Adversarial Attacks on Event-Based Pedestrian Detectors: A Physical Approach

Guixu Lin[1, 2], Muyao Niu[1], Qingtian Zhu[1], Zhengwei Yin[1], Zhuoxiao Li[1], Shengfeng He[2], Yinqiang Zheng[1]*

[1]The University of Tokyo, Japan
[2]Singapore Management University, Singapore

Abstract

Event cameras, known for their low latency and high dynamic range, show great potential in pedestrian detection applications. However, while recent research has primarily focused on improving detection accuracy, the robustness of event-based visual models against physical adversarial attacks has received limited attention. For example, adversarial physical objects, such as specific clothing patterns or accessories, can exploit inherent vulnerabilities in these systems, leading to misdetections or misclassifications. This study is the first to explore physical adversarial attacks on event-driven pedestrian detectors, specifically investigating whether certain clothing patterns worn by pedestrians can cause these detectors to fail, effectively rendering them unable to detect the person. To address this, we developed an end-to-end adversarial framework in the digital domain, framing the design of adversarial clothing textures as a 2D texture optimization problem. By crafting an effective adversarial loss function, the framework iteratively generates optimal textures through backpropagation. Our results demonstrate that the textures identified in the digital domain possess strong adversarial properties. Furthermore, we translated these digitally optimized textures into physical clothing and tested them in real-world scenarios, successfully demonstrating that the designed textures significantly degrade the performance of event-based pedestrian detection models. This work highlights the vulnerability of such models to physical adversarial attacks.

Introduction

In the 30 milliseconds between frames captured by a traditional camera, a car traveling at 60 kilometers per hour can move approximately 0.5 meters. This significant frame delay makes traditional cameras unsuitable for applications requiring real-time perception and rapid response, such as pedestrian detection in traffic scenarios. In contrast, event cameras operate with an asynchronous triggering mechanism, offering ultra-low latency in the microsecond range and a wide dynamic range ($\geq$ 120 dB). These characteristics make event cameras a superior solution for pedestrian detection tasks, providing lower latency and faster response times compared to traditional cameras.

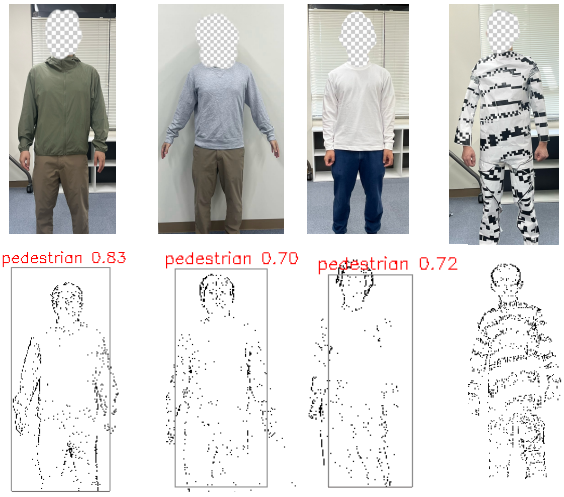


Figure 1: Demonstration of a physical adversarial attack: A person wearing adversarial clothing evades detection by an event-based pedestrian detector during movement, while a pedestrian in normal clothing is accurately detected. Bounding boxes indicate successful pedestrian detection.

In recent years, research on event-based pedestrian detection using deep learning methods has garnered significant attention, particularly with the introduction of large-scale datasets like 1MpX (Perot et al. 2020) and Gen1 (De Tournemire et al. 2020). Notably, RVT (Gehrig and Scaramuzza 2023) has highlighted the advantages of event cameras in pedestrian detection within traffic scenarios, achieving impressive results by processing event data alone. Further advancements in detection accuracy have been realized through innovative network architectures in models such as HMNet (Hamaguchi et al. 2023), GET (Peng et al. 2023), and SAST (Peng et al. 2024). However, the primary focus of current research has been on improving detection performance, with limited attention given to the potential vulnerabilities of event-based pedestrian detectors.

Given that these models are built on deep learning techniques, and considering the well-documented vulnerabilities of deep learning models in the RGB domain (Xu et al. 2021; Cai et al. 2022; Carlini and Wagner 2017), it is plausible

*Corresponding author.

that event-based pedestrian detectors may exhibit similar security weaknesses. In response, we conducted an unexplored investigation into the security of event-driven pedestrian detection models. In this way, event-based vision systems can enhance convenience without compromising human safety. Unlike previous studies (Marchisio et al. 2021; Lee and Myung 2022), which primarily focus on digital adversarial attacks by modifying event data, our work is the first to explore the impact of physical adversarial attacks on these detection models. Specifically, we examine how the clothing style of pedestrians can affect the performance of event-driven pedestrian detectors.

Figure 1 illustrates an example of a physical event attack. Our objective is to design clothing textures that can deceive event-based pedestrian detectors, rendering the wearer undetectable. To better simulate real-world scenarios, we developed an end-to-end digital adversarial attack framework. This framework utilizes 3D differentiable rendering techniques to transform the challenge of designing adversarial textures for physical attacks into a 2D texture optimization problem. By crafting a loss function tailored for adversarial attacks, the framework iteratively generates optimal 2D textures through backpropagation, facilitating the execution of the attack. To validate the effectiveness of the physical attack, we transferred the optimal texture derived in the digital domain into the physical world and conducted experiments, successfully executing the physical adversarial attack. Our goal in identifying the adversarial texture is to better develop defenses. This includes designing robust defense algorithms, integrating other types of sensors, and avoiding the fabrication of certain clothing that could interfere with the detection.

In summary, the contributions of this paper are threefold:

1. We present the first exploration of physical adversarial attacks in event-based vision, specifically validating these attacks in event-based pedestrian detection tasks.
2. We develop an end-to-end digital adversarial attack framework that transforms the design of 3D clothing textures into a 2D texture optimization problem using 3D differentiable rendering techniques, thereby enabling physical adversarial attacks on target detectors.
3. We demonstrate the effectiveness of the proposed method by successfully executing attacks on event-based pedestrian detectors, both in the digital domain and in real-world scenarios, highlighting the vulnerability of these detectors to physical adversarial attacks.

Related Work

Physical Adversarial Attacks

Deep learning-based vision models have been shown to be vulnerable to adversarial attacks, which typically involve subtle perturbations in the digital domain that can drastically alter model outputs (Xu et al. 2021; Cai et al. 2022; Carlini and Wagner 2017). As research in this area has evolved, there has been growing interest in exploring adversarial attacks within the physical world, where the perturbations are applied to objects or environments rather than digital inputs. These physical adversarial attacks are designed to be robust against real-world variations, such as changes in lighting, viewing angles, and distances, enabling them to effectively compromise vision AI models in practical scenarios (Wei et al. 2022).

Common forms of physical attacks include adversarial patches and stickers, which can be placed on objects or worn by people to fool detection systems. These attacks can be categorized into two main types: white-box and black-box. White-box attacks (Hu et al. 2021, 2023; Tan et al. 2021) require detailed knowledge of the target model, including its architecture and parameters, allowing attackers to craft highly effective adversarial perturbations. Black-box attacks (Li et al. 2021; Wei et al. 2020), on the other hand, do not require such knowledge and instead rely on probing the model's responses to various inputs to generate effective perturbations. In this work, we adopt a white-box approach to conduct physical adversarial attacks, leveraging our understanding of the model's inner workings to design precise adversarial patterns.

Adversarial Attacks on Event-based Vision

While most research on adversarial attacks has focused on RGB-based models, the vulnerabilities of visual AI systems that operate in other modalities, such as thermal infrared vision (Zhu et al. 2021, 2022; Wei et al. 2023), near-infrared vision (Niu et al. 2023), and event-based vision (Marchisio et al. 2021; Lee and Myung 2022), remain relatively underexplored. Event-based vision, in particular, offers unique challenges and opportunities for adversarial attack research due to its asynchronous, high-temporal-resolution nature.

Although research (Hao et al. 2023; Bu et al. 2023) on adversarial attacks in spiking neural networks (SNNs) is applicable to event data, this paper focuses specifically on adversarial attacks targeting deep learning models based on event-based data, extending beyond SNNs. For instance, Marchisio et al. (Marchisio et al. 2021) explored the creation of adversarial examples by projecting event-based data onto 2D images, generating 2D adversarial images. However, this approach indirectly targets event data and does not directly manipulate the raw input of event cameras, thereby limiting its effectiveness and applicability in real-world scenarios. Building on this, Lee et al. (Lee and Myung 2022) developed an adversarial attack algorithm specifically designed for event-based models. Their approach involved shifting the timing of events and generating additional adversarial events to deceive the model. This method was tested successfully on the N-Caltech101 dataset (Orchard et al. 2015), demonstrating the potential for digital adversarial attacks on event-based systems. However, these studies have primarily been confined to the digital domain, where the adversarial perturbations are applied to the event data directly rather than through physical means. In this paper, we advance the field by extending adversarial attacks on event-based vision into the physical domain, with a focus on pedestrian detection tasks. We design adversarial clothing textures that can deceive event-based pedestrian detectors in real-world scenarios, demonstrating the feasibility and effectiveness of physical adversarial attacks on event-based vision systems.

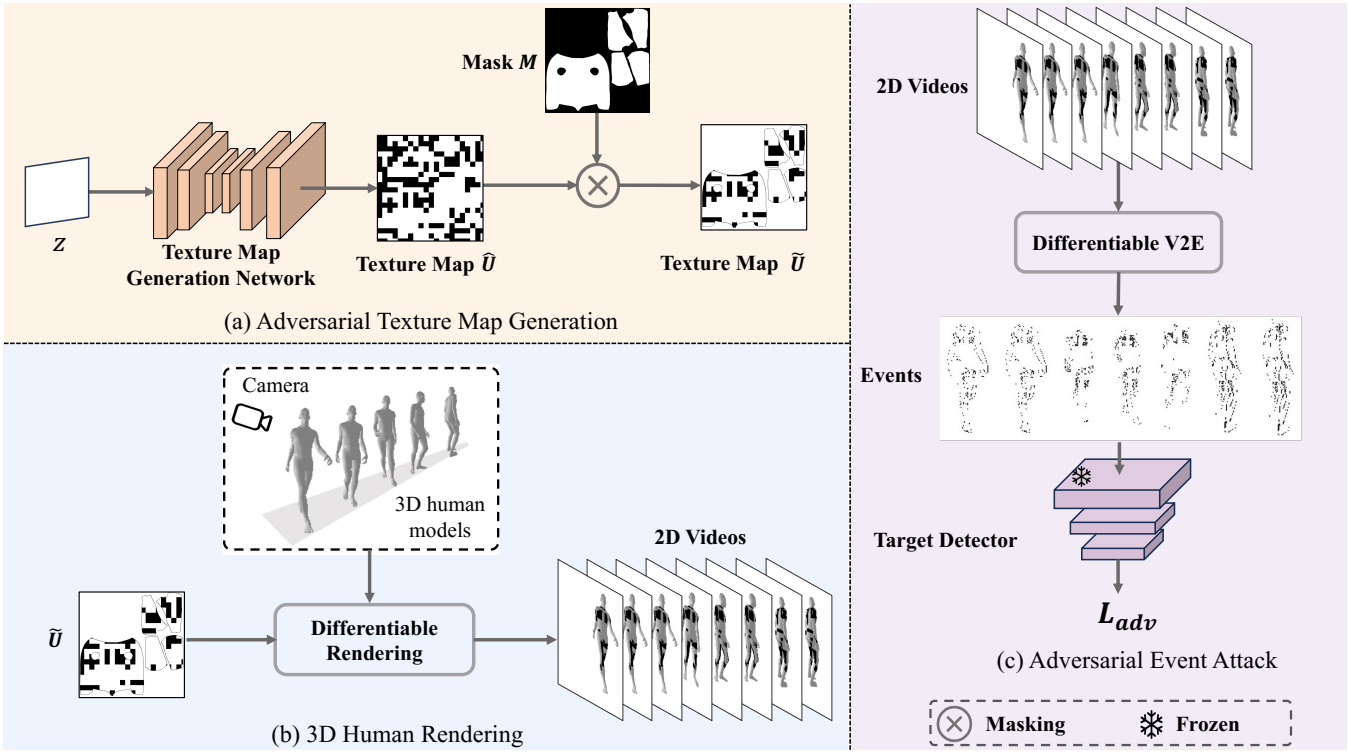


Figure 2: Demonstration of our method. (a) Adversarial Texture Map Generation: The input z is fed into a texture map generation network, producing a grayscale texture map $\hat{U}$. After applying a masking operation, the adversarial texture map $\tilde{U}$ is obtained. (b) 3D Human Rendering: The adversarial texture map $\tilde{U}$ is combined with 3D human model shape and pose parameters, and a differentiable renderer is used to generate 2D videos of continuous human motion. (c) Adversarial Event Attack: From these generated 2D videos, events $\tilde{E}$ are created using the differentiable V2E method and are used to attack event-based pedestrian detectors f, where the neural network parameters of f remain frozen. By applying the adversarial loss L_{adv}, the entire end-to-end pipeline is updated through backpropagation, ultimately resulting in the optimal adversarial texture map $\tilde{U}$.

Methodology

Formulation

Event Representation. An event stream consists of a series of events, with each event e_i characterized by the following elements: position (x_i, y_i), timestamp t_i, and polarity p_i. A positive polarity $p_i = +1$ indicates that the event is triggered when the logarithm of the pixel's intensity increases beyond a positive contrast threshold. Conversely, a negative polarity $p_i = -1$ indicates that the event is triggered when the logarithm of the pixel's intensity decreases beyond a negative contrast threshold. In this paper, we treat the positive and negative contrast thresholds equally, denoting them collectively as θ. Following the approach used in RVT (Gehrig and Scaramuzza 2023), we represent the event stream E over the time interval $[t_1, t_2)$ as follows:

$$E(p,\tau,x,y) = \sum_{e_i \in E} \delta(p-p_i)\delta(x-x_i, y-y_i)\delta(\tau-\tau_i), \quad (1)$$

where $\tau_i = \left\lfloor \frac{t_i - t_1}{t_2 - t_1} \cdot B \right\rfloor$ with B as the discretized time bins, and $\delta(x)$ stands for Dirac delta function.

Problem Definition. Let f denotes the pre-trained event-based pedestrian detection model. Given the events E as input, the outputs Y of the model consist of the bounding box position $f_{\text{pos}}(E)$, the object probability $f_{\text{obj}}(E)$, and the class score $f_{\text{cls}}(E)$, similar to most object detectors. This relationship is formulated as:

$$Y = f(E) = [f_{\text{pos}}(E), f_{\text{obj}}(E), f_{\text{cls}}(E)]. \quad (2)$$

Our goal is to fool the detector so that it fails to detect pedestrians. Specifically, we aim to simultaneously reduce both the object probability and the class score of pedestrians:

$$\min f_{\text{conf}}(E) = \min(f_{\text{obj}}(E) + f_{\text{cls}}(E)). \quad (3)$$

To achieve this, we develop a neural network G that generates a 2D texture map $\tilde{U}$ containing adversarial patches. This process involves first creating an initial texture map $\hat{U}$ from a parameter z, and then applying a mask M to obtain the expected texture map:

$$\tilde{U} = Filter(G(z), M). \quad (4)$$

Utilizing the 3D human model shape β, continuous pose parameters set $\boldsymbol{\phi}$, the texture map mask M, and the camera

extrinsic parameters $[R|t]$, we generate a sequence of 2D frames $\mathbf{I}$ through a differentiable rendering process $\mathcal{R}$. These frames represent a human figure wearing clothes with adversarial patches:

$$\mathbf{I} = \{I_k\}_{k=1}^{N} = \mathcal{R}(\tilde{U}, \beta, \boldsymbol{\phi}, [R|t]), \tag{5}$$

where N is the number of frames. The corresponding adversarial events $\tilde{E}$ are then generated using a differentiable video-to-event (V2E) method T, based on the event rendering times $\mathbf{t} = \{t_k\}_{k=1}^{N}$:

$$\tilde{E} = T(\mathbf{I}, \mathbf{t}). \tag{6}$$

Therefore, the objective function from Equation 3 can be reformulated as:

$$\arg\min f_{\text{conf}}(\tilde{E}). \tag{7}$$

Here, $\tilde{E}$ is defined as:

$$\tilde{E} = T(\mathcal{R}(Filter(G(z), M), \beta, \phi, [R|t]), \mathbf{t}). \tag{8}$$

The proposed method aims to identify the optimal adversarial texture map $\tilde{U}$, which generates adversarial events $\tilde{E}$ to deceive the event-based pedestrian detector f.

Adversarial Attack Framework

In this section, we provide a detailed overview of our adversarial attack pipeline, as illustrated in Figure 2. Our method consists of three key components: adversarial texture map generation, 3D human model rendering, and adversarial event attack. The adversarial texture map generation involves designing the texture pattern and developing the texture map generation network. The adversarial event attack includes the use of a differentiable V2E conversion method and the formulation of an adversarial loss function. By integrating these components, our approach systematically incorporates adversarial patches into event sequences, successfully misleading the event-based pedestrian detector.

Design of Texture Map Pattern. Since event cameras capture only changes in brightness, we simplify the design of clothing patterns for the attack by focusing on high-contrast color blocks, specifically black and white, without the need for colored grids. The pedestrian model's clothing texture is represented using two color blocks: black for low brightness areas and white for high brightness areas. The texture map $\hat{U}$ shown in Figure 3 (b) illustrates the pattern we designed.

Texture Map Generation Network. The texture map used in this paper has a resolution of $H \times W$, where $H = W$. It consists of an $n \times n$ grid of blocks, each $c \times c$ pixels, where $c = \frac{H}{n}$. As shown in Figure 3, we develop a texture map generation network that takes z as input, where z is a single-channel $n \times n$ image initialized to white (i.e., value = 1). After passing through the generation block, it outputs an $n \times n$ grayscale matrix u with values ranging from 0 to 1. This matrix is then binarized, resulting in an $n \times n$ texture map u_b composed solely of black and white (i.e., values are either 0 or 1). The final step involves an upsampling operation by a factor of c, producing the texture map $\hat{U}$ with a

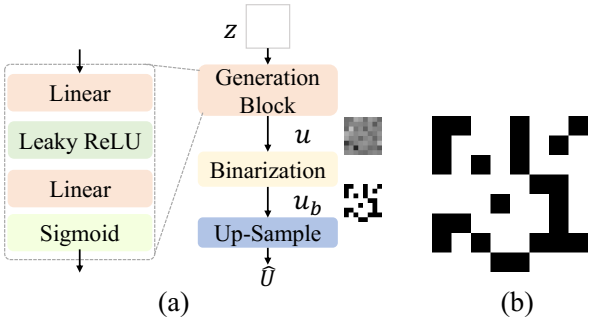


Figure 3: Illustration of the texture map generation network. (a) represents the network structure of the texture map generation network. (b) shows a demo of a texture map $\hat{U}$. The output $\hat{U}$ consists of an $n \times n$ grid of white or black blocks, where each block is $c \times c$ pixels in size.

resolution of $H = W$. The binarization operation is implemented using the Straight-Through Estimator (STE) (Bengio, Léonard, and Courville 2013), which allows the threshold function to be applied during the forward pass, while using approximate gradients during the backward pass. In the forward pass of STE, we perform the binarization using the following formula:

$$u_{bj} = \begin{cases} 1, & \text{if } u_j > 0.5 \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where $j \in [1, n \times n]$, u_j is the j-th value in u, and u_{bj} is the corresponding j-th value in u_b. In practice, to ensure training stability, we avoided direct hard binarization in the STE process. Instead, we adopted a soft binarization approach, gradually transitioning to hard binarization as the training progressed.

To increase rendering flexibility, such as excluding specific regions like the head, feet and fingers from texture rendering, we apply masking to occlude the corresponding areas in the $\hat{U}$, resulting in the texture map $\tilde{U}$, as shown in Figure 2. The masking process is formalized as follows:

$$\tilde{U}(x, y) = M(x, y) \cdot \hat{U}(x, y) + (1 - M(x, y)) \cdot \mathbf{1}, \tag{10}$$

where $\mathbf{1}$ represents a white image of the same size as the mask M.

3D Human Model Rendering. One of the key ideas of the proposed attack is to seek for a feasible pattern bounded to a 3D human with the adversarial loss based on 2D detection, in an end-to-end manner. To this end, we employ differentiable rendering to generate consecutive frames of one SMPL-based 3D human parameterized model (Bogo et al. 2016), enabling the back-propagation of gradients from 2D coordinates to the texture map. The 3D poses at different timestamps are animated from one canonical SMPL model so that the temporal consistency of UV mapping can be guaranteed. Specifically, we employ the implementation of PyTorch3D (Ravi et al. 2020) for differentiable rendering for its native compatibility with PyTorch.

Video to Event. To convert a 2D video clip into event data, we begin by transforming a sequence of N continuous images $\{I_k\}_{k=1}^{N}$ from the RGB color space to the YUV color space. We then extract the image sequence of the Y channel, which represents the luminance values, and convert these values into logarithmic space. Given a series of rendering times $\{t_k\}_{k=1}^{N}$ and a predefined contrast threshold, we calculate the differences between each pair of consecutive frames (e.g., I_k and I_{k+1}). Subsequently, we compute, in parallel, the number of all positive events $N_{t_{k+1}}^{p}$ and negative events $N_{t_{k+1}}^{n}$ between these frames. The differentiable event rendering mechanism we employ is adapted from (Gu et al. 2021), which makes the process of V2E differentiable by using a near-parallel event rendering reformulation. In contrast to the original paper, we fix the contrast threshold at $\theta = 0.2$ rather than using learnable thresholds.

Adversarial Loss. Following Equation 8, we develop the adversarial loss $\mathcal{L}_{adv}$ to simultaneously minimize both the object confidence score and the class confidence score. Specifically, the L_{obj} object confidence score is:

$$L_{obj} = \frac{1}{M} \sum_{i=1}^{M} f_{obj}^{(i)}(\tilde{E}), \tag{11}$$

where M is the number of attacked event sequences. The L_{cls} class confidence score is:

$$L_{cls} = \frac{1}{M} \sum_{i=1}^{M} f_{cls}^{(i)}(\tilde{E}). \tag{12}$$

The total loss L_{adv} is the sum of these two losses,

$$L_{adv} = \lambda_1 L_{obj} + \lambda_2 L_{cls}, \tag{13}$$

where λ_1 = λ_2 = 10,000 in the experiments. Using this adversarial loss, we employ back-propagation to iteratively update the adversarial texture maps.

Experiments

Settings

Metrics. We rely on two metrics to evaluate the effectiveness of the adversarial attack. The first is Average Precision (AP), a standard metric in detection tasks, where a lower AP indicates stronger adversarial performance. Since the target detector processes event sequences that record pedestrians, we are particularly interested in its ability to correctly identify pedestrians from these sequences. To capture this, we introduce a sequence-focused metric, the Sequence Attack Success Rate (SeqASR), defined as:

$$\text{SeqASR} = 1 - \frac{N}{M}, \tag{14}$$

where M is the total number of event sequences, and N is the number of sequences in which pedestrians are successfully detected. A higher SeqASR indicates a greater likelihood that the pedestrian detector fails to correctly identify pedestrians, demonstrating a more effective attack.

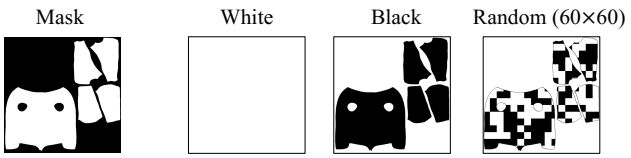


Figure 4: The masks used for all textures and the basic texture maps for comparison.

Implementation Details. We use RVT (Gehrig and Scaramuzza 2023) as the target detector, known for its excellent detection performance with event sequences. Specifically, we select the official pre-trained baseline model RVT-B for our adversarial attack experiments. This detection model was trained on the Gen1 automotive detection dataset (De Tournemire et al. 2020), which consists of real-world event data captured by event cameras at a resolution of 304 $\times$ 240. During both the training and evaluation phases, we keep the target detector frozen. To ensure compatibility with the target detector, the rendered 2D videos and events are also set to a resolution of 304 $\times$ 240. The texture map used in this study has a resolution of 1024 $\times$ 1024. The time bins B are set to 10. We train the network for 11,500 iterations using the Adam optimizer with an initial learning rate of 10^{-4}. Our framework is trained and validated with a batch size of 1 on an NVIDIA 3090 GPU.

Dataset in the Digital Space. We utilize the CMU motion capture dataset provided by AMASS (Mahmood et al. 2019), which includes various subjects in different poses. Each subject comprises several trials, and we select a subset of these trials for our experiments. Our training dataset consists of 47 trials, encompassing 114,173 poses, while the test dataset includes 13 trials, with a total of 10,323 poses. To enhance the robustness of our attack model, we apply data augmentation by introducing randomness to the camera's extrinsic parameters, varying the angles and sizes of the rendered human figures during training. For comprehensive evaluation, the test set includes 3D human renderings at three different scales.

Results

Digital Attack. Since the evaluation metrics are tied to the confidence thresholds used in the detector's post-processing stage, which can vary depending on the specific detection task, we selected four thresholds between 0.001 and 0.25—common values in detection tasks—to determine the optimal pattern and evaluate attack performance at different confidence levels. As shown in Table 1, these thresholds are 0.001, 0.01, 0.1, and 0.25, and the sequence length of the input event is 3. The mask selected for validation is shown in Figure 4. To better illustrate performance, we selected three texture maps for comparison: white, black, and a random pattern (60 $\times$ 60 pixels), as also illustrated in Figure 4.

We trained six different grid sizes, ranging from 60$\times$60 to 10$\times$10. The test results, shown in Table 1, indicated that the 20$\times$20 and 10$\times$10 grids achieved better AP and SeqASR scores. We selected the 10$\times$10 grid, which performed better

Metrics	Confidence Thresholds	Compared Texture White	Black	Random	Ours Adversarial Texture (*size of grids*) 60×60	50×50	40×40	30×30	20×20	10×10
AP ↓	0.001	46.8%	40.3%	41.4%	37.2%	39.3%	25.5%	17.1%	**11.1%**	11.8%
SeqASR ↑		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	**0.2%**
AP ↓	0.01	49.7%	44.8%	45.3%	41.0%	43.4%	32.6%	23.2%	**15.7%**	15.9%
SeqASR ↑		0.0%	0.0%	0.0%	0.1%	0.2%	0.0%	0.4%	**1.4%**	**1.4%**
AP ↓	0.1	50.6%	45.4%	46.2%	42.2%	44.5%	34.4%	22.2%	**12.2%**	12.5%
SeqASR ↑		0.0%	0.9%	0.0%	1.5%	1.4%	9.3%	18.2%	29.1%	**29.8%**
AP ↓	0.25	50.5%	43.8%	45.2%	39.4%	41.7%	28.0%	14.4%	6.1%	**5.7 %**
SeqASR ↑		1.2%	3.7%	4.7%	8.0%	9.0%	24.7%	43.2%	59.5%	**63.1%**

Table 1: Result of digital adversarial attack. The best performance is highlighted in **bold**.

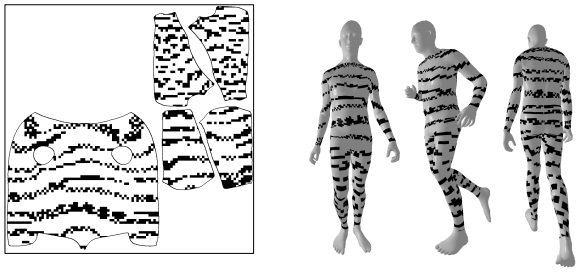
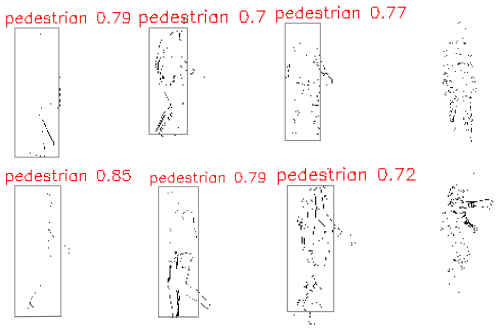


Figure 5: Visualization of the optimal texture map (grid size 10×10 pixels) and the corresponding rendered human.



(a) White (b) Black (c) Random (d) Optimal Texture

Figure 6: Visualization of digital attacks. Bounding boxes indicate the successful detection of pedestrians.

in the SeqASR metric, as the optimal texture pattern for subsequent experiments. The optimal 10×10 texture pattern is shown in Figure 5. The visualization of prediction results for different texture maps is shown in Figure 6, demonstrating that the optimal texture can successfully deceive the event-based pedestrian detector in the digital domain.

To further investigate the impact of textures on different body parts (including the upper body, legs, and arms) on overall attack performance, we selected the 10×10 grid size and trained with various mask combinations to identify the optimal pattern for each combination. The confidence threshold was set at 0.25 during training and testing. The

Body Parts in Texture Map			Metrics	
upper body	arms	legs	AP ↓	SeqASR ↑
✓			44.6%	4.7%
	✓		49.7%	1.2%
		✓	30.1%	21.9%
✓	✓		40.8%	5.6%
✓		✓	11.9%	52.0%
	✓	✓	21.6%	30.3%
✓	✓	✓	**5.7%**	**63.1%**

Table 2: Result of digital adversarial attack with different rendered body parts.

Metrics	Confidence Thresholds	Indoor Case1	Case2	Ours	Outdoor Ours
AP ↓	0.001	9.1%	15.1%	**0.0%**	7.6%
SeqASR ↑		0.0%	0.0%	**0.0%**	0.0%
AP ↓	0.01	12.6%	16.7%	**0.0%**	7.8%
SeqASR ↑		0.0%	0%	**2.1%**	1.4%
AP ↓	0.1	15.1%	19.2%	**0.0%**	6.0%
SeqASR ↑		0.0%	0.0%	**37.9%**	21.6%
AP ↓	0.25	14.5%	18.0%	**0.0%**	4.6%
SeqASR ↑		4.3%	0.0%	**74.3%**	46.0%

Table 3: Result of physical adversarial attack.

final evaluation results are presented in Table 2. The quantitative results indicate that the lower body (i.e., legs) exhibits the best attack performance among the three regions. Performance generally improves as more body parts are covered by the adversarial texture, with full-body coverage yielding the most significant impact, which aligns with our expectations.

Physical Attack. For the physical attack experiments, we used an INIVATION DAVIS346 MONO event camera to capture real-world event data, with a spatial resolution of 346×260 pixels. The output was cropped to 304×240 pixels to ensure compatibility with the target detector. The input event sequence length for the detector was set to 10. We then expanded the optimal texture pattern obtained from the dig-

Figure 7: Visualization of the clothes with optimal texture and the compared clothes for validation.

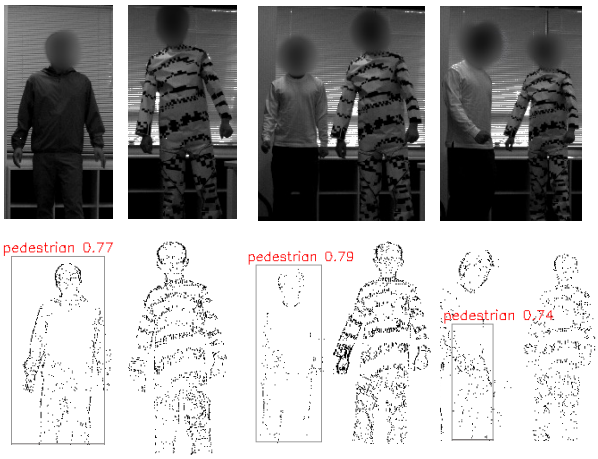


Figure 8: Visualization of physical attacks in indoor scenes. Bounding boxes indicate the successful detection of pedestrians. Images are cropped for better visualization.

ital attack experiments and printed it on paper. The printed textures were cut and assembled into clothing pieces according to body parts, as shown in Figure 7. For comparison, we selected two additional sets of clothing to serve as case studies, also illustrated in Figure 7. The event camera was fixed in place to record subjects wearing different textured clothing, with each subject performing the same set of actions. Since the event camera captures grayscale images alongside event data, we used YOLOv7 (Wang, Bochkovskiy, and Liao 2023) to annotate bounding boxes and classify these grayscale images, providing ground truth labels. For each texture, we collected 1,400 consecutive images and evaluated the AP and SeqASR metrics, as detailed in Table 3. In indoor scenes, we observed that the ordinary clothing in Case 1 and Case 2 exhibited lower SeqASR, while the optimal texture demonstrated superior adversarial performance in the physical attack. This indicates that the pattern remains effective when transitioning from the digital domain to physical attacks. Figure 8 illustrates the impact of clothing textures on event-based detection. In Figure 8 (a), the pedestrian wearing normal textures is successfully detected by the event-based detector. Conversely, in Figure 8 (b), the adversarial texture effectively evades detection. Panel (c) highlights the difference between normal and adversarial textures: the former is detected by the detector, while the latter bypasses it entirely.

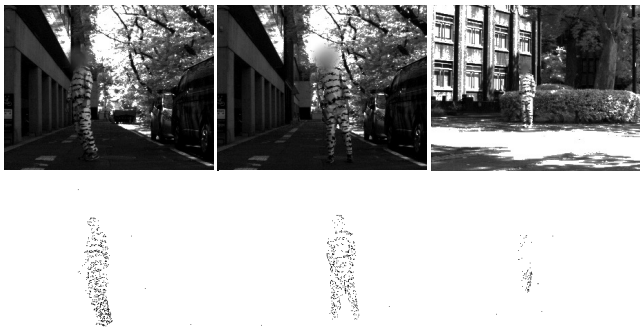


Figure 9: Visualization of the detection results in outdoor scenes. None of the pedestrians is detected.

When considering the optimal texture across both indoor and outdoor scenes, the AP and SeqASR metrics reveal a decrease in the effectiveness of our adversarial clothing in outdoor environments. This reduction can be attributed to the more complex, dynamic backgrounds and fluctuating lighting conditions present in outdoor settings. Despite these challenges, the physical adversarial attack success rate remains notably high even in outdoor scenarios. Figure 9 visualizes the detection results in outdoor scenes, showing that a pedestrian wearing adversarial textured clothing effectively evades detection by the event-based pedestrian detector.

Conclusion

This paper presents an end-to-end method for creating adversarial clothing designed to attack event-based pedestrian detectors in the physical domain. We address the challenge of identifying the most effective adversarial clothing by formulating it as a task of optimizing a 2D texture map in the digital domain. Through the use of 3D rendering techniques, this optimized texture pattern is mapped onto a 3D human model, where it demonstrates strong adversarial effectiveness in the digital space. We then successfully translate this texture pattern into the physical domain, achieving comparable attack results. Our findings indicate that event-based pedestrian detectors, much like their RGB-based counterparts, are vulnerable to security breaches.

Limitations and Future Work. This study focuses on designing black-and-white adversarial patches for texture mapping, but real-world clothing typically features a wider range of colors and more complex textures. While effective, our approach will be extended to incorporate more realistic and diverse clothing styles, aligning better with real-world scenarios. Additionally, we will develop advanced defense mechanisms to counter various physical adversarial attacks.

Acknowledgments

We express our gratitude to Mingze Ma and Yifan Zhan for their contributions to conducting the experiments. This research was supported in part by JSPS KAKENHI Grant Numbers 24K22318, 22H00529, 20H05951, JST-Mirai Program JPMJMI23G1, and ROIS NII Open Collaborative Research 2023-23S1201, JST SPRING (Grant Number JPMJSP2108), Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the AI Singapore Programme under the National Research Foundation Singapore (Grant AISG3-GV-2023-011), and the Lee Kong Chian Fellowships.

References

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 561–578. Springer.

Bu, T.; Ding, J.; Hao, Z.; and Yu, Z. 2023. Rate gradient approximation attack threats deep spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7896–7906.

Cai, Z.; Xie, X.; Li, S.; Yin, M.; Song, C.; Krishnamurthy, S. V.; Roy-Chowdhury, A. K.; and Asif, M. S. 2022. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 149–157.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

De Tournemire, P.; Nitti, D.; Perot, E.; Migliore, D.; and Sironi, A. 2020. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*.

Gehrig, M.; and Scaramuzza, D. 2023. Recurrent vision transformers for object detection with event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Gu, D.; Li, J.; Zhang, Y.; and Tian, Y. 2021. How to learn a domain-adaptive event simulator? In *Proceedings of the 29th ACM International Conference on Multimedia*, 1275–1283.

Hamaguchi, R.; Furukawa, Y.; Onishi, M.; and Sakurada, K. 2023. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22867–22876.

Hao, Z.; Bu, T.; Shi, X.; Huang, Z.; Yu, Z.; and Huang, T. 2023. Threaten spiking neural networks through combining rate and temporal information. In *The Twelfth International Conference on Learning Representations*.

Hu, Y.-C.-T.; Kung, B.-H.; Tan, D. S.; Chen, J.-C.; Hua, K.-L.; and Cheng, W.-H. 2021. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7848–7857.

Hu, Z.; Chu, W.; Zhu, X.; Zhang, H.; Zhang, B.; and Hu, X. 2023. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16975–16984.

Lee, W.; and Myung, H. 2022. Adversarial attack for asynchronous event-based data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1237–1244.

Li, S.; Aich, A.; Zhu, S.; Asif, S.; Song, C.; Roy-Chowdhury, A.; and Krishnamurthy, S. 2021. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34: 2085–2096.

Mahmood, N.; Ghorbani, N.; F. Troje, N.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *The IEEE International Conference on Computer Vision (ICCV)*.

Marchisio, A.; Pira, G.; Martina, M.; Masera, G.; and Shafique, M. 2021. Dvs-attacks: Adversarial attacks on dynamic vision sensors for spiking neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Niu, M.; Li, Z.; Zhan, Y.; Nguyen, H. H.; Echizen, I.; and Zheng, Y. 2023. Physics-Based Adversarial Attack on Near-Infrared Human Detector for Nighttime Surveillance Camera Systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8799–8807.

Orchard, G.; Jayawant, A.; Cohen, G. K.; and Thakor, N. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9: 437.

Peng, Y.; Li, H.; Zhang, Y.; Sun, X.; and Wu, F. 2024. Scene Adaptive Sparse Transformer for Event-based Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16794–16804.

Peng, Y.; Zhang, Y.; Xiong, Z.; Sun, X.; and Wu, F. 2023. Get: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6038–6048.

Perot, E.; De Tournemire, P.; Nitti, D.; Masci, J.; and Sironi, A. 2020. Learning to detect objects with a 1 megapixel event camera. In *Adv. Neural Inform. Process. Syst.*

Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501*.

Tan, J.; Ji, N.; Xie, H.; and Xiang, X. 2021. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *Proceedings of the 29th ACM international conference on multimedia*, 5307–5315.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, H.; Tang, H.; Jia, X.; Wang, Z.; Yu, H.; Li, Z.; Satoh, S.; Van Gool, L.; and Wang, Z. 2022. Physical adversarial attack meets computer vision: A decade survey. *arXiv preprint arXiv:2209.15179*.

Wei, H.; Wang, Z.; Jia, X.; Zheng, Y.; Tang, H.; Satoh, S.; and Wang, Z. 2023. Hotcold block: Fooling thermal infrared detectors with a novel wearable design. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 15233–15241.

Wei, Z.; Chen, J.; Wei, X.; Jiang, L.; Chua, T.-S.; Zhou, F.; and Jiang, Y.-G. 2020. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12338–12345.

Xu, Q.; Tao, G.; Cheng, S.; and Zhang, X. 2021. Towards feature space adversarial attack by style perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10523–10531.

Zhu, X.; Hu, Z.; Huang, S.; Li, J.; and Hu, X. 2022. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13317–13326.

Zhu, X.; Li, X.; Li, J.; Wang, Z.; and Hu, X. 2021. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3616–3624.