Multi-Expert Adversarial Attack Detection in Person Re-identification Using Context Inconsistency

Xueping Wang[1,2], Shasha Li[3], Min Liu *[1,2], Yaonan Wang[1,2] and Amit K. Roy-Chowdhury[3]

[1]College of Electrical and Information Engineering, Hunan University, China
[2]National Engineering Laboratory for Robot Visual Perception and Control Technology, China
[3]University of California, Riverside

Abstract

*The success of deep neural networks (DNNs) has promoted the widespread applications of person re-identification (ReID). However, ReID systems inherit the vulnerability of DNNs to malicious attacks of visually inconspicuous adversarial perturbations. Detection of adversarial attacks is, therefore, a fundamental requirement for robust ReID systems. In this work, we propose a Multi-Expert Adversarial Attack Detection (***`MEAAD`***) approach to achieve this goal by checking context inconsistency, which is suitable for any DNN-based ReID systems. Specifically, three kinds of context inconsistencies caused by adversarial attacks are employed to learn a detector for distinguishing the perturbed examples, i.e., a) the embedding distances between a perturbed query person image and its top-K retrievals are generally larger than those between a benign query image and its top-K retrievals, b) the embedding distances among the top-K retrievals of a perturbed query image are larger than those of a benign query image, c) the top-K retrievals of a benign query image obtained with multiple expert ReID models tend to be consistent, which is not preserved when attacks are present. Extensive experiments on the Market1501 and DukeMTMC-ReID datasets show that, as the first adversarial attack detection approach for ReID,* **`MEAAD`** *effectively detects various adversarial attacks and achieves high ROC-AUC (over 97.5%).*

1. Introduction

The success of DNNs has benefited a wide range of computer vision tasks, such as image classification [13, 16], object detection [12, 36], face recognition [34, 20], video classification [21, 47], and person ReID [50, 23, 14, 50, 33]. Person ReID is a critical task aiming to retrieve pedestrians across multiple non-overlapping cameras. By learning the discriminative feature embedding and adaptive distance metric models, DNNs-based ReID models, in recent years, have extensive applications in video surveillance or criminal identification for public safety. However, recent research has found that these models inherit the vulnerability of DNNs to adversarial examples [42, 45, 2, 8] which are slightly perturbed input images but lead DNNs to make wrong predictions [11, 39]. Detection of adversarial examples is, therefore, a fundamental requirement for robust ReID systems because the insecurity of ReID systems may cause severe losses. However, ReID is defined as a ranking problem rather than a classification problem and thus existing defense methods for image classification [6, 18, 30, 26, 28, 41] do not fit the person ReID problem.

In addition, the top-K retrievals output by person ReID systems, compared to the prediction label in classification task, contain richer information and can be potentially employed to detect adversarial attacks. To illustrate, let's consider the top-10 retrievals obtained with five different state-of-the-art person ReID systems (LSRO [53], AlignedReID [48], PCB [38], HACNN [23] and Mudeep [33]) of a query sample before and after an adversarial attack in Fig. 1(a). *When considering the retrievals returned by a single ReID system, e.g. LSRO, they are visually more similar to the query image before attack than that after attack, and they are visually more similar to each other before attack. When considering the retrievals returned by different expert models, they are consistent before attack but certainly not after attack.* We did an empirical study as shown in Fig. 1(b) and found that the embedding distance is able to reflect visual similarity; specifically, the retrievals of the benign query tend to gather together in the embedding space while the retrievals of the perturbed query tend to spread over.

Inspired by these observations, we propose Multi-Expert

*Corresponding author: liu_min@hnu.edu.cn.
X. Wang was a visiting student at UCR in 2019-20.

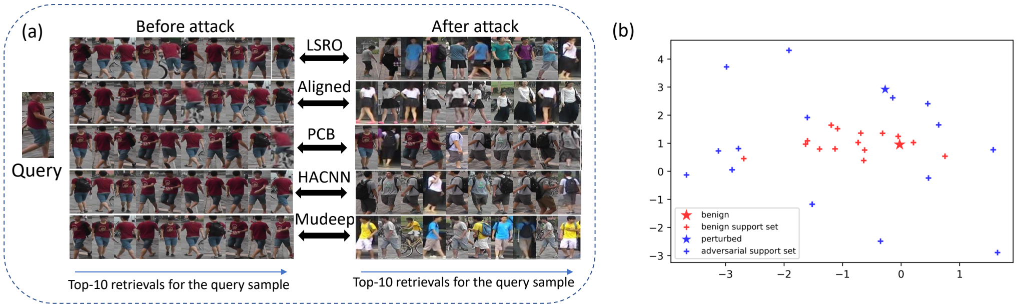

Figure 1. (a) shows the top-10 retrievals for a query sample before and after an adversarial attack. Five state-of-the-art person ReID models, i.e., LSRO [53], AlignedReID [48], PCB [38], HACNN [23] and Mudeep [33] are used as the expert models. Deep Mis-Ranking attack [42] is used to generate the adversarial perturbations and AlignedReID is the attack target model. The top-10 retrievals of a benign query (before attack) are consistent across multiple expert models, while they are messy for a perturbed query sample (after attack). (b) presents (using PCA) the embedding space of an expert model (AlignedReID). The original query sample is marked with red star and its retrievals are marked with red plus marker. The perturbed query sample is marked with blue star and its retrievals are marked with blue plus marker. We observe that the retrievals of the benign query sample gather tightly around the benign query sample in the embedding space. In comparison, the retrievals of the perturbed query sample spread over the space. We have quantitative and more detailed results in Fig. 2.

Adversarial Attack Detection which detects adversarial attacks for person ReID systems by detecting context inconsistency. To the best of our knowledge, this is the first strategy to detect adversarial attacks against person ReID systems. To make use of the heterogeneity of different ReID models, we use multiple ReID networks with different architectures as expert models in **MEAAD**. We define support set as the top-K retrievals output by a single expert model. Context used in **MEAAD** accounts for three types of relations: 1) the relations between the query and its support samples returned by a single expert (*Query-Support Affinity*); 2) the relations among the support samples returned by a single expert (*Support-Support Affinity*); 3) the relations between the support samples returned by one expert and those returned by another (*Cross-Expert Affinity*). We then train a detector with the context features of both benign and perturbed query samples and use it to detect adversarial attacks during testing. The contributions are as below,

- To the best of our knowledge, this is the first adversarial attack detection strategy for the defense of ReID systems.
- We empirically study the side effect brought by adversarial attacks, i.e., context inconsistency in the retrieval results. We then propose **MEAAD** which aims to detect adversarial attacks by checking context inconsistency of a query sample to be detected.
- Extensive experiments on the Market1501 and DukeMTMC-ReID datasets show that, **MEAAD** effectively detects various adversarial attacks and achieves high ROC-AUC (over 97.5% in all cases).

2. Related works

2.1. Person re-identification

Person re-identification is a cross-camera instance retrieval problem, which aims at searching persons across multiple cameras. With the advancement of deep learning, person ReID has achieved inspiring performance on the widely used benchmarks. In this field, deep learning-based feature representation methods which focus on developing the feature construction strategies have been widely used [48, 23, 33]. In [14, 50, 44], the authors proposed to employ deep metric learning models to address the person ReID task, which aim at designing the training objectives with different loss functions or sampling strategies. In recent years, using GAN to transfer the source domain images to target-domain style is a popular approach for ReID [7, 55, 54]. With the generated images, this enables using supervised ReID models in the unlabeled target domain. Another direction is to learn ReID models from limited labeled data [35, 43]. These methods have achieved impressive performance. In our framework, we adopt the state-of-the-art person ReID models with different network architectures as our expert models and extract context features from the outputs of these experts for adversarial attack detection.

2.2. Adversarial attacks

Adversarial attacks have achieved remarkable success in fooling DNN-based systems, e.g., image classification [11, 19, 17, 27, 29, 32, 21] and object detection [5, 49], etc. A few adversarial attacks have been proposed for attacking ReID models. Wang et al. [42] proposed a learning-to-misrank formulation to perturb the ranking of the ReID system outputs. Ding et al. [8] proposed an effective method to train universal adversarial perturbations (UAPs) against person ReID models from the global list-wise perspective. [3, 2] proposed adversarial metric attack to perturb the ReID systems. Instead of the previous digital perturbations, Wang et al. [45] implemented robust physical-world attacks against deep ReID for generating adversarial patterns on clothes, which learns the variations of image pairs

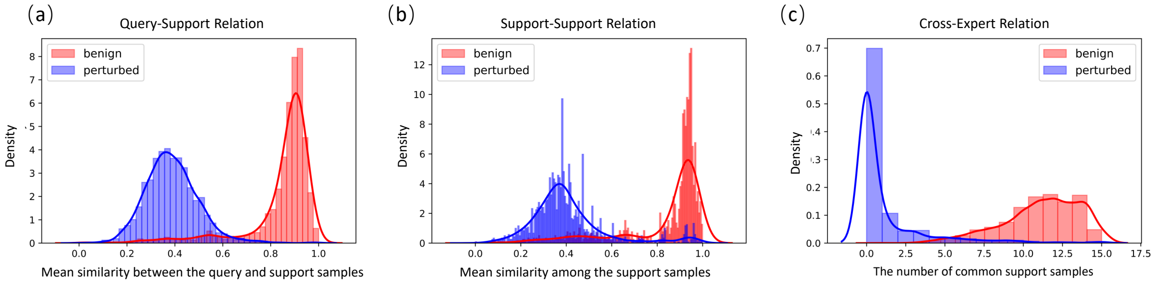

Figure 2. Empirical study results: (a) plots the query-support relation distribution for both benign and perturbed query samples. The query-support relation is defined to be the average of cosine similarity between the embedding feature of the query sample and the embedding features of the support samples. (b) plots the support-support relation distribution. The support-support relation is defined to be the average of cosine similarity among the embedding features of the support samples for each query image in the same support set. (c) plots the cross-expert relation distribution. We use the same number of support samples across all support sets to describe cross-expert relation.

across cameras to pull closer the image features from the same camera, while pushing features from different cameras farther. Our defense strategy is dependent on the contextual information, and therefore does not rely heavily on the mechanism to generate the perturbations.

2.3. Adversarial defense

To address the vulnerability of DNNs to adversarial attacks, some adversarial training-based defense approaches have been proposed [11, 18, 28, 41]. However, adversarial training-based defense methods degrade the natural performance of the target models and they can be evaded by the optimization-based attack [4], either wholly or partially. Recent works have focused on detection-based defense methods which aim at distinguishing adversarial examples from benign ones [24, 22, 30, 6, 26]. Li et al. [22] proposed a context inconsistency-based adversarial perturbation detection method for object detection task. They constructed a fully connected graph for each detected object by accounting for four types of region relationships, and trained a classifier for each category. Yin et al. recently presented how to use language descriptions to detect adversarial attacks through context inconsistencies [46]. [24] applied steganalysis techniques to model the dependence between adjacent pixels; adversarial perturbations in most cases alter the dependence between pixels and thus can be detected by their method. However, most of these methods are developed for the classification task and they are not suitable for defending the person ReID models (ranking systems), because in a classification task, the training and testing set share the same categories, while in ReID, there is no category overlap between them.

3. Methodology

3.1. Threat model

The attacker's goal is to cause the target ReID system to retrieve person images of wrong identities. In this paper, we assume the attacker is able to launch attacks against the target ReID system by perturbing the query images, but not poisoning the gallery images. The same threat model has been used in [42, 25, 8, 3, 52, 40], and is reasonable as galleries are very large (usually secured) and attacking a large number of gallery images is very time-consuming [52]. We assume the attack target ReID model is white-box to the attacker because even through in practical attacks it could be black-box, works [31, 9] have shown that attackers could estimate the function of a black-box model by making queries and reasoning on the query results. This assumption favors the attacker, and thus makes the defense systems more robust. The other expert models used for consistency check never output retrieval results to the users, thus we assume the attacker is not aware of their existence. Note that we also explore two adaptive attacks where the attacker is aware of all experts and our defense scheme in Section 4.5.

3.2. Empirical study on context inconsistency

While the adversarial examples fool the ReID system to retrieve wrong images from the gallery set, they have a side effect, i.e., causing "messy" retrieval results as shown in Fig. 1. We define three relations to describe “messy". Before introducing the three relations, we define the top-K retrievals returned by the ReID system as the support set and each retrieval in it as the support sample. We refer to the support set of a benign query sample as benign support set and refer to that of a perturbed query sample as adversarial support set for simplicity. The empirical study is done with 2,000 benign query samples and 2,000 perturbed query samples obtained with the Deep Mis-Ranking attack [42] and top-15 retrievals of each query are used as the support set. Two person ReID systems (LSRO [53] and AlignedReID [48]) are used in the empirical study, and we call each system an expert model.

Query-Support Relation. The retrieved images in the benign support set tend to be similar to the query image. We validate whether the embedding feature similarity between the query image and the support set reflects the same trend and the results are shown in Fig. 2(a). We define Query-Support Relation as the average of cosine similarity between the embedding feature of the query sample and the embedding features of the support samples. We observe that compared to benign query samples, perturbed queries generally have lower similarity to its support samples in the embedding space. This implies that we could distinguish benign and attack by the Query-Support Relation.

Support-Support Relation. The retrieved images in the benign support set tend to be similar to each other. We validate whether the embedding feature similarity among the support samples reflects the same trend and the results are shown in Fig. 2(b). We define Support-Support Relation as the average of cosine similarity among the embedding features of the support samples for each query image. We observe that compared to benign support samples, adversarial support samples have lower similarity to each other in the embedding space. This implies that we could distinguish benign and attack by the Support-Support Relation.

Cross-Expert Relation. We observe from Fig. 1 that the benign support sets returned by different expert models overlap with each other a lot. We use the number of the common support samples returned by all expert models to describe Cross-Expert Relation. Fig. 2(c) shows that for a benign query sample, different expert models tend to return the same retrievals, which implies that the Cross-Expert Relation could be used to distinguish benign and attack.

3.3. Multi-expert adversarial attack detection

Inspired by the above empirical studies, we propose multi-expert adversarial attack detection illustrated in Fig. 3 to distinguish the perturbed samples from benign ones by checking context inconsistency of the query samples.

3.3.1 Context feature

To formulate the problem, we use I to denote the query image and use $F_i(\cdot), i = 1, 2, ..., N$ to denote the functions of the N expert models. We denote the support set (top-K retrievals) retrieved by the i^{th} expert model as $\mathbf{S}_i = \{S_{i,j} | j = 1, ..K\}$. Each model learns a mapping from the image space to its latent feature embedding space. Therefore the embedding feature of I with the i^{th} expert model can be represented as $F_i(I)$. The heterogeneity of these expert models provides multi-view information for each query sample. The context feature is composed of three parts: query-support affinity, support-support affinity and cross-expert affinity, and we describe each in details.

Query-Support Affinity. We extract the query-support affinity feature for each expert model in the same way. Therefore, we use $\mathbf{S} = \{S_j | j = 1, ..K\}$ instead of $\mathbf{S}_i = \{S_{i,j} | j = 1, ..K\}$ for simplicity afterwards. Similarly, we use $F(\cdot)$ instead of $F_i(\cdot)$. If we use A_{q-s} to denote the query-support affinity feature for the current expert model, then A_{q-s} is a vector of K dimension and the j^{th} element is defined as the cosine similarity between the embedding feature of I and the embedding feature of the support sample S_j as shown in Eq. 1.

$$A_{q-s}[j] = CosSimilarity(F(I), F(S_j)) \tag{1}$$

We calculate A_{q-s} for all the expert models and stack them together, which is the final query-support affinity feature with dimension $N * K$.

Support-Support Affinity. We extract support-support affinity feature for each expert model in the same way. If we use A_{s-s} to denote the support-support affinity feature for the current expert model, then A_{s-s} is a matrix of $K * K$ dimension and the element on (i, j) is defined as the cosine similarity between the embedding feature of S_i and the embedding feature of S_j as shown in Eq. 2.

$$A_{s-s}[i, j] = CosSimilarity(F(S_i), F(S_j)) \tag{2}$$

Note that A_{s-s} is a symmetric matrix and the diagonal elements are always 1 (suppose the embedding feature is normalized). Therefore, instead of keeping all the elements, we keep the $K * (K - 1)/2$ elements in the upper-right (or lower-left) matrix and A_{s-s} becomes a vector of $K' = K * (K - 1)/2$ dimension. We calculate A_{s-s} for all the expert models and stack them together, which is the final support-support affinity feature with dimension $N * K'$.

Cross-Expert Affinity. To calculate the cross-expert affinity, we need the support sets of all the N expert models. At each time, we choose an expert model as the base model and other $N - 1$ expert models are called member models. We choose the base model in turns, and thus if we use A_{c-e} to denote the cross-expert affinity feature then A_{c-e} is a matrix and the i^{th} row of the matrix is the feature calculated when the i^{th} expert model is chosen as the base model. The element on (i, j) is defined as the frequency that the j^{th} support sample output by the base model (denoted as $S_{i,j}$) appears in the support sets output by the member expert models as shown in Eq. 3.

$$A_{c-e}[i, j] = \frac{\sum_{l \in \{1,..,N\} - \{i\}} \mathbb{1}(S_{i,j} \in \mathbf{S}_l)}{N - 1} \tag{3}$$

$\mathbb{1}(\cdot)$ is an indicator function which gives a value of 1 when the argument is true. Therefore, the cross-expert affinity feature is a matrix with dimension $N * K$.

In summary, there are three parts of the context feature, i.e., query-support affinity feature $A_{q-s} \in \mathbb{R}^{N*K}$, support-support affinity feature $A_{s-s} \in \mathbb{R}^{N*K'}$ and cross-expert affinity feature $A_{c-e} \in \mathbb{R}^{N*K}$. We flatten all the matrices into vectors and concatenate them together as the final context feature for one query sample. We use x to denote the context feature, thus $x \in \mathbb{R}^d, d = N*K + N*K' + N*K$.

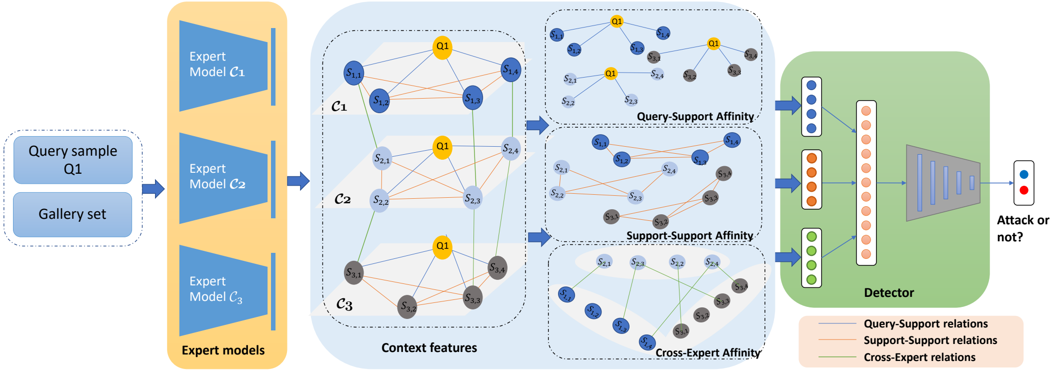

Figure 3. The pipeline of the proposed Multi-Expert Adversarial Attack Detection system. We employ multiple state-of-the-art ReID networks with different architectures as the expert models. The top-K retrievals of a query sample are defined as a support set and each retrieval is a support sample. Based on query-support affinity, support-support affinity and cross-expert affinity, we define context feature for each query image and its support sets. A detector with context features as input is then learnt to distinguish attack from benign. It may be noted that in this figure we use three expert models and top-4 retrievals as an example to illustrate our framework.

3.3.2 Adversarial attack detector

With the context feature defined, the next step is to learn an adversarial attack detector. As shown in Fig. 3, the detector is basically a binary classifier which takes context features as inputs and outputs whether the query image is perturbed or not. Since the input is of relative low dimension, we use Multi-Layer Perceptron (MLP) classifier as the detector.

To train the detector, we extract context features with benign query samples and assign classification label $y = 0$ to them, and also extract context features with perturbed query samples and assign classification label $y = 1$ to those context features. Therefore, the training set is $\{(x_i, y_i)|i = 1, 2, ..M\}$ and M is the size of the training set. During testing, given a query sample, we first extract the context feature from the query sample and its support sets retrieved by multiple expert models, and then input the context feature into the detector to decide if the query sample is perturbed.

4. Experiments

4.1. Implementation details

Datasets. We validate the adversarial attack detection performance of **MEAAD** on both Market1501 [51] and DukeMTMC-ReID [37] datasets. Market1501 is captured by six cameras. The training dataset contains 12,936 cropped images of 751 identities, while the testing set contains 19,732 cropped images of 750 identities. DukeMTMC-ReID dataset is captured by eight cameras. There are 16,522 bounding boxes of 702 identities for training and another 702 identities of 17,661 images for testing. We follow the standard training and testing splits for these two datasets in our experiments.

Attack implementations. We evaluate our defense strategy against two state-of-the-art adversarial attack approaches (Deep Mis-Ranking [42] and advPattern [45]) that are specifically designed against ReID systems, four attacks (FGSM [11], CW [32], Deepfool [29] and PGD [27]) that are designed for general DNNs, and two adaptive attacks (adaptive CW and multi-model targeted attack) against **MEAAD**. The two attacks specific to ReID are described:

- *Deep Mis-Ranking* [42] is a digital attack that perturbs the ranking of the ReID system's outputs by proposing a learning-to-misrank formulation.

- *advPattern* [45] is a physical-world attack against ReID systems which adds printable adversarial patterns on clothes. Note that to do evaluation on a large scale, we do not print the generated patterns and add them physically. Instead, we add the patterns digitally onto the person images. This favors attackers since they can control how their physical perturbations are captured.

Defense implementations. Top-15 retrievals are used as the support set for each query. To create an expert system with high heterogeneity, person ReID models with different network architectures are used during evaluation. Due to their superior performance on the Market1501 dataset, PCB [38], AlignedReID (AR) [48], HACNN [23], LSRO [53] and Mudeep (MD)[33] are the five candidates to serve as expert models, and similarly, AlignedReID [48], LSRO [53] HHL [54], CamStyle (CS) [55] and SPGAN [7] are the five candidates to serve as expert models for evaluation on the DukeMTMC-ReID dataset. For all the eight models, we use the author-released models with trained parameters. The ReID performance of the methods on both datasets is presented in the Supplementary Material. The detector in **MEAAD** is an MLP classifier with 2 hidden layers that contain 512 and 256 nodes respectively. ReLU function is used as the activation function. In addition to collecting benign context features for the detector training, we collect adversarial context features by perturbing the query samples in the training set with Deep Mis-Ranking [42], adv-

Table 1. Comparison with the state-of-the-art adversarial attack detection methods on the Market1501 and DukeMTMC-ReID datasets against Deep Mis-Ranking and advPattern attacks.

Defense Methods	Market1501						DukeMTMC-ReID					
	Deep Mis-Ranking			advPattern			Deep Mis-Ranking			advPattern		
	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
DkNN [30]	87.9	93.8	86.9	98.9	99.1	99.0	90.2	97.8	91.1	98.5	99.0	98.6
LID [26]	90.6	96.2	91.1	99.3	99.7	99.4	87.4	95.2	88.1	99.4	99.6	**99.7**
SRM [24]	94.3	98.2	94.1	99.2	99.8	99.5	91.2	97.2	92.1	99.6	99.7	**99.7**
MEAAD (Voting)	91.7	91.7	91.0	98.6	98.6	98.6	88.7	88.7	88.1	96.8	96.8	96.7
MEAAD (Detector)	**98.5**	**99.8**	**98.6**	**99.6**	**100**	**99.6**	**95.3**	**99.2**	**95.5**	**99.7**	**99.7**	**99.7**

Pattern [45] and other attacking methods. Note that there is no query/gallery separation in the training set; we randomly choose one person image as the query sample and use all the others as the gallery samples. SGD optimizer with momentum 0.9 is used for training. The learning rate is 1e-4. The detector training is finished after 5,000 iterations and batch size is set to 1,024. Our experiments are conducted on a NVIDIA GTX 2080TI GPU using Pytorch.

Evaluation metric. To tell if a query image input into the ReID system is perturbed, we first get the retrieval results from the chosen expert models and extract the context feature; the context feature is then input into the detector to be classified to attack or benign. Therefore, one metric used to evaluate the detection performance is the classification accuracy or called detection accuracy (*Acc*). We keep the number of benign and perturbed samples balanced in our testing set. In addition to using probability threshold 0.5 to decide perturbed or not, we can flexibly adjust the threshold and get the Receiver Operating Characteristic (ROC) curve, for which, we report area under the ROC curve, i.e., *ROC-AUC*, as another detection performance metric. Similarly, we also use *F1 score* which is the harmonic mean of the dection precision and recall as one metric.

4.2. Attack detection performance

In this section, we evaluate the proposed adversarial detection method against both Deep Mis-Ranking attack and advPattern attack on the Market1501 and DukeMTMC-ReID datasets. Three state-of-the-art detection methods are extended to deal with ReID systems; they are used as the baseline methods which are described below. More details can be found in the Supplementary Material.

• *Local Intrinsic Dimensionality (LID)* [26] characterizes the intrinsic dimensionality of adversarial regions which is a property of datasets [1]. Adversarial perturbation affects the LID characteristics of adversarial regions, and thus they are used to detect adversarial examples.

• *Deep k-Nearest Neighbors (DkNN)* [30] combines the k-NN algorithm with feature representations of samples: an input is compared to its neighbors in the metric space. Labels of these neighbors afford confidence estimates for inputs outside the model's training manifold, e.g. adversarial examples, which are used to detect adversarial attacks.

Table 2. Adversarial attack detection with different number of expert models on the Market1501 dataset. * indicates the attack target model known to the attackers.

Expert models	Acc	AUC	F1
AR*	95.2	99.1	95.5
AR*+PCB	97.8	99.7	97.9
AR*+PCB+LSRO	98.4	99.8	98.4
AR*+PCB+LSRO+HACNN	98.5	99.8	98.6

• *Spatial Rich Model (SRM)* [10, 24] detects adversarial examples from steganalysis point of view and proposes enhanced steganalysis features which are sensitive to small perturbations. Therefore, it can be used to distinguish the perturbed samples from the benign ones.

Moreover, instead of extracting the complete context feature and training a data-driven detector, we simply use the number of common support samples across all expert models as the feature and threshold over it to decide attack or benign. This is used as the forth baseline (**MEAAD** (Voting)).

For the evaluation on Market1501, AlignedReID is chosen as the attack target model which is known to the attacker, and AlignedReID, LSRO, PCB and HACNN are used as the experts. For the evaluation on DukeMTMC-ReID, LSRO is the attack target model, and LSRO, SPGAN, AlignedReID and HHL are selected as the experts.

The detection performance is shown in Tab. 1. We observe that advPattern attack is easier to be detected compared to Deep Mis-Ranking attack; the baseline methods and ours all have an F1 score over 96.7% but ours (**MEAAD** (Detector)) performs better consistently. To detect the Deep Mis-Ranking attack, our method clearly outperforms the baseline methods on both datasets. For example, the F1 score on the Market1501 dataset of the DkNN method is 86.9%; that of the LID method is 91.1%; that of the SRM method is 94.1%; that of **MEAAD** (Voting) is 91.0% with $threshold = 5$; and **MEAAD** (Detector) achieves 98.6%, which is 4.5% better than the best baseline.

4.3. Ablation study

In this section, we do ablation studies to understand a) how the number of expert models affects the detection performance; b) whether the detection performance is sensitive to the different choices of expert models; c) how the

Table 3. Adversarial attack detection with/without using the attack target model as one of the expert models on Market1501.

Expert models	Acc	AUC	F1
AR*	95.2	99.1	95.5
AR*+PCB+LSRO+HACNN	**98.5**	**99.8**	**98.6**
PCB	88.2	95.1	88.7
PCB+LSRO	93.7	98.5	93.9
PCB+LSRO+HACNN	94.2	98.5	94.2

size of the support set affects the detection performance; d) the importance of the three types of relations (query-support relation, support-support relation and cross-expert relation) in attack detection. Deep Mis-Ranking is used to attack AlignedReID model for the evaluation on Market1501.

Number of expert models. In this section, we study whether more expert models improve the detection performance of **MEAAD** on the Market1501 dataset. As shown in Tab. 2, with more expert models, the detection performance is better. We suppose this is because more expert models bring more context information and thus the extracted context features are more discriminative between benign and perturbed samples. Note that when the number of expert models is one, there is no cross-expert affinity feature, only query-support affinity feature and support-support affinity feature are used, in which case, however, we still get very good performance: F1 score on Market1501 is 95.5%. Combining four expert models (AlignedReID, LSRO, PCB and HACNN), we achieve the best detection performance: 98.5% detection accuracy on the Market1501 dataset. The results on the DukeMTMC-ReID dataset can be found in the Supplementary Material.

Choices of the expert models. In this section, we explore the detection performance of **MEAAD** with different expert model choices. Two choice strategies are compared, that is, including the attack target model as one of the expert models, and not using the attack target model as one of the expert models. The results are shown in Tab. 3. We observe that on the Market1501 dataset, the F1 score of using the attack target model (AR model) as the only expert model is 95.5%, which is higher than 94.2% when using other three expert models (PCB+LSRO+HACNN). This indicates that it is beneficial to include the attack target model as one of the expert models. The same conclusion can be drawn on the DukeMTMC-ReID dataset and the details can be found in the Supplementary Material. A potential reason is that the attack is tuned to the target model and this creates a larger variance with the other experts in the retrievals.

Size of the support set. Note that all the previous evaluations are done with the size of support set equal to 15, basically top-15 retrievals are used to extract the context. We explore how the size of the support set affects the attack detection performance. Specially, we evaluate the attack detection performance when $K = 1, 5, 10, 15, 20, 30$ as shown in Tab. 4. Note that $K = 1$ means there is no support-support affinity feature, and only query-support affinity feature and cross-expert affinity feature are used. We observe that in general using a larger support set gives better attack detection rate, when $K = 15$, we achieve 98.5% detection accuracy - 6.2% improvement comparing to the result of $K = 1$. It can be seen that thereafter ($K > 15$), with the increase of the support samples, the performance is almost stable.

Table 4. Adversarial attack detection with different sizes of the support set on the Market1501 dataset.

Top-K	Acc	AUC	F1
$K = 1$	92.3	99.2	92.9
$K = 5$	94.4	99.7	94.7
$K = 10$	97.5	99.8	97.6
$K = 15$	98.5	99.8	98.6
$K = 20$	98.5	99.8	98.5
$K = 30$	98.5	99.8	98.6

Table 5. Ablation test: adversarial attack detection with different context features on the Market1501 dataset.

A_{c-e}	A_{q-s}	A_{s-s}	Acc	AUC	F1
✓			94.9	99.4	95.1
	✓		93.6	99.6	94.0
		✓	90.9	97.8	91.6
✓	✓		97.2	99.5	97.1
✓		✓	95.6	99.6	95.7
	✓	✓	97.0	99.5	97.0
✓	✓	✓	**98.5**	**99.8**	**98.6**

Importance of different relations. The context feature used in **MEAAD** is composed of three parts, that is, query-support affinity feature (A_{q-s}), support-support affinity feature (A_{s-s}) and cross-expert affinity feature (A_{c-e}). To quantify the contribution of each relation, we conduct an ablation study on the Market1501 dataset. As shown in Tab. 5, when using only one relation for detecting attacks, we have already achieved over 90.5% detection accuracy in all cases, especially 94.9% for cross-expert affinity. All the three features are complementary, that is, combining two of them improves the detection performance consistently, such as the detection accuracy is increased from 90.9% (support-support affinity) to 97.2% (cross-expert affinity and query-support affinity). When using all of them, our method achieves the best attack detection performance 98.5% for F1 score, 99.8% AUC score and 98.6% detection accuracy.

4.4. Defense against new attack methods

Recall that we need both the benign and perturbed samples to train our detector in **MEAAD**. Therefore, the detector is able to detect the attacks which have appeared in its training set. However, as new attack methods are proposed, it is not feasible to exhaustively cover all the attack methods in the training set. In this section, we evaluate how our defense method transfers to new unknown attacks. We extend

Table 6. Adversarial attack detection against unseen new attacks on the Market1501 dataset.

Settings	Attacks	Testing											
		CW			Deepfool			FGSM			PGD		
		Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
Training	CW	96.1	98.7	96.2	96.1	98.6	96.2	96.2	98.6	96.3	96.9	99.0	97.0
	Deepfool	95.8	98.6	95.9	96.0	98.6	96.1	95.8	98.6	95.9	96.6	98.9	96.7
	FGSM	96.1	98.7	96.2	96.0	98.7	96.1	96.1	98.8	96.2	96.8	99.0	96.9
	PGD	94.0	98.4	93.8	93.7	98.3	93.6	94.1	98.4	93.9	97.4	98.8	97.4

four state-of-the-art adversarial attack methods against general DNNs to attack ReID systems; basically, we regard the last layer of the ReID model as the identity prediction layer, and on top of that the adversarial query examples can be generated for training and testing. The four attack methods: FGSM [11], CW [32], Deepfool [29] and PGD [27] are implemented with Torchattacks [15].

We generate the training set with one attack method, and test the trained detector on all the attack methods. The experiment is done on the Market1501 dataset. AlignedReID is used as the attack target model. AlignedReID, LSRO and PCB are used as the three expert models. The results are shown in Tab. 6. When the detector is tested on the same attack method as that used in its training set, the detection performance is very good, for example, F1 score equals to 96.2% when detecting CW attack. When the detector is asked to detect unseen attacks, the performance remains or drops just by a little bit, for example, F1 score drops from 96.2% to 95.9% when detecting unknown Deepfool attack compared to known CW attack. This indicates that although different attack methods generate the perturbations in different ways, the perturbations tend to always cause messed up retrieval results, and thus our defense strategy of detecting context inconsistency transfers well across different attack methods and is effective to unknown new attacks.

4.5. Adaptive Attacks against **MEAAD**

To further evaluate **MEAAD**'s robustness towards adaptive attacks, we extend an existing adaptive attack method, i.e., adaptive CW attack, and adopt a new adaptive attack method, i.e., multi-model targeted attack to evade **MEAAD**. More details can be found in the Supplementary Material.

Adaptive CW attack. We extend the adaptive CW algorithm [4] by introducing a new loss item (associated with three kinds of context defined in **MEAAD**) to the loss function and the new loss term is as below:

$$l_*(\mathbf{MEAAD}(x_{adv})) = -\sum(A_{qs} + A_{ss} + A_{ce}) \qquad (4)$$

The new item is defined to reduce the retrieval inconsistency of the adversarial examples. The rationale for the minimization of the added term in Eq. 4 is that adversarial examples have lower context affinity than benign examples. Experiments show that such adaptive attack decreases **MEAAD**'s detection accuracy by 1.3% when only the attack target ReID model is white-box to the attacker, and 3.5% when all the ReID models are white-box to the attacker. In either way, the ROC-AUC score of **MEAAD** is still high, over 95%.

Multi-model targeted attack. As shown in Fig. 1, the retrieval results of non-targeted attack are messy and not consistent across different models, and thus such attacks are detected by **MEAAD**. If we assume all expert ReID models are white-box to the attacker, then the attacker could do targeted attack against all the models simultaneously. In other words, this adaptive attack generates adversarial examples that fool all the ReID models used in **MEAAD** (both the target model and the expert models) to retrieve the same wrong identity and thus context would be more consistent. We extend the adversarial metric attack in [2] to a multi-model targeted attack for attacking **MEAAD**. However, aligned with previous works [56], we find that targeted attack against multiple ReID models is hard and only 211 (6.2%) such adversarial examples from all the 3,368 testing samples. **MEAAD**'s detection accuracy on the 211 adversarial examples is 88.6%.

5. Conclusions

In this paper, we propose a Multi-Expert Adversarial Attack Detection framework that detects adversarial attacks against ReID systems by checking context inconsistency, a side effect of the adversarial attacks. Empirical studies show that query-support affinity, support-support affinity and cross-expert affinity are able to distinguish the perturbed ones. Therefore, we propose to leverage the three relations to form the context feature for each query sample. A detector is then trained on the context features of both benign and perturbed samples and are then used to detect adversarial attacks. Experiments on the Market1501 and DukeMTMC-ReID datasets show that **MEAAD** effectively detects various adversarial attacks, that is, Deep Mis-Ranking, advPattern, Deepfool, CW, FGSM and PGD, and **MEAAD** can effectively detect unknown new attacks.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62073126 and Grant 61771189, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2020JJ2008, and in part by the Office of Naval Research (ONR) at UCR under Grant N00014-19-1-2264.

References

[1] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38, 2015. 6

[2] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip H.S. Torr. Adversarial metric attack and defense for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2119–2126, 2021. 1, 2, 8

[3] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 794–795, 2020. 2, 3

[4] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 3, 8

[5] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 2

[6] Gilad Cohen, Guillermo Sapiro, and Raja Giryes. Detecting adversarial samples using influence functions and nearest neighbors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14453–14462, 2020. 1, 3

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018. 2, 5

[8] Wenjie Ding, Xing Wei, Xiaopeng Hong, Rongrong Ji, and Yihong Gong. Universal adversarial perturbations against person re-identification. *arXiv preprint arXiv:1910.14184*, 2019. 1, 2, 3

[9] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 3

[10] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 6

[11] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *the International Conference on Learning Representations*, 2015. 1, 2, 3, 5, 8

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2

[15] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020. 8

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 2

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 3

[19] Shasha Li, Mustafa Arslan, Amir Khojastepour, Srikanth V Krishnamurthy, and Sampath Rangarajan. Deeptrack: Grouping rfid tags based on spatio-temporal proximity in retail spaces. In *the IEEE International Conference on Computer Communications*, pages 1271–1280, 2020. 2

[20] Shasha Li, Karim Khalil, Rameswar Panda, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and Ananthram Swami. Measurement-driven security analysis of imperceptible impersonation attacks. *arXiv preprint arXiv:2008.11772*, 2020. 1

[21] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*, 2018. 1, 2

[22] Shasha Li, Shitong Zhu, Sudipta Paul, Amit Roy-Chowdhury, Chengyu Song, Srikanth Krishnamurthy, Ananthram Swami, and Kevin S Chan. Connecting the dots: Detecting adversarial perturbations using context inconsistency. *arXiv preprint arXiv:2007.09763*, 2020. 3

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 1, 2, 5

[24] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. Detection based defense against adversarial examples from the steganalysis point of view. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4825–4834, 2019. 3, 6

[25] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Who's afraid of adversarial queries? the impact of image modifications on content-based image retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 306–314, 2019. 3

[26] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018. 1, 3, 6

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 5, 8

[28] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*, 2015. 1, 3

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 2, 5, 8

[30] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018. 1, 3, 6

[31] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*, pages 506–519, 2017. 3

[32] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *the IEEE Symposium on Security and Privacy*, pages 582–597, 2016. 2, 5, 8

[33] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5399–5408, 2017. 1, 2, 5

[34] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3781–3790, 2017. 1

[35] Dripta S Raychaudhuri and Amit K Roy-Chowdhury. Exploiting temporal coherence for self-supervised one-shot video re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 258–274, 2020. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1

[37] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision*, pages 17–35. Springer, 2016. 5

[38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, pages 480–496, 2018. 1, 2, 5

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[40] Giorgos Tolias, Filip Radenovic, and Ondrej Chum. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5037–5046, 2019. 3

[41] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 1, 3

[42] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep misranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2020. 1, 2, 3, 5

[43] Xueping Wang, Min Liu, Dripta S. Raychaudhuri, Sujoy Paul, Yaonan Wang, and Amit K. Roy-Chowdhury. Learning person re-identification models from videos with weak supervision. *IEEE Transactions on Image Processing*, 30:3017–3028, 2021. 2

[44] Xueping Wang, Rameswar Panda, Min Liu, Yaonan Wang, and Amit K. Roy-Chowdhury. Exploiting global camera network constraints for unsupervised video person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2

[45] Zhibo Wang, Siyan Zheng, Mengkai Song, Qian Wang, Alireza Rahimpour, and Hairong Qi. advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8341–8350, 2019. 1, 2, 5, 6

[46] Mingjun Yin, Shasha Li, Zikui Cai, Chengyu Song, M. Salman Asif, Amit K. Roy-Chowdhury, and Srikanth V. Krishnamurthy. Exploiting multi-object relationships for detecting adversarial attacks in complex scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[47] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-excited sampler: Video adversarial attack with sparked prior. *arXiv preprint arXiv:2003.07637*, 2020. 1

[48] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 1, 2, 3, 5

[49] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1989–2004, 2019. 2

[50] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019. 1, 2

[51] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification:

A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 5

[52] Zhedong Zheng, Liang Zheng, Zhilan Hu, and Yi Yang. Open set adversarial examples. *arXiv preprint arXiv:1809.02681*, 2018. 3

[53] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017. 1, 2, 3, 5

[54] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision*, pages 172–188, 2018. 2, 5

[55] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018. 2, 5

[56] Shitong Zhu, Shasha Li, Zhongjie Wang, Xun Chen, Zhiyun Qian, Srikanth V Krishnamurthy, Kevin S Chan, and Ananthram Swami. You do (not) belong here: detecting dpi evasion attacks with context learning. In *Proceedings of the International Conference on Emerging Networking EXperiments and Technologies*, pages 183–197, 2020. 8