Invisible Triggers, Visible Threats! Road-Style Adversarial Creation Attack for Visual 3D Detection in Autonomous Driving

Jian Wang, Lijun He, Yixing Yong, Haixia Bi, Fan Li*

School of Information and Communications Engineering, Xi'an Jiaotong University
wj851329121@stu.xjtu.edu.cn, lijunhe@mail.xjtu.edu.cn, yongyx@stu.xjtu.edu.cn, haixia.bi@mail.xjtu.edu.cn, lifan@mail.xjtu.edu.cn

Abstract

Modern autonomous driving (AD) systems leverage 3D object detection to perceive foreground objects in 3D environments for subsequent prediction and planning. Visual 3D detection based on RGB cameras provides a cost-effective solution compared to the LiDAR paradigm. While achieving promising detection accuracy, current deep neural network-based models remain highly susceptible to adversarial examples. The underlying safety concerns motivate us to investigate realistic adversarial attacks in AD scenarios. Previous work has demonstrated the feasibility of placing adversarial posters on the road surface to induce hallucinations in the detector. However, the *unnatural appearance* of the posters makes them easily noticeable by humans, and their *fixed content* can be readily targeted and defended. To address these limitations, we propose the AdvRoad to generate diverse road-style adversarial posters. The adversaries have naturalistic appearances resembling the road surface while compromising the detector to perceive non-existent objects at the attack locations. We employ a two-stage approach, termed Road-Style Adversary Generation and Scenario-Associated Adaptation, to maximize the attack effectiveness on the input scene while ensuring the natural appearance of the poster, allowing the attack to be carried out stealthily without drawing human attention. Extensive experiments show that AdvRoad generalizes well to different detectors, scenes, and spoofing locations. Moreover, physical attacks further demonstrate the practical threats in real-world environments.

Code — https://github.com/WangJian981002/AdvRoad

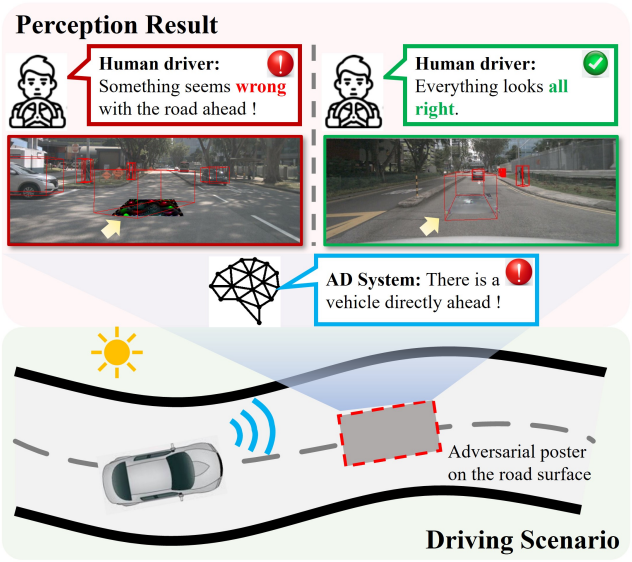


Figure 1: Illustration of the adversarial FP attacks on the road. The 3D detection system will perceive a ghost object near the poster. Compared with previous work (left), our poster (right) is harder to attract human attention, making it more likely to pose a real threat.

Introduction

Visual 3D object detection (Ma et al. 2024; Hu et al. 2023; Chen et al. 2017) has emerged as a pivotal technology in autonomous driving systems (Mao et al. 2023; Chen et al. 2024; Gulino et al. 2023), offering cost-effective environmental perception through widely accessible RGB cameras. Despite its computational efficiency and hardware affordability compared to LiDAR-dependent methods, the reliability of deep neural networks (DNNs) in safety-critical scenarios remains questionable due to their vulnerability to adversarial attacks (Yang et al. 2025; Lin et al. 2024; Han et al. 2023; Wang et al. 2023b). Recent studies have revealed that carefully designed inputs, such as additive perturbations (Zhang et al. 2024, 2021; Goodfellow, Shlens, and Szegedy 2014; Athalye et al. 2018) and local patches (Guesmi et al. 2024; Thys, Van Ranst, and Goedemé 2019; Hu et al. 2024, 2025), can catastrophically alter the output of DNNs, and these adversarial examples can be successfully implemented in physical scenarios (Sato et al. 2021). This security threat poses unpredictable consequences given the life-critical nature of 3D perception systems, which motivates us to investigate realistic adversarial attacks for 3D object detection in real-world environments.

Adversarial attacks on 3D object detectors can be divided into two types based on model errors: false negatives (FN) where real objects evade detection, and false positives (FP) where non-existent targets are identified. Most existing studies (Zhu et al. 2023; Abdelfattah et al. 2021; Cheng et al.

*Corresponding author.

Victim Model	Attack Way	Consequence
LiDAR-based	Malicious laser signal injection (Cao et al. 2023; Jin et al. 2023; Hau et al. 2021)	FN
	Malicious laser signal injection (Jin et al. 2023; Sun et al. 2020; Cao et al. 2019a; Wang et al. 2023a)	FP
	3D adversarial mesh (Tu et al. 2020; Cao et al. 2019b)	FN
Camera-based	Adversarial camouflage (Li, Lian, and Chen 2024)	FN
	Adversarial patch (Zhu et al. 2023; Wang, Li, and He 2025; Cheng et al. 2023; Xie et al. 2023)	FN
	Adversarial poster (Wang et al. 2025)	FP

Table 1: The summarization of physical adversarial attacks targeting 3D object detection in AD.

2023; Tu et al. 2020; Zhang, Hou, and Yuan 2024; Wang, Li, and He 2025) focus on FN attacks — for example, attaching adversarial patches to vehicles to make them invisible to detectors, which may result in a rear-end collision. However, implementing FN attacks typically requires physical access to the target object, limiting their practical application. FP attacks, on the other hand, aim to make detectors "see" imaginary obstacles, potentially triggering sudden braking or dangerous lane changes by autonomous vehicles. Despite posing similar safety risks as FN attacks, FP attacks for visual 3D detection remain poorly studied in the current attack literature.

Recently, Wang *et al.* (Wang et al. 2025) pioneered physical FP attack targeting visual 3D detectors by placing a carefully optimized poster on the road, thereby inducing the detector to perceive a ghost object near the poster (as shown in Fig. 1 left). Since the poster is 2D and lacks thickness, it is flexible to print and launch the attack. Moreover, the generated poster has strong generalization ability, allowing it to be effective in various scenarios. They learn the poster by proposing an image-3D applying algorithm that can differentiably render the 3D space poster onto the image, and directly optimizing the poster's pixel values. **However**, the following weaknesses remain: ❶ *The content of the poster significantly differs from the road surface, making it easily noticed by humans.* Directly optimizing poster pixels cannot constrain the learned content, which often has patterns with non-naturalistic appearances. ❷ *Each training session can only generate one poster, making it easy to be targeted and defended.* They generate a single adversarial poster that is effective across all scene images. Despite achieving a high attack success rate, the single poster can be easily exploited and defended against (e.g., by fine-tuning the model with data containing this adversary).

In response to limitations ❶ and ❷, we propose a novel FP attack pipeline that generates diverse road-style adversarial posters. All posters have patterns similar to the background road surface, making them harder to attract human attention, allowing the attack to proceed silently. Our framework employs a two-stage approach, Road-Style Adversary Generation and Scenario-Associated Adaptation, to generate diverse naturalistic adversaries. In the first stage, we utilize GAN-based techniques to train an adversarial generator that maps latent noise vectors to road-style adversaries. First, we use drones to take aerial photos of ground scenes and construct a road image collection for training the style discriminator. Then, we gradually update the generator by iteratively rendering the mapped poster onto the scene image and backpropagating the gradients based on the adversarial objective and style discriminator. The first stage ensures that the generated posters are similar to the source collection images while being able to compromise the detector. In the second stage, we derive a local-optimal poster tailored to a specific input, aiming to enhance its deceptive effectiveness within the given scenario. Concretely, we initialize a poster from the generator trained in the first stage, randomly sample and place it at various locations in the current scene, and then backpropagate the gradients to the latent space to optimize the noise vector toward a locally optimal solution. Note that in this stage, the generator is frozen. Therefore, the found adversary not only has naturalistic road styles but also achieves the strongest attack effectiveness in the current scene.

In short, our contributions are:

- We present a naturalistic FP attack pipeline for inducing the ghost object on the road. The crafted posters can significantly evade human perception while compromising the 3D models, thereby increasing the practical threat to the AD system.
- We introduce Road-Style Adversary Generation and Scenario-Associated Adaptation to maximize the attack capability of the adversarial poster. Moreover, all adversaries are effective across various scenes and at certain observation distances (e.g., $\leq 10m$).
- Extensive experiments in both the digital and physical worlds demonstrate the effectiveness of our approach with improved stealthiness. In addition, our posters are harder to defend against using existing defense techniques.

Related Work

3D object detection is the most crucial perception task in modular autonomous driving systems, which identifies and locates surrounding traffic participants like vehicles and pedestrians in 3D space. Errors in detection results gradually accumulate, thereby affecting subsequent predictions and planning.

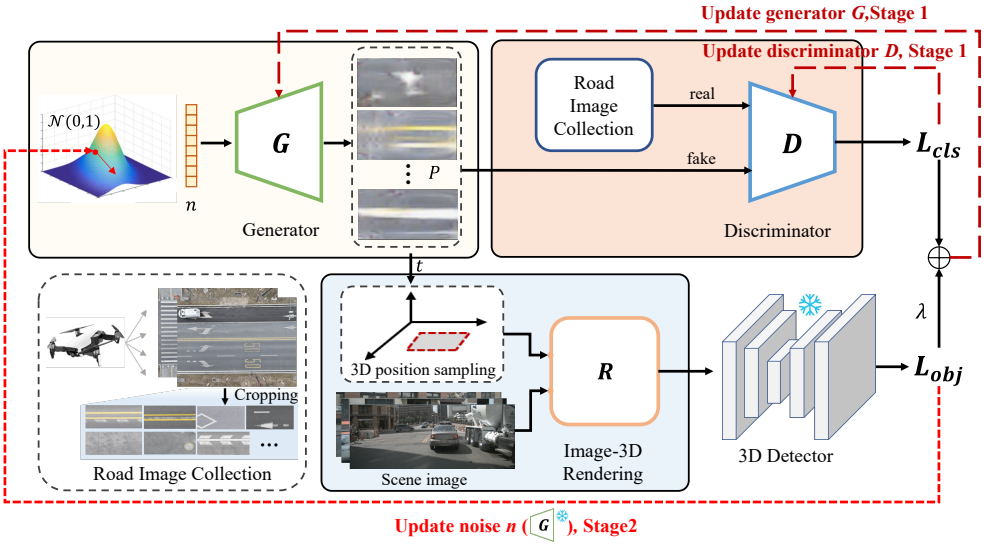


Figure 2: The AdvRoad framework. Stage 1 trains an adversarial generator that outputs universal road-style posters; Stage 2 updates the poster (the latent vector) to enhance the attack capability for the given scene.

Recent research efforts have developed various physical attack methods targeting LiDAR-based and camera-based detection algorithms. We provide a summary of these works in terms of physical attack ways and attack consequences in Table 1. For LiDAR-based approaches, attackers must alter the captured LiDAR point clouds. This can be achieved by strategically emitting laser pulses toward the target LiDAR sensor (Cao et al. 2019a; Jin et al. 2023; Cao et al. 2023) or placing 3D adversarial objects in the environment (Tu et al. 2020; Cao et al. 2019b). However, altering LiDAR data typically requires complex equipment like laser diodes, photodiodes, or industrial-grade 3D printers. This limits the attack's flexibility in dynamically changing AD scenarios. For camera-based approaches, attackers can utilize more affordable adversarial patches to perturb the captured images, leading to various detection errors. Extending Wang *et al.*'s framework (Wang et al. 2025) of using road surface posters to create fake 3D objects, we further study how to hide these patterns from human drivers. The spoofing capability to detectors and stealthiness to humans make the attack more dangerous in real driving scenarios.

Problem Definition

We investigate adversarial FP attacks against visual 3D object detection models by placing the learned poster on the road surface. Specifically, attackers apply the adversarial poster to a benign image x by firstly sampling the poster location in 3D space and then rendering the poster onto the image. This results in an adversarial input $\mathcal{R}(x, \delta, t)$, where $\mathcal{R}(\cdot)$ is the rendering function that applies the adversary δ to the input x according to the sampled transformation parameter t. Our goal is to induce the 3D detector, denoted as F_θ, to produce a desired response y^* at the poster position, which is formally defined by maximizing the likelihood of the response y^* under the adversarial input:

$$\max_{\delta} \log p(y^*|\mathcal{R}(x, \delta, t)) \tag{1}$$

where p is the probability function; δ is the adversary to be optimized, which can either be an explicit representation of the poster (e.g., a pixel array $P \in [0, 255]^{3\times H\times W}$) or an implicit representation (e.g., a generator $G : n \in \mathbb{R}^d \rightarrow P \in [0, 255]^{3\times H\times W}, n \sim \mathcal{N}(0, 1)$).

Continuous Attack Goals. When driving, the observation of the scene changes continuously, such as varying distances and viewing angles. The practical threat posed by a road-surface poster is substantially reduced if its effectiveness is confined to a specific scenario or location. Because the AD system is likely to classify targets appearing in merely one or two frames as sensor noise. Consequently, these posters must exhibit universality to remain effective across varying scenarios and viewing distances.

To this end, we leverage the Expectation of Transformation (EoT) (Athalye et al. 2018) across all training samples and a wide range of spoofing locations to enhance the physical robustness and generality of the attack. Formally, the optimization objective becomes:

$$\delta = \arg\max \sum_{x\in X} \sum_{t\in T} \log p(y^*|\mathcal{R}(x, \delta, t)) \tag{2}$$

where X comprises all possible image inputs and T is the set of position transformation parameters.

Proposed AdvRoad Framework

Road-Style Adversary Generation

Fig. 2 presents the framework of our road-style creation attack. The first stage aims to learn an adversarial generator that outputs diverse road-style posters while containing critical foreground features, enabling them to induce FP predictions in the detector. We employ a standard GAN pipeline to

integrate road surface style and spoofing information to the generator.

Road Image Collection. We utilize a DJI drone to capture aerial photography of traffic scenes, obtaining a series of raw images containing diverse road surfaces. We then crop authentic road patches from the collected aerial images, with each patch dimension approximating vehicle size (2m×4m). The built collection comprises over 2,000 road surface images covering various road patterns and styles (as shown in Fig. 2 left). This collection serves as the real reference for training the style discriminator, while the synthetic counterparts are generated by the generator. The inclusion of authentic imagery facilitates learning the realistic road patterns from the generator, thereby enhancing the visual indistinguishability of adversarial outputs to human drivers.

Adversarial Objective. The adversarial generator G is trained on the whole detection dataset X. Given a frame of input image $x \in X$, we first sample the poster locations in the 3D space then render the poster $G(n), n \in \mathbb{R}^d$ to the x through differentiable Image-3D rendering (discussed later). Then, we prepare the adversarial label y^* for the input $\mathcal{R}(x, \delta, t)$. We mask out the regions of real objects on the input image using ground truth (GT) bounding box annotations, and replace the GT labels with spoofing bounding boxes introduced by the posters. This approach provides dual advantages: First, masking surrounding objects prevents the gradient being vanishing by averaging. Second, we can reuse the detector's native loss function to directly calculate the adversarial loss. Formally, the adversarial loss is:

$$L_{obj} = J(F_\theta(\mathcal{R}(x, G(n), t)), y^*) \tag{3}$$

where $J(\cdot)$ is the original loss function for detector F_θ. The other training objective comes from style discriminator D. We freeze the D and maximize the discriminator's confidence in classifying the generator's outputs as real. The final training objective for the generator is:

$$L_G = L_{cls} + \lambda \cdot L_{obj} \tag{4}$$

We alternately train the generator and the discriminator, gradually injecting spoofing and style information into the generated posters.

Scenario-Associated Adaptation

After the first stage training, the posters output by the generator both resemble the road surface and maintain a certain degree of deception. They can serve as a universal trap to trigger FP predictions in the 3D detector. However, the generated posters rely on sampling noise from the latent space—a process that involves inherent randomness and may lead to unstable attack effects. Therefore, we further introduce Scenario-Associated Adaptation to enhance the attack capability of the posters in specific scenarios.

Specifically, given the input scenario x, we first *randomly initialize* the noise vector n from the Gaussian distribution $\mathcal{N}(0, 1)$ and fed n into the *frozen* generator to get the spoofing poster. Then, we randomly sample the poster locations in the current scene and perform rendering. Finally, we back-propagate the gradient to the latent space based on the adversarial loss L_{obj} and update vector n. This process is formulated as:

$$\begin{aligned} n_0 &= n \\ n_{i+1} &= n_i - \alpha \cdot \nabla_{n_i} J(F_\theta(\mathcal{R}(x, G(n_i), t)), y^*) \end{aligned} \tag{5}$$

We repeat this process until reach the maximum iteration number. To preserve realism, we ensure that the updated noise n_{i+1} falls into the hypersphere of radius η centered at initial noise n_0 in each iteration. The final poster exhibits strongest deceptive capability in the current scene while being authentic.

Image-3D Rendering

Conventional 2D patch rendering (Thys, Van Ranst, and Goedemé 2019; Lee and Kolter 2019; Hu et al. 2021) solely performs scaling and rotating the patch according to the objects' 2D bounding boxes. Although convenient, the physical size of the patch and its position in 3D space are not considered, which are critical for realistic physical attacks on 3D detectors. Following (Wang et al. 2025; Zhu et al. 2023), we briefly introduce how to render the poster from the 3D road surface onto the image.

First, we sample the 3D spatial positions of posters to place them on the road surface. Within a sector spanning $\pm\Delta_\theta$ relative to the vehicle's heading direction and a distance range of d_{min} to d_{max} meters, we randomly sample placement locations while avoiding overlapping with existing scene objects. Second, we project the four poster corner points onto the image plane using the camera's intrinsic and extrinsic matrices, thereby determining the adversarial region in the image (the quadrangle region defined by projected corner points). Third, for each pixel within this region, we inversely calculate its 3D coordinates aided by road height information (approximated from the bottom face height of the nearest scene object to the poster). Finally, each pixel's RGB value is calculated via bilinear interpolation based on its 3D position relative to the poster. For more details refer to supplementary-A.

Experiment

Experimental Setup

Victim Model. The vision-based BEV space 3D detector BEVDet (Huang et al. 2021), BEVDet4D (Huang and Huang 2022), and BEVFormer (Li et al. 2024) are selected as victim models for the attack, considering their representativeness and the fact that BEV space inherently supports most downstream perception tasks for AD (Hu et al. 2023). For each detector, the ResNet50 (He et al. 2016) and SwinTransformer-Tiny (Liu et al. 2021) are used as image backbone respectively.

Dataset. For the digital attack, we use nuScenes dataset (Caesar et al. 2020) to train the adversary and perform the attack. nuScenes is a large-scale, multi-modal dataset specifically designed for AD and 3D object detection. The training and validation set contains 28,130 and 6,019 frames respectively, with each frame including image data from six cameras and 360° 3D object annotations. We train all detectors

(a) AdvRoad

(b) AdvPoster (c) Real Picture

Figure 3: Visualizations of attack results in the digital domain. We place the spoofing poster on the road surface to launch the attack. (1) AdvRoad, our road-style naturalistic adversarial poster; (2) AdvPoster (Wang et al. 2025), generated by directly optimizing the pixel values; (3) Real Picture, use images of real vehicles as posters.

on the training set following their official settings and detection performances are given in supplementary-B.1. The confidence scores for the detected objects are over 0.1 for all models following (Wang et al. 2025; Tu et al. 2020).

Evaluation Metric. The attack success rate (ASR) is used to evaluate the creation attack, which measures the proportion of successfully detected fake objects among all spoofing attempts. We consider a fake object is successfully detected when the minimum distance between the detector's predictions and the center of the poster is less than d_{thr}. Multiple center distance thresholds, $\{2.0m, 1.5m, 1.0m, 0.5m\}$, are adopted for comprehensive evaluation. Unlike using the IoU as an indicator, precise alignment of the detected bounding box with the y^* in terms of size and orientation is less critical. Since the AD system may lead to dangerous consequences as long as a fake obstacle is perceived near the poster.

Moreover, we use the Learned Perceptual Image Patch Similarity (LPIPS) score (Zhang et al. 2018) to assess the environmental consistency of the attack, which measures perceptual similarity between benign and attacked images.

Implementation Details. We set the category of the spoofing object to the most common *vehicle*, with the poster's physical size being $2m \times 4m$. For spoofing locations, we aim to induce FP predictions either in front of or behind the self-vehicle. Therefore, the posters are placed at a distance of 7 to 10 meters from the self-vehicle with $\Delta_\theta = 5°$. Within this range, the AD system's misjudgment leaves little time for the driver to react. We sample 1,000 validation frames and place posters at two locations per frame, yielding 2,000 attacks for ASR computation. See the supplementary materials for more training details.

Main Result

We verify the effectiveness of our road-style adversarial poster (AdvRoad) in comparison with Benign, Random, and Real picture. Specifically, *Benign:* original scene without the adversarial pattern, however, we still sample and record the attack locations (with a fixed random seed) considering the miscalculation cases, where the detection results for the real objects are incorrectly counted to spoofed ones. *Random:* randomly initialize a poster for the attack. *Real picture:* use images of real vehicles as posters for the attack (Fig. 3(c)).

Model	Attack	LPIPS ↓	ASR (%, ↑)			
			2.0m	1.5m	1.0m	0.5m
BEVDet -R50	Benign	-	1.5	0.3	0.1	0
	Random	0.2136	8.0	4.9	2.9	0.8
	Real picture	0.2066	30.4	22.8	15.7	6.3
	AdvRoad	0.1472	62.6	55.6	42.7	23.3
BEVDet -SwinT	Benign	-	1.2	0.2	0	0
	Random	0.2138	1.7	0.7	0.4	0.1
	Real picture	0.2066	25.7	19.0	12.1	4.9
	AdvRoad	0.1337	60.2	56.3	47.6	28.8
BEVDet4D -R50	Benign	-	1.2	0.1	0	0
	Random	0.2137	3.4	1.9	1.2	0.3
	Real picture	0.2066	45.1	38.1	29.0	14.9
	AdvRoad	0.1331	49.1	42.9	32.7	17.7
BEVDet4D -SwinT	Benign	-	1.2	0.3	0	0
	Random	0.2136	1.4	0.3	0	0
	Real picture	0.2066	23.4	19.3	13.7	7.3
	AdvRoad	0.1370	39.1	35.1	27.8	15.7
BEVFormer -R50	Benign	-	1.4	0.3	0	0
	Random	0.1304	1.4	0.3	0	0
	Real picture	0.1260	6.3	3.5	1.8	0.7
	AdvRoad	0.0822	44.5	32.7	20.7	8.2
BEVFormer -SwinT	Benign	-	1.5	0.3	0	0
	Random	0.1306	1.5	0.5	0.1	0
	Real picture	0.1260	20.9	16.6	10.4	4.3
	AdvRoad	0.0818	37.3	30.6	21.0	8.9

Table 2: Digital attack results of adversarial creation attack in the nuScenes dataset.

Table 2 shows the digital attack results in nuScenes dataset. We achieve good attack performance across six different 3D detectors, successfully inducing FP predictions near the poster locations (see Fig. 3(a) for the visualization results). This holds regardless of whether the target model uses a CNN-based or Transformer-based backbone, a geometry-based or network-based PV-to-BEV transformation, or an anchor point-based or query-based detection head. For AD systems, such an error rate (exceeding 40%) is catastrophic. A suddenly appearing object within 10 meters in front of the ego-vehicle can trigger emergency braking or lane-changing decisions, potentially leading to severe safety incidents. Moreover, since our posters resemble the road surface, drivers have limited time to react and effectively intervene.

Since we avoid overlapping with scene objects when sampling attack locations, miscalculation cases are nearly negligible, e.g., the ASR for Benign consistently $< 1.5\%$ under $CD_{2.0m}$. Moreover, we observe an interesting result that taking a real vehicle image as the spoofing poster may also lead to FP errors. Specifically, Real picture can achieve an

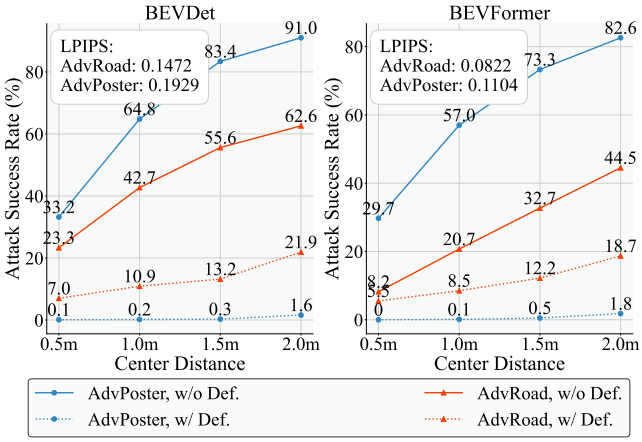


Figure 4: Comparison with AdvPoster $w/$ and w/o defense. All victim models use ResNet50 as image backbone.

ASR of up to 45.1% under $CD_{2.0m}$ for BEVDet4D-R50. Although the picture poster is 2D and lacks thickness, it can still provide intrinsic foreground visual cues that the detector captures and recognizes. However, similar to a real object, a 'flat' vehicle is also likely to attract the driver's attention and has higher LPIPS scores compared with AdvRoad.

Comparison with Current Work

We compare our attack with AdvPoster (Wang et al. 2025) in terms of attack performance, naturality, and defense difficulty, which explicitly learns the adversary by directly optimizing the pixel values of the poster.

Attack Performance. The results are given in Fig. 4 with aligned experimental settings (e.g., poster's physical size, attack distances). The solid lines represent the original attack performance without defense. As shown, AdvPoster achieves superior attack performance compared to our method, attaining ASR of 91% and 82.6% under $CD_{2.0m}$ on BEVDet and BEVFormer respectively. This demonstrates the effectiveness of directly optimizing the explicit representation of adversaries.

Naturality. AdvPoster optimizes a single deceptive poster across entire input scenes. While demonstrating high attack efficacy, the learned patterns exhibit uncontrollable and often abstract characteristics. As shown in Fig. 3(b), AdvPoster appears visually distinct from road surfaces and is very attention-grabbing. In contrast, our AdvRoad injects style information into the generator through implicit adversarial representations, producing natural-looking posters that blend seamlessly into road textures and achieving a lower LPIPS score. This enables stealthy attacks while amplifying practical security threats. More quantitative visual comparisons are provided in supplementary-B.2.

Defense Difficulty. For AdvPoster, single training generates only one unique adversarial example with salient discriminative features, it is more likely to be targeted and defended by the AD system. Therefore, we employ adversarial augmentation as the defense. Specifically, we incorporate the learned posters into the training set and fine-tune the detector for extra 2 epochs. The attack results after defense are shown in Fig. 4 (dash lines). We observe that AdvPoster only achieves less than 2% ASR against the defended detectors. It is easy for the models to filter out these adversarial patterns inside the image. However, AdvRoad still achieves $\sim 20\%$ ASR under $CD_{2.0m}$, which demonstrates the strong resilience of our attack when facing the defense. This is because AdvRoad can generate a large number of diverse adversaries, and these posters exhibit textures similar to the road surface, thus further increasing the difficulty of defense. More defense results are given in supplementary-B.3.

	Stage		ASR ($\%, \uparrow$)			
	1th	2nd	2.0m	1.5m	1.0m	0.5m
AdvRoad@1	✓		23.4	19.2	13.1	6.8
AdvRoad@2		✓	26.7	21.9	16.3	7.3
AdvRoad	✓	✓	62.6	55.6	42.7	23.3
AdvRoad*	✓	✓	67.0	60.5	48.6	27.2

Table 3: Ablations for Road-Style Adversary Generation and Scenario-Associated Adaptation. The results (ASR-$\%$) are given on BEVDet-R50.

Physical Size	LPIPS ↓	ASR ($\%, \uparrow$)			
		2.0m	1.5m	1.0m	0.5m
$1.5m \times 3.0m$	0.1123	31.3	26.5	19.9	10.4
$2.0m \times 3.0m$	0.1292	49.4	42.1	32.5	17.7
$2.0m \times 3.5m$	0.1384	56.6	50.4	38.6	21.2
$2.0m \times 4.0m$	0.1472	62.6	55.6	42.7	23.3
$2.0m \times 4.5m$	0.1555	66.1	58.1	44.0	23.5
$2.0m \times 5.0m$	0.1633	67.6	59.3	45.0	23.1

Table 4: Ablations for the physical size of road posters. The results are given on BEVDet-R50.

Ablation

Two Stage Approach. We conduct an ablation study to validate the effectiveness of each component in the AdvRoad. The results are shown in Table 3. We first briefly introduce the ablation settings. AdvRoad@1 directly uses the outputs from Stage 1 for the attack; AdvRoad@2 follows the typical GAN paradigm by first training a road poster generator *without the supervision of the adversarial objective* L_{obj}. Then, we freeze the generator using Scenario-Associated Adaptation to search the latent vector to perform the attack; AdvRoad* extends the update iteration to 50 in Stage 2. AdvRoad@1 achieves 23.4% ASR after the Stage 1 training. This reflects the attack performance when randomly selecting a poster from the generator and placing it within the 7–10m range. Additionally, even with a naturally trained generator without injecting adversarial information, searching the latent space can still discover deceptive content, achieving an ASR of 26.7%. Combining the Road-Style Adversary Generation and Scenario-Associated Adaptation, AdvRoad boosts the ASR to 62.6%, which highlights the effectiveness of our two-stage attack. Also, increasing the

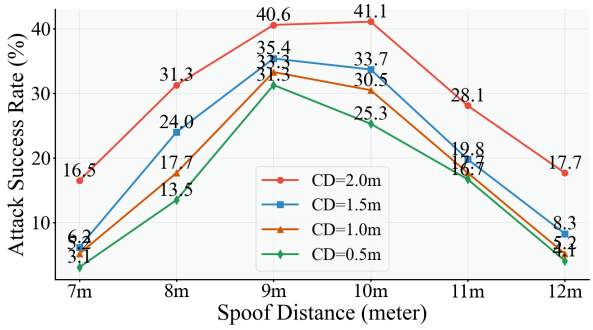


Figure 5: Attack results on the KITTI dataset. The ASRs (%) at different spoofing distances are given.

update iterations in Stage 2 can further improve the ASR.

Physical Size. Theoretically, the more pixels an attacker can manipulate in an image, the stronger attack capability becomes. However, since posters placed on road surfaces undergo perspective projection when captured by cameras, increasing the physical size of posters yields diminishing marginal returns in pixel gains for the adversarial region. On the other hand, since AdvRoad performs an instance-level attack that aims to induce a ghost vehicle near the poster center, excessively large posters would actually hinder models from achieving precise localization. Our most lenient evaluation metric (Center distance $\leq 2m$) exactly matches the edge position of a 4m-long poster. As shown in Table 4, increasing the physical size can improve the ASR. However, when the poster length exceeds 4m, the incremental gains become limited. Specifically, a 5m-long poster shows a decrease in ASR under the strict evaluation criterion of $CD_{0.5m}$. However, all the posters have textures similar to the road surface, making them difficult for humans to detect.

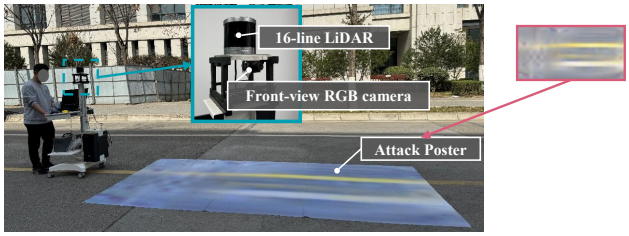
Attack to Broader Dataset

To verify the generalization ability of AdvRoad across different datasets, we further conduct attack experiments on KITTI scenes (Geiger, Lenz, and Urtasun 2012). We use BEVDet as the victim model and the ASRs at different spoofing distances are shown in Fig. 5. We observe that the poster achieves the strongest attack effectiveness at distances between 9 and 10 meters. When the distance is too short, the poster may not be fully captured by the camera, while at longer distances, the adversarial region in the image becomes limited, leading to a decline in attack performance.

Further discussions on AdvRoad can be found in the supplementary-B.4.

Physical Attack Experiment

To assess the practicality of our AdvRoad, we conduct attack experiments in physical-world environments. Since visual 3D detectors are typically camera-dependent (e.g., the PV2BEV transformation is related to the camera's intrinsic parameters), physical experiments first require training a 3D detector adapted to the custom scene and camera. Therefore, we built a physical detection platform with a front-view RGB camera and a 16-line LiDAR (Fig. 6(a)). The LiDAR



(a) Physical experiment setup

(b) Sunlit area (c) Shaded area (d) Poster wrinkled

Figure 6: Physical attack environment and results.

Condition	ASR
Sunlit area	49.4%(170/344)
Shaded area	28.3%(78/276)
Poster wrinkled	40.2%(103/256)
Partial occlusion	43.8%(92/210)
Indoor	19.5%(57/292)

Table 5: Quantitative attack results under different conditions.

sensor is only used to annotate the scene objects for training the custom 3D detector. Then, we train the generator, select a spoofing poster, and print it for the attack.

Fig. 6(b-d) illustrates some physical attack results. It can be seen that the poster can successfully induce the detector to generate false predictions at its location. Despite some color deviations in the printing process (we use cost-effective banner fabric for printing), the poster remains effective. This is because, during training, we apply random brightness and contrast adjustments, as well as inject random noise, to enhance the robustness of posters. The quantitative results under different physical conditions are given in Table 5. In addition, the texture similar to the road surface can further reduce the attention from human drivers. Consider using fabrics with better color fidelity, such as canvas, to reduce the color difference with the background road surface, thereby making the attack more covert. Further analysis of the physical attacks is provided in the supplementary-B.5.

Conclusion

This paper introduces AdvRoad, a naturalistic FP attack pipeline for visual 3D object detection in AD. AdvRoad leverages Road-Style Adversary Generation and Scenario-Associated Adaptation to perform stealthy adversarial attacks for human drivers, inducing 'ghost' objects for the perception system and potentially causing real-world threats such as emergency braking. Extensive experiments on both digital and physical domains demonstrate the effectiveness of AdvRoad. Compared with previous work, our attack is harder to detect and defend against, highlighting significant security risks to AD systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62531012, in part by the National Key Research and Development Program of China under Grant 2022YFA1003800, the Key Research and Development Program of Shaanxi Province under Grant 2025CY-YBXM-040, and in part by the XJTU Research Fund for Al Science under Grant 2025YXYC004.

References

Abdelfattah, M.; Yuan, K.; Wang, Z. J.; and Ward, R. 2021. Adversarial attacks on camera-lidar models for 3d car detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2189–2194. IEEE.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, 284–293. PMLR.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Cao, Y.; Bhupathiraju, S. H.; Naghavi, P.; Sugawara, T.; Mao, Z. M.; and Rampazzi, S. 2023. You can't see me: Physical removal attacks on {lidar-based} autonomous vehicles driving frameworks. In *USENIX Security Symposium*, 2993–3010.

Cao, Y.; Xiao, C.; Cyr, B.; Zhou, Y.; Park, W.; Rampazzi, S.; Chen, Q. A.; Fu, K.; and Mao, Z. M. 2019a. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2267–2281.

Cao, Y.; Xiao, C.; Yang, D.; Fang, J.; Yang, R.; Liu, M.; and Li, B. 2019b. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*.

Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1907–1915.

Cheng, Z.; Choi, H.; Liang, J.; Feng, S.; Tao, G.; Liu, D.; Zuzak, M.; and Zhang, X. 2023. Fusion is not enough: Single modal attacks on fusion models for 3D object detection. *arXiv preprint arXiv:2304.14614*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3354–3361. IEEE.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guesmi, A.; Ding, R.; Hanif, M. A.; Alouani, I.; and Shafique, M. 2024. Dap: A dynamic adversarial patch for evading person detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24595–24604.

Gulino, C.; Fu, J.; Luo, W.; Tucker, G.; Bronstein, E.; Lu, Y.; Harb, J.; Pan, X.; Wang, Y.; Chen, X.; et al. 2023. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36: 7730–7742.

Han, S.; Lin, C.; Shen, C.; Wang, Q.; and Guan, X. 2023. Interpreting adversarial examples in deep learning: A review. *ACM Computing Surveys*, 55(14s): 1–38.

Hau, Z.; Co, K. T.; Demetriou, S.; and Lupu, E. C. 2021. Object removal attacks on lidar-based 3d object detectors. *arXiv preprint arXiv:2102.03722*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, C.; Shi, W.; Yao, W.; Jiang, T.; Tian, L.; Chen, X.; and Li, W. 2024. Adversarial infrared curves: An attack on infrared pedestrian detectors in the physical world. *Neural Networks*, 178: 106459.

Hu, C.; Shi, W.; Yao, W.; Jiang, T.; Tian, L.; and Li, W. 2025. Two-stage optimized unified adversarial patch for attacking visible-infrared cross-modal detectors in the physical world. *Applied Soft Computing*, 112818.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

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.

Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Jin, Z.; Ji, X.; Cheng, Y.; Yang, B.; Yan, C.; and Xu, W. 2023. Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle. In *IEEE Symposium on Security and Privacy*, 1822–1839. IEEE.

Lee, M.; and Kolter, Z. 2019. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*.

Li, L.; Lian, Q.; and Chen, Y.-C. 2024. Adv3D: Generating 3D Adversarial Examples for 3D Object Detection in Driving Scenarios with NeRF. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 10813–10820.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2024. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, C.; Ji, X.; Yang, Y.; Li, Q.; Zhao, Z.; Peng, Z.; Wang, R.; Fang, L.; and Shen, C. 2024. Hard Adversarial Example Mining for Improving Robust Fairness. *IEEE Transactions on Information Forensics and Security*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Ma, Y.; Wang, T.; Bai, X.; Yang, H.; Hou, Y.; Wang, Y.; Qiao, Y.; Yang, R.; and Zhu, X. 2024. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8): 1909–1963.

Sato, T.; Shen, J.; Wang, N.; Jia, Y.; Lin, X.; and Chen, Q. A. 2021. Dirty road can attack: Security of deep learning based automated lane centering under Physical-World attack. In *USENIX security symposium*, 3309–3326.

Sun, J.; Cao, Y.; Chen, Q. A.; and Mao, Z. M. 2020. Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX Security Symposium*, 877–894.

Thys, S.; Van Ranst, W.; and Goedemé, T. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13716–13725.

Wang, J.; Li, F.; and He, L. 2025. A Unified Framework for Adversarial Patch Attacks against Visual 3D Object Detection in Autonomous Driving. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, J.; Li, F.; Lv, S.; He, L.; and Shen, C. 2025. Physically Realizable Adversarial Creating Attack against Vision-based BEV Space 3D Object Detection. *IEEE Transactions on Image Processing*.

Wang, J.; Li, F.; Zhang, X.; and Sun, H. 2023a. Adversarial obstacle generation against lidar-based 3d object detection. *IEEE Transactions on Multimedia*, 26: 2686–2699.

Wang, Z.; Yang, H.; Feng, Y.; Sun, P.; Guo, H.; Zhang, Z.; and Ren, K. 2023b. Towards transferable targeted adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20534–20543.

Xie, S.; Li, Z.; Wang, Z.; and Xie, C. 2023. On the adversarial robustness of camera-based 3d object detection. *arXiv preprint arXiv:2301.10766*.

Yang, B.; Zhang, H.; Wang, J.; Yang, Y.; Lin, C.; Shen, C.; and Zhao, Z. 2025. Adversarial Example Soups: Improving Transferability and Stealthiness for Free. *IEEE Transactions on Information Forensics and Security*.

Zhang, J.; Lou, Y.; Wang, J.; Wu, K.; Lu, K.; and Jia, X. 2021. Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. *IEEE Internet of Things Journal*, 9(5): 3443–3456.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, T.; Wang, L.; Zhang, X.; Zhang, Y.; Jia, B.; Liang, S.; Hu, S.; Fu, Q.; Liu, A.; and Liu, X. 2024. Visual Adversarial Attack on Vision-Language Models for Autonomous Driving. *arXiv preprint arXiv:2411.18275*.

Zhang, Y.; Hou, J.; and Yuan, Y. 2024. A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. *International Journal of Computer Vision*, 132(5): 1592–1624.

Zhu, Z.; Zhang, Y.; Chen, H.; Dong, Y.; Zhao, S.; Ding, W.; Zhong, J.; and Zheng, S. 2023. Understanding the robustness of 3D object detection with bird's-eye-view representations in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21600–21610.