ApDepth-G: Multi-Step Geometric Diffusion Refinement for Robust Monocular Depth Estimation

Jiawei Wang[a,b,*,1], Mingbo Lei[a,b,1], Haoze Shou[a,b], Yusu Liang[a,b] and Jingxuan Wang[c]

[a]School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang, China
[b]National Special Computer Engineering Technology Research Center (Shenyang Branch), Shenyang, China
[c]School of Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China

ARTICLE INFO

Keywords:
Diffusion Model
Monocular Depth Estimation
Sky Collapse
Depth Refinement
Geometric Diffusion

ABSTRACT

Diffusion-based monocular depth estimation methods have shown strong generalization ability through generative priors and iterative denoising. However, they often suffer from sky-depth collapse in outdoor scenes, where sky regions that should be distant and smooth are incorrectly estimated as finite, distorted, or affected by pseudo-texture artifacts. This issue also appears in our previous ApDepth framework and is common in diffusion-based depth models due to the ambiguity of textureless and cloud-textured regions. To address this problem, we propose **ApDepth-G**, a multi-step geometric diffusion refinement framework for sky-stable monocular depth estimation. ApDepth-G progressively refines latent depth representations through multi-step denoising and improves training with offset noise regularization, SNR-based loss reweighting, and latent gradient consistency. In addition, sky-aware refinement is introduced to suppress unreliable sky-region responses and reduce pseudo-depth structures. Experimental observations show that ApDepth-G effectively alleviates sky-depth collapse in challenging outdoor scenes while maintaining robust depth estimation performance. Codes are available at: https://github.com/Haruko386/ApDepth-G.

1. Introduction

Monocular depth estimation (MDE) aims to recover a dense geometric representation from a single RGB image and serves as a fundamental component in autonomous driving, robotic perception, three-dimensional reconstruction, and augmented reality. Despite substantial progress, MDE remains inherently ill-posed because different three-dimensional scenes may produce similar two-dimensional observations. A reliable estimator must therefore infer not only local appearance cues, but also globally coherent scene geometry and plausible long-range structures [5, 14].

*Corresponding author
Haruko386@outlook.com (J. Wang)
ORCID(s): 0009-0006-0793-6698 (J. Wang); 0009-0001-3198-8045 (M. Lei)
[1]These authors contributed equally to this work.

Modern discriminative methods address this ambiguity by directly learning image-to-depth mappings from large-scale annotated, synthetic, or pseudo-labeled datasets. Representative foundation models, such as Depth Anything V2, achieve strong zero-shot generalization through powerful visual encoders and extensive supervision [14]. Nevertheless, their performance remains dependent on the coverage and quality of the training distribution. Difficult regions such as transparent objects, reflective surfaces, low-texture areas, distant backgrounds, and the sky provide weak or misleading visual evidence and therefore remain challenging.

The emergence of large-scale diffusion models provides an alternative approach to monocular geometry estimation. Latent diffusion models such as Stable Diffusion encode rich semantic and structural knowledge learned from internet-scale image–text data [18, 21]. Recent methods have repurposed these generative priors for dense prediction by formulating depth estimation as an image-conditioned denoising process [4, 7, 9, 12]. Marigold demonstrates that a pretrained latent diffusion model can be efficiently adapted to affine-invariant depth estimation using only synthetic data while retaining strong zero-shot generalization [12].

Subsequent studies have explored different ways to improve the reliability, efficiency, and geometric fidelity of diffusion-based depth estimation. Pixel-Perfect Depth performs diffusion directly in pixel space to avoid edge artifacts and flying pixels introduced by VAE compression, while using semantic prompts and a cascaded diffusion transformer to preserve both global structure and fine details [28]. CDPR introduces physically grounded polarization cues and a confidence-aware latent fusion mechanism to improve prediction in textureless, transparent, and reflective regions, demonstrating the importance of reliable auxiliary information when RGB observations are ambiguous [30]. Lotus-2 further decouples global geometric prediction and detail reconstruction through a deterministic two-stage framework, where a constrained multi-step process progressively refines fine-grained geometry without altering the established global structure [10]. These studies reveal that diffusion-based depth estimation still depends critically on the quality of conditioning priors, the representation space, and the design of the refinement trajectory.

Meanwhile, reducing the iterative denoising process to a single forward pass has become an important direction for improving inference efficiency [8, 23, 26]. Our previous work, ApDepth, follows this direction by converting stochastic multi-step depth generation into deterministic single-step prediction and introducing a pretrained Depth Anything V2 estimator as a geometric prior [24]. Although this design substantially improves efficiency and preserves robust global structures, further inspection reveals a persistent failure mode in outdoor scenes: predicted sky regions may collapse toward finite or spatially inconsistent depth.

Sky regions should generally appear distant and geometrically smooth. However, diffusion-based estimators may produce local depressions, cloud-shaped pseudo-depth structures, or discontinuous responses near the horizon. This problem arises because sky appearance contains few reliable geometric cues. Illumination gradients, cloud textures, overexposure, and weak boundaries can be incorrectly interpreted as depth variations. Moreover, the pretrained depth prior itself may assign unstable finite values to the sky. When the entire correction process is compressed into one deterministic step, the model has limited opportunity to revise these erroneous responses.

This observation motivates us to reconsider iterative denoising as a geometric correction mechanism rather than merely a stochastic generation process. Multiple denoising steps allow the model to repeatedly revise an uncertain latent depth representation, progressively suppressing unreliable local structures while maintaining the global scene layout. Based on this perspective, we propose **ApDepth-G**, a multi-step geometric diffusion refinement framework designed to improve the stability of sky and long-range depth estimation.

ApDepth-G retains the depth-prior-guided formulation of ApDepth and restores a progressive reverse diffusion trajectory. The RGB latent, pretrained depth-prior latent, and noisy depth latent are jointly processed by the denoising U-Net, allowing each step to correct inconsistent geometry under semantic guidance. To stabilize optimization across different noise levels, we introduce offset-noise regularization and signal-to-noise-ratio-based loss reweighting. We additionally reconstruct the clean depth latent during training and impose latent gradient consistency to preserve meaningful structural boundaries while suppressing irregular local responses. Together, these components transform the reverse process into a stable geometric refinement trajectory for ambiguous outdoor regions.

ApDepth-G is not intended to universally replace the efficient single-step ApDepth model. Instead, the two frameworks provide complementary operating regimes: ApDepth emphasizes fast deterministic inference, whereas ApDepth-G trades part of this efficiency for stronger progressive correction and improved robustness in difficult outdoor scenes.

The main contributions of this work are summarized as follows:

- We identify and study *sky-depth collapse*, a characteristic failure mode in which diffusion-based monocular depth models map visually ambiguous sky regions to finite, locally distorted, or pseudo-textured depth structures.
- We propose ApDepth-G, a depth-prior-guided multi-step diffusion framework that progressively corrects uncertain latent depth representations and improves the stability of sky and long-range regions.
- We introduce a stabilized training strategy combining offset-noise regularization, SNR-based loss reweighting, and latent gradient consistency to balance diffusion optimization and preserve geometric boundaries.
- We establish a complementary extension to ApDepth, providing a multi-step geometric refinement regime for applications where outdoor robustness is more important than minimum inference latency.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation aims to recover dense scene geometry from a single RGB image. Since the projection from three-dimensional scenes to two-dimensional observations inevitably discards geometric information, the task is fundamentally ill-posed. Early learning-based methods directly regress depth from images using convolutional neural networks. Eigen et al. [5] introduced a multi-scale architecture that jointly models global scene layout and local depth structures. Subsequent works improved depth representation and optimization through ordinal regression [6], adaptive depth bins [2], local planar guidance [13], conditional random fields [31], and transformer-based dense prediction [16].

To improve cross-dataset generalization, several methods learn affine-invariant or relative depth representations from heterogeneous datasets. MiDaS [17] combines multiple datasets with scale- and shift-invariant objectives, while DPT [16] leverages vision transformers to capture long-range dependencies. More recently, Depth Anything [29] and Depth Anything V2 [14] exploit large-scale visual foundation models and extensive pseudo-labeled or synthetic training data, substantially improving zero-shot generalization. MoGe [25] directly predicts affine-invariant point maps from open-domain images, providing a unified representation of monocular geometry.

Despite their strong performance, discriminative models primarily rely on deterministic regression. At depth discontinuities, minimizing regression losses can encourage intermediate predictions between foreground and background, producing over-smoothed boundaries and flying pixels in reconstructed point clouds [28]. Their robustness also remains dependent on the diversity and coverage of supervised data. Regions with weak visual evidence, including transparent surfaces, specular materials, textureless areas, distant structures, and the sky, remain particularly difficult.

2.2. Diffusion Models for Monocular Depth Estimation

Diffusion models have emerged as an alternative paradigm for monocular depth estimation because of their powerful generative priors and ability to model complex conditional distributions. DiffusionDepth [3] formulates metric depth estimation as an iterative denoising process conditioned on monocular image features. DepthGen [20] extends conditional diffusion to multiple dense prediction tasks, while DDVM [19] investigates large-scale diffusion pretraining for visual perception.

A major development is the adaptation of pretrained text-to-image diffusion models to dense prediction. Marigold [12] fine-tunes Stable Diffusion using synthetic image–depth pairs and demonstrates strong zero-shot affine-invariant depth estimation. Its success shows that visual priors learned from internet-scale image–text data can be transferred to geometric prediction with relatively limited task-specific supervision. GeoWizard [7] extends this paradigm to joint depth and surface-normal estimation and introduces scene-distribution separation to improve geometric consistency. DepthFM [9] replaces the curved diffusion trajectory with flow matching, enabling more efficient transport between image and depth distributions while preserving the advantages of generative modeling.

However, directly retaining the original stochastic, multi-step image-generation formulation may be suboptimal for deterministic dense prediction. GenPercept [27] systematically studies the adaptation of diffusion models and finds that stochastic sampling can negatively affect deterministic perception tasks. It therefore proposes a single-step deterministic fine-tuning framework with task-specific image-space supervision. E2E-FT [8] identifies an inefficient inference configuration in earlier diffusion depth pipelines and demonstrates that a corrected single-step schedule, followed by end-to-end task-specific fine-tuning, can achieve highly competitive accuracy with substantially lower inference cost. Lotus [11] further argues that noise prediction and iterative denoising are not necessary for dense prediction, and reformulates the model to directly predict annotations in a single step.

Several works focus on recovering the details lost by compressed or deterministic inference. DepthMaster [23] introduces feature alignment to reduce texture overfitting and Fourier enhancement to balance global structures and high-frequency details. FiffDepth [1] transforms a diffusion generator into a feed-forward estimator while preserving useful generative features and introducing semantic supervision from DINOv2. Our previous ApDepth framework [24] similarly adopts a depth-prior-guided single-step formulation, using a pretrained Depth Anything V2 estimator to provide a stable global geometric reference.

The representation space also affects prediction quality. Most diffusion-based depth models operate in the latent space of a pretrained variational autoencoder for computational efficiency. Pixel-Perfect Depth [28] points out that VAE compression can degrade depth discontinuities and introduce flying pixels. It instead performs diffusion directly in pixel space and proposes Semantics-Prompted Diffusion Transformers and a cascaded token design to jointly preserve global semantic coherence and fine-grained geometry. Although pixel-space diffusion alleviates latent compression artifacts, it introduces substantially higher optimization and computational complexity.

2.3. Reliable Priors and Geometric Refinement

Ambiguous image regions remain a central challenge for monocular depth estimation. Recent works therefore incorporate additional semantic, physical, or geometric priors to constrain the depth solution space. Iris [32] integrates language descriptions into diffusion-based depth estimators. Text conditions provide semantic and spatial information about objects and scene layouts, improving depth prediction in small, occluded, poorly illuminated, or visually ambiguous regions. CDPR [30] introduces polarization observations as physically grounded auxiliary cues. It encodes RGB and polarization information into a shared latent space and employs a confidence-aware gating mechanism to adaptively suppress unreliable polarization signals while retaining useful information around transparent, reflective, and textureless surfaces.

Other works combine discriminative predictions with generative refinement. SharpDepth [15] uses diffusion distillation to combine the metric accuracy of discriminative models with the sharp structural boundaries produced by generative priors. Lotus-2 [10] separates global structure estimation from detail recovery. Its first stage produces a deterministic and globally coherent geometric prediction, while a constrained multi-step rectified-flow stage progressively enhances fine-grained details within the manifold established by the initial predictor. This design demonstrates that iterative computation can be used as a controlled refinement process rather than as unrestricted stochastic generation.

The accuracy of far-range structures has recently received increasing attention. VistaDepth [33] observes that diffusion-based depth models are biased toward near-range regions because of the long-tailed depth distribution and may fail to preserve high-frequency details in distant structures. It introduces latent frequency modulation to enhance distant geometric details and an adaptive BiasMap to reweight the diffusion objective across timesteps. These findings are closely related to outdoor sky estimation, where the target depth lies at the extreme end of the depth distribution and contains few reliable local cues.

Unlike these methods, our work focuses on the characteristic *sky-depth collapse* observed in diffusion-based monocular depth estimation. The sky provides neither stable texture nor finite geometric structure, while cloud patterns and illumination gradients can easily be misinterpreted as depth variations. ApDepth-G retains the pretrained depth prior introduced in ApDepth but employs a multi-step reverse process as progressive geometric correction. Offset-noise regularization, SNR-based loss reweighting, and latent gradient consistency are jointly used to stabilize the denoising trajectory and suppress irregular responses in sky and long-range regions.

3. Method

3.1. Overview

Given an RGB image $X \in \mathbb{R}^{H\times W\times 3}$, our goal is to estimate an affine-invariant depth map $\hat{d}$. ApDepth-G builds on the latent diffusion formulation of Marigold [12] and the depth-prior conditioning introduced in ApDepth [24], but restores a multi-step reverse trajectory for progressive geometric correction. The central motivation is that a single deterministic pass has limited opportunity to revise unreliable responses in weakly constrained regions. In contrast, repeated denoising allows the model to update an uncertain depth representation while repeatedly consulting both image evidence and a stable monocular depth prior.

The overall framework is shown in Fig. 1. A frozen foundation depth estimator M_{FFD}, instantiated with Depth Anything V2 [14], first predicts a prior depth map $d_p = M_{\mathrm{FFD}}(X)$. The input image, prior depth, and ground-truth

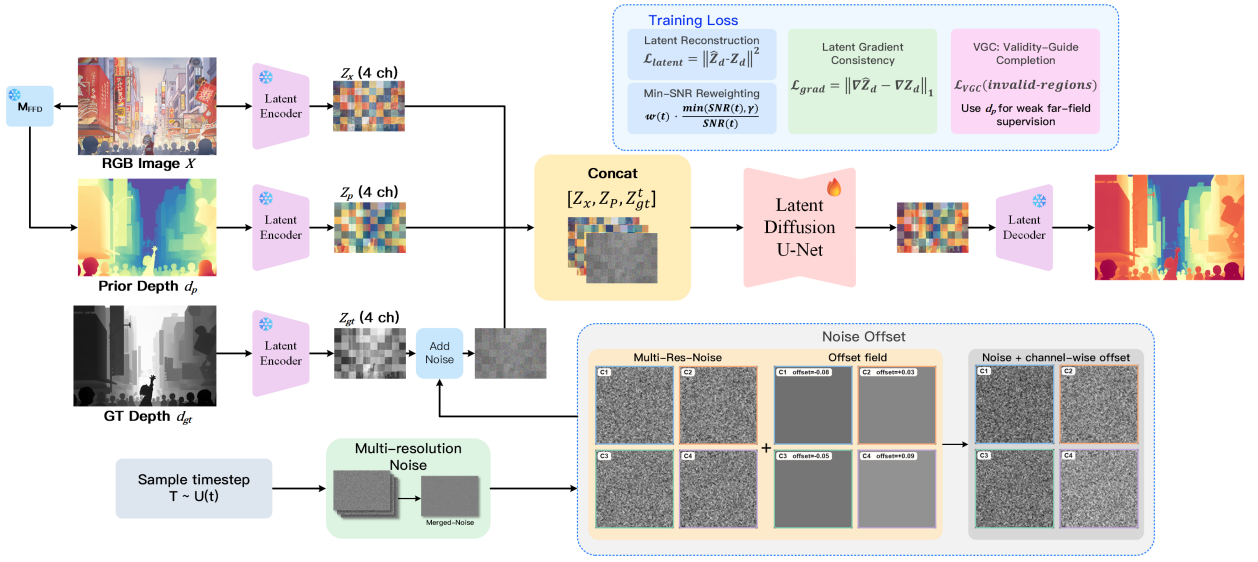

Figure 1: Overview of ApDepth-G. A frozen foundation depth model predicts the prior depth d_p, and a frozen VAE maps the RGB image, prior depth, and ground-truth depth to four-channel latent representations. The noisy depth latent is concatenated with the RGB and prior latents and progressively refined by the trainable latent diffusion U-Net. During training, annealed multi-resolution noise is augmented with a channel-wise offset, and the model is optimized using a Min-SNR-reweighted latent objective, latent gradient consistency, and validity-guided completion in regions without valid depth supervision. Snowflake and flame symbols denote frozen and trainable modules, respectively.

depth are encoded by the frozen variational autoencoder (VAE) of Stable Diffusion v2 [18]. At a randomly sampled diffusion timestep, the noisy depth latent is concatenated with the RGB and prior latents and processed by a trainable denoising U-Net. Training is stabilized by multi-resolution noise and channel-wise offset noise, while the optimization objective combines Min-SNR reweighting, latent gradient consistency, and validity-guided completion (VGC). The latter supplies weak, prior-based supervision only where the original depth annotation is invalid, which commonly includes sky and out-of-range regions.

3.2. Depth-Prior-Guided Latent Diffusion

3.2.1. Latent representation

We follow the affine-invariant depth representation used by latent diffusion depth estimators. For each ground-truth depth map d_{gt}, the 2nd and 98th percentiles of its valid values, denoted by d_2 and d_{98}, define the normalized depth

$$\tilde{d}_{gt} = 2\left(\frac{d_{gt} - d_2}{d_{98} - d_2} - \frac{1}{2}\right), \tag{1}$$

which is clipped to $[-1, 1]$. Invalid values are excluded from the training objectives through a binary validity mask M. Since the pretrained VAE expects a three-channel input, a depth map is replicated across three channels before encoding. With the frozen encoder $\mathcal{E}$ and decoder $\mathcal{D}$, we obtain

$$z_x = \mathcal{E}(X), \qquad z_p = \mathcal{E}(d_p), \qquad z_d = \mathcal{E}(\tilde{d}_{gt}), \tag{2}$$

where $z_x, z_p, z_d \in \mathbb{R}^{4\times h\times w}$. The same VAE is used for all three branches, keeping their feature resolutions and latent statistics compatible.

3.2.2. Prior-conditioned denoiser

At training time, a timestep t is sampled uniformly from the diffusion horizon $\{0, \dots, T-1\}$. Let $\bar{\alpha}_t$ denote the cumulative noise schedule and let ϵ_{aug} be the augmented noise described in Sec. 3.3. The forward process constructs

$$z_{d,t} = \sqrt{\bar{\alpha}_t}\, z_d + \sqrt{1-\bar{\alpha}_t}\, \epsilon_{\text{aug}}. \tag{3}$$

The denoiser receives the channel-wise concatenation

$$c_t = \mathrm{Cat}\left[z_x, z_p, z_{d,t}\right] \in \mathbb{R}^{12\times h\times w}. \tag{4}$$

Thus, the RGB latent provides appearance and semantic evidence, the prior latent supplies a stable global geometric reference, and the noisy depth latent represents the state to be refined. No sky segmentation map or additional semantic channel is required.

The input convolution of the pretrained U-Net originally accepts four channels. We expand it to 12 channels by repeating its weight tensor three times and dividing the resulting weights by three. This preserves the scale of the initial activations while allowing all three latent sources to contribute from the start of fine-tuning. The VAE and text encoder remain frozen, an empty text embedding is used, and only the denoising U-Net is optimized. This design retains the generative and structural priors of Stable Diffusion while limiting the number of trainable components.

3.3. Robust Noise Augmentation

Standard independent Gaussian noise is dominated by high-frequency variations and provides limited coverage of slowly varying perturbations. Such perturbations are nevertheless important for depth, where global layout, distant regions, and textureless sky occupy large spatial extents. We therefore combine annealed multi-resolution noise with a channel-wise offset. Both components are used only during training.

3.3.1. Annealed multi-resolution noise

For each latent sample, we draw a full-resolution Gaussian field $\epsilon^{(0)}$ and a set of independent Gaussian fields $\{\epsilon^{(i)}\}_{i=0}^{K-1}$ at progressively lower spatial resolutions. Each low-resolution field is bilinearly upsampled by $\mathcal{U}_i$ to $h \times w$ and superimposed on the base noise:

$$\tilde{\epsilon}_{\mathrm{MR}} = \epsilon^{(0)} + \sum_{i=0}^{K-1} \lambda_t^{\,i}\, \mathcal{U}_i\big(\epsilon^{(i)}\big), \qquad \epsilon_{\mathrm{MR}} = \frac{\tilde{\epsilon}_{\mathrm{MR}}}{\mathrm{Std}(\tilde{\epsilon}_{\mathrm{MR}})}. \tag{5}$$

Following the implementation, the spatial downscale factor at each level is randomly selected between two and four, at most ten levels are accumulated, and construction stops when either spatial dimension reaches one. The base strength is $\lambda_0 = 0.9$. We anneal the effective strength according to the sampled diffusion timestep,

$$\lambda_t = \lambda_0 \frac{t}{T}. \tag{6}$$

Consequently, coarse noise components are emphasized at highly corrupted timesteps and gradually vanish as t approaches zero. Normalizing the composite noise to approximately unit variance prevents the number of scales from unintentionally changing the overall diffusion magnitude. This construction follows the multi-resolution training principle used in Marigold [12], while serving here as part of the robust geometric augmentation.

3.3.2. Channel-wise offset noise

Multi-resolution noise expands the spatial frequency coverage, but its realization can still remain close to zero mean in every latent channel. We additionally sample a channel-wise offset

$$\delta \sim \mathcal{N}\big(0, \sigma_{\mathrm{off}}^2 I\big), \qquad \delta \in \mathbb{R}^{4\times 1\times 1}, \qquad \sigma_{\mathrm{off}} = 0.1, \tag{7}$$

and broadcast it over the spatial dimensions. The final training noise is

$$\epsilon_{\mathrm{aug}}(c,u,v) = \epsilon_{\mathrm{MR}}(c,u,v) + \delta_c. \tag{8}$$

Unlike another spatial noise map, δ_c is constant over all positions of channel c. It therefore perturbs channel means without changing the local noise texture. This exposes the denoiser to low-frequency and mean-shift variations that are difficult to represent using zero-mean pixel-wise noise alone, reducing sensitivity to global latent bias in smooth or weakly textured regions.

3.4. Geometry-Aware Training Objectives

The training objective is defined in the latent space and is evaluated only on reliable support unless otherwise stated. We conservatively downsample the image-space validity mask: a latent cell is considered valid only if all pixels in its corresponding 8×8 image block are valid. The resulting mask is repeated across the four latent channels and is denoted by M_z.

3.4.1. Clean latent reconstruction with Min-SNR reweighting

Stable Diffusion v2 uses the velocity parameterization. For the augmented noise in Eq. (8), the velocity target is

$$v_t = \sqrt{\bar{\alpha}_t}\,\epsilon_{\text{aug}} - \sqrt{1-\bar{\alpha}_t}\,z_d. \tag{9}$$

The U-Net predicts $\hat{v}_t = f_\theta(c_t, t)$, from which a clean depth latent is reconstructed as

$$\hat{z}_d = \sqrt{\bar{\alpha}_t}\,z_{d,t} - \sqrt{1-\bar{\alpha}_t}\,\hat{v}_t. \tag{10}$$

Different diffusion timesteps have substantially different noise levels and can produce competing optimization magnitudes. We use Min-SNR weighting [**?**] to cap the influence of high-SNR timesteps. Specifically,

$$\text{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}, \qquad w(t) = \frac{\min(\text{SNR}(t), \gamma)}{\text{SNR}(t)}, \qquad \gamma = 5. \tag{11}$$

The primary latent objective is then

$$\mathcal{L}_{\text{latent}} = \frac{1}{\|M_z\|_1}\left\| M_z \odot \sqrt{w(t)}\,\left(\hat{v}_t - v_t\right)\right\|_2^2. \tag{12}$$

Although Eq. (12) is evaluated in velocity space, Eq. (10) makes the corresponding clean latent explicit for the geometric regularizers below.

3.4.2. Latent gradient consistency

The point-wise diffusion objective does not explicitly preserve local depth transitions. To maintain geometric discontinuities while discouraging irregular latent structures, we match the magnitudes of first-order horizontal and vertical differences between $\hat{z}_d$ and z_d. Let ∇_x and ∇_y denote forward finite differences, and let M_x and M_y be the validity masks cropped to the corresponding difference supports. We define

$$\begin{aligned}\mathcal{L}_{\text{grad}} = &\frac{\left\| M_x \odot \left(|\nabla_x \hat{z}_d| - |\nabla_x z_d|\right)\right\|_1}{\|M_x\|_1} \\ &+ \frac{\left\| M_y \odot \left(|\nabla_y \hat{z}_d| - |\nabla_y z_d|\right)\right\|_1}{\|M_y\|_1}.\end{aligned} \tag{13}$$

Matching gradient magnitudes rather than only latent values encourages the reconstructed representation to retain meaningful boundaries. Because the term is restricted to valid support, it does not force uncertain sky pixels to imitate arbitrary invalid ground truth.

3.4.3. Validity-guided completion

Validity masks are necessary for sparse, out-of-range, or otherwise unreliable sensor measurements. However, masking the reconstruction loss also leaves all invalid regions unconstrained. In outdoor data, large invalid components frequently correspond to the sky or distant surfaces, allowing their latent responses to drift or develop pseudo-texture. We address this issue with VGC, a training-only regularizer that uses the detached prior latent z_p as a weak target on invalid support. Importantly, VGC uses only the existing depth validity mask and does not require a sky segmentation model or an additional U-Net input channel.

Let $U_z = 1 - M_z$ be the invalid latent mask and let

$$r_b = \frac{\|U_z^{(b)}\|_1}{4hw}, \qquad g_b = \mathbf{1}\left[r_b \geq \rho\right], \qquad \rho = 0.08, \tag{14}$$

denote the invalid ratio and sample gate. The gate prevents small missing-depth holes, which are common in otherwise valid indoor samples, from activating the completion objective. We write $\bar{z}_U^{(b,c)}$ for the spatial mean of latent channel c over the gated invalid region. The region-level anchor is

$$\mathcal{L}_{\text{anchor}} = \frac{1}{C\sum_b g_b}\sum_{b,c} g_b\left(\bar{\hat{z}}_{d,U}^{(b,c)} - \bar{z}_{p,U}^{(b,c)}\right)^2. \tag{15}$$

Only the regional mean is aligned. This choice supplies a coarse far-field reference without copying high-frequency artifacts from the foundation-model prior into the predicted sky.

We further suppress spurious structures strictly inside invalid regions. Defining U_x and U_y as masks whose two adjacent pixels are both invalid, the smoothness term is

$$\mathcal{L}_{\text{smooth}} = \frac{1}{2}\left(\frac{\|U_x \odot \nabla_x \hat{z}_d\|_1}{\|U_x\|_1} + \frac{\|U_y \odot \nabla_y \hat{z}_d\|_1}{\|U_y\|_1}\right). \tag{16}$$

Since finite differences crossing the validity boundary are excluded, this term smooths sky-like interiors without blurring valid object silhouettes or the horizon boundary. The VGC objective is

$$\mathcal{L}_{\text{VGC}} = \lambda_a \mathcal{L}_{\text{anchor}} + \lambda_s \mathcal{L}_{\text{smooth}}, \qquad \lambda_a = 0.02, \quad \lambda_s = 0.005. \tag{17}$$

3.5. Overall Objective and Multi-Step Refinement

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{latent}} + \lambda_g \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{VGC}}, \qquad \lambda_g = 0.1. \tag{18}$$

The latent and gradient terms learn the conditional denoising trajectory from reliable depth supervision, whereas VGC constrains regions omitted by that supervision. Together with robust noise augmentation, the objective exposes the model to perturbations ranging from fine stochastic texture to global channel shifts while retaining explicit geometric guidance.

At inference time, the depth state is initialized with standard Gaussian noise $z_{d,T} \sim \mathcal{N}(0, I)$. The RGB latent z_x and prior latent z_p are computed once and concatenated with the current depth latent at every reverse step. A DDIM scheduler [22] then updates $z_{d,t}$ to $z_{d,t-1}$ for 50 steps. Multi-resolution noise, offset noise, and all auxiliary losses are disabled during inference. Finally, the refined latent $z_{d,0}$ is decoded by $\mathcal{D}$; the three decoded channels are averaged and mapped from $[-1, 1]$ to $[0, 1]$ to obtain the affine-invariant prediction $\hat{d}$. In this way, the pretrained depth prior establishes the global scene layout, while the multi-step trajectory progressively corrects inconsistent local geometry under the image condition.

4. Experiments

5. Conclusion

A. Acknowledgments

This work was supported by National Natural Science Foundation of China (62073227) and Liaoning Provincial Science andTechnologyDepartment Foundation(2023JH2/101300212).

References

[1] Bai, Y., Huang, Q., 2024. Fiffdepth: Feed-forward transformation of diffusion-based generators for detailed depth estimation. arXiv preprint arXiv:2412.00671 URL: https://arxiv.org/abs/2412.00671.

[2] Bhat, S.F., Alhashim, I., Wonka, P., 2021. Adabins: Depth estimation using adaptive bins, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4009–4018.

[3] Duan, Y., Guo, X., Zhu, Z., 2023. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. arXiv preprint arXiv:2303.05021 URL: https://arxiv.org/abs/2303.05021.

[4] Duan, Y., Guo, X., Zhu, Z., 2024. Diffusiondepth: Diffusion denoising approach for monocular depth estimation, in: European Conference on Computer Vision, Springer. pp. 432–449.

[5] Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27.

[6] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2002–2011.

[7] Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X., 2025. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image, in: European Conference on Computer Vision, Springer. pp. 241–258.

[8] Garcia, G.M., Abou Zeid, K., Schmidt, C., De Geus, D., Hermans, A., Leibe, B., 2025. Fine-tuning image-conditional diffusion models is easier than you think, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 753–762.

[9] Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B., 2024. Depthfm: Fast monocular depth estimation with flow matching. arXiv:2403.13788.

[10] He, J., Li, H., Sheng, M., Chen, Y.C., 2026. Lotus-2: Advancing geometric dense prediction with powerful image generative model. URL: https://arxiv.org/abs/2512.01030, arXiv:2512.01030.

[11] He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.C., 2024. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124 .

[12] Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K., 2025. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence URL: https://arxiv.org/abs/2505.09358, arXiv:2505.09358.

[13] Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H., 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 .

[14] Li, X., Deng, R., Fan, Y., Chen, P., Chen, S., Liu, Z., Han, L., Zhu, S., Sun, H., Lu, Y., Li, Q., 2024. Depth anything v2. arXiv preprint arXiv:2404.14442 URL: https://arxiv.org/abs/2404.14442.

[15] Pham, D.H., Do, T., Nguyen, P., Hua, B.S., Nguyen, K., Nguyen, R., 2025. Sharpdepth: Sharpening metric depth predictions using diffusion distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17060–17069. URL: https://arxiv.org/abs/2411.18229.

[16] Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 12179–12188.

[17] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence 44, 1623–1637.

[18] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

[19] Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J., 2024. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. Advances in Neural Information Processing Systems .

[20] Saxena, S., Kar, A., Norouzi, M., Fleet, D.J., 2023. Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 .

[21] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems 35, 25278–25294.

[22] Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 .

[23] Song, Z., Wang, Z., Li, B., Zhang, H., Zhu, R., Liu, L., Jiang, P.T., Zhang, T., 2025. Depthmaster: Taming diffusion models for monocular depth estimation. arXiv preprint arXiv:2501.02576 .

[24] Wang, J., Yuan, S., Lei, M., Chen, Y., 2026. Apdepth: Aiming for precise monocular depth estimation based on diffusion models. Preprint Manuscript.

[25] Wang, R., et al., 2024. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. arXiv preprint arXiv:2410.19115 URL: https://arxiv.org/abs/2410.19115.

[26] Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C., 2024a. What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 .

[27] Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C., 2024b. What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 URL: https://arxiv.org/abs/2403.06090.

[28] Xu, G., Lin, H., Luo, H., Wang, X., Yao, J., Zhu, L., Pu, Y., Chi, C., Sun, H., Wang, B., Chen, G., Ye, H., Peng, S., Yang, X., 2025. Pixel-perfect depth with semantics-prompted diffusion transformers. Advances in Neural Information Processing Systems URL: https://arxiv.org/abs/2510.07316, arXiv:2510.07316. neurIPS 2025.

[29] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything: Unleashing the power of large-scale unlabeled data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition .

[30] Yu, R., Jia, T., Wang, H., Li, X., Yang, X., Zhang, Z., Liu, C., 2026. Cdpr: Cross-modal diffusion with polarization for reliable monocular depth estimation. URL: https://arxiv.org/abs/2604.11097, arXiv:2604.11097.

[31] Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P., 2022. New crfs: Neural window fully-connected crfs for monocular depth estimation. arXiv preprint arXiv:2203.01502 .

[32] Zeng, Z., Ni, J., Wang, D., Rim, P., Chung, Y., Yang, F., Hong, B.W., Wong, A., 2024. Iris: Integrating language into diffusion-based monocular depth estimation. arXiv preprint arXiv:2411.16750 URL: https://arxiv.org/abs/2411.16750.

[33] Zhan, M., Zhang, L., Chu, X., Wang, B., 2025. Vistadepth: Frequency modulation with bias reweighting for enhanced long-range depth estimation. arXiv preprint arXiv:2504.15095 URL: https://arxiv.org/abs/2504.15095.