
DepthVanish: Optimizing Adversarial Interval Structures for Stereo-Depth-Invisible Patches

Yun Xing^{1,2,5*} **Yue Cao**^{5,6*} **Nhat Chung**⁵ **Jie Zhang**⁵ **Ivor Tsang**^{5,6}

Ming-Ming Cheng^{2,3} **Yang Liu**⁶ **Lei Ma**^{1,4} **Qing Guo**^{2,5†}

¹ University of Alberta, Canada ² VCIP, CS, Nankai University, China

³ NKIARI, Shenzhen Futian, China ⁴ The University of Tokyo, Japan

⁵ CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore

⁶ Nanyang Technological University, Singapore

Abstract

Stereo depth estimation is a critical task in autonomous driving and robotics, where inaccuracies (such as misidentifying nearby objects as distant) can lead to dangerous situations. Adversarial attacks against stereo depth estimation can help reveal vulnerabilities before deployment. Previous works have shown that repeating optimized textures can effectively mislead stereo depth estimation in digital settings. However, our research reveals that these naively repeated textures perform poorly in physical implementations, *i.e.*, when deployed as patches, limiting their practical utility for stress-testing stereo depth estimation systems. In this work, for the first time, we discover that introducing regular intervals among the repeated textures, creating a grid structure, significantly enhances the patch’s attack performance. Through extensive experimentation, we analyze how variations of this novel structure influence the adversarial effectiveness. Based on these insights, we develop a novel stereo depth attack that jointly optimizes both the interval structure and texture elements. Our generated adversarial patches can be inserted into any scenes and successfully attack advanced stereo depth estimation methods of different paradigms, *i.e.*, RAFT-Stereo and STTR. Most critically, our patch can also attack commercial RGB-D cameras (Intel RealSense) in real-world conditions, demonstrating their practical relevance for security assessment of stereo systems. The code is officially released at: <https://github.com/WiWiN42/DepthVanish>

1 Introduction

Depth estimation is a crucial component in safety-critical embodied systems like autonomous driving [6] and robotics [3], where accurate perception of the 3D environment is essential for reliable operation. Investigating the errors in depth estimation, such as mistaking nearby objects as distant ones in safety-critical embodied systems [30, 5, 38, 8, 20, 42, 41], can provide critical insights for safety practices. Most existing works focus on the security vulnerabilities of monocular depth estimation, which relies heavily on scene priors from single images. Stereo depth estimation, on the other hand, utilizes geometric constraints and typically provides more robust and metrically accurate results, making it attractive for high-stakes applications.

However, despite this inherent advantage, recent studies revealed that DNN-based stereo pipelines remain vulnerable to adversarial attacks, as carefully crafted pixel-level perturbations [31, 1] can cause substantial disparity estimation errors. Nevertheless, previous works have primarily addressed digital

*indicates equal contribution. This work was done during Yun Xing was an intern at CFAR & IHPC, A*STAR and Nankai University. † Corresponding author, email: tsingqguo@ieee.org.

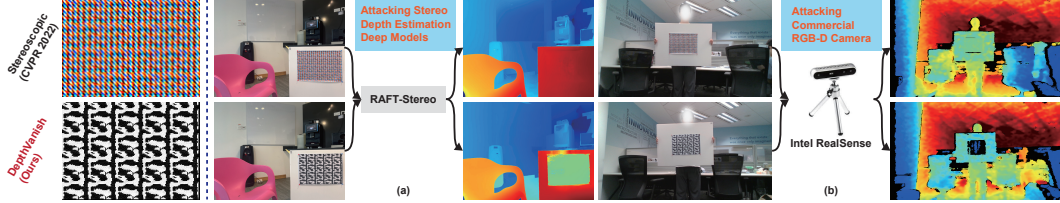


Figure 1: Baseline (Stereoscopic [1]) vs. our DepthVanish on attacking RAFT-Stereo [19] and Intel RealSense.

attacks utilizing full-image noise, which are impractical in real-world contexts due to constraints like limited patch size, varying viewing angles, and dynamic lighting conditions, *etc.* As illustrated in the first row of Fig. 1, when applied as a physical patch, the existing Stereoscopic [1] fails to effectively attack RAFT-Stereo and Intel RealSense. This lack of physically realizable and generalizable attack methods presents a significant limitation in evaluating the robustness of stereo systems, particularly as stereo estimation continues to be deployed in real-world, safety-critical applications.

In this study, we address these limitations by introducing the first adversarial patch attack that is effective in both digital and physical settings against widely deployed deep stereo depth estimation models (Fig. 1 second row). Fundamentally, we discover that adding regular intervals among repeated textures to form a spatial structure shows great potential for improving the attack effectiveness and enables digital-to-physical transferability. Through systematic analysis, we show how interval spacing influences the attack success. These insights inform a novel optimization pipeline that jointly designs patches’ texture and structure to achieve high attack effectiveness across models and deployment settings. Thus, we propose a novel optimization pipeline that co-designs both texture elements and interval structure for generating adversarial patches that ❶ remain effective when physically printed and inserted into real scenes, ❷ work across diverse datasets and environments and ❸ generalize across different stereo depth estimation models, including commercial RGB-D sensors, *i.e.*, Intel RealSense. In summary, our contributions are as follows,

- We introduce the first adversarial attack that is both digitally and physically effective for deep stereo estimation models including the advanced RAFT-Stereo and Stereo Transformer.
- By conducting a comprehensive empirical study, we discover that regular interval spacing among repeated textures significantly improves the patch attack effectiveness and its real-world transferability over naive texture repetitions.
- We develop a joint optimization algorithm, *i.e.* DepthVanish, that co-designs the texture and its spatial structure within the patch to maximize the digital and physical attack effectiveness.
- By physically evaluating our patch, we expose severe safety concerns of existing stereo depth estimation systems and highlight the emergency of practical model robustness enhancement.

2 Related Work

Stereo depth estimation. Stereo-based depth estimation is a technique that infers scene depth from visual correspondences, which captured as disparity maps, between pairs of stereo images in various applicable settings [23, 35, 2, 25, 24, 15]. Traditional methods typically follow a multi-stage pipeline involving the computation of matching costs, cost aggregation, and optimization to predict and refine disparities [26, 4, 33, 10]. In contrast, recent advances have incorporated deep neural networks [29], enabling end-to-end learning of feature representations for correspondence matching and direct prediction of disparity and/or depth. In particular, CNN-based methods [4, 12, 34, 39, 22, 27] typically build 3D cost volumes from shared-weight feature encoders, attention-based models [18, 11, 28, 14] employ vision transformers to model global correspondences and disambiguate difficult regions, and iterative refinement methods [19, 16, 32] apply a recurrent update operator to progressively converge on the final disparity, avoiding the memory-intensive 3D cost volumes. Compared to monocular methods [36, 37], stereo offers improved robustness by leveraging geometric constraints from dual viewpoints, but still faces challenges in low-texture and repetitive areas [40, 21].

Depth estimation attack. Due to their effectiveness and capability for real-time performance, depth estimation systems have become essential components of safety-critical applications such as autonomous driving [6] and robotic navigation [3]. Monocular depth estimation models, in particular, have been extensively studied under both digital [30, 5, 38] and physical adversarial

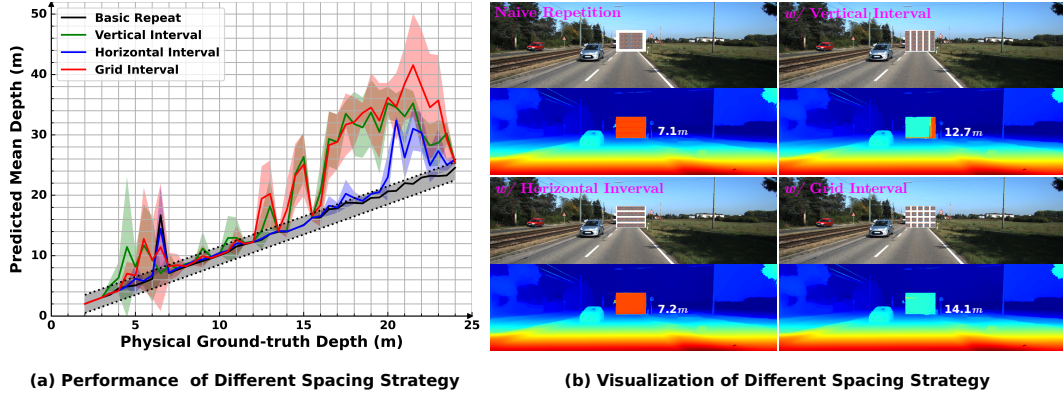


Figure 2: Adversarial effect of interval spacing on depth prediction. (a) Mean predicted depth (solid lines) and variance (shaded regions) for different interval spacing strategies, averaged over interval widths of $2 - 10 px$. The gray dashed band indicates $\pm 1.5 m$ from the ground truth. (b) Visualization of depth prediction results for typical different interval spaced patches where the ground truth depth is $7 m$.

attacks [8, 20, 42, 41]. These evaluations have revealed various system vulnerabilities and led to the development of tailored defense strategies [13, 9, 7], including adversarial training and robust feature learning. In contrast, despite their geometric soundness and widespread deployment, stereo depth estimation systems [19, 17] have received limited attention in adversarial research. Existing research has focused primarily on digital, white-box attacks [1, 31], overlooking potential vulnerabilities in the physical world. This gap is particularly concerning, as stereo systems rely on precise correspondence between left and right images. Failures in such systems can lead to serious consequences, especially in autonomous applications where accurate and reliable 3D perception [35] is critical.

3 Motivation

3.1 Naive Repetition Fails in Realistic Patch Attacks

Stereo depth estimation recovers 3D structure by identifying correspondences between left and right images [43], typically formulated as a pixel-wise optimization along epipolar lines:

$$d^*(x) = \arg \min_d C(x, d), \quad (1)$$

where $d \in \mathbb{Z}$ represents the horizontal disparity between pixel x in the left image and pixel $x - d$ in the right image, and $C(x, d)$ denotes the matching cost between them. When repetitive patterns are presented, the cost volume exhibits periodic ambiguity [26]:

$$C(x, d) \approx C(x, d + ns), \quad \forall n \in \mathbb{Z}, \quad (2)$$

where s denotes the spatial repetition period. This periodicity produces multiple equally plausible matches, thereby increasing the likelihood of incorrect or unstable depth estimations.

Previous adversarial attacks [1, 31] inject repetitive optimized noise over the entire image to exploit such periodic ambiguities. Since global injection is impractical in real-world scenarios, we instead explore attacks using localized adversarial patches. As shown by the black curve in Fig. 2(a), we deploy the repetitive noise from [1] as patch into a real-world scene at different ground-truth depth and plot the corresponding predicted mean depth. It can be seen that simply repeating the noise within patches results in predicted depth that remains same to the ground truth, indicating limited adversarial effectiveness. This is visually confirmed in Fig. 2(b) (top left), where a naive repeated patch yields a predicted depth of $7.1 m$, which is almost identical to the ground truth of $7 m$. This observation reveals a key limitation of existing studies: naive repetition fails to generate sufficient ambiguity within practical patches, which motivates the need for more structured pattern designs.

3.2 Structured Intervals: Enhancing Patch Adversarial Effectiveness

To address the limitation, we propose introducing regular intervals into the repetitive pattern to amplify the matching ambiguity as Eq.(2), thereby enhancing the adversarial effect of the patch.

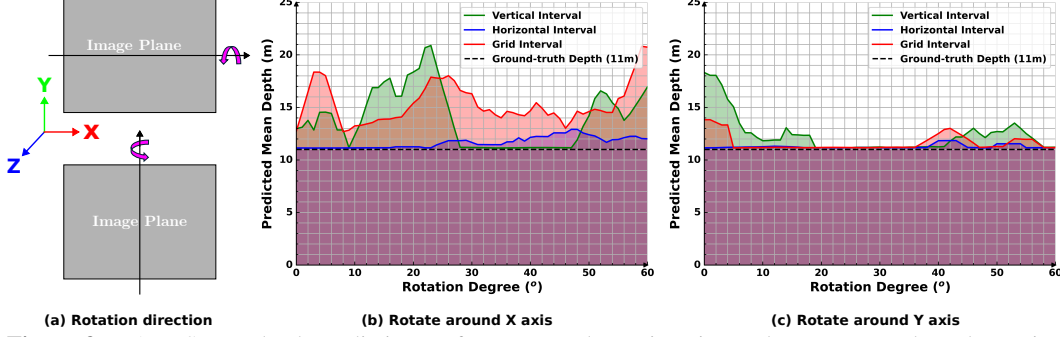


Figure 3: RAFT-Stereo depth prediction performance under various interval structures and patch rotation degrees. (a) Illustration of rotation around the X and Y axes. (b) Depth prediction performance at different rotation degrees around X axis. (c) Depth prediction performance at different rotation degrees around Y axis.

As demonstrated in Fig. 2(b), given the patch with basic repetitive pattern from [1] (top left), we add vertical (top right), horizontal (bottom left) and grid (bottom right) space to form patches with structured intervals. We systematically evaluate the impact of different interval configurations on the RAFT-Stereo model using the KITTI dataset. As shown in Fig. 2(a), structured intervals notably enhance attack effectiveness. ❶ Basic Repeat (**black**): the predicted patch depth remain close to the ground truth depth indicating minimal adversarial influence, which is also verified by the visualization in Fig. 2(b) (top left). ❷ Horizontal Interval (**blue**): moderate overestimation beyond 15 m (e.g., a 20 m true depth yields a ~ 24 m prediction). Visual results (Fig. 2(b), bottom left) confirm a slight increase to 7.2 m. ❸ Vertical Interval (**green**): produces larger errors, frequently reaching ~ 30 m at a 20 m ground truth. In Fig. 2(b) (top right), the predicted depth surges to 12.7 m. ❹ Grid Interval (**red**): combining intervals in both directions produces the strongest adversarial effect, with depth predictions surpassing 40 m at a 23 m ground truth. In the visual result (Fig. 2(b), bottom right), the predicted depth reaches 14.1 m, demonstrating a significant adversarial effectiveness.

In summary, structuring the patch with both horizontal and vertical intervals (i.e., grid spacing) greatly increases the adversarial effect of patches, far exceeding the impact of simple repetition. However, we also observe two critical limitations: ❶ the overall attack performance remains limited, especially when the patch is placed close to the camera. ❷ the significant variation in performance across different interval configurations suggests that a single fixed interval structure is insufficient.

3.3 Structured Intervals: Improving Attack Robustness across Viewpoints

A practical adversarial patch must maintain its effectiveness even when the patch is rotated or viewed from different orientations. This is particularly important under real-world deployment conditions, where precise placement is difficult to control. To this end, we systematically evaluate the impact of interval structure on attack robustness against patch rotation. As shown in Fig. 3(a), we rotate the patch along two axes (i.e., X and Y) and summarize the predicted mean depth in Fig. 3(b) and (c).

For X-axis rotation (Fig. 3(b)): ❶ the horizontal (**blue**) and vertical (**green**) intervals exhibit angle-dependent performance, succeeding only at certain angles; ❷ the grid interval (**red**) is more robust, demonstrating more consistent effectiveness across different angles. For Y-axis rotation (Fig. 3(c)), although all configurations show moderate attack robustness across viewpoints, adding intervals still yields improvements. These results show that structured intervals improve attack robustness to patch rotation, which is essential for reliable adversarial attacks in real-world scenarios.

In summary, the above findings underscore the promise of structured intervals but also reveal their limitations under varying depths and configurations. These observations highlight the need for further optimization of the patch’s texture and structure to achieve more effective attacks.

4 Problem Formulation

To formalize the stereo depth estimation task and define our attack objective, we begin with the following setup. Given a stereo image pair $(\mathbf{I}_l, \mathbf{I}_r)$ where $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{3 \times H \times W}$ of a specific scene, a pretrained stereo depth estimation model $\mathcal{F}(\cdot)$ predicts the pixel-wise disparity map $\mathbf{d}_{pred} =$

$\mathcal{F}(\mathbf{I}_l, \mathbf{I}_r) \in \mathbb{R}^{H \times W}$. The corresponding depth map is computed as $\mathbf{z} = \frac{f \times B}{\mathbf{d}_{pred}}$ where f and B denote the focal length and baseline of the stereo camera rig respectively.

In general, the objective of adversarial patch attack is to construct a patch $\mathbf{P} \in \mathbb{R}^{3 \times h_p \times w_p}$ such that the stereo depth estimation model $\mathcal{F}(\cdot)$ produces an incorrect disparity output for the patch:

$$\mathcal{F}^p(\hat{\mathbf{I}}_l, \hat{\mathbf{I}}_r) \neq \mathcal{F}^p(\mathbf{I}_l, \mathbf{I}_r) \quad (3)$$

where $\hat{\mathbf{I}}_l$ and $\hat{\mathbf{I}}_r$ denote the stereo images with the adversarial patch $\mathbf{P}$, and p indicates the corresponding pixel region occupied by the patch within the prediction results. As analyzed in Sec. 3, interval spacing can trigger critical depth estimation failures, *i.e.*, the disappearance attack. To expose the severity of such vulnerability, we define a more destructive attack objective:

$$\mathcal{F}^p(\hat{\mathbf{I}}_l, \hat{\mathbf{I}}_r) = \mathbf{0}, \quad s.t. \mathbf{d}_{gt}^p = \mathbf{c}, \quad (4)$$

where the model predicts zero disparity for the patch region (*i.e.*, infinite depth), despite the ground truth disparity of the patch, $\mathbf{d}_{gt}^p$, indicating a fixed, close distance $(f \times B)/\mathbf{c}$. This attack objective reveals more severe vulnerabilities than Eq. (3) and poses substantial safety risks, particularly when the patch is physically realizable and effective in real-world deployments.

5 Methodology

In this work, we build upon our novel findings in Sec. 3 and propose realizing the attack goal in Eq. (4) by exploiting the attack capability of interval spacing. However, this is a non-trivial problem since ❶ Eq. (4) requires the patch’s ground-truth depth to be close but Fig. 2 (a) indicates that interval spacing exerts only a limited adversarial effect when the patch is deployed closely. Moreover, ❷ the robustness against rotation is a critical requirement for the patch to be physically attack effective. Yet we observed in Fig. 3 that the robustness provided by the naive interval strategy is rather limited especially against the rotation of Y axis. As a result, it is obvious that an advanced interval spacing strategy is required to realize our attack goal as defined in Eq. (4).

Fundamentally, interval spacing induces a mask $\mathbf{M}$ that partitions the patch $\mathbf{P}$ into interval structure $\mathbf{P}_s = \mathbf{M} \odot \mathbf{P}$ and texture content $\mathbf{P}_t = (1 - \mathbf{M}) \odot \mathbf{P}$, such that $\mathbf{P} = \mathbf{P}_s + \mathbf{P}_t$. Hence, we propose to optimize these components to reveal their adversarial effects. Beginning with the naive interval spacing strategy, and thus the mask $\mathbf{M}$, in Sec. 3, we first focus on optimizing the texture content $\mathbf{P}_t$, which composed of tiled texture elements $\mathbf{E}$, forming the basis of our Grid-based Attack. We then introduce the DepthVanish Attack, which jointly optimizes both $\mathbf{P}_s$ and $\mathbf{P}_t$ for maximal effect.

5.1 Grid-based Attack

In general, it is straightforward to setup an optimization pipeline for optimizing the texture element with grid intervals, where the patch is formed by repeating the texture elements over the grid. Fundamentally, there two main aspects that need to be considered: ❶ the physical constraint required for the texture element to form a patch and ❷ the objective function adopted for optimization.

Given our primary goal is to achieve physical attack effectiveness, the patch must comply with the physical geometry constraints during the optimization. Specifically, given a user pre-defined physical patch size (u, v) and physical distance to the camera e in meters, we first find the corresponding pixel size of the patch (h_p, w_p) with the help of stereo calibration information (See details in Sec. 6.1). Then, we empirically adopt the optimal interval width o and number of repetition k from Sec. 3 to determine the texture element $\mathbf{E}$ size (h_t, w_t) as

$$h_t = \frac{h_p - k \cdot o}{k + 1}, \quad w_t = \frac{w_p - k \cdot o}{k + 1}. \quad (5)$$

Based on the size of the texture element, the texture component $\mathbf{P}_t$, and consequently the full patch $\mathbf{P}$, is constructed by tiling the base texture unit $\mathbf{E}$ in a regular grid pattern as illustrated in Fig. 2(b) (bottom right). With the correctly assembled and deployed grid-based patch, we set the optimizing objective function as regional mean square error (rMSE) which is formulated as

$$\mathcal{L}_{rMSE} = \frac{1}{h_p \cdot w_p} \sum_{i=1}^{h_p} \sum_{j=1}^{w_p} (\mathcal{F}(\hat{\mathbf{I}}_l, \hat{\mathbf{I}}_r) - \mathcal{F}(\mathbf{I}_l, \mathbf{I}_r))^2. \quad (6)$$

Let $\mathcal{R}$ be the set of $k \times k$ grid locations on where $\mathbf{E}$ is repeated. The texture element is updated with average gradients: $\mathbf{E} \leftarrow \mathbf{E} - \eta \cdot \frac{1}{|\mathcal{R}|} \sum_{(i,j) \in \mathcal{R}} \nabla \mathbf{E} \mathcal{L}_{rMSE}^{(i,j)}$, where η is the learning rate. Gradients are only applied to the repeated texture regions while the interval areas remain untouched.

5.2 DepthVanish Attack

As we will see in Fig. 5, the above grid-based optimization can successfully mount an attack against various stereo systems but the results are still far from our attack goal defined in Eq. (4). Thus we further consider optimizing the interval structure $\mathbf{P}_s$ simultaneously during the updating of the texture element $\mathbf{E}$. Practically, optimizing the interval structure on patch level will break the texture repetitions as the interval will be updated to have irregular size. To keep the repetitions and incorporate the interval’s attack capability, we propose to jointly optimize the interval structure within the texture element and, following [1], tile the optimized texture element $\mathbf{E}$ to form the final patch.

Same to grid-based attack, given a user pre-defined patch physical size (u, v) and physical distance to the camera e in meters, we first find the corresponding pixel size of the patch (h_p, w_p) . Then we calculate the texture element size (h_t, w_t) by simply dividing (h_p, w_p) to the repetition times k . In order to optimize the texture element so that the interval structure integrated as part of the texture, we propose to regularize the texture element $\mathbf{E}$ during optimization with two objectives. First, we directly cast entropy constraint on the texture element for regularizing its values to be binary, so that a crisp separation is formed to serve as the required interval structure:

$$\mathcal{L}_{entropy} = \frac{1}{h_t \cdot w_t} \sum_{i=1}^{h_t} \sum_{j=1}^{w_t} -\mathbf{E}_{ij} \log(\mathbf{E}_{ij} + \epsilon) - (1 - \mathbf{E}_{ij}) \log(1 - \mathbf{E}_{ij} + \epsilon). \quad (7)$$

However, we experimentally found that the texture element cannot form a clear pattern with only entropy regularization. As a result, we further integrate the total variation loss to penalizes local pixel-level variation, encouraging the formation of smooth areas:

$$\mathcal{L}_{tv} = \frac{1}{h_t \cdot w_t} \sum_{i=1}^{h_t} \sum_{j=1}^{w_t} |\mathbf{E}_{i+1,j} - \mathbf{E}_{ij}| + |\mathbf{E}_{i,j+1} - \mathbf{E}_{ij}|. \quad (8)$$

With the entropy and total variant constraints, we arrived at an objective function that can shape a clearly interval pattern for the texture element. In summary, the overall objective function adopted for optimization is formulated as

$$\mathcal{L} = \mathcal{L}_{rMSE} + \alpha * \mathcal{L}_{entropy} + \beta * \mathcal{L}_{tv}, \quad (9)$$

where α and β are the hyper-parameters balancing the sharp border and coherent region requirements. Hence, we update with $\mathbf{E} \leftarrow \mathbf{E} - \eta \cdot \frac{1}{|\mathcal{R}|} \sum_{(i,j) \in \mathcal{R}} \nabla \mathbf{E} \mathcal{L}^{(i,j)}$ where η is the learning rate.

5.3 Implementation Details.

During the optimization for the Grid-based and DepthVanish attacks, we use a patch with a physical size $(u, v) = (0.891 \text{ m}, 1.26 \text{ m})$ and specify the physical ground-truth depth $e = 5 \text{ m}$. To assemble the texture element into a patch, we empirically set the number of repetition as 5 for horizontal and 4 for vertical, i.e., $k = (4, 5)$. For the optimization and corresponding evaluation results with different patch physical setup, we provide them in the supplemental material. When the patch is optimized as grid-based attack, the optimal interval size $o = 10 \text{ px}$ from Sec. 3 is applied. As for the loss weights adopted during the depth vanish attack, we keep setting $\alpha = 0.1$ and $\beta = 10$. Please find more details of implementation for both Grid-based and DepthVanish attack in the supplemental material.

6 Experiments

6.1 Experimental Setup

Dataset. For the evaluation of digital attack effectiveness, we adopt the stereo images from KITTI scene flow (KITTI-scene) [23] and DrivingStereo [35] datasets. Both datasets are composed of stereo images of urban traffic scenes where the image size of KITTI-scene is (1242, 375) and DrivingStereo

Table 1: Statistical attack performance of our DepthVanish, grid-based patch and existing baselines for PSMNet, DeepPruner, AANet, RAFT-Stereo and STTR on KITTI-scene dataset. The best results are highlighted in **bold**.

KITTI-scene	PSMNet		DeepPruner		AANet		RAFT-Stereo		STTR	
	D1	EPE	D1	EPE	D1	EPE	D1	EPE	D1	EPE
Stereoscopic Patch	6.23 $\pm$ 1.13	5.28 $\pm$ 0.88	8.29 $\pm$ 10.23	3.29 $\pm$ 2.05	6.79 $\pm$ 2.30	3.69 $\pm$ 0.39	5.79 $\pm$ 9.88	3.58 $\pm$ 6.73	4.58 $\pm$ 2.83	1.30 $\pm$ 5.37
Stereopagnosia Patch	2.17 $\pm$ 0.09	2.18 $\pm$ 0.58	5.40 $\pm$ 11.16	1.62 $\pm$ 2.33	3.42 $\pm$ 2.59	1.96 $\pm$ 0.44	4.18 $\pm$ 11.27	2.09 $\pm$ 12.66	3.02 $\pm$ 3.00	1.28 $\pm$ 8.79
Grid-based Patch (ours)	3.35 $\pm$ 1.09	48.21 $\pm$ 8.24	55.39 $\pm$ 10.77	38.60 $\pm$ 3.03	60.59 $\pm$ 8.90	53.84 $\pm$ 4.93	40.09 $\pm$ 5.87	67.24 $\pm$ 7.90	5.23 $\pm$ 7.49	45.34 $\pm$ 7.30
DepthVanish (ours)	55.30 $\pm$ 6.85	50.71 $\pm$ 9.71	97.07 $\pm$ 12.42	67.19 $\pm$ 4.85	66.42 $\pm$ 10.10	56.54 $\pm$ 5.26	89.31 $\pm$ 6.56	66.01 $\pm$ 6.18	92.38 $\pm$ 8.76	69.25 $\pm$ 6.62

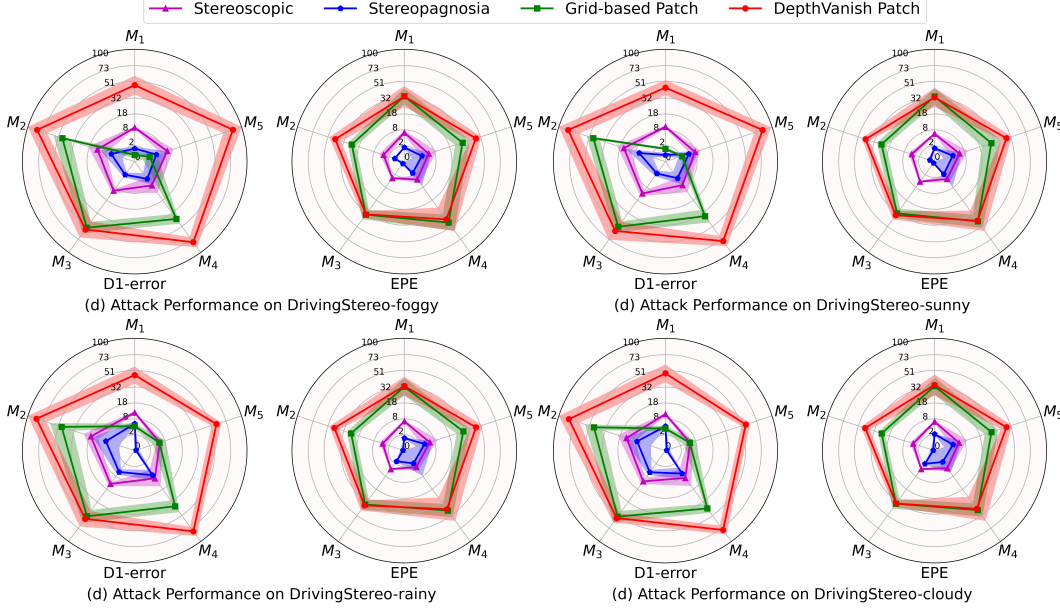


Figure 4: Attack performance of our DepthVanish, grid-based patch and existing baselines for PSMNet (M1), DeepPruner (M2), AANet (M3), RAFT-Stereo (M4) and STTR (M5) on the sub-sets of DrivingStereo dataset.

is (1758, 800). In more detail, we adopt the four sub-sets of DrivingStereo that were captured under different weather conditions (*i.e.*, sunny, foggy, rainy, cloudy) where we report the attack performance for each of them respectively. Following [1], 40 stereo image pairs for each (sub-)dataset are selected to verify the effectiveness of different patches. For the physical evaluation, we manually capture stereo images with i3DStroid² where various safety critical situations are considered. We refer readers to the supplemental material for the details of how the physical stereo images are captured and the pipeline we adopted for physical deployment.

Attack targets. Following [31, 1], we apply our attack method to PSMNet [4], DeepPruner [10] and AANet [33] for validating the general attack effectiveness. Moreover, we empirically found that they are out-of-date and can be easily disturbed, thus we further select RAFT-Stereo [19] and Stereo TRansformer (STTR) [17] which represent the promising iterative optimization-based methods and transformer-base methods as our main attack targets. For the detail hyper-parameter setting and the pretrained checkpoint adopted during the attack, please find all of them in supplemental material.

Digital deployment. During the digital optimization and evaluation, the patch needs to be placed inside the scene according to physical constraints. To achieve this, we apply the calibration information provided by the KITTI and DrivingStereo dataset. In specific, given a patch with a predefined physical size in meters, we first set the homogeneous 3D coordinates of the patch’s corners with respect to the reference camera coordinate system. Then we calculate the corresponding pixel coordinates with the help of the rectified projection and rotation matrix. The full calculation is detailed in supplement.

Evaluation metrics. Following the convention, we adopt bad pixel error (D1-error) and End-Point Error (EPE) for evaluating the prediction performance which are calculated as follows:

$$D1 = \frac{\# \text{ of bad pixels}}{\# \text{ of total pixels}} \times 100\%, \quad EPE = \frac{1}{N} \sum_{i=1}^N |\mathbf{d}_{pred}^i - \mathbf{d}_{gt}^i|, \quad (10)$$

²<http://stereo.jpn.org/eng/iphone/help/index.html>

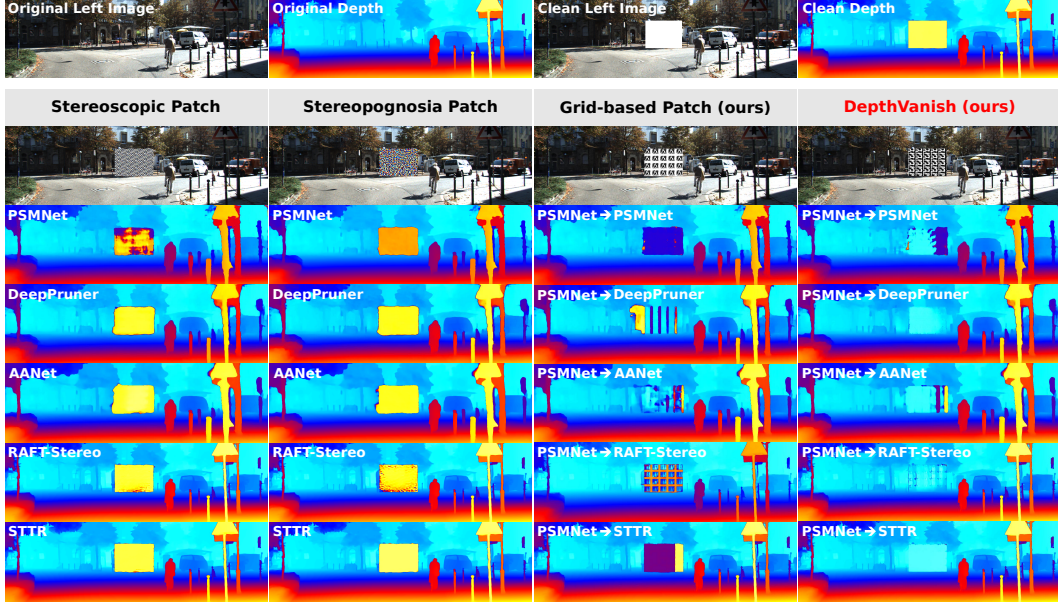


Figure 5: Visualization of different digital patch attack baselines and our DepthVanish patch against different target models on KITTI-scene dataset. Note that the original and clean depth are estimated by RAFT-Stereo.

where the bad pixel is one that satisfy $|\mathbf{d}_{pred} - \mathbf{d}_{gt}| > \max(3, 0.05 \cdot \mathbf{d}_{gt})$. To evaluate patch attack effectiveness, we first follow Eq. (4) to set the ground-truth disparity of the patch as $\mathbf{d}_{gt} = \mathbf{c}$. Then, we define the bad pixels as those satisfying $|\mathbf{d}_{pred} - \mathbf{c}| > \max(3, 0.05 \cdot \mathbf{d}_{gt})$ and $|\mathbf{d}_{pred} - \mathbf{0}| < \frac{\mathbf{c}}{n}$, where n defines how many times deeper than the actual depth will a patch be considered to be attack effective. In summary, we report the average D1-error and EPE with standard deviation where higher values indicate better attack performance.

6.2 Digitally Attack Stereo Estimator

We first conduct digital attack experiments with our proposed DepthVanish patch on KITTI-scene dataset and the four sub-sets of DrivingStereo, *i.e.*, sunny, foggy, rainy, cloudy.

Setting: Due to the lack of existing works on attacking stereo matching using patches, we use the results from existing digital attack studies (*i.e.*, Stereoscopic [1] and Stereopagnosia [31]) as patches and deploy them into the scene as the first set of baselines. However, it should be noted that such comparison is not fair enough as existing works [1, 31] are not specifically designed for patch attack. Thus we further setup our own baseline (*i.e.*, grid-based patch from Sec. 5.1) for a fair comparison.

Results: ① We report the attack results for the five attack target models on KITTI-scene dataset in Tab. 1. It can be seen that existing digital attacks are ineffective under the patch attack setup, while our Grid-based Patch significantly outperforms them. Notably, DepthVanish achieves strong attack performance, especially against DeepPruner, RAFT-Stereo and STTR. We illustrate the results on DrivingStereo dataset in Fig. 4. It is evident that similar attack performance can be observed on all four sub-sets. ② In addition to the standard evaluation, Fig. 5 shows a KITTI sample comparison. Compared to the Clean Depth, we first note that existing attack works fail to mislead all the five target models, where only the Stereoscopic Patch shows limited influence against PSMNet. However, as the results shown in the last column, our DepthVanish patch casts strong influence where it almost disappeared within the depth results. More surprisingly, our patch enjoys significant transferability over models where the patch optimized with PSMNet shows strong attack effect on other four models. Based our experimental experience, all patches with such clear interval patterns are transferable across models, a capability we attribute to the insights analyzed in Sec. 3. Please refer to the supplement for the comprehensive experimental results of attack transferability.

6.3 Physically Attack Stereo Estimator

In this section, we conduct physical evaluation for our DepthVanish patches that optimized with different stereo estimators to highlight the importance and emergency of research on stereo matching

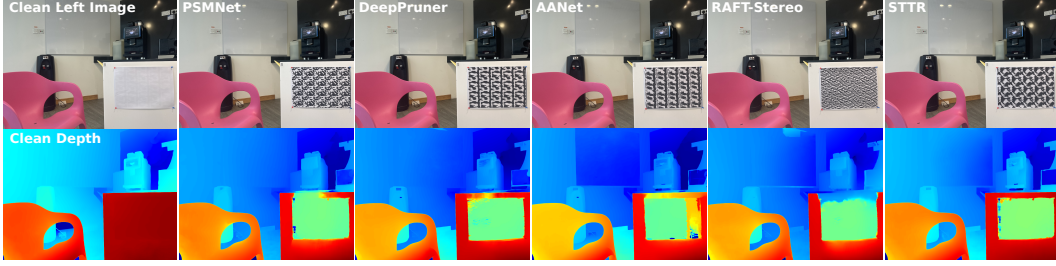


Figure 6: Visualization of physical attack results of our DepthVanish patches against different stereo depth estimators. Note that the clean depth is estimated with RAFT-Stereo.

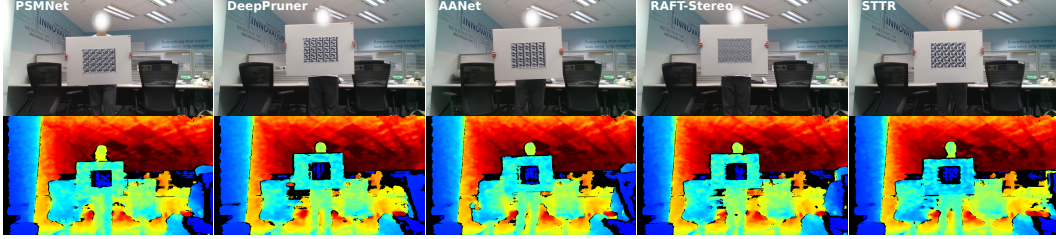


Figure 7: Visualization of DepthVanish attack performance against Intel RealSense depth camera (D435i).

reliability. As shown in Fig. 6, we host our DepthVanish patches on a white board for the purpose of highlighting the depth inconsistency. ❶ It can be observed from the results that our DepthVanish patch consistently preserves its attack effectiveness after deployed into the physical environment. Compared to the Clean Depth, the board region occupied by our DepthVanish patches are predicted as far away in general. ❷ However, it should be noted that the induced depth error are limited compared to the digital effectiveness in Fig. 5. We ascribe such performance degradation to the lighting variation and imprecise photo-capturing process, where the left and right images are captured manually and separately. Therefore, we further conduct evaluation for our patch against a commercial stereo depth camera in the next section. In summary, despite of the imprecise stereo image capturing process, our DepthVanish patches successfully attack advanced DNN-based stereo estimators with consistency.

6.4 Attack Commercial Stereo System

To further assess the practicality and robustness of our DepthVanish patch, we evaluate its performance on a commercial stereo camera system, specifically the Intel RealSense D435i depth camera. We focus on evaluating the patch’s robustness from three aspects: model generalization, viewing orientation, and distance variation. ❶ **Model generalization:** we deploy the patches that optimized with five stereo models over KITTI dataset and evaluate their attack effective against D435i camera. As shown in Fig. 7, the patch consistently disrupts D435i predictions regardless of which model is optimized for, demonstrating strong attack transferability. ❷ **Orientation robustness:** we physically rotate the patch along the X and Y axes (see Fig. 3(a)). As visualized in Fig. 8, the patch (optimized with PSMNet on KITTI) remains effective under different viewing angles, confirming its robustness to rotation. ❸ **Distance robustness:** our method also shows robustness under varying distances. Corresponding visual results are provided in the supplementary material.

6.5 Ablation Study

In this section, we conduct ablation analysis on the DepthVanish attack to assess the impact of the hyperparameters α and β in the objective function of Eq. (9). As shown in Fig. 9, both parameters are critical for optimal attack performance. Specifically, it can be seen that the performance degraded significantly when $\alpha = 0$, i.e., the $\mathcal{L}_{entropy}$ is removed from Eq. (9), which highlights the importance

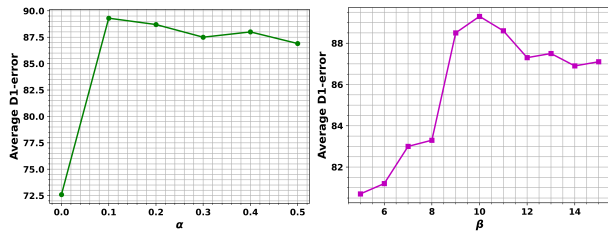


Figure 9: Attack performance of DepthVanish against RAFT-Stereo under different α and β on KITTI dataset.

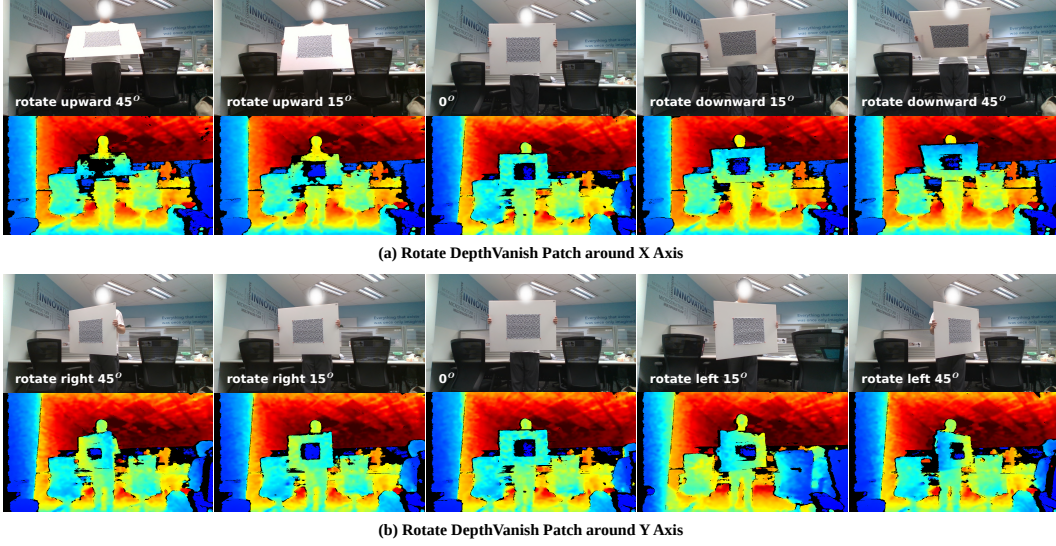


Figure 8: Visualization of DepthVanish attack performance with different rotation degrees around both X and Y axes against Intel RealSense depth camera (D435i).

of the clear interval spacing for attack effectiveness. Moreover, the total variation constraint $\mathcal{L}_{tv}$ is also important where a clear performance degradation can be observed when β decreases below 9. In summary, the synergistic combination of entropy and total variation regularization effectively ensures that our DepthVanish patches achieve the maximal attack performance

7 Conclusion

In this work, we present DepthVanish, a significant advancement in physical adversarial attack that jointly optimizes both texture element and interval structure of a patch to fool stereo depth estimation systems. By thoroughly analyzing the influence of regular spacing on naive texture repetition, we introduce a novel insight into enhancing the attack effectiveness and digital-to-physical transferability of the patch. To demonstrate the potentially dangerous consequences of depth estimation failure, we design the patch to be "disappear", where the patch is estimated as far away despite being physically close. Unlike previous methods limited to digital environments, our approach succeeds in both digital and physical settings, when evaluated against widely applied depth estimation models and commercial RGB-D cameras. These findings reveal critical vulnerabilities in current depth estimation technologies and raise concerns about their reliability in safety-critical autonomous systems.

Acknowledgments and Disclosure of Funding

This research was supported by Shenzhen Science and Technology Program (No. JCYJ20240813114237048), "Science and Technology Yongjiang 2035" key technology breakthrough plan project (No. 2025Z053). This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG4-GC-2023-008-1B), and National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore, and Infocomm Media Development Authority. This work is also supported in part by Canada CIFAR AI Chairs Program, the Natural Sciences and Engineering Research Council of Canada, and JST-Mirai Program Grant No.JPMJMI20B8, JSPS KAKENHI Grant No.JP21H04877, No.JP23H03372, No.JP24K02920.

References

- [1] Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, and Alex Wong. Stereoscopic universal perturbations across different architectures and datasets. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition*, pages 15180–15190, 2022.
- [2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *CoRR*, abs/2001.10773, 2020.
 - [3] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.
 - [4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
 - [5] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Adversarial attacks on monocular pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12500–12505. IEEE, 2022.
 - [6] Junda Cheng, Wei Yin, Kaixuan Wang, Xiaozhi Chen, Shijie Wang, and Xin Yang. Adaptive fusion of single-view and multi-view depth for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10138–10147, June 2024.
 - [7] Zhiyuan Cheng, Cheng Han, James Liang, Qifan Wang, Xiangyu Zhang, and Dongfang Liu. Self-supervised adversarial training of monocular depth estimation against physical-world attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9084–9101, 2024.
 - [8] 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 *European Conference on Computer Vision (ECCV)*, 2022.
 - [9] Zhiyuan Cheng, James Liang, Guanhong Tao, Dongfang Liu, and Xiangyu Zhang. Adversarial training of self-supervised monocular depth estimation against physical-world attacks. In *The Eleventh International Conference on Learning Representations, (ICLR)*, 2023.
 - [10] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4384–4393, 2019.
 - [11] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H. Taylor, Mathias Unberath, Alan L. Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision (ECCV)*, 2022.
 - [12] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [13] Junjie Hu and Takayuki Okatani. Analysis of deep networks for monocular depth estimation through adversarial attacks with proposal of a defense method. *arXiv preprint arXiv:1911.08790*, 2019.
 - [14] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
 - [15] Dehong Kong, Siyuan Liang, Xiaopeng Zhu, Yuansheng Zhong, and Wenqi Ren. Patch is enough: naturalistic adversarial patch against vision-language pre-training models. *Visual Intelligence*, 2(1):33, 2024.
 - [16] Jie Li, Peidong Wang, Pengfei Xiong, Tao Cai, Zeguo Yan, Lei Yang, Jiawei Liu, Huan Fan, and Shuang Liu. Crestereo: Practical depth from stereo via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
 - [17] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021.
 - [18] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy S. Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

- [19] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.
- [20] Hangcheng Liu, Zhenhu Wu, Hao Wang, Xingshuo Han, Shangwei Guo, Tao Xiang, and Tianwei Zhang. Beware of road markings: A new adversarial patch attack to monocular depth estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [21] Baoli Lu, Liang Sun, Lina Yu, and Xiaoli Dong. An improved graph cut algorithm in stereo matching. *Displays*, 2021.
- [22] Yamin Mao, Zhihua Liu, Weiming Li, Yuchao Dai, Qiang Wang, Yun-Tae Kim, and Hong-Seok Lee. Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [23] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018.
- [24] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [25] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proceedings of German Conference on Pattern Recognition*, 2014.
- [26] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.
- [27] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision (ECCV)*, 2022.
- [28] Qing Su and Shihao Ji. Chitransformer: Towards reliable stereo from cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [29] Fabio Tosi, Luca Bartolomei, and Matteo Poggi. A survey on deep stereo matching in the twenties. *International Journal of Computer Vision*, 2025.
- [30] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020.
- [31] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2879–2888, 2021.
- [32] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatoughi, Fisher Yu, and Dacheng Tao. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [33] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1959–1968, 2020.
- [34] Gengshan Yang, Joshua Manela, Michael Hapgood, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019.
- [35] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [36] 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*, 2024.
- [37] 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 (NeurIPS)*, 2024.
- [38] Gyungeun Yun, Kyungho Joo, Wonsuk Choi, and Dong Hoon Lee. Poster: Unveiling the impact of patch placement: Adversarial patch attacks on monocular depth estimation. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS*, 2023.

- [39] Feihu Zhang, Victor Adrian Prisacariu, Ruigang Yang, and Philip H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [40] Tong Zhao, Mingyu Ding, Wei Zhan, Masayoshi Tomizuka, and Yintao Wei. Depth-aware volume attention for texture-less stereo matching, 2024.
- [41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2024.
- [42] Tianyue Zheng, Jingzhi Hu, Rui Tan, Yinqian Zhang, Ying He, and Jun Luo. pi-jack: Physical-world adversarial attack on monocular depth estimation with perspective hijacking. In *33rd USENIX Security Symposium, USENIX Security*, 2024.
- [43] Ce Zhou, Qiben Yan, Yan Shi, and Lichao Sun. {DoubleStar}:{Long-Range} attack towards depth estimation based obstacle avoidance in autonomous systems. In *31st USENIX security symposium (USENIX Security 22)*, pages 1885–1902, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims in the abstract and introduction are fully supported by both theoretical and experimental results throughout the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of our approach are discussed in the supplementary material.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not contain formal theorems or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 5.3 and Sec. 6.1 provide all critical information needed to reproduce the main experimental results. Additional instructions and settings are included in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the anonymized code and sufficient instructions for reproducing our main experimental results in the supplementary material upon submission. This ensures that reviewers and future readers can faithfully replicate and verify the findings.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all relevant training, evaluation, and test details in Sec. 5.3 and Sec. 6.1. Additional information is provided in supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports error bars (mean $\pm$ standard deviation) for the quantitative results and describes how they are calculated. Variability is reported, and figures or tables are referenced in the text to clarify the sources of error.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The information about compute resources is detailed in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research follows the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The potential positive and negative societal impacts of this work are discussed in detail in the supplementary material.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no high risk of dual use or misuse upon release, as it focuses on methods, not on releasing highly sensitive models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All used datasets and code are properly cited and are released under their respective public licenses as referenced in the paper and supplement.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new datasets or code assets. All methods are described for reproduction only.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or research on human subjects is involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No experiments requiring IRB approval are conducted or reported.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The research, methodology, and results do not make use of LLMs as a core component.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.