experiments show the effectiveness of the proposed attack method against person re-ID.

**Conclusion**

In this paper, we attempt to attack state-of-the-art re-ID models using universal adversarial perturbations. To effectively disorder the ranking list of re-ID models, we propose to minimize the average precision ($AP$) from the global ranking perspective. Extensive experiments show that our UAP attack method can significantly decrease the performance of re-ID models. Specifically, we achieve $97\% + mDR$ on all the five models trained on two widely used public datasets. In addition, we further observe that the attack performance of UAP falls down noticeably when applied to models with different CNN architectures or trained on different datasets. Thereafter we propose a $TV$ minimization based approach to enhance the generalization ability of UAP attack. Extensive experiments show the effectiveness of the proposed attack method against person re-ID.
References

Bai, S.; Li, Y.; Zhou, Y.; Li, Q.; and Torr, P. H. 2019. Adversarial metric attack for person re-identification. arXiv preprint arXiv:1901.10650.

Beck, A., and Teboulle, M. 2009. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE transactions on image processing 18(11):2419–2434.

Chan, T. F., and Wong, C.-K. 1998. Total variation blind deconvolution. IEEE transactions on Image Processing 7(3):370–375.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018a. Boosting adversarial attacks with momentum. In CVPR, 9185–9193.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018b. Boosting adversarial attacks with momentum. In CVPR, 9185–9193.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

He, K.; Cakir, F.; Adel Bargal, S.; and Sclaroff, S. 2018. Hashing as tie-aware learning to rank. In CVPR, 4023–4032.

Hendrik Metzen, J.; Chaitanya Kumar, M.; Brox, T.; and Fischer, V. 2017. Universal adversarial perturbations against semantic image segmentation. In ICCV, 2755–2764.

Ho, C.-H.; Leung, B.; Sandstrom, E.; Chang, Y.; and Vasconcelos, N. 2019. Catastrophic child’s play: Easy to perform, hard to defend adversarial attacks. In CVPR, 9229–9237.

Huang, J., and Mumford, D. 1999. Statistics of natural images and models. In CVPR, volume 1, 541–547. IEEE.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into transferable adversarial examples and black-box attacks. In ICLR 2017.

Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; and Hu, J. 2018. Pose transferable person re-identification. In CVPR, 4099–4108.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In CVPRW, 0–0.

Mei, S., and Zhu, X. 2015. Using machine teaching to identify optimal training-set attacks on machine learners. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In CVPR, 1765–1773.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In CVPR, 2574–2582.

Mopuri, K. R.; Ganeshan, A.; and Radhakrishnan, V. B. 2018. Generalizable data-free objective for crafting universal adversarial perturbations. IEEE transactions on pattern analysis and machine intelligence.

Revaud, J.; Almazan, J.; de Rezende, R. S.; and de Souza, C. R. 2019. Learning with average precision: Training image retrieval with a listwise loss. arXiv preprint arXiv:1906.07589.

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, 17–35. Springer.

Rudin, L. I.; Osher, S.; and Fatemi, E. 1992. Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena 60(1-4):259–268.

Saquib Sarfraz, M.; Schumann, A.; Eberle, A.; and Stiefelhagen, R. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In CVPR, 420–429.

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, 480–496.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018a. Learning discriminative features with multiple granularities for person re-identification. In 2018 ACM Multimedia Conference on Multimedia Conference, 274–282. ACM.

Wang, H.; Fan, Y.; Wang, Z.; Jiao, L.; and Schiele, B. 2018b. Parameter-free spatial attention network for person re-identification. arXiv preprint arXiv:1811.12150.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In CVPR, 79–88.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2019. Densely semantically aligned person re-identification. In CVPR, 667–676.

Zhao, G.; Zhang, M.; Liu, J.; and Wen, J.-R. 2019. Unsupervised adversarial attacks on deep feature-based retrieval with gan. arXiv preprint arXiv:1907.05793.

Zhedong Zheng, Liang Zheng, Z. H. Y. 2019. Open set adversarial examples. arXiv preprint arXiv:1809.02681.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In ICCV.

Zheng, F.; Deng, C.; Sun, X.; Jiang, X.; Guo, X.; Yu, Z.; Huang, F.; and Ji, R. 2019a. Pyramidal person re-identification via multi-loss dynamic training. In CVPR, 8514–8522.

Zheng, M.; Karanam, S.; Wu, Z.; and Radke, R. J. 2019b. Re-identification with consistent attentive siamese networks. In CVPR, 5735–5744.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019c. Joint discriminative and generative learning for person re-identification. In CVPR, 2138–2147.