OpenDepth: Repurposing Image Editing Priors for Robust Open-World Monocular Depth Estimation

Mengke Zhang
zmk1460997475@bupt.edu.cn
International School, Beijing University of Posts and Telecommunications
Beijing, China

Tianyi Yang
3662748738@bupt.edu.cn
International School, Beijing University of Posts and Telecommunications
Beijing, China

Xinzhuo Li
xinzhuo.li@se23.qmul.ac.uk
International School, Queen Mary University of London
Beijing, China

Anlong Ming*
mal@bupt.edu.cn
School of Computer Science, Beijing University of Posts and Telecommunications
Beijing, China

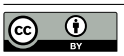
Abstract

Recently, the integration of generative priors has revolutionized Monocular Depth Estimation (MDE), achieving state-of-the-art (SOTA) performance with remarkably high data efficiency. However, although methods based on Text-to-Image (T2I) models excel in general scenes, their efficacy degrades sharply when confronted with open-world environmental perturbations, such as illumination variations and adverse weather. Since T2I priors are optimized for semantic synthesis rather than preserving pixel-level structure, they struggle to decouple geometry from appearance under severe visual shifts, easily triggering structural hallucinations. Therefore, our core insight is to model open-world depth estimation as consistent geometric inference across multi-style images, fully leveraging the powerful structure-preserving editing priors of Image-to-Image (I2I) models. Based on this, we propose OpenDepth, a robust open-world MDE framework that introduces "Intra-Domain Awareness"—the capability to map diverse stylistic inputs sharing identical geometry into a unified latent space. To achieve this, we first align shallow features within the Diffusion Transformer (DiT) to anchor degraded observations to unified geometric representations. Additionally, we introduce a joint RGB latent reconstruction task, requiring the model to recover clean RGB information from degraded inputs. This multi-stage strategy effectively decouples environmental style interference, compelling the model to focus strictly on underlying geometric recovery. Extensive evaluations show that OpenDepth sets new SOTA records for robust zero-shot depth estimation across standard and extreme weather benchmarks, demonstrating exceptional structural integrity and adaptability.

CCS Concepts

• **Computing methodologies** → **Computer vision**; *Machine learning*; Artificial intelligence.

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836413

Keywords

Monocular Depth Estimation, Open-World Robustness, Image Editing Priors, Latent Feature Alignment, Diffusion Transformer

ACM Reference Format:
Mengke Zhang, Tianyi Yang, Xinzhuo Li, and Anlong Ming. 2026. OpenDepth: Repurposing Image Editing Priors for Robust Open-World Monocular Depth Estimation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836413

1 Introduction

Monocular Depth Estimation [1, 13, 35, 69] serves as a cornerstone for a wide range of downstream applications, including autonomous driving [61], robotic navigation [52], and augmented reality [11]. Despite its undeniable importance, achieving high-fidelity depth maps remains a formidable challenge due to the task's inherent ill-posed nature and the scarcity of large-scale, precisely labeled datasets. Recently, the field has witnessed a paradigm shift with the integration of generative priors. Pre-trained Text-to-Image diffusion models, such as Stable Diffusion [43], have demonstrated a remarkable capacity to capture complex scene statistics. By treating depth as a specialized image generation modality, these models have achieved SOTA performance in zero-shot geometric estimation.

However, existing models struggle with complex environmental perturbations (e.g., adverse weather, motion blur) in open-world tasks. Directly fine-tuning them on degraded samples remains suboptimal due to the scarcity of comprehensively annotated real-world data. While recent methods leverage generative priors to achieve SOTA performance with limited data, relying on T2I models introduces new bottlenecks. Optimized primarily for semantic synthesis, T2I priors frequently produce severe structural hallucinations under visual degradations. Fundamentally, monocular depth estimation is an Image-to-Image translation task, making T2I priors intrinsically misaligned with its strict geometric objectives. To isolate this, we conduct a matched-data comparison (Figure 1) training both a T2I model (Lotus-D [21]) and our I2I baseline on the identical 59k samples. Under varying perturbations, the I2I baseline successfully preserves scene geometry, contrasting sharply with the T2I model's structural failures. This highlights the superior

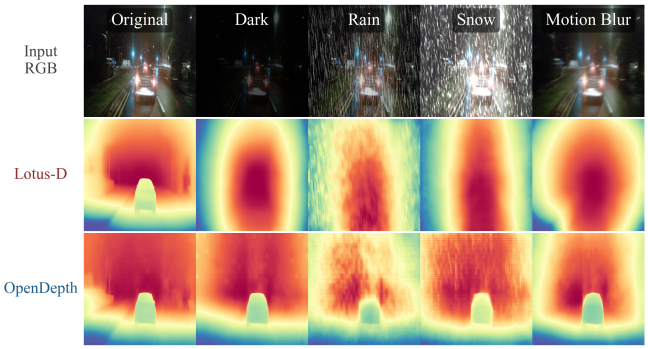


Figure 1: Standalone comparison of T2I and I2I priors under perturbations (zero-shot). In a separate controlled experiment, both models are trained on the same 59K samples from Hypersim [42] and Virtual KITTI [6]. The T2I model (Lotus-D [21]) suffers from severe structural hallucinations under visual degradations, while the I2I baseline better preserves scene geometry, suggesting stronger inherent robustness for depth estimation.

inherent robustness of I2I priors for geometric tasks, motivating our I2I-based foundation.

To overcome this fundamental limitation, our core insight is to reframe open-world depth estimation as a unified geometric inference process across diverse visual styles. In this context, Image-to-Image editing models naturally align with the task, as they excel at accommodating various stylistic transformations while strictly preserving the underlying structural integrity. Nevertheless, the mere injection of editing priors is insufficient to fully resolve the challenge. Under extreme environmental degradations, severe input distribution shifts still trigger alignment deviations within the feature space, hindering simple, direct depth regression from extracting pure geometric cues. Consequently, we argue that a robust depth estimator must possess "Intra-Domain Awareness"—the capability to actively map and anchor observations from diverse degradation conditions into a unified, clean latent domain prior to executing core geometric reasoning. This mechanism is indispensable for truly unleashing the geometric perception potential of editing priors in complex open-world environments.

Driven by this philosophy, we introduce **OpenDepth**, a robust open-world MDE framework. Moving beyond existing approaches that focus exclusively on regressing depth from degraded features, our method prioritizes global alignment within the feature space. Specifically, we implement a multi-stage alignment and reconstruction mechanism: First, we align image features from the early Single Stream blocks to establish robust correspondence between perturbed observations and their unified geometric structures. Subsequently, by leveraging the inherent bipartite structure of the DiT output sequence, we introduce unified RGB latent reconstruction as a joint task alongside depth estimation. This joint optimization compels the model to reconstruct unperturbed RGB information from diverse conditional inputs, thereby ensuring the latent features conform to the unified latent domain established by the editing prior. This strategy effectively enables the model to decouple environmental style interference and focus exclusively on recovering high-fidelity geometry.

The contributions of this work are threefold: • **Systematic Analysis of Generative Priors:** We provide a comprehensive analysis of the limitations of T2I-based generative priors in MDE, revealing that their stochastic nature and lack of structural constraints lead to performance degradation in complex environments. We demonstrate the superiority of shifting toward I2I editing models as the foundation for robust geometric tasks. • **Unified Latent Alignment Framework:** We propose an innovative alignment mechanism that transcends direct regression by operating in a constrained latent space. By integrating unified RGB reconstruction and initial DiT block alignment, we ensure that diverse visual observations are projected into a unified latent domain, establishing a robust "many-to-one" mapping from perturbed inputs to ground-truth depth. • **SOTA Performance & High Data Efficiency:** OpenDepth sets new SOTA records for robust zero-shot depth estimation across standard and degraded benchmarks. Notably, it achieves exceptional detail preservation in complex open-world scenes using only 79K training samples, demonstrating remarkable data efficiency.

2 Related Work

2.1 Foundation MDE Model

MDE has rapidly evolved from early supervised and self-supervised methods [13, 19, 35, 85] into powerful foundation models with formidable zero-shot generalization. Discriminative approaches established robust multi-source training paradigms [12, 39, 40, 73, 75], while subsequent works improved metric accuracy and structural fidelity via transformer decoding, internal discretization, and scale decomposition [1–3, 23, 29, 33, 37, 38, 60, 74, 76, 86]. The Depth Anything series [69, 70] pushed performance boundaries via large-scale semi-supervised learning, while its latest iteration [30] further extended this paradigm to unified multi-view and monocular geometry recovery with a plain transformer architecture. Lightweight discriminative frameworks also proved that efficiency and accuracy can coexist [41]. Concurrently, generative models repurposed T2I diffusion priors for dense prediction [14, 27, 82], with subsequent works optimizing inference via Flow Matching and deterministic single-step paradigms [20, 21, 48, 63], improving conditioning and detail refinement [15, 32, 36, 71, 79], unifying generation and dense perception [83], extending to video sequences [24], and adapting diffusion priors for self-supervised learning [56]. However, as argued by FE2E [58], T2I models exhibit inherent stochasticity and lack pixel-level structural constraints, motivating a shift toward image editing models [4, 31, 62] whose intrinsic "editing priors" naturally maintain structural integrity. This has given rise to a new class of I2I-based dense geometry estimators, represented by FE2E [58] and Edit2Perceive [46], which demonstrate that repurposing image editing models yields superior structural fidelity over their T2I counterparts for depth and normal estimation.

Compared to existing I2I schemes (e.g., FE2E [58], Edit2Perceive [46]), OpenDepth addresses a critical gap: while existing methods leverage editing models to regress depth directly, severe input shifts in extreme environments still cause feature space misalignments. Rather than simply migrating editing features, OpenDepth

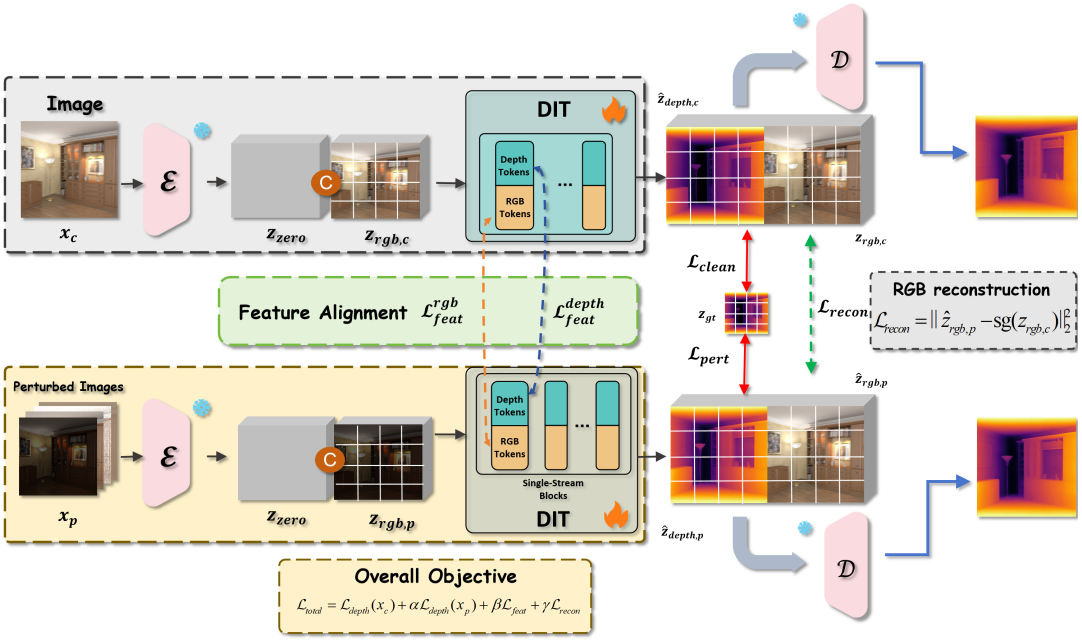


Figure 2: Overview of the OpenDepth framework. To harness editing priors for robust open-world depth estimation, we introduce Latent Feature Alignment ($\mathcal{L}_{feat}$) and Unified RGB Reconstruction ($\mathcal{L}_{recon}$). This dual-constraint strategy projects diverse degraded inputs into a unified latent domain, strictly decoupling environmental styles from geometric structures.

introduces an "Intra-Domain Awareness" alignment strategy. By incorporating a unified RGB reconstruction task and shallow latent space alignment at the earliest feature extraction stages, we compel the model to project diverse degraded observations into a unified latent domain defined by the editing prior, enabling superior consistency and noise resistance.

2.2 Robust Monocular Depth Estimation

Achieving robustness under open-world perturbations is critical, as systematic benchmarks [22, 26, 28] reveal severe degradation in SOTA models under adverse weather and illumination. Early domain adaptation bridged the synthetic-to-real gap via adversarial alignment or image translation [16, 18, 53, 59]. Subsequently, self-supervised methods tackled photometric consistency violations through joint nighttime enhancement [84], frequency-aware regularization [9], masked image modeling [8], curriculum contrastive learning [44, 57], depth consistency regularization [49], dynamic video training [51], IMU-guided learning [78], dual-branch all-day estimation [81], distribution compensation [68], foundation-model-assisted stereo [54], stabilized adversarial generalization [72], and motion-structure priors [64]. In the foundation model era, parameter-efficient adaptation has emerged as the dominant strategy for weather robustness via knowledge distillation [77] and LoRA variants [65, 67]. Generative approaches also explore clean-image synthesis [25] or joint dehazing [80]. Despite this progress, current state-of-the-art open-world models, such as DepthAnything-AC [50], primarily rely on unsupervised output-consistency and spatial distance constraints.

Diverging from current SOTA methods like DepthAnything-AC [50], which primarily rely on discriminative architectures and massive data-level augmentations, OpenDepth pioneers the integration of generative Image-to-Image (I2I) editing priors for robust open-world perception. By leveraging the inherent structural preservation capabilities of these foundation models, our approach achieves superior robustness with remarkably high data efficiency, utilizing merely 79K training samples. Mechanistically, rather than enforcing complex output-level consistency, we propose a concise "Latent Alignment" pathway. By projecting diverse degraded inputs into a unified, style-invariant latent domain, this mechanism fundamentally eliminates environmental interference at the feature source, ensuring exceptional structural integrity in complex scenarios.

3 Methodology

In this section, we introduce **OpenDepth**, a robust Monocular Depth Estimation framework built upon the STEP1X-Edit foundation model (overview shown in Figure 2). We first review the Multi-Modal Diffusion Transformer (MMDiT) preliminaries and analyze the geometric advantages of I2I editing priors. Finally, we detail our core contributions—Latent Feature Alignment and Unified RGB Reconstruction—which construct a unified latent space for robust geometric perception.

3.1 Preliminaries: Multi-Modal Diffusion Transformer

Unlike conventional U-Nets, the MMDiT backbone in OpenDepth processes serialized patch embeddings to handle RGB conditions, depth targets, and text prompts.

Specifically, the input RGB image x and a target depth d are encoded via a pre-trained VAE into latents z_{rgb} and z_{depth}, which are patchified and concatenated into a joint spatial input $\mathcal{T}_{in}$. The MMDiT processes $\mathcal{T}_{in}$ alongside the text sequence $\mathcal{T}_{txt}$ through two stages:

1. Double Stream Blocks: The spatial and text sequences maintain independent representations but interact via shared attention:

$$\mathcal{T}_{img}^{(k)}, \mathcal{T}_{txt}^{(k)} = \text{DoubleStreamBlock}(\mathcal{T}_{img}^{(k-1)}, \mathcal{T}_{txt}^{(k-1)}, \mathcal{E}_{time}) \quad (1)$$

where $\mathcal{E}_{time}$ denotes the timestep embedding, and $\mathcal{T}_{img}^{(0)} = \mathcal{T}_{in}$.

2. Single Stream Blocks: After K double stream blocks, the sequences are concatenated for deep semantic fusion:

$$\mathcal{T}_{joint}^{(l)} = \text{SingleStreamBlock}(\mathcal{T}_{joint}^{(l-1)}, \mathcal{E}_{time}) \quad (2)$$

where $\mathcal{T}_{joint}^{(0)} = [\mathcal{T}_{txt}^{(K)} \parallel \mathcal{T}_{img}^{(K)}]$. Finally, the output sequence is un-patchified into the predicted depth latent $\hat{z}_{depth}$ and the reconstructed RGB latent $\hat{z}_{rgb}$.

To clarify the diffusion formulation, OpenDepth operates in a deterministic, discriminative mode. Instead of iterative denoising, the target depth d is strictly zero-initialized, and the MMDiT performs a single-step direct regression at a fixed timestep ($t = 1.0$). Consequently, all training objectives—including $\mathcal{L}_{feat}$ and $\mathcal{L}_{recon}$—are computed exclusively at this fixed timestep.

3.2 Advantages of I2I Editing Priors: A Motivational Analysis

Existing generative MDE methods primarily rely on T2I models (e.g., Stable Diffusion [43]). However, T2I models are optimized for semantic synthesis driven by text prompts and lack explicit constraints to maintain pixel-level structural integrity. When input images are corrupted by environmental perturbations (e.g., rain, haze, low light, or noise), T2I priors tend to produce structural hallucinations or blurry boundaries, as the degraded visual features deviate from their expected distribution. Conversely, I2I editing models possess inherent transformation priors: they are trained to alter specific attributes (style or domain) while strictly preserving the underlying geometry of the source image.

Motivational Experiment and Analysis. To validate this hypothesis, we evaluate a representative T2I-based depth model (Lotus-D [21]) against our I2I baseline (without robustness fine-tuning) by calculating the cosine similarity of intermediate spatial features between clean and perturbed images across different layers. As shown in Figure 3, the T2I model struggles to recognize the geometric essence, exhibiting extremely low similarity in initial layers. Conversely, the I2I model penetrates the perturbations early on, capturing geometric features highly consistent with the clean input and maintaining this high similarity throughout deeper layers. This phenomenon demonstrates that I2I models possess an inherent "intra-domain awareness" that resists visual degradation,

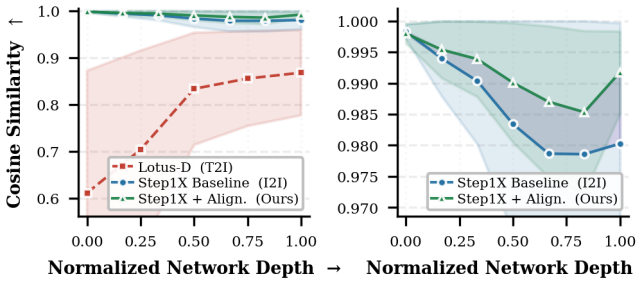


Figure 3: Cosine similarity analysis in the feature space across normalized network depths. (Left) Comparison between T2I (Lotus-D) and I2I (OpenDepth) priors. The I2I baseline exhibits significantly higher and more stable similarity than the T2I model, demonstrating inherent intra-domain awareness. (Right) Effect of our Latent Feature Alignment.

serving as the logical cornerstone for adopting an I2I editing model as our foundation for robust MDE.

3.3 Latent Feature Alignment

Although I2I priors are inherently structure-preserving, severe perturbations can still cause feature drift after cross-modal fusion. We therefore apply Latent Feature Alignment to image tokens in the first L Single Stream blocks, where features already capture preliminary cross-modal context while remaining close to the input structure.

Our goal is to establish a rigid mapping ensuring that a perturbed image x_p yields identical, fine-grained geometric features to its clean version x_c. As introduced in Section 3.1, during the Single Stream Block phase, the network has completed preliminary cross-modal interactions and begins extracting deep spatial features. Let $\mathcal{T}_{joint}^{(l)} \in \mathbb{R}^{B\times(N_{txt}+N_{img})\times C}$ denote the output sequence of the l-th Single Stream Block. We first slice the sequence to isolate the spatial tokens corresponding to the image and depth:

$$F^{(l)} = \mathcal{T}_{joint}^{(l)}[:, N_{txt}:, :] \in \mathbb{R}^{B\times N_{img}\times C} \quad (3)$$

For each sample b in the batch, we flatten its spatial features into a vector $f_b^{(l)} \in \mathbb{R}^{N_{img}\cdot C}$ and compute a per-sample cosine similarity loss. We enforce this feature consistency in the first L shallow Single Stream Blocks:

$$\mathcal{L}_{feat} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{B}\sum_{b=1}^{B}\left[1 - \frac{\text{sg}(f_{c,b}^{(l)})\cdot f_{p,b}^{(l)}}{\|\text{sg}(f_{c,b}^{(l)})\|_2 \|f_{p,b}^{(l)}\|_2}\right] \quad (4)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. Similar to the design in Section 3.4, the clean features serve as immutable anchors, ensuring that only the perturbed branch is updated to align with the canonical geometric representation.

By aligning features in the early Single Stream blocks (we use $L = 4$ in practice), the network learns to suppress environmental noise early and extract cleaner depth-relevant features. As shown in Figure 3, this alignment improves feature similarity throughout the network, especially in deeper layers, yielding more consistent

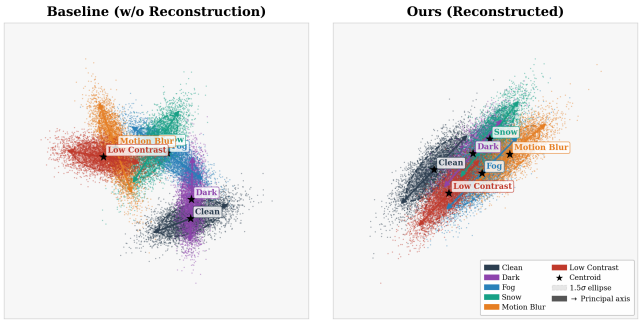


Figure 4: t-SNE visualization of deep features across diverse environmental styles. (Left) The Baseline exhibits scattered feature clusters dominated by environmental styles. (Right) Our RGB Reconstruction perfectly aligns these diverse perturbation clusters into a unified latent domain, demonstrating true style-invariance for consistent geometric perception.

geometric representations. Ablations on different alignment depths are provided in the Appendix.

3.4 Unified RGB Reconstruction in Latent Space

While feature alignment ensures the purity of shallow features, we further expect the model to "see through" appearance variations to reach a unified semantic understanding of scenes with identical geometry. To this end, we introduce an implicit restoration task: Unified RGB Reconstruction in the latent space.

By leveraging the inherent bipartite structure of the MMDiT output sequence, the architecture naturally produces a reconstructed RGB latent $\hat{z}_{rgb}$ alongside the depth estimate. Specifically, for a perturbed input, the final spatial output sequence $\mathcal{T}_{img}^{(out)}$ is first split along the channel dimension into depth prediction tokens $\mathcal{T}_{depth,p}$ and RGB reconstruction tokens $\mathcal{T}_{rgb,p}$. Subsequently, through an inverse unpatchify operation, we restore $\mathcal{T}_{rgb,p}$ into a 2D latent feature map:

$$\hat{z}_{rgb,p} = \text{Unpatchify}(\mathcal{T}_{rgb,p}) \tag{5}$$

Rather than discarding this output, we utilize it to unify the perturbed and clean branches at the feature level. When the network receives a perturbed latent $z_{rgb,p}$, we supervise its RGB output $\hat{z}_{rgb,p}$ using the clean RGB latent $z_{rgb,c}$ (extracted from the unperturbed image x_c via the VAE). The cross-domain reconstruction loss is defined as:

$$\mathcal{L}_{recon} = \|\hat{z}_{rgb,p} - \text{sg}(z_{rgb,c})\|_2^2 \tag{6}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. The stop-gradient operation ensures that the clean RGB latent serves strictly as an immutable anchor. This prevents erroneous gradient flow from degrading the clean representations, forcing the perturbed branch to actively align with the canonical domain rather than allowing the two branches to collapse toward a sub-optimal middle state.

Feature Clustering Analysis. To verify that the reconstruction module forces the model to ignore visual styles and extract pure geometric cues, we perform t-SNE visualization (Figure 4) on deep features (MMDiT Single Stream Block 37) across 20 scenes under 6 distinct visual perturbations. Without the reconstruction constraint, feature clusters for various perturbation types scatter diversely across the latent space with misaligned extension directions, revealing that the **Baseline**'s internal representation remains heavily entangled with surface-level visual noise. By integrating the reconstruction mechanism, **OpenDepth** achieves perfectly aggregated feature clusters across all perturbation types with fully aligned principal distributional axes, demonstrating that our model effectively filters out appearance interference (e.g., fog, blur, or low-light) and extracts consistent underlying geometry, achieving true perturbation invariance.

3.5 Overall Objective

During training, each batch contains a clean image x_c and its randomly perturbed counterpart x_p, sharing the same ground-truth depth d_{gt}. The depth regression loss $\mathcal{L}_{depth}$ computes the masked Mean Squared Error (MSE) between the predicted and ground-truth depth latents. The overall robust training objective is a weighted sum of the clean depth loss, perturbed depth loss, feature alignment loss, and cross-domain RGB reconstruction loss:

$$\mathcal{L}_{total} = \mathcal{L}_{depth}(x_c) + \alpha\mathcal{L}_{depth}(x_p) + \beta\mathcal{L}_{feat} + \gamma\mathcal{L}_{recon} \tag{7}$$

where α, β, and γ are hyperparameters that balance perturbation robustness, feature consistency, and the contribution of latent reconstruction. Empirically, we set $\alpha = 1$, $\beta = 1$, and $\gamma = 1$ as the optimal configuration, with detailed sensitivity ablations provided in the Appendix. This rigorous, multi-stage alignment strategy effectively capitalizes on editing priors to complete dense geometric tasks in complex scenarios.

4 Experiments

4.1 Experiment Setup

Implementation details. We build our method upon the Step1X-Edit [31] v1.0 framework. During finetuning, all parameters except for the DiT module are frozen, and the language control input is left blank. The process employs LoRA with rank = 64 and scale factor $\alpha = 32$. We trained for 20 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 8. With gradient checkpointing enabled, the entire training was conducted on NVIDIA RTX 4090 GPUs.

Training datasets. We train our model on a mixture of three datasets: Hypersim [42], Virtual KITTI [6], and Cityscapes [10]. Following the data split and preprocessing protocol of Lotus [21], we use approximately 39k images from Hypersim (a photorealistic indoor synthetic dataset), 20k samples from Virtual KITTI (a synthetic street-view dataset with a maximum depth of 80m), and additionally incorporate 20k samples from Cityscapes, a real-world urban driving dataset. Emulating Marigold's [27] dataset probability sampling mechanism, we adopt a per-dataset sampling ratio of 8:1:1 for Hypersim, Virtual KITTI, and Cityscapes, respectively.

Perturbation strategies. To synthesize the perturbed counterparts x_p, we utilize visual degradations (e.g., low-light, fog, snow, blur) following DepthAnything-AC [50]. Unlike applying single random perturbations, we adopt a rigorous compounding strategy during training: low-light degradation with noise is consistently applied as a base perturbation to all x_p samples, while other environmental degradations are probabilistically injected on top in

Table 1: Zero-shot relative depth estimation results on general benchmarks. Comparison with general depth estimation models and robust depth estimation models. Best results are in bold, and second best are underlined. * indicates models based on pre-trained Stable Diffusion.

Method	Encoder	KITTI [17]		NYUv2 [47]		Sintel [5]		ETH3D [45]		DIODE(indoor) [55]		Avg Rank↓
		AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	
General Depth Estimation Model												
DepthPro [3]	ViT-S	6.8	94.7	5.2	97.2	24.0	68.2	5.7	96.3	6.0	95.2	4.3
DepthAnything V1 [69]	ViT-S	8.2	93.6	5.3	97.2	21.1	73.6	6.4	96.6	7.8	94.0	5.2
DepthAnything V2 [70]	ViT-S	8.3	93.4	5.1	97.3	24.0	71.2	6.4	96.9	7.3	94.3	4.9
Marigold$^*_{v1.1}$ [27]	VAE	9.9	91.6	5.5	96.4	42.5	53.5	6.5	95.9	31.0	87.5	10.2
Lotus-D* [21]	VAE	8.8	91.2	5.2	96.8	29.9	64.9	5.7	96.7	8.3	92.1	7.3
Diffusion-E2E-FT* [15]	VAE	9.6	92.1	5.4	96.5	45.7	57.1	6.4	95.9	28.6	87.7	9.3
FE2E [58]	VAE	6.6	96.0	4.1	97.7	21.0	73.1	3.8	**98.7**	6.2	94.8	2.2
Robust Depth Estimation Model												
DynaDepth [78]	ResNet	8.5	91.6	21.2	63.0	45.8	39.1	19.2	70.3	18.0	73.7	12.1
EC-Depth [49]	ViT-S	7.5	93.1	22.0	60.6	44.7	40.5	18.1	77.9	17.2	76.0	10.2
STEPS [84]	ResNet	20.4	68.3	22.0	60.6	45.8	37.5	19.7	69.9	21.4	65.2	13.7
robustdepth [44]	ViT-S	7.6	92.9	17.4	71.8	41.1	43.6	18.5	77.6	18.2	74.1	10.2
weather-depth [57]	ViT-S	7.5	93.2	16.7	73.3	41.9	44.3	18.1	77.1	17.0	76.0	9.2
DepthAnything-AC [50]	ViT-S	8.3	93.4	5.1	97.3	23.5	71.5	6.5	96.8	7.3	94.1	5.1
OpenDepth (Ours)	VAE	**6.3**	**96.2**	**3.9**	**98.0**	**19.9**	**76.2**	**3.7**	98.4	**5.5**	**95.3**	**1.1**

Table 2: Zero-shot relative depth estimation results on open-world benchmarks. Comparison with general depth estimation models and robust depth estimation models. * indicates models based on pre-trained Stable Diffusion.

Method	Encoder	NuScenes-night [7]		RobotCar-night [34]		DS-rain [66]		DS-cloud [66]		DS-fog [66]		Avg Rank↓
		AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	AbsRel↓	$\delta1$ ↑	
General Depth Estimation Model												
DepthPro [3]	ViT-S	21.8	66.9	23.7	53.4	12.4	84.1	15.8	77.9	10.2	89.2	5.1
DepthAnything V1 [69]	ViT-S	23.2	67.9	23.9	51.8	13.3	81.9	15.0	80.1	9.8	89.1	5.6
DepthAnything V2 [70]	ViT-S	20.0	72.5	23.9	51.8	12.5	84.0	15.1	79.8	10.3	89.0	5.4
MoGe[60]	ViT-L	32.9	41.5	37.9	27.6	23.7	53.2	24.5	52.7	24.4	47.9	13.4
Marigold$^*_{v1.1}$ [27]	VAE	32.2	49.2	24.4	60.2	16.0	79.7	18.5	74.9	13.4	85.9	11.2
Lotus-D* [21]	VAE	22.1	66.8	24.4	53.4	12.2	83.8	15.4	78.2	10.2	88.3	7.4
FE2E [58]	VAE	21.9	67.2	24.1	59.6	12.0	84.0	15.1	78.5	10.0	88.6	7.1
Diffusion-E2E-FT* [15]	VAE	31.7	50.9	23.3	64.4	15.0	82.8	16.6	79.5	12.7	87.4	9.3
Robust Depth Estimation Model												
DynaDepth [78]	ResNet	38.1	39.4	51.2	29.4	23.9	60.6	17.2	60.8	14.4	90.1	13.8
EC-Depth [49]	ViT-S	24.3	62.3	22.8	55.2	15.5	76.6	15.8	76.7	10.9	86.1	7.6
STEPS [84]	ResNet	25.2	58.8	35.0	36.7	30.1	48.0	25.2	58.8	21.6	64.1	14.8
robustdepth [44]	ViT-S	26.0	59.7	31.1	52.1	16.7	75.5	16.8	77.5	10.5	88.2	10.5
weather-depth [57]	ViT-S	-	-	-	-	15.8	76.4	16.0	76.7	10.5	87.9	12.5
DepthAnything-AC [50]	ViT-S	19.8	72.7	22.7	55.5	12.5	84.0	14.9	80.1	10.3	88.9	3.2
OpenDepth (Ours)	VAE	**14.5**	**81.2**	**17.0**	**73.3**	**9.7**	**89.3**	**12.5**	**84.2**	**8.5**	**92.0**	**1.0**

a mutually exclusive manner. This forces the model to learn resilient representations under multi-factor interference. Detailed perturbation configurations are provided in the Appendix.

Evaluation benchmarks. We evaluate the zero-shot performance of our method on both adverse-condition and standard benchmarks following DepthAnything-AC [50]. Specifically, we use NuScenes-night [7], RobotCar-night [34], and Driving-Stereo [66] for real-world adverse conditions, and KITTI-C [28] for synthetic corruptions. For standard generalization, we report results on KITTI [17], NYUv2 [47], Sintel [5], ETH3D [45], and DIODE [55]. For all relative-depth benchmarks, predictions are resized to the ground-truth resolution, aligned per image in disparity space on the cropped valid region, clipped to the dataset-specific depth range, and evaluated with AbsRel and δ_1, averaged over images.

4.2 Zero-Shot Relative Depth Estimation on General Benchmarks

Performance on general benchmarks. As demonstrated by the average rank in Table 1, our OpenDepth model achieves SOTA performance with a top rank of 1.1, successfully outperforming both the general depth estimation models and the specialized robust depth estimation models. Furthermore, the results indicate that I2I-based models, specifically FE2E [58] and ours, consistently surpass Text-to-Image models built on pre-trained Stable Diffusion. This

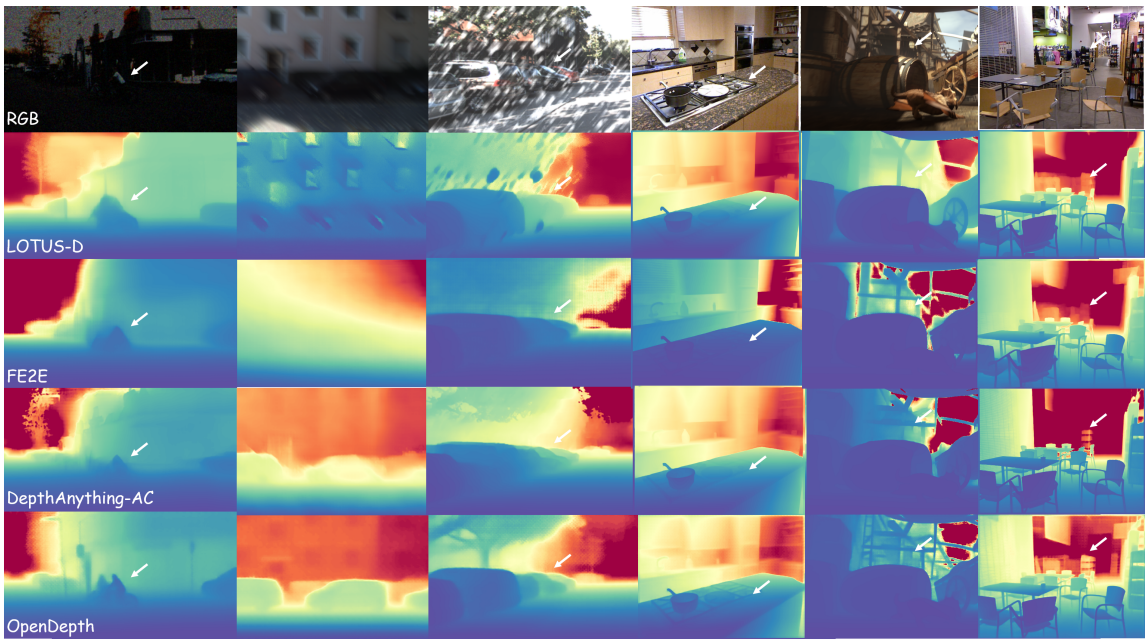


Figure 5: Qualitative comparison on zero-shot depth estimation. OpenDepth demonstrates superior detail preservation and structural consistency over Lotus-D [21], FE2E [58], and DepthAnything-AC [50] across both general and open-world conditions, as highlighted by white arrows.

clearly validates the superiority of editing priors over generative priors even in standard, general scenarios. Finally, while many foundation models rely on massive amounts of unlabeled training data to achieve strong zero-shot generalization, our method achieves leading performance using merely 79K labeled samples. This highlights the exceptional training data efficiency of our proposed framework.

4.3 Zero-Shot Relative Depth Estimation on Open-World Benchmarks

Performance on real complex lighting and weather scenarios. OpenDepth achieves SOTA results across all open-world adverse benchmarks (e.g., NuScenes-night, DS-rain, as detailed in Table 2), significantly outperforming both general and specialized robust models. This proves our architecture effectively leverages editing priors to maintain stable geometry perception under severe visual degradation (see visual comparisons in Figure 5).

Performance on KITTI-C benchmarks. As shown in Table 3, OpenDepth demonstrates exceptional resilience against synthetic corruptions, securing the best performance in three out of four categories (Dark, Motion, and Gaussian noise). This confirms our mechanisms successfully filter out severe appearance shifts for consistent depth prediction.

4.4 Ablation Study

To validate the effectiveness of our proposed components, we conduct an ablation study on four open-world adverse condition datasets (see Table 4). Our baseline employs a DiT-based image editing model whose inherent transformation priors naturally preserve geometric features in complex environments.

We first compare generative (G) and discriminative (D) formulations of our OpenDepth baseline. As shown in Table 4, OpenDepth(D) consistently outperforms OpenDepth(G) across all benchmarks (e.g., AbsRel 18.0 vs. 19.3 on NuScenes-night), corroborating the finding in Lotus [21] that deterministic prediction is more suitable than stochastic generation for dense geometric tasks. Consequently, all OpenDepth results reported in the main comparison tables refer to this discriminative formulation.

Effects of Perturbation Training ($\mathcal{L}_{depth}(x_p)$). Introducing perturbation training yields a consistent improvement, reducing the AbsRel error from 18.0 to 16.6 on NuScenes-night. Notably, even applying the same perturbation strategy to the T2I baseline (Lotus-D) only marginally improves its performance, still substantially lagging behind our OpenDepth(D)+P (16.6). This confirms that perturbation training alone cannot compensate for the inherent limitations of T2I priors, and the superiority of our I2I foundation.

Effects of Feature Alignment ($\mathcal{L}_{feat}$). Adding $\mathcal{L}_{feat}$ further reduces AbsRel from 16.6 to 14.8 on NuScenes-night and from 19.6 to 18.5 on RobotCar-night, demonstrating that our alignment mechanism effectively anchors perturbed inputs to the same fine-grained geometric representations as clean images.

Effects of Unified RGB Reconstruction ($\mathcal{L}_{recon}$). Integrating $\mathcal{L}_{recon}$ yields the most pronounced gains, with AbsRel decreasing from 9.4 to 8.5 on DS-fog and from 18.5 to 17.0 on RobotCar-night. This confirms its critical role in projecting diverse degraded observations into a unified latent domain, realizing robust "many-to-one" geometric translation in open-world scenarios.

Table 3: Zero-shot relative depth estimation results on synthetic KITTI-C [28] benchmarks.

Method	Encoder	Dark		Snow		Motion		Gaussian	
		AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
DynaDepth [78]	ResNet	16.3	75.2	33.8	39.3	23.4	60.9	27.4	50.1
STEPS [84]	ResNet	23.0	63.1	24.2	62.2	29.1	50.8	20.4	69.2
DepthPro [3]	ViT-S	14.5	79.3	19.7	68.5	17.0	74.6	17.0	74.5
DepthAnything V2 [70]	ViT-S	13.0	83.2	11.5	87.2	12.7	84.0	15.7	78.5
DepthAnything-AC [50]	ViT-S	13.0	83.4	**11.4**	**87.3**	12.6	84.1	15.3	79.3
OpenDepth (Ours)	VAE	**12.6**	**83.9**	11.7	86.9	**11.8**	**85.3**	**14.1**	**81.3**

Table 4: Ablation study of different components. T2I (Lotus-D [21]) and Ours(G) are included as baselines for comparison. (G) and (D) denote generative and discriminative models. P: perturbation, L_{feat}: feature alignment, L_{recon}: RGB reconstruction.

Baseline	Modules			DS-Rain [66]		DS-Fog [66]		NuScenes-night [7]		RobotCar-night [34]	
	P	L_{feat}	L_{recon}	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
T2I (Lotus-D) [21]				12.2	83.8	10.2	88.3	22.1	66.8	24.4	53.4
	✓			12.0	84.6	10.0	88.9	20.1	69.8	22.5	57.3
OpenDepth(G)				12.1	85.1	10.2	88.8	19.3	71.9	21.2	60.4
OpenDepth(D)				11.9	85.5	10.3	89.4	18.0	73.7	20.4	62.5
	✓			11.8	85.9	9.7	89.7	16.6	77.0	19.6	66.1
	✓	✓		11.5	86.2	9.4	90.4	14.8	79.8	18.5	71.1
	✓	✓	✓	**9.7**	**89.3**	**8.5**	**92.0**	**14.5**	**81.2**	**17.0**	**73.3**



Figure 6: Qualitative comparison of multi-scale edge responses under severe low-light (Dark) perturbations.

4.5 Qualitative Analysis

As shown in Figure 5, , we visualize multi-scale edge responses as geometric proxies under severe low-light (Dark) perturbations (Figure 6). Specifically, these responses are extracted by fusing Sobel gradients, Canny edges, and Laplacian responses from predicted depth maps. The **Baseline** suffers from severe structural degradation with blurred contours and noisy responses, indicating that distribution shifts can disrupt geometric consistency even with I2I priors. By introducing **Latent Feature Alignment**, the model suppresses noise and restores structural patterns by anchoring degraded observations to canonical geometric representations. Finally, the **Full Model (with RGB Reconstruction)** produces sharp and coherent feature maps, accurately recovering fine structures such as chair-leg intersections. This demonstrates that unified RGB reconstruction effectively enforces a style-invariant latent domain for high-fidelity geometric perception.

5 Conclusion

In this paper, we propose OpenDepth to elevate robust Monocular Depth Estimation in open-world environments. Through systematic analysis, we demonstrate the inherent superiority of Image-to-Image editing priors over Text-to-Image models in preserving structural integrity against environmental degradations. To resolve the representation misalignment caused by diverse perturbations, we equip our model with "Intra-Domain Awareness" via two core mechanisms: shallow feature alignment and joint RGB latent reconstruction. This strategy successfully projects varied degraded inputs into a unified latent space, strictly decoupling stylistic interference from geometric perception. Extensive evaluations confirm that OpenDepth achieves exceptional detail preservation and robust zero-shot generalization, setting new SOTA records across both standard and extreme weather benchmarks.

6 Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions. We also thank all members of our research group for their valuable discussions and support during the development of this work.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4009–4018.

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).

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

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*. Springer, 611–625.

[6] Yohann Cabon, Naila Murray, and Martin Humenberger. 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773* (2020).

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[8] Hemang Chawla, Kishaan Jeeveswaran, Elahe Arani, and Bahram Zonooz. 2023. Image masking for robust self-supervised monocular depth estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10054–10060.

[9] Xingyu Chen, Thomas H Li, Ruonan Zhang, and Ge Li. 2023. Frequency-aware self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5808–5817.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[11] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. 2020. DepthLab: Real-time 3D interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 829–843.

[12] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10786–10796.

[13] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

[14] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. 2024. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*. Springer, 241–258.

[15] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. 2025. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the Winter Conference on Applications of Computer Vision*. 753–762.

[16] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. 2023. Robust monocular depth estimation under challenging conditions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8177–8186.

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3354–3361.

[18] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 270–279.

[19] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3828–3838.

[20] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. 2025. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 3203–3211.

[21] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. 2024. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124* (2024).

[22] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).

[23] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 10579–10596.

[24] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. 2025. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2005–2015.

[25] Kui Jiang, Jing Cao, Zhaocheng Yu, Junjun Jiang, and Jingchun Zhou. 2025. Always clear depth: Robust monocular depth estimation under adverse weather. In *Proc. Int. Joint Conf. Artif. Intell.* 1251–1259.

[26] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 2022. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18963–18974.

[27] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9492–9502.

[28] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. 2023. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems* 36 (2023), 21298–21342.

[29] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. 2024. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10016–10025.

[30] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2025. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025).

[31] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. 2025. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761* (2025).

[32] Yu Liu, Kun Sun, Chang Tang, Yuhua Qian, and Xin Li. 2025. Tpdepth: Leveraging text prompts with controlnet to boost diffusion-based depth estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 4290–4299.

[33] Yihao Liu, Feng Xue, Anlong Ming, Mingshuai Zhao, Huadong Ma, and Nicu Sebe. 2024. Sm4depth: Seamless monocular metric depth estimation across multiple cameras and scenes by one model. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3469–3478.

[34] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 2017. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research* 36, 1 (2017), 3–15.

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

[36] Suraj Patni, Aradhye Agarwal, and Chetan Arora. 2024. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 28285–28295.

[37] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. 2023. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21477–21487.

[38] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. 2024. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10106–10116.

[39] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12179–12188.

[40] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.

[41] Zeyu Ren, Zeyu Zhang, Wukai Li, and Hao Tang. 2026. AnyDepth: Depth Estimation Made Easy. (2026).

[42] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In

Proceedings of the IEEE/CVF international conference on computer vision. 10912–10922.

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

[44] Kieran Saunders, George Vogiatzis, and Luis J Manso. 2023. Self-supervised monocular depth estimation: Let's talk about the weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8907–8917.

[45] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3260–3269.

[46] Yiqing Shi, Yiren Song, and Mike Zheng Shou. 2025. Edit2Perceive: Image Editing Diffusion Models Are Strong Dense Perceivers. *arXiv preprint arXiv:2511.18673* (2025).

[47] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 746–760.

[48] Ziyang Song, Zerong Wang, Bo Li, Hao Zhang, Ruijie Zhu, Li Liu, Peng-Tao Jiang, and Tianzhu Zhang. 2025. Depthmaster: Taming diffusion models for monocular depth estimation. *arXiv preprint arXiv:2501.02576* (2025).

[49] Ziyang Song, Ruijie Zhu, Chuxin Wang, Jiacheng Deng, Jianfeng He, and Tianzhu Zhang. 2023. EC-Depth: Exploring the consistency of self-supervised monocular depth estimation in challenging scenes. *arXiv preprint arXiv:2310.08044* (2023).

[50] Boyuan Sun, Modi Jin, Bowen Yin, and Qibin Hou. 2025. Depth anything at any condition. *arXiv preprint arXiv:2507.01634* (2025).

[51] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. 2023. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE transactions on pattern analysis and machine intelligence* 46, 1 (2023), 497–508.

[52] Lei Tai, Jingwei Zhang, Ming Liu, and Wolfram Burgard. 2018. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 1111–1117.

[53] Madhu Vankadari, Sourav Garg, Anima Majumder, Swagat Kumar, and Ardhendu Behera. 2020. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *European Conference on Computer Vision*. Springer, 443–459.

[54] Madhu Vankadari, Samuel Hodgson, Sangyun Shin, Kaichen Zhou, Andrew Markham, and Niki Trigoni. 2024. Dusk till dawn: Self-supervised nighttime stereo depth estimation using visual foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 17976–17982.

[55] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. 2019. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463* (2019).

[56] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. 2025. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905* (2025).

[57] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. 2024. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4976–4982.

[58] JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. 2025. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338* (2025).

[59] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. 2021. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16055–16064.

[60] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2025. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5261–5271.

[61] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. 2019. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8445–8453.

[62] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).

[63] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. 2024. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090* (2024).

[64] Weilong Yan, Ming Li, Haipeng Li, Shuwei Shao, and Robby T Tan. 2025. Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21880–21890.

[65] Weilong Yan, Xin Zhang, and Robby T Tan. 2025. ER-LoRA: Effective-Rank Guided Adaptation for Weather-Generalized Depth Estimation. *arXiv preprint arXiv:2509.00665* (2025).

[66] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. 2019. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 899–908.

[67] Guanglei Yang, Rui Tian, Yongqiang Zhang, Zhun Zhong, Yongqiang Li, and Wangmeng Zuo. 2024. Multi-modality driven LoRA for adverse condition depth estimation. *arXiv preprint arXiv:2412.20162* (2024).

[68] Haolin Yang, Chaoqiang Zhao, Lu Sheng, and Yang Tang. 2024. Self-supervised monocular depth estimation in the dark: Towards data distribution compensation. *arXiv preprint arXiv:2404.13854* (2024).

[69] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10371–10381.

[70] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *Advances in Neural Information Processing Systems* 37 (2024), 21875–21911.

[71] Wenzheng Yang, Songwei Pei, Bingfeng Liu, Qian Li, and Shangguang Wang. 2025. OGDepth: Leveraging Object Guidance in Diffusion Models for Enhanced Monocular Depth Estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 4378–4387.

[72] Yuanqi Yao, Gang Wu, Kui Jiang, Siao Liu, Jian Kuai, Xianming Liu, and Junjun Jiang. 2024. Improving domain generalization in self-supervised monocular depth estimation via stabilized adversarial training. In *European Conference on Computer Vision*. Springer, 183–201.

[73] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. 2020. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569* (2020).

[74] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9043–9053.

[75] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. 2021. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 204–213.

[76] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. 2022. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3916–3925.

[77] Longjian Zeng, Zunjie Zhu, Rongfeng Lu, Ming Lu, Bolun Zheng, Chenggang Yan, and Anke Xue. 2025. DepthDark: robust monocular depth estimation for low-light environments. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 11239–11248.

[78] Sen Zhang, Jing Zhang, and Dacheng Tao. 2022. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics. In *European Conference on Computer Vision*. Springer, 143–160.

[79] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. 2024. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. *arXiv preprint arXiv:2407.17952* (2024).

[80] Yafei Zhang, Shen Zhou, and Huafeng Li. 2024. Depth information assisted collaborative mutual promotion network for single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2846–2855.

[81] Zezheng Zhang, Ryan KY Chan, and Kenneth KY Wong. 2024. GlocalFuse-Depth: Fusing transformers and CNNs for all-day self-supervised monocular depth estimation. *Neurocomputing* 569 (2024), 127122.

[82] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5729–5739.

[83] Shuhong Zheng, Zhipeng Bao, Ruoyu Zhao, Martial Hebert, and Yu-Xiong Wang. 2024. Diff-2-in-1: Bridging generation and dense perception with diffusion models. *arXiv preprint arXiv:2411.05005* (2024).

[84] Yupeng Zheng, Chengliang Zhong, Pengfei Li, Huan-ang Gao, Yuhang Zheng, Bu Jin, Ling Wang, Hao Zhao, Guyue Zhou, Qichao Zhang, et al. 2023. Steps: Joint self-supervised nighttime image enhancement and depth estimation. *arXiv preprint arXiv:2302.01334* (2023).

[85] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. 2017. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1851–1858.

[86] Ruijie Zhu, Chuxin Wang, Ziyang Song, Li Liu, Tianzhu Zhang, and Yongdong Zhang. 2024. Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation. *arXiv preprint arXiv:2407.08187* (2024).