In: 2023 IEEE 9th International Conference on Big Data Computing Service and Applications (BigDataService),
17-20 July 2023, Athens Greece
DOI:10.1109/BigDataService58306.2023.00032

Soft Body Pose-Invariant Evasion Attacks against Deep Learning Human Detection

1st Chen Li
School of Aerospace, Transport and Manufacturing
Cranfield University
Bedford, United Kingdom
c.li.21@cranfield.ac.uk

2nd Weisi Guo
School of Aerospace, Transport and Manufacturing
Cranfield University
Bedford, United Kingdom
Weisi.guo@cranfield.ac.uk

***Abstract*—Evasion attacks on deep neural networks (DNN) use artificial data to manipulate common neural network layers (e.g., convolution operations) to create higher losses. This allows targets to evade detection and/or classification across a wide range of DNNs without the need for a backdoor attack or to know specific type used. Most of the existing work in evasion attacks have focused on planar images (e.g., photo, satellite imaging) in relatively consistent lighting conditions. More recent work have recognised the need to create patterns that are more easily printed or work in diverse lighting environments.**

Here, we build printable evasion patterns for fabric clothing to highlight the risks to autonomous systems and provide data for future adversarial training. These novel evasion attacks are for soft body human stakeholders, where patterns are designed to take into account body rotation, fabric stretch, printable, and lighting variations. We show that these are effective and robust to different human poses. This poses a significant threat to safety of autonomous vehicles and adversarial training should consider this new area.

***Index Terms*—Deep Learning, Human Detection, Evasion Attack, Safety, Security**

I. Introduction

Increased autonomy in transportation means that the interaction between autonomous piloting and humans is a critical safety area. Examples include but are not limited to detecting pedestrians [1], monitoring driver presence for human-in-loop autonomy [2], and ensuring safe social distancing [3]. Most autonomous platforms use a range of sensors (e.g., vision, IR, lidar) to fuse data and create a holistic understanding the surrounding environment. However, low-end platforms and some commercial operators (e.g., Tesla) have plans to use a camera only system [4]. Here, the back-end analytics is often performed by convolution neural networks (CNNs), where deep layers of convolution operations extract meaningful deep features in image sequences [1].

Evasion attacks are a class of attacks that use small data manipulations to create false results or null results (nothing detected) for general classes of DNNs, such as CNNs [5]. The data manipulations can manifest themselves in digital changes (e.g., a digital display) or physically (e.g., graffiti, spray camouflage, or printed patterns). These attacks do not need to know the specific CNN architecture being used, nor any extraction or back-door access. As shown in Fig. 1, a vision-based CNN human detection model uses camera data as input to determine the existence of objects according to the object confidence and detection threshold. However, physical evasion patterns could easily interfere with the model decision by changing the colour information locally in camera data. In practice, these evasion patterns can arise due to: (1) unintended use (urban evasion of CCTV for privacy enhancement) [6], (2) intended malicious use on or by others, and (3) accidental patterns emerging as part of art or fashion. In all these cases, humans maybe missed and serious considerations to how we can include evasion attacks in CNN training is needed.

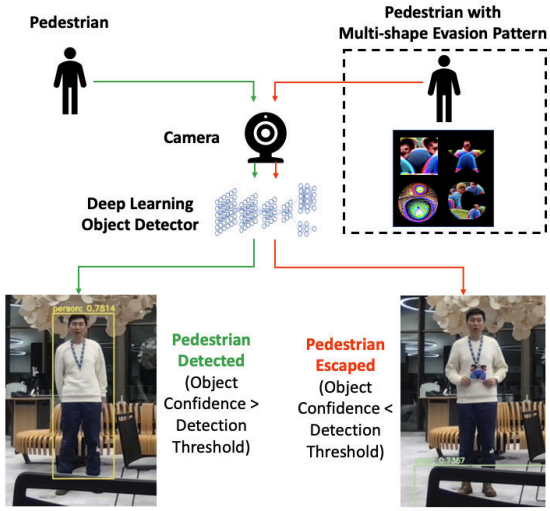

Fig. 1. Deep Learning using CNN for human Detection. (Green) human without any evasion patterns is easily detected by deep learning object detector. (Red) human holding a physical evasion pattern can escape detection by deep learning object detector.

A. Review of Evasion Pattern

Adversarial evasion attack methods mislead the inference process of DNN models by adding visible or imperceptible perturbation noise to images. This modifies the latent features in images and leads to interfered model decision errors [7]–[9]. Adversarial patch methods [10], attack CNNs by applying

We acknowledge funding from EPSRC TAS-S: Trustworthy Autonomous Systems: Security (EP/V026763/1) and Department for Transport S-TRIG project.

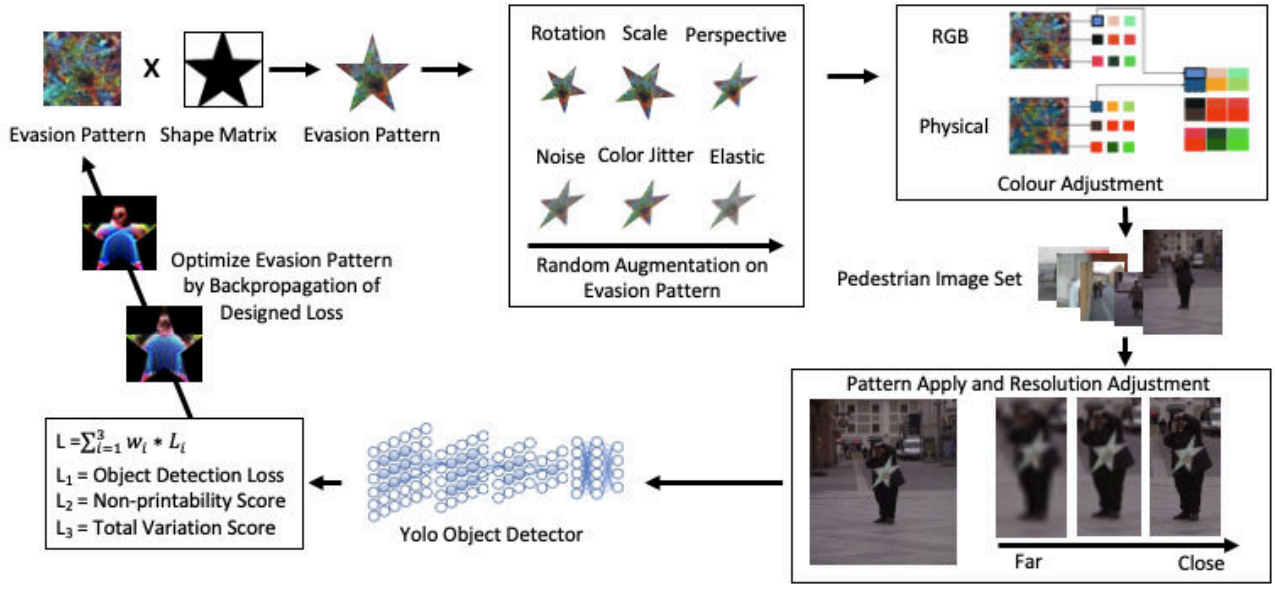

Fig. 2. The Process of Training Evasion Patterns for DL human Detection

a visible evasion pattern digitally or physically on an image or object, with different purposes (e.g., reduce the accuracy of classification, detection or both).

There are many studies in recent years related to evasion attacks. Authors in [5] propose a basic patch method to fool a YOLO v2 CNN detector, and provide a demo that a person escapes from detection by holding a physical patch. Authors in [11] design a practical neural stealth T-shirt which considers both the deformation of evasion patterns caused by certain human movements and the color difference between RGB color, printed color and recaptured color by cameras. Authors in [12] propose a method to discover the most efficient local evasion texture from patterns and use Toroidal Cropping to produce clothes that are full of textures. There are several novel variants of patch training such as using GAN [13] to generate multiple patches at the same time; generating butterfly-shape patches with GAN [14]; data independent method to generate patches [15]; train face-mask-shape patches [16]; train deformable patches [17]; train transferable patches with one white-box and several black-box models [18].

However, these methods ignore several factors in the physical world that influence the effectiveness of evasion patterns. Firstly, as studied in [19], the perspective of patches caused by object rotations influences the effectiveness of evasion patterns. Secondly, different shooting distances cause changes in image resolutions (e.g., clarity of optical features). High-resolution patches can easily reach a higher performance during virtual training compared with lower ones, but the meticulous features could easily destroy due to resolution changes. Thirdly, the deformation of clothes can lead to deformation in evasion patterns. This influences the distribution of pattern features and reduces the effectiveness of evasion patterns.

B. Contributions and Novelty

In this research, we wish to highlight the dangers of evasion patterns that can lead to missing humans by autonomous platforms. In particular, we study and discuss the feasibility of current evasion pattern training methods for real-world usage on humans that have *soft bodies*, *different postures*, and clothing that will *distort the evasion patterns via elastic deformation*. This had the novel advantage of being functionally effective in 3D on soft bodies with different postures, but also having different printing shapes that can make shape filtering detection ineffective.

To achieve the above, we propose a novel generation method to train diverse *shaped evasion patterns with high real-world usability*. We design experiments to verify the performance of different evasion patterns and from the results we show that our evasion patterns are with higher invariance and robustness to pattern deformation and perspectives compared to other techniques.

II. Method

We aim to generate evasion patterns in arbitrary shapes, and stick the printed physical patterns with a suitable size on clothes to attack the DL detector effectively. We first introduce the method to form the evasion pattern with a certain shape, and then introduce the augmentation tricks to enhance the usability of evasion patterns in physical use. The overall pattern training process is demonstrated in Fig. 2.

A. Shape the Evasion Pattern

As shown in the top-left Fig. 2. Firstly, the initialised format of the evasion pattern is a 3D matrix with random noise (size: $3\times n\times n$. 3 channels for R, G, B colours respectively. Each channel with a certain length and width n, express the resolution of evasion pattern). Then, a shape matrix will

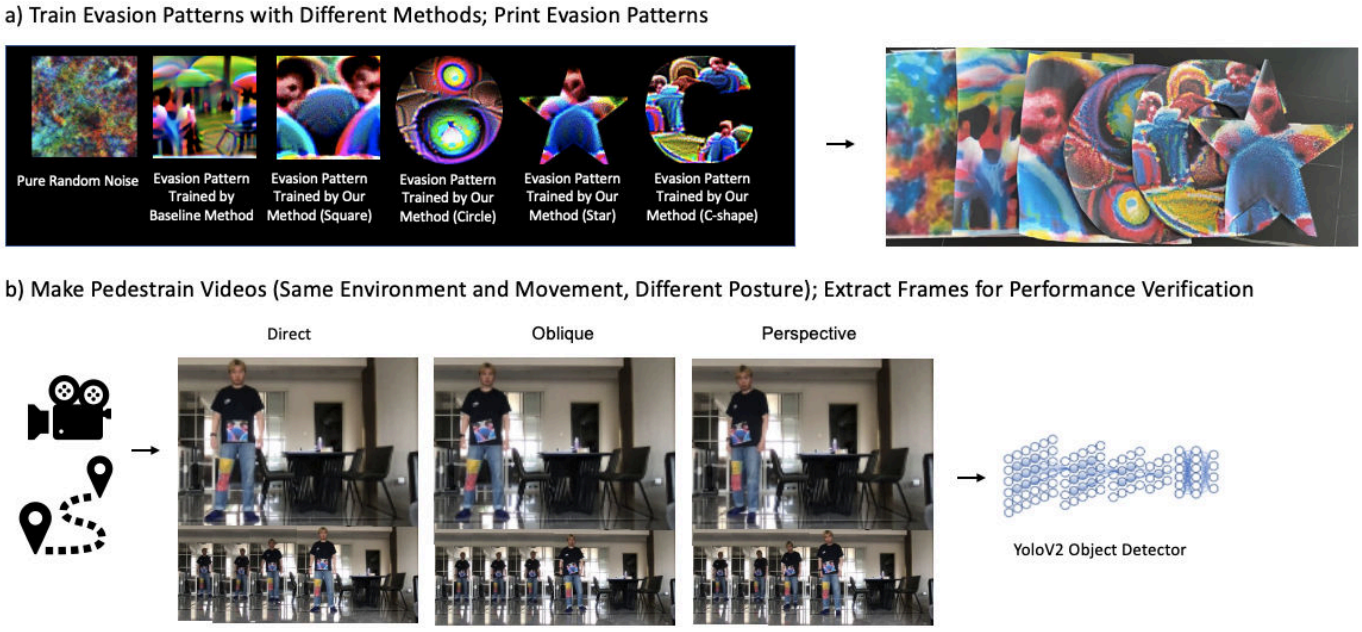

Fig. 3. The Training and Verification of Evasion Patterns: (a) training robust patterns with different shapes, (b) verification with different human poses.

be established with the same size as the evasion pattern. The shape matrix is conducted by only 0 and 1, which use black and white colours to express the expected shape of the evasion pattern. Then, multiply the square-shape evasion pattern with the shape matrix to eliminate the out-of-shape colour information, and only leave the evasion pattern within the expected pattern shape.

B. The Augmentation Tricks of Evasion Patterns

The baseline method used in this paper is proposed in [5]. This method optimises evasion patterns to maximise the detection loss of a pre-trained Yolo V2 object detector. Prior to training, we applied several augmentations to the evasion patterns to simulate the effects that a physical evasion pattern might encounter during use. This is to grant the evasion pattern more robustness through invariance to different environments, human postures, and camera perspectives.

As shown in Fig. 2, the evasion pattern in a certain shape will be applied with a set of random augmentation. The augmentation includes rotation, scaling, perspective, noise, color jitter, and elastic deformation respectively. Rotation, scaling and perspective enhancements are applied to increase the robustness of the evasion patterns to different pattern sizes and human postures (e.g., different human body postures will lead to changes in the relative position of the evasion patterns and the human body). Adding random noise and color jitter is to avoid the overfitting problem of the evasion pattern. At the same time, color jitter can also partially simulate the influence of different ambient lighting on the color performance of the pattern in the camera. This enhances the robustness of evasion patterns in different lighting environments. Normal elastic deformation augmentation only changes the color patterns within an image, without changing the image shape. Here, we put the evasion pattern on a larger black background and then do elastic augmentation. This can simulate the texture of evasion patterns on everyday clothing, and the deformation of evasion patterns caused by human posture and movements.

During the previous research and our experiments, there exist differences among RGB color, printed color, and camera-captured physical color. This will reduce the effectiveness of evasion patterns in real-world attacks. The authors in [20] proposed a method to use a regression model on the mapping relationship between RGB color and physical color. This method could simulate the color performance of a physical evasion pattern in real-world attacks. As we only focus on real-world attacks, we apply this method to evasion patterns to enhance their real-world attack effectiveness.

C. The Optimization of Evasion Patterns by Designed Loss

As shown in Fig. 2, the processed evasion pattern will be virtually stuck onto the humans in images with a suitable size guided by labels (indicate the coordinates of humans in images). Then we apply random resolution adjustment on the whole post-evasion-pattern image to simulate the dropping in resolutions caused by camera distance. Then, the processed images will be fed into DL human detection models, and the evasion pattern will be directly optimized by the backpropagation of the designed loss. In the design of loss, we consider the following aspects:

- Loss 1 (L_1): Object detection loss. This loss is the maximum confidence score for all detection boxes classified as humans in a picture labelled by DL object detectors (e.g. Yolo V2). In each detection box, if the confidence score is below a pre-defined confidence threshold in the DL detector, the detector does not report objects detected in that box. As the aim of evasion patterns is to let the

TABLE I
REAL-WORLD PERFORMANCE EVALUATION OF EVASION PATTERNS ON YOLO V2 GENERIC OBJECT DETECTORS (YOLO V2 THRESHOLD: 0.7, ACCURACY - HUMAN DETECTION SUCCESS RATE, CONFIDENCE SCORE - AVERAGE MAXIMUM CONFIDENCE SCORE OF ALL DETECTION BOXES CLASSIFIED AS HUMAN

	Posture: Direct		**Posture: Oblique**		**Posture: Perspective**	
Evasion Pattern	Accuracy	Confidence Score	Accuracy	Confidence Score	Accuracy	Confidence Score
None	100%	0.895	100%	0.895	99.7%	0.887
Pure Random Noise	100%	0.834	99.6%	0.828	99.6%	0.827
Baseline Method	75.8%	0.796	99.4%	0.808	99.6%	0.817
Our Method (Square)	**42.6%**	**0.757**	**67.4%**	**0.768**	**73.2%**	**0.775**
Our Method (Circle)	75.5%	0.798	73.7%	0.794	81.6%	0.801
Our Method (Star)	78.2%	0.801	75.1%	0.798	99.5%	0.816
Our Method (C-shape)	66.7%	0.774	70.2%	0.786	86.8%	0.805

human escape from CNN detection, the L_1 is expected to be minimized during the evasion pattern optimization.

- Loss 2 (L_2): Non-printability score. L_2 indicates to what extent the RGB evasion pattern could be printed correctly by common printers [21], as printers can only print a limited number of colours compared with RGB colours. Suppose each pixel in an evasion pattern P is represented by p, and the printable colour set of a common printer is C:

$$L_2 = \sum_{p \in P} \min_{c \in C} abs(p - c) \tag{1}$$

  To make sure the color could be printed accurately, the L_2 is expected to be reduced during the optimization of the evasion pattern.
- Loss 3 (L_3): Total variation score. L_3 indicates how smooth the colour transitions among close pixels [22]. This is to avoid noisy patterns, and also help to eliminate meticulous features. This is calculated by:

$$L_3 = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} \tag{2}$$

  The L_3 is expected to be reduced during the optimization of the evasion pattern.

The overall designed loss function could be seen as follows:

$$L = \sum_{i=1}^{3} w_i \times L_i \tag{3}$$

Each loss is granted with a weight to balance different evasion pattern capabilities. During the optimization of evasion patterns, the direction of gradient descent will lead to a reduction of L. In our settings, the weights are set to 1, 0.1, and 0.5 respectively.

III. VERIFICATION EXPERIMENTS SETTING

A set of verification experiments is designed to compare the effectiveness of evasion patterns trained with our method and the baseline method in real-world attacks. As shown in Fig. 3 a), we prepare a set of physical evasion patterns that are random noise pattern, baseline method pattern, and four different shape evasion patterns (square, circle, star and C-shape) trained with our method (DL model: Yolo V2 object detector; Evasion pattern resolution: 150*150; Training settings: start learning rate 1e-2, batch size 12, 300 epochs) for comparison. Then, as shown in Figure 3 b), we aim to test the performance of evasion patterns on three posture scenarios:

- Direct (The pattern is placed horizontally on the clothes, and the human directly facing the camera while moving)
- Oblique (the pattern is placed obliquely on the clothes, and the human directly facing the camera while moving)
- Perspective (the pattern is placed obliquely on the clothes, and the human facing the camera with perspective angles while moving)

to control the variables for the comparison experiment, we fixed the environment, character movement route and speed during the camera data collection. Video data are recorded with 2K resolution and 60 FPS. Videos are further processed into image sets with a sampling rate of every 3 frames. Then, we verify the evasion pattern performance with a pre-trained Yolo V2 object detector with the threshold set to 0.7.

IV. VERIFICATION RESULTS

The verification result could be seen in Table. I. Although pure random noise can steadily reduce the confidence score in all three postures (avg confidence drops from 0.89 to 0.83), it still exceeds the threshold which leads to non-effects on the accuracy of the Yolo detector. The baseline method is proven to be effective in direct posture experiments (accuracy drops from 100% to 75.8%), but fails in other postures due to the training process not considering the deformation and perspective of evasion patterns.

The square shape evasion pattern shows the best performance in all three postures. The best performance of square shape pattern is in direct posture experiments (accuracy drops from 100% to 43%), while in oblique and perspective experiments still fairly effective (average accuracy drops from 99.7% to 70%). The other three evasion patterns in circle, star and C-shape show similar performance in direct and oblique experiments (average accuracy drops from 100% to 70%), but not as good as the square shape evasion pattern in perspective experiments.

V. Conclusions and Future Work

Evasion attacks on deep neural networks (DNN) use artificial data to manipulate common CNN architectures to create higher losses. This allows targets to evade detection and/or classification and can have safety and security impact across transportation value chain. Most of the existing work in evasion attacks have focused on planar images in relatively consistent lighting conditions.

Here, we build printable evasion patterns for fabric clothing. These are for soft body humans, designed to be rotation, stretch, and lighting invariant. We show that these are effective and robust to different human poses. The preliminary research here show that diverse patterns can poses a significant threat to safety of autonomous vehicles by eroding detection accuracy from 100% to 42-77% depending on posture.

From an adversarial perspective, we will in the future develop more postures and crowd dynamics for training, with widespread applications in both counter-surveillance and providing adversarial training samples for surveillance. We believe defence strategies [23] should be considered in future research to improve transportation engineering's safety in this new area. Potential future work direction include using Topological Data Analysis [24] to identify evasion features, adversarial training, and detection strategies [25].

References

[1] G. Li, Y. Yang, and X. Qu, "Deep learning approaches on pedestrian detection in hazy weather," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8889–8899, 2020.

[2] Z. Hu, Y. Xing, W. Gu, D. Cao, and C. Lv, "Driver anomaly quantification for intelligent vehicles: A contrastive learning approach with representation clustering," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 37–47, 2023.

[3] B. Fraser, B. Copp, G. Singh, O. Keyvan, T. Bian, V. Sonntag, Y. Xing, W. Guo, and A. Tsourdos, "Reducing viral transmission through ai-based crowd monitoring and social distancing analysis," in *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2022, pp. 1–6.

[4] P. Cai, Y. Sun, Y. Chen, and M. Liu, "Vision-based trajectory planning via imitation learning for autonomous vehicles," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 2736–2742.

[5] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[6] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 994–15 003.

[7] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[8] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[9] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[10] A. Sharma, Y. Bian, P. Munz, and A. Narayan, "Adversarial patch attacks and defences in vision-based tasks: A survey," *arXiv preprint arXiv:2206.08304*, 2022.

[11] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *European Computer Vision Conference (ECCV)*. Springer, 2020, pp. 665–681.

[12] Z. Hu, S. Huang, X. Zhu, F. Sun, B. Zhang, and X. Hu, "Adversarial texture for fooling person detectors in the physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 307–13 316.

[13] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 1028–1035.

[14] H. Yakura, Y. Akimoto, and J. Sakuma, "Generate (non-software) bugs to fool classifiers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1070–1078.

[15] X. Zhou, Z. Pan, Y. Duan, J. Zhang, and S. Wang, "A data independent approach to generate adversarial patches," *Machine Vision and Applications*, vol. 32, no. 3, pp. 1–9, 2021.

[16] A. Zolfi, S. Avidan, E. Yuval, and A. Shabtai, "Adversarial mask: Real-world universal adversarial attack on face recognition models," *arXiv 2111.10759*, 2021.

[17] Z. Chen, B. Li, S. Wu, J. Xu, S. Ding, and W. Zhang, "Shape matters: deformable patch attack," in *European Conference on Computer Vision*. Springer, 2022, pp. 529–548.

[18] H. Huang, Z. Chen, H. Chen, Y. Wang, and K. Zhang, "T-sea: Transfer-based self-ensemble attack on object detection," *arXiv preprint arXiv:2211.09773*, 2022.

[19] M. Lennon, N. Drenkow, and P. Burlina, "Patch attack invariance: How sensitive are patch attacks to 3d pose?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 112–121.

[20] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *European conference on computer vision*. Springer, 2020, pp. 665–681.

[21] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *ACM SIGSAC conference on computer and communications security*, 2016, pp. 1528–1540.

[22] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[23] L. Liu, "Deception, robustness and trust in big data fueled deep learning systems," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3–3.

[24] C. Li, S. Sun, Z. Wei, A. Tsourdos, and W. Guo, "Scarce data driven deep learning of drones via generalized data distribution space," *Neural Computing and Applications*, 2023.

[25] E. Soares, P. Angelov, and N. Suri, "Similarity-based deep neural network to detect imperceptible adversarial attacks," in *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022, pp. 1028–1035.

2023-09-01

Soft body pose-invariant evasion attacks against deep learning human detection

Li, Chen

IEEE

Li C, Guo W. (2023) Soft body pose-invariant evasion attacks against deep learning human detection. In: 2023 IEEE 9th International Conference on Big Data Computing Service and Applications (BigDataService), 17-20 July 2023, Athens, Greece. pp. 155-156

https://doi.org/10.1109/BigDataService58306.2023.00032