Cheating Stereo Matching in Full-scale: Physical Adversarial Attack against Binocular Depth Estimation in Autonomous Driving

Kangqiao Zhao[1*], **Shuo Huai**[1*], **Xurui Song**[2], **Jun Luo**[1†]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]S-Lab, Nanyang Technological University, Singapore
{kangqiao.zhao, shuo.huai, junluo}@ntu.edu.sg, song0257@e.ntu.edu.sg

Abstract

Though deep neural models adopted to realize the perception of autonomous driving have proven vulnerable to adversarial examples, known attacks often leverage 2D patches and target mostly monocular perception. Therefore, the effectiveness of Physical Adversarial Examples (PAEs) on stereo-based binocular depth estimation remains largely unexplored. To this end, we propose the first texture-enabled physical adversarial attack against stereo matching models in the context of autonomous driving. Our method employs a 3D PAE with global camouflage texture rather than a local 2D patch-based one, ensuring both visual consistency and attack effectiveness across different viewpoints of stereo cameras. To cope with the disparity effect of these cameras, we also propose a new 3D stereo matching rendering module that allows the PAE to be aligned with real-world positions and headings in binocular vision. We further propose a novel merging attack that seamlessly blends the target into the environment through fine-grained PAE optimization. It has significantly enhanced stealth and lethality upon existing hiding attacks that fail to get seamlessly merged into the background. Extensive evaluations show that our PAEs can successfully fool the stereo models into producing erroneous depth information.

Introduction

Recent studies have revealed the vulnerabilities of deep neural perception models for *Autonomous Driving* (AD) against physical adversarial attacks (Lovisotto et al. 2021; Zhu et al. 2021b; Cao et al. 2021). These attacks rely on physically deployable artifacts, also known as *Physical Adversarial Example*s (PAEs) (Eykholt et al. 2018; Thys et al. 2019), to affect the vision features captured by the neural networks and then manipulate their predictions. Currently, most PAEs are in the form of 2D patches that affect only a limited fraction of an object, and they often target only *Monocular Depth Estimation* (MDE) models (Cheng et al. 2022; Guesmi et al. 2023; Wong, Mundhra, and Soatto 2021; Cheng et al. 2021). Consequently, these threats to AD might be readily bypassed in real-world 3D scenarios with the support of stereo-based *Binocular Depth Estimation* (BDE) (Xu and Zhang 2020; Wang et al. 2019; Li, Chen, and Shen 2019).

In fact, attacks on BDE driven by Stereo Matching (SM), a depth estimation method leveraging visual disparity, remain very limited. Sharply different from the PAEs for MDE, attacking SM models presents unique challenges due to its reliance on the physical 3D disparity between two images, rather than only on the 2D image pixels captured by a single camera; it is this uniqueness that invalidates most of MDE attacks under SM-BDE. Although a few proposals (Liu et al. 2024; Cheng et al. 2021) have explored attacks on stereo matching, they mainly rely on 2D patches overlaid in *image space*, which not only limits their effectiveness to narrow viewpoints, but also breaks the physical disparity consistency required by real-world systems, rendering them ineffective under physical deployment. Meanwhile, SM-BDE is increasingly adopted in AD, with dedicated datasets available to provide stereo data for training SM-BDE models (Geiger, Lenz, and Urtasun 2012; Mayer et al. 2016). Therefore, it makes sense to consider full-scale 3D PAEs that are effective against SM-BDE, because only such PAEs can substantiate their realistic threat to real-world AD. To generate such PAE, one would need a global 3D camouflage texture to cover the whole surface of a target object, because such a texture can significantly increase the affected range within camera views and thus achieve a full-scale adversarial influence on the target.

Adapting PAEs to SM-BDE may also enable more meaningful attacks. Existing attacks on depth estimation primarily focused on hiding objects by pushing their estimated depth to infinity (Guesmi et al. 2023; Zheng et al. 2024; Cheng et al. 2022). As a result, such attacks fail to really hide any objects, as the contour of the object remains, totally exposing the attackers' intention. A more meaningful attack, as one may intuitively expect, should involves adjustable factors for depth fine-tuning, so that attackers can completely adapt the depth of different parts of the target object (including contours) into the surroundings, which we refer to as *merging attack*. Nonetheless, realizing the 3D PAE faces a great challenge under SM-BDE, as each pair of rendered images must precisely align with the physical geometry of stereo cameras in the real world. In addition to making PAEs viable under SM-BDE, a qualified design should also remain robust to varying viewpoints (e.g., position and heading) of the SM camera pair, as well as to diversified external environments (e.g., lighting and weather conditions).

[*]These authors contributed equally.
[†]Corresponding author

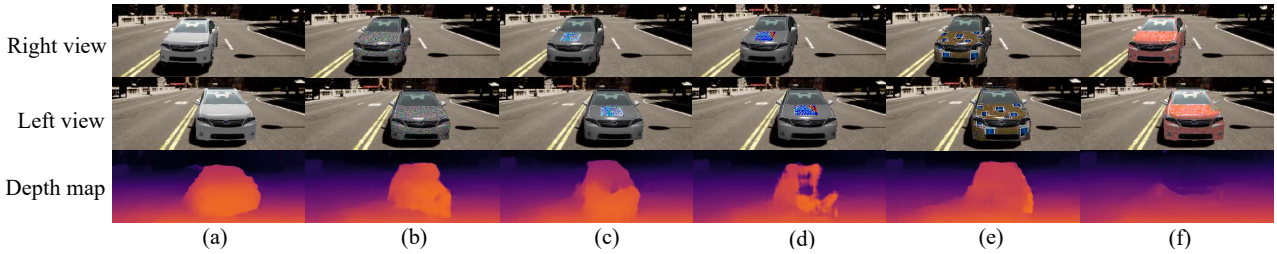

Figure 1: Comparison between existing PAEs and our 3D texture-enabled *merging attack* PAEs in CARLA simulator (Dosovitskiy et al. 2017). (a) Benign right and left camera images with the corresponding depth map; (b) Random texture; (c), (d) 2D adversarial patches from (Liu et al. 2024) and (Cheng et al. 2021), respectively, applied to 3D vehicle. (e) MDE Adversarial texture from (Zheng et al. 2024); (f) Our 3D *merging attack* adversarial texture.

To address the fundamental limitation of image-space attacks under stereo vision, we propose a novel stereo-consistent rendering module that enables physically realizable adversarial camouflage. We leverage 3D object detection to precisely anchor a full-surface adversarial model onto the real-world object with accurate pose alignment. This ensures that the rendered adversarial object produces disparity-consistent projections across both stereo views, a property essential for fooling depth perception in real-world SM-BDE systems. To realize *merging attack*, we incorporate depth around the contour of the target object to optimize our adversarial camouflage texture. This enables us to understand the contextual information around the object and separate the object into distinct regions with tailored optimization strategies. Compared to uniformly treating the entire object as an optimization target, this approach allows us to make more diversified adjustments.

Additionally, our PAE optimization accounts for diverse environmental variations, including changes in distances, rotations lighting, and weather conditions. Figure 1 presents our *merging attack* PAEs compared to prior works (Liu et al. 2024; Cheng et al. 2021; Zheng et al. 2024). The results demonstrate that, under real-world 3D stereo matching constraints, only our method remains effective. Our main contributions are summarized as follows:

- We introduce the first texture-enabled physical adversarial attack against stereo matching models in AD. Our PAE is in form of a 3D camouflage texture attached to the surface of the target object, which can affect object's projection on stereo cameras across various viewpoints.
- We design a novel stereo-aligned 3D rendering module that integrates our adversarial object precisely into the scene using 3D object detection, ensuring faithfully preserving real-world binocular disparity.
- We propose a novel and enhanced hiding attack method: *merging attack*, which uses fine-grained PAE optimization to seamlessly blend the target object into the background, achieving superior stealth and undetectability.

We evaluate our attack in both digital simulations and real-world stereo captures, demonstrating strong generalization under varying lighting, weather, and viewpoints. When deployed in a full AD perception stack, our textures cause critical downstream failures, including missed obstacles and faulty path planning, highlighting serious safety risks.

Related Work and Background

Deep Stereo Matching. SM-BDE has received significant attention for its reliability and efficiency in depth estimation for AD (Baidu 2025; Waymo 2025; Mobileye 2025). Early works (Zagoruyko and Komodakis 2015; Zbontar and LeCun 2016) leverage deep networks for separate feature extraction and matching. Subsequent advances include PSMNet (Chang and Chen 2018), which incorporates pyramid pooling and 3D CNNs; GA-Net (Zhang et al. 2019), which introduces attention-based cost aggregation; and RAFT-Stereo (Lipson, Teed, and Deng 2021), which uses recurrent units for multi-level refinement. Recent works explore cascaded recurrent networks, such as CREStereo (Li et al. 2022), and focus on robust generalization and arbitrary-scale disparity, as demonstrated by AnyStereo (Liang and Li 2024). Given the rapid development of SM-BDE in AD, addressing its security has become critical.

Physical Adversarial Attack. Adversarial attacks exploit vulnerabilities in DNN by introducing crafted perturbations to input data, causing the model to make incorrect predictions (Goodfellow et al. 2014; Madry et al. 2017). PAEs have shown real-world effectiveness (Kurakin et al. 2018; Eykholt et al. 2018), sparking broad interest in safety-critical domains such as AD. They have been studied extensively across key AD tasks, including 3D and 2D object detection (Tu et al. 2020; Abdelfattah et al. 2021; Cao et al. 2021), object tracking (Schmalfuss, Scholze, and Bruhn 2022; Ranjan et al. 2019), object segmentation (Zhu et al. 2021a; Xie et al. 2017), and MDE (Cheng et al. 2022; Guesmi et al. 2023; Wong, Mundhra, and Soatto 2021). Recent works (Liu et al. 2024; Cheng et al. 2021) also explored physical attacks on SM-BDE using 2D PAE but oversimplify stereo image relationships by assuming a direct shift between left and right views, neglecting stereo viewpoint differences. While this assumption may be efficient for digital 2D patch attacks, it fails to generalize to the complexities of real-world 3D scenarios, as shown in Figure 1. However, our 3D camouflage texture as PAE can conform to the parametric constraints of stereo cameras and achieve state-of-the-art attack effectiveness across diverse viewpoints on SM models.

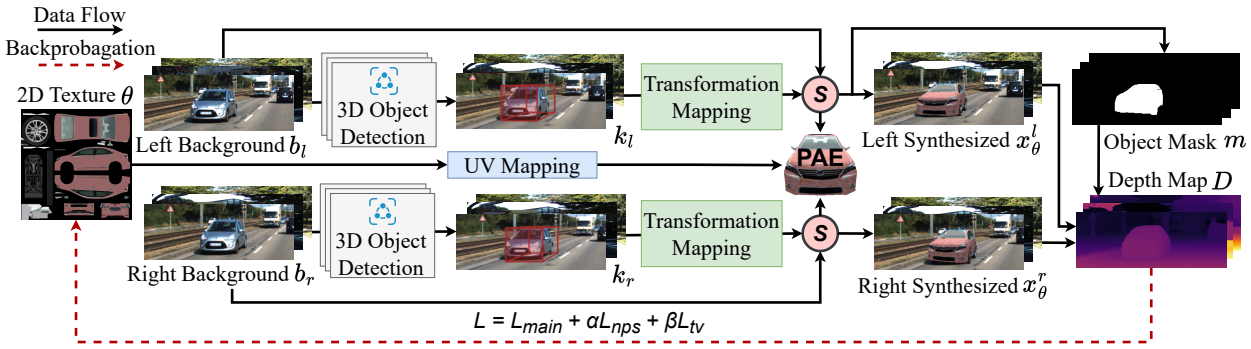

Figure 2: Overview of our attack against SM-BDE models. The adversarial camouflage texture θ is mapped on the PAE and synthesized with the backgrounds (b_l and b_r) using a differentiable renderer R with rendering configurations (k_l and k_r). The optimization of θ is driven by backpropagation, guided by a loss function tailored to the objectives of the *merging attack*.

Methodology

Problem Definition

In this paper, we aim to generate a physically deployable adversarial texture θ, which can be applied to a real-world object (e.g., a vehicle), such that when observed by a stereo camera, the object can not be perceived by SM-BDE models. This attack, which we refer to as the *merging attack*, misleads the SM-BDE model into assigning background depth to the object, reducing its visibility in the final depth map.

Unlike prior approaches that rely on physically unrealistic perturbations or 2D patches that fail under varying viewpoints, our method optimizes a static adversarial texture that can be physically applied (e.g., printed) to the vehicle surface. Once deployed, this texture remains effective across diverse scenes, viewpoints, and lighting conditions, without any online adjustment or access to stereo inputs.

To simulate real-world stereo perception during training, we use a differentiable rendering pipeline. Given a stereo background image pair $b = (b^l, b^r)$ and a detected object, we render a textured 3D mesh O with adversarial texture θ under a stereo camera configuration $k = (k_l, k_r)$, using a renderer R. The rendered object is then composited with the original background images through a synthesis module S:

$$x_\theta = \mathrm{S}(\mathrm{R}(O, \theta, k), b, m), \tag{1}$$

where m is the object mask used for blending. The resulting image pair $x_\theta = (x_\theta^l, x_\theta^r)$ simulates how the physical object would appear in both stereo views with θ. We then optimize θ such that the SM-BDE model F produces depth predictions close to a background-informed target depth d_t:

$$\theta = \arg\min_\theta \mathcal{L}(F(x_\theta^l, x_\theta^r), d_t). \tag{2}$$

After optimizing, the result is a geometry-aware, physically realizable adversarial texture θ that deceives stereo depth models under real-world deployment.

Stereo-Aligned 3D Rendering

While physical adversarial attacks have been widely explored in MDE, their extension to SM-BDE is far from straightforward. MDE models infer depth from single image, and adversarial rendering in such settings can be flexibly adjusted, as long as the generated image appears visually plausible. In contrast, SM models depend fundamentally on binocular disparity between two images captured from rigid, spatially coupled viewpoints, which imposes strict physical constraints on how adversarial examples must be rendered.

Specifically, a successful stereo physical attack must ensure that: i) the adversarial object maintains geometrically consistent appearances across both left and right views; ii) the surrounding background context remains coherent in both images to support reliable disparity computation; and iii) the two camera viewpoints follow physically accurate stereo baselines, rather than synthetic approximations. To address these challenges, we introduce a stereo rendering configuration framework, parameterized by k, which defines a complete physical stereo setup, inlcuding camera intrinsics, extrinsics, and object placement. This governs how adversarial scenes are rendered. Unlike monocular rendering pipelines, our formulation ensures disparity-consistent dual-view generation, forming the geometric foundation for physically realizable, stereo-consistent adversarial attacks. This design enables us to, for the first time, align adversarial optimization with the physical constraints of real-world stereo-based perception systems in autonomous driving.

To generate physically consistent stereo AEs, as shown in Figure 2, we begin by selecting a stereo image pair from a real-world scene that contains a visible vehicle. Using this pair, we first apply a 3D object detection model to obtain the vehicle's pose and size, represented by a 3D bounding box:

$$\mathrm{bbox} = \{t_x, t_y, t_z, t_l, t_w, t_h, t_r, t_c\}, \tag{3}$$

where (t_x, t_y, t_z) denotes the bounding box's center position in the world coordinate system, (t_l, t_w, t_h) represent the bounding box dimensions (length, width, and height), t_r is the heading angle (rotation around the y-axis), and t_c specifies the object category. As the render camera is set to always look toward the center of the adversarial object, we adopt a spherical coordinate system to parameterize the rendering

viewpoint as $k = \{\text{dist}, \text{elev}, \text{azim}\}$, computed as:

$$\begin{aligned} \text{dist} &= \sqrt{(c_x - t_x)^2 + (c_y - t_y)^2 + (c_z - t_z)^2}, \\ \text{elev} &= \arctan\left(\frac{c_y - t_y}{\sqrt{(c_x - t_x)^2 + (c_z - t_z)^2}}\right), \\ \text{azim} &= \arctan\left(\frac{c_x - t_x}{c_z - t_z}\right). \end{aligned} \tag{4}$$

We parameterize the stereo viewpoints as (k_l, k_r) for the left and right cameras, respectively, based on their 3D positions (c_x, c_y, c_z) obtained from stereo calibration. Using these configurations, we independently render a 3D adversarial vehicle with optimized texture θ into both views, ensuring disparity-consistent appearance. Our method replaces the original object with an adversarial mesh, preserving stereo alignment and depth consistency across the scene.

Merging Attack Texture Generation

In real-world images, objects farther from the camera appear smaller and exhibit lower texture detail. Leveraging this property, we aim to optimize the adversarial texture such that the vehicle visually and geometrically blends into its background. We focus on aligning the vehicle with the surrounding background rather than nearby foreground objects. However, this background similarity is not uniform across the adversarial vehicle: the lower regions of the vehicle, which are closer to the ground, tend to align well with the nearby background, while the upper parts show larger depth gaps. This motivates a region-aware optimization strategy that adapts different parts of the object to their respective depth contexts. Specifically, to ensure effective *merging attack*, we design a three-step process: boundary depth extraction, region segmentation, and texture optimization.

During boundary depth extraction, we obtain the surrounding background depth from the scene's depth map d by expanding the object mask m (see Figure 3). Applying max pooling with kernel size $p_k \times p_k$ enlarges m to cover nearby regions, and subtracting the original mask yields the boundary mask m_{bg}. We then extract the background depth d_{bg} via element-wise multiplication with d:

$$m_{bg} = \text{Maxpool}(m) - m, \quad d_{bg} = d \cdot m_{bg}. \tag{5}$$

Next, in the region segmentation stage, we divide the adversarial vehicle's depth map horizontally to optimize different regions for varying surrounding depths. To balance simplicity and effectiveness, we divide the map into upper and lower regions, as shown in Figure 3. To anchor this division, we compute the mean background depth from the previously extracted boundary map, $d_{bg}^{avg} = \sum d_{bg} / \sum m_{bg}$, to split the vehicle depth. We then locate two reference points on the left and right sides of the object boundary whose depths are closest to d_{bg}^{avg}. These points define a horizontal contour across the object surface that approximately aligns with the background depth plane, guiding the split between regions. This segmentation enables the upper and lower parts of the texture to be optimized toward distinct depth priors, enhancing the effectiveness of *merging attack*.

Finally, during texture optimization, we introduce a *region-aware depth alignment loss* that guides the adversarial texture to match the depth statistics of the surrounding background, aiming to achieve seamless blending between the adversarial object and its environment. Specifically, we compute the average depth of the object and the surrounding region in both the upper and lower segments defined by the horizontal partition. The main *merging loss* is formulated as:

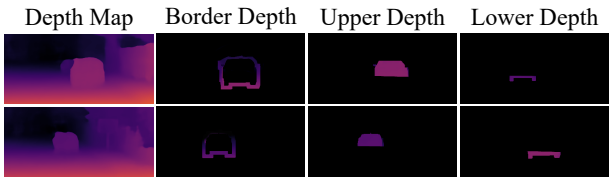

Figure 3: Depth results of boundary and each segmentation.

$$\mathcal{L}_{\text{merge}}(\theta) = \text{MSE}(d_{obj}^{up}, d_{bg}^{up}) + \text{MSE}(d_{obj}^{bt}, d_{bg}^{bt}), \tag{6}$$

where d_{obj}^{up} and d_{obj}^{bt} denote the average depth of the object's upper and bottom regions, respectively, and d_{bg}^{up}, d_{bg}^{bt} are the corresponding local background depths extracted around the object boundary. By independently optimizing each region to align with its local background depth, the attack blend the adversarial vehicle into the scene, resulting in a more concealed and less detectable adversarial effect.

To ensure that the adversarial texture is both physically printable and visually smooth, we further incorporate two regularization terms: a total variation (TV) loss ($\mathcal{L}_{\text{tv}}$) (Mahendran and Vedaldi 2015) that suppresses high-frequency noise, and a non-printability score (NPS) loss ($\mathcal{L}_{\text{nps}}$) (Sharif et al. 2016) that encourages the use of printer-reproducible colors. The final loss function is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{merge}}(\theta) + \alpha \mathcal{L}_{\text{nps}}(\theta) + \beta \mathcal{L}_{\text{tv}}(\theta), \tag{7}$$

where α and β are hyperparameters balancing the merging objective, printability, and smoothness.

Appearing Attack. Our *merging attack* represents an advanced form of hiding attack, designed to reduce the visibility of the target vehicle. Complementary to *merging attack*, we also introduce a strategy, *appearing attack*, which aims to make the target vehicle highly conspicuous. This could forces surrounding vehicles to engage in sudden braking maneuvers, potentially causing accidents and unexpected disruptions in traffic flow. To achieve this objective, we define the loss function $\mathcal{L}_{\text{appear}}$ for texture optimization, which seeks to minimize the perceived depth within the target mask m, thereby minimizing the visual distance between the adversarial object and the camera. This is achieved by minimizing the MSE between d_{obj} and D_{max}, where D_{max} is set according to model's configuration for upper limit:

$$\mathcal{L}_{\text{appear}}(\theta) = \text{MSE}(d_{obj}, D_{max}). \tag{8}$$

The total loss $\mathcal{L}(\theta)$ for our *appearing attack* follows a formulation similar to that of the *merging attack*, with $\mathcal{L}_{\text{merge}}(\theta)$ in Eq. (7) replaced by $\mathcal{L}_{\text{appear}}(\theta)$.

Experiments

Experiment Setup

Implementation Details. We train our adversarial textures using the KITTI2015 dataset (Geiger, Lenz, and Ur-

Method	PSMNet			GA-Net			RAFT-Stereo			CREStereo			AnyStereo		
	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{cover}} \uparrow$	$\mathcal{E}_{\text{sys}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{cover}} \uparrow$	$\mathcal{E}_{\text{sys}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{cover}} \uparrow$	$\mathcal{E}_{\text{sys}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{cover}} \uparrow$	$\mathcal{E}_{\text{sys}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{cover}} \uparrow$	$\mathcal{E}_{\text{sys}} \uparrow$
Benign	0.631	0.013	0	0.641	0.012	0	0.786	0.012	0	0.677	0.017	0	0.572	0.093	0
Random	0.677	0.052	0	0.680	0.031	0	0.793	0.024	0	0.684	0.029	0	0.580	0.022	0
PASM	0.475	0.154	0.13	0.411	0.088	0.12	0.502	0.148	0.07	0.431	0.094	0.15	0.471	0.124	0.15
Adv-DM	0.510	0.176	0.04	0.449	0.075	0.12	0.614	0.143	0.05	0.444	0.077	0.17	0.480	0.119	0.09
Ours	**0.058**	**0.553**	**0.74**	**0.069**	**0.588**	**0.69**	**0.082**	**0.571**	**0.62**	**0.071**	**0.598**	**0.70**	**0.056**	**0.576**	**0.76**

Table 1: Comparison results between existing attacks and our *merging attack* against various target models.

tasun 2012), a widely adopted benchmark for AD that provides stereo image pairs. The training samples cover diverse vehicle poses, scales, and backgrounds, promoting generalization across real-world scenes. To enhance the robustness of adversarial textures under diverse environmental conditions, we adopt Expectation over Transformation (EoT) (Athalye et al. 2018) during training. Specifically, we randomly perturb the position of a point light source within the range $[-3, 3]$ meters along each axis, and uniformly sample ambient light intensity from $[0.3, 0.9]$. To simulate challenging weather such as rain and fog, we additionally inject random Gaussian noise into the rendered images during training. For boundary depth extraction, the kernel size p_k is set to 40. We optimize the adversarial texture using a differentiable renderer and train for 100 epochs with the Adam optimizer (Kingma 2014), with an initial learning rate of 0.01, scheduled via cosine decay to a minimum of $1e-4$. The loss coefficients are set to $\alpha = 5$ and $\beta = 0.1$. All experiments are performed on a single NVIDIA RTX A5000 GPU.

Evaluation Metrics. We evaluate our 3D adversarial camouflage attack from both perception and system perspectives using three metrics: i) Hiding Error ($\mathcal{E}_{\text{blend}}$), which evaluates how well the adversarial object blends into its surroundings by measuring the proportion of object pixels whose depth significantly deviates from that of the background contour:

$$\mathcal{E}_{\text{blend}} = \frac{1}{\sum_m} \cdot \sum_{(i,j)\in m} \mathbb{I}\left(|d(i,j) - \bar{d}_{\text{bg}}(i)| > \tau\right), \quad (9)$$

where $d(i, j)$ is the predicted depth at pixel (i, j), and $\mathbb{I}(\cdot)$ is the indicator function. The contour average depth $\bar{d}_{\text{bg}}(i)$ is computed per row i using pixels selected by m_{bg}. τ is a predefined threshold used to determine various depth differences; we set $\tau = 20$ in our experiments. ii) Perturbation Coverage Ratio ($\mathcal{E}_{\text{cover}}$), which quantifies the fraction of the object region where depth has been altered by the :

$$\mathcal{E}_{\text{cover}} = \frac{1}{\sum_m} \cdot \sum_{(i,j)\in m} \mathbb{I}\left(d^{\text{adv}}(i,j) \neq d^{\text{benign}}(i,j)\right), \quad (10)$$

where d^{adv} and d^{benign} are the depth maps under adversarial and benign textures, respectively. iii) System-Level Collision Rate ($\mathcal{E}_{\text{sys}}$), which evaluates downstream safety. We use a full Apollo perception and planning stack. We define the system-level metric as the proportion of evaluation episodes where the ego vehicle fails to avoid the adversarial object:

$$\mathcal{E}_{\text{sys}} = \frac{\text{\# of collisions}}{\text{\# of total runs}}. \quad (11)$$

Higher values of $\mathcal{E}_{\text{cover}}$ and $\mathcal{E}_{\text{sys}}$ indicate stronger adversarial impact, while lower $\mathcal{E}_{\text{blend}}$ reflects better concealment and a more effective merging attack.

Metric	Tex.	Time of Day		View Angle		Distance (m)	
		Midday	Sunset	Side	Front	12	24
$\mathcal{E}_{\text{blend}} \downarrow$	Benign	0.481	0.536	0.557	0.513	0.517	0.402
	Adv.	**0.087**	**0.067**	**0.071**	**0.085**	**0.074**	**0.095**
$\mathcal{E}_{\text{cover}} \uparrow$	Benign	0.036	0.042	0.030	0.038	0.035	0.032
	Adv.	**0.519**	**0.577**	**0.581**	**0.520**	**0.504**	**0.467**

Table 2: Real-world evaluation results under varying time of day, viewing angle, and distances.

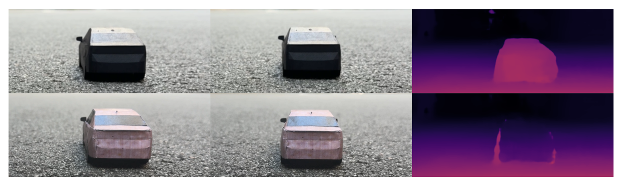

Figure 4: Visualization of our 3D PAEs in the real world. Top: Benign. Bottom: Adversarial.

For the *appearing attack* (Results in Table 5), we replace $\mathcal{E}_{\text{blend}}$ with the Mean Depth Shift ($\mathcal{E}_{\text{shift}}$), which measures how much the adversarial object appears closer to the camera compared to its benign counterpart. A higher value indicates a stronger effect caused by the adversarial texture:

$$\mathcal{E}_{\text{shift}} = \frac{1}{\sum_m} \cdot \sum_{(i,j)\in m} \left(d^{\text{adv}}(i,j) - d^{\text{benign}}(i,j)\right). \quad (12)$$

Compared Methods & Target SM Models. We compare our method with the two most relevant physical attacks designed against SM-BDE: PASM (Liu et al. 2024) and Adv-DM (Cheng et al. 2021). In addition, we include two baselines textures: *benign* (neutral textures) and *random* (random noise). We evaluate all textures on five representative SM-BDE models: PSMNet (Chang and Chen 2018), GA-Net (Zhang et al. 2019), RAFT-Stereo (Lipson, Teed, and Deng 2021), CREStereo (Li et al. 2022), and AnyStereo (Liang and Li 2024), covering diverse cost volume constructions and model designs.

Attack Effectiveness

Digital Evaluation. We evaluate the trained textures in the CARLA simulator (Dosovitskiy et al. 2017) using a world-aligned stereo setup. Adversarial vehicles are placed at distances ranging from 3 to 30 meters and orientations between

Figure 5: Our 3D texture-PAEs against SM-BDE under different viewpoint variations.

Method	Metric	Distance (m)			Heading Angle (°)											
		3–9	9–15	15–20	0	30	60	90	120	150	180	210	240	270	300	330
Benign		0.684	0.628	0.481	0.657	0.644	0.639	0.609	0.612	0.633	0.640	0.625	0.629	0.618	0.633	0.653
Random		0.704	0.681	0.489	0.684	0.667	0.654	0.656	0.641	0.648	0.659	0.654	0.651	0.640	0.649	0.659
PASM	$\mathcal{E}_{\text{blend}} \downarrow$	0.475	0.455	0.417	0.461	0.480	0.481	0.515	0.484	0.488	0.473	0.491	0.504	0.467	0.472	0.483
Adv-DM		0.510	0.461	0.423	0.516	0.534	0.541	0.578	0.536	0.524	0.520	0.542	0.554	0.585	0.547	0.536
Ours		**0.058**	**0.069**	**0.084**	**0.058**	**0.062**	**0.060**	**0.051**	**0.059**	**0.065**	**0.065**	**0.063**	**0.059**	**0.050**	**0.057**	**0.063**
Benign		0.037	0.029	0.024	0.028	0.025	0.029	0.015	0.017	0.027	0.030	0.025	0.020	0.017	0.021	0.026
Random		0.053	0.041	0.034	0.048	0.045	0.039	0.033	0.036	0.046	0.048	0.041	0.037	0.032	0.037	0.044
PASM	$\mathcal{E}_{\text{cover}} \uparrow$	0.183	0.165	0.093	0.132	0.120	0.111	0.094	0.107	0.114	0.144	0.121	0.105	0.099	0.112	0.126
Adv-DM		0.211	0.176	0.098	0.146	0.128	0.122	0.107	0.129	0.129	0.132	0.125	0.118	0.102	0.117	0.125
Ours		**0.585**	**0.532**	**0.468**	**0.496**	**0.533**	**0.548**	**0.595**	**0.543**	**0.536**	**0.501**	**0.523**	**0.544**	**0.562**	**0.547**	**0.528**

Table 3: Evaluation of attack robustness across various distances and a full range of heading angles. Our method consistently achieves higher $\mathcal{E}_{\text{cover}}$ and lower $\mathcal{E}_{\text{blend}}$ under varied viewpoints.

$0°$ and $360°$, under varying lighting and weather conditions in the "Town 10" map. Table 1 summarizes the evaluation results of our *merging attack* in comparison with state-of-the-art physical attacks across five representative SM-BDE models. All attacks are tested under identical camera baselines and environmental conditions in the CARLA simulator. Our adversarial texture consistently achieves a higher $\mathcal{E}_{\text{cover}}$, indicating a broader impact on predicted depth. This can be attributed to the larger effective surface coverage and better generalization across object orientations. Moreover, our approach attains the lowest $\mathcal{E}_{\text{blend}}$ among all methods, suggesting superior blending into the surrounding environment and reduced detectability in depth predictions, both key objectives of hiding attacks. Importantly, when integrated into a full commercial AD pipeline – Apollo, our attack leads to the highest $\mathcal{E}_{\text{sys}}$, indicating a significant increase in missed detection and path planning failures. This highlights the safety risk posed by our physically realizable texture.

Physical Evaluation. To assess real-world deployability, we 3D-print a 1:30 scale sedan and physically apply the optimized texture to its surface. A total of 300 stereo image pairs were collected across diverse conditions (e.g., viewing angles, distances, and environments) using two iPhone 14 Pro Max cameras, emulating the KITTI stereo setup. As shown in Table 2, our method maintains strong performance across all evaluation metrics under transformation and lighting variations. Notably, $\mathcal{E}_{\text{blend}}$ remains low, indicating that the adversarial vehicle seamlessly merges into the surrounding under physical conditions. As illustrated in Figure 4, while the benign-textured vehicle maintains clear geometric boundaries in depth maps, our adversarial texture causes the vehicle to disappear into the background. These results highlight the physical realizability and robustness of our method.

Attack Analysis

Attack Robustness. To evaluate the robustness of our attack under diverse real-world conditions, we conduct extensive tests across varying camera-object distances, viewing angles, and environmental factors such as weather and lighting. As shown in Figure 5, our adversarial textures consistently outperform prior methods across all settings, achieving higher affected region ratios $\mathcal{E}_{\text{cover}}$ and lower hiding errors $\mathcal{E}_{\text{blend}}$. These results are further validated by the quantitative metrics summarized in Table 3. While performance generally degrades at longer distances due to reduced visual resolution, our method remains notably more robust to angular variation. Unlike patch-based attacks that depend on frontal visibility, our full-surface adversarial texture maintains effectiveness from oblique viewpoints due to its geometry-aware design. In addition, our attack remains stable under challenging weather conditions such as rain and fog. Even in the worst case, it achieves a hiding error $\mathcal{E}_{\text{blend}}$ of 0.091 and an affected region ratio $\mathcal{E}_{\text{cover}}$ of 0.463, confirming its resilience in low-visibility scenarios. More settings and results are provided in the *supplementary material*.

Ablation Study. We conduct an ablation study to evaluate the effectiveness of the Stereo-Aligned 3D Rendering (SAR) module and the *merging attack* mechanism. We compare ad-

Metric	Losses			Modules			
	$\mathcal{L}_{\text{main}}$	$\mathcal{L}_{\text{nps}}$	$\mathcal{L}_{\text{tv}}$	*None*	*SAR*	*Merge*	*Full*
$\mathcal{E}_{\text{blend}} \downarrow$	✓			0.611	0.393	0.615	0.049
$\mathcal{E}_{\text{cover}} \uparrow$				0.017	0.556	0.029	0.607
$\mathcal{E}_{\text{blend}} \downarrow$	✓	✓		0.631	0.403	0.611	0.051
$\mathcal{E}_{\text{cover}} \uparrow$				0.015	0.541	0.024	0.587
$\mathcal{E}_{\text{blend}} \downarrow$	✓		✓	0.602	0.379	0.598	**0.042**
$\mathcal{E}_{\text{cover}} \uparrow$				0.020	0.582	0.031	**0.616**
$\mathcal{E}_{\text{blend}} \downarrow$	✓	✓	✓	0.685	0.411	0.665	0.058
$\mathcal{E}_{\text{cover}} \uparrow$				0.015	0.515	0.019	0.573

Table 4: Ablation results with different modules and losses.

Metric	PSMNet	GA-Net	RAFT	CRE	AnyStereo
$\mathcal{E}_{\text{cover}} \uparrow$	0.427	0.405	0.455	0.422	0.415
$\mathcal{E}_{\text{shift}} \uparrow$	16.48	15.87	15.64	15.57	15.10

Table 5: *Appearing attack* against different SM models.

versarial textures generated under four configurations: no module (*None*), *SAR* only, *merging* only, and the full method with both components. In the *merging only* setting, we define the main loss $\mathcal{L}_{\text{main}}$ as $\mathcal{L}_{\text{merge}}$ and reuse the same rendering parameters k from the left view for the right. For *None* and *SAR only*, we replace $\mathcal{L}_{\text{merge}}$ with a standard hiding loss (Zheng et al. 2024). As shown in Table 4, removing the physical constraint $\mathcal{L}_{\text{nps}}$ slightly improves performance due to the relaxation of printability limits. However, without enforcing stereo consistency, the texture has little effect on the model's predictions. In contrast, applying SAR significantly increases the affected region on the depth map (high $\mathcal{E}_{\text{cover}}$), while omitting the *merging* loss results in a clearly visible vehicle contour (low $\mathcal{E}_{\text{blend}}$).

System Evaluation. To assess the real-world impact of our attack, we conduct system-level experiments on the open-source autonomous driving platform Baidu Apollo. In each scenario, the adversarial vehicle is placed ahead of the victim AV to simulate realistic traffic. We select representative scenes and feed perception outputs, after decision-level fusion (Zhu et al. 2024), into Apollo's planning module under both benign and adversarial conditions. In the benign case, the AV detects the vehicle ahead, slows down, and initiates a lane change to avoid collision. In contrast, under attack, the adversarial texture disrupts scene understanding, causing the AV to maintain its speed and lane, ultimately resulting in a rear-end collision. As shown in Figure 7, the AV safely changes lane at a 10 m distance in the benign setting. However, in the adversarial case, it fails to respond and collides with the target in both tests. These results highlight the severe safety risks posed by stereo-consistent physical attacks.

Attack Transferability. We first evaluate cross-model transferability by testing whether adversarial textures trained

Figure 6: Depth map of appearing attack.

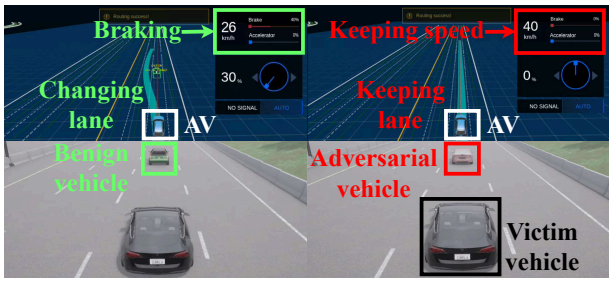

Figure 7: System-level effect of our attack.

Figure 8: Attack transferability on human model. Top: Benign. Bottom: Adversarial

on one stereo model can effectively attack others. Textures optimized on PSMNet transfer well to GANet ($\mathcal{E}_{\text{blend}} = 0.087$; $\mathcal{E}_{\text{cover}} = 0.524$), likely due to similar structures and disparity mechanisms. In contrast, transferability to models with iterative refinement paradigm like RAFT-Stereo ($\mathcal{E}_{\text{blend}} = 0.128$; $\mathcal{E}_{\text{cover}} = 0.451$) is relatively weaker. We further examine cross-category transferability by training adversarial textures on a human mesh. As shown in Figure 8, when an adversarial-textured human model is placed at the same depth as a benign vehicle in the simulation environment, the resulting depth map shows that the adversarial-textured human blends seamlessly into the background.

Conclusion

In this paper, we introduce the first texture-enabled physical adversarial attack against stereo matching-based binocular depth estimation in the context of autonomous driving. To precisely align with the disparity features utilized in stereo matching models, we develop a stereo-synchronized 3D rendering module that ensures view disparity remain consistent with the stereo cameras. This synchronization enables the generation of rendered objects that faithfully replicate real-world 3D transformations. We further propose a novel hiding strategy, termed merging attack, which can seamlessly blend the adversarial object into the background, enhancing stealth and reducing detectability. Extensive experiments at both the perception and system levels demonstrate that our attack generalizes across stereo matching models and environmental conditions, effectively compromising the full 3D structure of the target object.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG4-GC-2023-006-1B).

References

Abdelfattah, M.; Yuan, K.; Wang, Z. J.; and Ward, R. 2021. Towards Universal Physical Attacks on Cascaded Camera-Lidar 3D Object Detection Models. In *Proc. of the 28th IEEE ICIP*, 3592–3596. IEEE.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing Robust Adversarial Examples. In *International conference on machine learning*, 284–293. PMLR.

Baidu. 2025. Apollo: Open Autonomous Driving Platform. https://apollo.auto/. Accessed: 2025-03-02.

Cao, Y.; Wang, N.; Xiao, C.; Yang, D.; Fang, J.; Yang, R.; Chen, Q. A.; Liu, M.; and Li, B. 2021. Invisible for Both Camera and Lidar: Security of Multi-Sensor Fusion Based Perception in Autonomous Driving Under Physical-World Attacks. In *Proc. of the 42nd IEEE S&P*, 176–194. IEEE.

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid Stereo Matching Network. In *Proc. of the 31st IEEE/CVF CVPR*, 5410–5418.

Cheng, K.; et al. 2021. Towards Adversarially Robust and Domain Generalizable Stereo Matching by Rethinking DNN Feature Backbones. *arXiv preprint arXiv:2108.00335*.

Cheng, Z.; Liang, J.; Choi, H.; Tao, G.; Cao, Z.; Liu, D.; and Zhang, X. 2022. Physical Attack on Monocular Depth Estimation With Optimal Adversarial Patches. In *Proc. of the 17th ECCV*, 514–532. Springer.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. In *Proc. of the 31st IEEE/CVF CVPR*, 1625–1634.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. of the 25th IEEE/CVF CVPR*, 3354–3361.

Goodfellow, I. J.; et al. 2014. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*.

Guesmi, A.; Hanif, M. A.; Alouani, I.; and Shafique, M. 2023. Aparate: Adaptive Adversarial Patch for Cnn-Based Monocular Depth Estimation for Autonomous Navigation. *arXiv preprint arXiv:2303.01351*.

Kingma, D. P. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kurakin, A.; et al. 2018. Adversarial Examples in the Physical World. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; and Liu, S. 2022. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. In *Proc. of the 35th IEEE/CVF CVPR*, 9331–9340.

Li, P.; Chen, X.; and Shen, S. 2019. Stereo R-CNN Based 3D Object Detection for Autonomous Driving. In *Proc. of the 32nd IEEE/CVF CVPR*, 7644–7652.

Liang, Z.; and Li, C. 2024. Any-Stereo: Arbitrary Scale Disparity Estimation for Iterative Stereo Matching. In *Proc. of the 38th AAAI*, volume 38, 3333–3341.

Lipson, L.; Teed, Z.; and Deng, J. 2021. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In *International Conference on 3D Vision (3DV)*.

Liu, Y.; Zhai, J.; Ma, C.; Zeng, P.; Wang, X.; and Zhao, Y. 2024. Physical Attack for Stereo Matching. In *Proc. of the 1st ACM CVDL*, 1–5.

Lovisotto, G.; Turner, H.; Sluganovic, I.; Strohmeier, M.; and Martinovic, I. 2021. {Slap}: Improving Physical Adversarial Examples With {Short-Lived} Adversarial Perturbations. In *Proc. of the 30th USENIX*, 1865–1882.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083*.

Mahendran, A.; and Vedaldi, A. 2015. Understanding Deep Image Representations by Inverting Them. In *Proc. of the 28th IEEE/CVF CVPR*, 5188–5196.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *Proc. of the 29th IEEE/CVF CVPR*, 4040–4048.

Mobileye. 2025. Mobileye: Advanced Driver-Assistance Systems and Autonomous Driving Solutions. https://www.mobileye.com/. Accessed: 2025-05-11.

Ranjan, A.; Janai, J.; Geiger, A.; and Black, M. J. 2019. Attacking Optical Flow. In *Proc. of the 17th IEEE/CVF ICCV*, 2404–2413.

Schmalfuss, J.; Scholze, P.; and Bruhn, A. 2022. A Perturbation-Constrained Adversarial Attack for Evaluating the Robustness of Optical Flow. In *Proc. of the 17th ECCV*, 183–200. Springer.

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M. K. 2016. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proc. of the 23rd ACM CCS*, 1528–1540.

Thys, S.; et al. 2019. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. In *Proc. of the 18th IEEE/CVF CVPRW*, 0–0.

Tu, J.; Ren, M.; Manivasagam, S.; Liang, M.; Yang, B.; Du, R.; Cheng, F.; and Urtasun, R. 2020. Physically Realizable Adversarial Examples for Lidar Object Detection. In *Proc. of the 33rd IEEE/CVF CVPR*, 13716–13725.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2019. Pseudo-Lidar From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In *Proc. of the 32nd IEEE/CVF CVPR*, 8445–8453.

Waymo. 2025. Waymo: Autonomous Driving Technology. https://waymo.com/. Accessed: 2025-07-22.

Wong, A.; Mundhra, M.; and Soatto, S. 2021. Stereopagnosia: Fooling Stereo Networks With Adversarial Perturbations. In *Proc. of the 35th AAAI*, volume 35, 2879–2888.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *Proc. of the 16th IEEE/CVF ICCV*, 1369–1378.

Xu, H.; and Zhang, J. 2020. Aanet: Adaptive Aggregation Network for Efficient Stereo Matching. In *Proc. of the 33rd IEEE/CVF CVPR*, 1959–1968.

Zagoruyko, S.; and Komodakis, N. 2015. Learning to Compare Image Patches via Convolutional Neural Networks. In *Proc. of the 28th IEEE/CVF CVPR*, 4353–4361.

Zbontar, J.; and LeCun, Y. 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. In *Journal of Machine Learning Research*, volume 17, 1–32.

Zhang, F.; Prisacariu, V.; Yang, R.; and Torr, P. H. 2019. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching. In *Proc. of the 32nd IEEE/CVF CVPR*, 185–194.

Zheng, J.; Lin, C.; Sun, J.; Zhao, Z.; Li, Q.; and Shen, C. 2024. Physical 3D Adversarial Attacks Against Monocular Depth Estimation in Autonomous Driving. In *Proc. of the 37th IEEE/CVF CVPR*, 24452–24461.

Zhu, Y.; Miao, C.; Hajiaghajani, F.; Huai, M.; Su, L.; and Qiao, C. 2021a. Adversarial Attacks Against Lidar Semantic Segmentation in Autonomous Driving. In *Proc. of the 19th ACM SenSys*, 329–342.

Zhu, Y.; Miao, C.; Xue, H.; Yu, Y.; Su, L.; and Qiao, C. 2024. Malicious attacks against multi-sensor fusion in autonomous driving. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 436–451.

Zhu, Y.; Miao, C.; Zheng, T.; Hajiaghajani, F.; Su, L.; and Qiao, C. 2021b. Can We Use Arbitrary Objects to Attack Lidar Perception in Autonomous Driving? In *Proc. of the 28th ACM CCS*, 1945–1960.