3DOSA: UNIFIED RADIANCE FIELDS OF 3D OBJECTS AND SCENES FOR PHYSICALLY REALIZABLE ADVERSARIAL ATTACKS IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving multi-view consistency in physical adversarial attacks requires an accurate 3D representation of the target object that captures intricate surface geometry. Recently, 3D Gaussian Splatting (3DGS) provides a promising representation for adversarial optimization due to its explicit and differentiable modeling of 3D objects and scenes. However, existing 3DGS-based attacks typically optimize physical adversarial camouflages without any physical environmental context or rely on 2D masks to identify optimizable Gaussians in the 3D scene, which can cause gradients to leak beyond the intended perturbation boundaries. In addition, volumetric Gaussians are difficult to align precisely with object surfaces, which can produce floating artifacts and geometric distortions that make the resulting camouflage physically undeployable. To address these limitations, we propose "3DOSA", a general physical attack framework that unifies the 3D object and the scene within a single adversarial radiance field. 3DOSA introduces an explicit spatial decoupling mechanism to isolate target Gaussians from the surrounding scene and prevent gradient leakage into background and non-printable regions. We further incorporate 2D Gaussian Splatting (2DGS) to represent adversarial Gaussians as planar surfels that closely adhere to the object surface. This representation suppresses floating artifacts and preserves the geometric fidelity of the object to maintain the physical realizability of the adversarial camouflage. We primarily evaluate 3DOSA on stereo matching, which imposes cross-view consistency beyond the multi-view requirements of general physical attacks, and further evaluate its versatility to object detection. Extensive experiments in simulated and physical environments demonstrate strong attack performance and robustness across diverse viewpoints and models.

1 INTRODUCTION

The Perception module in Autonomous Driving Systems (ADS) processes data from onboard sensors and converts it into context-aware information for safe navigation. However, the backbone of the perception tasks is fundamentally driven by Deep Neural Networks (DNNs) and Transformers, which are inherently susceptible to adversarial attacks. To achieve real-world deception, general physical attacks have evolved beyond 2D perturbations, shifting toward modeling 3D objects to improve multi-view consistency. For perception models that strictly rely on rigid geometric constraints like stereo matching, 3D spatial modeling becomes an indispensable prerequisite.

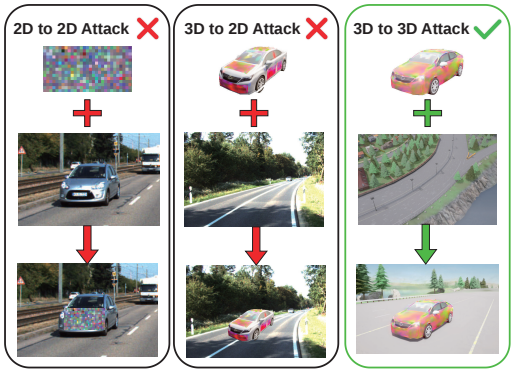


Figure 1: Dimensional gaps in attacks.

Currently, the most conventional approach to achieve 3D mesh-based attack is by optimizing "adversarial camouflage" via 3D differential rendering Huang et al. (2020); Wang et al. (2021; 2022); Suryanto et al. (2022; 2023). By bridging 3D mesh geometries with 2D camera views through dif-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ferentiable rasterization, these methods optimize 2D texture maps across multiple synthesized 3D transformations of the mesh in the image. However, these methods suffer from two critical limitations. First, although the target object is modeled in 3D, it is still pasted onto a static 2D planar background and physically decoupled from the surrounding scene. This leads to unnatural poses, distorted grounding, and unrealistic boundary occlusions across multi-view observations. Second, the optimized texture colors are static without considering any dynamic environmental lighting and view-dependent reflections in real-world scenes.

Recent studies have explored the use of Neural Radiance Fields (NeRF) to capture the 3D optical geometry of the target object Huang et al. (2024b); Lou et al. (2025); Zhang et al. (2025; 2024). 3DGS provides an option to improve physical rendering realism in adversarial optimization since it can simultaneously provide an accurate 3D spatial representation of the scene. Nevertheless, applying 3DGS to physical adversarial attacks introduces two fundamental challenges. First, it remains difficult to accurately retrieve and isolate the target Gaussians within a continuous radiance-based 3D scene representation. Existing 3DGS-based attack methods rely on 2D masks or background-removal assumptions to delineate target boundaries. Therefore, the object remains isolated from the surrounding scene context, leaving the cross-dimensional gap still unresolved. Without explicit target assignment, backpropagated adversarial gradients can contaminate ambient Gaussians near perturbation boundaries. Second, the volumetric Gaussian primitives in standard 3DGS cannot strictly adhere to the physical surface of the target object, resulting in spatial misalignment during adversarial optimization. Together, these limitations can produce geometric distortions, floating artifacts, boundary bleeding, and unnatural color distortions, rendering the final camouflage physically undeployable. As shown in Figure 2 (a), camouflage generated without explicit Gaussian decoupling and surface-constrained representation exhibits blurry boundary bleeding and unnatural color distortions. In contrast, camouflage generated via our method is clear and edge-cutting (Figure 2 (b)).

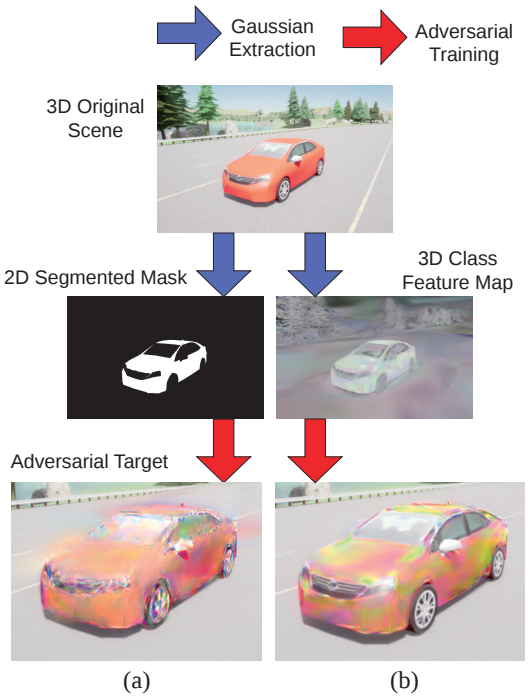


Figure 2: Rendering results of the 3D adversarial object generated using different Gaussian extraction methods in a CARLA simulated 3D scene.

To address these limitations, we propose 3DOSA, a general adversarial attack framework that integrates both the 3D object and the scene within a single adversarial radiance field. Specifically, we utilize 3D Gaussian Grouping Ye et al. (2024) to establish entity-level assignments directly within the 3D spatial domain, thereby isolating non-target and environmental Gaussians from backpropagated adversarial gradients. Simultaneously, we leverage 2DGS Huang et al. (2024a) to collapse volumetric ellipsoids into flat, disk-like surfels. This encourages the Gaussians to closely adhere to the object surface and eliminates spurious floating artifacts around surface boundaries, thereby preserving the physical fidelity of the object. This unified formulation achieves fully cross-dimensional continuity and enables multi-view consistent, physically realizable attacks against diverse autonomous driving perception tasks. Extensive experiments are conducted across both digital and physical setups, covering challenging 3D geometry-dependent perception tasks such as stereo matching, as well as conventional tasks such as object detection. The results demonstrate that 3DOSA successfully disrupts perception across different scenarios and task models, highlighting its attack effectiveness and versatility. To conclude, the main contributions of this paper are summarized as follows:

- We introduce 3DOSA, the first fully cross-dimensional, task-agnostic physical adversarial attack framework for autonomous driving. Unlike conventional 2D patches and 3D mesh-based attacks rendered onto detached 2D backgrounds, our framework unifies the 3D adversarial object and the continuous 3D scene representation within a single radiance

field. This provides a unified foundation for evaluating the robustness of diverse perception models under physically realizable attacks.

- We propose a novel explicit 3D Gaussian decoupling mechanism and a differentiable 2D surfel optimization objective to achieve view-consistent and physically realizable attacks. This formulation ensures strict gradient isolation for the target Gaussians while maintaining the physical fidelity of the object.
- We conduct extensive evaluations across diverse stereo matching models, complemented by object detection experiments in simulated environments and real-world physical evaluations. Adversarial camouflages generated by 3DOSA consistently demonstrate strong attack performance and multi-view robustness compared with baseline attacks.

2 Related Work

2.1 Physical Adversarial Attack

Early attacks designed 2D patches to mislead traffic sign recognition Eykholt et al. (2018); Sitawarin et al. (2018), pedestrian detection Thys et al. (2019); Xu et al. (2020); Hu et al. (2022), 2D/3D vehicle bounding box detectors Chen et al. (2019); Zhang et al. (2019b); Tu et al. (2020); Cao et al. (2021), and monocular depth estimation Wong et al. (2021); Cheng et al. (2022). In contrast, physical attacks on stereo matching remain relatively underexplored despite rapid advances from cost-volume aggregation Chang & Chen (2018); Zhang et al. (2019a) and recurrent refinement Lipson et al. (2021); Wang et al. (2023) to recent methods that exploit monocular foundation-model priors Wen et al. (2025); Bartolomei et al. (2025). Existing physical attacks on stereo matching mainly rely on planar patches based on surrogate gradient approximations Cheng et al. (2021) or rigid disparity shifts using Expectation over Transformation (EoT) Athalye et al. (2018); Liu et al. (2024). However, these 2D patch-based formulations fail to preserve multi-view spatial and geometric consistency under dynamic camera poses. To overcome this limitation, researchers transitioned to optimizing 3D mesh-based adversarial camouflages via 3D differentiable rendering pipelines to improve multi-view consistency Huang et al. (2020); Wang et al. (2021; 2022); Suryanto et al. (2022); Zhou et al. (2023). Nevertheless, these methods suffer from cross-dimensional fragmentation by rendering 3D surfaces onto detached 2D backgrounds without complex lighting interactions between the object and the scene. Subsequent work leveraged NeRF Huang et al. (2024b) to bypass CAD mesh dependencies. Preliminary explorations adopted 3DGS for physical camouflage generation. Yet, these methods heavily rely on 2D masks Lou et al. (2025) or idealized background assumptions to identify targets without explicit 3D spatial decoupling Zhang et al. (2024; 2025); Meng et al. (2025).

2.2 3D Radiance Fields and Gaussian Splatting

Neural Radiance Fields (NeRF) Mildenhall et al. (2021) represent 3D scenes with implicit neural functions and synthesize novel views through volume rendering. Despite subsequent improvements in training efficiency Müller et al. (2022) and scalability Tancik et al. (2022), their computationally intensive ray sampling remains a bottleneck for real-time rendering. In contrast, 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) adopts explicit Gaussian primitives and differentiable rasterization for efficient, high-quality rendering. Gaussian Grouping Ye et al. (2024) further assigns identity embeddings to individual Gaussians for joint scene reconstruction and entity-level segmentation. 2D Gaussian Splatting (2DGS) Huang et al. (2024a) replaces volumetric ellipsoids with oriented planar surfels and introduces ray-splat intersection constraints to improve multi-view geometric consistency and surface reconstruction.

3 Methodology

3.1 Problem Definition

3DGS explicitly represents a 3D physical scene using a collection of N parameterized 3D Gaussian primitives $\mathcal{G} = \{g_i\}_{i=1}^{N}$. Each Gaussian $g_i = \{\mu_i, \mathbf{s}_i, q_i, \alpha_i, \mathbf{c}_i\}$ is defined by its spatial center position $\mu_i \in \mathbb{R}^3$, 3D scaling vector $\mathbf{s}_i \in \mathbb{R}^3$, rotation quaternion $q_i \in \mathbb{R}^4$, opacity $\alpha_i \in [0, 1]$, and spherical harmonics (SH) coefficients $\mathbf{c}_i$ encoding view-dependent radiance color. Given a camera

viewpoint parameterized by an extrinsic pose matrix $\mathbf{T}$ and an intrinsic calibration matrix K, the 3D Gaussians are projected onto the image plane, and the synthesized color at pixel p is rendered via differentiable point-based α-blending:

$$I(p) = \sum_{i \in \mathcal{N}} c_i(\mathbf{T}) \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{1}$$

where $\mathcal{N}$ denotes the depth-ordered sequence of Gaussians overlapping pixel p, and $c_i(\mathbf{T})$ is the directional color computed from the SH coefficients $\mathbf{c}_i$ under the given camera viewpoint. We denote the complete forward rendering across all pixels compactly as $I = \mathcal{R}(\mathcal{G}; \mathbf{T}, K)$.

A stereo matching model predicts dense disparity from a rectified stereo pair captured under stereo camera poses $\mathbf{T} = (\mathbf{T}_l, \mathbf{T}_r)$ and intrinsics K, where $\mathbf{T}_l$ and $\mathbf{T}_r$ maintain a fixed relative pose. In the 3D space, we decouple the physical scene into an immutable background $\mathcal{G}_{\text{bg}}$ and the target object $\mathcal{G}_{\text{obj}}(\theta)$, where $\theta = \{\mathbf{c}_i \mid g_i \in \mathcal{G}_{\text{obj}}\}$ denotes the optimizable SH coefficients of the target-object Gaussians, while their geometric and opacity parameters remain fixed. The synchronous stereo image pair (I_l, I_r) is differentiably rendered from this unified 3D representation:

$$(I_l, I_r) = \mathcal{R}\big(\mathcal{G}_{\text{bg}} \cup \mathcal{G}_{\text{obj}}(\theta); \mathbf{T}, K\big), \tag{2}$$

from which a stereo matching model F predicts the disparity map $d = F(I_l, I_r)$. Under the Expectation over Transformation (EoT) across camera viewpoints $\mathbf{T} \sim \mathcal{P}$, the primary objective of our attack is to optimize θ such that the predicted disparity matches an adversarial target disparity d_t:

$$\theta^* = \arg\min_{\theta \in \Omega} \mathbb{E}_{\mathbf{T} \sim \mathcal{P}} \Big[\mathcal{L}\big(F(I_l, I_r), d_t\big) \Big], \tag{3}$$

where Ω denotes the target optimization domain, and $\mathcal{L}$ represents the main adversarial loss. This formulation imposes two key requirements for physically realizable adversarial optimization. First, the target optimization domain must be accurately identified and isolated from the surrounding scene Gaussians, such that adversarial gradients are strictly restricted to the target object. Second, the optimized Gaussian representation should faithfully adhere to the physical surface of the target object to preserve its geometric fidelity during adversarial optimization. Furthermore, to demonstrate the generalizability of our framework across different downstream tasks, this formulation can be readily extended to object classification/detection models F_{det} by optimizing θ to suppress the detection of the target object via $\mathcal{L}_{\text{det}}\big(F_{\text{det}}(I), y\big)$, where y represents the true bounding box and class label.

3.2 3D Geometry-Preserving Representation

3.2.1 3D Explicit Gaussian Decoupling

The first challenge arises from the absence of explicit primitive boundaries in standard radiance fields. When generating physical adversarial camouflage, perturbation gradients must be strictly confined to the printable surfaces of the target while leaving other areas unaltered. To eliminate this spatial ambiguity, we introduce an explicit 3D primitive-level decoupling mechanism that partitions the Gaussians within the scene into two mutually exclusive classes $k \in \{0, 1\}$. $k = 1$ denotes the optimizable subset consisting solely of the printable surfaces of the target object, and $k = 0$ denotes a frozen complement encompassing the surroundings of the target as well as non-printable surface components such as windows, lights, and wheels of a vehicle.

Specifically, during digital-twin scene reconstruction, pre-filtered multi-view 2D masks of the optimizable region are incorporated into the Gaussian Grouping framework Ye et al. (2024) to supervise a 3D identity feature field. Each Gaussian primitive is augmented with a learnable identity feature vector $\mathbf{f}_i \in \mathbb{R}^{16}$ in addition to its existing geometric and photometric attributes, yielding an extended representation $g_i = \{\mu_i, \mathbf{s}_i, q_i, \alpha_i, \mathbf{c}_i, \mathbf{f}_i\}$. A linear classification head then projects this feature vector into binary classification logits $\mathbf{z}_i = W\mathbf{f}_i + b \in \mathbb{R}^2$. For our adversarial optimization, the identity labels are formulated into two classes $k \in \{0, 1\}$, where $k = 1$ denotes the optimizable target and $k = 0$ denotes the frozen complement. Accordingly, the scene primitive set $\mathcal{S}$ is deterministically partitioned into the optimizable target subset $\mathcal{S}_{\text{target}}$ and the frozen complement $\mathcal{S}_{\text{frozen}}$:

$$\mathcal{S}_{\text{target}} = \left\{ g_i \in \mathcal{S} \;\middle|\; \arg\max_{k \in \{0,1\}} (\mathbf{z}_i)_k = 1 \right\}, \quad \mathcal{S}_{\text{frozen}} = \mathcal{S} \setminus \mathcal{S}_{\text{target}} \tag{4}$$

The $\mathcal{S}_{\text{frozen}}$ initially retains all ambient background Gaussians across the entire reconstructed scene. Propagating and rendering redundant distant Gaussians during subsequent adversarial optimization imposes unnecessary computational overhead and introduces irrelevant dependencies on distant scene content. To retain the locally relevant scene context while preserving the unified object-scene representation, we compute the spatial centroid of the target object $\bar{\mu} = \frac{1}{|\mathcal{S}_{\text{target}}|} \sum_{g_i \in \mathcal{S}_{\text{target}}} \mu_i$ and prune distant Gaussians residing outside a local Euclidean sphere $\mathcal{B}(\bar{\mu}, R) = \{x \in \mathbb{R}^3 \mid \|x - \bar{\mu}\|_2 \leq R\}$ with cutoff radius R. Consequently, the frozen set is refined as $\mathcal{S}'_{\text{frozen}} = \{g_i \in \mathcal{S}_{\text{frozen}} \mid \mu_i \in \mathcal{B}(\bar{\mu}, R)\}$, discarding distant background Gaussians from the computational graph while preserving nearby frozen Gaussians for accurate local occlusion modeling.

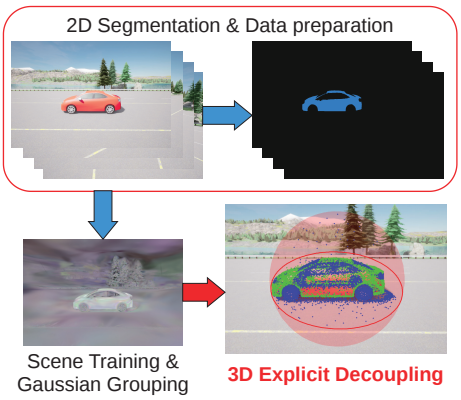


Figure 3: 3D Explicit Decoupling via Gaussian Grouping. (Green: Target, Blue: Frozen, Red: Cutoff Sphere)

3.2.2 Surfel-based adversarial optimization

The second challenge lies in accurately representing the target object for physical adversarial optimization. Standard 3DGS provides an effective representation of the overall 3D scene. However, physical adversarial attacks require substantially higher representation accuracy for the target object in the scene, as the optimized camouflage must closely conform to its physical surface for real-world deployment. Off-surface Gaussian primitives can introduce floating artifacts and geometric inconsistencies during optimization, compromising the physical fidelity of the resulting camouflage. To achieve more precise surface-constrained adversarial optimization, we build the scene and represent the target $\mathcal{S}_{\text{target}}$ using 2DGS Huang et al. (2024a).

Each surfel g_i is parameterized by a unit normal $\mathbf{n}_i \in \mathbb{R}^3$ and orthogonal in-plane tangent base vectors $(\mathbf{u}_i, \mathbf{v}_i)$. For a rectified binocular camera setup, camera rays from view $s \in \{l, r\}$ intersect the surfels at $\mathbf{x}_{i,s}$. Based on the 2D Gaussian profile $G(\mathbf{x}_{i,s}) = \exp\left(-\frac{u(\mathbf{x}_{i,s})^2 + v(\mathbf{x}_{i,s})^2}{2}\right)$, the synthesized color at pixel p in each stereo view is formulated as:

$$I_s(p) = \sum_{i \in \mathcal{N}_s} c_i(\mathbf{T}_s) \alpha_i G(\mathbf{x}_{i,s}) \prod_{j=1}^{i-1} (1 - \alpha_j G(\mathbf{x}_{j,s})) \tag{5}$$

During adversarial optimization, the geometric and opacity parameters of $g_i \in \mathcal{S}_{\text{target}}$, together with all parameters of $g_i \in \mathcal{S}'_{\text{frozen}}$, are frozen. Only the surface radiance parameters θ are updated. To increase the stealthiness of the attack, we apply the "Merging Attack" strategy Zhao et al. (2026), which aims to align the depth of the target object with its surrounding background through mask expansion and background depth extraction. Unlike the original "Merging Attack", which extracts the surrounding depth from the benign prediction of the target model, we directly derive the surrounding disparity d_{bg} from the geometric ground-truth disparity d_t rendered by the reconstructed 2DGS scene. This scene-derived reference is independent of the victim model, ensuring a consistent geometric prior across different target models. Following the merging strategy, the surrounding disparity d_{bg} and the target mask $\mathcal{M}$ are correspondingly partitioned into top and bottom regions, yielding $(d_{\text{bg}}^{\text{top}}, d_{\text{bg}}^{\text{bottom}})$ and $(\mathcal{M}_{\text{top}}, \mathcal{M}_{\text{bottom}})$, respectively. The predicted disparity of each target region is then optimized toward its corresponding surrounding disparity to enhance attack stealthiness. Given the predicted disparity $d = F(I_l, I_r)$, we denote d_{obj}^k as the average predicted disparity within target region $\mathcal{M}_k$. The resulting surfel-based adversarial loss is formulated as:

$$\mathcal{L}_{\text{main}}(\theta) = \mathbb{E}_{\mathbf{T} \sim \mathcal{P}} \left[\sum_{k \in \{\text{top,bottom}\}} \text{MSE}\left(d_{\text{obj}}^k, d_{\text{bg}}^k\right) \right] \tag{6}$$

To extend 3DOSA to other vision tasks such as object detection, the main objective can be readily adapted to a "Hiding Attack":

$$\mathcal{L}_{\text{main}}(\theta) = \mathbb{E}_{\mathbf{T} \sim \mathcal{P}} \left[F_{\text{conf}}(I, y_{\text{target}}) \right], \tag{7}$$

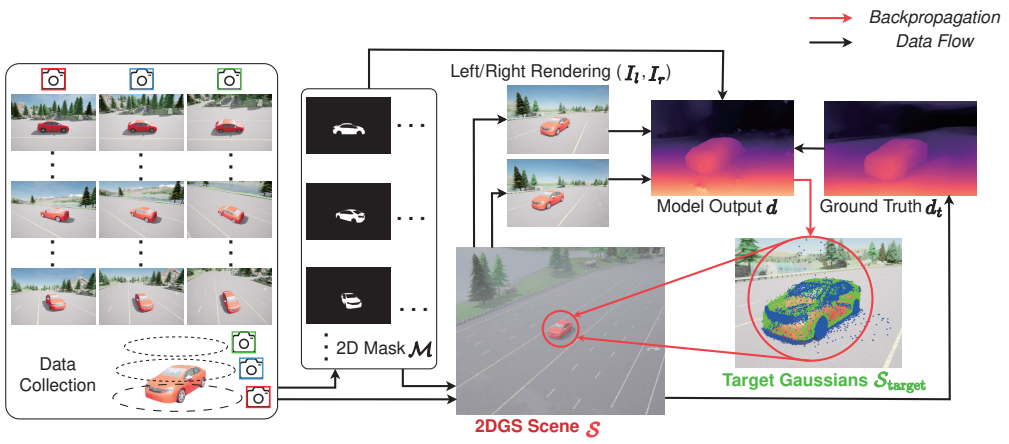


Figure 4: Overview of our attack against stereo matching models using 3DOSA. The initial adversarial 2DGS scene with explicit Gaussian identity features $\mathcal{S}$ is reconstructed from multi-view observations at different heights. The left and right images I_l and I_r are rendered from the current scene and fed into the target stereo model to obtain the predicted disparity d. The scene-rendered geometric ground truth d_t, together with the corresponding target mask $\mathcal{M}$, provides the reference for adversarial optimization. Finally, the target Gaussian radiance parameters are updated through the surfel-based adversarial loss.

where $F_{\text{conf}}(I, y_{\text{target}})$ denotes the mean confidence of the top-K target-class predictions associated with the target object specified by y_{target}.

Finally, Total Variation (TV) and Non-Printability Score (NPS) losses are incorporated with weights α and β for better physical deployability, respectively:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{main}}(\theta) + \alpha\mathcal{L}_{\text{nps}}(\theta) + \beta\mathcal{L}_{\text{tv}}(\theta) \tag{8}$$

4 Experiments

4.1 Experimental Setup

Data Collection. We construct our training environment using the "Town04" map in CARLA Simulator. To capture multi-view geometric and photometric properties of the target vehicle, virtual camera rigs are distributed at three distinct elevation ($1.4m$, $3.2m$, $5m$) tiers surrounding the vehicle ($r = 9m$), with 100 viewpoints uniformly sampled along a circular trajectory at each elevation, yielding a total of 300 multi-view images. The cameras are oriented directly toward the vehicle's geometric center. The 2D foreground masks $\mathcal{M}$ covering vehicle surfaces are extracted via a material-ID rendering pass in CARLA.

To generate the digital test data, we position the stereo cameras at three distinct distances and seven heading angles relative to the target vehicle within the "Town10" map of CARLA Simulator to assess multi-view and scene robustness. To assess real-world deployability of our attack, we utilize a 1:30 scale 3D-printed vehicle model and apply the adversarial camouflage to its surface. A total of 100 stereo image pairs were collected across two distances (50cm, 100cm) and three viewing perspectives (front, side, back) using an iPhone 14 Pro Max camera with a rescaled KITTI stereo-baseline setup.

Implementation Details. For scene representation, the Gaussian Grouping module uses a 16-dimensional identity feature $\mathbf{f}_i \in \mathbb{R}^{16}$. The reconstructed 2DGS scene contains approximately 20,000 surfels in the full scene $\mathcal{S}$, including approximately 4,000 surfels belonging to the target vehicle $\mathcal{S}_{\text{target}}$. We train the 2DGS scene $\mathcal{S}$ for 30,000 iterations following the standard densification and pruning schedule, during which all surfel parameters are optimized until high-fidelity visual and geometric convergence is achieved. Following scene reconstruction, we randomly sample stereo camera extrinsics from the viewpoint distribution $\mathcal{T}$ to render training image pairs for adversarial optimization. To enhance the physical robustness of the generated camouflage under varying environmental conditions, we apply EoT with random ambient illumination intensity scaling factors



Figure 5: Depth outputs of the target model from different viewpoints in CARLA simulator.

Table 1: Attack comparison results against different target models.

Method	PSMNet		UniMatch		RAFT-Stereo		CREStereo		FoundationStereo	
	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{bad}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{bad}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{bad}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{bad}} \uparrow$	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{bad}} \uparrow$
Benign	0.752	0.013	0.916	0.011	0.828	0.012	0.785	0.017	0.953	0.011
Random	0.677	0.052	0.894	0.015	0.793	0.024	0.684	0.029	0.921	0.011
Liu et al.	0.475	0.154	0.824	0.040	0.502	0.148	0.431	0.094	0.861	0.015
Cheng et al.	0.510	0.176	0.799	0.038	0.614	0.143	0.444	0.077	0.882	0.014
StereoAlign	0.058	0.553	0.356	0.270	0.082	0.571	0.071	0.598	0.509	0.253
3DOSA	**0.041**	**0.835**	**0.093**	**0.702**	**0.070**	**0.768**	**0.058**	**0.792**	**0.109**	**0.676**

sampled from $\mathcal{U}(0.8, 1.2)$. For "Merging Attack", the dilation kernel size for foreground expansion is set to 50 pixels. For "Hiding Attack", K is set to 20. During adversarial optimization, the geometric and opacity parameters of the target surfels are frozen, and the adversarial camouflage is optimized for 100 epochs using the Adam optimizer with an initial learning rate of 0.01. Finally, the TV and NPS loss weights α and β are both set to 0.01.

Evaluation Metrics. To comprehensively assess attack efficacy, stealthiness, and model transferability, we adopt three complementary metrics with a pixel-wise threshold $\tau = 3$: ① **Bad Pixel Ratio:** $\mathcal{E}_{\text{Bad}} = \frac{1}{|m|}\sum_{(i,j)\in m} \mathbb{I}\left(|d(i,j) - d_{\text{t}}(i,j)| > \tau\right)$, which quantifies the proportion of target pixels whose predicted disparity deviates from the ground-truth disparity d_{t} by more than τ; ② **Blending Error:** $\mathcal{E}_{\text{blend}} = \frac{1}{|m|}\sum_{(i,j)\in m} \mathbb{I}\left(|d(i,j) - \bar{d}_{\text{bg}}(i)| > \tau\right)$, which measures how well the target disparity blends into the background reference disparity $\bar{d}_{\text{bg}}$; and ③ **Average Precision (AP_{50}):** the standard detection metric at IoU $= 0.5$ for evaluating cross-task transferability.

Target Models & Baseline Attacks. We evaluate five representative stereo matching models covering three major architectures: ① **Cost Volume-based**: PSMNet Chang & Chen (2018); ② **Recurrent Iterative-based**: RAFT-Stereo Lipson et al. (2021) and IGEV-Stereo Wang et al. (2023); and ③ **Transformer-based**: UniMatch Yang et al. (2023) and Foundation-Stereo Wen et al. (2025). To assess the cross-task transferability of 3DOSA to 2D object detection, we further evaluate the attack against three representative object detectors: Faster R-CNN Ren et al. (2017), YOLOv5 Moulouel & Fu (2024), and DETR Carion et al. (2020). We compare our attack against the most relevant physical adversarial attacks targeting stereo matching, including Cheng et al. Cheng et al. (2021), Liu et al. Liu et al. (2024), and the stereo-aligned camouflage attack Zhao et al. (2026). Furthermore, we compare 3DOSA against other 3D Gaussian-based physical attacks on 2D object detection, including PGA Lou et al. (2025), Meng et al. (2025) and 3DGAA Zhang et al. (2025).

4.2 Experiment Results

4.2.1 Attack Effectiveness

Digital Attack. Table 1 presents the quantitative attack performance against various stereo matching models, with both metrics averaged across all evaluated viewpoints. Our 3DOSA framework consistently outperforms all baseline attacks across the evaluated target models and metrics. This performance gain is particularly evident compared with dimensionally decoupled attacks. Although conventional 3D-to-2D attacks remain effective against traditional cost-volume-based models, their

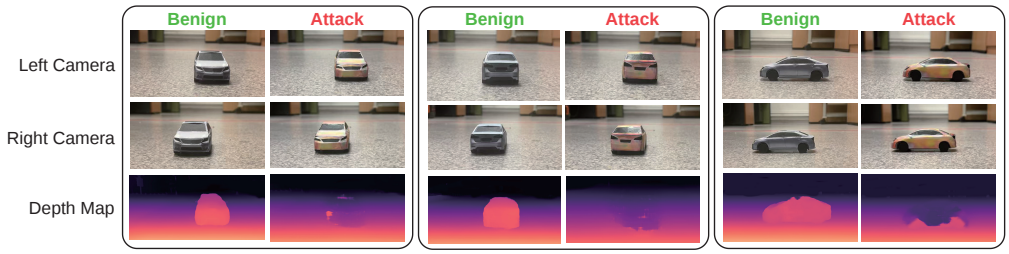


Figure 6: Depth outputs of the target model from different viewpoints in real world.

Table 3: Attack comparison results from different viewpoints and weathers in CARLA Simulator.

Method	Metric	Distance		Heading Angle (°)							Weather	
		5m	15m	0	30	60	90	120	150	180	Sunset	Rainy
Benign	$\mathcal{E}_{\text{blend}} \downarrow$	0.852	0.668	0.783	0.772	0.761	0.757	0.742	0.737	0.715	0.654	0.672
Random		0.823	0.656	0.767	0.759	0.743	0.736	0.728	0.720	0.710	0.643	0.666
Liu et al.		0.811	0.646	0.732	0.727	0.720	0.711	0.703	0.696	0.689	0.631	0.638
Cheng et al.		0.813	0.650	0.748	0.735	0.721	0.714	0.707	0.699	0.692	0.645	0.650
StereoAlign		0.346	0.427	0.321	0.328	0.334	0.347	0.346	0.352	0.359	0.457	0.470
3DOSA		**0.074**	**0.123**	**0.082**	**0.083**	**0.095**	**0.117**	**0.104**	**0.100**	**0.091**	**0.112**	**0.129**
Benign	$\mathcal{E}_{\text{bad}} \uparrow$	0.010	0.024	0.013	0.015	0.018	0.019	0.020	0.020	0.018	0.029	0.250
Random		0.033	0.051	0.044	0.046	0.048	0.051	0.050	0.053	0.049	0.067	0.065
Liu et al.		0.091	0.074	0.080	0.084	0.081	0.079	0.085	0.076	0.080	0.112	0.106
Cheng et al.		0.089	0.071	0.081	0.083	0.078	0.075	0.078	0.080	0.083	0.117	0.125
StereoAlign		0.366	0.220	0.317	0.320	0.315	0.319	0.316	0.316	0.317	0.392	0.388
3DOSA		**0.702**	**0.643**	**0.680**	**0.674**	**0.641**	**0.632**	**0.639**	**0.668**	**0.675**	**0.562**	**0.550**

performance degrades considerably against Transformer-based architectures (e.g., UniMatch and FoundationStereo). This degradation may be attributed to perspective distortions and boundary discontinuities introduced by fixed 3D-to-2D texture projections, which can weaken cross-view consistency under architectures with global contextual modeling. In contrast, 3DOSA directly optimizes explicit 3D Gaussians in the global physical space to preserve cross-view geometric consistency, effectively avoiding such cross-dimensional fragmentation. In terms of stealthiness, the adversarial camouflage generated by 3DOSA and StereoAlign blends more effectively into the background, achieving substantially lower $\mathcal{E}_{\text{blend}}$ than the other attacks.

Physical Attack. Table 2 presents the quantitative results of 3DOSA in the real world. Consistent with the digital evaluation, the attack maintains its effectiveness and stealthiness across different viewpoints and distances. Figure 6 illustrates the adversarial texture applied to the scaled vehicle model. The benign disparity maps accurately recover the vehicle geometry, validating the physical setup. In contrast, the adversarial disparity maps cause the vehicle disparity to blend substantially with the surrounding background across different viewpoints, demonstrating its physical realizability and robustness.

Table 2: Real-world evaluation results under varying viewing angles and distances.

Metric	Tex.	View Angle			Distance (cm)	
		Front	Side	Back	50	100
$\mathcal{E}_{\text{blend}} \downarrow$	Benign	0.749	0.774	0.768	0.741	0.617
	Adv.	**0.085**	**0.121**	**0.094**	**0.087**	**0.102**
$\mathcal{E}_{\text{bad}} \uparrow$	Benign	0.028	0.020	0.24	0.025	0.032
	Adv.	**0.520**	**0.491**	**0.549**	**0.574**	**0.467**

4.2.2 Attack Analysis

Attack Robustness. Table 3 presents a comprehensive evaluation of attack robustness across diverse viewpoints and illumination conditions. Notably, our 3DOSA-based attack maintains consistent performance across all physical variations. As expected, attack effectiveness slightly degrades as the distance increases or ambient illumination decreases, primarily due to lower spatial resolution and reduced contrast that weaken high-frequency camouflage features. Furthermore, attack performance reaches its lowest point at a heading angle of 90° ($\mathcal{E}$blend $= 0.117$, $\mathcal{E}$bad $= 0.632$), which can be attributed to the flatter side of the vehicle containing fewer optimizable Gaussians than the geometrically complex front and rear surfaces. Nevertheless, 3DOSA still outperforms all baseline attacks by a substantial margin, demonstrating its robustness across diverse physical conditions.

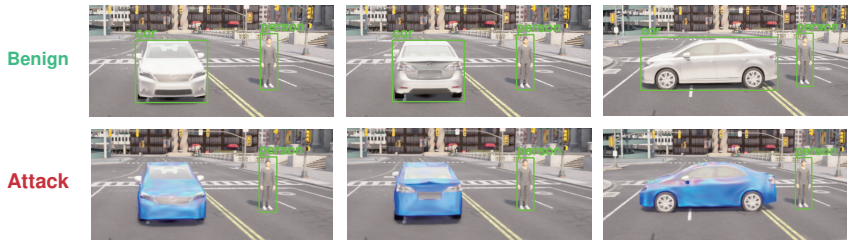


Figure 7: Task-agnostic transferability of 3DOSA against object detection.

Ablation Study. To investigate the individual contribution of each module in our framework, we conduct ablation experiments under four distinct configurations: ① **Vanilla 3DGS**: baseline 3D scene representation with 2D mask-based Gaussian extraction; ② **3DGS w/ Decouple**: incorporating explicit Gaussian decoupling into the baseline 3DGS; ③ **Vanilla 2DGS**: adopting 2D Gaussian surfels for adversarial optimization; and ④ **3DOSA**: our integrated 3D-to-3D attack framework.

Table 4: Ablation study results.

Configuration	$\mathcal{E}_{\text{blend}} \downarrow$	$\mathcal{E}_{\text{bad}} \uparrow$
2D Texture Map	0.541	0.166
Vanilla 3DGS	0.327	0.272
3DGS w/ Decouple	0.114	0.520
Vanilla 2DGS	0.209	0.359
3DOSA	**0.074**	**0.702**

As shown in Table 4, each module contributes to improving camouflage quality and physical feasibility in 3DOSA. In both *Vanilla 3DGS* and *Vanilla 2DGS*, Gaussian extraction through 2D masks without explicit 3D spatial constraints causes gradient leakage across adjacent Gaussians. While both configurations exhibit improved digital attack efficacy compared to conventional attacks with *2D Texture Map*, corrupted boundaries and disordered surface patterns make the resulting camouflage difficult to realize in the physical world. With explicit Gaussian grouping (*3DGS w/ Decouple*), the optimization gradients are strictly confined to the designated Gaussians. Nevertheless, volumetric 3D Gaussians lack precise surface alignment, leading to noticeable floating artifacts along the vehicle contours. By combining 2D Gaussian surfels with explicit Gaussian grouping, 3DOSA produces physically feasible and high-fidelity adversarial camouflage with strong attack efficacy.

Task-Agnostic Transferability. To demonstrate that 3DOSA is task-agnostic, we further extend our pipeline to optimize adversarial camouflage against 2D object detection models by modifying the main loss to suppress the confidence score of the "Car" class (Eq. 7). Table 5 presents the average precision comparison across different 3DGS-based attacks. The results show that 3DOSA retains high attack efficacy when transferred to 2D detectors while consistently outperforming existing baselines

Table 5: 3DGS-based attack comparison results against object detection. (mAP@0.5 $\downarrow$)

Method	Target Detector			Avg.
	Faster	YOLO	DETR	
Clean	85.40	89.11	91.68	88.73
PGA	32.31	34.77	39.95	35.68
3DGAA	54.22	57.86	60.01	57.36
Meng et al.	57.23	58.09	63.42	59.58
3DOSA	**18.45**	**22.62**	**24.11**	**21.73**

Figure 7 illustrates the detection results under benign and adversarial scenarios from different viewpoints. In the benign scenario, both the car and person are successfully detected. When the adversarial camouflage is applied to the vehicle, the car is no longer detected, while the person remains successfully detected.

5 Conclusion

In this work, we presented 3DOSA, the first unified object-scene radiance field framework for physically realizable adversarial attacks in autonomous driving. To address adversarial gradient leakage and non-physical floating artifacts, our framework integrates explicit spatial decoupling with surface-aligned 2D Gaussian surfels for adversarial optimization. Both simulated and real-world experiments demonstrate that 3DOSA achieves strong attack effectiveness, multi-view consistency, and task transferability. These results establish 3DOSA as a viable framework and testing platform for future physical adversarial attacks and defenses.

AI USE STATEMENT

In this work, we acknowledge the use of generative AI tools exclusively for writing assistance, limited to polishing grammar and style. No part of the research design, experiments, analysis, or core contributions relied on generative AI tools, and all AI-assisted text was critically reviewed and finalized by the authors. We take full responsibility for the final content of this work.

REFERENCES

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018.

Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Liu, and Bo Zhou. Invisible for both camera and lidar: Security properties of multi-modal 3d object detection in autonomous driving. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 998–1015, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5418, 2018.

Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pp. 52–68, 2019.

Yulong Cheng, Xiao Guo, Jiakai Shen, Bo Wang, and Xuhong Lv. Towards robust physical-world backdoor attacks on stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16180–16189, 2021.

Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiqiang Shen, and Xiangzhe Liu. Physical attack on monocular depth estimation with optimal adversarial patches. In *European Conference on Computer Vision (ECCV)*, pp. 514–532, 2022.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1625–1634, 2018.

Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Hang Su. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13307–13316, 2022.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *ACM Transactions on Graphics (ToG)*, 43(4):1–12, 2024a.

Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 717–726. IEEE, 2020.

Tao Huang et al. Transferable targeted physical adversarial attack via implicit neural representation. In *European Conference on Computer Vision (ECCV)*, 2024b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pp. 218–227, 2021.

Yang Liu, Jucai Zhai, Chihao Ma, Pengcheng Zeng, Xinan Wang, and Yong Zhao. Physical Attack for Stereo Matching. In *Proc. of the 1st ACM CVDL*, pp. 1–5, 2024.

Tianrui Lou, Xiaojun Jia, Siyuan Liang, Jiawei Liang, Ming Zhang, Yanjun Xiao, and Xiaochun Cao. 3d gaussian splatting driven multi-view robust physical adversarial camouflage generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28752–28762. IEEE, 2025.

Lingzhuang Meng, Mingwen Shao, Yuanjian Qiao, Wenjie Liu, and Xiang Lv. Gaussian splitting attack: Gaussian splatting-based multi-view 3d adversarial attack. *Pattern Recognition*, pp. 112466, 2025.

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. *Communications of the ACM*, 65(1):99–106, 2021.

Moulouel and Skylar Fu. What is yolov5: A deep look into the internal features of the popular object detector, 2024. URL `https://arxiv.org/abs/2407.20892`.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149, 2017.

Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mungan, Daniel Delbridge, and David Wagner. Darts: Deceiving autonomous cars with toxic signs. In *arXiv preprint arXiv:1802.06430*, 2018.

Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Yeonsik Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Dta: Physical camouflage attacks using differentiable transformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15305–15314, 2022.

Naufal Suryanto, Yongsu Kim, Harashta Tatimma Larasati, Hyoeun Kang, Thi-Thu-Huong Le, Yoonyoung Hong, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Active: Towards highly transferable 3d physical camouflage for universal and robust vehicle evasion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4282–4291. IEEE, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8248–8258, 2022.

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

James Tu, Mengye Ren, Josef Sivic, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13716–13725, 2020.

Donghua Wang, Tianshu Yao, Xiao Chen, Mengye Ren, Donglin Zou, and Jinhui Tang. Dual attention suppression attack on dilated convolutional networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8412–8421, 2021.

Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaoya Zhang, Wen Yao, and Xiaoqian Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2414–2422, 2022.

Gangwei Wang, Yun Wang, Jiayu Pan, Lixing Pan, Baichen Sun, and Hong Ding. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21919–21928, 2023.

Bowei Wen, Jonathan Tremblay, Kevin Lin, Yu Xiang, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Alex Wong, Igor Cvisic, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13065–13074, 2021.

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 *European Conference on Computer Vision (ECCV)*, pp. 665–681, 2020.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 2023.

Mingqiao Ye, Martin Danelljan, Matthew Fisher, Nanxuan Li, Yue Hu, and Fisher Yu. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision (ECCV)*, 2024.

Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 185–194, 2019a.

Yang Zhang, Hani Hassan, Eline Shi, and Benjamin IP Rubinstein. Camou: Learning physical vehicle camouflages to fool neural network object detectors. In *arXiv preprint arXiv:1912.00196*, 2019b.

Yiming Zhang et al. Bitaa: A bi-task adversarial attack for object detection and depth estimation via 3d gaussian splatting. *arXiv preprint arXiv:2408.06732*, 2024.

Yixun Zhang, Lizhi Wang, Junjun Zhao, Wending Zhao, Feng Zhou, Yonghao Dang, and Jianqin Yin. 3dgaa: Realistic and robust 3d gaussian-based adversarial attack for autonomous driving. *arXiv preprint arXiv:2507.09993*, 2025.

Kangqiao Zhao, Shuo Huai, Xurui Song, and Jun Luo. Cheating stereo matching in full-scale: Physical adversarial attack against binocular depth estimation in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13190–13198, 2026.

Zhaoyu Zhou, Xinyi Jiang, Jun Li, Xianglong Liu, and Rongrong Ji. Physical adversarial attack on vehicle detection with robust and natural camouflage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8244–8253, 2023.

A Appendix

A.1 Training Data

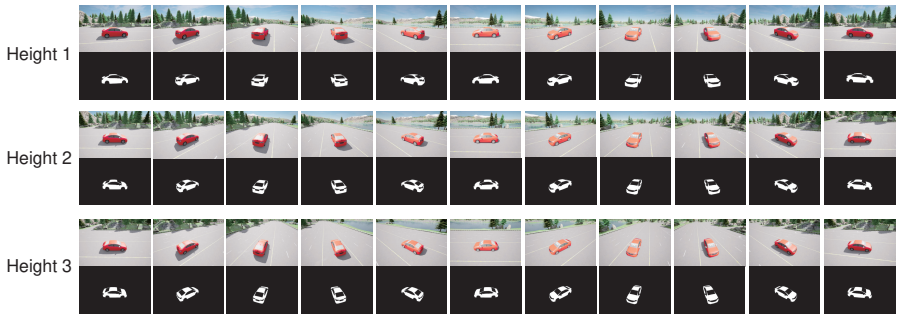


Figure 8: Examples of multi-view training images and their corresponding printable-region masks.

Figure 8 visualizes a subset of the training data collected in "Town04" of CARLA Simulator and their corresponding 2D masks of the printable regions on the vehicle. These multi-view training data are used to reconstruct the initial radiance field for subsequent adversarial optimization.

A.2 Adversarial Radiance Field

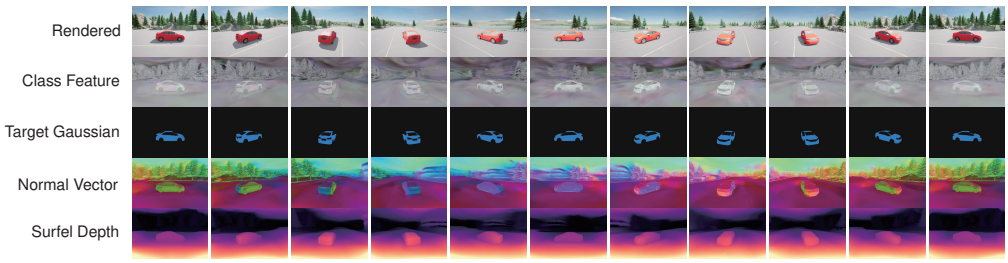


Figure 9: Different representations of the target object in the reconstructed 2DGS scene.

Figure 9 illustrates different representations of the target object in the scene from multiple viewpoints. From top to bottom: ① **Rendered**: RGB images rendered from the reconstructed scene; ② **Class Feature**: visualization of the 16-dimensional identity features learned through Gaussian Grouping; ③ **Target Gaussian**: target Gaussians identified through explicit Gaussian decoupling; ④ **Normal Vector**: surface normal maps of the 2D Gaussian surfels and ⑤ **Surfel Depth**: depth maps rendered from the reconstructed 2DGS scene.

A.3 Additional Experimental Results

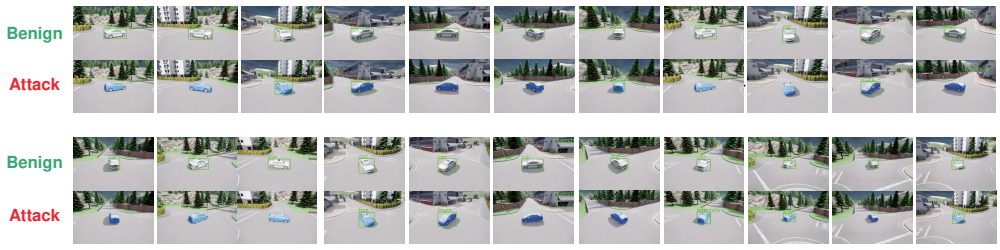


Figure 10: Attack results across different viewing angles and distances in "Town02".

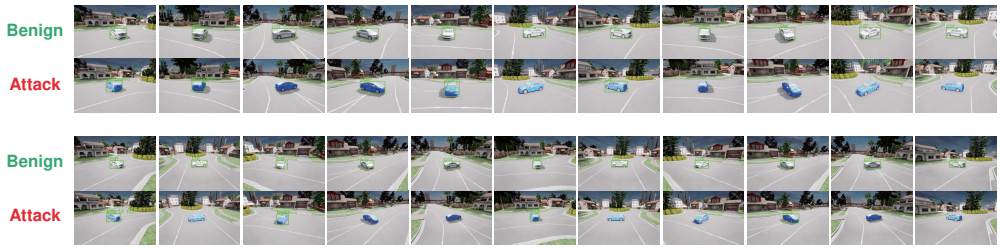


Figure 11: Attack results across different viewing angles and distances in "Town03".

Figures 10 and 11 demonstrate the attack results in different scenes. The results show that the adversarial vehicle is either not detected or misclassified as other classes across most viewpoints. Although the adversarial camouflage is optimized in a single scene, it maintains strong attack performance and robustness within unseen environments.