The 3D Mirage: Probing and Taming 3D Hallucinations

Hoang Nguyen*, Xiaohao Xu*†, and Xiaonan Huang

University of Michigan, Ann Arbor, USA
* Equal Contribution † Project Lead

Abstract. Monocular depth foundation models achieve remarkable generalization by learning large-scale semantic priors, but this creates a critical vulnerability: they hallucinate illusory 3D structures from planar/low-curvature but perceptually ambiguous inputs. We term this failure the **3D Mirage**. This paper introduces a novel end-to-end framework to **probe**, **score**, and **tame** this under-quantified safety risk *in monocular depth under context variation*. To **probe**, we present **3D-Mirage**, the first benchmark to combine context variation and precise annotation for real-world illusions with real object exclusions, multi-surface support; *purpose-built to stress-test monocular depth on real-world illusions*. To **score**, we propose a second-order magnitude-based evaluation with two metrics: the **Deviation Composite Score (DCS)** for high second-order 3D structure and the **Confusion Composite Score (CCS)** for contextual instability. To **tame** this failure, we introduce **Grounded Self-Distillation**, a parameter-efficient strategy on Depth-Anything-V2 baseline that surgically targets and resolves hallucination on illusion ROIs while preserving background knowledge, avoiding catastrophic forgetting. Our work provides an innovative pipeline for diagnosing and addressing this phenomenon, urging a necessary shift in the evaluation of MDE from pixel-wise accuracy to structural and contextual robustness.

Code: https://github.com/hdnndh/The-3D-Mirage-Probing-and-Taming-3D-Hallucinations
Dataset: https://huggingface.co/datasets/3dmirage/3D-Mirage

1 Introduction

Enabling reliable perception and reconstruction of 3D scenes is critical for safe and robust visual intelligence [32–36,44,45,61] and autonomous driving experience [37]. Driven by this necessity, Monocular Depth Estimation (MDE) has transitioned from a challenging academic problem to a core perception component in real-world systems. This rapid adoption is fueled by powerful foundation models such as Depth-Anything V2 [63], Zoe-Depth [3], and MiDaS/DPT [47,48], which are trained on massive, diverse datasets. However, their remarkable zero-shot generalization obscures a critical and unexamined vulnerability: an over-reliance on large-scale statistical priors, trading geometric fidelity for semantic consistency, making them susceptible to perceptual ambiguity and adversarial attack.

In this work, we identify and analyze a critical failure mode we term the **3D Mirage**. We find that SOTA depth foundation models fail in two common, safety-critical scenarios: 1) when presented with perceptually ambiguous 2D patterns, such as 3D street art, and 2) when operating under a restricted field-of-view (FOV) that removes broad contextual cues. Fig. 1 shows this: the same flat road section reads as planar in the full scene, yet yields a significant non-existent 3D structure once the view is restricted to it. The trigger of this phenomenon is the illusory texture on the low-curvature carrier surface, causing hallucination of 3D structure. Furthermore, a faithful predictor should report the same flat geometry however the context varies, and the change in context exposes a core dependency of hallucination. This is a failure of contextual grounding: the model's depth prediction is not anchored in local geometric reality, but is instead a fragile artifact of priors shaped by large-scale training.

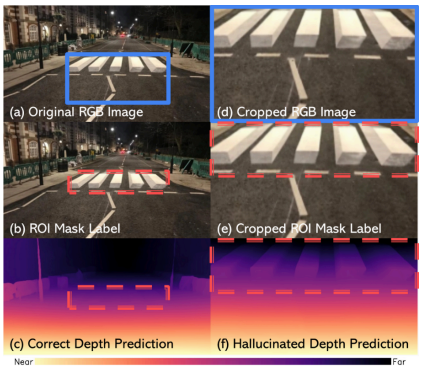

Fig. 1: The 3D Mirage: Hallucinations induced by illusory texture patterns. (**a**) A driving scene featuring a deceptive non-existent 3D structure. (**b**) With full global context, the depth foundation model [63] correctly identifies the road as planar(**c**). (**d-f**) However, when the view is restricted to the local region, the model fails to disambiguate the texture from geometry. It hallucinates significant non-existent 3D obstacles (**f**) from the phantom pattern, illustrating a critical vulnerability in reliable 3D perception for autonomous driving scenarios.

This failure is not an isolated anecdote. We demonstrate that this vulnerability is systemic across the current generation of leading models. As shown in Fig. 2, we subjected a wide range of architectures, from transformer-based (Depth-Anything V2 [63]) and diffusion-based (Marigold [28]) to generative (DepthFM [20]) and commercially-developed (Depth Pro [6]), to these 3D mirage inputs. All models exhibited similar failures, unstably predicting spurious 3D structures from low-curvature surfaces.

This collective failure exposes a critical gap in *how* we evaluate these models. Standard metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) [8] are *perceptually-blind* to these structural failures. By averaging pixel-wise errors, they cannot differentiate between a slight, uniform mis-calibration and a massive, hallucinated obstacle. We posit that MDE evaluation must evolve to assess ***structural integrity*** and ***contextual stability***, which are far more critical for real-world deployment than pure pixel accuracy.

To address this, our work provides the first end-to-end framework to systematically **probe**, **score**, and **tame** 3D hallucinations. Our contributions are threefold:

– We **probe** this vulnerability by introducing **3D-Mirage**, a benchmark purpose-built for paired full vs. context-varied evaluation of monocular depth on planar/low-curvature illusion carriers, with manually annotated ROI poly-

gons supporting ***multiple illusions*** per scene, ***real object exclusions***, and ***illusion spanning surfaces***.
- We **score** these failures by proposing a novel **Second-order magnitude-based evaluation framework**, introducing two metrics: the **Deviation Composite Score (DCS)** to measure spurious second-order structure (hallucination intensity) and the **Confusion Composite Score (CCS)** to measure contextual instability (*i.e.*, the mirage effect).
- We **tame** these hallucinations with a novel **Grounded Self-Distillation** (GSD) strategy. By applying low-parameter adapters to the model's encoder, we use our benchmark to suppress spurious curvature on illusion ROIs while using the frozen teacher model to enforce alignment on stable background and border regions. This efficiently grounds the model, mitigating hallucinations without catastrophic forgetting of its core pre-trained knowledge.

Ultimately, our contributions provide the essential tools: a targeted benchmark, perceptually-aware metrics, and an efficient mitigation strategy to advance MDE from simple geometric accuracy to the structural and contextual robustness demanded by safety-critical applications.

2 Related Works

2.1 2D and 3D Visual Hallucination

Visual hallucination, predicting content not present in the input, is a known failure in 2D. This includes classifying nonsense images [43], detecting objects in empty locations [27, 41], or segmenting non-existent structures [10, 30]. Such failures, perilous in safety-critical domains [31, 67], are linked to over-parameterization and models overly relying on context over evidence [10, 29, 51, 52]. In 3D, this problem is less studied but more complex. We define 3D hallucination as predicting depth variations on geometrically flat or low-curvature surfaces [42]. This is exacerbated by the ill-posed nature of MDE: the 3D-to-2D projection discards depth [49], forcing networks to use learned priors to resolve ambiguity. This can yield multiple valid reconstructions [4, 9] or overfitting to dataset- or camera-specific biases like texture cues [11]. Consequently, most MDE literature has focused on geometric accuracy rather than characterizing these structural failures.

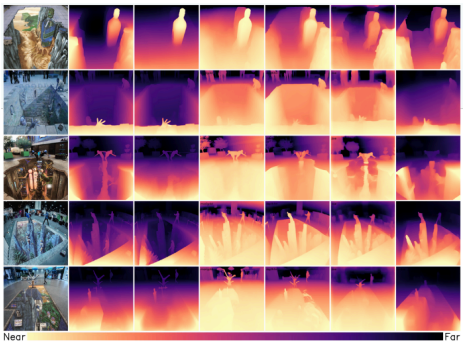

Fig. 2: Hallucinations across SOTA monocular depth models on images. Given an optical-illusion region or a view with restricted context, all tested monocular depth foundation models (DAv2 [63], Depth Pro [6]), Marigold [28], DepthFM [20], ZoeDepth [3], MiDaS [48], predict *spurious depth variation/3D structure.*

2.2 MDE Models and Benchmarks

Monocular Depth Estimation (MDE) seeks to recover 3D structure from a single RGB image. Early methods evolved from supervised [15] to self-supervised using geometric constraints [18, 19, 68]. The field has recently shifted toward large foundation models like Depth Anything (DAv2) [62, 63], ZoeDepth [3], MiDaS/DPT [48], and Marigold [28], Depth Pro [6], DepthFM [20]. Trained on broad data, these models achieve remarkable zero-shot generalization but rely heavily on statistical priors. This reliance enables them to fill in depth in ambiguous or deceptive regions [58, 60, 67], trading geometric fidelity for semantic robustness.

However, existing MDE benchmarks are insufficient for probing this failure mode. Mainstream datasets (KITTI [17] [40] [16], NYUv2 [50], ScanNet [14], Depth-in-the-Wild [11]) emphasize *geometrically-consistent* scenes, lacking the "perceptual traps" to trigger 3D hallucinations. Adversarial datasets are also limited: TartanAir-Adv [56] uses synthetic motion. Other illusion datasets have different emphases: Booster [46] focuses on specular/transparent surfaces, while MonoTrap [2] contains a limited indoor stereo setting (26 scenes). These illusion datasets lack systematic FOV variation. Yao et al. [64] recently introduced the 3D Visual Illusion Depth Estimation dataset and a VLM-driven framework that fuses binocular disparity with monocular depth. While this fusion improves robustness against visual illusions, it incurs substantial computational overhead, primarily from VLM. Furthermore, their results indicate that standard stereo baselines already perform exceptionally well on texture-rich, non-specular/transparent illusions (e.g., inpainting, pictures, replays, and holography) because of cross-view texture consistency. Ultimately, their primary advantage lies in using learned monocular priors to resolve mirror geometries. Their dataset focuses on 2D planar surface and evaluates on small-scale table-top and indoor wall scenes [64]. In contrast, our benchmark consists of 468 indoor and outdoor texture illusions, both small and large scale scenes. We also specifically target a complementary failure mode in *monocular* foundation models: *context variation* (paired full-context and context-altered views) that induces unstable 3D hallucinations on low-curvature regions. To our knowledge, **3D-Mirage** is the first benchmark centered on real-world low-curvature-carrier optical-illusion scenes that introduces controlled context variation (paired full-context and context-altered views) together with precise annotation for complex illusion cases (Fig. 3) to systematically evaluate hallucination and contextual stability in monocular depth.

2.3 Probing and Mitigating 3D Hallucination

Early probes of MDE hallucination used textured transparent surfaces [13] or scored failures from a semantic angle [30, 39, 67]. However, these methods rarely localize the hallucination or quantify it systematically. Existing defenses are often model-specific and generalize poorly [21, 23, 26, 31]. To date, 3D hallucination remains difficult to label and formally quantify [42], limiting systematic study.

While hallucinated 3D content can be viewed as a form of 3D anomaly or out-of-distribution 3D content, current approaches focus mainly on geometric 3D anomaly detection. Semantic 3D anomalies remain underexplored, and this work aims to address such semantic anomalies. Given the scale of modern foundation models, full fine-tuning to correct such failures is prohibitive and risks catastrophic forgetting. Parameter-Efficient Fine-Tuning [24] (PEFT) methods like Low-Rank Adaptation (LoRA) [25] offer an alternative. LoRA freezes the model and injects small, trainable low-rank matrices, allowing efficient adaptation. We are the first to explore PEFT to tame 3D hallucinations with Depth-Anything-V2 as our baseline. We hypothesize that using a targeted benchmark, we can employ LoRA to ground a depth model, teaching it to ignore illusory 2D cues while preserving its pre-trained knowledge.

3 The 3D-Mirage Benchmark

To systematically probe the '3D Mirage' vulnerability, we introduce **3D-Mirage**, a benchmark purpose-built to elicit and measure 3D hallucinations in monocular depth models under *illusory* and *context-varied* conditions. The benchmark is designed not to test average-case accuracy, but to specifically target the failure modes where learned priors override geometric reality.

3.1 Dataset: Curation and Properties

The creation of 3D-Mirage involved a three-stage pipeline:

Data Collection. We first collected 468 real-world RGB images featuring *painting* and *street-art* 3D illusions across varied scenes. These include chalk anamorphoses, forced-perspective murals, and large-format advertisements that create a strong perceptual suggestion of 3D geometry on a 2D plane. Approximately 80% of scenes are outdoor, 40% lie on pedestrian walkways, 8% are billboard/advertisement-like cases, 16% span multiple support surfaces, and about 40% contain real objects or people near or on top of the illusion.

ROI Annotation. After filtering, we manually annotated precise polygonal Region of Interest (ROI) masks for each illusion. If real objects are inside illusion, nested ROIs are used to mark and exclude them. These masks delineate regions that are low-curvature support surfaces yet *suggest 3D structure in appearance*. This mask is the key component for our geometry-based evaluation.

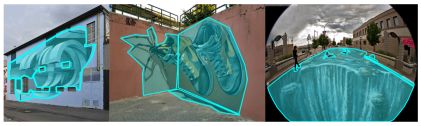

Fig. 3: Visualization of ROI annotation. We annotate the hallucination regions with individual polygon per surface ***(2,3)*** and nested rule to exclude real objects ***(1,3)***. This works well with our plane-mixture and gating process in training.

Context-Variation Augmentation. To emulate the limited FOV and partial occlusions common in autonomous driving, we generated four random crops for each sample. These crops are centered on the

ROI and retain at least 40% of the ROI diagonal, ensuring the illusion is present but the surrounding scene context is partially or fully absent.

Statistics. Each benchmark instance consists of a full-context image, its illusion ROI mask(s), and one of its four associated context-varied crops. The final benchmark contains **1,872** full-crop instances, all annotated and verified by human annotators. The dataset is designed to provide a challenging test of model robustness. As shown in Fig. 4, the illusion ROIs are a significant part of the image, covering an average of 49% of the total area. The context-varied crops are tighter, covering an average of 41% of the original image.

3.2 Evaluation: Quantifying Hallucinations

A core component of our benchmark is an evaluation framework that complements standard metrics, especially when ***GT is not available or hard to collect at scale***, to quantify hallucination and confusion. DCS/CCS are not intended to replace ground-truth depth metrics; they are reference-free structural probes for the annotated low-curvature carrier regions and are used together with standard depth and preservation evaluations in Sec. 5.

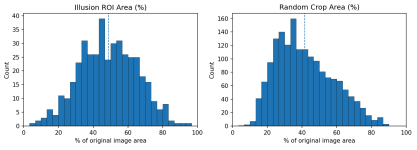

Fig. 4: Statistics of illusion regions in the 3D-Mirage dataset. Area distributions for illusion regions (left) and their corresponding random crops (right), as a percentage of the original image area. The dotted vertical line denotes the average value.

Shortcomings of Standard Metrics. Standard metrics (MAE, RMSE, REL) dilute ROI-specific failures and evaluate views independently. *Even when restricted to the ROI, they require GT depth and do not measure full–crop consistency for the same physical region.* Scale-invariant (SI-log) and AbsRel objectives share this gap; they score pixel-wise agreement with ground-truth depth, not the *second-order curvature* that defines a 3D mirage, so a hallucinated bulge that preserves ordinal depth can satisfy them; DCS and CCS instead quantify ROI curvature directly and without GT.

Dual-View Projection Space. Let f_θ be an MDE model. For each benchmark instance i with full image x_{full}, crop x_{crop} defined by crop box b_i, and ROI polygons, we compute $D_{\text{full}} = f_\theta(x_{\text{full}})$ and $D_{\text{crop}} = f_\theta(x_{\text{crop}})$. Using b_i, we crop and resize the full-view depth prediction to the crop frame, obtaining $\tilde{D}_{\text{full}\to\text{crop}}$, and rasterize the union of ROI polygons into a crop-frame mask M_i. For context variation, the full-image branch D_{full} is itself a *context-expansion baseline* of the crop view. An image-edited, non-illusory counterpart at matched scene context and crop support scores only DCS ≈ 54 and CCS $\approx 1.4 \times 10^{-4}$, with added Gaussian noise lifting DCS to at most ≈ 65, whereas the illusion reaches DCS ≈ 861 and CCS $\approx 7.4 \times 10^{-4}$ (a $\sim$16$\times$ DCS gap). DCS and CCS therefore stay insensitive to alignment artifacts, crop geometry, and mild pixel-level noise, and the full control study is in the supplementary material.

ROI-normalized Second-order magnitude-based responses. Let $\text{QN}(\cdot\,; M)$ denote ROI-conditional 1–99% quantile normalization (with quantiles computed on ROI pixels) and let $L(\cdot)$ be the second-order magnitude operator. We define

$$R_i^{\text{full}} = L(\text{QN}(\tilde{D}_{\text{full}\to\text{crop}}; M_i)), \qquad R_i^{\text{crop}} = L(\text{QN}(D_{\text{crop}}; M_i)).$$

For $v \in \{\text{full}, \text{crop}\}$, R_i^v denotes the ROI-normalized second-order response map for branch v. From R_i^v, we compute two ROI scalars within M_i: **top10sum** $t_{v,i}$, sum of the largest 10% response values; and **mean10** $\mu_{v,i}$, a 10%-trimmed high-response mean computed over the largest 90% response values.

Deviation Composite Score (DCS). Let $\mathbf{t}_i = [t_{\text{full},i},\, t_{\text{crop},i}]^\top \in \mathbb{R}^2$ and $\bar{\mathbf{t}} = \frac{1}{N}\sum_i \mathbf{t}_i$. We define

$$d_{\text{cluster}} = \|\bar{\mathbf{t}}\|_2, \qquad d_{\text{avg}} = \frac{1}{N}\sum_i \|\mathbf{t}_i\|_2, \qquad \text{DCS} = d_{\text{cluster}} + d_{\text{avg}}. \tag{1}$$

Confusion Composite Score (CCS). Let $\boldsymbol{\mu}_i = [\mu_{\text{full},i},\, \mu_{\text{crop},i}]^\top$ and $\bar{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \boldsymbol{\mu}_i$. Let $\mathbf{u} = \frac{1}{\sqrt{2}}[1, -1]^\top$ be the off-diagonal unit direction. We define

$$D_{\text{cluster}} = |\mathbf{u}^\top \bar{\boldsymbol{\mu}}|, \qquad D_{\text{avg}} = \frac{1}{N}\sum_i |\mathbf{u}^\top \boldsymbol{\mu}_i|, \qquad \text{CCS} = D_{\text{cluster}} + D_{\text{avg}}. \tag{2}$$

A low DCS indicates little spurious second-order structure inside the annotated low-curvature carrier, while a low CCS indicates that the same physical region remains stable between full and context-restricted views.

4 Methodology: Taming 3D Mirages

4.1 Problem Definition

Let f_θ be a pre-trained monocular depth estimation foundation model with weights θ. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, the model produces a dense depth map $D = f_\theta(x)$, where $D \in \mathbb{R}^{H\times W}$. We define a "3D Mirage" as a failure mode characterized by two conditions, identified using a benchmark dataset $\mathcal{D}$. Each sample in $\mathcal{D}$ consists of a full-context image x_{full}, a context-restricted crop x_{crop}, and a binary mask m defining a Region of Interest (ROI) that serves as a *locally low-curvature carrier* (i.e., physically smooth or piecewise-smooth).

Failure modes. We consider the following two failure modes of models:

1. **Geometric Hallucination (Deviation):** The model predicts spurious high-curvature 3D structure inside the low-curvature ROI. We diagnose this using a fixed second-order operator $\mathcal{L}$ (separable second-difference magnitude), where a geometrically consistent prediction should satisfy $\mathcal{L}(D) \odot m \approx 0$. A high response indicates a geometric hallucination.
2. **Contextual Instability (Confusion):** The model's prediction for the *same* physical region changes significantly when surrounding context is altered. Let $D_{\text{full}} = f_\theta(x_{\text{full}})$ and $D_{\text{crop}} = f_\theta(x_{\text{crop}})$. From known crop box b_i, we resample the full-view depth prediction into the crop coordinate frame and compare it to D_{crop} on the ROI pixels (mask M_i in the crop frame). We call the model contextually unstable when the predicted non-existent 3D structure inside the ROIs differ substantially.

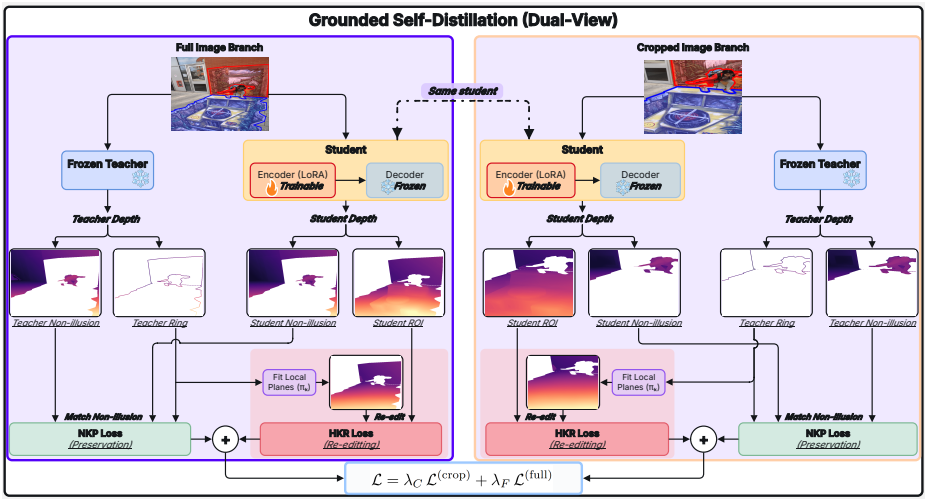

Fig. 5: Grounded Self-Distillation (Dual-View). For each illusion sample, we process both the **full image** (left) and a **context-restricted crop** (right). A **frozen teacher** predicts depth for each view, while a single **student** (shared weights across both branches) predicts depth using a **trainable LoRA-augmented encoder** and a **frozen decoder**. In each branch, we separate the prediction into a **non-illusion/background region** (and an ROI-adjacent **ring**) and the **illusion ROI**. The **Non-hallucination Knowledge Preservation (NKP)** loss distills the teacher's stable geometry onto the student on non-illusion/ring regions to prevent catastrophic forgetting. The **Hallucination Knowledge Re-editing (HKR)** loss suppresses spurious structure inside the ROI by penalizing second-order structure, and re-targeting each ROI instance toward **local surface hypotheses** fitted from the teacher's ring neighborhood. The final objective combines both views.

Desired Properties. Our goal is to learn a parameter-efficient adaptation $\Delta\theta$ for f_θ, producing an adapted model $f_{\theta'}$ with $\theta' = \theta + \Delta\theta$. Rather than directly optimizing the evaluation metrics, we target the following three *conceptual* properties:

1. **ROI grounding:** suppress spurious curvature inside the ROI, i.e., encourage low $\mathcal{L}(f_{\theta'}(x))$ on pixels in m.
2. **Context stability:** make the ROI prediction invariant to context removal, so that full-image and aligned crop predictions agree on m.
3. **Knowledge preservation:** preserve the teacher's behavior outside the ROI to avoid catastrophic forgetting, i.e., keep $f_{\theta'}(x)$ close to $f_\theta(x)$ on $(1 - m)$.

Our proposed metrics DCS and CCS (Sec. 3.2) are reference-free structural *measurement tools* aligned with deviation and confusion. They complement standard metrics and serve as relative probes to use with our benchmark.

4.2 Grounded Self-Distillation (GSD) Pipeline

The ROI second-order magnitude-based projection in Sec. 3.2 isolates two failure modes on illusion ROIs: (i) spurious curvature inside low-curvature carriers

(read out radially as DCS) and (ii) context-driven drift between full and crop views (off-diagonal shift as CCS). We leverage a strong pretrained depth model (Depth-Anything v2) and adapt it so that the network *learns to suppress illusory curvature* and *remains less sensitive to missing context*, while preserving its background/ordinal behavior. Concretely, the objective mirrors the axes of our evaluation: reduce curvature artifacts inside the ROI and stabilize full/crop predictions without sacrificing non-ROI structure.

We illustrate our whole system pipeline in Fig. 5. The 3D Mirage vulnerability largely stems from global context priors in the model's ViT encoder, which we tune directly. We use LoRA for this adaptation because it surgically modifies encoder behavior in a low-rank subspace, preserving the frozen backbone weights (our "teacher") and preventing catastrophic forgetting [1, 25].

Let T be the frozen teacher (*e.g.*, DAv2) and S the student obtained by inserting LoRA adapters into the teacher's *encoder*. Only LoRA parameters (and a small gating MLP) are trainable; the frozen teacher is used only for training supervision, so inference requires a single student forward pass with the LoRA adapters enabled. Training is dual-view with shared weights (Fig. 5): a crop branch receives x_{crop} and a full branch receives x_{full}. Denote student depths by s (crop) and s^F (full), teacher depths by t and t^F. Importantly, the re-editing objective does *not* force the illusion to stay on a single flat surface: we suppress hallucinated second-order structure while allowing piecewise-smooth geometry supported locally via a teacher-guided mixture of simple surface hypotheses and a learned gate (see Fig. 7).

4.3 Composite Loss Function

We optimize a weighted sum of terms per branch (crop/full) and then sum branches.

Notation. We use a masked mean operator $\overline{u}_m = \frac{\langle u,m\rangle}{\langle \mathbf{1},m\rangle}$, with $\langle\cdot,\cdot\rangle$ denoting element-wise inner product over pixels. We write the magnitude of our fixed separable second-difference operator as $\mathcal{L}(\cdot)$.

Normalization and Masks (Per Branch). Each branch is normalized to the teacher's background statistics (μ_B, σ_B) computed over a branch-specific background mask m_{bg} (Eq. 3), yielding normalized depths (z, z_T) for the crop view and (z^F, z_T^F) for the full view.

From the binary ROI mask m we form an ROI-adjacent ring r by dilation and subtraction, and a thin guard ring r_g by one additional dilation step. We then split r into a high-gradient edge subset r_e *(top 10% by $|\mathcal{L}(z_T)|$ within r)* and the complementary low-gradient seam $r_f = r \setminus r_e$. The background mask excludes the ROI and both rings:

$$m_{\text{bg}} = (1-m)(1-r)(1-r_g), \tag{3}$$

(and analogously m_{bg}^F for the full branch). On the seam r_f we also compute a locally smoothed teacher depth $\tilde{z}_T$ using ring-restricted local averaging (masked smoothing on r_f); we use the analogous $\tilde{z}_T^F$ for the full branch.

Let $\mathcal{R}$ denote the set of ROI instances (polygons) in a given view, with per-instance masks $\{m_j\}_{j\in\mathcal{R}}$ and union mask $m = \bigcup_{j\in\mathcal{R}} m_j$. Around each ROI instance j we fit up to K simple surface hypotheses $\{\pi_{j,k}\}_{k=1}^{K}$ to the teacher depth on a thin ROI-adjacent band (the ring), and record residual scales $\{\sigma_{j,k}\}$ (lower residual indicates a more plausible local surface explanation). For the student we define per-instance ROI deviations:

$$\ell_{j,k} = \overline{|\, z - \pi_{j,k} \,|}_{\,m_j}, \qquad k = 1, \ldots, K. \tag{4}$$

A compact gating network G maps ROI/ring statistics to logits over the K fitted hypotheses. We set $w_j = \mathrm{softmax}(G(\cdot))$ to get mixture weights $\{w_{j,k}\}_{k=1}^{K}$. Soft targets q_j are derived from $\{\sigma_{j,k}\}$ (lower residual $\Rightarrow$ higher target weight), and we add a cross-entropy regularizer $\mathrm{CE}(w_j, q_j)$ (with temperature/label-smoothing) and an anchor term $\min_k \ell_{j,k}$.

Hallucination Knowledge Re-editing (HKR) Loss. This term directly targets the radial axis (DCS) by collapsing second-order structure inside the ROI toward zero. We apply a curvature penalty over the union ROI mask m, and a teacher-guided mixture loss accumulated over ROI instances $\mathcal{R}$:

$$\mathcal{L}_{\mathrm{HKR}} = \alpha_1 \overline{|\mathcal{L}(s)|}_{m} + \alpha_2 \sum_{j\in\mathcal{R}} \sum_{k=1}^{K} w_{j,k} \ell_{j,k} \tag{5}$$

We apply the same HKR form to both crop and full views (with view-specific masks, teacher statistics, and fitted hypotheses), and downweight the full-view contribution in the total objective. For regularizer images with no illusion ROI annotation, $\mathcal{R} = \varnothing$ and the ROI-local HKR terms are inactive by construction.

Non-hallucination Knowledge Preservation (NKP) Loss. This `self-distillation` term preserves the teacher's geometry on stable *background* regions using m_{bg}. To stabilize the transition across the ROI boundary, preserve edge detail and suppress halo artifacts, we use a compact *ring* regularizer that (i) tethers the student's depth to a locally smoothed teacher on the low-gradient seam r_f, and (ii) matches second-order structure (via $\mathcal{L}(\cdot)$) on the high-gradient edge subset r_e and the protective guard ring r_g. The resulting loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{NKP}} = {} & \alpha_3 \overline{|z - z_T|}_{m_{\mathrm{bg}}} + \alpha_4 \overline{|\mathcal{L}(z) - \mathcal{L}(z_T)|}_{m_{\mathrm{bg}}} + \alpha_5 \overline{|z - \tilde{z}_T|}_{r_f} \\ & + \alpha_6 \overline{|\mathcal{L}(z) - \mathcal{L}(z_T)|}_{r_e} + \alpha_7 \overline{|\mathcal{L}(z) - \mathcal{L}(z_T)|}_{r_g} \end{aligned} \tag{6}$$

We apply the same NKP structure to the full branch.

Total Objective Loss and Regularization. Per branch, the objective is the weighted sum of $\mathcal{L}_{\mathrm{HKR}} + \mathcal{L}_{\mathrm{NKP}}$ plus gating regularizers ($\mathrm{CE}(w, q)$ and the anchor term). Crop and full branches are combined to bias against context drift while keeping the crop branch dominant: $\mathcal{L} = \lambda_C \mathcal{L}_{\mathrm{crop}} + \lambda_F \mathcal{L}_{\mathrm{full}}$. The full branch uses the same HKR/NKP structure.

To avoid degenerate over-smoothing, illusion batches are interleaved with non-illusion data during the optimization steps. We incorporate two real-image datasets during training: The Penn–Fudan dataset provides 170 urban street

images with 345 upright pedestrians, offering diverse occlusions and pedestrian scales [54]. The CamVid collection contributes 701 raw still frames of urban driving scenes widely used in autonomous-driving research [7]. For these regularizer images, training reduces to the preservation objective for the scenes. This regularization helps suppress over-flattening and edge drift without weakening training supervision, and targets safety-critical deployments of depth models.

5 Experiments

To validate our framework, we first establish the vulnerability of SOTA models on our **3D-Mirage** benchmark. We then demonstrate the effectiveness of our **Grounded Self-Distillation** method in taming these hallucinations.

5.1 Experimental Setup

Baselines We compare against a comprehensive suite of SOTA monocular depth foundation models using their official weights. This includes the **Depth Anything families** (DA-{S,B,L} [62] and DAv2-{S,B,L} [63], including the indoor (DAv2-I) and outdoor (DAv2-O) specialized variants) and **other foundation models** (DepthPro [6], Marigold [28], DepthFM [20], ZoeDepth [3], and MiDaS [48]).

Our primary baseline for adaptation is `Depth-Anything-V2-Large-hf` **(DAv2-L)** [63], which serves as the frozen **teacher model** (T) and the initial backbone for our **student model** (S).

Implementation Details We implement our method in PyTorch, using the PEFT library [24] for LoRA adaptation.

1) Data. We use a custom sampler with a 4:1 ratio of 3D-Mirage (positive) samples to regularizer (negative) samples. Negative samples are drawn from Penn-Fudan [54] and CamVid [7] to prevent catastrophic forgetting/flattening on standard street scenes. We apply 50% horizontal flip and 5% photometric jitter augmentations. We split the benchmark at the *source-image* level to avoid leakage between a full image and its context-restricted crops. All crops derived from the same

Table 1: Quantitative comparison on the 3D-Mirage benchmark. We evaluate SOTA foundation models and our method (Grounded Self-Distillation) using our proposed metrics. **Lower is better.** Our method (Ours) drastically reduces both geometric deviation (DCS) and contextual instability (CCS) compared to all baselines, including its own teacher model (DAv2-L). Δ denotes the relative improvement of our model over the DAv2-L baseline.

Model	$d_{\text{cluster}}\downarrow$	$d_{\text{avg}}\downarrow$	DCS↓	$D_{\text{cluster}}\downarrow$	$D_{\text{avg}}\downarrow$	CCS↓
DepthPro [6]	317.8	331.4	649.1	6.680e-4	9.290e-4	1.597e-3
Marigold [28]	701.1	726.2	1.427e3	2.294e-3	2.402e-3	4.696e-3
DepthFM [20]	1.020e3	1.063e3	2.083e3	4.914e-3	5.215e-3	1.013e-2
ZoeDepth [3]	291.5	297.8	589.3	7.486e-4	7.560e-4	1.505e-3
MiDaS [48]	330.2	340.0	670.2	4.120e-4	5.090e-4	9.220e-4
DA-S [62]	225.9	233.9	459.8	3.190e-4	3.570e-4	6.760e-4
DA-B [62]	236.0	246.7	482.7	2.710e-4	3.520e-4	6.230e-4
DA-L [62]	243.3	251.7	495.0	2.730e-4	3.290e-4	6.030e-4
DAv2-IS [63]	415.6	424.5	840.1	1.452e-3	1.473e-3	2.924e-3
DAv2-IB [63]	347.5	359.5	706.9	1.133e-3	1.183e-3	2.315e-3
DAv2-IL [63]	406.1	418.4	824.5	1.161e-3	1.187e-3	2.348e-3
DAv2-OS [63]	685.4	698.4	1.384e3	3.101e-3	3.102e-3	6.203e-3
DAv2-OB [63]	713.4	726.1	1.439e3	2.901e-3	2.902e-3	5.804e-3
DAv2-OL [63]	537.0	547.5	1.085e3	1.959e-3	1.961e-3	3.920e-3
DAv2-S [63]	495.9	511.2	1.007e3	7.210e-4	7.900e-4	1.512e-3
DAv2-B [63]	431.7	449.3	881.0	6.320e-4	7.270e-4	1.359e-3
DAv2-L [63] (**Baseline**)	488.8	505.8	994.6	6.840e-4	7.820e-4	1.466e-3
Ours	**28.55**	**30.09**	**58.64**	**9.174e-5**	**9.894e-5**	**1.907e-4**
Δ (%)	(**-94.16**%)	(**-94.05**%)	(**-94.10**%)	(**-86.59**%)	(**-87.35**%)	(**-86.99**%)

original image are assigned to the same split. We use a 90/10 train/test split, yielding 421/47 scenes and 1684/188 paired full-crop instances.
2) Model. We inject LoRA adapters (rank $r = 16$, $\alpha = 32$, dropout 0.05, 'bias=none') into the DINOv2 encoder's patch embedding layer and all MLP linear layers ('fc1', 'fc2') within the 24 transformer blocks. This results in only **4M trainable parameters** (≈1.2% of the DAv2-L backbone).
3) Training. We use the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 0.01, and global gradient clipping of 1.0. For training stability, all student and teacher depth outputs are z-normalized over background pixels before loss computation. The model is trained for only **1 epoch** with a batch size of 8 on an NVIDIA A100 GPU. Extending training to 4 epochs yields only marginal gains on DCS/CCS, while the other evaluation metrics show mixed behavior. Detailed results are provided in the supplementary.

For the losses, we fix the loss weights to α_1=1.0, α_2=0.4, α_3=1.0, α_4=0.5, α_5=0.3, α_6=0.8, and α_7=0.3 to keep the different losses numerically comparable.

Evaluation We evaluate models on two fronts. First, we test for **hallucination robustness** using our **3D-Mirage** benchmark with the proposed **DCS** (hallucination intensity) and **CCS** (contextual instability) metrics, where lower is better. Second, we test for **knowledge preservation** using an ordinal pairwise accuracy protocol on **NYU-v2** [50] (as detailed in Sec. 5.3) to ensure our method does not catastrophically forget general depth estimation.

5.2 Main Results: Taming 3D Mirages

Table 1 presents the quantitative results on our 3D-Mirage benchmark. The results are decisive: **Across all evaluated SOTA foundation models, we observe consistently elevated DCS/CCS on 3D-Mirage, suggesting these models are highly vulnerable to 3D mirages.** The failure is systemic, afflicting all tested architectures (transformer, diffusion-based, etc.). This suggests that their training condition has inadvertently created powerful, dataset-level priors (e.g., complex 2D patterns often imply high second-order 3D structure) that override local geometric evidence when faced with ambiguous, out-of-distribution perceptual traps. Our baseline, DAv2-L, scores a high 994.6 on DCS, confirming it perceives significant, spurious 3D geometry.

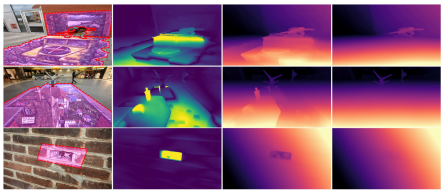

Fig. 6: Qualitative results of our Grounded Self-Distillation. Each row compares our model to the baseline on a 3D-Mirage sample. (1) Input RGB. (2) Error heatmap (Ours vs. Baseline), showing changes are confined to the ROI. (3) Baseline (DAv2-L) depth, which hallucinates non-existent 3D structures. (4) Our model's depth, which correctly perceives the planar surface. Our method tames the 3D mirage without distorting the background, including illusions existing on separate planes (first row).

In contrast, our GSD method achieves a DCS of only **58.64** and a CCS of **1.907e-4**, the best scores by a large margin. This represents a massive **94% reduction in geometric deviation (DCS)** and an **87% reduction in contextual instability (CCS)** compared to the DAv2-L teacher. This demonstrates not only that the hallucination is removed, but that the model is much less confused by the removal of context. It has learned to ground its prediction in local geometric evidence (low-curvature ROI) rather than being swayed by fragile semantic priors.

This illusion taming is visualized in Fig. 6, 7. Our model (column 4) successfully identifies and resolve hallucination, including cases on multiple surfaces, and, in some cases, curved surfaces. The baseline (column 3) dangerously hallucinates large obstacles, caverns, and non-existent 3D structure. Crucially, the difference heatmap (column 2) confirms that our model's corrections are surgically confined to the illusion ROI. This provides strong evidence that our $\mathcal{L}_{\text{NKP}}$ (knowledge preservation) loss is working as intended, preventing the flattening objective from leaking and destroying valid geometry in the background.

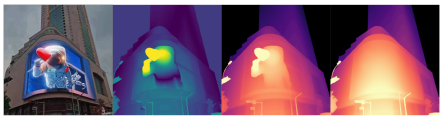

Fig. 7: Impact of plane-mixture and gating. Our HKR loss does not force a single fronto-parallel plane. Our model learns to approximate curvature and can partially recover curved surface from hallucination.

Generalization to 3D visual illusions (3DVI). We additionally evaluate our model zero-shot on the real-world test split of the 3D-Visual-Illusion [64, 65] (3DVI) dataset from Yao et al. (NeurIPS'25), following their evaluation protocol (Table 2), which reports disparity-space metrics (EPE, bad-t) and depth-space metrics (AbsRel, RMSE, δ_1) over the provided illusion masks under a single global scale-and-shift alignment.

Table 2: Results on illusion regions of the 3D-Visual-Illusion (3DVI) test set. "align" denotes globally shared affine (scale+shift) alignment computed from ground truth, following Yao et al. (NeurIPS'25).

Method	FT	Disparity Space				Depth Space		
		EPE↓	bad2↓	bad3↓	bad5↓	AbsRel↓	RMSE↓	δ_1 ↑
Stereo or multi-view input models (not our focus)								
Dust3R [55]	×	6.74	52.89	45.31	36.61	0.25	0.22	87.09
VGGT [53]	×	6.16	53.32	44.89	37.20	0.13	0.12	78.46
RAFT-Stereo [38]	×	1.62	24.32	13.20	2.97	0.04	0.06	99.18
Selective-RAFT [57]	×	1.58	23.46	12.65	2.57	0.03	0.07	99.60
Selective-IGEV [59]	×	1.67	24.06	13.11	2.99	0.04	0.10	99.26
MochaStereo [12]	×	1.75	25.49	14.11	3.54	0.04	0.11	98.76
StereoAnything [22]	×	2.41	29.00	16.15	6.54	0.11	0.32	96.23
3DVI [64]	✓	**1.77**	**26.72**	**15.73**	**3.60**	**0.03**	**0.08**	**99.60**
Monocular or single-view input models (our focus)								
DAv2 [63]	×	5.81	61.45	43.18	30.57	0.14	0.15	92.86
Metric3D [66]	×	12.46	94.11	91.14	82.05	0.34	0.29	48.97
DAv2 metric [63]	×	16.24	92.53	87.43	75.25	0.52	0.39	48.75
DepthPro [5]	×	12.26	87.08	80.60	62.43	0.28	0.25	65.92
Marigold [28]	×	21.16	65.67	59.67	53.19	0.45	0.37	63.65
DAv2 metric(align)	×	5.23	56.82	45.50	28.89	0.17	0.15	93.70
Metric3D(align)	×	5.70	66.26	50.92	40.43	0.17	0.17	94.80
DepthPro(align)	×	4.36	44.98	34.98	24.70	0.09	0.10	93.83
Ours	×	**1.75**	**26.67**	**15.52**	**6.59**	**0.03**	**0.06**	**99.50**

Cross-architecture generalization. The GSD objective is not tied to the DAv2 encoder. Applying the same recipe to a second transformer backbone (ZoeDepth) and to a diffusion backbone (Marigold) yields consistent reductions on both 3D-Mirage and 3DVI: on ZoeDepth (Table 5.2), DCS drops 589 → 335; on Marigold (Fig. 8), a LoRA

(r=16) adaptation cuts DCS and CCS by **44%** and **40%** while preserving DA-2K/DIW depth, with full results in the supplement.

Table 3: Transfer to ZoeDepth (a second transformer backbone): 3D-Mirage (DCS/CCS) and 3DVI metrics. Lower is better except δ_1.

Model	DCS↓	CCS↓	EPE↓	bad2↓	AbsRel↓	RMSE↓	δ_1 ↑
ZoeDepth	589.27	1.505e-3	9.555	67.66	0.1899	0.2255	76.16
Our ZD	**335.12**	**1.136e-3**	**7.286**	**61.70**	**0.1273**	**0.1400**	**84.76**

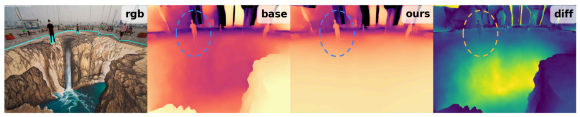

Fig. 8: Qualitative depth on the Marigold diffusion backbone (from left: RGB, baseline, ours, difference). GSD removes the hallucinated 3D structure on the illusion surface (dashed) while preserving fine detail such as the standing person.

5.3 Ablation Study

We ablate both terms of our objective: *Hallucination Knowledge Re-editing* ($\mathcal{L}_{\text{HKR}}$) suppresses the mirage, and *Non-hallucination Knowledge Preservation* ($\mathcal{L}_{\text{NKP}}$) prevents over-flattening and preserves general depth. We compare the full model against two loss-removal variants on 3D-Mirage (DCS/CCS), NYUv2, KITTI 15, and DA-2K [63] (Table 4). Removing $\mathcal{L}_{\text{HKR}}$ preserves standard depth but leaves the 3D mirage largely intact; removing $\mathcal{L}_{\text{NKP}}$ reaches the lowest DCS/CCS but degrades the standard metrics, indicating over re-editing. Our full model gives the best overall trade-off; the qualitative effect of each term is shown in Fig. 9.

Table 4: Component ablation.

Benchmark: Metric	No Halluc. Re-editing	No Knowl. Preserv.	Ours
DCS↓	988.60	42.83	58.64
CCS↓ ($\times 10^{-3}$)	1.470	0.141	0.191
NYUv2 AbsRel↓	0.1623	0.1533	0.1597
NYUv2 δ_1 (%)↑	79.15	80.25	79.58
KITTI15 [40] AbsRel↓	0.3409	0.3510	0.3424
KITTI15 [40] δ_1 (%)↑	39.33	38.27	39.21
DA-2K: Rel. Acc. (%)↑	97.05	94.00	96.08

We further compare GSD against simpler adaptation strategies (Table 5): a direct ROI *Simple L1*, naive encoder finetuning (*Ftune Enc.*), and decoder finetuning (*Ftune Dec.*). The Simple L1 variant trains the LoRA student with only an ROI L1 objective pulling the prediction toward a plane fitted from the frozen teacher's ROI-adjacent ring. Our full pipeline gives the best balance of illusion robustness and general-depth performance, while encoder finetuning degrades generalization most and Simple L1 yields smaller gains.

6 Conclusion

We identified, diagnosed, and mitigated a critical vulnerability in SOTA monocular depth models: the **3D Mirage**, a systemic failure where models hallucinate

Table 5: Ablation: our full GSD pipeline vs simpler adaptation strategies. Lower is better except DA-2K, δ_1, and BG R^2.

Model	DIW (%)	DA-2K (%)	NYUv2	KITTI15	3DVI				3D-Mirage		
	WHDR↓	pairwise↑	RMSE ↓	RMSE ↓	EPE ↓	bad2 ↓	RMSE ↓	δ_1 (%) ↑	BG R^2 ↑	DCS ↓	CCS ↓
Ours	11.48	96.08	0.5239	6.85	1.750	26.67	0.06383	99.50	0.9395	58.64	1.907e-4
Simple L1	12.03	93.04	0.5299	7.10	2.055	32.80	0.06970	98.77	0.8965	59.68	2.118e-4
Ftune Enc.	32.15	58.51	0.9014	8.54	7.629	69.93	0.1787	81.64	0.6296	27.85	9.389e-5
Ftune Dec.	14.46	87.48	0.6487	6.93	4.430	50.93	0.1137	93.15	0.8992	70.26	1.986e-4

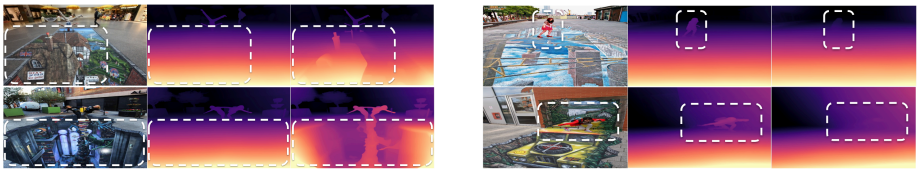

Fig. 9: Qualitative ablation of the two loss terms (each panel: input RGB, our full model, ablated variant). *Left* (w/o $\mathcal{L}_{\text{HKR}}$): the background is preserved but the 3D mirage on the road is left intact. *Right* (w/o $\mathcal{L}_{\text{NKP}}$): the mirage is removed but the flattening leaks into the background and distorts real objects. This mirrors the quantitative trade-off in Table 4.

spurious 3D structure from ambiguous texture, posing a significant risk to safety-critical applications. To our knowledge, this is the first end-to-end framework to **probe** the failure with our GT-free **3D-Mirage** benchmark, **score** it with novel second-order magnitude-based metrics, **DCS** (structural deviation) and **CCS** (contextual stability), and **tame** it with a parameter-efficient **Grounded Self-Distillation** strategy. Guided by a composite $\mathcal{L}_{\text{HKR}}$ (re-editing) and $\mathcal{L}_{\text{NKP}}$ (preservation) loss, our method reduces hallucinations by **over 94%** and instability by **87%**, and our ablations confirm the adaptation is surgically precise, avoiding the catastrophic forgetting of naive finetuning. This work provides the essential tools to advance MDE from simple pixel accuracy toward the structural and contextual robustness required for real-world deployment.

Limitations and Future Work. Our **3D-Mirage** benchmark is primarily focused on low-curvature surfaces perceived as 3D structure, which does not encompass the full spectrum of perceptual ambiguity, such as texture-less surfaces, reflections, glass, shadows, or adverse weather. Our LoRA-based mitigation was demonstrated on a transformer-based MDE architecture; we further validate its transfer to a second transformer backbone and a diffusion model (Marigold) in the supplement. We do not yet explicitly evaluate on real road hazards such as potholes, curbs, or speed bumps, which would be a valuable safety-oriented complement to our current preservation benchmarks.

Broader Impact. Monocular depth models are increasingly deployed in safety-critical perception, where a hallucinated bump on a flat road or wall can trigger unnecessary braking or unsafe maneuvers. Our contributions target this risk end-to-end: the **3D-Mirage** benchmark exposes the failure, the reference-free **DCS**/**CCS** metrics let practitioners audit for spurious curvature without ground-truth depth (rarely available at deployment scale), and **Grounded Self-Distillation** mitigates it with an adapter that folds into the backbone at inference with no added latency or memory. Removing hallucinated structure must not be mistaken for a guarantee of metric depth accuracy: DCS and CCS certify structural and contextual stability on the low-curvature carrier, not absolute scale, and should complement rather than replace task-level validation before deployment.

References

1. Aghajanyan, A., Gupta, S., Zettlemoyer, L.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 7319–7328. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.568, https://aclanthology.org/2021.acl-long.568/
2. Bartolomei, L., Tosi, F., Poggi, M., Mattoccia, S.: Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In: CVPR. pp. 1013–1027 (June 2025)
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023), https://arxiv.org/abs/2302.12288
4. Bian, J.W., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
5. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint **2410.02073** (2024), https://arxiv.org/abs/2410.02073
6. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025)
7. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters (2009), http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/, accessed: November 7, 2025
8. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? arguments against avoiding RMSE in the literature. Geoscientific Model Development **7**(3), 1247–1250 (2014). https://doi.org/10.5194/gmd-7-1247-2014, https://gmd.copernicus.org/articles/7/1247/2014/
9. Chawla, J., Thakurdesai, N., Godase, A., Reza, M.A., Crandall, D.J., Jung, S.H.: Error diagnosis of deep monocular depth estimation models. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8400–8407. IEEE (2021). https://doi.org/10.1109/IROS51168.2021.9636673
10. Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1108–1118 (2023)
11. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems (NeurIPS) (2016)
12. Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J.: Mocha-stereo: Motif channel attention network for stereo matching. In: CVPR. pp. 27768–27777 (June 2024)
13. Costanzino, A., Zama Ramirez, P., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Learning depth estimation for transparent and mirror surfaces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9244–9255. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00386

14. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2432–2443 (2017)
15. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
16. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361 (2012)
17. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
18. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left–right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6602–6611 (2017)
19. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3827–3837 (2019)
20. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast generative monocular depth estimation with flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
21. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2485–2494 (2020)
22. Guo, X., Zhang, C., Zhang, Y., Wang, R., Nie, D., Zheng, W., Poggi, M., Zhao, H., Ye, M., Zou, Q., Chen, L.: Stereo anything: Unifying zero-shot stereo matching with large-scale mixed data. arXiv preprint arXiv:2411.14053 (2024)
23. Heo, M., Lee, J., Kim, K.R., Kim, H.U., Kim, C.S.: Monocular depth estimation using whole strip masking and reliability-based refinement. In: European Conference on Computer Vision (ECCV). pp. 39–55 (2018)
24. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2790–2799. PMLR (2019), https://proceedings.mlr.press/v97/houlsby19a.html
25. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR). OpenReview.net (2022), https://openreview.net/forum?id=nZeVKeeFYf9
26. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
27. Kayhan, O.S., Vredebregt, B., van Gemert, J.C.: Hallucination in object detection: A study in visual part verification. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2234–2238. IEEE (2021). https://doi.org/10.1109/ICIP42928.2021.9506670

28. Ke, B., Obukhov, A., Huang, S., Metzger, N., Caye Daudt, R., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9492–9502 (June 2024)
29. Kim, S., Park, C., Jeon, G., Kim, S., Kim, J.H.: Automated audit and self-correction algorithm for seg-hallucination using meshcnn-based on-demand generative ai. Bioengineering (Basel) **12**(1), 81 (2025). https://doi.org/10.3390/bioengineering12010081
30. Lee, S., Park, S.H., Jo, Y., Seo, M.: Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 391–404. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.naacl-long.23, https://aclanthology.org/2024.naacl-long.23/
31. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
32. Li, X., Qiu, K., Wang, J., Xu, X., Singh, R., Yamazaki, K., Chen, H., Huang, X., Raj, B.: R 2-bench: Benchmarking the robustness of referring perception models under perturbations. In: European Conference on Computer Vision. pp. 211–230. Springer (2024)
33. Li, X., Wang, J., Xu, X., Li, X., Raj, B., Lu, Y.: Robust referring video object segmentation with cyclic structural consensus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22236–22245 (October 2023)
34. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3402–3413 (June 2024)
35. Li, X., Wang, J., Xu, X., Raj, B., Lu, Y.: Online video instance segmentation via robust context fusion. arXiv preprint arXiv:2207.05580 (2022)
36. Li, X., Wang, J., Xu, X., Yang, M., Yang, F., Zhao, Y., Singh, R., Raj, B.: Towards noise-tolerant speech-referring video object segmentation: Bridging speech and text. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2283–2296. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.140, https://aclanthology.org/2023.emnlp-main.140/
37. Li, Y., Kong, L., Hu, H., Xu, X., Huang, X.: Is your LiDAR placement optimized for 3D scene understanding? In: Advances in Neural Information Processing Systems. vol. 37, pp. 34980–35017 (2024)
38. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: Proceedings of the International Conference on 3D Vision (3DV). pp. 218–227 (2021). https://doi.org/10.1109/3DV53792.2021.00032
39. Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., Fung, P.: Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. In: Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR). pp. 37–58. Association for Computational Linguistics, Bangkok, Thailand (aug 2024). https://doi.org/10.18653/v1/2024.alvr-1.4, https://aclanthology.org/2024.alvr-1.4/

40. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3061–3070 (2015)
41. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 86–94 (2017)
42. Muhovič, J., Koporec, G., Perš, J.: Hallucinating hidden obstacles for unmanned surface vehicles using a compositional model. In: Proceedings of the 26th Computer Vision Winter Workshop (CVWW). University of Ljubljana (2023), http://prints.vicos.si/publications/436, accessed: November 7, 2025
43. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 427–436. IEEE (2015). https://doi.org/10.1109/CVPR.2015.7298640
44. Qiu, K., Li, X., Chen, H., Kuen, J., Xu, X., Gu, J., Luo, Y., Raj, B., Lin, Z., Savvides, M.: Image tokenizer needs post-training. arXiv preprint arXiv:2509.12474 (2025)
45. Qiu, K., Li, X., Kuen, J., Chen, H., Xu, X., Gu, J., Luo, Y., Raj, B., Lin, Z., Savvides, M.: Robust latent matters: Boosting image generation with sampling error synthesis. arXiv preprint arXiv:2503.08354 (2025)
46. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Stefano, L.D.: Booster: A benchmark for depth from images of specular and transparent surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 85–102 (2024). https://doi.org/10.1109/TPAMI.2023.3323858
47. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (2021), https://openaccess.thecvf.com/content/ICCV2021/html/Ranftl_Vision_Transformers_for_Dense_Prediction_ICCV_2021_paper.html
48. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1623–1637 (2022)
49. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems (NIPS) (2005)
50. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 746–760 (2012)
51. Singh, K.K., Mahajan, D., Grauman, K., Lee, Y.J., Feiszli, M., Ghadiyaram, D.: Don't judge an object by its context: Learning to overcome contextual bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11070–11078 (2020)
52. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1521–1528 (2011)
53. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (June 2025)
54. Wang, L., Shi, J., Song, G., Shen, I.F.: Penn-fudan database for pedestrian detection and segmentation (2007), https://www.cis.upenn.edu/~jshi/ped_html/, accessed: November 7, 2025

55. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (June 2024)
56. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual SLAM. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916 (2020)
57. Wang, X., Xu, G., Jia, H., Yang, X.: Selective-stereo: Adaptive frequency information selection for stereo matching. In: CVPR. pp. 19701–19710 (June 2024)
58. Wong, A., Cicek, S., Soatto, S.: Targeted adversarial perturbations for monocular depth prediction. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
59. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: CVPR. pp. 21919–21928 (June 2023)
60. Xu, X., Xue, F., Li, X., Li, H., Yang, S., Zhang, T., Johnson-Roberson, M., Huang, X.: Towards ambiguity-free spatial foundation model: Rethinking and decoupling depth ambiguity. arXiv preprint arXiv:2503.06014 (2025)
61. Xu, X., Zhang, T., Zhao, S., Li, X., Wang, S., Chen, Y., Li, Y., Raj, B., Johnson-Roberson, M., Scherer, S., Huang, X.: Scalable benchmarking and robust learning for noise-free ego-motion and 3d reconstruction from noisy video. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=Pz9zFea4MQ
62. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2024)
63. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Advances in Neural Information Processing Systems (NeurIPS) (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/26cfdcd8fe6fd75cc53e92963a656c58-Paper-Conference.pdf
64. Yao, C., Liu, Z., Zeng, J., Yu, L., Wu, Y., Jia, Y.: 3d visual illusion depth estimation. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), https://arxiv.org/abs/2505.13061
65. Yao, C., et al.: 3d-visual-illusion-depth-estimation (official code and dataset release). https://github.com/YaoChengTang/3D-Visual-Illusion-Depth-Estimation (2025), accessed: 2026-02-21
66. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: ICCV. pp. 9043–9053 (October 2023)
67. Zheng, J., Lin, C., Sun, J., Zhao, Z., Li, Q., Shen, C.: 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)
68. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1851–1858 (2017)