CAMOUPURIFIER: ZERO-SHOT AND CAMOUFLAGE-AGNOSTIC DEFENSE AGAINST ADVERSARIAL CAMOUFLAGE ATTACKS ON MONOCULAR DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial camouflage attacks pose a significant threat to monocular depth estimation (MDE) systems in autonomous driving due to their physical realizability and robustness across diverse viewpoints. Existing adversarial defenses primarily focus on imperceptible perturbations or localized adversarial patches, rendering them ineffective against full-surface adversarial camouflage textures and potentially introducing additional geometric distortions. To address these limitations, we propose CamouPurifier, a zero-shot and camouflage-agnostic adversarial purification framework that incorporates adversarial camouflage localization and texture reconstruction. Specifically, we first employ a low-confidence-threshold detector to identify candidate vehicles and then integrate semantic-aware localization (SAL) with multi-domain texture anomaly localization (MTAL) to generate precise adversarial camouflage masks. Subsequently, we remove the localized adversarial patterns and reconstruct the missing vehicle textures using a geometry-guided diffusion model conditioned on vehicle edge information. To integrate the reconstructed textures into the original image, we further apply boundary-aware blending to maintain visual consistency with the surrounding scene. Comprehensive experiments demonstrate that CamouPurifier reduces the vehicle-region depth discrepancy by 84.5% compared with the undefended setting and by 51.2% compared with the strongest baseline. Its fidelity-oriented configuration further reduces global and background discrepancies by 53.7% and 41.2%, respectively.

1 INTRODUCTION

Monocular depth estimation (MDE) is a fundamental component of autonomous driving systems, providing dense geometric information from a single RGB image for obstacle localization and collision avoidance. Despite recent advances, MDE models remain vulnerable to adversarial attacks Wong et al. (2020); Yamanaka et al. (2020); Guesmi et al. (2024). In particular, physically realizable adversarial camouflage attacks Zheng et al. (2024); Zhang et al. (2026) can map optimized textures onto vehicle surfaces and manipulate depth predictions under varying viewpoints, distances, weather conditions, and partial occlusions. Such attacks may cause a vehicle to appear substantially closer or farther than its actual position, posing direct risks to downstream driving decisions.

Existing adversarial defenses primarily address norm-bounded perturbations or localized adversarial patches. Input purification methods remove imperceptible perturbations through denoising or generative restoration Nie et al. (2022); Lin et al. (2025), whereas patch defenses localize and suppress suspicious regions using entropy Tarchoun et al. (2023), mutual information Jing et al. (2024); Wang et al. (2026), or spatial inconsistency Xiang et al. (2023). However, these methods do not transfer well to adversarial camouflage, as illustrated in Figure. **??**. First, adversarial camouflage typically covers most of a vehicle's visible surface and forms multiple disconnected regions, which breaks the compact-region assumption adopted by traditional patch-based defenses Xiang et al. (2023). Second, changes in viewpoint, illumination, distance, and weather substantially alter its visual appearance, making low-level texture features unreliable for localization. Third, monoc-

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.

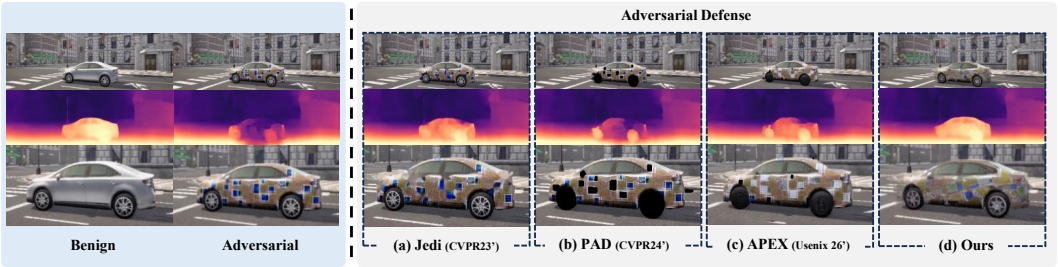


Figure 1: Qualitative comparison of adversarial defenses. Jedi leaves substantial camouflage residuals, while PAD excessively masks depth-relevant vehicle structures and APEX introduces reconstruction artifacts. CamouPurifier effectively removes the adversarial texture and restores the depth prediction while preserving the original vehicle geometry.

ular depth estimation relies heavily on object contours and local geometric structures. Aggressive masking or unconstrained restoration can therefore damage geometry-related cues and introduce new depth errors even after removing the adversarial texture, and affecting other perception tasks. These challenges make full-surface adversarial camouflage substantially different from conventional perturbations and patches, while its detection and purification for MDE remain underexplored.

To address these challenges, we propose CamouPurifier, a zero-shot and camouflage-agnostic framework for adversarial camouflage localization and geometry-preserving texture reconstruction. CamouPurifier first detects candidate vehicle regions to exclude irrelevant background content. Within each vehicle region, the semantic-aware localization module (SAL) uses frozen DINO Siméoni et al. (2025) representations to identify local features that are inconsistent with the global semantics of a natural vehicle. In parallel, the multi-domain texture anomaly localization module (MTAL) captures abnormal appearance and texture variations. The two localization features are adaptively thresholded and fused to obtain the final camouflage mask. Based on the detected mask, CamouPurifier removes the adversarial texture and reconstructs a benign vehicle appearance using a geometry-guided diffusion model. The camouflage mask is decomposed into connected components and processed within local crops. Canny edge maps are introduced through ControlNet to constrain the reconstructed content to follow the original vehicle contour and body structure. Finally, Gaussian boundary blending integrates the reconstructed regions into the input image while preserving benign background content. The entire framework operates as an input preprocessing module, requiring no modifications or retraining of the target MDE model.

Our main contributions are summarized as follows:

- We investigate the underexplored problem of detecting and purifying full-surface adversarial camouflage for monocular depth estimation, and identify the limitations of directly applying existing adversarial patch defenses.
- We propose CamouPurifier, a zero-shot and camouflage-agnostic framework that combines semantic inconsistency and multi-domain texture anomaly cues for distributed camouflage localization, followed by geometry-guided diffusion reconstruction for depth-relevant structure preservation.
- We conduct extensive experiments across four different attacks, six MDE models, and diverse physical conditions. The results demonstrate substantial improvements over existing defenses in vehicle-region purification, global depth consistency, and cross-condition generalization.

2 Related Work

2.1 Adversarial Camouflage Attacks

Adversarial camouflage is a highly threatening physical attack that maps adversarial textures onto 3D object surfaces, providing greater robustness to viewpoint changes, long distances, and partial occlusions than conventional 2D adversarial patches. Early methods, such as DTA Suryanto et al.

(2022), FCA Wang et al. (2022), and RAUCA Zhou et al. (2024), mainly attacked object detectors through differentiable rendering and full-surface texture optimization. 3D^{2}Fool Zheng et al. (2024) later introduced the first 3D texture-based attack against monocular depth estimation. Although APA Yamanaka et al. (2020) and SSAP Guesmi et al. (2024) were designed for 2D adversarial patches, their adversarial losses also perform well when used for 3D texture optimization. For black-box scenarios, DepthHack Sun et al. (2025) applied probabilistic sampling and score-based optimization to craft robust 3D adversarial textures. However, there are few defenses against these attacks.

2.2 Adversarial Purification Defenses

Existing adversarial defenses mainly target imperceptible perturbations through input transformations such as JPEG compression, Gaussian filtering, bit-depth reduction, and diffusion-based purification methods including DiffPure Nie et al. (2022) and ADBM Li et al. (2025). More recent studies have extended defense to adversarial patches, where SAC Liu et al. (2022), Jedi Tarchoun et al. (2023), PAD Jing et al. (2024), ObjectSeeker Xiang et al. (2023), and APEX Wang et al. (2026) typically detect, localize, and remove suspicious regions before prediction. However, these methods are primarily designed for localized patches, whereas adversarial camouflage distributes perturbations over the entire 3D object surface and exhibits strong variations across viewpoints, distances, and illumination conditions. Therefore, effective purification defenses against full-surface 3D adversarial camouflage, particularly for monocular depth estimation, remain largely unexplored.

3 Preliminaries

Given a background scene $\mathbf{I}_s$, a vehicle mesh $\mathcal{V}_{mesh}$, and a camera and environment configuration Φ, the clean and adversarial vehicle renderings are obtained as

$$(\mathbf{I}_v^{adv}, \mathbf{M}_v) = \mathcal{R}(\mathcal{V}_{mesh}, \mathbf{T}_{adv}, \Phi), \tag{1}$$

where $\mathbf{T}_{adv}$ denote the adversarial texture, and $\mathbf{M}_v$ is the visible vehicle mask. The corresponding clean and adversarial scene images are constructed as

$$\mathbf{I}_{adv} = \mathbf{M}_v \odot \mathbf{I}_v^{adv} + (1 - \mathbf{M}_v) \odot \mathbf{I}_s. \tag{2}$$

Unlike localized adversarial patches, adversarial camouflage can cover the entire visible object surface, and its image-space distribution varies substantially with viewpoint, distance, illumination, and weather conditions.

A camouflage defense consists of a detector $\mathcal{D}$ and a purifier $\mathcal{P}$. The detector predicts a suspicious-region mask, while the purifier reconstructs the detected regions to recover a benign input:

$$\widetilde{\mathbf{D}} = \mathcal{M}_\theta(\mathcal{P}(\mathbf{I}_{adv}, \mathcal{D}(\mathbf{I}_{adv}))), \tag{3}$$

where $\mathcal{M}_\theta$ is the target MDE model and $\widetilde{\mathbf{D}}$ is the purified depth prediction. The defense aims to suppress camouflage-induced depth errors while preserving the background and other benign image regions.

4 Methodology

4.1 Defense Pipeline

In this section, we introduce the pipeline of CamouPurifier. As illustrated in Figure. 2, CamouPurifier consists of two stages: camouflage localization and texture reconstruction. First, given an adversarial scene image, we employ an object detector with low-confidence threshold to identify vehicles in the scene. Then, Semantic-Aware Localization (SAL) and Multi-Domain Texture Anomaly Localization (MTAL) are used to detect semantically sensitive regions and texture-anomalous regions, respectively. The two types of regions are subsequently combined through intersection operations to obtain the final camouflage mask. Second, the adversarial camouflage within the localized mask is removed. Following PAD Jing et al. (2024), we replace the masked region with black pixels. Finally, a diffusion model conditioned on vehicle geometry is employed to generate the purified texture, thereby preserving the original geometric structure of the vehicle to the greatest extent possible.

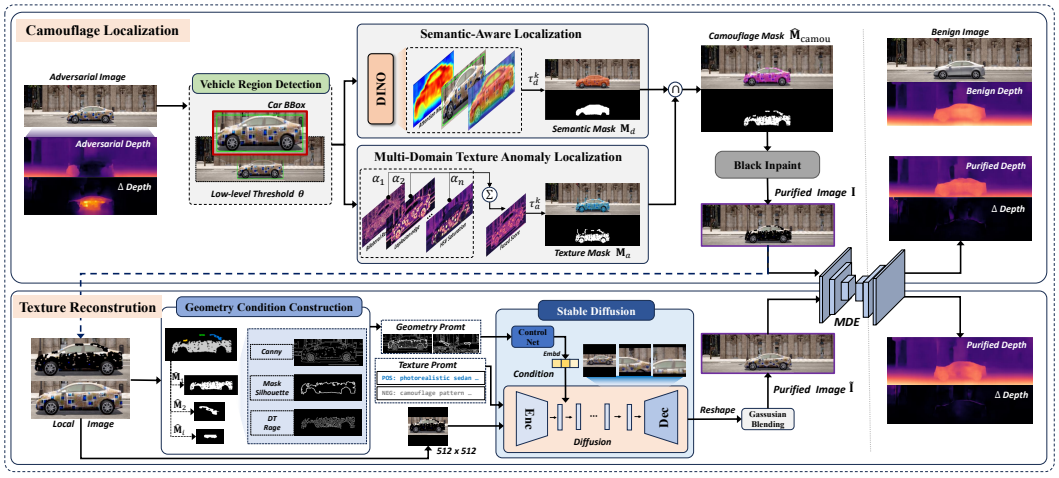


Figure 2: Framework of CamouPurifier defense. CamouPurifier first uses camouflage localization to identify and remove camouflage textures, then uses texture reconstruction to generate a clean vehicle texture and reapplies them to the scene image.

4.2 Camouflage Localization

For scene images, the first step is to locate the positions of vehicles with adversarial camouflage in the image. In CamouPurifier, this involves four parts: vehicle region detection, semantic-aware localization, multi-domain texture anomaly localization, and mask fusion.

Vehicle Region Detection. Adversarial camouflage is usually applied to vehicle surfaces, whereas most pixels in an autonomous-driving image belong to benign background regions. Directly analyzing the entire image may introduce unnecessary false positives and computational overhead. We first employ a pretrained YOLOv8 Varghese & Sambath (2024) detector to obtain a set of vehicle bounding boxes

$$\mathcal{B} = \mathcal{D}_{\text{yolo}}\left(\mathbf{I}_{\text{adv}}; \theta\right) = \{\mathbf{b}_k\}_{k=1}^{K}, \tag{4}$$

where $\mathbf{b}_k = (x_k^0, y_k^0, x_k^1, y_k^1)$ denotes the k-th detected vehicle region and θ is the detection threshold.

Semantic-Aware Localization (SAL). For each detected vehicle region $\mathbf{b}_k$, we employ a pretrained DINOv3 Siméoni et al. (2025) encoder $\mathcal{F}_{\text{DINO}}(\cdot)$ to extract a dense semantic response map:

$$\mathcal{S}_d^k = \mathcal{F}_{\text{DINO}}\left(\mathbf{I}_{\text{adv}}|_{\mathbf{b}_k}\right), \tag{5}$$

Although adversarial camouflage may contain visually complex colors and patterns, its local representations often exhibit abnormal semantic responses compared with the coherent semantics of natural vehicles. We therefore obtain a binary localization mask by applying a percentile-based threshold to $\mathcal{S}_d^k$:

$$\tau_d^k = Q_{q_d}\left(\mathcal{S}_d^k\right), \tag{6}$$

$$\mathbf{M}_d^k(p) = \mathcal{H}_\phi\left(\mathbb{I}\left[\mathcal{S}_d^k(p) \geq \tau_d^k\right]\right), \tag{7}$$

where p denotes a spatial location within $\mathbf{b}_k$, $Q_{q_d}(\cdot)$ denotes the q_d-th percentile operator, $\mathbb{I}[\cdot]$ is the indicator function, and $\mathcal{H}_\phi(\cdot)$ denotes the operation that removes connected components whose area is smaller than the given threshold ϕ. The resulting local binary masks $\{\mathbf{M}_d^k\}_{k=1}^{K}$ are projected back to the original image coordinates and merged to form the global semantic localization mask $\mathbf{M}_d$.

4.2.1 Multi-Domain Texture Anomaly Localization (MTAL).

The MTAL branch localizes adversarial camouflage by analyzing abnormal appearance and texture distributions within each detected vehicle region. For the k-th vehicle region, their corresponding texture anomaly score maps are uniformly formulated as

$$\mathcal{S}_a^k = \sum_{t \in \mathcal{T}} \lambda_t \mathcal{N}\left(\mathcal{A}_t\left(\mathbf{I}_{\text{adv}}|_{\mathbf{b}_k}\right)\right), \tag{8}$$

where $\mathcal{T}$ denotes the set of anomaly analysis methods. λ_i controls the contribution of the i-th anomaly analysis method and $\sum_{i\in\mathcal{T}}\lambda_i = 1$. For CamouPurifier, we employ four complementary anomaly analysis methods, including bilateral residual, Laplacian response, local standard deviation, and color saturation, $\mathcal{T} = \{\mathrm{bil}, \mathrm{lap}, \mathrm{std}, \mathrm{sat}\}$. Specifically, $\mathcal{A}_{\mathrm{bil}}$ computes the bilateral residual to capture local appearance deviations, $\mathcal{A}_{\mathrm{lap}}$ extracts high-frequency intensity variations, $\mathcal{A}_{\mathrm{std}}$ measures local texture dispersion, and $\mathcal{A}_{\mathrm{sat}}$ identifies abnormal color saturation. $\mathcal{N}(\cdot)$ denotes the normalization operation, due to the response maps have different numerical ranges. A larger value in $\mathcal{S}_a^k$ indicates a stronger deviation from the natural appearance and texture distribution of the target vehicle. We then apply a percentile-based threshold to obtain the local binary texture anomaly mask:

$$\tau_a^k = Q_{q_a}\left(\mathcal{S}_a^k\right), \tag{9}$$

$$\mathbf{M}_a^k(p) = \mathcal{H}_\phi\left(\mathbb{I}\left[\mathcal{S}_a^k(p) \geq \tau_a^k\right]\right), \tag{10}$$

where p denotes a spatial location within $\mathbf{b}_k$, $Q_{q_a}(\cdot)$ denotes the q_a-th percentile operator. Finally, the resulting local masks $\{\mathbf{M}_a^k\}_{k=1}^K$ are projected back to the original image coordinates and merged to form the global texture anomaly localization mask $\mathbf{M}_a$.

Mask Fusion. The SAL module captures semantic inconsistency, while the MTAL captures local texture abnormality. To retain regions supported by both cues and reduce false detections, we use intersection as the default fusion strategy:

$$\hat{\mathbf{M}}_{\mathrm{camou}} = \Gamma\left(\mathbf{M}_d \odot \mathbf{M}_a\right), \tag{11}$$

where $\odot$ denotes element-wise multiplication and $\Gamma(\cdot)$ represents small-component removal.

4.3 Texture Reconstrution (TR)

After localization, directly removing or masking the suspicious pixels destroys the visual structure of the vehicle and may itself introduce large depth errors. We therefore design a geometry-preserving style projection $\mathcal{P}$ to replace adversarial patterns with a natural vehicle appearance while preserving the object boundary, body shape, and surrounding scene context.

Geometry condition construction. The fused mask is decomposed into connected components $\{\hat{\mathbf{M}}_i\}_{i=1}^N = \mathcal{C}(\hat{\mathbf{M}}_{\mathrm{camou}})$, $N \leq N_{\max}$, where small noisy components are discarded and at most $N_{\max} = 3$ components are retained. Processing the mask component by component reduces memory consumption and prevents unrelated vehicle regions from interfering with one another during diffusion-based reconstruction. For the i-th component, its bounding box is expanded by a padding margin p to retain sufficient contextual information. To prevent Stable Diffusion from changing the vehicle shape, we extract the edge structure of the local crop using a Canny operator Canny (1986):

$$\mathbf{C}_i = \mathrm{Canny}\left(\mathbf{I}_i\right). \tag{12}$$

The local image, camouflage mask, and edge condition are resized to the standard diffusion input resolution $\mathbf{I}_i^{512\times512\times3}$.

ControlNet-guided diffusion inpainting. The square crop is reconstructed using a Stable Diffusion Rombach et al. (2022) inpainting model conditioned on both the camouflage mask and the Canny edge map:

$$\mathbf{I}_i^r = \mathrm{StableDiffusion}\left(\mathbf{I}_i^s, \mathbf{M}_i^s, \mathbf{C}_i^s, \mathbf{e}^+, \mathbf{e}^-; \boldsymbol{\eta}\right), \tag{13}$$

where $\mathbf{e}^+$ and $\mathbf{e}^-$ denote the text embeddings of the positive and negative prompts, respectively. The positive prompt encourages a clean, smooth, and factory-painted vehicle surface, while the negative prompt suppresses camouflage, stripes, and graffiti. The parameter set $\boldsymbol{\eta}$ contains the number of denoising steps, classifier-free guidance scale, ControlNet Zhang et al. (2023) conditioning scale, inpainting strength, and random seed. During reverse diffusion, the ControlNet branch injects multi-scale edge features into the denoising U-Net:

$$\mathbf{z}_{t-1} = \mathcal{U}_\theta\left(\mathbf{z}_t, t, \mathbf{e}^+, \mathcal{C}_\psi(\mathbf{C}_i^s)\right), \qquad t = T, \ldots, 1, \tag{14}$$

where $\mathbf{z}_t$ is the noisy latent at step t, $\mathcal{U}_\theta$ is the latent denoising U-Net, and $\mathcal{C}_\psi$ is the ControlNet encoder. The edge condition constrains the generated content to follow the original vehicle contour, while the inpainting mask restricts reconstruction to the detected camouflage region. The generated result is resized back to the original crop resolution $\mathbf{I}_i^r = \mathcal{R}_i^{-1}(\mathbf{I}_i^r)$.

Boundary-aware Gaussian blending. Directly replacing the crop may produce visible seams along the mask boundary. We therefore smooth the local mask using a Gaussian kernel:

$$\boldsymbol{\alpha}_i = \mathcal{G}_\sigma * \mathbf{M}_i^c, \tag{15}$$

where $*$ denotes convolution and $\mathcal{G}_\sigma$ is a Gaussian kernel with standard deviation σ. The reconstructed and original crops are blended as

$$\widetilde{\mathbf{I}}_i = \boldsymbol{\alpha}_i \odot \mathbf{I}_i^r + (1 - \boldsymbol{\alpha}_i) \odot \mathbf{I}_i. \tag{16}$$

Each purified crop is pasted back into its original image location. After all connected components have been processed, we obtain the final purified image $\widetilde{\mathbf{I}}$.

Algorithm 1: CamouPurifier Defense

Input: Adversarial image $\mathbf{I}_{\text{adv}}$, thresholds q_d and q_a, minimum area $A_{\min}$, maximum component number $N_{\max}$

Output: Purified image $\widetilde{\mathbf{I}}$

Detect vehicle boxes $\mathcal{B} \leftarrow \mathcal{D}_{\text{yolo}}(\mathbf{I}_{\text{adv}})$.
for each vehicle box $\mathbf{b}_k \in \mathcal{B}$ **do**
  Extract the vehicle crop $\mathbf{I}_k$.
  Compute the semantic mask $\mathbf{M}_d^k$ via Eqs. (5)–(7).
  Compute the texture anomaly mask $\mathbf{M}_a^k$ via Eqs. (8)–(10).
  Project $\mathbf{M}_d^k$ and $\mathbf{M}_a^k$ back to the original image.
end for
Fuse the two masks: $\widehat{\mathbf{M}}_{\text{camou}} \leftarrow \Gamma(\mathbf{M}_d \odot \mathbf{M}_a)$.
Remove the adversarial camouflage: $\mathbf{I}_{\text{blk}} \leftarrow (1 - \widehat{\mathbf{M}}_{\text{camou}}) \odot \mathbf{I}_{\text{adv}}$.
Obtain connected components $\{\mathbf{M}_i\}_{i=1}^N$.
for $i = 1$ to N **do**
  Extract the padded local image $\mathbf{I}_i$ and local mask $\mathbf{M}_i^c$.
  Construct the geometry condition $\mathbf{C}_i$.
  Reconstruct the local texture $\widetilde{\mathbf{I}}_i$.
  Construct the soft blending mask $\boldsymbol{\alpha}_i \leftarrow \mathcal{G}_\sigma * \mathbf{M}_i^c$.
  Blend the reconstructed texture $\widetilde{\mathbf{I}}_i$.
  Paste $\widetilde{\mathbf{I}}_i$ back into $\widetilde{\mathbf{I}}$.
end for
return $\widetilde{\mathbf{I}}$.

5 Experiment

5.1 Experimental Setup

Dataset. We constructed a physical-scene evaluation dataset in the CARLA Dosovitskiy et al. (2017) simulation environment, covering 36 viewing angles, 20 weather conditions, and viewing distances ranging from $5 \sim 15$ m. The dataset contains adversarial textures generated by four attack methods against six monocular depth estimation models. For each attack–model combination, we collected 500 images, resulting in a total of $12,000$ images for evaluation.

MDE Models. We select six representative MDE models covering three different architectures: (1) *CNN-based*: monodepth2 Godard et al. (2019), depth-hints Watson et al. (2019), and many-depth Watson et al. (2021). (2) *Transformer-based*: DepthAnythingV2 Yang et al. (2024). (3) *Robustness-based*: MIMDepth Chawla et al. (2023) and Robustdepth Saunders et al. (2023).

Attack Methods. We evaluate four adversarial camouflage attacks. 3D^{2}Fool Zheng et al. (2024) is the first 3D texture-based attack against MDE and improves physical robustness through texture transformation and weather-related augmentation. To compare with representative patch-based attacks, we extend APA Yamanaka et al. (2020) and SSAP Guesmi et al. (2024) to full-surface UV texture optimization using differentiable rendering, denoted as 3D-APA and 3D-SSAP, respectively. RobCamou is our proposed method, which combines complex physical augmentation with conflict-aware gradient aggregation to improve physical robustness and stabilize multi-view texture optimization. All methods were optimized for each MDE model and evaluated under the same CARLA environment configuration.

Baseline Defenses. We compare three attack-agnostic purification defenses. Jedi Tarchoun et al. (2023) localizes suspicious regions using entropy analysis and reconstructs them with an autoencoder. PAD Jing et al. (2024) is a training-free, patch-agnostic method that detects and removes patches by exploiting semantic independence and spatial heterogeneity. APEX Wang et al. (2026) adopts a zero-shot localization pipeline followed by structure-guided image inpainting. All defenses are applied as preprocessing modules before MDE inference to assess whether methods originally designed for imperceptible perturbations or localized patches can generalize to full-surface adversarial camouflage.

Evaluation Metrics. We evaluate the effectiveness of the defense in three metrics: global depth discrepancy $\mathcal{E}_{\mathbf{I}}$, vehicle-region depth discrepancy $\mathcal{E}_{\mathbf{v}}$, and background depth discrepancy $\mathcal{E}_{\mathbf{b}}$. The

MDE	Defense	Clean	3D^{2}Fool			3D-APA			3D-SSAP			RobCamou		
			$\mathcal{E}_{\mathbf{I}}\downarrow$	$\mathcal{E}_{\mathbf{V}}\downarrow$	$\mathcal{E}_{b}\downarrow$	$\mathcal{E}_{\mathbf{I}}\downarrow$	$\mathcal{E}_{\mathbf{V}}\downarrow$	$\mathcal{E}_{b}\downarrow$	$\mathcal{E}_{\mathbf{I}}\downarrow$	$\mathcal{E}_{\mathbf{V}}\downarrow$	$\mathcal{E}_{b}\downarrow$	$\mathcal{E}_{\mathbf{I}}\downarrow$	$\mathcal{E}_{\mathbf{V}}\downarrow$	$\mathcal{E}_{b}\downarrow$
monodepth2	*No Defense*	-	*1.095*	*5.336*	*0.556*	*1.951*	*25.986*	*0.508*	*2.275*	*26.219*	*0.531*	*2.918*	*33.274*	*0.674*
	Jedi *(CVPR 23')*	**0.090**	0.829	3.508	0.499	0.532	3.580	0.486	0.709	4.413	0.505	0.729	5.637	0.482
	PAD *(CVPR 24')*	0.530	0.749	1.649	0.697	0.725	3.915	0.725	0.909	5.975	0.716	1.337	10.656	0.806
	APEX *(USENIX 26')*	0.345	1.042	4.110	0.700	1.483	13.815	0.598	1.993	16.724	0.660	1.993	16.288	0.711
	CamouPurifier†	0.553	0.592	**1.016**	0.627	0.734	**1.291**	0.769	0.608	**1.178**	0.642	**0.608**	**1.321**	0.739
	CamouPurifier‡	0.172	**0.497**	1.447	**0.450**	**0.429**	2.009	**0.432**	**0.446**	2.054	**0.442**	0.681	4.498	**0.477**
depth-hints	*No Defense*	-	*0.416*	*2.092*	*0.368*	*1.554*	*6.297*	*0.732*	*1.961*	*17.558*	*0.449*	*2.136*	*22.202*	*0.765*
	Jedi *(CVPR 23')*	**0.096**	0.404	1.333	0.457	0.718	3.918	0.696	0.521	2.628	0.491	0.625	5.094	0.444
	PAD *(CVPR 24')*	0.515	0.618	1.649	0.648	0.798	2.372	0.724	0.853	4.185	0.735	1.368	12.960	0.824
	APEX *(USENIX 26')*	0.269	0.431	1.923	0.412	1.204	6.359	0.792	1.317	10.011	0.474	1.341	13.972	0.577
	CamouPurifier†	0.664	0.766	**1.041**	0.856	**0.673**	**1.782**	0.752	0.779	**0.925**	0.874	1.159	**1.346**	1.250
	CamouPurifier‡	0.165	**0.385**	1.901	**0.385**	0.842	4.377	**0.567**	**0.370**	1.599	**0.384**	**0.484**	2.983	**0.402**
manydepth	*No Defense*	-	*0.369*	*1.453*	*0.380*	*1.657*	*9.972*	*1.209*	*1.369*	*12.938*	*0.588*	*1.639*	*11.845*	*1.194*
	Jedi *(CVPR 23')*	**0.106**	**0.467**	1.306	0.625	0.834	4.509	0.701	0.759	2.468	**0.532**	**0.728**	4.754	0.768
	PAD *(CVPR 24')*	0.545	0.653	1.363	0.710	0.977	2.544	0.941	0.874	2.567	0.854	1.639	2.840	0.957
	APEX *(USENIX 26')*	0.307	0.538	1.969	0.641	1.498	9.310	1.110	1.235	9.832	0.706	1.446	9.989	1.126
	CamouPurifier†	0.632	0.604	**1.244**	0.657	**0.650**	**1.275**	0.669	**0.586**	**1.050**	0.619	0.838	**1.391**	0.857
	CamouPurifier‡	0.235	0.554	1.310	**0.610**	0.796	4.515	**0.632**	0.657	3.492	0.583	0.872	5.196	**0.751**
DepthAny-thingv2	*No Defense*	-	*0.755*	*0.957*	*0.834*	*5.963*	*6.811*	*6.117*	*11.393*	*10.458*	*11.677*	*1.403*	*5.718*	*1.335*
	Jedi *(CVPR 23')*	**0.157**	0.779	1.132	0.849	1.115	1.510	1.289	1.577	1.539	1.688	0.903	2.416	0.921
	PAD *(CVPR 24')*	1.067	1.974	1.486	2.155	2.680	3.410	2.873	3.102	2.294	3.371	2.138	5.020	2.211
	APEX *(USENIX 26')*	0.612	1.167	2.295	1.240	4.262	5.962	4.373	7.970	8.119	8.222	1.561	5.016	1.526
	CamouPurifier†	1.052	1.219	1.325	1.305	1.394	1.747	1.488	1.396	1.162	1.516	1.254	1.525	1.335
	CamouPurifier‡	0.342	**0.748**	**0.940**	**0.824**	**1.047**	**1.273**	**1.119**	**1.050**	**0.980**	**1.157**	**0.780**	**1.478**	**0.837**
MIMDepth	*No Defense*	-	*3.134*	*2.860*	*3.544*	*3.916*	*13.107*	*4.408*	*3.836*	*7.483*	*4.155*	*4.174*	*14.979*	*4.326*
	Jedi *(CVPR 23')*	**0.582**	3.472	2.434	3.708	3.752	7.342	3.909	3.949	4.289	4.211	3.972	7.380	**4.030**
	PAD *(CVPR 24')*	3.126	5.046	**2.018**	5.420	5.126	5.330	5.465	5.469	3.918	5.912	5.232	5.677	5.529
	APEX *(USENIX 26')*	2.348	4.001	4.013	4.250	4.785	13.495	4.899	4.159	8.268	4.431	4.985	14.280	5.100
	CamouPurifier†	3.153	3.771	3.164	4.037	4.137	**4.558**	4.393	4.221	**3.629**	4.516	4.186	**4.376**	4.395
	CamouPurifier‡	1.311	**3.209**	2.573	**3.426**	**3.655**	6.540	**3.785**	**3.209**	4.179	**3.815**	**3.881**	6.915	**4.030**

Table 1: Defense performance comparison across different MDE models and adversarial camouflage attacks. Bold is the best, and underline is the second best. Due to space limitations, the results on RobustDepth are in the Supplementary Materials.

lower the values of these three metrics, the better the defense. In addition, to evaluate the camouflage localization accuracy, we use the Precision, Recall, F1 score and the IoU metric of the located areas to the ground truth camouflage region. Higher values of the F1 and IoU indicate better performance. Details are shown in the Supplementary Materials.

Implementation details. For CamouPurifier, the $\theta = 0.15$, texture anomaly score use weights $\lambda_{\text{bil}} = 0.40$, $\lambda_{\text{lap}} = 0.25$, $\lambda_{\text{std}} = 0.20$, $\lambda_{\text{sat}} = 0.15$. $q_a = 70$, $q_d = 30$. min area is 200 pixel. For stable diffusion, $T = 40$, inpaint size is $512 \times 512 \times 3$, and max componet is 3. We denote the method without TR as CamouPurifier†, and with TR as CamouPurifier‡. All experiments are conducted on an A100 GPU, and the dataset is generated from CARLA v0.9.15, which is deployed on one NVIDIA GeForce RTX 3090-24G.

5.2 Main Result

We report the performance over six MDE models and four adversarial camouflage attacks. For CamouPurifier, we report both CamouPurifier† and CamouPurifier‡ defense results.

Overall Defense Performance. As shown in Table 4, CamouPurifier† achieves an average vehicle-region discrepancy $\mathcal{E}_{\mathbf{V}}$ of 1.683, compared with 10.875 without defense, corresponding to an error reduction of 84.5% of all the models and attacks. It also reduces $\mathcal{E}_{\mathbf{V}}$ by 51.2% relative to Jedi, which is the strongest baseline on average. Meanwhile, CamouPurifier‡ achieves the lowest average global and background discrepancies, reducing $\mathcal{E}_{\mathbf{I}}$ and $\mathcal{E}_{b}$ by 53.7% and 41.2%, respectively, compared with the undefended setting. Compared with Jedi, these two discrepancies are further reduced by 8.8% and 5.8%. For different attacks, CamouPurifier consistently suppresses the vehicle-region depth errors caused by all four attacks. For 3D^{2}Fool, it reduces $\mathcal{E}_{\mathbf{V}}$ by 40.5% relative to no defense and by 9.2% relative to the strongest baseline. The improvement becomes more pronounced under stronger full-surface attacks. Specifically, the vehicle-region discrepancy

is reduced by 83.3%, 89.3%, and 88.1% against 3D-APA, 3D-SSAP, and RobCamou, respectively, compared with the undefended setting. Relative to the strongest baseline for each attack, the corresponding reductions are 44.9%, 54.1%, and 60.3%. In particular, under RobCamou, CamouPurifier decreases the average $\mathcal{E}_{\mathbf{V}}$ from 15.578 to 1.850, demonstrating that the proposed localization and geometry-guided reconstruction remain effective against physically robust and multi-view optimized camouflage.

In addition, we compared the inference efficiency of different defense methods. As shown in Table 2, CamouPurifier† achieved the highest processing speed of 2.705 FPS, making it more suitable for latency-sensitive scenarios. In contrast, although CamouPurifier‡ incurs additional computational overhead, it performs more comprehensive purification and is therefore recommended for applications that prioritize restoration accuracy and overall depth quality over inference speed. These two modules offer a flexible trade-off between computational efficiency and defense performance. For a detailed analysis, such as quality of purified images, robustness under different weather, generalization across MDE Models, and defense against adaptive attacks, please refer to Appendix in the Supplementary Material.

Method	Jedi	PAD	APEX	**CamouPurifier†**	**CamouPurifier‡**
FPS ↑	1.717	0.02	0.026	2.705	0.098

Table 2: Efficiency comparison of different methods.

5.3 Camouflage-Region Detection

Figure 3 compares the pixel-level camouflage localization performance of different methods using Precision, Recall, F1, and IoU. CamouPurifier achieves the best overall performance, obtaining approximately 62.95% Precision, 70.8% Recall, 65.18% F1, and 49.51% IoU. In particular, its high Recall indicates that most adversarial camouflage regions can be successfully identified, while the competitive Precision shows that this improvement is not achieved by excessively expanding the predicted masks. The comparison also reveals different limitations of the existing methods. Jedi achieves relatively high Precision but substantially lower Recall, indicating a conservative localization behavior that misses a considerable portion of the camouflage texture. PAD obtains a Recall comparable to CamouPurifier, but its much lower Precision suggests severe over-localization, where benign vehicle or background pixels are incorrectly included in the predicted masks. APEX performs poorly in Recall 41.12%, F1$20.68\%$, and IoU26.12%, showing limited capability in capturing full-surface adversarial camouflage. Overall, CamouPurifier provides the most balanced localization results and achieves clear improvements in both F1 and IoU.

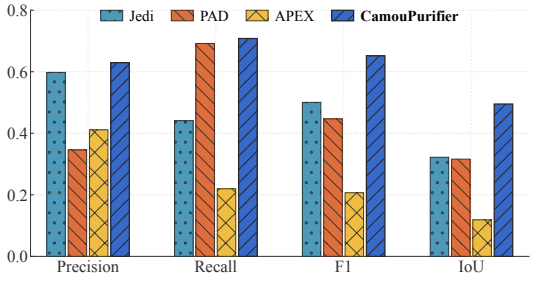


Figure 3: Pixel-level camouflage-region localization performance against the ground-truth texture masks.

5.4 Generalization across Different Vehicles

Figure 4 evaluates CamouPurifier‡ on three vehicles with distinct appearances and geometries: Lincoln, Audi e-tron, and Cybertruck. Adversarial camouflage causes severe depth distortions across all types of vehicles, leading to incomplete contours and inconsistent depth responses. After purification, CamouPurifier‡ effectively suppresses the adversarial textures and restores more continuous and structurally coherent vehicle shapes. The consistent improvements across sedans, SUVs, and vehicles with highly angular geometry demonstrate that CamouPurifier‡ is not restricted to a specific vehicle model and generalizes well across diverse body shapes and surface structures.



Figure 4: Evaluation of CamouPurifier on vehicles with different body shapes and geometric structures.

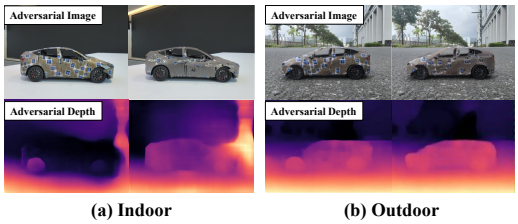


Figure 5: Physical-world evaluation of CamouPurifier in indoor and outdoor scenes.

5.5 Physical Case Study

To further evaluate the practical effectiveness of CamouPurifier, we conduct physical-world experiments using a vehicle model covered with printed adversarial camouflage. As shown in Figure. 5, the camouflage causes substantial depth distortions in both indoor and outdoor scenes, resulting in incomplete vehicle shapes and abnormal depth responses. After purification, CamouPurifier effectively suppresses the adversarial texture and recovers more continuous and structurally coherent depth predictions. Specifically, it reduces the average depth error from 26.53 to 3.57, corresponding to an error reduction of 86.5%. These results demonstrate that CamouPurifier generalizes remains effective under variations in illumination and scene context.

5.6 Ablation Study

We evaluate the contributions of SAL, MTAL, and TR in Table 3. SAL and MTAL individually reduce the $\mathcal{E}_{\mathbf{V}}$ by 94.6% and 93.6%, respectively, compared with no defense, while SAL achieves better overall performance. Combining SAL and MTAL further decreases $\mathcal{E}_{\mathbf{I}}$ and $\mathcal{E}_{\mathbf{V}}$ to 0.665 and 1.202, demonstrating the complementarity of semantic and texture information. Adding TR to SAL reduces $\mathcal{E}_{\mathbf{I}}$, $\mathcal{E}_{\mathbf{V}}$, and $\mathcal{E}_b$ by 24.9%, 7.7%, and 26.3%, respectively. The complete model achieves the lowest clean-image, global, and background errors of 0.172, 0.513, and 0.450. Compared with no defense, these results correspond to reductions of 75.1% in $\mathcal{E}_{\mathbf{I}}$ and 20.6% in $\mathcal{E}_b$. Although its vehicle-region error is higher than that of SAL+TR, the full model provides a better balance between attack removal and global structural preservation.

SAL	MTAL	TR	Clean	$\mathcal{E}_{\mathbf{I}} \downarrow$	$\mathcal{E}_{\mathbf{V}} \downarrow$	$\mathcal{E}_b \downarrow$
No Defense			-	2.060	22.704	0.567
✔			0.674	0.732	1.227	0.767
	✔		0.827	0.926	1.452	0.948
✔	✔		0.553	0.665	1.202	0.694
✔		✔	0.323	0.550	**1.132**	0.565
	✔	✔	0.231	0.515	2.532	0.459
✔	✔	✔	**0.172**	**0.513**	2.502	**0.450**

Table 3: Ablation study of SAL, MTAL, and TR. Lower values indicate better performance.

6 Conclusion

In this work, we proposed CamouPurifier, a zero-shot and camouflage-agnostic defense framework. CamouPurifier combines semantic-aware localization and multi-domain texture anomaly localization to identify distributed camouflage regions, and further employs geometry-guided diffusion reconstruction to remove adversarial textures while preserving depth-relevant vehicle structures. Extensive experiments across different attacks, MDE models, viewing conditions, and physical scenes demonstrate the robustness, and generalizability of the proposed framework. Future work will focus on improving reconstruction efficiency and enhancing robustness against adaptive adversarial camouflage attacks.

AI USE STATEMENT

Large Language Models (LLMs) were employed to assist with manuscript writing and polishing. Specifically, an LLM was used to refine language expressions, enhance readability, and clarify the presentation across different sections of this paper. Its support included sentence rephrasing, grammar correction, and optimization of textual coherence. Notably, the LLM did not participate in research ideation, methodology formulation, or experimental design. All research concepts, ideas and data analyses were conceived and performed by the authors. The LLM's contribution was limited exclusively to linguistic improvements of the manuscript; it was not involved in any scientific reasoning or data analysis.

ETHICS STATEMENT

This paper describes a research effort aimed at enhancing the safety of autonomous driving (AD) systems. The proposed method for filtering out adversarial textures helps improve the robustness of autonomous driving systems, thereby ensuring their safety.

REFERENCES

John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

Hemang Chawla, Kishaan Jeeveswaran, Elahe Arani, and Bahram Zonooz. Image masking for robust self-supervised monocular depth estimation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pp. 10054–10060. IEEE, 2023.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 3827–3837. IEEE, 2019.

Amira Guesmi, Muhammad Abdullah Hanif, Ihsen Alouani, Bassem Ouni, and Muhammad Shafique. SSAP: A shape-sensitive adversarial patch for comprehensive disruption of monocular depth estimation in autonomous navigation applications. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, pp. 2786–2793. IEEE, 2024.

Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, and Cong Zou. PAD: patch-agnostic defense against adversarial patch attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 24472–24481. IEEE, 2024.

Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yingzhe He, Jie Shi, and Xiaolin Hu. ADBM: adversarial diffusion bridge model for reliable adversarial purification. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Guang Lin, Zerui Tao, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Adversarial guided diffusion models for adversarial purification. *Neural Networks*, 191:107705, 2025.

Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14953–14962. IEEE, 2022.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Kieran Saunders, George Vogiatzis, and Luis J. Manso. Self-supervised monocular depth estimation: Let's talk about the weather. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 8873–8883. IEEE, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yibo Sun, Yucheng Shi, Lei Shi, Lin Wei, Yufei Gao, Qiushi Li, and Wenwen Li. Breaking monocular depth estimation with depthhack: A black-box 3d physical adversarial attack for autonomous driving. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 5920–5926. IEEE, 2025.

Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. DTA: physical camouflage attacks using differentiable transformation network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15284–15293. IEEE, 2022.

Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael B. Abu-Ghazaleh, and Ihsen Alouani. Jedi: Entropy-based localization and removal of adversarial patches. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 4087–4095. IEEE, 2023.

Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pp. 1–6. IEEE, 2024.

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 *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 2414–2422. AAAI Press, 2022.

Jiayimei Wang, Tao Ni, Guowen Xu, Qingchuan Zhao, and Cong Wang. Adversarial patch exterminator: Zero-shot and patch-agnostic defense framework against adversarial patch attacks. In *35th USENIX Security Symposium (USENIX Security 26)*, 2026.

Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 2162–2171. IEEE, 2019.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel J. Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 1164–1174. Computer Vision Foundation / IEEE, 2021.

Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Proc. Int. Conf. Learn. Represent.*, 33:8486–8497, 2020.

Chong Xiang, Alexander Valtchanov, Saeed Mahloujifar, and Prateek Mittal. Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pp. 1329–1347. IEEE, 2023.

Koichiro Yamanaka, Ryutaroh Matsumoto, Keita Takahashi, and Toshiaki Fujii. Adversarial patch attacks on monocular depth estimation networks. *IEEE Access*, 8:179094–179104, 2020.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Yixun Zhang, Feng Zhou, and Jianqin Yin. Bitaa: A bi-task adversarial attack on 3d gaussian fields for cross-task perception. *IEEE Signal Processing Letters*, 2026.

Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, and Chao Shen. Physical 3d adversarial attacks against monocular depth estimation in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 24452–24461. IEEE, 2024.

Jiawei Zhou, Linye Lyu, Daojing He, and Yu Li. RAUCA: A novel physical adversarial attack on vehicle detectors via robust and accurate camouflage generation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235, pp. 62076–62087. PMLR / OpenReview.net, 2024.

A Appendix

A.1 Dataset

Evaluation dataset. We construct our evaluation dataset from the CARLA Dosovitskiy et al. (2017) simulator, which contains $1,043$ clean driving scenes rendered at a resolution of 1024×320. The scenes cover diverse camera viewpoints, vehicle distances, illumination conditions, and weather conditions. Each image is associated with a pixel-level vehicle mask generated by the simulator. We evaluate six representative monocular depth estimation models: monodepth2 Godard et al. (2019), depth-hints Watson et al. (2019), manydepth Watson et al. (2021), DepthAnythingV2 Yang et al. (2024), RobustDepth Saunders et al. (2023), and MIMDepth Chawla et al. (2023).

For each victim MDE model, we consider four physical adversarial camouflage attacks: 3D^{2}Fool Zheng et al. (2024), 3D-APA Yamanaka et al. (2020), 3D-SSAP Guesmi et al. (2024), and RobCamou. Clean and adversarial images are aligned by scene identifier, and the clean vehicle mask is used consistently when computing regional depth errors. To avoid diluting the evaluation with ineffective attack instances, we rank the $1{,}043$ candidates for each model–attack pair according to their undefended vehicle-region depth error $\mathcal{E}_{\mathbf{V}}$ and retain the strongest 500. The ranking is performed before applying any defense and is therefore independent of the evaluated defense method. This results in 500 samples for each of the 6×4 model–attack combinations, totally $12,000$ scene images.

Clean evaluation set. We additionally construct a clean set to quantify the benign distortion introduced by each defense. Specifically, we use the 500 clean counterparts of the monodepth2–3D^{2}Fool evaluation split and apply the same set of scene identifiers to all six MDE models. For this evaluation, the defense input and reference image are identical before processing. Consequently, any non-zero depth error reflects distortion introduced solely by the defense rather than by an adversarial attack. This protocol enables a direct comparison between attack suppression and clean-image preservation.

Adaptive-attack set. For adaptive evaluation, we generate an additional $1{,}043$ adversarial images using a Lexus vehicle and monodepth2 as the victim model. The adaptive objective jointly preserves the depth attack and suppresses the localization response of CamouPurifier. These images cover vehicle distances from approximately 5 to 15 m and retain the same scene, weather, and viewpoint diversity as the clean CARLA-MDE data. Each adversarial image is paired with its stem-aligned clean image and simulator-derived clean vehicle mask. For reporting adaptive robustness, we recompute the undefended $\mathcal{E}_{\mathbf{V}}$ and select the strongest 500 samples. We then report the defended errors over the vehicle region ($\mathcal{E}_{\mathbf{V}}$), full image ($\mathcal{E}_{\mathbf{I}}$), and non-vehicle background ($\mathcal{E}_b$).

MDE	Defense	Clean	3D^{2}Fool			3D-APA			3D-SSAP			RobCamou		
			$\mathcal{E}_\mathbf{I}$	$\mathcal{E}_\mathbf{V}$	$\mathcal{E}_b$	$\mathcal{E}_\mathbf{I}$	$\mathcal{E}_\mathbf{V}$	$\mathcal{E}_b$	$\mathcal{E}_\mathbf{I}$	$\mathcal{E}_\mathbf{V}$	$\mathcal{E}_b$	$\mathcal{E}_\mathbf{I}$	$\mathcal{E}_\mathbf{V}$	$\mathcal{E}_b$
RobustDepth	*No Defense*	-	*1.037*	*1.923*	*0.682*	*2.490*	*8.029*	*1.269*	*2.354*	*8.048*	*1.162*	*1.961*	*5.448*	*1.206*
	Jedi *(CVPR 23')*	**0.081**	0.723	1.123	**0.558**	1.198	3.703	**0.883**	1.200	3.985	**0.863**	**0.856**	2.707	**0.773**
	PAD *(CVPR 24')*	0.394	0.899	1.411	0.881	1.291	3.715	1.071	1.304	3.821	1.065	1.050	3.045	0.942
	APEX *(USENIX 26')*	0.232	1.058	2.681	0.882	1.898	6.756	1.303	1.912	7.062	1.244	1.488	5.226	1.224
	CamouPurifier†	0.492	**0.718**	**0.903**	0.750	**0.885**	**1.086**	0.928	**0.983**	**0.919**	1.046	0.977	**1.139**	1.009
	CamouPurifier‡	0.169	0.783	1.534	0.734	1.224	3.111	1.010	1.008	2.491	0.868	0.974	2.657	0.880

Table 4: Defense performance on RobustDepth model and adversarial camouflage attacks. Lower values are better for all metrics. Bold and underlined values denote the best and second-best results among defense methods, respectively.

A.2 Attack Methods

RobCamou is our proposed method, which combines complex physical augmentation with conflict-aware gradient aggregation to improve physical robustness and stabilize multi-view texture optimization. All methods were optimized for each MDE model and evaluated under the same CARLA environment configuration.

A.3 Evaluation Metrics

We evaluate the effectiveness of the defense in three metrics: global depth discrepancy $\mathcal{E}_\mathbf{I}$, vehicle-region depth discrepancy $\mathcal{E}_\mathbf{v}$, and background depth discrepancy $\mathcal{E}_\mathbf{b}$.

$$\mathcal{E}_\mathbf{I} = \frac{sum(|\mathbf{D} - \widetilde{\mathbf{D}}|)}{sum(\mathbf{I})}, \tag{17}$$

$$\mathcal{E}_\mathbf{v} = \frac{sum(|\mathbf{D} - \widetilde{\mathbf{D}}| \odot \mathbf{M}_v)}{sum(\mathbf{M}_v)}, \tag{18}$$

$$\mathcal{E}_\mathbf{b} = \frac{sum(|\mathbf{D} - \widetilde{\mathbf{D}}| \odot \mathbf{M}_b)}{sum(\mathbf{M}_b)}, \tag{19}$$

where $\mathbf{D}$ is the true depth, $\widetilde{\mathbf{D}}$ is the purified depth. $\mathbf{M}_v$ is the mask of the vehicle, $\mathbf{M}_b$ is the background mask, $\mathbf{M}_b = 1 - \mathbf{M}_v$. The lower the values of these three metrics, the better the defense.

To evaluate camouflage localization accuracy, we adopt pixel-level Precision ($\mathcal{P}$), Recall ($\mathcal{R}$), F1 score ($\mathcal{F}_1$), and intersection over union (IoU):

$$\mathcal{P} = \frac{\left\|\widehat{\mathbf{M}}_c \odot \mathbf{M}_c\right\|_1}{\left\|\widehat{\mathbf{M}}_c\right\|_1}, \tag{20}$$

$$\mathcal{R} = \frac{\left\|\widehat{\mathbf{M}}_c \odot \mathbf{M}_c\right\|_1}{\left\|\mathbf{M}_c\right\|_1}, \tag{21}$$

$$\mathcal{F}_1 = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}}, \tag{22}$$

$$\text{IoU} = \frac{\left\|\widehat{\mathbf{M}}_c \odot \mathbf{M}_c\right\|_1}{\left\|\widehat{\mathbf{M}}_c + \mathbf{M}_c - \widehat{\mathbf{M}}_c \odot \mathbf{M}_c\right\|_1}, \tag{23}$$

where $\widehat{\mathbf{M}}_c$ and $\mathbf{M}_c$ denote the predicted and ground-truth camouflage masks, respectively. Higher values indicate more accurate camouflage localization.

A.4 CamouPurifier configuration.

Unless otherwise stated, we use the same configuration across all attacks and MDE models. We first detect vehicle regions using YOLOv8-n Varghese & Sambath (2024) with a confidence threshold of

0.15, considering the COCO vehicle classes car, motorcycle, bus, and truck. Each detected bounding box is expanded by 8 pixels. If no vehicle is detected, the full image is used as the fallback region. The texture-deviation score combines bilateral residual, Laplacian response, local standard deviation, and saturation with weights 0.40, 0.25, 0.20, and 0.15, respectively. Each response is normalized using its 2nd and 98th percentiles. We threshold the combined score at the 70th percentile within each vehicle region, followed by 5×5 morphological closing and opening. Connected components smaller than 200 pixels are removed. For semantic localization, we use a DINOv3 ViT-S/16 model Siméoni et al. (2025) pretrained on LVD-1689M. Input crops are resized to 224×224, and the localization heatmap is computed from the cosine similarity between the CLS token and patch tokens. We use the 30th percentile of the DINO heatmap as the default threshold and intersect the resulting mask with the texture-deviation mask. Small enclosed holes of at most 500 pixels are filled, while large dark regions are retained to avoid modifying vehicle windows. We additionally apply a 3×3 morphological closing operation. For the adaptive-attack experiment, we lower the DINO percentile to 20 to increase localization sensitivity, while keeping all other parameters unchanged.

Appearance purification. We implement the appearance projection using Stable Diffusion Inpainting (`runwayml/stable-diffusion-inpainting`) with a Canny-conditioned ControlNet (`lllyasviel/control_v11p_sd15_canny`). Each localized component is expanded by 48 pixels and resized to 512×512. We process at most three components per image and discard components smaller than 150 pixels. Diffusion is performed for 40 steps with an inpainting strength of 0.55, classifier-free guidance scale of 8.5, ControlNet conditioning scale of 0.90, and random seed 42. The generated region is blended with the original image using a 7×7 Gaussian-smoothed mask. The positive prompt is:

> *"black car, Unreal Engine 4, CARLA style, solid silver metallic factory paint, clean body no pattern, natural glass windows and tires, soft daylight, highly detailed."*

The negative prompt is:

> *"camouflage, camouflage pattern, adversarial texture, dazzle paint, zebra stripes, graffiti, multicolor patches, sketch, cartoon, blurry, deformed car, black hole, silhouette."*

We denote the localization-only variant, which replaces localized pixels with black values without diffusion-based reconstruction, as CamouPurifier†. The complete localization-and-purification pipeline is denoted as CamouPurifier‡.

B Experiments Result

B.1 Additional Analysis of Defense Baselines

Results on RobustDepth. RobustDepth remains vulnerable to adversarial camouflage despite its improved robustness. Without defense, the vehicle region error $\mathcal{E}_{\mathbf{V}}$ ranges from 1.923 under 3D^{2}Fool to 8.048 under 3D-SSAP. CamouPurifier† achieves the lowest $\mathcal{E}_{\mathbf{V}}$ under all four attacks. It reduces the error to 0.903, 1.086, 0.919, and 1.139 under 3D^{2}Fool, 3D-APA, 3D-SSAP, and RobCamou, respectively. Compared with no defense, the corresponding reductions are 53.0%, 86.5%, 88.6%, and 79.1%. CamouPurifier† also achieves the lowest full image error under the first three attacks and remains competitive under RobCamou.

CamouPurifier‡ places greater emphasis on preserving unaffected image regions. Its vehicle region errors are higher than those of CamouPurifier†, but it achieves the second lowest clean error of 0.169 and maintains competitive background errors. The background errors are 0.734, 0.868, and 0.880 under 3D^{2}Fool, 3D-SSAP, and RobCamou, respectively. In contrast, JPEG compression and Gaussian filtering frequently increase both the vehicle region error and the full image error on RobustDepth. These results show that the proposed localization based defense remains effective when applied to a robust MDE model, while uniform input transformations provide limited and inconsistent protection.

Generalization across MDE Models. CamouPurifier provides consistent vehicle-region protection across different MDE models. Compared with no defense, it reduces the average $\mathcal{E}_{\mathbf{V}}$ by 94.7%,

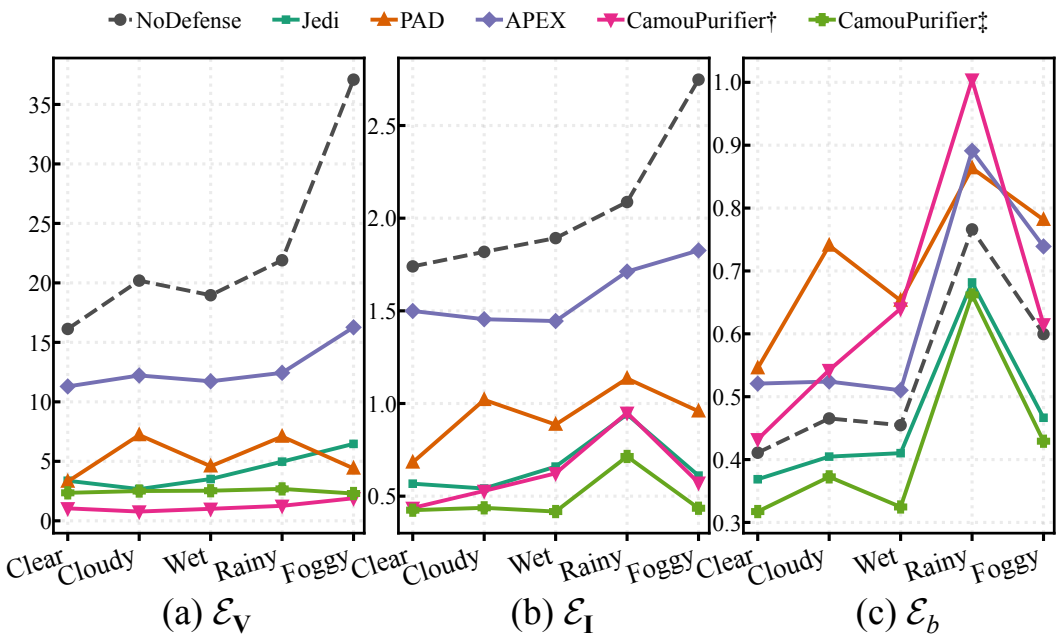


Figure 6: Defense performance under different weather conditions. Lower values indicate better performance.

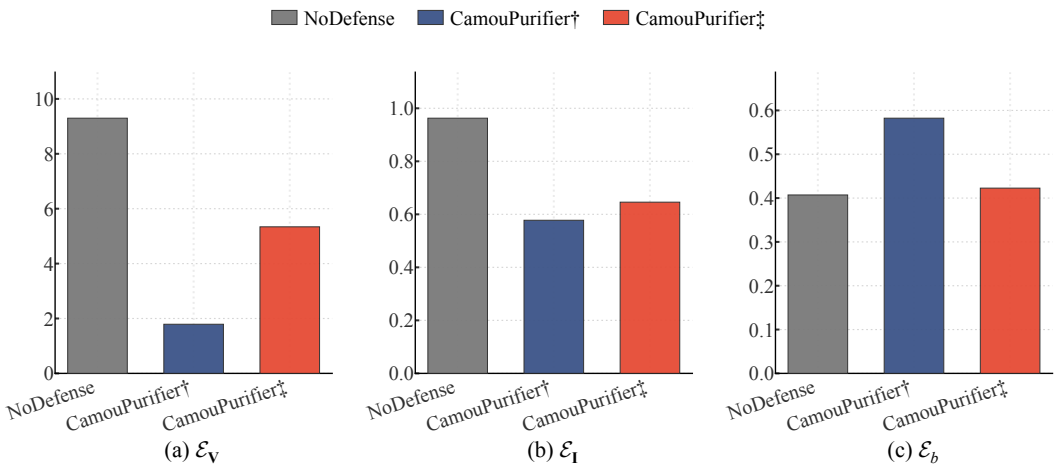


Figure 7: Defense performance against the adaptive attack on the top 500 samples ranked by the undefended vehicle-region error. We report the error over the vehicle region ($\mathcal{E}_{\mathbf{V}}$), full image ($\mathcal{E}_{\mathbf{I}}$), and non-vehicle background ($\mathcal{E}_b$). CamouPurifier† denotes localization-only intervention, and CamouPurifier‡ denotes the complete locate-and-purify pipeline. Lower values indicate better performance.

89.4%, and 86.3% on monodepth2, depth-hints, and manydepth, respectively. The reductions remain substantial on DepthAnythingv2, reaching 75.9%. Compared with the strongest baseline for each model, CamouPurifier further reduces the vehicle-region discrepancy by 72.0% on monodepth2, 60.7% on depth-hints, 46.7% on manydepth, and 12.7% on DepthAnythingv2. Although MIMDepth remains the most challenging model, CamouPurifier still reduces its vehicle-region discrepancy by 59.1% relative to no defense and by 7.2% relative to the strongest baseline.

Defense–Fidelity Trade-off. The two CamouPurifier configurations exhibit complementary operating characteristics. The standard CamouPurifier configuration achieves a 42.3% lower average

Method	Full Image					Vehicle Region		
	PSNR↑	SSIM↑	MSE↓	MAE↓	LPIPS↓	PSNR_{veh} ↑	SSIM_{veh} ↑	$\text{LPIPS}_{\text{veh}}$ ↓
Jedi	26.09	**0.8840**	0.003165	0.018075	0.0936	15.65	**0.5684**	0.0504
PAD	18.06	0.8370	0.021408	0.044944	0.1373	8.08	0.2788	0.0696
APEX	25.14	0.8581	0.003783	0.022990	0.1105	14.95	0.5295	0.0540
Ours	**26.45**	0.8805	**0.002867**	**0.018056**	**0.0920**	**15.90**	0.5309	**0.0489**

Table 5: Image quality comparison of different defense methods. ↑ and ↓ indicate that higher and lower values are better, respectively. The best results are highlighted in bold.

vehicle-region discrepancy than CamouPurifier*, indicating stronger removal of adversarial vehicle textures. In contrast, CamouPurifier* reduces the global discrepancy, background discrepancy, and clean-image discrepancy by 15.3%, 20.6%, and 63.4%, respectively, relative to the standard configuration. Its clean-image discrepancy is also 61.2% lower than PAD and 41.8% lower than APEX, although Jedi retains the lowest clean-image discrepancy. These results indicate that the standard configuration prioritizes aggressive vehicle-region purification, whereas CamouPurifier* provides a more balanced trade-off between adversarial removal and clean-image fidelity.

B.2 Robustness under Different Weather

Figure 6 evaluates the defense performance under five representative weather conditions, including Clear, Cloudy, Wet, Rainy, and Foggy scenes. In terms of vehicle-region discrepancy $\mathcal{E}_{\mathbf{V}}$, CamouPurifier† achieves the lowest error under all five conditions, reducing the average error from approximately 22.8 without defense to 1.18, corresponding to a reduction of 94.8%. Compared with Jedi, the strongest baseline on average, it further reduces $\mathcal{E}_{\mathbf{V}}$ by approximately 71.5%. The improvement is particularly evident under foggy conditions, where the undefended error increases to approximately 37.0, whereas CamouPurifier maintains an error of only 1.8, achieving a reduction of 95.1%. After introducing texture reconstruction, CamouPurifier‡ provides better global and background consistency. Specifically, CamouPurifier‡ reduces the average global discrepancy $\mathcal{E}_{\mathbf{I}}$ by approximately 76.5% relative to no defense and by 28.5% relative to Jedi. It also reduces the average background discrepancy $\mathcal{E}_b$ by approximately 21.6% and 9.2%, respectively. Although rainy weather causes higher global and background errors for all methods due to degraded visibility and complex reflections, CamouPurifier still achieves the lowest $\mathcal{E}_{\mathbf{I}}$ and $\mathcal{E}_b$. These results demonstrate that CamouPurifier remains effective under diverse weather disturbances. These results demonstrate that CamouPurifier remains effective under diverse weather disturbances, while texture reconstruction improves global structural consistency and background preservation at the cost of a slight increase in vehicle-region discrepancy.

B.3 Defense Against Adaptive Attack

We further evaluate CamouPurifier under an adaptive attack that jointly preserves the depth estimation effect and weakens the response of the localization module. Following the main evaluation protocol, we select the 500 samples with the largest undefended vehicle region errors. Figure 7 reports the vehicle region error $\mathcal{E}_{\mathbf{V}}$, full image error $\mathcal{E}_{\mathbf{I}}$, and background error $\mathcal{E}_{\mathbf{b}}$. CamouPurifier† applies direct intervention to the localized regions, while CamouPurifier‡ uses the complete localization and purification pipeline.

Both variants remain effective despite the attack being optimized against the localization mechanism. CamouPurifier† reduces $\mathcal{E}_{\mathbf{V}}$ from 9.30 to 1.79, yielding a reduction of 80.7%. CamouPurifier‡ reduces the error to 5.34, corresponding to a reduction of 42.6%. The full image error also decreases from 0.96 to 0.58 and 0.65, respectively. The background error increases moderately because part of the localization response extends beyond the vehicle region and the defense modifies the input image. Nevertheless, it remains far below the undefended vehicle region error. CamouPurifier† performs better in this setting, indicating that direct intervention on the localized regions is more effective against the adaptive attack, while reconstruction may retain some depth discrepancies. These results show that CamouPurifier remains effective even when the attacker has access to its localization strategy.

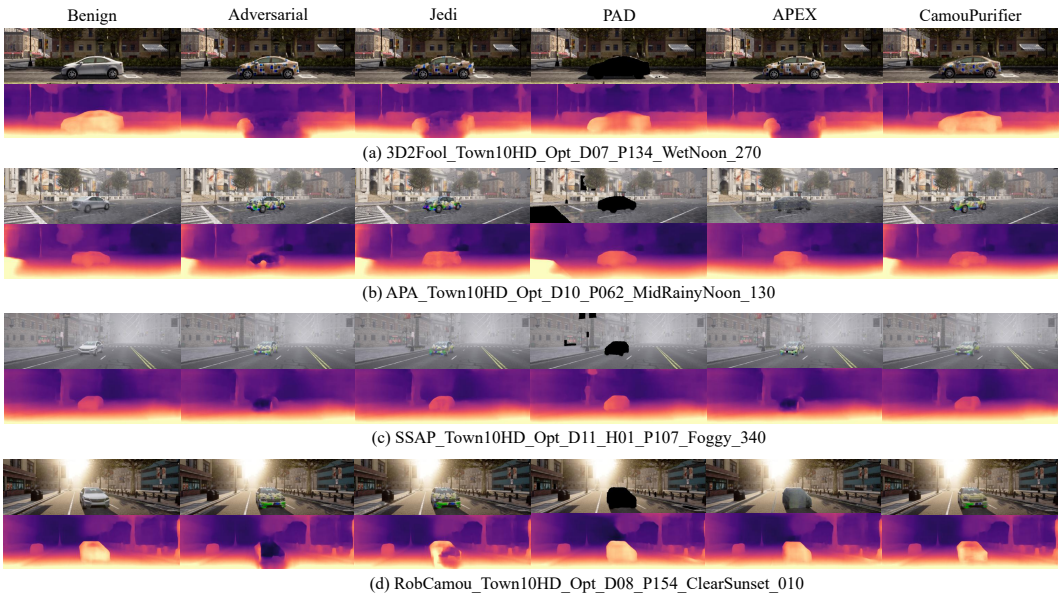


Figure 8: Visualizations of defenses under different attack methods and attack scenarios. The subtitles represent Attack_Town_Distance_Location_Weather_Angle, respectively.

B.4 Quality of Purified Images

Beyond depth restoration, we evaluate the visual quality of the purified images against their clean counterparts. Table 5 reports metrics over both the full image and the vehicle region. CamouPurifier achieves the best results on six out of eight metrics. For the full image, it obtains the highest PSNR of 26.45 dB and the lowest MSE, MAE, and LPIPS. Within the vehicle region, where most modifications are applied, it also achieves the highest PSNR of 15.90 dB and the lowest LPIPS of 0.0489.

Jedi achieves the highest SSIM for both evaluation regions. However, its full image SSIM of 0.8840 is only slightly higher than the 0.8805 achieved by CamouPurifier, while CamouPurifier performs better on all other full image metrics. It also obtains a lower LPIPS within the vehicle region, showing better perceptual similarity to the clean image after camouflage removal. In contrast, PAD introduces substantial visual distortion, as reflected by its lower PSNR and higher MSE, MAE, and LPIPS. These results show that CamouPurifier suppresses adversarial camouflage while preserving the visual content of the input.

B.5 Visualization

We randomly sample several defense results covering different attack methods, weather conditions, viewing distances, and camera angles. As shown in Figure. 8, Jedi partially suppresses the adversarial camouflage and restores some depth responses, but noticeable depth distortions remain around the vehicle. PAD generally recovers a recognizable vehicle shape in the depth map; however, it introduces large black regions in the purified image and frequently masks benign textured areas, such as road surfaces and billboards. Such artifacts may interfere with other downstream perception tasks. APEX preserves more image content but achieves relatively limited depth recovery, as it mainly removes salient textures on the vehicle surface without explicitly preserving geometry-related cues. In contrast, CamouPurifier consistently localizes adversarial camouflage under diverse conditions while preserving vehicle contours and surrounding scene content. It consequently produces more complete and geometrically coherent depth predictions, with substantially fewer artifacts that could affect other perception tasks.