An Effective Adversarial Attack on Person Re-Identification in Video Surveillance via Dispersion Reduction

YU ZHENG (Student Member, IEEE), YANTAO LU (Member, IEEE), AND SENEM VELIPASALAR (Senior Member, IEEE)

Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, USA

Corresponding author: Yu Zheng (yzheng04@syr.edu)

ABSTRACT

Person re-identification across a network of cameras, with disjoint views, has been studied extensively due to its importance in wide-area video surveillance. This is a challenging task due to several reasons including changes in illumination and target appearance, and variations in camera viewpoint and camera intrinsic parameters. The approaches developed to re-identify a person across different camera views need to address these challenges. More recently, neural network-based methods have been proposed to solve the person re-identification problem across different camera views, achieving state-of-the-art performance. In this paper, we present an effective and generalizable attack model that generates adversarial images of people, and results in very significant drop in the performance of the existing state-of-the-art person re-identification models. The results demonstrate the extreme vulnerability of the existing models to adversarial examples, and draw attention to the potential security risks that might arise due to this in video surveillance. Our proposed attack is developed by decreasing the dispersion of the internal feature map of a neural network to degrade the performance of several different state-of-the-art person re-identification models. The experimental results show that our proposed attack outperforms the state-of-the-art attack models on the best performing person re-identification approaches by a large margin, and results in the most drop in the mean average precision values.

INDEX TERMS

Adversarial attack, adversarial examples, person re-identification.

I. INTRODUCTION

In order to continuously track targets across multiple cameras with disjoint views, it is essential to re-identify the same target across different cameras. However, this is a very challenging task due to several reasons including changes in illumination and target appearance, and variations in camera intrinsic parameters and viewpoint.

There has been great interest and significant progress in person re-identification (ReID) [1]–[6], which is important for security and wide-area surveillance applications as well as human computer interaction systems. Fueled by the new models, including the neural network-based approaches, proposed in recent years, the performance of person ReID approaches has improved significantly. For instance, the rank-1 accuracy of the state-of-the-art method on the Market 1501 dataset [7] is 94.8% [1], which has increased from 44.4% when the dataset was initially released in 2015.

In this paper, we demonstrate the effectiveness of an attack model in generating adversarial examples (AEs) for the person ReID application, attack multiple state-of-the-art person ReID models, and also compare the performance of the presented attack approach with other state-of-the-art attack models via an extensive set of experiments on various person ReID benchmark datasets. One of our goals is to demonstrate the extreme vulnerability of multiple state-of-the-art person ReID approaches to this attack, and draw the attention of the research community to the existing security risks. In person ReID, the paired probe and gallery images are...
expected to have high similarity. However, by adding human-imperceptible perturbations to the probe images, the models are easily fooled even when the probe images appear the same as the original images.

Adversarial examples [8]–[10] have been extensively investigated recently in image classification [10], [11], object detection [12]–[14] and semantic segmentation [12], [15], etc. However, relatively less attention has been paid to the robustness of person ReID models. Bai et al. [16] proposed an adversarial metric attack, which targets on fooling the distance metrics in person ReID systems. An early attempt for the defense shows that a metric-preserving network can be applied to defend against such attack. Zheng et al. [17] propose Opposite Direction Feature Attack (OFDA) to generate adversarial examples/queries for retrieval tasks such as person ReID. The idea is to push away the feature of the adversarial query in the opposite direction of the original feature.

In this paper, we present and employ an effective approach to generate adversarial examples targeting person ReID methods. Our approach [18] is referred to as the Dispersion Reduction (DR), and it is a black-box attack. The main idea behind our approach is reducing the “contrast” of the internal feature map of a neural network. The intuition is that just like reducing the contrast of an image would make the objects less recognizable or distinguishable, reducing the contrast of an internal feature map would have a similar effect on recognizability of objects by the neural network. In our previous work [18], we showed the transferability of the DR attack across different tasks including object detection, classification and text recognition. The contribution of this work includes the following: We adapt the DR attack for the person ReID problem, and perform an extensive comparison and evaluation on different state-of-the-art methods and multiple benchmarks. In addition, we compare the performances of multiple attack methods. We show that making a feature map “featureless”, through dispersion reduction, is very-well suited to fool any state-of-the art ReID model. Moreover, we use different network models (different from the model used by the victim ReID networks) as the source model to generate the adversarial examples and show the effectiveness and generalizability of our attack approach. We also analyze the effect of the perturbation budget on the attack performance.

The rest of this paper is organized as follows: The related works on both person ReID and attack models are summarized in Section II. The proposed dispersion reduction-based attack approach, and the methodology are described in Section III. The experimental results are presented in Section IV, and the paper is concluded in Section V.

II. BACKGROUND AND RELATED WORK

A. PERSON ReID METHODS

Various person re-identification (ReID) approaches have been proposed in the past [19], which can be classified into different categories. There have been methods based on distance learning [20]–[26], on feature design and selection [27]–[33], and on mid-level feature learning [34]–[38].

Many works relied on color transformation and statistic models for person re-identification. Cheng and Piccardi [39] applied a cumulative color histogram transformation and employed an incremental major color spectrum histogram representation. Trajectory matching, height estimation and illumination-tolerant color representation were used by Madden and Piccardi [40]. Chae and Jo [41] employed a Gaussian Mixture Model (GMM) for the segmented regions in a person, and used a ratio of the GMMs to identify the same person. The Brightness Transfer Function (BTF) and its variants have been introduced to improve the matching performance. Porikli [42] proposed BTF for inter-camera color calibration. Later, Javed et al. [43] and Posser et al. [44] proposed the Mean Brightness Transfer Function (mBTF) and the Cumulative Brightness Transfer Function (cBTF), respectively. Datta et al. [45] presented the Weighted-BTF (wBTF), and Bhuiyan et al. [46] presented the Minimum Multiple Brightness Transfer Function (Min-MCBTF) to model the appearance variation by using a learning approach. However, it was assumed that multiple consecutive images are available for training, which is not the case for the commonly used benchmark datasets.

Researchers then focused on combining the features and distance metrics at the same time. Liao et al. [47] proposed Local Maximal Occurrence (LOMO) and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) for person ReID. Chen et al. [48] formulated a new view-specific person ReID framework, referred to as camera correlation aware feature augmentation (CRAFT). In this framework, cross-view feature adaptation is performed by measuring cross-view correlation from visual data distribution and carrying out adaptive feature augmentation. Matsukawa et al. [49] proposed the hierarchical Gaussian Of Gaussian (GOG) descriptor, which generates discriminative and robust features that describe color and textural information simultaneously. An image is first divided into horizontal strips. Then, local patches in the strips are modeled using a Gaussian distribution. Köstinger et al. [50] proposed the KISSME, which is a statistical inference perspective to address the problem of metric learning.

More recent works employ neural networks and achieve state-of-the-art performance in person ReID. Zheng et al. [1] proposed DG-Net that encompasses a generative module, which separately encodes a specific person into both appearance and structure. It also integrates a discriminative module that shares the appearance encoder with the generative module. As a result, the high-quality cross-id composed images are fed back to the appearance encoder online and used to improve the model for discriminative module. Zhang et al. [3] proposed AlignedReID performing automatic part alignment during learning, without requiring extra supervision or pose estimation. By learning jointly on global and local features, it aims to address existing drawbacks. Xie et al. [51] proposed PLR-OSNet, which introduces Part-level resolution (PLR)
into Omni-Scale Network (OSNet) [52]. It has two branches including both global and local feature representations. The global branch adopts a global-max-pooling layer, while the local branch employs a part-level feature resolution scheme for producing only a single ID-prediction loss, which is in contrast to existing part-based methods.

B. ADVERSARIAL ATTACK METHODS

Szegedy et al. [9] introduced the adversarial images, which can fool the Convolutional Neural Network (CNN)-based models, and cause misclassification by adding small perturbations to the original images. In one of the earlier works, Goodfellow et al. [53] proposed fast gradient sign method (FGSM), which generates AEs in one step. Several works extended this by iteratively updating the AEs with multi-step attacks including the basic iterative method (BIM) [10], deep fool [54], momentum iterative method [11], Diverse Inputs Method (DIM) [55] and Translation-Invariant (TI) attacks [56]. Compared with FGSM, the iterative methods generate a smaller perturbation, which makes the adversarial examples even more imperceptible to human eyes.

The transferability property of adversarial examples motivated research on black-box adversarial attacks. To perform black-box attacks, methods have been introduced [57], [58], which employ a substitute model that is trained to mimic the target model. Gradient-free attacks use feedback on query data, i.e., soft predictions [59], [60] or hard labels [61]. However, these aforementioned approaches require feedback from the target model, which is not practical in some scenarios. More recently, several methods have been proposed, which study the attack generation process itself. In general, an iterative attack [8], [62], [63] achieves a higher attack success rate than a single-step attack [53] in a white-box setting, but performs worse when transferred to other models. Below, we will summarize some of these attack methods.

1) GRADIENT-BASED ADVERSARIAL ATTACK METHODS

Fast Gradient Sign Method (FGSM) [53] generates the adversarial example \( x_{adv} \) by linearizing the loss function in the input space and performing one-step update as follows

\[
\begin{align*}
  x_{adv} &= x_{real} + \varepsilon \cdot \text{sign}(\nabla_x J(x_{real}, y)), \\
  \nabla_x J(x_{real}, y) &= \text{gradient of the loss function w.r.t. } x, \text{ sign(·) is the sign function that constrains the perturbation in } L_\infty \text{ norm bound. FGSM can generate more transferable adversarial examples, however, it may not be as effective in white-box attacks [10].}
\end{align*}
\]

Basic Iterative Method (BIM) [10] extends FGSM by updating the gradient in a multi-step manner with a small step size \( \alpha \), which can be expressed as

\[
\begin{align*}
  x_{adv}^{t+1} &= x_{adv}^t + \alpha \cdot \text{sign}(\nabla_x J(x_{adv}^t, y)),
\end{align*}
\]

where \( x_{adv}^0 = x_{real} \). BIM clips \( x_{adv}^t \) after each update, or sets \( \alpha = \varepsilon / T \), with T being the number of iterations to ensure that they are in an \( \varepsilon \)-neighbourhood of the real image.

Mutual Iterative Fast Gradient Sign Method (MI-FGSM) [11] integrates momentum term into the iterative attack process. The update procedure is as follows

\[
\begin{align*}
  g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x J(x_{adv}^t, y)}{\| \nabla_x J(x_{adv}^t, y) \|_1}, \\
  x_{adv}^{t+1} &= x_{adv}^t + \alpha \cdot \text{sign}(\nabla_x J(x_{adv}^t, y)),
\end{align*}
\]

where \( g_t \) collects the gradient information up to the \( t \)-th iteration, and \( \mu \) is the decay factor.

Diverse Inputs Method (DIM) [55] applies random and differentiable transformations to the input images with probability \( p \) and maximizes the loss function with respect to these transformed inputs. The transformed images are fed into the classifier for gradient calculation. Such transformation includes random resizing and padding with a given probability \( p \). This method can be combined with the momentum-based method to further improve the transferability.

2) TRANSLATION-INARIANT ATTACK METHODS

Translation-Invariant (TI) attack methods have been proposed by Dong et al. [56] to further improve the transferability on white-box models. The authors notice the difference between the discriminative regions used by defenses to identify object categories and the normal trained models. Rather than optimizing the objective function, TI attack method uses a set of translated images to optimize the adversarial examples as

\[
\begin{align*}
  \arg \max_{x_{adv}} \sum_{i,j} w_{ij} J(T_{ij}(x_{adv}), y), \\
  \text{s.t. } \| x_{adv} - x_{real} \|_\infty \leq \varepsilon
\end{align*}
\]

where \( T_{ij}(x) \) is the translation operation that shifts image \( x \) by \( i \) and \( j \) pixels along the two-dimensions, respectively, and \( w_{ij} \) is the weight for the loss \( J(T_{ij}(x_{adv}), y) \).

Note that the TI can be integrated into any gradient-based attack such as FGSM or DIM. For example, the translation-invariant method for fast gradient sign method (TI-FGSM) updates as

\[
\begin{align*}
  x_{adv} &= x_{real} + \varepsilon \cdot \text{sign}(W \ast \nabla_x J(x_{real}, y)),
\end{align*}
\]

Also, the translation-invariant method for diverse inputs method (DIM) can also be obtained by a similar approach.

III. PROPOSED APPROACH

In this section, we will describe the dispersion reduction-based attack on the person ReID application.

A. NOTATION

We use \( x_{real} \) to denote the original query image, and \( f(·) \) to denote a deep neural network classifier. The output feature map at layer \( k \) is denoted by \( \tilde{x} \), where \( \tilde{x} = f(x_{real}) \) at the first time step. For each step afterwards, we calculate the dispersion, which is denoted as \( g(·) \), and the gradient of dispersion as \( \nabla_{x_{real}} g(\tilde{x}) \) to update the adversarial examples \( x_{adv} \). More details will be provided in the following section.
B. DISPERSION REDUCTION

For person ReID, the existing models are trained with various benchmark datasets, which have different labeling schemes. Thus, compared to the image classification problem, person ReID is more complicated. More specifically, treating and attacking the person ReID models as black-boxes require an approach that is highly transferable and effective at attacking different training datasets and model architectures. The aforementioned existing black-box attacks, however, use a pre-trained model as surrogate, which shares the same training dataset and same labeling scheme with the targeted models. Moreover, most existing attack methods rely on task-specific loss functions, which greatly limits their transferability across tasks and different network models.

In our previous work [18], we showed that Dispersion Reduction (DR) has good transferability properties, and is successful in across task attack scenarios. DR employs a publicly available classification network as the surrogate source model, and attacks models that are used in different computer vision tasks, such as object detection, semantic segmentation and cloud API applications. DR is a black-box attack. Conventional black-box attacks establish a source model as the surrogate, for which the inputs are paired with the labels generated from the target model instead of the ground truth labels. In this way, the source model mimics the behavior of the target model. Our proposed DR attack, on the other hand, does not rely on the labeling system or a task-specific loss function, since DR only accesses top part of the model. Although a source model is still required, there is no need for training with new target models or querying the target model for labels. Instead, a pre-trained public model could simply serve as the source model due to the strong transferability of the proposed DR attack. As shown in Fig. 1, the DR attack reduces the contrast of an internal feature map, by reducing its dispersion, so that the information that is in the feature map becomes indistinguishable, and the following layers are not able to extract any useful information regardless of what kind of computer vision task is at hand. The adversarial example, shown in the second column of Fig. 1, was generated by attacking (reducing the dispersion of) the conv3.3 layer of VGG16 surrogate model. This also results in the distortion of the feature maps of the subsequent layers (e.g. conv5.3) compared to the original image feature maps.

![FIGURE 1. The DR attack reduces the dispersion of an internal feature map. This adversarial example was generated by attacking (reducing the dispersion of) the conv3.3 layer of VGG16 model, which also results in the distortion of the feature maps of the subsequent layers (e.g. conv5.3) compared to the original image feature maps.](image1.png)

![FIGURE 2. Effect of the DR attack when different layers of VGG16 is attacked. Attacking the middle layers results in higher drop in the performance compared to attacking the top or bottom layers.](image2.png)

 infer that for initial layers, the budget $\epsilon$ constrains the loss function to reduce the standard deviation, while for the layers near the output, the standard deviation is already relatively small, and cannot be reduced too much further. Based on this observation, we choose one of the middle layers as the target of the DR attack. More specifically, in our experiments, we attack conv3-3 for VGG16, the last layer of group - A for inception-v3, and the last layer of 2nd group of bottlenecks (conv3-8-3) for ResNet152.

The DR attack is defined as the following optimization problem:

$$\min_x g(f(x_{adv}, \theta))$$

$$s.t. \quad \|x_{adv} - x_{real}\|_\infty \leq \epsilon,$$

(6)
where \( f(\cdot) \) is a deep neural network classifier, \( \theta \) denotes the network parameters, and \( g(\cdot) \) computes the dispersion. As shown in Alg. 1, our proposed DR takes iterative steps that create adversarial examples by reducing the dispersion of an intermediate feature map at layer \( k \). Dispersion describes the extent to which a distribution is stretched or squeezed, and there can be different measures of dispersion such as variance, standard deviation, and gini coefficient [64]. We have chosen to use the standard deviation (denoted by \( g(\cdot) \)) as the dispersion metric due to its simplicity.

Given any feature map, DR iteratively adds perturbation to \( x_{\text{real}} \) along the direction of decreasing standard deviation, and maps it to the vicinity of \( x_{\text{adv}} \) by clipping at \( x \pm \epsilon \). Denote the feature map at layer \( k \) as \( \tilde{g} = f(x_{\text{adv}}) \). DR attack solves the following equation

\[
x_{\text{adv}}^{t+1} = x_{\text{adv}}^t - \nabla_{x_{\text{adv}}} g(\tilde{g}_k) = x_{\text{adv}}^t - \frac{dg(t)}{dt} \cdot \frac{df(x_{\text{adv}})}{dx_{\text{adv}}},
\]

(7)

The code is provided in [65].

**Algorithm 1** Dispersion Reduction Attack

**Input**: classifier \( f \), real image \( x_{\text{real}} \), feature map at layer \( k \), perturbation \( \epsilon \), iteration \( T \) and learning rate \( l \)

**Output**: adversarial example \( x_{\text{adv}} \), s.t. \( \|x_{\text{adv}} - x_{\text{real}}\|_\infty \leq \epsilon \)

1. **procedure** Dispersion reduction
2. \( x_{\text{adv}}^0 \leftarrow x_{\text{real}} \)
3. **for** \( t = 0 \) to \( T-1 \) **do**
4. \( \tilde{g}_k = f(x_{\text{adv}}^t) \)
5. Compute std \( g(\tilde{g}_k) \)
6. Compute gradient \( \nabla_{x_{\text{real}}} g(\tilde{g}_k) \)
7. Update \( x_{\text{adv}}^t \) by:
   8. \( x_{\text{adv}}^{t+1} = x_{\text{adv}}^t - \text{Adam}(\nabla_{x_{\text{real}}} g(\tilde{g}_k), l) \)
9. Project \( x_{\text{adv}}^t \) to the neighbour of \( x_{\text{real}} \)
10. \( x_{\text{adv}}^{t+1} = \text{clip}(x_{\text{adv}}^{t+1}, x_{\text{real}} - \epsilon, x_{\text{real}} + \epsilon) \)

**return** \( x_{\text{adv}}^{T-1} \)

C. VICTIM ReID MODELS AND IMPLEMENTATION DETAILS OF ATTACKS

In order to evaluate the effectiveness of our proposed adversarial DR attack, we adapt it for the person ReID problem, and attack three different state-of-the-art person ReID approaches, namely DG-Net [1], AlignedReID [3] and PLR-OSNet [51]. For person re-identification, both DG-Net and AlignedReID use ResNet-50 [66] as the backbone model, while PLR-OSNet employs the Omni-Scale Network as the backbone. DG-Net reaches 94.8% and 86.0% on the rank-1 accuracy and mean average precision (mAP), respectively, on Market 1501 dataset [7]. AlignedReID achieves 92.6% and 82.3% accuracy [67] on the rank-1 and mAP, respectively, and PLR-OSNet achieves 95.6% and 88.9% accuracy on the rank-1 and mAP, respectively, on the Market 1501 dataset.

We used the pre-trained models for these ReID approaches, provided by the authors on their Github pages [68]–[70]. During training, the images are resized to 256 x 128, which is a strong baseline that can achieve higher accuracy. We reduce the mini-batch size from 16 to 4 to save GPU memory usage on all models and all datasets. The learning rate for DG-Net, AlignedReID and PL-OSNet is 0.0001, 0.0002 and 0.0003, respectively. All models use a decay rate of \( \gamma = 0.1 \), which reduces the learning rate by a factor of 1/10 after \( T \) steps during the training. For DG-Net \( T \) is set to 60000. For AlignedReID and PL-OSNet, \( T \) is set to and 20, respectively.

For each dataset, the images are separated into training and testing folders. We follow the data preparation process described in [68], [69]. After pre-processing, we apply the TI-FGSM and TI-DIM attacks as described in [56], and detailed in the source code on Github page [71].

For our dispersion reduction (DR) attack, we first used the pre-trained ResNet-152 as the source model. The values of the parameters, listed in Algorithm 1, are as follows: \( \epsilon = 4 \), \( l \) (learning rate) = 0.05, \( T = 100 \). The adversarial examples are generated on the test images, and used for testing on the victim ReID models. As mentioned above, both DG-Net and AlignedReID use ResNet-50 [66] as the backbone model. Thus, in order to generate the adversarial examples with different surrogate models, we have also used VGG-16 and InceptionV3, as our source models. As discussed above, we used conv3-3 for VGG16, the last layer of group - A for inception-v3, and the last layer of 2nd group of bottlenecks (conv3-8-3) for ResNet152, as the attack layers. We also analyzed the effects of using different \( \epsilon \) values, and a detailed discussion is provided in the following section.

IV. EXPERIMENTS, RESULTS AND DISCUSSION

As mentioned above, we have used three state-of-the-art ReID methods as victim models, attacked them with the proposed DR attack, and evaluated the performance drop on four different datasets. Moreover, we attacked the same victim models with two other state-of-the-art attack approaches, namely TI-FGSM and TI-DIM [56], [71]. We compared the effectiveness of our DR attack with these other attack methods as well. Moreover, we have used three different network models as the surrogate source model to evaluate and compare the performance drop and the attack effectiveness.

A. DATASETS

We have employed four challenging and commonly used benchmark datasets to demonstrate the effectiveness of the proposed attack. These datasets are Market-1501 [7], CUHK 03 [37], DukeMTMCReID [72] and MSMT 17 [73], which are briefly described below.

1) MARKET-1501

Market-1501 [7] dataset contains 32,217 images of 1501 labeled persons from six camera views. There are 751
TABLE 1. mAP values of victim models, before and after attacks, on different datasets ($\epsilon = 4$). Row 1 to 3 are the results from state-of-the-art baseline methods. Row 4-12 are the performances after attacked by different attack approaches.

| Approach       | Market1501 | CUHK03 | DukeMTMC-reID | MSMT17 |
|----------------|------------|--------|---------------|--------|
| DG-Net (DG) [1] | 86.0       | 61.1   | 74.8          | 52.3   |
| AlignedReID (AR) [3] | 82.3       | 70.7   | 82.8          | 43.7   |
| PLR-OSNet(OS) [51] | 88.9       | 77.2   | -             | -      |
| TI-FGSM-DG [1], [53], [56] | 51.2       | 31.9   | 48.3          | 32.1   |
| TI-FGSM-AR [3], [53], [56] | 58.7       | 29.4   | 49.5          | 30.8   |
| TI-FGSM-OS [51], [53], [56] | 55.9       | 32.1   | 51.2          | -      |
| TI-DIM-DG [1], [53], [56] | 45.4       | 14.2   | 32.7          | 23.4   |
| TI-DIM-AR [3], [53], [56] | 44.9       | 16.5   | 29.4          | 22.6   |
| TI-DIM-OS [51], [53], [56] | 47.7       | 19.1   | 35.6          | -      |
| DR-DG (Proposed) | 28.9       | 7.8    | 20.2          | 15.3   |
| DR-AR (Proposed) | 20.2       | 8.3    | 12.3          | 12.8   |
| DR-OS (Proposed) | 29.3       | 9.5    | 19.6          | -      |

identities in the training set and 750 identities in the testing set. In the original study on this proposed dataset, mAP is the evaluation criteria used to compare the algorithm performances.

2) CUHK03
CUHK03 [37] dataset contains 8765 images of 1467 labeled persons. In this paper, we use a new protocol, in which the training set and test set have 767 and 700 identities, respectively. We select the detected bounding boxes instead of labeled bounding box results. It is a more difficult evaluation protocol for CUHK 03.

3) DukeMTMC-ReID
DukeMTMC-ReID [72] dataset is composed of 36,411 images of 1812 persons captured from eight cameras. There are 702 identities in the training set and 1110 identities in the testing set. The evaluation criteria is mAP, same with the Market-1501 dataset.

4) MSMT17
MSMT17 [73] is the largest image-based person ReID dataset introduced in 2018. It contains 124,069 labeled images of 4101 person IDs captured from 12 different outdoor or indoor cameras. The evaluation protocol/criteria is also same as the Market-1501 dataset, and uses mAP.

B. EVALUATION METRIC
With the same image perturbation ($\epsilon = 4$), we compare performances of all the attack methods while attacking the victim ReID approaches. The lower number indicates more drop in ReID accuracy, and thus, better attack performance. Mean average precision (mAP) is used as the evaluation metric. The effects of using different $\epsilon$ values are discussed in Section IV-D.

C. RESULTS AND DISCUSSION
In the first set of experiments, we used ResNet-152 as the source model of the attacks. The results are summarized in Table 1, wherein the first three rows show the mAP values for the baseline victim models, namely DG-Net, AlignedReID and PLR-OSNet, on different benchmark datasets. The mAP values for DG-Net are 86.0%, 61.1%, 74.8% and 52.3%, and the mAP values for AlignedReID are 82.3%, 70.7%, 82.8% and 43.7% for Market1501, CUHK03, DukeMTMC-ReID and MSMT17 datasets, respectively. For PLR-OSNet, the mAP values are 88.9%, 77.2% and 81.2% for Market1501, CUHK03 and DukeMTMC-ReID datasets, respectively. These three models are regarded as the state-of-the-art ReID approaches based on their performance.

The fourth to sixth rows in Table 1 show the value of the mAP after the victim models are attacked with TI-FGSM, which is a state-of-the-art attack method. For instance, for the Market-1501 dataset, the mAP value of DG-Net drops by 34.8 from 86.0 to 51.2, the mAP value of AlignedReID drops by 23.6 from 82.3 to 58.7, and the mAP value of PLR-OSNet drops by 33 from 88.9 to 55.9. For the CUHK03 dataset, the mAP value of DG-Net drops by 29.2 from 61.1 to 31.9, the mAP value of AlignedReID drops by 41.3 from 70.7 to 29.4, and the mAP value of PLR-OSNet drops by 45.1 from 77.2 to 32.1. The seventh to ninth rows in Table 1 show the value of mAP after the victim models are attacked with TI-DIM, which is another state-of-the-art attack method. Compared to TI-FGSM, this attack is more effective since it causes more drops in the mAP values for all four datasets.

The last three rows of Table 1 show the mAP value after the victim models are attacked with the proposed DR approach. As can be seen, our proposed approach is the most effective attack compared to TI-FGSM and TI-DIM, and causes the most drop in the mAP values for all victim models and for all four datasets. For instance, for the CUHK03 dataset, the mAP value of DG-Net drops by 46.9 from 61.1 to 14.2, the mAP value of AlignedReID drops by 54.2 from 70.7 to 16.5, and the mAP value of PLR-OSNet drops by 58.1 from 77.2 to 19.1. The last three rows of Table 1 show the mAP value after the victim models are attacked with the proposed DR approach. As can be seen, our proposed approach is the most effective attack compared to TI-FGSM and TI-DIM, and causes the most drop in the mAP values for all victim models and for all four datasets. For instance, for the CUHK03 dataset, the mAP value of DG-Net drops by 46.9 from 61.1 to 14.2, the mAP value of AlignedReID drops by 54.2 from 70.7 to 16.5, and the mAP value of PLR-OSNet drops by 58.1 from 77.2 to 19.1.

Fig. 3 shows some example images and query results for the Market-1501 dataset. The first column shows the query images, and columns 2 through 11 show the Rank 1 to
Rank 10 returned images for that query, respectively. The first and third rows are for the original query images, while the second and fourth rows are for the adversarial query images. The perturbations between the query images of first versus second row and third versus fourth row are imperceptible to the human eye, but the person ReID performances have been significantly impacted by the proposed attack. Similar results for CUHK03 and DukeMTMC-ReID datasets are shown in Fig. 4 and Fig. 5, respectively. We report the overall results for the MSMT17 dataset in Table 1, and are not able to provide example images due to the release agreement.

The examples in Figures 3, 4 and 5 show the effectiveness of the proposed DR attack. In these figures, the adversarial examples, although imperceptible to human eye, result in no matches even in Rank 10 returns. As a quantitative measure, we computed the peak signal to noise ratio (PSNR) as well as the structural similarity index measure (SSIM) between the adversarial images (generated by the TI-FGSM, TI-DIM and the proposed DR attack) and the original images, and calculated the average on the Market1501 dataset. The average SSIM value is 0.70, 0.72 and 0.72 for TI-FGSM, TI-DIM and the DR attacks, respectively. The average PSNR is 26, 28 and 27 for TI-FGSM, TI-DIM and the DR attacks, respectively. Since the perturbation budget is kept the same ($\epsilon = 4$) for all the attack methods, their average SSIM and PSNR values are similar. Some example adversarial images generated by these attacks are shown in Fig. 6 for qualitative comparison.

In the second set of experiments, we used two other network models, namely VGG-16 and InceptionV3, as our surrogate source models. The goal here was to use different network models, other than ResNet, to generate adversarial examples and show the generalizability of the proposed DR approach. We have generated AEs by using these different networks as the source models with the proposed DR approach and with TI-DIM. We then used the AEs to attack DG-Net and AlignedReID. In this experiment, we chose to use TI-DIM, since it has better attack performance than TI-FGSM based on Table 1. The results obtained with our proposed DR attack are summarized in Tables 2 and 3 when the victim ReID method is AlignedReID and DG-Net, respectively. As can be seen, when Resnet-152 is used as the surrogate model, it results in the highest drop in the mAP values. This is mostly because most of the ReID approaches use ResNet as their backbone network. However, even when we use VGG-16 or InceptionV3 as the surrogate source model, the proposed DR attack still causes significantly more drop in the mAP values compared to the state-of-the-art attack.
methods when they use ResNet as their surrogate model (please see Tables 1, 2 and 3).

The results obtained with the TI-DIM are summarized in Tables 4 and 5, which show the results obtained with the TI-DIM attack, with using different surrogate models, when the victim ReID method is AlignedReID and DG-Net, respectively. When we compare Table 2 with Table 4, and Table 3 and Table 5, it can be seen that the proposed DR attack still outperforms the TI-DIM as a blackbox attack even when the surrogate model is different from the target model.
**D. EFFECT OF $\epsilon$ ON THE PERFORMANCE**

In literature, it is a common practice to fix the value of $\epsilon$, and then compare the performance degradation for different attack methods. In the experiments above, we set $\epsilon = 4$, since it results in less change in the original image, and better demonstrates the difference between the attack methods. When $\epsilon$ is increased, more budget is given to each attack method to make changes on the original images, and they start to provide similar performance. A better attack should be able to provide more performance degradation with a smaller $\epsilon$ budget. As shown in Table 6 and Fig. 7, our proposed DR attack can reach a given attack effectiveness by using the least budget. For instance, the proposed DR attack drops the mAP value of DG-Net to 20.3 with an $\epsilon$ budget of 8.
whereas TI-DIM needs a budget of 12 to drop the mAP to 21.8.

### V. CONCLUSION

Neural network-based methods have achieved state-of-the-art performance on the person re-identification problem across different camera views. In this paper, we have presented a black-box and effective attack model, which is based on dispersion reduction, and does not rely on task-specific loss functions and label queries. We have used the adversarial examples generated by this approach to attack three different state-of-the-art person ReID models. We have also compared the performance of our attack approach with two other state-of-the-art attack models. The results demonstrate the effectiveness and generalizability of the proposed dispersion reduction attack on three state-of-the-art person ReID models. It also outperforms other state-of-the-art attack models by a large margin, and results in the most drop in the mean average precision values.

### ACKNOWLEDGMENT

The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

### REFERENCES

[1] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2138–2147.

[2] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, “AlignedeReID++: Dynamically matching local information for person re-identification,” Pattern Recognit., vol. 94, pp. 53–61, Oct. 2019.

[3] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, “AlignedeReID: Surpassing human-level performance in person re-identification,” 2017, arXiv:1711.08184. [Online]. Available: http://arxiv.org/abs/1711.08184

[4] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, “Deep-person: Learning discriminative deep features for person re-identification,” 2017, arXiv:1711.10658. [Online]. Available: http://arxiv.org/abs/1711.10658

[5] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, “Part-based deep hashing for large-scale person re-identification,” IEEE Trans. Image Process., vol. 26, no. 10, pp. 4806–4817, Oct. 2017.

[6] 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 Computer Vision—ECCV, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 501–518.

[7] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1116–1124.

[8] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Proc. IEEE Symp. Secur. Privacy (SP), May 2017, pp. 39–57.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in Proc. Int. Conf. Learn. Represent., 2014, pp. 2–11.

[10] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in Proc. ICLR Workshop 2017, pp. 1–13.

[11] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, “Boosting adversarial attacks with momentum,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9185–9193.

[12] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Vailhe, “Adversarial examples for semantic segmentation and object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1378–1387.

[13] S. Dong, K. Eykholt, I. Evtimov, F. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, and T. Kohn, “Physical adversarial examples for object detectors,” in Proc. 12th USENIX Workshop Offensive Technol. (WOT), Baltimore, MD, USA, Aug. 2018, pp. 1–10.

[14] X. Wei, S. Liang, N. Chen, and X. Cao, “Transferable adversarial attacks for image and video object detection,” in Proc. 25th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 954–960.

[15] A. Arnab, O. Miksik, and P. H. S. Torr, “On the robustness of semantic segmentation models to adversarial attacks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 888–897.

[16] S. Bai, Y. Li, Y. Zhou, Q. Li, and H. S. P. Torr, “Metric attack and defense for person re-identification,” 2019, arXiv:1901.10650. [Online]. Available: https://arxiv.org/abs/1901.10650

[17] Z. Zheng, L. Zheng, Z. Hu, and Y. Yang, “Open set adversarial examples,” 2018, arXiv:1809.02681. [Online]. Available: https://arxiv.org/abs/1809.02681

[18] Y. Lu, Y. Jia, J. Wang, B. Li, W. Chai, L. Carin, and S. Velipasalar, “Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 940–949.

[19] X. Wang and R. Zhao, “Person Re-identification: System Design and Evaluation Overview,” London, U.K.: Springer, 2014, pp. 351–370.

[20] W.-S. Zheng, S. Gong, and T. Xiang, “Person re-identification by probabilistic relative distance comparison,” in Proc. CVPR, Jun. 2011, pp. 649–656.

[21] M. Dikmen, E. Akbas, S. Thomas Huang, and N. Ahuja, Pedestrian Recognition with a Learned Metric, Berlin, Germany: Springer, 2011, pp. 501–512.

[22] W. Li, R. Zhao, and X. Wang, Human Reidentification with Transferred Metric Learning, Berlin, Germany: Springer, 2013, pp. 31–44.

[23] A. Mignon and F. Jurie, “PCCA: A new approach for distance learning from sparse pairwise constraints,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2666–2672.

[24] W.-S. Zheng, S. Gong, and T. Xiang, “Reidentification by relative distance comparison,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 3, pp. 653–668, Mar. 2013.

[25] C. Liu, C. C. Loy, S. Gong, and G. Wang, “POP: Person re-identification post-ranking optimisation,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 441–448.

[26] Z. Li, S. Chang, F. Liang, T. S. H. Huang, L. Cao, and J. R. Smith, “Learning locally-adaptive decision functions for person verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 3610–3617.

[27] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2360–2367.

### TABLE 6. \(\epsilon\)AP values obtained with different \(\epsilon\) values on Market-1501 dataset while attacking DG-Net and AlignedReID with TI-DIM and the proposed DR attack.

| Methods/mAP | \(\epsilon = 2\) | \(\epsilon = 4\) | \(\epsilon = 8\) | \(\epsilon = 12\) | \(\epsilon = 16\) |
|-------------|---------------|---------------|---------------|---------------|---------------|
| TI-DIM-DG   | 62.5          | 45.4          | 32.5          | 21.8          | 12.9          |
| TI-DIM-AR   | 64.2          | 44.9          | 29.4          | 21.0          | 18.5          |
| DR-DG       | 47.4          | 28.9          | 20.3          | 13.5          | 12.3          |
| DR-AR       | 48.1          | 20.2          | 21.8          | 15.9          | 11.2          |

### FIGURE 7. Effect of using different perturbation budget (\(\epsilon\)) on the attack performance.
YANTAO LU (Member, IEEE) received the B.S. degree in electrical engineering from Xian Jiaotong University, Shaanxi, China, in 2013, and the M.S. and Ph.D. degrees in electrical engineering from Syracuse University, Syracuse, NY, USA, in 2015 and 2020, respectively. His research interests include activity recognition from wearable cameras and object detection using neural networks.

SENEM VELIPASALAR (Senior Member, IEEE) received the B.S. degree in electrical and electronic engineering from Bogazici University, Istanbul, Turkey, in 1999, the M.S. degree in electrical sciences and computer engineering from Brown University, Providence, RI, USA, in 2001, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 2004 and 2007, respectively. From 2001 to 2005, she was with the Exploratory Computer Vision Group, IBM T. J. Watson Research Center, NY, USA. From 2007 to 2011, she was an Assistant Professor with the Department of Electrical Engineering, University of Nebraska–Lincoln. She is currently a Professor with the Department of Electrical Engineering and Computer Science, Syracuse University. Her research has been on mobile camera applications, wireless embedded smart cameras, multi-camera tracking and surveillance systems, and automatic event detection from videos.

Dr. Velipasalar received a Faculty Early Career Development Award (CAREER) from the National Science Foundation, in 2011. She was a recipient of the 2014 Excellence in Graduate Education Faculty Recognition Award. She is the coauthor of the paper, which received the 2017 IEEE Green Communications and Computing Technical Committee Best Journal Paper Award. She received the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo, in 2006. She was also a recipient of the EPScR First Award, IBM Patent Application Award, and Princeton and Brown University Graduate Fellowships. She is a member of the Editorial Board of the IEEE Transactions on Image Processing and Springer Journal of Signal Processing Systems.