ADV3D: GENERATING 3D ADVERSARIAL EXAMPLES FOR 3D OBJECT DETECTION IN DRIVING SCENARIOS WITH NERF

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks (DNNs) have been proven extremely susceptible to adversarial examples, which raises special safety-critical concerns for DNN-based autonomous driving stacks (*i.e.*, 3D object detection). Although there are extensive works on image-level attacks, most are restricted to 2D pixel spaces, and such attacks are not always physically realistic in our 3D world. Here we present Adv3D, the first exploration of modeling adversarial examples as Neural Radiance Fields (NeRFs) in driving scenarios. Advances in NeRF provide photorealistic appearances and 3D accurate generation, yielding a more realistic and realizable adversarial example. We train our adversarial NeRF by minimizing the surrounding objects' confidence predicted by 3D detectors on the training set. Then we evaluate Adv3D on the unseen validation set and show that it can cause a large performance reduction when rendering NeRF in any sampled pose. To enhance physical effectiveness, we propose primitive-aware sampling and semantic-guided regularization that enable 3D patch attacks with camouflage adversarial texture. Experimental results demonstrate that our method surpasses the mesh baseline and generalizes well to different poses, scenes, and 3D detectors. Finally, we provide a defense method to our attacks that improves both the robustness and clean performance of 3D detectors.

1 INTRODUCTION

The perception system of self-driving cars heavily rely on DNNs to process input data and comprehend the environment. Although DNNs have exhibited great improvements in performance, they have been found vulnerable to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015; Kurakin et al., 2017; Athalye et al., 2018). These adversarial examples crafted by adding imperceptible perturbations to input data, can lead DNNs to make wrong predictions. Motivated by the safety-critical nature of self-driving cars, we aim to explore the possibility of generating physically effective adversarial examples to disrupt 3D detectors in driving scenarios, and further improve the robustness of 3D detectors through adversarial training.

The 2D pixel perturbations (digital attacks) (Goodfellow et al., 2015; Szegedy et al., 2014) have been proven effective in attacking DNNs in various computer vision tasks (Xie et al., 2017; Xiang et al., 2019; Dong et al., 2020). However, these 2D pixel attacks are restricted to digital space and are difficult to realize in our 3D world. To address this challenge, several works have proposed physical attacks. For example, Athalye et al. (2018) propose the framework of Expectation Over Transformation (EOT) to improve the attack robustness over 3D transformation. Other researchers generate adversarial examples beyond image space through differentiable rendering, as seen in (Xiao et al., 2019; Zeng et al., 2019). These methods show great promise for advancing the field of 3D adversarial attacks and defense but are still limited in synthetic environments.

Given the safety-critical demand for self-driving cars, several works have proposed physically realizable attacks and defense methods in driving scenarios. For example, Cao et al. (2019; 2021) propose to learn 3D adversarial attacks capable of generating adversarial mesh to attack 3D detectors. However, their works only consider learning a 3D adversarial example for a few specific frames. Thus, the learned example is not universal and may not transfer to other scenes. To mitigate this problem,

Methods	Transferability	Adv. Type	Additional Requirements
Cao et al. (2019; 2021)	Poses	3D Mesh	Model, Annotation
Tu et al. (2020; 2021)	Poses, Scenes	3D Mesh	Model, Annotation
Xie et al. (2023)	Scenes, Categories	2D Patch	Model, Annotation
Adv3D	Poses, Scenes, Categories	3D NeRF	Model

Table 1: Comparison with prior works of adversarial attack in autonomous driving.

Tu et al. (2020; 2021) propose to learn a transferable adversary that is placed on top of a vehicle. Such an adversary can be used in any scene to hide the attacked object from 3D detectors. However, reproducing their attack in our physical world can be challenging since their adversary must have direct contact with the attacked object. We list detailed comparisons of prior works in Tab. 1.

To address the above challenges and generate 3D adversarial examples in driving scenarios, we build Adv3D upon recent advances in NeRF (Mildenhall et al., 2020) that provide both differentiable rendering and realistic synthesis. In order to generate physically effective attacks, we model Adv3D in a patch-attack (Sharma et al., 2022) manner and use an optimization-based approach that starts with a realistic NeRF object (Li et al., 2023) to learn its 3D adversarial texture. We optimize the adversarial texture to minimize the predicted confidence of all objects in the scenes, while keeping shape unchanged. During the evaluation, we render the input agnostic NeRF in randomly sampled poses, then we paste the rendered patch onto the unseen validation set to evaluate the attack performance. Owing to the transferability to poses and scenes, our adversarial examples can be executed without prior knowledge of the scene and do not need direct contact with the attacked objects, thus making for more feasible attacks compared with (Tu et al., 2020; 2021; Zhu et al., 2023; Xie et al., 2023). Finally, we provide thorough evaluations of Adv3D on camera-based 3D object detection with the nuScenes (Caesar et al., 2020) dataset. Our contributions are summarized as follows:

- We introduce Adv3D, the first exploration of formulating adversarial examples as NeRF to attack 3D detectors in autonomous driving. Adv3D provides photorealistic synthesis and demonstrates better attack performance than mesh-based adversarial examples.
- Incorporating the proposed primitive-aware sampling and semantic-guided regularization, Adv3D generates adversarial examples with enhanced physical realism and effectiveness.
- We conduct extensive real-world experiments to validate the transferability of our adversarial examples across unseen environments and detectors. Additionally, the analysis of these experiments provides valuable insights for developing more robust detectors.
- We show that by employing adversarial training with a trained adversarial NeRF, we can enhance the robustness and clean performance of 3D detectors.

2 Related Work

2.1 Adversarial Attack

DNNs are known to be vulnerable to adversarial attacks. Szegedy et al. (2014) first discovered that adversarial examples, generated by adding visually imperceptible perturbations to the original images, make DNNs predict a wrong category with high confidence. These vulnerabilities were also discovered in object detection and semantic segmentation (Liu et al., 2018; Xie et al., 2017). Moreover, DPatch (Liu et al., 2018) proposes transferable patch-based attacks by compositing a small patch to the input image. However, perturbing image pixels alone does not guarantee that adversarial examples can be created in the physical world. To address this issue, several works have performed physical attacks (Chen et al., 2019a; Xu et al., 2020; Brown et al., 2017; Komkov & Petiushko, 2021; Huang et al., 2020; Wu et al., 2020; Zhang et al., 2019; Wang et al., 2021a; Athalye et al., 2018) and exposed real-world threats. For example, Cheng et al. (2022) developed an adversarial patch with physical-oriented transformations to attack a depth estimation network. AdvPC (Hamdi et al., 2020b) investigate adversarial perturbations on 3D point clouds. SADA (Hamdi et al., 2020a) proposes semantic adversarial diagnostic attacks in various autonomous applications. ViewFool (Dong et al., 2022) and VIAT (Ruan et al., 2023) evaluate the robustness of DNNs to

adversarial viewpoints by using NeRF's differentiability. In our work, we mainly aim to generate 3D adversarial examples for 3D object detection in driving scenarios.

2.2 Robustness in Autonomous Driving

With the safety-critical nature, it is necessary to pay special attention to robustness in autonomous driving systems (Wang et al., 2021b). LiDAR-Adv (Cao et al., 2019) proposes to learn input-specific adversarial point clouds to fool LiDAR detectors. Tu et al. (2020) produces generalizable point clouds that can be placed on a vehicle roof to hide it. Furthermore, several work (Cao et al., 2021; Abdelfattah et al., 2021; Tu et al., 2021) try to attack a multi-sensor fusion system by optimizing 3D mesh through differentiable rendering. We compare our method with prior works in Tab. 1. Our method demonstrates stronger transferability and fewer requirements than prior works.

2.3 Image Synthesis using NeRF

NeRF (Mildenhall et al., 2020) enables photorealistic synthesis in a 3D-aware manner. Recent advances (Zhang et al., 2021) in NeRF allow for control over materials, illumination, and 6D pose of objects. Additionally, NeRF's rendering comes directly from real-world reconstruction, providing more physically accurate and photorealistic synthesis than previous mesh-based methods that relied on human handicrafts. Moreover, volumetric rendering (Kajiya & Von Herzen, 1984) enables NeRF to perform accurate and efficient gradient computation compared with dedicated renderers in mesh-based differentiable rendering (Kato et al., 2018; Chen et al., 2019b; Liu et al., 2019).

Recently, there has been tremendous progress in driving scene simulation using NeRF. Block-NeRF (Tancik et al., 2022) achieves city-scale reconstruction by modeling the blocks of cities with several isolated NeRFs to increase capacity. FEGR (Wang et al., 2023) learns to intrinsically decompose the driving scene for applications such as relighting. Lift3D (Li et al., 2023) use NeRF to generate new objects and augment them to driving datasets, demonstrating the capability of NeRF to improve downstream task performance. The driving scene simulation provides a perfect test bed to evaluate the effectiveness of self-driving cars.

3 Preliminary

3.1 Camera-based 3D Object Detection in Autonomous Driving

Camera-based 3D object detection is the fundamental task in autonomous driving. Without loss of generality, we focus on evaluating the robustness of camera-based 3D detectors.

The 3D detectors process image data and aim to predict 3D bounding boxes of all surrounding objects. The parameterization of a 3D bounding box can be written as $\mathbf{b} = \{\mathbf{R}, \mathbf{t}, \mathbf{s}, c\}$, where $\mathbf{R} \in SO(3)$ is the rotation of the box, $\mathbf{t} = (x, y, z)$ indicate translation of the box center, $\mathbf{s} = (l, w, h)$ represent the size (length, width, and height) of the box, and c is the confidence of the predicted box.

The network structure of camera-based 3D object detectors can be roughly categorized into FoV-based (front of view) and BEV-based (bird's eye view). FoV-based methods (Wang et al., 2021c;e;d) can be easily built by adding 3D attribute branches to 2D detectors. BEV-based methods (Philion & Fidler, 2020; Reading et al., 2021) typically convert 2D image feature to BEV feature using camera parameters, then directly detect objects on BEV planes. We refer readers to recent surveys (Ma et al., 2022; Li et al., 2022a) for more detail.

3.2 Differentiable Rendering using NeRF

Our method leverages the differentiable rendering scheme proposed by NeRF. NeRF parameterizes the volumetric density and color as a function of input coordinates. NeRF uses multi-layer perceptron (MLP) or hybrid neural representations (Fridovich-Keil et al., 2022; Müller et al., 2022) to represent this function. For each pixel on an image, a ray $\mathbf{r}(t) = \mathbf{r}_o + \mathbf{r}_d \cdot t$ is cast from the camera's origin $\mathbf{r}_o$ and passes through the direction of the pixel $\mathbf{r}_d$ at distance t. In a ray, we uniformly sample K points from the near plane t_{near} to the far plane t_{far}, the k^{th} distance is thus calculated as $t_k = t_{near} + (t_{far} - t_{near}) \cdot k/K$. For any queried point $\mathbf{r}(t_k)$ on the ray, the network takes its

position $\mathbf{r}(t_k)$ and predicts the per-point color $\mathbf{c}_k$ and density τ_k with:

$$(\mathbf{c}_k, \tau_k) = \text{Network}(\mathbf{r}(t_k)). \tag{1}$$

Note that we omit the direction term as suggested by (Gu et al., 2022). The final predicted color of each pixel $\mathbf{C}(\mathbf{r})$ is computed by approximating the volume rendering integral using numerical quadrature (Max, 1995):

$$\begin{aligned} \mathbf{C}(\mathbf{r}) &= \sum_{k=0}^{K-1} T_k \left(1 - \exp\left(-\tau_k \left(t_{k+1} - t_k\right)\right)\right) \mathbf{c}_k, \\ \text{with} \quad T_k &= \exp\left(-\sum_{k'<k} \tau_{k'} \left(t_{k'+1} - t_{k'}\right)\right). \end{aligned} \tag{2}$$

We build our NeRF upon Lift3D (Li et al., 2023). Lift3D is a 3D generation framework that generates photorealistic objects by fitting multi-view images synthesized by 2D generative modes (Karras et al., 2020) using NeRF. The network of Lift3D is a conditional NeRF with additional latent code input, which controls the shape and texture of the rendered object. The conditional NeRF in Lift3D is a tri-plane parameterized (Chan et al., 2022) generator. With its realistic generation and 3D controllability, Lift3D has demonstrated that the training data generated by NeRF can help to improve downstream task performance. To further explore and exploit the satisfactory property of NeRF, we present a valuable and important application in this work: we leverage the NeRF-generated data to investigate and improve the robustness of the perception system in self-driving cars.

4 METHOD

We illustrate the pipeline of Adv3D in Fig. 1. We aim to learn a transferable adversarial example in 3D detection that, when rendered in any pose (*i.e.*, location and rotation), can effectively hide surrounding objects from 3D detectors in any scenes by lowering their confidence. In Sec. 4.1, to improve the physical realizability of adversarial examples, we propose (1) Primitive-aware sampling to enable 3D patch attacks. (2) Disentangle NeRF that provides feasible geometry, and (3) Semantic-guided regularization that enables camouflage adversarial texture. To enhance the transferability across poses and scenes, we formulate the learning paradigm of Adv3D within the EOT framework (Athalye et al., 2018) in Sec. 4.3.

4.1 3D ADVERSARIAL EXAMPLE GENERATION

We use a gradient-based method to train our adversarial examples. The training pipeline involves 4 steps: **(i)** randomly sampling the pose of an adversarial example, **(ii)** rendering the example in the sampled pose, **(iii)** pasting the rendered patch into the original image of the training set, and finally, **(iv)** computing the loss and optimizing the latent codes. During inference, we discard the **(iv)** step.

4.1.1 POSE SAMPLING

To achieve adversarial attack in arbitrary object poses, we apply Expectation of Transformation (EOT) (Athalye et al., 2018) by randomly sampling object poses. The poses of adversarial examples are parameterized as 3D boxes $\mathbf{b}$ that are restricted to a predefined ground plane in front of the camera. We model the ground plane as a uniform distribution $\mathcal{B}$ in a specific range that is detailed in the supplement. During training, we independently sample the rendering poses of adversarial examples, and approximate the expectation by taking the average loss over the whole batch.

4.1.2 PRIMITIVE-AWARE SAMPLING

We model the primitive of adversarial examples as NeRF tightly bound by 3D boxes, in order to enable non-contact and physically realistic attacks. During volume rendering, we compute the intersection of rays $\mathbf{r}(t)$ with the sampled pose $\mathbf{b} = \{\mathbf{R}, \mathbf{t}, \mathbf{s}\} \in \mathcal{B}$, finding the first hit point and the last hit point of box (t_{near}, t_{far}) by the AABB-ray intersection algorithm (Majercik et al., 2018). We then sample our points inside the range (t_{near}, t_{far}) to reduce large unnecessary samples and avoid

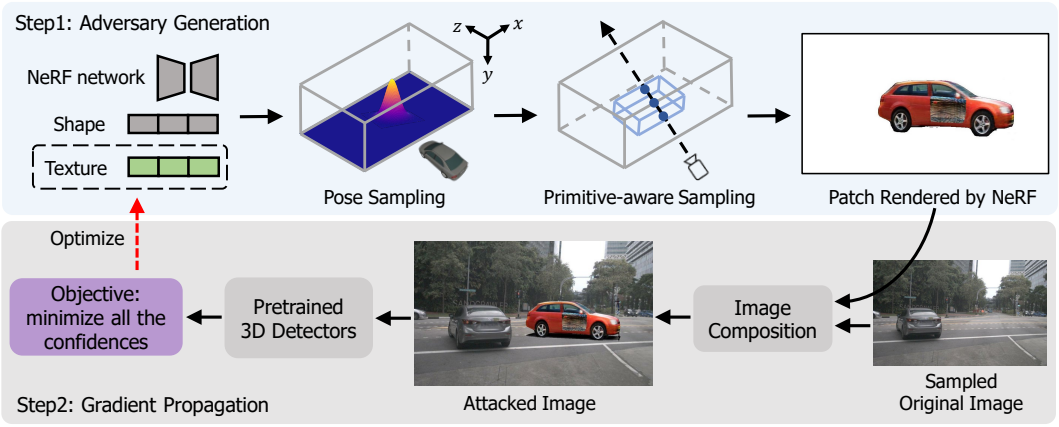

Figure 1: **Adv3D** aims to generate 3D adversarial examples that consistently perform attacks under different poses during rendering. We initialize adversarial examples from Lift3D (Li et al., 2023). During training, we optimize the texture latent codes of NeRF to minimize the detection confidence of all surrounding objects. During inference, we evaluate the performance reduction of pasting the adversarial patch rendered using randomly sampled poses on the validation set.

contact with the environment.

$$(t_{near}, t_{far}) = Intersect(\mathbf{r}, \mathbf{b}), \tag{3}$$

$$\mathbf{r}'(t_k) = \tilde{\mathbf{r}}(t_{near}) + (\tilde{\mathbf{r}}(t_{far}) - \tilde{\mathbf{r}}(t_{near})) \cdot k/K, \tag{4}$$

$$\tilde{\mathbf{r}}(t) = Transform(\mathbf{r}(t), \mathbf{b}), \tag{5}$$

where $\tilde{\mathbf{r}}(t)$ is the sampled points with additional global to local transformation. Specifically, we use a 3D affine transformation to map original sampled points $\mathbf{r}(t) = \mathbf{r}_o + \mathbf{r}_d \cdot t$ into a canonical space $\tilde{\mathbf{r}} = \{x, y, z\} \in [-1, 1]$. This ensures that all the sampled points regardless of their distance from the origin, are transformed to the range $[-1, 1]$, thus providing a compact input representation for NeRF network. The transformation is given by:

$$Transform(\mathbf{r}, \mathbf{b}) = \mathbf{s}^{-1} \cdot (\mathbf{R}^{-1} \cdot \mathbf{r} - \mathbf{t}), \tag{6}$$

where $\mathbf{b} = \{\mathbf{R}, \mathbf{t}, \mathbf{s}\}$, $\mathbf{R} \in SO(3)$ is rotation matrix of the box, $\mathbf{t}, \mathbf{s} \in \mathbb{R}^3$ indicate translation and scale vector that move and scale the unit cube to desired location and size. The parameters of $\mathbf{b}$ are sampled from a pre-defined distribution $\mathcal{B}$ detailed in the supplement.

Then, the points lied in $[-1, 1]$ are projected to exactly cover the tri-plane features $\mathbf{z}$ for interpolation. Finally, a small MLP takes the interpolated features as input and predicts RGB and density:

$$(\mathbf{c}_k, \tau_k) = \text{MLP}(Interpolate(\mathbf{z}, \mathbf{r}'(t_k))). \tag{7}$$

The primitive-aware sampling enables patch attacks (Sharma et al., 2022) in a 3D-aware manner by lifting the 2D patch to a 3D box, enhancing the physical realizability by ensuring that the adversarial example only has a small modification to the original 3D environment.

4.1.3 Disentangled NeRF Parameterization

The original parameterization of NeRF combines the shape and texture into a single MLP, resulting in an entangled shape and texture generation. Since shape variation is challenging to reproduce in the real world, we disentangle shape and texture generation and only set the texture as adversarial examples. We obtain texture latents $\mathbf{z}_{\mathbf{tex}.}$ and shape latents $\mathbf{z}_{\mathbf{shape}}$ from the Lift3D. During volume rendering, we disentangle shape and texture generation by separately predicting RGB and density:

$$\mathbf{c}_k = \text{Network}(\mathbf{z}_{\mathbf{tex}.}, \mathbf{r}'(t_k)), \quad \tau_k = \text{Network}(\mathbf{z}_{\mathbf{shape}}, \mathbf{r}'(t_k)), \tag{8}$$

where $\mathbf{z}_{\mathbf{shape}}$ is fixed and $\mathbf{z}_{\mathbf{texture}}$ is being optimized. Our disentangled parametrization can also be seen as a geometry regularization in (Tu et al., 2021; 2020) but keeps geometry unchanged as a usual vehicle, leading to a more realizable adversarial example.

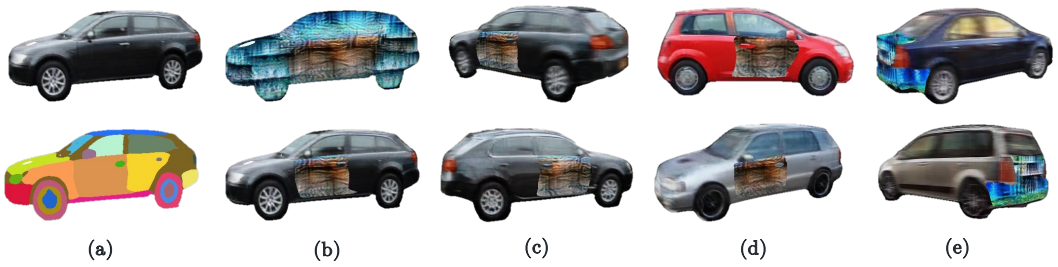

Figure 2: Rendered results of our adversarial examples. **(a)** Image and semantic label of an instance predicted by NeRF. **(b)** Top: our example without semantic-guided regularization. Bottom: our example with semantic-guided regularization. **(c)** Multi-view consistent synthesis of our examples. **(d,e)** The texture transfer results of side and back part adversary to other vehicles.

4.1.4 SEMANTIC-GUIDED REGULARIZATION

Setting the full part of the vehicle as adversarial textures is straightforward, but not always feasible in the real world. To improve the physical realizability, we propose to optimize individual semantic parts, such as doors and windows of a vehicle. Specifically, as shown in Fig. 2 (d, e)), we only set a specific part of the vehicle as adversarial texture while maintaining others unchanged. This semantic-guided regularization leads to a camouflage adversarial texture that is less likely spotted in the real world and improves physical effectiveness.

To achieve this, we add a semantic branch to Lift3D to predict semantic part labels of the sampled points. We re-train Lift3D by fitting multi-view images and semantic labels generated by EditGAN (Ling et al., 2021). Using semantic-guided regularization, we maintain the original texture and adversarial part texture at the same time but only optimize the adversarial part texture while leaving the original texture unchanged. This approach allows us to preserve a large majority of parts as usual, but to alter only the specific parts that are adversarial (see Fig. 2 (b, c)). In our implementation, we query the NeRF network twice, one for the adversarial texture and the other for the original texture. Then, we replace the part of original texture with the adversarial texture indexed by semantic labels in the point space.

Owing to this property, these adversarial textures can be printed and pasted on certain parts of vehicles to perform attacks. We provide real-world reproduction in the supplementary materials.

4.2 GRADIENT PROPAGATION

After rendering the adversarial examples, we paste the adversarial patch into the original image through image composition. The attacked image can be expressed as $I_1 \times M + I_2 \times (1 - M)$ where I_1 and I_2 are the patch and original image, M is foreground mask predicted by NeRF. Next, the attacked images are fed to pretrained and fixed 3D detectors to compute the objective and back-propagate the gradients. Since both the rendering and detection pipelines are differentiable, Adv3D allows gradients from the objective to flow into the texture latent codes during optimization.

4.3 LEARNING PARADIGM

We formulate our learning paradigm as EOT (Athalye et al., 2018) that finds adversarial texture codes by minimizing the expectation of a binary cross-entropy loss over sampled poses and scenes:

$$\mathbf{z_{tex.}} = \arg\min_{\mathbf{z_{tex.}}} \mathbb{E}_{\mathbf{b}\sim\mathcal{B}}\mathbb{E}_{\mathbf{x}\sim\mathcal{X}}[-\log(1 - P(I(\mathbf{x}, \mathbf{b}, \mathbf{z_{tex.}}))], \tag{9}$$

where $\mathbf{b}$ is the rendering pose sampled from the predefined distribution of ground plane $\mathcal{B}$, $\mathbf{x}$ is the original image sampled from the training set $\mathcal{X}$, $I(\mathbf{x}, \mathbf{b}, \mathbf{z_{tex.}})$ is the attacked image that composited by the original image $\mathbf{x}$ and the adversarial patch rendered using pose $\mathbf{b}$ and texture latent code $\mathbf{z_{tex.}}$, and $P(I(\cdot))$ represents the confidence of all proposals predicted by detectors. We approximate the expectation by averaging the objective of the independently sampled batch. The objective is a binary cross-entropy loss that minimizes the confidence of all predicted bounding boxes, including adversarial objects and normal objects.

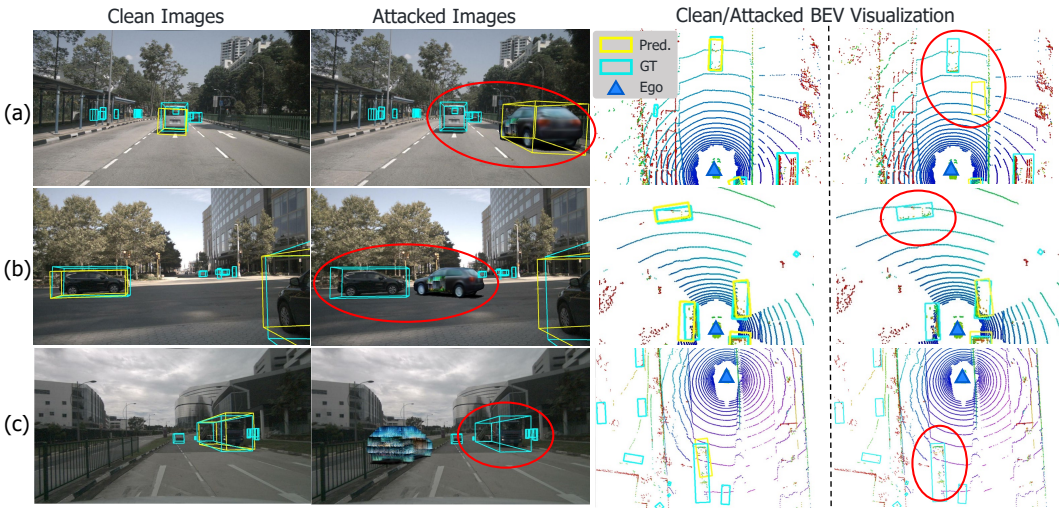

Figure 3: Visualization of BEVDet prediction on nuScenes validation set under our attacks. The visualization threshold is set at 0.6. The adversarial NeRF can hide surrounding objects by minimizing their predicted confidence in a non-contact manner (making the yellow boxes disappear).

Built within the framework of EOT, Adv3D helps to improve the transferability and robustness of adversarial examples over the sampling parameters (poses and scenes here). This means that the attack can be performed without prior knowledge of the scene and are able to disrupt models across different poses and times in a non-contact manner.

4.4 Adversarial Defense by Data Augmentation

Toward defenses against our adversarial attack, we also study adversarial training to improve the robustness of 3D detectors. Adversarial training is typically performed by adding image perturbations using a few PGD steps (Madry et al., 2017; Xie et al., 2020) during the training of networks. However, our adversarial example is too expensive to generate for the bi-level loop of the min-max optimization objective. Thus, instead of generating adversarial examples from scratch at every iteration, we directly leverage the transferable adversarial examples to augment the training set. We use the trained adversarial example to locally store a large number of rendered images to avoid repeated computation. During adversarial training, we randomly paste the rendered adversarial patch into the training images with a probability of 30%, while remaining others unchanged. We provide experimental results in Sec. 5.4.

5 Experiments

In this section, we first describe the training details of our adversarial attacks, and provide comparison with the mesh baseline in Sec. 5.1. Then we present the experiments of semantic-guided regularization in Sec. 5.2, the analysis of 3D attack in Sec. 5.3, and our adversarial defense method in Sec. 5.4. We evaluate the transferability across different detectors in the supplementary materials.

Models	Backbone	Type	Clean NDS	Adv NDS	Clean mAP	Adv mAP
FCOS3D (Wang et al., 2021c)	ResNet101	FoV	0.3770	0.2674	0.2980	0.1272
PGD-Det (Wang et al., 2021d)	ResNet101	FoV	0.3934	0.2694	0.3174	0.1321
DETR3D (Wang et al., 2021e)	ResNet101	FoV	0.4220	0.2755	0.3470	0.1336
BEVDet (Huang et al., 2021)	ResNet50	BEV	0.3822	0.2247	0.3076	0.1325
BEVFormer-Tiny (Li et al., 2022b)	ResNet50	BEV	0.3540	0.2264	0.2524	0.1217
BEVFormer-Base (Li et al., 2022b)	ResNet101	BEV	0.5176	0.3800	0.4167	0.2376

Table 2: Comparison of different detectors under our attack. Clean NDS and mAP denote evaluation using original validation data. Adv NDS and mAP denote evaluation using attacked data.

Semantic Part	NDS	mAP
Clean	0.382	0.307
No Part	0.302	0.234
Full Parts	0.224	0.132
Part of Front	0.267	0.148
Part of Side	0.265	0.149
Part of Rear	0.268	0.151

Table 3: Ablations of semantic parts.

Data	Adv train	NDS	mAP
Clean val	✗	0.304	0.248
Clean val	✓	**0.311**	**0.255**
Adv val †	✗	0.224	0.132
Adv val †	✓	**0.264**	**0.181**
Adv val §	✓	0.228	0.130

Table 4: Results of adversarial training.

Dataset We conduct our experiments on the nuScenes dataset (Caesar et al., 2020). This dataset is collected using 6 surrounded-view cameras that cover the full 360° field of view around the ego-vehicle. It contains 700 scenes for training and 150 scenes for validation. In our work, we train our adversarial examples on the training set and evaluate performance drop on the validation set.

Target Detectors and Metrics As shown in Tab. 2, we evaluate the robustness of six representative detectors. Three are FoV-based, and three are BEV-based. Following prior work (Xie et al., 2023), we evaluate the performance drop on the validation set after the attack. Specifically, we use the Mean Average Precision (mAP) and nuScenes Detection Score (NDS) (Caesar et al., 2020) to evaluate the performance of 3D detectors.

Quantitative Results We provide the experimental results of adversarial attacks in Tab. 2. The attacks are conducted in a full-part manner without semantic-guided regularization to investigate the upper limit of attack performance. We found that, in spite of FoV-based or BEV-based, they display similar robustness. Meanwhile, we see a huge improvement of robustness by utilizing a stronger backbone (ResNet101 versus ResNet50) when comparing BEVFormer-Base with BEVFormer-Tiny. We hope these results will inspire researchers to develop 3D detectors with enhanced robustness.

Rendering Results We visualize our attack results with semantic-guided regularization in Fig. 3 (a,b), and without regularization in Fig. 3 (c). The disappearance of detected objects is caused by their lower confidence scores. For example, the confidence predicted by detectors in Fig. 3 (a) have declined from 0.6 to 0.4, and are therefore filtered out by the threshold of 0.6. In Fig. 3 (a), we find that our adversarial NeRF is realistic enough to be detected by a 3D detector if it doesn't display much of the adversarial texture. However, once the vehicle shows a larger area of the adversarial texture as seen in Fig. 3 (b), it will hide all objects including itself due to our untargeted objective.

5.1 Compare with Mesh Attack

We compare our method with the mesh baseline, which uses a randomly picked ShapeNet car model (Chang et al., 2015) as an adversarial example. We use PyTorch3D's differentiable renderer (Ravi et al., 2020) and optimize the vertex color as an adversarial example to attack BEVDet. Similar to the setting of the NeRF counterpart, we randomly render the single mesh model and paste the patch onto the original images. In Tab. 5, we show that our method achieves better attack performance than the mesh baseline. This improvement can be attributed to the latent space of NeRF weights having a higher dimensional representation than vertex color and providing much more solutions for attacking, which results in a better attack performance.

Method	NDS	mAP
Clean	0.382	0.307
Mesh	0.301	0.218
NeRF	**0.264**	**0.189**

Table 5: Comparison with Mesh.

5.2 Semantic Parts Analysis

In Tab. 3, we provide experiments on the impact of different semantic parts on attack performance. Specifically, we focused on three salient parts of the vehicle: the front, side, and rear. Our results show that compared with adversarial attacks using full parts, the semantic-guided regularization leads to a slightly lower performance drop, but remains a realistic appearance and less likely spotted adversarial texture as illustrated in Fig. 2 (b).

Since we do not have access to annotation during training, we additionally conduct "No Part" experiment that no part of the texture is adversarial, to evaluate the impact of the collision and occlusion.

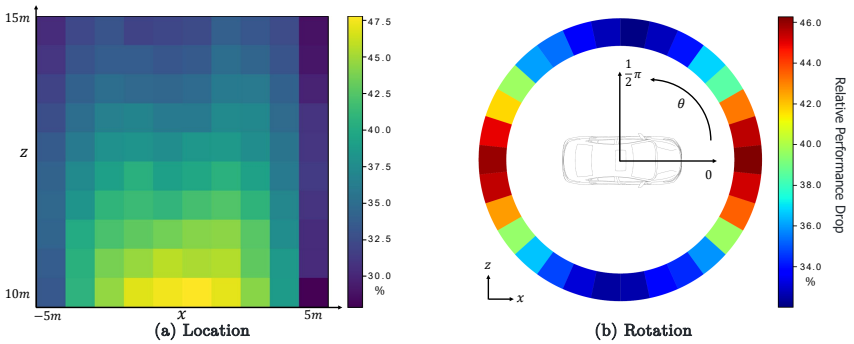

Figure 4: To examine the 3D-aware property of our adversarial examples, we ablate the relative performance drop by sampling adversarial examples within different bins of location and rotation.

We acknowledge that part of performance degradation can be attributed to the occlusion to original objects and the false positive prediction of adversarial objects (see Fig. 3 (a)), since we do not update the ground truth of adversarial objects to the validation set.

5.3 Effectiveness of 3D-aware attack

To validate the effectiveness of our 3D attacks, we ablate the impact of different poses on the attack performance. In Fig. 4 (a), we divide the BEV plane into 10×10 bins ranging from $x \in [-5m, 5m]$ and $z \in [10m, 15m]$. We then evaluate the relative mAP drop (percentage) of BEVDet (Huang et al., 2021) by sampling one adversarial example inside the bin per image, while keeping rotation randomly sampled. Similarly, we conduct experiments of 30 uniform rotation bins ranging from $[0, 2\pi]$ in Fig. 4 (b). The experimental results demonstrate that all aspects of location and rotation achieve a valid attack (performance drop $> 30\%$), thereby proving the transferability of poses in our 3D-aware attack.

A finding that contrasts with prior work (Tu et al., 2020) is the impact of near and far locations in z axis. Our adversarial example is more effective in the near region compared with the far region, while Tu et al. (2020) display a roughly uniform distribution in all regions. We hypothesize that the attack performance is proportional to the area of the rendered patch, which is highly related to the location of objects. Similar findings are also displayed in rotation. The vehicle that poses vertically to the ego vehicle results in a larger rendered area, thus better attack performance.

5.4 Adversarial Defense by Data Augmentation

We present the results of adversarial training in Tab. 4. The symbol † indicates attacks using the same adversarial example used in adversarial training, while § indicates a different example. We observe that incorporating adversarial training improves not only the robustness against the seen adversarial examples, but also the clean performance. However, we also note that our adversarial training is not capable of transferring to unseen adversarial examples trained in the same way, mainly due to the fixed adversarial example during adversarial training. Furthermore, we hope that future work can conduct in-depth investigations and consider handling the bi-level loop of adversarial training in order to better defend against adversarial attacks.

6 Conclusion

In this paper, we propose **Adv3D**, the first attempt to model adversarial examples as NeRF in driving scenarios. Adv3D enhances the physical realizability of attacks through our proposed primitive-aware sampling and semantic-guided regularization. Compared with prior works of adversarial examples in autonomous driving, our examples are more threatening in practice as we carry non-contact attacks, have feasible 3D shapes as usual vehicles, and display camouflage adversarial texture. Extensive experimental results also demonstrate that Adv3D achieves better attack performance and transfers well to different poses, scenes, and detectors. We hope our work provides valuable insights for creating more realistic evaluations to investigate and improve the robustness of autonomous driving systems.

REFERENCES

Mazen Abdelfattah, Kaiwen Yuan, Z Jane Wang, and Rabab Ward. Adversarial attacks on camera-lidar models for 3d car detection. In *IROS*, 2021.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. 2018.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*, 2019.

Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *IEEE Symposium on Security and Privacy (SP)*, 2021.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *ECML PKDD*, 2019a.

Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *NeurIPS*, 32, 2019b.

Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *ECCV*, 2022.

Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *CVPR*, 2020.

Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in Neural Information Processing Systems*, 35:36789–36803, 2022.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*, 2022.

Abdullah Hamdi, Matthias Müller, and Bernard Ghanem. Sada: semantic adversarial diagnostic attacks for autonomous applications. In *AAAI*, 2020a.

Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *ECCV*, 2020b.

Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *CVPR*, 2020.

James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3), 1984.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018.

Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *ICPR*. IEEE, 2021.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2017.

Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Enze Xie, Zhiqi Li, Hanming Deng, Hao Tian, Xizhou Zhu, Li Chen, Yulu Gao, Xiangwei Geng, Jia Zeng, Yang Li, Jiazhi Yang, Xiaosong Jia, Bohan Yu, Yu Qiao, Dahua Lin, Si Liu, Junchi Yan, Jianping Shi, and Ping Luo. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *arXiv preprint arXiv:2209.05324*, 2022a.

Leheng Li, Qing Lian, Luozhou Wang, Ningning Ma, and Ying-Cong Chen. Lift3d: Synthesize 3d training data by lifting 2d gan to 3d generative radiance field. In *CVPR*, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *ECCV*, 2022b.

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *NeurIPS*, 2021.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019.

Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018.

Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Y. Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Alexander Majercik, Cyril Crassin, Peter Shirley, and Morgan McGuire. A ray-box intersection algorithm and efficient dynamic voxel rendering. *Journal of Computer Graphics Techniques Vol*, 7(3), 2018.

Nelson Max. Optical models for direct volume rendering. *IEEE TVCG*, 1995.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM ToG*, 2022.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.

Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distribution-network for monocular 3d object detection. *CVPR*, 2021.

Shouwei Ruan, Yinpeng Dong, Hang Su, Jianteng Peng, Ning Chen, and Xingxing Wei. Improving viewpoint robustness for visual recognition via adversarial training. *arXiv preprint arXiv:2307.11528*, 2023.

Abhijith Sharma, Yijun Bian, Phil Munz, and Apurva Narayan. Adversarial patch attacks and defences in vision-based tasks: A survey. *arXiv preprint arXiv:2206.08304*, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022.

James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *CVPR*, 2020.

James Tu, Huichen Li, Xinchen Yan, Mengye Ren, Yun Chen, Ming Liang, Eilyan Bitar, Ersin Yumer, and Raquel Urtasun. Exploring adversarial robustness of multi-sensor perception systems in self driving. *arXiv preprint arXiv:2101.06784*, 2021.

Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, 2021a.

Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. Advsim: Generating safety-critical scenarios for self-driving vehicles. *CVPR*, 2021b.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *ICCV Workshop*, 2021c.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and Geometric Depth: Detecting objects in perspective. In *CoRL*, 2021d.

Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, 2021e.

Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *CVPR*, June 2023.

Tong Wu, Xuefei Ning, Wenshuo Li, Ranran Huang, Huazhong Yang, and Yu Wang. Physical adversarial attack on vehicle detector in the carla simulator. *arXiv preprint arXiv:2007.16118*, 2020.

Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *CVPR*, 2019.

Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *CVPR*, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *CVPR*, 2017.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *CVPR*, 2020.

Shaoyuan Xie, Zichao Li, Zeyu Wang, and Cihang Xie. On the adversarial robustness of camera-based 3d object detection. *arXiv preprint arXiv:2301.10766*, 2023.

Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *ECCV*, 2020.

Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *CVPR*, 2019.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ToG*, 2021.

Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019.

Zijian Zhu, Yichi Zhang, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Zhong, and Shibao Zheng. Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving. *arXiv preprint arXiv:2303.17297*, 2023.