Supplementary Material for "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

In the supplementary material, 1) we provide the recognition performance of person ReID models that we used as the experts in **MEAAD** on the Market1501 and DukeMTMC-ReID datasets. 2) we give more details for the choices of the expert models. 3) we show the detection performance of the proposed adversarial detection method with different number of expert models on the DukeMTMC-ReID dataset. 4) we report the detection performance of **MEAAD** on the DukeMTMC-ReID dataset with/without using the attack target model as one of the expert models. 5) we explore the detection performance of **MEAAD** on the adaptive CW attack which is aware of the defense scheme and has white-box access to the expert models used in **MEAAD**. 6) we also propose another adaptive attack method, named multi-model targeted attack, and evaluate **MEAAD**'s robustness towards it. 7) we present the implementation details of the three state-of-the-art adversarial attack detection baseline methods: Local Intrinsic Dimensionality (LID) [10], Deep k-Nearest Neighbors (DkNN) [11] and Spatial Rich Model (SRM) [9] which we used in the main paper.

1. ReID performance of the expert models

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 [13], AlignedReID (AR) [14], HACNN [8], LSRO [15] and Mudeep (MD) [12] are the five candidates to serve as expert models for evaluation on the Market1501 dataset, and similarly, AlignedReID (AR) [14], LSRO [15], HHL [16], CamStyle (CS) [17] and SPGAN [5] 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 these methods on both datasets is presented in Table 1. The common cumulative matching characteristic (CMC) and the mean average precision (mAP) metrics are utilized to demonstrate the performance of each method.

Table 1. Recognition performance with different expert ReID models on the Market1501 and DukeMTMC-ReID dataset.

Methods	Rank-1	Rank-10	mAP
Performance on Market1501			
PCB [13]	88.6	97.3	70.7
AlignedReID [14]	91.8	98.1	79.1
HACNN [8]	90.6	97.4	75.3
LSRO [15]	89.9	97.4	77.2
Mudeep [12]	73.0	93.1	49.9
Performance on DukeMTMC-ReID			
AlignedReID [14]	72.0	89.5	55.2
LSRO [15]	72.0	89.5	55.2
HHL [16]	71.4	87.7	51.8
CamStyle (CS) [17]	76.5	90.0	58.1
SPGAN [5]	73.6	88.9	54.6

2. Choices of expert models

From Tab. 1, it can be seen that the recognition performance of different expert models is various, e.g. there is 18.8% rank-1 accuracy difference between AlignedReID [14] model and Mudeep [12]. MEAAD is based on the context features which are extracted from the embedding features of the query samples and the corresponding support samples, so poor ReID models may affect the adversarial attack detection performance. Therefore, in this section, we provide more analysis on how the ReID performance of the expert models affect the adversarial attack detection performance. We do evaluation on the Market1501 dataset. Deep Mis-Ranking attack method is used to generate the perturbations against the target attack model (AlignedReID model). The results can be found in Tab. 2. It can be seen

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

Table 2. Adversarial attack detection performance with different 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*+Mudeep	93.2	99.4	93.7
AR*+LSRO	97.5	99.7	97.6
AR*+PCB	97.8	99.7	97.9
AR*+PCB+LSRO	98.4	99.8	98.4
AR*+HACNN+LSRO	98.2	99.8	98.2
AR*+PCB+HACNN	98.2	99.7	98.3
AR*+LSRO+Mudeep	96.4	99.7	96.5
AR*+HACNN+Mudeep	97.5	99.7	97.6
AR*+PCB+LSRO+HACNN	98.5	99.8	98.6
AR*+PCB+LSRO+Mudeep	97.9	99.8	97.9
AR*+LSRO+HACNN+Mudeep	98.2	99.8	98.2
AR*+PCB+HACNN+Mudeep	98.3	99.7	98.3
AR*+PCB+LSRO+HACNN+Mudeep	98.5	99.8	98.6

that when we use two ReID models as the experts, the adversarial attack detection performance will be affected by the poor model, i.e., 4.6% detection accuracy drops when using Mudeep as one of the experts. However, we find with the increase of the expert models, the side effect caused by the poor experts will be reduced gradually, such as 2.0% decreasement when using three experts and 0.6% decreasement for four experts. We conclude that 1) it is better to use ReID models with higher ReID performance for attack detection; 2) MEAAD is robust against poor expert models when using multiple ReID models as the experts.

3. Detection with different number of experts

In this section, we report the detection performance of the proposed defense method on DukeMTMC-ReID with different number of expert models. We get the same conclusion as the results on the Market1501 dataset (Tab. 2 in the main paper): the performance is better when we use more expert models because more expert models bring more context information and thus the extracted context features are more discriminative between benign and perturbed samples. As shown in Tab. 3, when using only the attack target model (LSRO), we still get very good performance: F1 score is 93.2%. Combining five expert models (LSRO+AR+SPGAN+HHL+CS), we achieve the best detection performance: 95.3% detection accuracy on the DukeMTMC-ReID dataset.

4. Detection with/without the target model

In this section, we report the detection performance of **MEAAD** with/without using attack target model as one of the experts on DukeMTMC-ReID. The results are shown in Tab. 4. We observe that the F1 score of using the attack target model (LSRO) as the only expert model is 93.2%,

Table 3. Adversarial attack detection performance with different number of expert models on the DukeMTMC-ReID dataset. * indicates the attack target model known to the attackers.

Expert models	Acc	AUC	F1
LSRO*	92.6	98.4	93.2
LSRO*+AR	93.2	99.1	93.5
LSRO*+AR+SPGAN	94.3	99.2	94.5
LSRO*+AR+SPGAN+HHL	95.3	99.2	95.5
LSRO*+AR+SPGAN+HHL+CS	95.3	99.2	95.5

Table 4. Adversarial attack detection performance with/without using the attack target model as one of the expert models on the DukeMTMC-ReID dataset. * indicates the attack target model.

Expert models	Acc	AUC	F1
LSRO*	92.6	98.4	93.2
LSRO*+AR+SPGAN	94.3	99.2	94.5
AR	88.6	98.8	89.7
AR+SPGAN	89.7	98.2	90.4
AR+SPGAN+CS	93.0	98.8	93.4
AR+SPGAN+CS+HHL	93.0	98.7	93.3

Table 5. Adversarial attack detection performance on the adaptive CW attack on the Market1501 dataset.

Attack method	Acc	AUC	F1
Non-adaptive CW attack	96.1	98.7	96.2
Adaptive CW with single model	94.5	97.6	94.9
Adaptive CW with all models	92.6	95.7	92.8

which is very close to 93.3% when using other four expert models (AR+SPGAN+CS+HHL). This indicates that it is beneficial to include the attack target model as one of the expert models.

5. Adaptive CW attack against MEAAD

To further evaluate the proposed adversarial detection method against adaptive attacks where the attacker is assumed to be aware of the consistency check and even have white-box access to all the expert models used in **MEAAD**, we extend the adaptive attack strategy, adaptive CW attack proposed in [2], to attack **MEAAD**, and evaluate **MEAAD**'s detection performance on the extended adaptive CW attack. Specifically, instead of minimizing density to evade kernel density-based adversarial detectors, here we modify the last term of the adaptive CW loss related to context consistency check used in **MEAAD** as below:

$$\text{minimize}||x - x_{adv}||_2^2 + \alpha \cdot (l_{cw}(x_{adv}) + l_*(\textbf{MEAAD}(x_{adv}))) \quad (1)$$

where x is a benign query sample to be attacked and x_{adv} is its corresponding perturbed version. α is a constant balancing between the amount of perturbation and the adversarial strength. $l_{cw}(x_{adv})$ is the original adversarial loss term used in [3, 2] to make the adversarial example classified to the target class. $\textbf{MEAAD}(x_{adv})$ is the sum over the three kinds of context affinity and $l_*(\texttt{MEAAD}(x_{adv}))$ is introduced to maximize the affinity defined in **MEAAD**. The rationale is that adversarial examples have lower context affinity than benign examples and thus we need to increase the affinity to evade **MEAAD**. We define it as below:

$$l_*(\texttt{MEAAD}(x_{adv})) = -\sum(A_{qs} + A_{ss} + A_{ce}) \quad (2)$$

For testing, LSRO is used as the attack target ReID model, and LSRO and PCB are the expert models. The results are in Tab. 5. We test the proposed method under two adaptive-attack scenarios. In the first scenario, we assume the attacker is aware of our context consistency-based defense scheme and only the attack target model is white-box to the attacker, i.e. $l_*(\texttt{MEAAD}(x_{adv})) = -\sum(A_{qs} + A_{ss})$ is defined with the query-support affinity and support-support affinity. We observe that the proposed the adaptive CW attack only decreases the detection accuracy of **MEAAD** by 1.6%. In the second scenario, we assume the attacker knows our defense strategy and has white-box access to all the ReID models used in **MEAAD**, i.e., $l_*(\texttt{MEAAD}(x_{adv}))$ is defined with all the three affinities. As shown Tab. 5, the detection accuracy drops by 3.5% compared to that against the original non-adaptive CW attack. Therefore, we may conclude that our **MEAAD** defense algorithm is robust to the adaptive CW attack.

6. Multi-model targeted attack against MEAAD

As shown in Fig. 1 in the main paper, the retrieval results of the non-targeted attack are messy and not consistent across different expert models, and thus such attacks are detected by **MEAAD**. If we assume all expert models are white-box to the attacker, the attacker could do targeted attack against all expert 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 is more consistent. We name this attack method as multi-model targeted attack. We extend the adversarial metric attack in [1] to a multi-model targeted attack as below. Given expert models $F_i(\cdot), i = 1, 2, ..., N$, N is the number of expert models used in **MEAAD**, we solve the following optimization problem to generate adversarial query examples.

$$\underset{x}{\text{minimize}} \frac{1}{N} \sum_i ||F_i(x) - F_i(g_t)||_2^2 \quad (3)$$

where x is a query image to be attacked and g_t is the gallery images with the pre-determined target person identity. Following the settings in [1], we use the Euclidean distance as the distance metric for attack. For testing, LSRO and PCB are used as the expert models. We use MI-FGSM [6] as the attacking method to generate the adversarial query examples on the Market1501 dataset.

Table 6. Adversarial attack detection performance on the multi-model targeted attack on the Market1501 dataset.

	$C = 5$	$C = 6$	$C = 7$	$C = 8$
# attacked samples	211	154	114	76

A successful multi-model targeted attack is defined as at least C samples of the targeted person identity are retrieved in top-15 retrievals by each expert model in **MEAAD**. However, aligned with previous works [18], we find that targeted attack against multiple models is hard. As shown in Tab. 6, only 211 (6.2%) such adversarial examples are found from all the 3,368 tests when $C = 5$. We evaluate **MEAAD**'s detection performance against such 211 adversarial examples and the detection accuracy is 88.6%. In summary, firstly, such adaptive adversarial examples do not always exist; second, **MEAAD** is still able to detect such examples with decent performance.

7. Implementations of the LID, DkNN and SRM

In this paper, we compare the proposed method with three state-of-the-art adversarial attack detection methods, Local Intrinsic Dimensionality (LID)[10], Deep k-Nearest Neighbors (DkNN) [11] and Spatial Rich Model (SRM) [7, 9] . In this section, we demonstrate the implementation details of these three methods for adversarial examples detection in person ReID.

The **LID** associated with each query example (either benign or perturbed) is estimated from its support set (top-K retrievals). For any new unknown test query example, a support set consisting its top-15 retrieved samples is used to estimate LID. The outputs of the feature embedding layer are used to calculate an LID estimate. They are then used as feature values to train a classifier (logistic regression (LR) is used, like [10]). Test examples are then classified by the LID-based classifier to either the positive (perturbed) or negative (benign) class by means of its LID-based feature values.

The **DkNN** algorithm is proposed to better estimate the prediction, confidence, and credibility for a given test sample, in which a test query input is compared to its top-15 retrievals (support set) according to the distance that separates them in the representations. Following the settings in [4], we convert the original DkNN algorithm [11] to an adversarial attack detection method. This is done by collecting the empirical p-values calculated in the DkNN strategy and formulating a reactive adversarial detector by training a LR model on these features. Note that since the DkNN method requires a calibration set, we randomly select 10% of the query examples for calibrating it and present all results by features from the embedding space alone.

SRM [7, 9] can effectively detect modifications caused by adversarial attack via modeling the dependence between adjacent pixels in natural images. Following the same settings in [9], 45 pixel predictors from the pixel's immediate neighborhood are used to obtain a residual which is an estimate of the image noise component. Then, we extract 34,671 steganalysis features and utilize them to train a classifier to distinguish the perturbed samples from the benign ones.

References

[1] 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. 3

[2] 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. 2

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 2

[4] 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. 3

[5] 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. 1

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 3

[7] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 3, 4

[8] 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

[9] 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. 1, 3, 4

[10] 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

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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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. 1

[17] 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. 1

[18] 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. 3