Phy-DRILL: 2.5D PHYSICAL DEPTH REVERSAL ILLUSION AGAINST MONOCULAR DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable monocular depth estimation (MDE) is crucial for the operational safety of autonomous systems, making it essential to evaluate its robustness in the physical world. Yet existing physical attacks, whether textures optimized on 2D patches or 3D meshes, alter only appearance and treat physical geometry as a fixed constraint. In this work, we reveal that parameterized geometry unlocks adversarial effects that appearance alone cannot reach. We exploit the **Depth Reversal Illusion** (DepthRI), a perceptual phenomenon in which protruding structures are visually misinterpreted as receding surfaces due to conflicting geometric and texture cues. We show that DepthRI deceives not only human vision but also SOTA MDE models and trace its effect to two factors: ❶ *the perspective structure defined by vanishing points, which grounds the illusion*, and ❷ *the texture gradient that activates it*. Guided by these findings, we propose *Phy*-DRILL (**Phy**sical **D**epth **R**eversal **ILL**usion), which constructs 2.5D adversarial patches grounded in perspective geometric relationships and further enhances their adversarial effectiveness through targeted surface texture optimization. Our approach causes physically protruding structures to be systematically misperceived as concave regions by SOTA MDE models. Across six MDE models, the experiments show that *Phy*-DRILL outperforms existing attacks in both digital and physical settings and transfers from a single surrogate model to five black-box models.

1 INTRODUCTION

Monocular Depth Estimation (MDE) (Ming et al., 2021; Eigen et al., 2014; Yang et al., 2024a) provides dense depth from a single RGB camera and is widely studied for robotics and other camera-based perception applications. As MDE is increasingly deployed in safety-critical applications, adversarial attacks have become an essential tool for exposing its vulnerabilities and informing the design of robust systems (Wei et al., 2024). Among existing studies, physical attacks are of the greatest practical importance for real-world system robustness. However, existing physical adversarial attacks against MDE models predominantly rely on 2D patches (Brown et al., 2017; Cheng et al., 2022; Yamanaka et al., 2020). Optimized for a specific projected appearance, these patches lack multi-view geometric consistency and degrade sharply as the viewpoint changes (Fig. 1 (c)). To address viewpoint degradation, full-coverage 3D texture attacks (Zhang et al., 2019; Zheng et al., 2024a; Wang et al., 2022) have recently emerged, but they essentially perform texture mapping on fixed meshes, which limits deployment flexibility. In both cases, the attacker remains in a passive role, constrained to adapt to the pre-defined geometry of an existing object (e.g., the surface of a car) and overlook a crucial insight: *physical geometry itself can be an adversarial variable*.

In this work, we take a substantially different approach by exploiting the **DepthRI (Depth Reversal Illusion)**. DepthRI is a well-documented perceptual phenomenon in human vision where protruding structures are misinterpreted as receding surfaces due to conflicting geometric and texture cues (Hill & Johnston, 2007; Gregory, 1997). This illusion arises from the inherent ambiguity in monocular depth perception: *the visual system relies heavily on perspective cues*, such as converging lines and texture gradients, *to infer depth*, and these cues can be deliberately manipulated to induce systematic misperception. Crucially, we discover that SOTA MDE models (Ranftl et al., 2022; 2021; Yang et al., 2024b; Bhat et al., 2023; Bochkovskii et al., 2024; Yin et al., 2023) are equally susceptible to DepthRI. Such an observation exposes a new attack surface grounded not in model-specific optimization artifacts, but in fundamental principles of visual perception.

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.

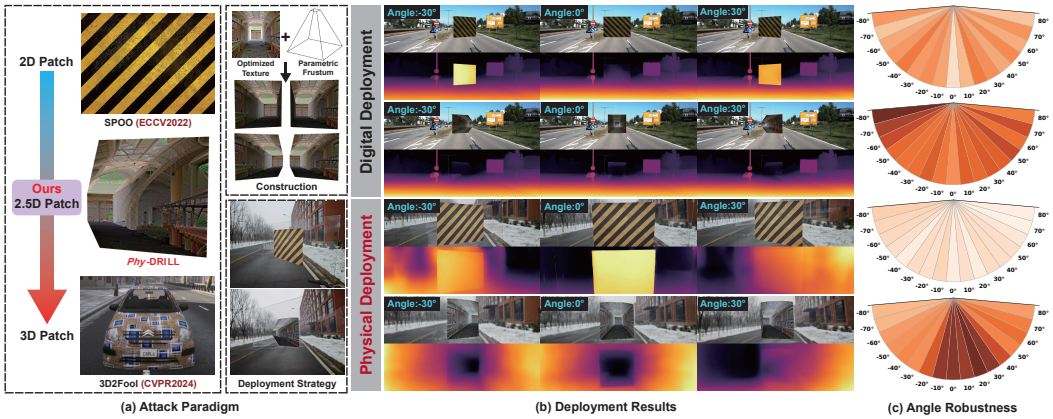

Figure 1: (a) Attack paradigm of our *Phy*-DRILL. Unlike existing 2D and 3D patches deployed on flat or complicated surfaces, our 2.5D patch consists of optimized texture and parametric geometry inherently capable of inducing depth reversal illusion. (b) Deployment results (DepthAnything v2) of our proposed *Phy*-DRILL (bottom rows) and SPOO baseline (top rows) across digital and physical environments. (c) Attack success rates across 17 angular conditions, where color intensity indicating the magnitude of the effect (higher intensity corresponds to a greater adversarial effect).

However, exploiting DepthRI to build a physically effective attack is a non-trivial task, as the geometric and texture configurations that sustain the illusion have never been characterized. We therefore conduct extensive empirical analyses to characterize DepthRI from the ground up. By systematically varying the perspective structure and the texture appearance of candidate objects, we isolate two governing factors: ❶ the number of vanishing points grounds the illusion, and ❷ the texture gradient activates it. These two factors place the attack between existing paradigms: a planar patch cannot supply the perspective structure, while a fully optimizable 3D mesh is neither necessary nor freely deployable. We thus introduce our designing as the **2.5D adversarial attack**, a parametric geometry carrying conforming texture (Fig. 1 (a)). Realizing this paradigm requires deciding what geometry to build and how to optimize the texture on it. We answer both with ***Phy*-DRILL** (**Phy**sical **D**epth **R**eversal **ILL**usion). It instantiates the geometry as a frustum, whose slanted surfaces form converging lines that trigger the perspective depth cues learned by MDE models, and co-optimizes the surface texture to drive the illusion to its full strength. The result is a physical trap: a convex obstacle in reality, systematically perceived as a concave depression. Across various MDE models, our *Phy*-DRILL outperforms existing attacks in both digital and physical environments (Fig. 1 (b)). In summary, our main contributions are as follows:

- We reveal the DepthRI deceives not only human vision but also SOTA MDE models, and we characterize it through an extensive empirical analysis, isolating the number of vanishing points and the texture gradient as the two governing factors.
- We introduce the 2.5D adversarial attack, a paradigm that pairs a parametric geometry with a perspective-conforming texture, and realize it as *Phy*-DRILL, which jointly optimizes the frustum geometry and its surface texture.
- We evaluate *Phy*-DRILL on six MDE models in both digital and physical environments, where it outperforms existing patch-based attacks and transfers from a single surrogate to five black-box models.

2 Related Work

Monocular depth estimation. Monocular Depth Estimation (MDE) predicts dense depth from a single RGB image, evolving from supervised and self-supervised CNNs (Eigen et al., 2014; Godard et al., 2017) to transformer-based dense prediction (Ranftl et al., 2021; 2022). Recent depth foundation models push accuracy further, including Depth Anything V2 (Yang et al., 2024b) for relative depth and ZoeDepth (Bhat et al., 2023), Metric3D (Yin et al., 2023) and Depth Pro (Bochkovskii et al.,

2024) for metric depth. These models form the victim set in our experiments, yet their robustness to adversarial perturbations remains underexplored especially under real-world physical deployment.

Visual illusions and depth perception. Human depth perception relies on monocular cues including linear perspective, texture gradients, and shading (Howard & Rogers, 2012; Cutting & Vishton, 1995). The Hollow-Face Illusion (Hill & Johnston, 2007; Gregory, 1997) is a classic depth reversal effect, which Gregory (1997) attributes to top-down priors overriding bottom-up sensory signals. The depth reversal illusion generalizes it: protruding structures can be misperceived as receding surfaces when geometric and texture cues conflict with typical viewing expectations (Eagleman, 2001). Recent studies revealed that DNNs exhibit similar susceptibilities to visual illusions as humans (Watanabe et al., 2018; Benjamin et al., 2019), suggesting that networks learn implicit priors about perspective geometry from natural images, much as the human visual system does.

Adversarial attacks on depth estimation. Adversarial attacks against MDE induce erroneous depth predictions through crafted perturbations (Wei et al., 2024). Digital attacks optimize pixel-level perturbations (Wong et al., 2020), while physical attacks must survive real capture conditions. Following the seminal adversarial patch (Brown et al., 2017), Yamanaka et al. (2020) first attacked MDE with printed patches, and Cheng et al. (2022) introduced an optimization framework with systematic evaluation metrics. Such planar patches are optimized for a specific projected appearance and degrade under perspective distortion; EoT (Athalye et al., 2018) mitigates this through data augmentation but cannot overcome the planar form itself. π-Jack (Zheng et al., 2024b) manipulates perspective cues by strategically placing common 3D objects to alter the estimated depth of target pixels. Full-coverage texture attacks (Zhang et al., 2019; Wang et al., 2022; Zheng et al., 2024a) improve viewpoint robustness by mapping textures onto an existing object mesh, which constrains deployment. Our 2.5D attack instead treats the geometry itself as an optimizable variable, decoupling the attack from any pre-defined object mesh.

3 METHODOLOGY

Observing the natural deceptiveness of Depth Reversal Illusion (DepthRI), we propose to exploit it to reveal the vulnerable priors within SOTA monocular depth estimators. However, even DepthRI is observed to consistently mislead human visual perception, whether and how this effect transfers to SOTA MDE models remains unclear. Moreover, despite its prevalence, parameterizing DepthRI for an optimization pipeline remains an open challenge. ❶ We therefore, for the first time, formally define DepthRI following the conventional adversarial attack pipeline and conduct an extensive empirical study to uncover the underlying principles (Sec. 3.1). ❷ Following the insights derived from the empirical study, we then deconstruct DepthRI into its fundamental components and design a tailored pipeline, *i.e.*, *Phy*-DRILL, that parameterizes it as an optimizable adversary (Sec. 3.2). Moreover, while existing adversaries struggle to realize digital-to-physical transferability, we *discuss the merits that DepthRI innate for physical deployment* thus enabling superior practical performance.

3.1 REVERSAL ILLUSION AGAINST DEPTH ESTIMATION

DepthRI formulation. As the case illustrated in Fig. 1 (a), the real-world objects exhibiting DepthRI effect are generally characterized by a combination of geometric structure and surface texture appearance. Let Λ represent a typical object showing DepthRI effect that is instantiated by a mesh rendering function $\mathcal{M}(\cdot)$ with geometry $\mathcal{G}$ and texture $\mathcal{T}$:

$$\Lambda = \mathcal{M}(\mathcal{G}_\theta, \mathcal{T}_\phi), \quad \theta \in \Theta, \phi \in \Phi \tag{1}$$

where Θ and Φ define the parameter spaces of geometry and texture configurations capable of inducing the DepthRI phenomenon. We deploy Λ into the target scene with a RGB camera set up for capturing the view as

$$\mathbf{I} = \Pi(\Lambda, \epsilon; \mathbf{c}, \mathbf{r}), \tag{2}$$

where $\Pi(\cdot)$ is the camera projection function parameterized by the camera position $\mathbf{c} \in \mathbb{R}^3$ and the rotation $\mathbf{r} \in SO(3)$ with SO being the special orthogonal group. ϵ encompasses environmental factors, such as lighting, shadow and background clutters, *etc*.

To assess the adversarial effect of Λ, let $f : \mathbf{I} \to \mathbf{D}$ denote a monocular depth estimator mapping an input RGB image $\mathbf{I}$ to a pixel-wise depth map $\mathbf{D}$. Motivated by the observation that DepthRI

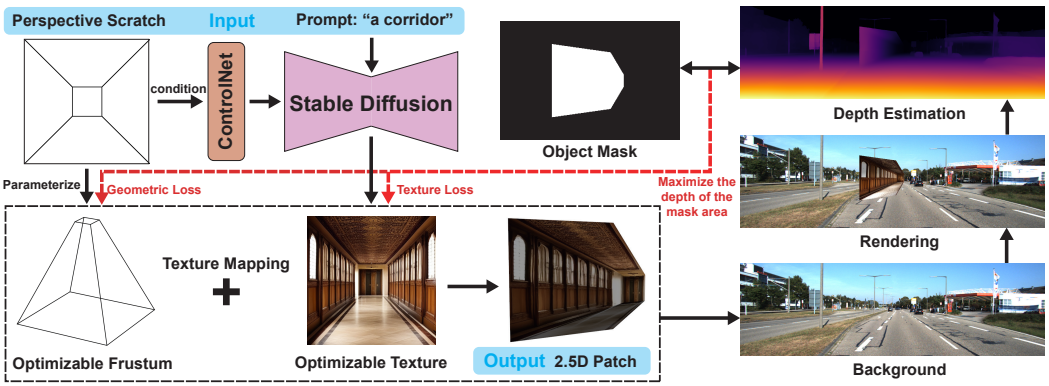

Figure 2: *Phy*-DRILL optimization pipeline of our 2.5D adversarial attack against MDE. Note that the black arrows denote the forward pass for constructing the adversary, while red arrows denote the gradient flow to the frustum geometry and surface texture during optimization.

confounds the perception of concavity and convexity in human vision, we posit that an effective DepthRI object induces an analogous inversion effect against algorithmic estimators:

$$f(\mathbf{I})[x, y] \propto \frac{d}{\mathbf{D}^{gt}[x, y]}, \quad d \in \mathbb{R}, [x, y] \in \mathbf{I}_{\Lambda} \tag{3}$$

where d denotes the intensity of depth reversal illusion and $[x, y]$ indexes pixels within the region occupied by Λ, *i.e.*, $\mathbf{I}_{\Lambda}$. Fundamentally, Eq. (3) characterizes the DepthRI object causes the estimator to predict depths inversely proportional to ground truth within the affected region, posing potential risks in safety-critical applications. To the best of our knowledge, whether and under what conditions Eq. (3) holds for SOTA monocular depth estimators remains an open question. To uncover the underpinnings, we first conduct empirical investigations as follows.

Empirical analysis. Given that DepthRI objects are composed of geometric structure and texture appearance as Eq. (1), we investigate the role of each in achieving the reversal illusion effect. As the forward pipeline depicted in Fig. 2, we employ Stable Diffusion v1.5 (Rombach et al., 2022)[1] as the backbone for texture appearance generation, where the prompt, inspired by Barbet & Fagot (2007), is empirically fixed as "a corridor." For the geometric structure, it is controlled by feeding a perspective scratch into a ControlNet (Zhang et al., 2023) [2]. Depth Anything v2 (Yang et al., 2024b) is adopted as the target estimator throughout the evaluation of the adversarial effect.

Geometric structure setup. For the geometric structure, we observe that DepthRI objects generally share a characteristic that their 2D projections on the image exhibit perspective ambiguity. For example, the 2D frontal view of the object in Fig. 1 can be recognized as either the projection of a concave surface or a convex one. Based on this observation, we examine the effects of different types of perspective projections for investigating the role of object geometry in producing the reversal illusion effect. Specifically, as shown by the first column in Fig. 3, we construct various types of linear perspective scratches following the principle of vanishing points (VPs) [3], *i.e.*, linear perspective with 1/2/3/4 vanishing points. Through the pipeline depicted in Fig. 2, the 2D frontal projection of the target object is generated where the texture is fixed by the text prompt. Then we follow Eq. (2) and Eq. (3) to get the scene image with the object deployed and the corresponding depth predictions.

Geometric structure investigation. In Fig. 3, we present predicted depth maps for two configurations: perspective structure alone (Perspective-only) and perspective structure with texture (Perspective w/ Texture), both evaluated under digital deployment. Two clear observations emerge regarding the role of geometric structure in achieving the reversal illusion effect. ❶ *Vanishing points (VP) strengthen the illusion, up to a limit.* Among the four configurations in Fig. 3, the illusion intensifies as the number

[1] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5
[2] https://huggingface.co/lllyasviel/control_v11p_sd15_lineart
[3] https://en.wikipedia.org/wiki/Vanishing_point

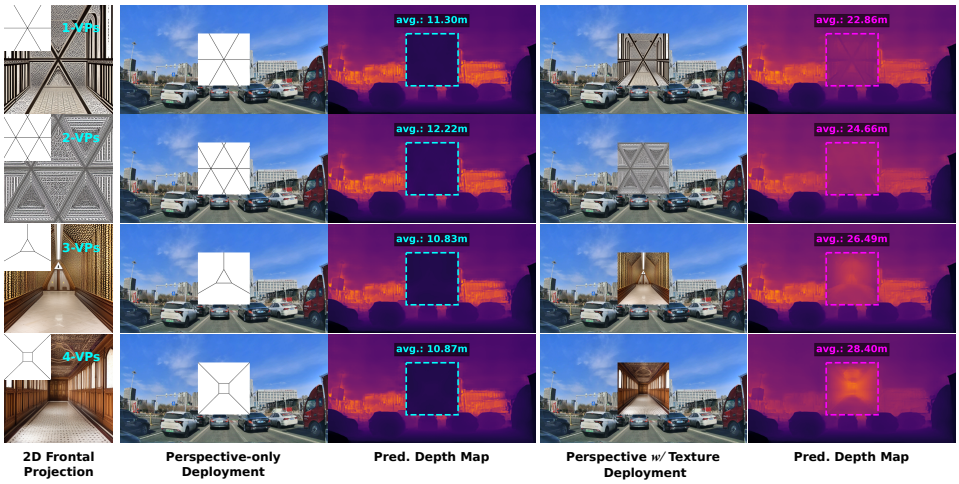

Figure 3: Depth Anything V2-Metric estimation results for the 2D frontal projection of DepthRI objects which are constructed from perspective scratches with various Vanishing Points (VPs). Note that the ground truth depth of the deployed DepthRI object is 10m.

of VPs increases. To test whether this trend continues, we extend the evaluation to 20 VPs. As shown in Fig. 4, the effect peaks at VP = 4 and remains satisfactory up to VP = 6, then decays as VPs grow further. ❷ *Geometry alone is not sufficient.* Perspective structure establishes the spatial foundation of the illusion, but the effect only materializes once a matching texture is applied. This suggests that geometric structure establishes the necessary spatial foundation for the reversal illusion, while texture is the activating factor that realizes its full adversarial effect, motivating further empirical studies in the following part.

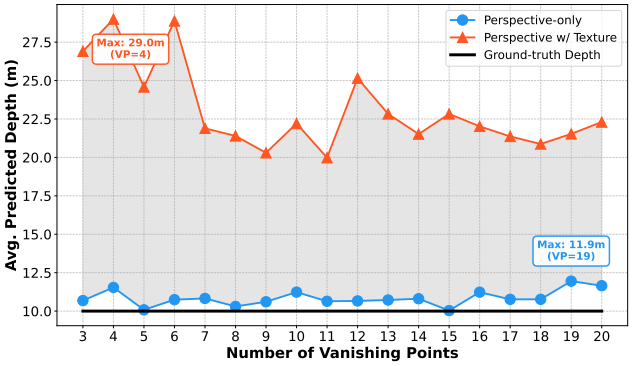

Figure 4: Trend of the average depth prediction as the increasing of vanishing points in perspective scratch. The best results and corresponding VPs are highlighted in boxes.

Texture appearance setup. Existing studies Barbet & Fagot (2007); Yildiz et al. (2021) have shown that it was the gradient of texture, namely texture elements appear smaller and more densely packed with increasing viewing distance, provides critical information for perceiving surface orientation and depth. Therefore, we apply modifications to the gradients of the texture for DepthRI object to investigate texture's influences over the illusion effects. As 4-VPs geometric structure achieved the best results (Fig. 4), we adopt the 4-VPs object as the geometric basis for this investigation. In specific, we adopt the following anti-perspective operation to non-linearly magnify the texture's central region while compress the periphery, thereby creating a "smaller-near-larger-far" visual effect that violates normal texture gradient. Given the pixel coordinates $[u, v]$ of the texture, we set the normalized distance from a texture pixel to center be denoted as $r \in [0, \sqrt{2}]$ and the anti-perspective strength as $s \in [0, 1)$. We have the anti-perspective transformation formulated as

$$u' = c_u + (u - c_u) \cdot r^s, \quad v' = c_v + (v - c_v) \cdot r^s \tag{4}$$

where $[u', v']$ are the transformed pixel coordinates, and $[c_u, c_v]$ represents the center pixel.

Texture appearance investigation. As the prediction results visualized in Fig. 5, we evaluated the influence of texture gradient under five levels of anti-perspective strength, *i.e.*, $s = \{0, 0.3, 0.5, 0.7, 0.9\}$. With the increasing of anti-perspective strength, we observe a gradual degradation of reversal illusion effect induced by the digitally deployed object inside the scene. This trend supports the conclusion that the texture appearance has to match the perspective projection defined by geometric structure in order to replicate the reversal illusion effect.

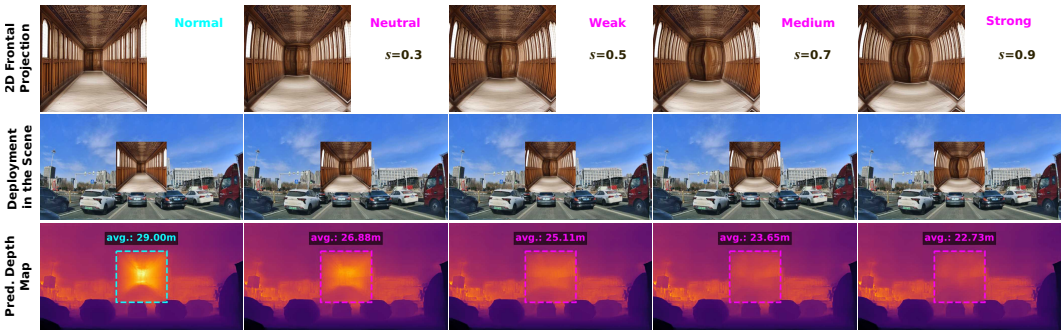

Figure 5: Depth Anything V2-Metric estimation results for the 2D frontal projection of 4-VPs DepthRI object from Fig. 3 under various anti-perspective strength.

Motivation. Our empirical analysis above shows that DepthRI objects are fully capable of posing intense adversarial effect against SOTA MDE model. ❶ Most importantly, we see that DepthRI effect can be fully replicated by constructing a geometric structure with dedicated texture appearance that maintaining a perspective projection confusion. Generally, ❷ we see the geometric structure constructed from 4-VPs linear perspective stands out as the most potential one for reversal illusion effect. ❸ As for the texture, the gradient performs as the central activator while the detail content requires further optimization to maximize the adversarial effect. The impact of texture semantics is evaluated in Appendix C. These findings together inspired us to propose a dedicated geometry-constrained texture optimization pipeline for building the DepthRI as an adversarial attack.

3.2 2.5D Adv. Reversal Illusion via Geo-Texture Co-optimization

Based on the above empirical investigations of DepthRI, we propose our *Phy*-DRILL as illustrated in Fig. 2, namely Geo-Texture Co-optimization to construct optimizable objects that realize DepthRI. As shown in Fig. 1(a), our approach differs from existing 2D patches and 3D adversarial objects. *Phy*-DRILL jointly designs texture and geometric structure, extending beyond texture-only 2D patches, while constraining its geometry to a predefined frustum instead of a fully optimizable 3D object.

Geo-Texture Co-optimization. We design a joint optimization framework in Fig. 2 where the parameters of geometry $\mathcal{G}_\theta$ and texture $\mathcal{T}_\phi$ are updated alternately. Following the definition of Eq. (3), our optimization goal is to synthesize a DepthRI object Λ that maximizes the visual illusion of depth (concavity), while maintaining its physical protrusion (convexity). The total objective function over the parameter spaces Θ and Φ is defined as:

$$\min_{\theta,\phi} \mathcal{L}_{\text{total}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{tex}}\mathcal{L}_{\text{tex}} \tag{5}$$

where λ_{adv}, λ_{geo}, and λ_{tex} are weighting hyper-parameters.

Adversarial depth maximization. To create a general illusion effect, we aim to maximize the depth values predicted by the victim estimator f. Recall that the rendered view is $\mathbf{I} = \Pi(\Lambda, \epsilon; \mathbf{c}, \mathbf{r})$. We minimize the negative mean predicted depth within the region occupied by the object, denoted as $\mathbf{I}_\Lambda$:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\mathbf{I}}\left[\frac{1}{|\mathbf{I}_\Lambda|}\sum_{[x,y]\in\mathbf{I}_\Lambda} f(\mathbf{I})[x,y]\right] \tag{6}$$

By minimizing $\mathcal{L}_{\text{adv}}$, the general reversal of depth prediction is forced.

Geometric structure parametrization. Following the empirical finding at Sec. 3.1, we build a 4-VPs object (See Optimizable Frustum in Fig. 2) and encourage the frustum to grow as tall as possible without breaking the visual illusion. Specifically, given the perimeter of base b and top t for the frustum, we parameterize a 4-VPs object's geometric structure as

$$\mathcal{G}_\Theta = G(r, h) \tag{7}$$

where the $r = t/b$ and h control the frustum's perimeter ratio and height respectively. $G(\cdot)$ is the frustum formation function. Then, we have the following loss for building the optimal geometric

structure

$$\mathcal{L}_{\text{geo}} = r - h \tag{8}$$

Note that we empirically fix the perimeter of base b during the actual optimization, since the size of the frustum will vary in different viewing distance anyway. Therefore, minimizing $\mathcal{L}_{\text{geo}}$ will find an optimal frustum height and perimeter ratio to form a steep and aggressive protrusion that challenges the consistency of the depth estimation model.

Texture appearance optimization. As discussed in Sec. 3.1, perspective-consistent texture gradients are essential for achieving the reversal illusion. We leverage Stable Diffusion with ControlNet conditioned on perspective scratches to initialize the texture. Since the pretrained diffusion model lacks priors of reversal illusion, we further optimize the generated texture with

$$\mathcal{L}_{\text{tex}} = \mathbb{E}_{\tau,\varepsilon}\left[w(\tau)\|\epsilon_\omega(z_\tau(\mathbf{I}); \mathbf{t}, \tau, \mathbf{C}_{\text{line}}) - \varepsilon\|_2^2\right] \tag{9}$$

where τ represents the timestep, ε is the Gaussian noise, $z_\tau(\mathbf{I})$ denotes the noisy latent encoded from the rendered view $\mathbf{I}$, and ϵ_ω refers to the frozen diffusion model parameterized by ω.

Physical deployment. To evaluate real-world reliability, we deploy our optimized 2.5D Adversarial Reversal Illusion object in a physical environment. Given the optimized geometry $\mathcal{G}$ and texture $\mathcal{T}$, we unwrap and print them on regular paper to construct the physical frustum Λ, which is placed at position p in the environment E. The fabrication details are provided in Appendix D. Images of Λ captured from different viewpoints are then fed into various MDE models for depth estimation.

4 Experiments

4.1 Implementation Details

Datasets and setup. We use KITTI (Geiger et al., 2012) images as backgrounds during optimization and evaluate on held-out KITTI validation images through digital compositing. Furthermore, we conduct controlled near-field physical evaluation to validate real-world efficacy. We adopt Depth Anything V2 (Yang et al., 2024b) as the white-box surrogate model during training, and evaluate the transferability of the generated adversarial objects on other black-box victim models. For physical evaluation, D^{gt} is approximated by the measured camera-to-object placement distance, and I_Λ is obtained by post-hoc mask calibration. Appendix A gives the complete optimization details and computational cost.

Baselines. We conduct comparative experiments against three baseline methods 3D2Fool (Zheng et al., 2024a), APA (Yamanaka et al., 2020), and SPOO (Cheng et al., 2022). For each baseline method, we utilize their respective optimal adversarial texture patches as reported in the original works. Notably, since 3D2Fool employs repetitive textures, we use the patch seed for a controlled comparison to ensure a fair evaluation. All other experimental settings and parameters remain consistent across all methods. In addition, we also re-optimize 3D2Fool using the same pipeline, and the comparison results are reported in Appendix B.

Victim models. To evaluate the universality and transferability of our attack, we select a comprehensive set of SOTA victim MDE models: Depth Anything V2 (and its Metric version) (Yang et al., 2024b), the zero-shot metric estimator Depth Pro (Bochkovskii et al., 2024), Metric3D (Yin et al., 2023), ZoeDepth (Bhat et al., 2023), and the classic self-supervised Monodepth2 (Godard et al., 2019). During the experiments, images of fixed resolution are fed into each of these models to obtain the corresponding predicted depth maps.

Optimal height ratio. A critical physical parameter is the frustum geometry relative to the observation distance. With the base edge length fixed at $1m$, we simulate observation distances $d \in [2m, 20m]$ and search for the optimal height h_{opt} that maximizes perceived depth. We find that the optimal ratio $\gamma = h_{\text{opt}}/d$ is scale-invariant across distances, with $\gamma_{\text{avg}} \approx 20.58\%$. Thus, during deployment, the frustum geometry should be uniformly scaled according to the target distance, maintaining $h_{\text{opt}} \approx 0.2058 \times d$ and the corresponding base-to-height ratio.

Table 1: Quantitative comparison of E_{d}, R_{a} (%), and A_{ade} across six victim MDE models in digital environment. ↑: higher values are better attacks.

Method	DA-V2			DA-V2-Metric			Depth Pro			Metric3D			ZoeDepth			Monodepth2		
	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$
Original	1.17	91.08	1.36	8.72	93.70	9.24	10.52	84.55	12.21	13.69	88.11	12.68	1.60	92.70	1.20	0.61	89.05	0.28
Noise	0.56	65.93	0.93	2.05	54.60	3.57	3.66	54.60	4.57	2.98	62.27	7.70	0.37	55.31	0.41	0.08	71.62	0.07
SPOO(Cheng et al., 2022)	0.61	82.91	1.95	2.54	46.16	2.80	3.22	84.30	5.63	15.59	79.41	6.79	0.18	31.39	0.65	0.58	87.79	0.22
3D2Fool(Zheng et al., 2024a)	0.86	68.51	1.03	1.73	77.14	4.01	4.56	63.21	6.74	4.95	81.85	20.75	0.23	40.95	0.55	0.22	89.76	0.30
APA(Yamanaka et al., 2020)	1.08	87.05	0.97	1.81	40.38	3.34	4.65	80.65	4.86	2.54	64.43	5.70	0.61	72.32	0.92	0.18	73.36	0.22
Ours	**1.80**	**99.29**	**2.32**	**12.10**	**95.37**	**16.79**	**16.28**	**91.80**	**20.59**	**47.74**	**99.64**	**45.47**	**2.96**	**98.00**	**2.42**	**0.75**	**97.39**	**0.33**

Table 2: Quantitative comparison of E_{d}, R_{a} (%), and A_{ade} across six victim MDE models in physical environment. ↑: higher values are better attacks.

Method	DA-V2			DA-V2-Metric			Depth Pro			Metric3D			ZoeDepth			Monodepth2		
	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$	$E_{\text{d}}\uparrow$	$R_{\text{a}}\uparrow$	$A_{\text{ade}}\uparrow$
Original	1.75	91.56	0.60	6.42	93.94	3.93	4.13	96.61	1.91	8.80	82.60	7.32	0.91	94.17	0.67	0.14	97.10	0.11
SPOO(Cheng et al., 2022)	0.15	21.42	0.43	0.86	74.47	0.96	1.13	88.16	0.98	0.19	49.95	1.62	0.08	47.86	0.07	0.05	91.09	0.04
3D2Fool(Zheng et al., 2024a)	0.35	30.40	0.45	1.61	71.68	1.59	0.70	64.97	1.92	7.06	88.16	12.28	0.05	28.88	0.09	0.14	96.84	0.12
APA (Yamanaka et al., 2020)	0.85	75.01	0.66	0.93	65.41	0.53	1.72	87.99	1.74	0.28	43.70	8.34	0.14	78.81	0.08	0.02	65.32	0.02
Ours	**4.61**	**99.97**	**2.38**	**6.94**	**97.62**	**4.83**	**9.62**	**99.90**	**6.44**	**41.31**	**99.85**	**48.33**	**1.92**	**99.50**	**1.24**	**0.26**	**99.25**	**0.15**

Table 3: Results under different patch portions and environmental conditions.

Portion	30%	15%	10%	5%	30% angle avg.	Condition	Daylight	Cloudy	Fluorescent	Occlusion	Motion Blur
$E_{\text{d}}\uparrow$	15.87	15.81	8.25	3.56	5.99	$E_{\text{d}}\uparrow$	16.24	13.49	7.65	13.25	9.34
$R_{\text{a}}(\%)\uparrow$	99.09	98.14	90.38	85.55	90.71	$R_{\text{a}}(\%)\uparrow$	94.91	90.66	92.97	97.46	97.59

4.2 Evaluation metrics.

To quantitatively evaluate the efficacy and robustness of the adversarial patch Λ, we employ three metrics. Let $\mathbf{I}_\Lambda$ denote the set of pixels belonging to the object in the captured view $\mathbf{I}$, and compare the estimated depth $f(\mathbf{I})$ with the ground truth depth $\mathbf{D}^{gt}$.

Mean depth estimation error (E_{d}): This metric quantifies the average absolute depth distortion within $\mathbf{I}_\Lambda$ as $E_{\text{d}} = \frac{1}{|\mathbf{I}_\Lambda|}\sum_{[x,y]\in\mathbf{I}_\Lambda} |f(\mathbf{I})[x,y] - \mathbf{D}^{gt}[x,y]|$.

Ratio of affected region (R_{a}): This metric measures the percentage of pixels in $\mathbf{I}_\Lambda$ where the depth error exceeds 10% of the ground truth depth: $R_{\text{a}} = \frac{1}{|\mathbf{I}_\Lambda|}\sum_{[x,y]\in\mathbf{I}_\Lambda} \mathbb{1}(|f(\mathbf{I})[x,y] - \mathbf{D}^{gt}[x,y]| > 0.1\cdot\mathbf{D}^{gt}[x,y])$, where $\mathbb{1}(\cdot)$ is the indicator function.

Angle-aware depth error (A_{ade}): To assess the robustness across different viewpoints, A_{ade} calculates the mean E_{d} over sampled viewing angles $\mathcal{A}$: $A_{\text{ade}} = \frac{1}{|\mathcal{A}|}\sum_{\alpha\in\mathcal{A}} E_{\text{d}}(\alpha)$, where $E_{\text{d}}(\alpha)$ denotes the error measured at a specific camera pose corresponding to angle α.

4.3 Experimental Results

Digital comparative results. We conducted attack experiments in digital environment. The MDE model Depth Anything V2 is denoted as DA-V2. We utilize PyTorch3D (Ravi et al., 2020) to render the adversarial patches and superimpose them onto background images sampled from KITTI. The rendering distance is fixed at $10m$. Here, "Original" refers to the texture without any optimization, and "Noise" denotes texture composed of random noise. The quantitative results are presented in Table 1. Across all evaluated victim models, *Phy*-DRILL consistently delivers superior overall digital attack performance. Furthermore, we evaluate the robustness of *Phy*-DRILL under varying viewing angles and observation distances, see Appendix E.

Physical comparative results. For physical deployment, patches were printed on standard office paper and mounted on transparent acrylic boards. Images were captured at a fixed distance of 0.5m, with the frustum height set to 0.1m. As shown in Table 2, even in physical environments, our *Phy*-DRILL achieves the best E_{d} and R_{a}, while maintaining competitive A_{ade}. We visualize the attack results in Figure 6. Notably, even against black-box MDE model Metric3D, our *Phy*-DRILL demonstrates robust attack effectiveness. In contrast, 2D patch-based methods are recognized as having depth values consistent with the ground truth in most scenarios.

Physical results across different distances and conditions We conduct comprehensive physical experiments (Tab. 3, Fig. 7) to demonstrate the robust effectiveness of *Phy*-DRILL under diverse

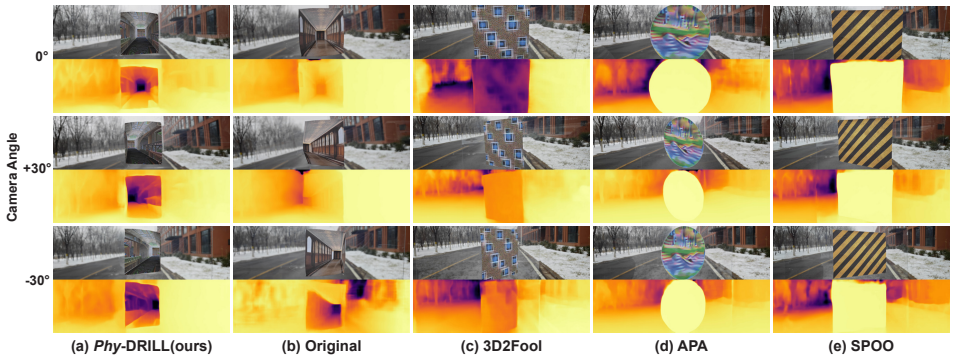

Figure 6: Physical attacks of patches against the black-box MDE model Metric3D

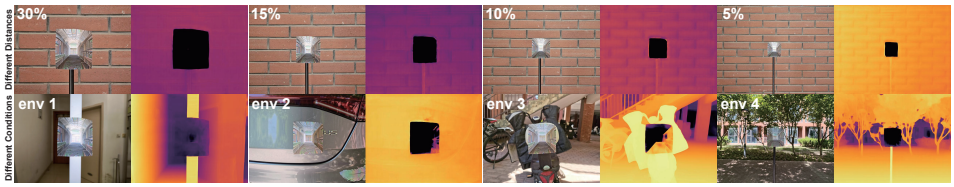

Figure 7: Physical robustness at image occupancies from 30% to 5% (top) and in varied environments (bottom), the victim MDE model is Metric3D.

real-world conditions. Specifically, we evaluate the attack in four different indoor and outdoor scenarios. These experiments cover varying lighting conditions, viewpoints, distances (i.e., different patch proportions in the image), and seamless integration onto plausible structures such as walls and vehicles. Under these dynamic physical constraints, *Phy*-DRILL consistently induces significant depth errors, demonstrating its robustness in practical deployment scenarios.

4.4 Ablation Study

To decouple the contribution of geometry and texture, we conduct an ablation study with four settings: ❶ Original Texture: The 2D original corridor image. ❷ Original Texture + Mesh: The 2D original corridor image is mapped onto the optimized frustum via texture mapping. ❸ Optimized Texture: The 2D optimized texture. ❹ Optimized Texture + Mesh: Our complete 2.5D adversarial patch. Results in Table 4 demonstrate the complementary effects of geometry and texture. The "Original Texture" achieves limited attack performance, indicating that perspective cues alone can induce depth errors but are insufficient for a strong DepthRI effect. Mapping it onto the optimized frustum further improves the attack, while the jointly optimized texture and geometry in our *Phy*-DRILL achieve the best performance, confirming that both components contribute to DepthRI attacks.

Table 4: Ablation study on geometry and texture.

Method	Depth Anything V2			Metric3D		
	E_d↑	R_a↑	A_{ade}↑	E_d↑	R_a↑	A_{ade}↑
Original Texture (2D)	1.05	84.27	1.07	9.69	77.11	7.68
Original Texture + Mesh (2.5D)	1.17	91.08	1.36	13.69	88.11	12.68
Optimized Texture (2D)	1.50	93.05	1.16	34.51	92.61	34.23
Optimized Texture + Mesh (2.5D)	**1.80**	**99.29**	**2.32**	**47.74**	**99.64**	**45.47**

5 Conclusion

In this paper, we introduced *Phy*-DRILL, a novel physical attack against MDE that exploits the Depth Reversal Illusion (DepthRI) effect. By grounding adversarial patches in perspective geometry and optimizing surface texture, we cause protruding 2.5D structures to be systematically misperceived as concave regions by SOTA MDE models. Extensive experiments demonstrate that *Phy*-DRILL significantly outperforms existing methods, offering a robust and viewpoint-consistent physical attack. Our work reveals a critical vulnerability in depth perception systems and opens new directions for geometry-aware adversarial attacks inspired by human visual phenomena.

AI USE STATEMENT

In this work, we used Stable Diffusion v1.5 with Lineart ControlNet for generating the required initial textures and participating in the subsequent optimization process. Additionally, we used ChatGPT for language polishing, modifying part of the code, and auditing. We checked the LLM-generated code and the corresponding revised text and confirmed their correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a physical attack on camera-only monocular depth estimation as an authorized robustness stress test. Our evaluation covers digital and controlled near-field physical settings; full-size public-road tests, interference with deployed systems, and effects on downstream vehicle control remain outside the present study's scope. Because the method has dual-use potential, we state the attacker assumptions and empirical limits explicitly and discuss stereo, temporal, active-depth, and cross-modal consistency checks as potential mitigation directions. The materials should be used only for authorized robustness evaluation and defensive research. All participants provided informed consent, and the study was approved by institutional ethics review.

REPRODUCIBILITY STATEMENT

Section 4 reports the victim models, data, metrics, and evaluation protocol. Appendix A specifies the optimization steps, learning rates, loss weights, model version, augmentations, update schedule, parameter bounds, hardware, and runtime. Appendix C−G provide the prompt sensitivity, fabrication procedure, complete angle/distance curves, transferred-defense evaluation, and stereo extension.

REFERENCES

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, pp. 284–293, 2018.

Isabelle Barbet and Joël Fagot. Control of the corridor illusion in baboons (papio papio) by gradient and linear-perspective depth cues. *Perception*, 36(3):391–402, 2007.

Ari S. Benjamin, Cheng Qiu, Li Q. Zhang, Konrad P. Kording, and Alan A. Stocker. Shared visual illusions between humans and artificial neural networks. *Journal of Vision*, 19(10):57, 2019.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. In *CVPR*, pp. 9585–9594, 2023.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

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

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pp. 5410–5418, 2018.

Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *ECCV*, pp. 514–532, 2022.

James E. Cutting and Peter M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. *Perception of Space and Motion*, pp. 69–117, 1995.

David M. Eagleman. Visual illusions and neurobiology. *Nature Reviews Neuroscience*, 2(12): 920–926, 2001.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, volume 27, 2014.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pp. 3354–3361, 2012.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pp. 270–279, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pp. 3828–3838, 2019.

Richard L. Gregory. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1121–1127, 1997.

Harold Hill and Alan Johnston. The hollow-face illusion: Object-specific knowledge, general assumptions or properties of the stimulus? *Perception*, 36(2):199–223, 2007.

Ian P. Howard and Brian J. Rogers. *Perceiving in Depth, Volume 2: Stereoscopic Vision*. Oxford University Press, 2012.

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 *CVPR*, pp. 14953–14962. IEEE, 2022.

Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12179–12188, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44 (3):1623–1637, 2022.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with PyTorch3D. *arXiv preprint arXiv:2007.08501*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. IEEE, 2022.

Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaofeng Zhang, Wenqian Yao, and Xiaowen Chen. FCA: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *AAAI*, pp. 2414–2422, 2022.

Eiji Watanabe, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka. Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology*, 9:345, 2018.

Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng Wang. Physical adversarial attack meets computer vision: A decade survey. *IEEE TPAMI*, 46(12):9797–9817, 2024.

Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. In *NeurIPS*, volume 33, pp. 8486–8497, 2020.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, pp. 21919–21928, 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, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *NeurIPS*, volume 37, pp. 21875–21911, 2024b.

Gizem Y Yildiz, Irene Sperandio, Christine Kettle, and Philippe A Chouinard. Interocular transfer effects of linear perspective cues and texture gradients in the perceptual rescaling of size. *Vision Research*, 179:19–33, 2021.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, pp. 9043–9053, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pp. 3813–3824. IEEE, 2023.

Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*, 2019.

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 *CVPR*, pp. 24452–24461, 2024a.

Tianyue Zheng, Jingzhi Hu, Rui Tan, Yinqian Zhang, Ying He, and Jun Luo. π-Jack: Physical-World adversarial attack on monocular depth estimation with perspective hijacking. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 7321–7338, 2024b.

A Reproducibility and Optimization Details

A.1 Generative backbone and environmental augmentation

We use Stable Diffusion v1.5 with the corresponding Lineart ControlNet as the generative prior. During Score Distillation Sampling (SDS) (Poole et al., 2022), the classifier-free guidance scale is 7.5, and the diffusion timestep is sampled uniformly from $[0.02, 0.98]$ of the full schedule. During optimization, we incorporate a differentiable augmentation pipeline. Specifically, we randomize camera poses by varying the distance, elevation, and azimuth angles, and simulate varying illumination conditions by randomly positioning point light sources within the scene. At each optimization step, the camera distance is sampled from $[3, 8]$ m, the elevation and azimuth are sampled from $[-60°, 60°]$, and the point-light coordinates are sampled with $x, y \in [-6, 6]$ and $z \in [-7, -2]$. With probability 0.5, we apply image-level corruptions comprising ColorJitter (brightness, contrast, and saturation 0.2; hue 0.1) and Gaussian blur (kernel size 5 and $\sigma \in [0.1, 2.0]$).

A.2 Alternating optimization and computational cost

PyTorch3D provides differentiable rendering. Gradients from $\mathcal{L}_{\text{adv}}$ pass through the rasterizer to the texture map $\mathcal{T}_\phi$ and frustum parameters $\mathcal{G}_\theta$. The diffusion backbone is frozen and supplies only the SDS regularizer. Geometry is updated once every four optimization steps, while texture is updated on the remaining three steps. We optimize parameter logits and apply a sigmoid in the forward pass, constraining texture RGB values to $[0, 1]$ and the frustum inner scale to $[0.05, 0.95]$. The geometry gradient norm is clipped at 1.0. Optimization uses Adam for $50{,}000$ steps, with learning rates 0.01 for texture and 0.05 for geometry. The loss weights are $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{tex}} = 2.0$, and $\lambda_{\text{geo}} = 0.1$.

A complete *Phy*-DRILL optimization on one NVIDIA RTX 3090 GPU (24 GB) takes approximately 3–5 hours. As a rough reference, APA (Yamanaka et al., 2020) takes 3–8 hours on an RTX 3090, SPOO (Cheng et al., 2022) takes 8–18 hours on an RTX 3090/4090, and 3D2Fool (Zheng et al., 2024a) takes 24–72 hours on an A5000/A100; these values are descriptive because the hardware and implementations differ across methods.

B MATCHED COMPARISON WITH RE-OPTIMIZED 3D2FOOL

To reduce optimization-pipeline confounding, we re-optimize 3D2Fool (Zheng et al., 2024a) using the same setup as *Phy*-DRILL, *i.e.*, Depth Anything V2 as white-box victim model and same augmentation pipeline. APA and SPOO are evaluated with their original attack parameterizations, as they cannot be readily adapted to our attack setup. It can be seen from Tab. B.1 that *Phy*-DRILL still outperforms the re-optimized 3D2Fool in black-box transferability, while the performance of 3D2Fool lags behind its original attack texture. This matched comparison reduces the concern that *Phy*-DRILL's transferability advantage arises solely from a stronger optimizer.

Table B.1: Matched comparison with the re-optimized 3D2Fool method in digital environment.

Method	Depth Anything V2 Metric			Metric3D		
	$E_{\mathrm{d}}\uparrow$	$R_{\mathrm{a}}(\%)\uparrow$	$A_{\mathrm{ade}}\uparrow$	$E_{\mathrm{d}}\uparrow$	$R_{\mathrm{a}}(\%)\uparrow$	$A_{\mathrm{ade}}\uparrow$
3D2Fool	1.27	30.61	2.32	2.54	52.62	4.99
***Phy*-DRILL**	**11.86**	**94.47**	**15.77**	**48.21**	**99.27**	**47.51**

C PROMPT SENSITIVITY

After fixing the optimal 4-VP geometry, we further investigate how the semantic content of the generated texture affects DepthRI. Since the texture gradient provides the fundamental perspective cue, we focus here on the influence of different text prompts used in Stable Diffusion. We fix the geometry to the 4-VP construction, use Depth Anything V2-Metric as the evaluated estimator, and set the object ground-truth depth to 10 m. As shown in Figure C.1, textures generated from different prompts consistently induce depth overestimation, with predicted depths generally concentrated between 20 and 30 m. This suggests that the attack remains effective across different texture semantics as long as the required perspective gradient is preserved. Among the tested prompts, the concise prompt "a corridor" achieves the largest average predicted depth of approximately 29 m. We therefore adopt "a corridor" as an empirical design choice for our texture generation pipeline.

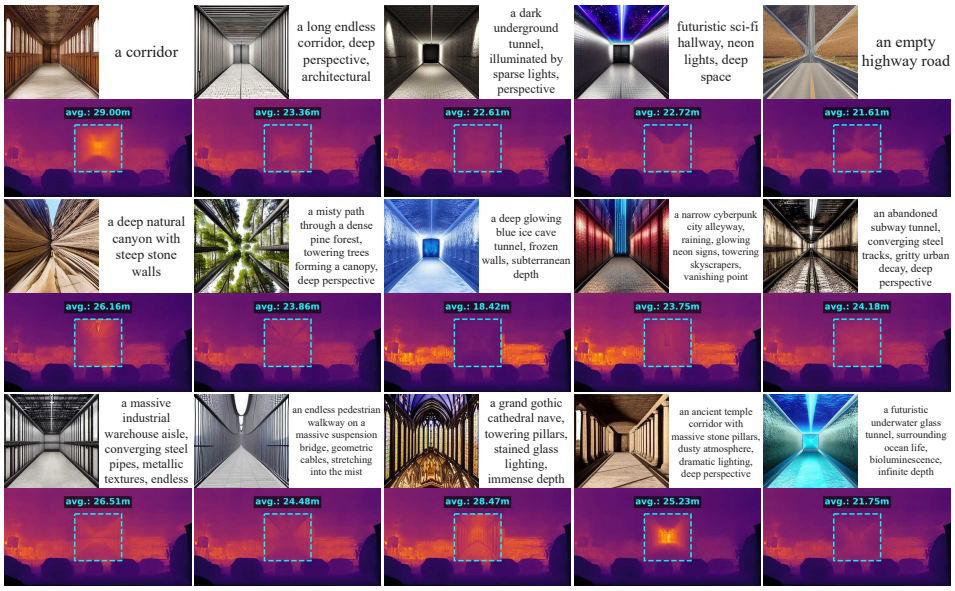

Figure C.1: Depth estimates for textures generated from different prompts. All objects use the same 4-VP geometric structure with a ground-truth depth of 10 m. Different prompts consistently induce depth overestimation, while the concise prompt "a corridor" achieves the highest average predicted depth (29 m).

D PHYSICAL DEPLOYMENT AND THREAT MODEL

We consider a physical attack scenario where the attacker fabricates and places a pre-optimized 2.5D object in the camera view to induce depth overestimation in MDE models. The attacker cannot modify model parameters or sensor streams. We assume white-box access to Depth Anything V2 for optimization and black-box transfer to other MDE models. The victim system is a monocular RGB-frame MDE model without stereo or cross-modal geometric consistency checks. The simulated ratio $\gamma \approx 20.58\%$ is scale-invariant across distances. The attack targets model-level depth overestimation, not human imperceptibility. We invited 10 participants to observe the physically deployed 2.5D patch under the tested viewing condition and report whether its central surface appeared protruding or receding. All 10 participants reported a receding interpretation, although the presence of the patch itself remained visually noticeable. This descriptive observation concerns perceived depth polarity only; it does not establish imperceptibility, naturalness, concealment, or population-level prevalence.

Unlike purely digital attacks or fully 3D adversarial objects, our 2.5D adversarial patch requires minimal manufacturing effort while maintaining a rigid geometric structure. D.1 illustrates the straightforward fabrication process, which resembles simple paper crafting. After the optimization converges, the texture mapped on the frustum is unwrapped into a 2D planar template and directly printed onto standard office A4 paper using a conventional color printer. The printed template is then cut along the predefined outlines and assembled along the corresponding edges to construct the physical 2.5D frustum. This fabrication process requires only common materials and tools, demonstrating the low-cost and accessible nature of our physical attack. As revealed by the back view in D.1(d), the fabricated patch is intrinsically a hollow, convex geometric structure. When viewed from the front during deployment, the optimized texture gradient aligns with the physical perspective cues, inducing a strong depth reversal illusion.

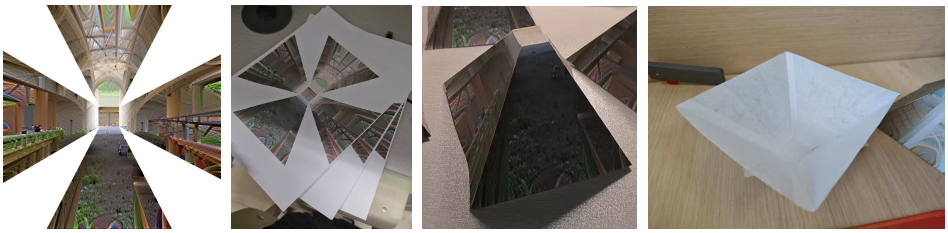

Figure D.1: Fabrication process: texture unwrapping, printing, assembly, and the back view of the resulting frustum.

E ADDITIONAL RESULTS ON ANGLE AND DISTANCE ROBUSTNESS

To further analyze the impact of viewing angle and distance on attack performance, we conduct additional evaluations in a digital environment. We render the adversarial patches at different viewing angles and observation distances against various victim MDE models. Specifically, the viewing angle ranges from -80° to 80° with the observation distance fixed at $10m$, while the observation distance varies from $3m$ to $15m$ with the viewing angle fixed at 0°.

The Mean depth estimation error (E_d) results under varying viewing angles and observation distances are shown in E.1. Our *Phy*-DRILL achieves consistently strong attack performance across different victim MDE models and maintains robustness under diverse observation conditions. Notably, some MDE models such as Depth Anything V2 exhibit relatively low E_d at frontal views, possibly because frontal views primarily rely on texture while structural contribution becomes more pronounced under angular variations. Moreover, E_d increases as distance grows, which can be explained by perspective projection: at greater distances, the patch occupies a smaller image region with reduced contextual interaction, making it more likely to be predicted as farther away.

As a complement to the E_d analysis, we further evaluate robustness using the Ratio of affected region (R_a). While E_d measures the absolute magnitude of depth distortion, R_a reflects the spatial consistency and coverage of the adversarial effect across the target patch. As illustrated in E.2, we

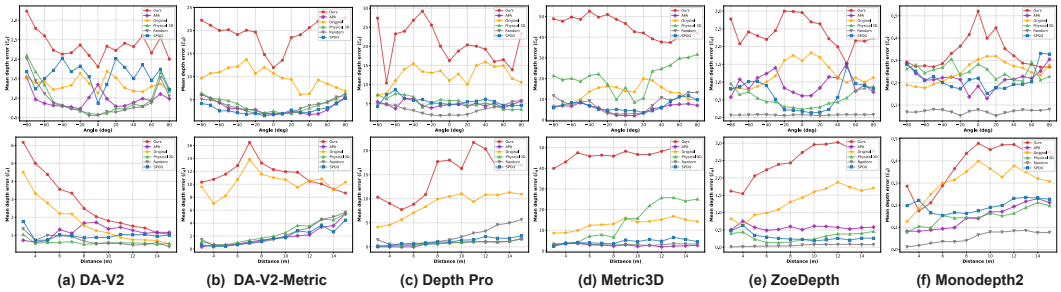

Figure E.1: Mean Depth Estimation Error (E_{d}) under varying viewing angles and observation distances across different victim MDE models.

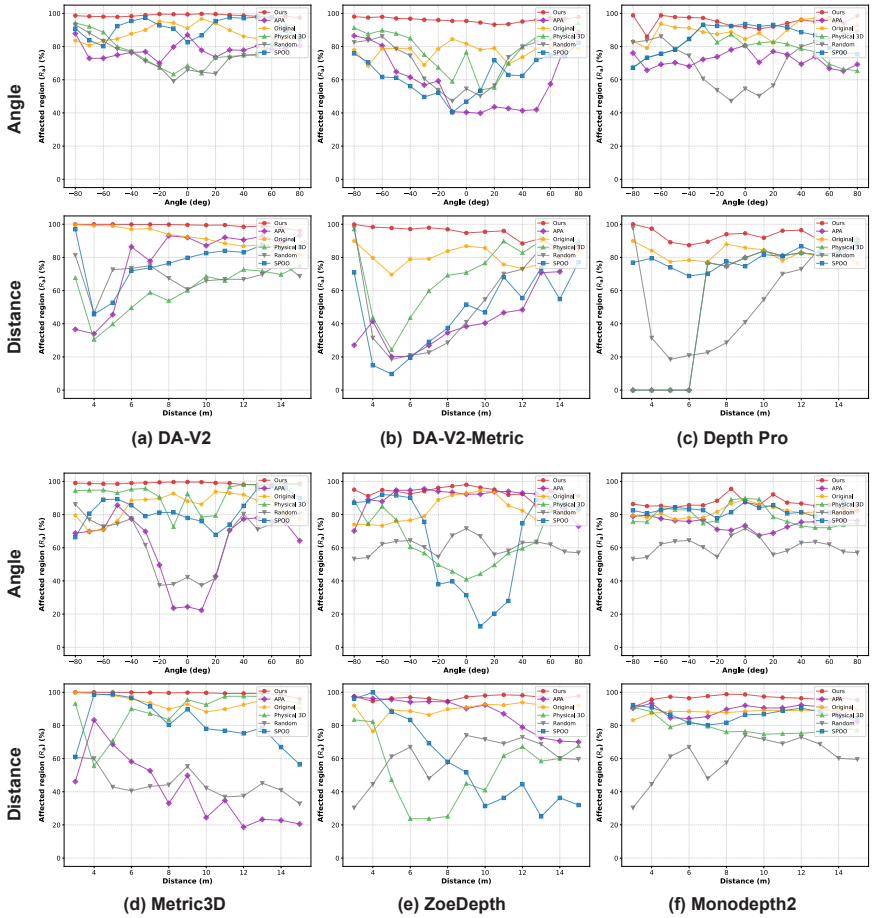

Figure E.2: Affected-region ratio R_{a} under varying viewing angles and rendering distances.

visualize the R_{a} curves across six SOTA MDE models under varying viewing angles and observation distances.

It is evident from the curves that our 2.5D adversarial patch consistently maintains an overwhelmingly high R_{a}, approaching 100% across almost all testing conditions. Regarding the viewing angle robustness, 2D baseline methods (*e.g.*, APA and SPOO) exhibit performance degradation and fluctuations across different perspectives, while our method preserves a stable attack coverage across the entire -80° to 80° range. Similarly, in the distance evaluation, while baseline performances heavily fluctuate at certain distances, our scheme remains highly stable and continuously outperforms others.

F TRANSFERRED EVALUATION WITH SAC

To provide a preliminary evaluation against existing adversarial defenses, we further test SAC (Liu et al., 2022). SAC was originally developed to detect and complete adversarial patches for object detectors rather than monocular depth estimation (MDE). Therefore, we treat it only as a transferred defense and do not assume that it is specifically designed for the characteristics of DepthRI.

Table F.1 reports the Mean depth estimation error (E_{d}) before and after applying SAC. In the digital setting, SAC changes E_{d} only slightly, from 14.73 m to 14.69 m. In the physical setting, the measured E_{d} remains unchanged at 11.12 m. These results indicate that, under our evaluated setup, directly transferring SAC provides only limited mitigation against DepthRI.

Table F.1: Mean depth error of Depth Anything V2-Metric before and after applying the transferred SAC defense.

Setting	Digital E_{d} (m)	Physical E_{d} (m)
Without SAC	14.73	11.12
With SAC	14.69	11.12

G EXTENSION TO STEREO MATCHING

Stereo matching is widely deployed in safety-critical domains due to its robustness to monocular appearance cues. We extend *Phy*-DRILL to stereo matching via black-box transfer and evaluate two representative models, IGEV (Xu et al., 2023) and PSMNet (Chang & Chen, 2018). Using PyTorch3D (Ravi et al., 2020), we construct a virtual stereo camera and synchronously render the 2.5D object into both left and right views for disparity inference.

The quantitative results are reported in Table G.1, with representative outputs shown in Figures G.1 and G.2. Phy-DRILL achieves the highest Ed and Ra on both stereo models, while its angle-averaged error is competitive rather than uniformly best. Since stereo systems estimate depth from binocular disparity, pure 2D patches cannot provide physically consistent geometric discontinuities across the two views. In contrast, the 2.5D structure of *Phy*-DRILL produces consistent perspective variations in both views and therefore exhibits stronger transferability than 2D baselines. However, the stereo models still largely interpret the object according to its convex physical geometry rather than as the intended receding cavity. This limitation is expected because our current texture optimization does not explicitly incorporate binocular matching consistency, leaving robust physical stereo attacks for future work.

Table G.1: Digital black-box transfer to stereo matching models.

Method	IGEV			PSMNet		
	$E_{\mathrm{d}}\uparrow$	$R_{\mathrm{a}}(\%)\uparrow$	$A_{\mathrm{ade}}\uparrow$	$E_{\mathrm{d}}\uparrow$	$R_{\mathrm{a}}(\%)\uparrow$	$A_{\mathrm{ade}}\uparrow$
Noise	13.90	75.01	17.77	7.38	71.12	7.16
Original	13.20	77.15	19.70	10.48	70.01	7.83
3D2Fool	10.24	67.64	13.77	9.41	74.95	6.88
APA	15.59	71.53	19.54	6.11	59.36	1.67
SPOO	16.00	69.94	12.73	11.94	77.65	8.48
***Phy*-DRILL**	**17.60**	**79.63**	**18.75**	**13.36**	**79.96**	**8.62**

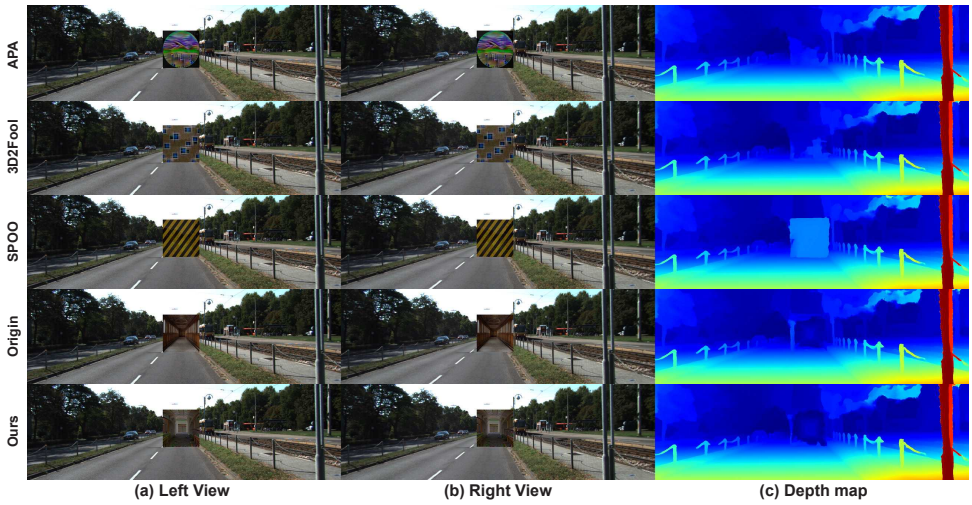

Figure G.1: Qualitative black-box transfer results on IGEV: left view, right view, and predicted depth.

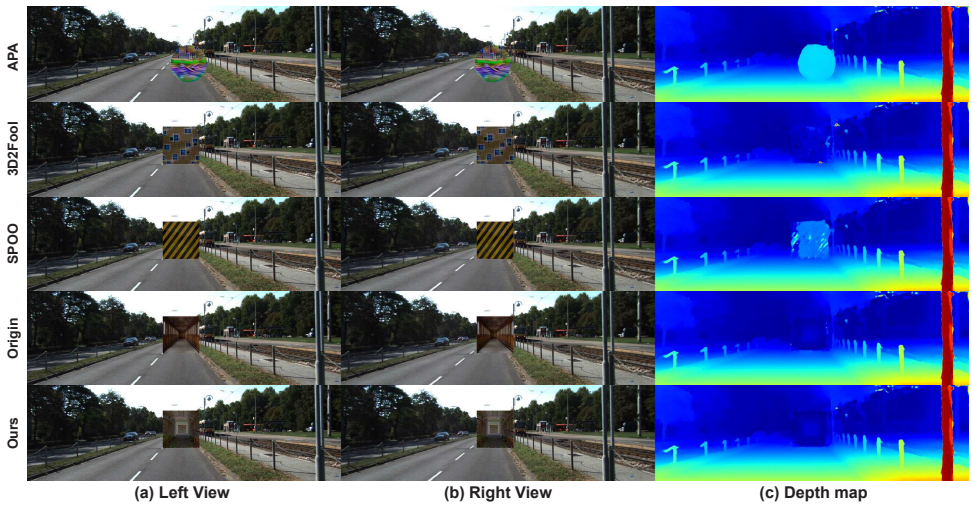

Figure G.2: Qualitative black-box transfer results on PSMNet: left view, right view, and predicted depth.