D3SS: DEPTH RECOVERY COVERING DIVERSE 3D SCENE SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing depth perception methods overemphasize 2D features of depth maps, introducing diverse 3D distortions caused by the scale problem, creating the demand for depth recovery in 3D scenes. Conventional depth recovery approaches only simulate raw depth at fixed scales in limited scenes, causing a significant simulation-to-reality gap that hinders the generalization of depth recovery models. To address the problem, we propose a diffusion-based raw depth generation pipeline for simulating diverse scene-scale distortions in both global and local patterns. To enhance the generalization capability in diverse RGB conditions, we first integrate the frozen depth foundation model (DFM), which is trained on large-scale RGB-D data, into the proposed framework. We further propose loss functions based on different views of 3D scenes to refine the 3D structure. On the one hand, our loss functions enforce the model to learn 3D features; on the other hand, our losses mitigate overfitting of the 3D distortions caused by DFM. Extensive studies on the recovery of different conditions in multiple datasets demonstrate the effectiveness and progressiveness of the proposed method.

1 INTRODUCTION

Due to advances in network architectures and RGB-D datasets, depth perception based on RGB images has been widely studied in recent years, including depth estimation Ranftl et al. (2022); Yin et al. (2023); Yang et al. (2024); Bochkovskiy et al. (2025); Piccinelli et al. (2026), depth completionPark et al. (2020); Wang et al. (2025a), and depth recoveryHe et al. (2021); Kim et al. (2021); Miangoleh et al. (2021); Wang et al. (2024b); Lin et al. (2025b). However, most existing methods treat depth maps solely as 2D images, overemphasizing 2D features (i.e., image gradients Ranftl et al. (2022); Wang et al. (2024c); Kang et al. (2025), the image resolution Miangoleh et al. (2021); Dai et al. (2023); Lin et al. (2025b)) while neglecting the 3D attributes of depth maps. As illustrated in Fig. 1, state-of-the-art (SOTA) depth estimation methods still demonstrate poor 3D performance because of the famous scale-ambiguity problem Eigen et al. (2014); Ranftl et al. (2022); Piccinelli et al. (2026). For the same reason, classical depth recovery approaches based on 2D depth images learning also struggle to recover these 3D distortions.

Recovering depth maps addressing 3D features is a natural idea Wang et al. (2021); Wen et al. (2025); Wang et al. (2025b). However, existing methods exhibit limited generalization capability for diverse 3D distortions. One fundamental bottleneck lies that a learning-based depth recovery model demands the raw depth and ground-truth (GT) depth at the same time. While public RGB-D datasets like Middlebury Scharstein et al. (2014), NYUv2 Silberman et al. (2012), and KITTI Geiger et al. (2012) provide GT depth, raw depth maps are typically simulated since few datasets capture both concurrently. Distortions in raw depth maps can be regarded as *scale distortions* that span a global-to-local spectrum: global scale ambiguity shifts the overall depth distribution Ranftl et al. (2022), while local artifacts affect fine-grained structures Hirschmuller (2007); He et al. (2021). Consequently, there are infinite conditions of scale distortions in depth maps, and it is impossible to simulate covering all conditions during training. Naively simulated raw depth leads to the fixed scene scale of depth maps, causing gaps on unseen scene scales for diverse input raw depth maps.

On the other hand, the generalization capability is strong in the foundation model within the task of depth estimation, benefiting from a large amount of RGB-D data and well-designed training strategies Yang et al. (2024); Hu et al. (2025); Lin et al. (2025a). However, there is still no foundation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

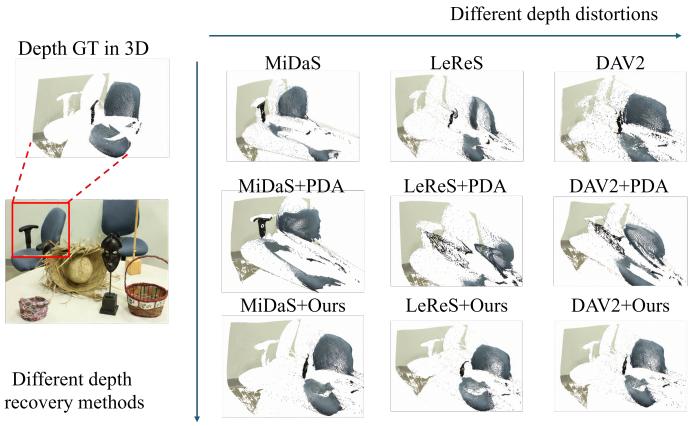


Figure 1: Recent depth estimation methods (i.e., MiDaSRanftl et al. (2022), LeReSYin et al. (2023), DAV2Yang et al. (2024)) show limited performance on 3D scenes, while SOTA depth recovery methods (i.e. PDALin et al. (2025b)) also lack of capability in recovering diverse 3D distortions.

model in task of depth recovery because of the lack of suffcient paired RGB-raw depth-GT depth data. It is an alternative way to leverage DFM to gain strong generalization capability Yu et al. (2024); Lin et al. (2025b); Xu et al. (2025); Zhou et al. (2026); however, the 3D distortions are still not fully solved even in DFM. To our knowledge, this work is the first to systematically investigate generalization limitations for 3D depth recovery.

In the proposed method, we first extensively utilize a diffusion model to simulate diverse scene scales of raw depth in the real world. With the basic principle and the observation of different effects caused by diffusion, we control the diffusion steps for simulating diverse scene scales in global patterns, and different initial stochastic noises are set for simulating scene scales in local patterns. Furthermore, an interpolation method is utilized to simulate diverse intermediate conditions. In order to improve the generalization capability on 3D scenes, we integrate the frozen DFM as part of our model for strong generalization on unseen scenes. Furthermore, a loss function based on the 3D consistency across different views (view consistency loss) is designed to force the model to learn the 3D features of depth maps. Since DFMs still exhibit limited sensitivity to scale distortions in 3D scenes, an exclusive constraint is further developed in the view consistency loss to enhance 3D generalization.

Our main contributions are highlighted as follows:

- We propose a diffusion-based raw depth simulation pipeline that diversifies scene-scale distortions of raw depth maps in global and local patterns.
- We propose a method to integrate DFM into our model for strong generalization on unseen scenes. Based on DFM, we propose a view consistency loss, not only forcing the model to learn 3D features, but also addressing the 3D distortions caused by DFM.
- Extensive results on the 3D recovery of different models in multiple datasets demonstrate the effectiveness and progressiveness of the proposed method.

2 Related Works

2.1 Depth Recovery Focusing on 2D Features

Most depth recovery approaches process depth maps as conventional 2D images within neural networks. This paradigm overemphasizes the 2D features of the depth map, such as the image gradient Zhu et al. (2020); Xu et al. (2022); Yuan et al. (2023); Wang et al. (2024c); Kang et al. (2025) and the image resolution Dai et al. (2023); Tang et al. (2021); Miangoleh et al. (2021); Lin et al. (2025b). Though 2D features contribute to depth maps, the underlying 3D structure is still not adequately captured. An alternative way is to consider the structure alignments between depth and RGB image Liu et al. (2021); Wang et al. (2023b), primarily enhancing the shape of the object in 3D detection

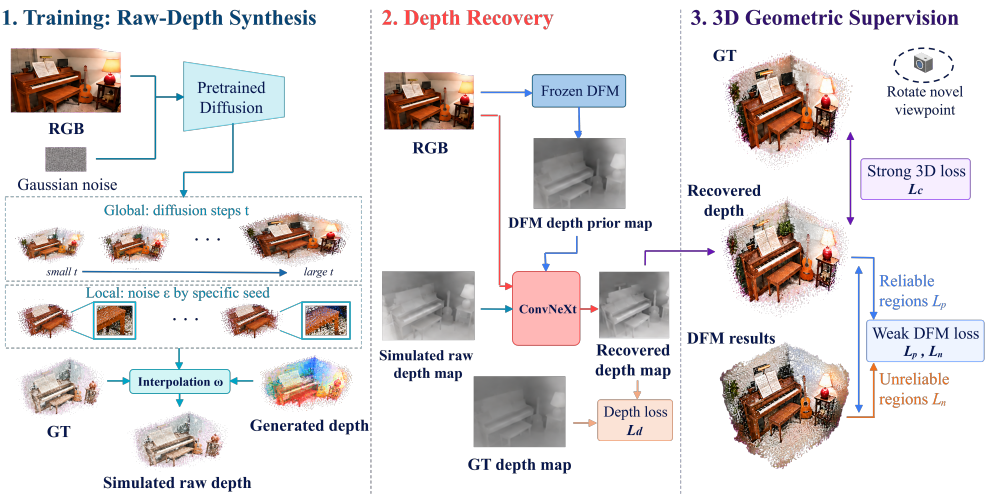


Figure 2: The framework of the proposed method. A pretrained diffusion model first simulates diverse raw depth from RGB and noise under global and local 3D distortions. A ConvNeXt then recovers depth from the RGB image, simulated raw depth, and a frozen DFM prior, supervised by $\mathcal{L}_d$; and novel-view 3D consistency is enforced with a strong loss $\mathcal{L}_c$ and a weak DFM loss split into reliable $\mathcal{L}_p$ and unreliable $\mathcal{L}_n$ terms to avoid overfitting 3D distortions in DFM.

Wang et al. (2019); Li et al. (2023). However, such alignment remains predominantly constrained near edges, leaving the interior structure not adequately addressed.

2.2 Depth Recovery Focusing on 3D Features

Recovering depth through explicit 3D feature modeling represents an established research direction. Several approaches leverage surface normal priors to regularize the structure of depth maps Wang et al. (2021); Wen et al. (2025); Wang et al. (2025b; 2026); Hu et al. (2025); Bochkovskiy et al. (2025), yet this introduces cascaded errors due to the ill-posed nature of normal estimation. There are some alternative strategies for recovering depth in 3D. LeReS Yin et al. (2023) proposes to recover the 3D scene shape by refining the focal length and shift of depth map. VitualNormal Yin et al. (2021) utilizes non-planar geometric constraints in 3D space. DeCoTr Shi et al. (2024) proposes hybrid 2D-3D attentions to refine the 3D structure of depth. Critically, these methods insufficiently model the diverse real-world 3D distortions, overfitting the fixed depth scale of the simulation in finite scenes. This architectural limitation inherently constrains their generalization capability for 3D depth recovery in the real world.

3 Method

Given an RGB image $\mathbf{C}$ and its corresponding raw depth map $\mathbf{D_r}$ as input, our method aims to recover the refined depth map $\mathbf{D}$. The whole framework is illustrated in Fig. 2. To address the diverse scene-scale distortions in the real world, we first simulate $\mathbf{D_r}$ based on a pretrained diffusion model Ke et al. (2024) conditioned on $\mathbf{C}$ and randomly sampled Gaussian noise ε. The proposed raw depth simulation pipeline effectively diversifies the scene-scale of raw depth in both global and local patterns. To enhance the generalization capability on unseen scenes, the DFM is integrated into our model. However, acknowledging that DFM outputs $\mathbf{D_f}$ may still exhibit 3D inaccuracies, we further design novel view consistency loss functions that not only force the model to learn 3D features, but also address the 3D distortions caused by the DFM.

3.1 Diffusion-based Raw Depth Simulation

Learning-based depth recovery model rigidly demands the raw depth $\mathbf{D_r}$ and GT depth $\mathbf{D}^*$ at the same time. While public RGB-D datasets like NYUv2 Silberman et al. (2012) and KITTI Geiger

et al. (2012) provide GT depth, they lack authentic and sufficient raw depth counterparts due to sensor limitations. This necessitates synthetic $\mathbf{D_r}$ simulation. To avoid overfitting the fixed scale of the raw depth, we employ a pretrained diffusion model Ke et al. (2024) for generating plausible raw depth in diverse scales. With the basic principle and the observation of effects in the diffusion model, we systematically address scene-scale distortions in both global and local patterns.

We leverage the reverse process of a latent diffusion model Rombach et al. (2022); Ke et al. (2024) to generate diverse distorted raw depth maps $\mathbf{D}_r$. The reverse diffusion step is formulated as:

$$\begin{aligned}\mathbf{Z}_{\mathbf{t-1}}^{(\mathbf{D}_r)} = &\sqrt{\bar{\boldsymbol{\alpha}}_{\mathbf{t-1}}}\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{Z}_{\mathbf{t}}^{(\mathbf{D}_r)}, \mathbf{Z}^{(\mathbf{C})}, \mathbf{t}\right) \\ &+\sqrt{1-\bar{\boldsymbol{\alpha}}_{\mathbf{t-1}}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{Z}_{\mathbf{t}}^{(\mathbf{D}_r)}, \mathbf{Z}^{(\mathbf{C})}, \mathbf{t}\right),\end{aligned} \tag{1}$$

where $\mathbf{Z}^{(\mathbf{x})} = \mathcal{E}(\mathbf{x})$ denotes the latent encoding of input $\mathbf{x}$ via a variational autoencoder (VAE) Rombach et al. (2022), $\mathbf{t}$ indexes the diffusion step, and $\bar{\boldsymbol{\alpha}}_{\mathbf{t}} = \prod_{\mathbf{i}=1}^{\mathbf{t}} \boldsymbol{\alpha}_{\mathbf{i}}$ represents the cumulative noise schedule coefficient. The denoising network $\boldsymbol{\theta}$ simultaneously predicts both the base state $\mathbf{f}_{\boldsymbol{\theta}}$ and noise component $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.

The synthesis process initiates from Gaussian noise $\mathbf{Z}_0^{(\mathbf{D_r})} \approx \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, iteratively applying Eq. 9 for $\mathbf{t}$ steps to obtain the denoised latent $\mathbf{Z}_{-\mathbf{t}}^{(\mathbf{D_r})}$. The final raw depth is reconstructed through the VAE decoder: $\mathbf{D_r} = \mathcal{D}(\mathbf{Z}_{-\mathbf{t}}^{(\mathbf{D_r})})$.

This stochastic formulation enables diverse distortion synthesis: varying the total diffusion steps $\mathbf{t} \sim \mathcal{U}(0, T)$ generates global distortions, while sampling different noise vectors $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ produces heterogeneous local artifacts, as illustrated in Fig. 2. For ease of understanding, the main text only presents macro phenomena and generation schemes. Specific theoretical explanations can be found in Appendix A.1.

To further expand distortion diversity, we introduce stochastic interpolation between the generated and GT depth:

$$\hat{\mathbf{D}}_{\mathbf{r}} = \boldsymbol{\omega} sf[\mathcal{D}(\mathbf{Z}_{-\mathbf{t}}^{(\mathbf{D_r})})] + (1-\boldsymbol{\omega})\mathbf{D}^*, \tag{2}$$

where $sf[\cdot]$ denotes the scale and shift is aligned with ground truth (GT) depth Ranftl et al. (2022); Yin et al. (2023), $\mathbf{D}^*$ denotes GT depth, $\boldsymbol{\omega} \sim \mathcal{U}(0, 1)$ is a uniformly sampled mixing coefficient, and $\hat{\mathbf{D}}_{\mathbf{r}}$ represents the final simulated raw depth. This combination generates intermediate distortions not captured by either source alone, significantly diversifying the plausible scene-scale distortions in raw depth maps.

As a result, the diffusion-based method effectively simulate diverse 3D distortions in raw depth maps in both global scale and local scale, thanks to random Gaussian noise and different diffusion steps.

3.2 Improving the Generalization on 3D Scenes

While our diffusion-based pipeline diversifies scene-scale distortions in raw depth maps, a simulation-to-reality domain gap still persists, limiting generalization capability. For this problem, we first integrate the frozen DFM Yang et al. (2024) as a feature extractor, leveraging its pretrained depth priors for robust scene understanding. Second, we introduce novel loss functions based on the view consistency, which not only explicitly enforces the model to learn 3D features but is further designed to correct residual 3D distortions in DFM outputs. The co-design well improves the generalization capability on 3D scenes.

3.2.1 Integrating the frozen DFM.

Taking advantage of large-scale RGB-D datasets and carefully designed network architecture, DFM Ranftl et al. (2022); Yin et al. (2023); Yang et al. (2024); Lin et al. (2025a) has demonstrated exceptional generalization to real-world scenes. We harness this capability by integrating the frozen output of DAV2 Yang et al. (2024) as a prior within our framework. The forward process of our recovery model is formulated as follows:

$$\mathbf{D} = \mathcal{R}_{\varphi}\left(\mathbf{C}, \hat{\mathbf{D}}_{\mathbf{r}}, \mathbf{D}_{\mathbf{f}}\right), \tag{3}$$

where $\mathcal{R}_\varphi$ denotes our recovery network with parameters φ, which adopts the common-used ConvNeXt Yang et al. (2024), $\mathbf{D_f}$ represents DAV2's depth estimation, $\hat{\mathbf{D}}_\mathbf{r}$ is the simulated raw depth from Eq. 2, and $\mathbf{C}$ is the RGB image. These three components are input into the ConvNeXt in a simple form of channel concating for joint feature extraction. This integration provides a robust depth perception while preserving flexibility for distortion correction.

3.2.2 3D CONSTRAINT VIA DIFFERENT VIEWS.

While prior works leverage DFM as auxiliary components Lin et al. (2025b); Yu et al. (2024), however, the persistent 3D distortions in DFM are still unaddressed. In addition to the standard depth supervision, we propose a novel 3D constraint to explicitly guide the model in capturing the 3D-aware characteristics of depth maps.

We adopt the commonly-used pinhole camera Yin et al. (2021; 2023) to unproject 2D depth maps $\mathbf{D} = \{\mathbf{d_1}, \mathbf{d_2}, ...\}$ into 3D point clouds $\mathbf{P} = [\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{R}, \mathbf{G}, \mathbf{B}] = \{(\mathbf{x_1}, \mathbf{y_1}, \mathbf{z_1}, \mathbf{r_1}, \mathbf{g_1}, \mathbf{b_1}), (\mathbf{x_2}, \mathbf{y_2}, \mathbf{z_2}, \mathbf{r_2}, \mathbf{g_2}, \mathbf{b_2}), ...\}$. The process is defined as follows:

$$\begin{cases} \mathbf{x_i} = (\mathbf{u_i} - \mathbf{u_0})\mathbf{d_i}/\mathbf{f}, \\ \mathbf{y_i} = (\mathbf{v_i} - \mathbf{v_0})\mathbf{d_i}/\mathbf{f}, \\ \mathbf{z_i} = \mathbf{d_i}. \end{cases} \tag{4}$$

where $(\mathbf{u_i}, \mathbf{v_i})$ denotes the coordinate of the pixel, $(\mathbf{u_0}, \mathbf{v_0})$ denotes the optical center of the camera, $\mathbf{f}$ denotes the focal length.

Our 3D constraint is motivated by a key observation: 3D distortions may be imperceptible in the original camera view, they manifest clearly in alternative perspectives (e.g., top-down views). We further this insight into a differentiable 3D loss. For the output $\mathbf{D}$ and GT depth $\mathbf{D}^*$, they are first unprojected into 3D space as $\mathbf{P}$ and $\mathbf{P}^*$, respectively. Then, the new views $\tilde{\mathbf{V}} = \{\tilde{\mathbf{v}}_\mathbf{1}, \tilde{\mathbf{v}}_\mathbf{2}, ...\}$ and $\mathbf{V}^* = \{\tilde{\mathbf{v}}^*_\mathbf{1}, \tilde{\mathbf{v}}^*_\mathbf{2}, ...\}$ of $\mathbf{P}$ and $\mathbf{P}^*$ are generated by random extrinsic parameters of the camera, respectively. The process of generating a new view is defined in Algorithm 1. The extrinsic parameters containing the rotation $\mathbf{R}$ and the translation $\mathbf{T}$ are generated by the predefined object position and eye position, which ensures rationality of the view for $\mathbf{P}$. Then the new camera coordinates can be generated by $\mathbf{R}$, $\mathbf{T}$, and $\mathbf{P}$. With the predefined intrinsic parameters $\mathbf{K}$, a new view is generated by depth buffers Jia & Evans (2014). Note that a differentiable sorting scheme is adopted, this enables the gradient of the network to be returned normally and ensures that occluded areas do not contribute to the gradient in loss optimization.

As a result, the 3D constraint between $\mathbf{D}$ and $\mathbf{D}^*$ is defined as follows:

$$\mathcal{L}_c = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\mathcal{V}(\mathbf{D}, \mathbf{C}) - \mathcal{V}(\mathbf{D}^*, \mathbf{C})\|_1, \tag{5}$$

where $|\Omega|$ denotes the sum of valid pixels in $\mathbf{D}^*$.

3.3 LOSS FUNCTIONS

For improving the generalization ability on 3D scenes, the 3D constraint is divided into 2 parts: conventional constraint between $\mathbf{D}$ and $\mathbf{D}^*$ ($\mathcal{L}_c$ in equation 5), and specific weak constraint between $\mathbf{D}$ and metric $\mathbf{D_f}$, encouraging the model to learn from DFM but not overfitting the existing 3D distortions. The weak constraint is simple but effective, dividing into 2 parts: The first part encourages learning from the consistency between $\mathbf{D_f}$ and $\mathbf{D}^*$, defined as follows:

$$\mathcal{L}_p = \frac{1}{|\Phi|} \sum_{p \in \Phi} \|\mathcal{V}(\mathbf{D}, \mathbf{C}) - \mathcal{V}(\mathbf{D_f} \cdot \mathbf{M}, \mathbf{C})\|_1, \tag{6}$$

where $|\Phi|$ denotes the number of valid pixels in $\mathbf{D_f} \cdot \mathbf{M}$, $\mathbf{M} = \{\mathbf{m_i} = 1 | \mathbf{D_{f,i}} - \mathbf{D^*_i}| / \mathbf{D^*_i} < \boldsymbol{\tau}, else\ \mathbf{m_i} = 0\}$ denotes the mask in $\mathbf{D_f}$, The second part punishes the inconsistency between D_f and $\mathbf{D}^*$, defined as follows:

$$\mathcal{L}_n = -\frac{1}{|\Theta|} \sum_{p \in \Theta} \|\mathcal{V}(\mathbf{D}, \mathbf{C}) - \mathcal{V}(\mathbf{D_f} \cdot (1 - \mathbf{M}), \mathbf{C})\|_1, \tag{7}$$

Algorithm 1 Generating a random new view of the point cloud

Input: The point cloud $\mathbf{P} = [\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{R}, \mathbf{G}, \mathbf{B}]$
Parameter: the focal length $\mathbf{f}$, the size $(\mathbf{h}, \mathbf{w})$ of view image, random weight $\mathbf{w_x}, \mathbf{w_y}, \mathbf{w_z} \sim \mathcal{U}$, random direction $\mathbf{h_x}, \mathbf{h_y} \in \{-\mathbf{1}, \mathbf{1}\}$, up matrix $\hat{\mathbf{U}} = [\mathbf{0}, \mathbf{1}, \mathbf{0}]$
Output: The new view $\tilde{\mathbf{V}}$

Object position $\mathbf{O} \leftarrow [\mathbf{w_x} \cdot \mathbf{mean}(\mathbf{X}), \mathbf{w_y} \cdot \mathbf{mean}(\mathbf{Y}), \mathbf{w_z} \cdot \mathbf{mean}(\mathbf{Z})]^T$
Eye position $\mathbf{E} \leftarrow [\mathbf{w_z} \cdot \mathbf{mean}(\mathbf{Z}) \cdot \mathbf{h_x}, \mathbf{w_z} \cdot \mathbf{mean}(\mathbf{Z}) \cdot \mathbf{h_y}, \mathbf{w_z} \cdot \mathbf{mean}(\mathbf{Z})]^T$
Sight matrix $\mathbf{S} \leftarrow \mathbf{Uniform}(\mathbf{O} - \mathbf{E})$ where $\mathbf{Uniform}(\cdot)$ denotes standardize data into $\mathcal{N}(0, 1)$
Rotation matrix $\mathbf{R} \leftarrow [\mathbf{Uniform}(\mathbf{S} \times \hat{\mathbf{U}}), \mathbf{Uniform}(\mathbf{S} \times \hat{\mathbf{U}} \times \mathbf{S}), -\mathbf{S}]$
Translation matrix $\mathbf{T} \leftarrow -\mathbf{E}$
New camera coordinates $\mathbf{C_c} \leftarrow [[\mathbf{R}, \mathbf{T}], [\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}]] \times [\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{I}]$ where $\mathbf{I} = \mathbf{ones_like}(\mathbf{Z})$
$\mathbf{K} \leftarrow [[\mathbf{f}, \mathbf{0}, \mathbf{w}/\mathbf{2}], [\mathbf{0}, \mathbf{f}, \mathbf{h}/\mathbf{2}], [\mathbf{0}, \mathbf{0}, \mathbf{1}]]$
$\mathbf{U_c} \leftarrow \mathbf{K}[\mathbf{0}, \mathbf{0}] \cdot \mathbf{C_c}[\mathbf{0}]/\mathbf{C_c}[\mathbf{2}] + \mathbf{K}[\mathbf{0}, \mathbf{2}]$
$\mathbf{V_c} \leftarrow \mathbf{K}[\mathbf{1}, \mathbf{1}] \cdot \mathbf{C_c}[\mathbf{1}]/\mathbf{C_c}[\mathbf{2}] + \mathbf{K}[\mathbf{1}, \mathbf{2}]$
for $\mathbf{i} \in \mathbf{sorted}(\mathbf{C_c}[\mathbf{2}], -\mathbf{1})$ where $\mathbf{sorted}(\cdot, -\mathbf{1})$ denotes the index of sorting with the reverse order **do**
  if $\mathbf{U_c}[\mathbf{i}] \geq \mathbf{0}$ and $\mathbf{U_c}[\mathbf{i}] < \mathbf{w}$ and $\mathbf{V_c}[\mathbf{i}] \geq \mathbf{0}$ and $\mathbf{V_c}[\mathbf{i}] < \mathbf{h}$ **then**
    $\mathbf{x_i}, \mathbf{y_i} \leftarrow \mathbf{U_c}[\mathbf{i}], \mathbf{V_c}[\mathbf{i}]$
  end if
  $\tilde{\mathbf{V}}[\mathbf{y_i}, \mathbf{x_i}] \leftarrow \mathbf{P}[\mathbf{i}, \mathbf{3}:]$
end for
return The new view $\tilde{\mathbf{V}}$

where $|\Theta|$ denotes the number of valid pixels in $\mathbf{D_f} \cdot (1 - \mathbf{M})$. In this way, we encourage the proposed model to maintain consistency with the DFM in M, while ensuring model is not learning from DFM in $1 - M$. The total loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_c + \mathcal{L}_{DFM}. \tag{8}$$

where $\mathcal{L}_d = \|\mathbf{D} - \mathbf{D}^*\|_1$, $\mathcal{L}_{DFM} = \mathcal{L}_p + \mathcal{L}_n$. The whole framework containing main architecture and mentioned loss functions is illustrated in Fig. 2.

4 Experiments

4.1 Experimental settings

Datasets and baselines. Following Marigold Ke et al. (2024), we train our model in the mixed dataset containing Hypersim Roberts et al. (2021) and VKITTI Gaidon et al. (2016). For testing our model and baselines, all comparing methods and our methods are tested generally on commonly used Middlebury 2014 Scharstein et al. (2014), NYUv2 Silberman et al. (2012), KITTI Geiger et al. (2012), and RGBDD He et al. (2021). We compare our method with recent state-of-the-art methods, DKN Kim et al. (2021), DCT Zhao et al. (2022), LeReS Yin et al. (2023), SGN Wang et al. (2024b), G2 Wang et al. (2024a), ASDN Guo et al. (2025), and PDA Lin et al. (2025b). All methods are tested by their public code and pretrained weights. We first recover the affine-invariant results of monocular depth estimation models containing MiDaS Ranftl et al. (2022), LeReS Yin et al. (2023), and Marigold Ke et al. (2024) on Middlebury 2014, NYUv2, and KITTI, respectively. We further recover recent 3D perception models MoGe Wang et al. (2025b) and Moge-2 Wang et al. (2026) to show the progressiveness of the proposed method. We also validate the generalization recovery effect of RGBDD benchmark He et al. (2021) on raw depth with real-world scale distortions.

Metrics and implementation details. We report our method with the commonly used metrics to evaluate 2D depth maps Hu et al. (2022), containing AbsRel, RMSE, iRMSE, $\delta_{1.25}$, and Chamfer-Distance (CD) to evaluate 3D performance. During training, the RGB images and corresponding depth maps are resized into 320×320. Adam optimizer with an initial learning rate of 1×10^{-4} is adopted to train our model with a single RTX 3090 GPU. Due to the fact that DFM may suffer from 3D distortions, the results of DAV2 are also fed into the input of networks with a probability of 0.3 to replace the simulation of the diffusion model. The number of training epochs is set as 40, and the

Table 1: Quantitative results of recovering depth estimation results on Middlebury 2014, NYUv2, KITTI. The best metrics are in bold, and the method leveraging DAV2 are in italic.

		Middlebury 2014				NYUv2				KITTI			
		AbsRel↓	RMSE↓	iRMSE↓	$\delta_{1.25}$ ↑	AbsRel↓	RMSE↓	iRMSE↓	$\delta_{1.25}$ ↑	AbsRel↓	RMSE↓	iRMSE↓	$\delta_{1.25}$ ↑
DAV2		0.041	0.158	24.8	98.9	0.059	0.233	46.8	96.1	0.129	3.28	34.9	85.7
MiDaS	+DKN	0.076	0.288	43.4	93.6	0.121	0.427	81.1	86.3	0.345	4.83	193	39.4
	+DCT	0.079	0.296	44.2	93.9	0.118	0.448	82.9	86.3	0.319	4.88	215	44.5
	+SGN	0.077	0.288	43.6	93.8	0.119	0.432	**80.6**	86.5	0.331	4.84	195	41.9
	+G2	0.077	0.286	43.5	93.7	0.121	0.427	81.0	86.3	0.343	4.82	184	39.6
	+ASDN	0.098	0.366	57.3	89.7	0.202	0.626	142	75.7	0.447	5.81	236	34.3
	+PDA	0.097	0.346	52.5	91.1	0.149	0.489	86.2	81.4	0.393	5.11	143	34.0
	+Ours	**0.038**	**0.140**	**23.4**	**99.4**	**0.048**	**0.179**	82.8	**98.1**	**0.105**	**1.89**	**31.8**	**90.6**
LeReS	+DKN	0.070	0.270	38.9	95.7	0.078	0.287	**59.1**	93.6	0.153	2.78	44.4	77.4
	+DCT	0.072	0.282	40.4	95.2	0.083	0.309	61.0	93.2	0.164	2.97	57.2	71.0
	+SGN	0.071	0.274	39.5	95.6	0.079	0.292	59.7	93.4	0.143	2.83	38.3	79.7
	+G2	0.071	0.269	39.2	95.8	0.078	0.285	**59.1**	93.5	0.152	2.78	44.3	77.3
	+ASDN	0.106	0.405	58.4	88.3	0.207	0.697	138	72.9	0.533	7.41	213	22.7
	+PDA	0.093	0.337	48.4	92.1	0.112	0.386	76.5	88.5	0.246	4.21	40.6	58.8
	+Ours	**0.038**	**0.141**	**23.5**	**99.5**	**0.047**	**0.177**	69.0	**98.2**	**0.097**	**1.87**	**19.1**	**92.0**
Marigold	+DKN	0.053	0.222	30.8	98.0	0.064	0.254	51.9	95.2	0.115	2.70	25.1	88.1
	+DCT	0.057	0.233	32.0	97.4	0.071	0.286	56.2	94.6	0.131	2.82	30.9	83.3
	+SGN	0.055	0.225	31.4	97.7	0.067	0.265	53.8	94.7	0.108	2.69	22.0	89.0
	+G2	0.052	0.208	29.2	98.3	0.063	0.247	**50.7**	95.3	0.114	2.65	24.0	88.2
	+ASDN	0.110	0.441	64.4	86.5	0.212	0.727	153	71.5	0.499	7.03	157	24.0
	+PDA	0.075	0.283	41.0	94.9	0.109	0.377	73.9	89.2	0.198	4.20	31.6	67.8
	+Ours	**0.038**	**0.140**	**23.4**	**99.4**	**0.047**	**0.178**	73.6	**98.1**	**0.096**	**1.87**	**18.8**	**92.3**

Table 2: Quantitative results of recovering 3D perception results MoGe and MoGe-2 on Middlebury 2014, NYUv2, KITTI.

Method	Middlebury		NYUv2		KITTI	
	RMSE↓	CD↓	RMSE↓	CD↓	RMSE↓	CD↓
MoGe	0.156	0.064	0.285	0.121	2.23	28.9
MoGe+Ours	**0.137**	**0.056**	**0.181**	**0.099**	**1.89**	**25.4**
MoGe2	0.134	0.058	0.169	0.087	1.74	21.9
MoGe2+Ours	**0.133**	**0.054**	**0.164**	**0.077**	**1.71**	**19.4**

batch size is set as 32. To avoid the discrete points in the point cloud generated by recovery results, a simple weighted median filter Ma et al. (2013) based on DAV2 Yang et al. (2024) is applied to refine the recovery results.

4.2 Comparison with the State-of-the-Art

Enhancing different depth estimation models. While some depth estimation methods overemphasize 2D features, they underperform in 3D scenes. We comprehensively evaluate depth recovery approaches using three base models: MiDaS, LeReS, and Mairgold. The quantitative results in Table 1 demonstrate our method achieves unified and significant accuracy improvements over raw depth maps compared to recent methods. Though competing methods (i.e., PDA) adopt DFM for generalization, our approach outperforms all methods by addressing diverse scene-scale distortions, not only in raw depth but also within DFM outputs, which are frequently overlooked. Note that more visual results can be found in Appendix A.5. The results in Table 2 demonstrate that our approach still achieves certain improvements in depth maps and 3D evaluation, even for the latest 3D perception models.

Real-world depth recovery. RGBDD provides raw depth maps with real-world scale distortions and GT depth maps simultaneously. Following SGN Wang et al. (2024b), we bicubically upsample low-resolution raw depth to match the RGB image resolution. Quantitative results in Table 3 and

Table 3: Quantitative results of recovering real-world low-resolution depth maps on the RGBDD dataset.

method	AbsRel↓	RMSE↓	iRMSE↓	$\delta_{1.25}$ ↑
DKN	0.027	0.086	33.9	98.4
DCT	0.022	**0.059**	25.6	99.1
LeRes	0.153	0.297	112	89.5
SGN	**0.021**	**0.059**	26.4	99.0
G2	0.026	0.076	31.7	98.7
ASDN	0.027	0.080	33.5	98.4
PDA	0.065	0.151	54.4	96.0
Ours	**0.021**	0.060	**24.0**	**99.5**

Table 4: Ablation study of components in diffusion-based data simulation.

			Mid.2014		NYUv2		KITTI	
global	local	inter.	AbsRel	RMSE	AbsRel	RMSE	AbsRel	RMSE
✓			0.042	0.152	0.052	0.186	0.108	1.89
	✓		0.043	0.155	0.051	0.183	0.106	1.88
		✓	0.038	0.145	**0.046**	**0.175**	0.102	1.88
✓	✓		0.042	0.151	0.050	0.181	0.107	1.88
✓		✓	0.039	0.145	**0.046**	0.176	0.106	1.89
	✓	✓	0.040	0.148	0.047	0.177	0.105	1.88
✓	✓	✓	**0.038**	**0.140**	**0.046**	0.178	**0.096**	**1.87**

Table 5: Ablation study of leveraging DFM during training and testing.

		Mid.2014		NYUv2		KITTI	
train	test	AbsRel	RMSE	AbsRel	RMSE	AbsRel	RMSE
		0.061	0.240	0.068	0.258	0.141	3.14
✓		0.059	0.226	0.066	0.254	0.174	3.29
	✓	0.042	0.155	0.054	0.189	0.099	**1.87**
✓	✓	**0.038**	**0.140**	**0.047**	**0.178**	**0.096**	**1.87**

visual comparisons in appendix demonstrate our method's state-of-the-art performance in recovering real-world low-resolution depth.

4.3 Ablation Studies

All ablations evaluate Marigold recovery, with hyperparameters tuned here. For reproducibility, we set Marigold's initial random seed to 1 and interference steps to 10, as recommended. Additional ablations are in the Appendix A.2.

Components of diffusion-based raw depth simulation. In the module of diffusion-based data simulation, we separately explore the effects of global distortion simulation (denoted as "global"), local distortion simulation (denoted as "local"), and interpolation-based strategy (denoted as "inter."). Results in Table 4 show the effectiveness of each component in the simulation module, and combining three components together achieves the best results.

The leverage of DFM. To validate DFM's contribution to generalization, we conduct ablation studies removing it during either training or testing in Table 5. Testing without DFM loses large-scale data priors; training without it means DFM results are fed as raw depth into an data-simulation-trained model, the proposed method could recovers 3D distortion but is limited by missing DFM distribution during training. Thus, adopting DFM in both phases is optimal.

The weak supervision based on DFM. To mitigate 3D distortions of DFM, we introduce weak supervision (Eqs. equation 6, equation 7) that prevents overfitting to DFM's distortion estimates. Table 6 confirms each component's contribution in the proposed weak supervision. Note that while

Table 6: Ablation study of components in the proposed weak supervision.

$\mathcal{L}_p$	$\mathcal{L}_n$	Mid.2014 AbsRel	Mid.2014 RMSE	NYUv2 AbsRel	NYUv2 RMSE	KITTI AbsRel	KITTI RMSE
✓		0.040	0.148	0.047	0.178	0.103	1.88
	✓	0.042	0.153	0.050	0.181	0.103	1.88
✓	✓	**0.038**	**0.140**	**0.047**	**0.178**	**0.096**	**1.87**

Table 7: Ablation study of the proposed 3D constraint.

type	Mid.2014 AbsRel	Mid.2014 RMSE	NYUv2 AbsRel	NYUv2 RMSE	KITTI AbsRel	KITTI RMSE
w/o Ours	0.041	0.153	0.049	0.180	0.104	1.88
with VNL	0.040	0.149	**0.046**	0.176	0.102	1.88
with SGRL	0.039	0.146	**0.046**	**0.175**	0.104	1.88
with Ours	**0.038**	**0.140**	0.047	0.178	**0.096**	**1.87**

Table 8: Ablation study of the proposed 3D constraint.

supervison	Mid.2014 AbsRel	Mid.2014 RMSE	NYUv2 AbsRel	NYUv2 RMSE	KITTI AbsRel	KITTI RMSE
depth	0.041	0.151	0.048	**0.177**	0.101	1.88
rgb	**0.038**	**0.140**	**0.047**	0.178	**0.096**	**1.87**

DFM Yang et al. (2024) handles mirror surfaces prevalent in NYUv2, their absence in our training sets (Hypersim+VKITTI) limits verification efficacy.

The effectiveness of 3D constraint. Motivated by the observation that point clouds from different depth maps exhibit high consistency in the main view but diverge in alternative viewpoints, we propose multi-view consistency losses ($\mathcal{L}_c$, $\mathcal{L}_{DFM}$ in Eq. equation 8) for 3D scene optimization. Unlike existing constraints based on the main view (VNLYin et al. (2021), SGRL Xian et al. (2020)), our losses enforce cross-view consistency on depth maps. Table 7 demonstrates their superiority: ablating our losses ("w/o Ours") or substituting them with VNL/SGRL degrades performance, while our loss function achieves optimal accuracy. There is also a variant of this loss, which is designed as $\tilde{\mathbf{V}}[\mathbf{y_i}, \mathbf{x_i}] \leftarrow \mathbf{C_c}[2]$ in Algorithm 1 in the paper. The meaning is to generate and supervise corresponding depth maps in the new view instead of RGB supervision; however, the results in Table 8 show that the common RGB supervision achieves better results. We hold the opinion that the supervision based on depth maps may suffer from the scale problem, while supervision on RGB can mitigate the problem.

5 Conclusion

2D-based depth perception methods are inherently prone to 3D distortions, rendering accurate 3D recovery a critical challenge. Existing depth recovery approaches, which typically simulate raw depth maps within fixed scales and constrained environments, often struggle to generalize to real-world scenarios. To address this, we first simulate raw depth maps across a wide range of scales, encompassing both global and local patterns, facilitated by a diffusion model. A frozen depth foundation model (DFM) provides an initial scene estimation, yet remains susceptible to 3D distortions in practice. To overcome this, we design loss functions grounded in multi-view 3D consistency, which effectively guide the proposed model to learn intrinsic 3D features of depth maps rather than relying on superficial 2D cues. These losses are further formulated under a weak-supervision paradigm to enhance generalization across diverse 3D scenes. Extensive experiments on various real-world raw depth recovery tasks demonstrate the effectiveness and advancement of the proposed method.

6 STATEMENTS

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to [formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist with translation, source/search for information, create or modify scientific figures or images, and edit a research paper to improve readability].

We have not used generative AI tools to [generate synthetic datasets, help develop theoretical models or conceptual frameworks, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, formulate questions for surveys or interviews, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, transcribe recordings of research material, summarize or analyse existing literature, discover research topics or identify gaps, brainstorm, identify relevant literature, format references, suggest a structure for a research paper, and propose a title or keywords for a research paper].

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

First, for novel models or algorithms, a link to an anonymous downloadable source code is submitted as supplementary materials.

Second, for theoretical results, clear explanations of any assumptions and a complete proof of the claims are included in the appendix.

Third, all datasets drawn from the existing literature are accompanied by appropriate citations and publicly available. The data processing steps are the same with existing benchmarks.

REFERENCES

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pp. 75602–75637, 2025.

Yaqiao Dai, Renjiao Yi, Chenyang Zhu, Hongjun He, and Kai Xu. Multi-resolution monocular depth map fusion by self-supervised gradient-based composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 488–496, 2023.

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

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4340–4349, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361. IEEE, 2012.

Chuchen Guo, Weijie Zhou, Zheng Liu, and Ying He. You should learn to stop denoising on point clouds in advance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3212–3219, 2025.

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 9229–9238, 2021.

Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Junjie Hu, Chenyu Bao, Mete Ozay, Chenyou Fan, Qing Gao, Honghai Liu, and Tin Lun Lam. Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8244–8264, 2022.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 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*, 2025.

Chao Jia and Brian L. Evans. Constrained 3d rotation smoothing via global manifold regression for video stabilization. *IEEE Trans. Signal Process.*, 62(13):3293–3304, 2014.

Jiahui Kang, Qing Cai, Runqing Tan, Yimei Liu, and Zhi Liu. C2pd: Continuity-constrained pixel-wise deformation for guided depth super-resolution. *arXiv preprint arXiv:2501.07688*, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.

Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1477–1485, 2023.

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

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17070–17080, 2025b.

Wei Liu, Pingping Zhang, Yinjie Lei, Xiaolin Huang, Jie Yang, and Michael Ng. A generalized framework for edge-preserving and structure-preserving image smoothing. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6631–6648, 2021.

Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *Proceedings of the IEEE international conference on computer vision*, pp. 49–56, 2013.

S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9685–9694, 2021.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 120–136. Springer, 2020.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proceedings of the 36th German Conference on Pattern Recognition, Münster, Germany, September 2-5, 2014*, pp. 31–42. Springer, 2014.

Yunxiao Shi, Manish Kumar Singh, Hong Cai, and Fatih Porikli. Decotr: Enhancing depth completion with 2d and 3d attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10736–10746, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 746–760. Springer, 2012.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th acm international conference on multimedia*, pp. 4390–4399, 2021.

Haotian Wang, Meng Yang, Ce Zhu, and Nanning Zheng. Rgb-guided depth map recovery by two-stage coarse-to-fine dense crf models. *IEEE Transactions on Image Processing*, 32:1315–1328, 2023a.

Haotian Wang, Meng Yang, Ce Zhu, and Nanning Zheng. Rgb-guided depth map recovery by two-stage coarse-to-fine dense crf models. *IEEE transactions on image processing*, 32:1315–1328, 2023b.

Haotian Wang, Meng Yang, and Nanning Zheng. G2-monodepth: A general framework of generalized depth inference from monocular rgb+x data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3753–3771, 2024a.

Haotian Wang, Meng Yang, Xinhu Zheng, and Gang Hua. Scale propagation network for generalizable depth completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 1908–1922, 2025a.

Jin Wang, Longhua Sun, Ruiqin Xiong, Yunhui Shi, Qing Zhu, and Baocai Yin. Depth map super-resolution based on dual normal-depth regularization and graph laplacian prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3304–3318, 2021.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 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*, pp. 5261–5271, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.

Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. 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*, pp. 8445–8453, 2019.

Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5823–5831, 2024b.

Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5823–5831, 2024c.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5249–5260, 2025.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 611–620, 2020.

Mutian Xu, Chongjie Ye, Haolin Liu, Yushuang Wu, Jiahao Chang, and Xiaoguang Han. Stablesim2real: Exploring simulation of real-captured 3d data with two-stage depth diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2609–2619, 2025.

Wei Xu, Qing Zhu, and Na Qi. Depth map super-resolution via joint local gradient and nonlocal structural regularizations. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (12):8297–8311, 2022.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. volume 37, pp. 21875–21911, 2024.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2021.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6480–6494, 2023.

Shangshu Yu, Meiqing Wu, and Siew-Kei Lam. Vfm-depth: Leveraging vision foundation model for self-supervised monocular depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Structure flow-guided network for real depth super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3340–3348, 2023.

Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5697–5707, 2022.

Zhiyuan Zhou, Ruofeng Liu, Taichi Liu, Weijian Zuo, Shanshan Wang, Zhiqing Hong, and Desheng Zhang. Any to full: Prompting depth anything for depth completion in one stage. *arXiv preprint arXiv:2603.05711*, 2026.

Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13116–13125, 2020.

A Appendix

A.1 Theoretical basis of the diffusion-based data generation

Problem Formulation. Learning-based depth recovery model rigidly demands the raw depth $\mathbf{D}_r$ and GT depth $\mathbf{D}^*$ at the same time. While public RGB-D datasets like NYUv2 Silberman et al. (2012) and KITTI Geiger et al. (2012) provide ground-truth (GT) depth, they lack authentic and sufficient raw depth counterparts due to sensor limitations. This necessitates synthetic $\mathbf{D}_r$ simulation. To avoid overfitting the fixed scale of the raw depth, we employ a pretrained diffusion model Ke et al. (2024) for generating plausible raw depth in diverse scales. With the basic principle and the observation of effects in the diffusion model, we systematically address scene-scale distortions in both global and local patterns.

We leverage the reverse process of a latent diffusion model Rombach et al. (2022); Ke et al. (2024) to generate diverse distorted raw depth maps $\mathbf{D}_r$ with the condition of RGB image $\mathbf{C}$. The reverse diffusion step is formulated as:

$$\begin{aligned}\mathbf{Z}_{\mathbf{t-1}}^{(\mathbf{D}_r)} = &\sqrt{\bar{\boldsymbol{\alpha}}_{\mathbf{t-1}}}\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{Z}_{\mathbf{t}}^{(\mathbf{D}_r)}, \mathbf{Z}^{(\mathbf{C})}, \mathbf{t}\right) \\ &+\sqrt{1-\bar{\boldsymbol{\alpha}}_{\mathbf{t-1}}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{Z}_{\mathbf{t}}^{(\mathbf{D}_r)}, \mathbf{Z}^{(\mathbf{C})}, \mathbf{t}\right),\end{aligned} \tag{9}$$

where $\mathbf{Z}^{(\mathbf{x})} = \mathcal{E}(\mathbf{x})$ denotes the latent encoding of input $\mathbf{x}$ via a variational autoencoder (VAE) Rombach et al. (2022), $\mathbf{t}$ indexes the diffusion step, and $\bar{\boldsymbol{\alpha}}_{\mathbf{t}} = \prod_{\mathbf{i}=1}^{\mathbf{t}} \boldsymbol{\alpha}_{\mathbf{i}}$ represents the cumulative noise schedule coefficient. The denoising network $\boldsymbol{\theta}$ simultaneously predicts both the base state $\mathbf{f}_{\boldsymbol{\theta}}$ and noise component $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.

The synthesis process initiates from Gaussian noise $\mathbf{Z}_0^{(\mathbf{D}_r)} \approx \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, iteratively applying Eq. 9 for $\mathbf{t}$ steps to obtain the denoised latent $\mathbf{Z}_{-\mathbf{t}}^{(\mathbf{D}_r)}$. Note that the output $\mathbf{Z}_{-\mathbf{t}}^{(\mathbf{D}_r)}$ is not unique; it is a function of both the initial noise $\boldsymbol{\varepsilon}$ and the discrete step size $\Delta\mathbf{t}$. The final raw depth is reconstructed through the VAE decoder: $\mathbf{D}_{\mathbf{r}}^{\mathbf{t}} = \mathcal{D}(\mathbf{Z}_{-\mathbf{t}}^{(\mathbf{D}_r)})$.

Global Depth Scale vs. Inference Steps. We first establish how the number of inference steps N affects the global scale of the output depth. The global scale is defined as the mean value of the depth map, i.e., $S_{\text{global}} = \mathbb{E}[\hat{\mathbf{D}}_r]$.

Let $\mathbf{Z}_{-T}^{*(\mathbf{D}_r)}$ be the exact solution of the ordinary differential equation in Eq. (2) at $\mathbf{t} = T$, and let $\hat{\mathbf{Z}}_{-T}^{(\mathbf{D}_r)}$ be the numerical solution obtained via an ordinary differential equation (ODE) solver with step size $\Delta\mathbf{t} = 1$. For a Lipschitz-continuous drift function Song et al. (2020) with constant L_μ, the expected squared error between the exact and numerical latent solutions is bounded by:

$$\mathbb{E}\left[\|\mathbf{Z}_{-T}^{*(\mathbf{D}_r)} - \hat{\mathbf{Z}}_{-T}^{(\mathbf{D}_r)}\|_2^2\right] \leq \frac{C_1^2}{T^2} + \frac{C_2}{T} \leq \frac{C}{T}, \tag{10}$$

where C is a constant dependent on L_μ and the total diffusion time T. Taking the square root yields:

$$\|\mathbf{Z}_{-T}^{*(\mathbf{D}_r)} - \hat{\mathbf{Z}}_{-T}^{(\mathbf{D}_r)}\|_2 \leq \frac{C}{\sqrt{T}}. \tag{11}$$

Since the VAE decoder $\mathcal{D}$ is frozen and globally Lipschitz continuous with constant L_D, the stability of the decoder ensures:

$$\begin{aligned}&\|\mathbf{D}_r^{-T*} - \hat{\mathbf{D}}_r^{-T}\|_2 = \\ &\|\mathcal{D}(\mathbf{Z}_{-T}^{*(\mathbf{D}_r)}) - \mathcal{D}(\hat{\mathbf{Z}}_{-T}^{(\mathbf{D}_r)})\|_2 \leq L_D \cdot \frac{C}{\sqrt{T}}.\end{aligned} \tag{12}$$

To isolate the global scale, we apply the expectation operator $\mathbb{E}[\cdot]$, which is a linear functional with unit norm. Consequently, the global scale deviation satisfies:

$$\left|\mathbb{E}[\mathbf{D}_r^{-T*}] - \mathbb{E}[\hat{\mathbf{D}}_r^{-T}]\right| \leq \frac{L_D \cdot C}{\sqrt{T}}. \tag{13}$$

Specifically, dropping the square root for the dominant ODE convergence (Heun/DDIM exhibits first-order convergence in $1/N$), the dominant global bias decays as:

$$\left|S^*_{\text{global}} - S_{\text{global}}\right| \propto \mathcal{O}\left(\frac{1}{T}\right). \tag{14}$$

This proves that increasing the inference steps N monotonically reduces the global offset (baseline shift) of the estimated depth, aligning it with the true scale.

Local Depth Scale vs. Initial Input Noise. We now analyze the sensitivity of local structures, such as edges and texture details, to the initial noise $\boldsymbol{\varepsilon}$. The local scale S_{local} is measured by the magnitude of the spatial gradient, $S_{\text{local}} = \|\nabla \hat{\mathbf{D}}_r\|_2$.

The reverse ODE defines a deterministic dynamical system $\Phi : \mathbf{Z}_{-T} \mapsto \mathbf{Z}_0$ parameterized by the conditional input $\mathbf{Z}^{(\mathbf{C})}$. However, due to the inherent ambiguity of depth estimation, this system exhibits sensitivity to initial conditions.

Consider two different initial noise samples $\mathbf{Z}_0^{(a)}$ and $\mathbf{Z}_0^{(b)}$, with difference $\delta\mathbf{Z}_T = \mathbf{Z}_0^{(a)} - \mathbf{Z}_0^{(b)}$. The evolution of this perturbation through the reverse dynamics is governed by the Jacobian matrix Ho et al. (2020):

$$\frac{\partial \mathbf{Z}_{-T}}{\partial \mathbf{Z}_0} = \prod_{k=1}^{N} \frac{\partial \mathbf{f}_\theta(\mathbf{Z}_{\mathbf{t}_k}, \mathbf{t}_k)}{\partial \mathbf{Z}_{\mathbf{t}_k}} + \prod_{k=1}^{N} \frac{\partial \epsilon_\theta(\mathbf{Z}_{\mathbf{t}_k}, \mathbf{t}_k)}{\partial \mathbf{Z}_{\mathbf{t}_k}}, \tag{15}$$

where $\mathbf{f}_\theta$ and ϵ_θ together represent a single reverse step. Let $\lambda_{\max} > 0$ be the maximum Lyapunov exponent of this system over the interval $[0, T]$. The spectral norm of the Jacobian is bounded below by:

$$\left\| \frac{\partial \mathbf{Z}_{-T}}{\partial \mathbf{Z}_0} \right\|_2 \geq c \cdot e^{\lambda_{\max} T}, \tag{16}$$

for some positive constant c. Hence, the difference in the final latent states satisfies:

$$\|\mathbf{Z}_{-T}^{(a)} - \mathbf{Z}_{-T}^{(b)}\|_2 \geq c \cdot e^{\lambda_{\max} T} \|\mathbf{Z}_0^{(a)} - \mathbf{Z}_0^{(b)}\|_2. \tag{17}$$

To examine local details, we apply the spatial gradient operator ∇ to the decoded depth. By the chain rule, the gradient of the difference is:

$$\nabla \hat{\mathbf{D}}_r^T(a) - \nabla \hat{\mathbf{D}}_r^T(b) = \nabla \mathcal{D}(\mathbf{Z}_{-T}^{(a)}) - \nabla \mathcal{D}(\mathbf{Z}_{-T}^{(b)}). \tag{18}$$

Since the decoder $\mathcal{D}$ is a convolutional neural network that amplifies high-frequency components, its local gradient operator is Lipschitz continuous with respect to $\mathbf{Z}_0$ in the high-frequency subspace. Therefore, combining with the exponential growth of the latent perturbation:

$$\left\| \nabla \hat{\mathbf{D}}_r^T(a) - \nabla \hat{\mathbf{D}}_r^T(b) \right\|_2 \propto e^{\lambda_{\max} T} \cdot \|\mathbf{Z}_0^{(a)} - \mathbf{Z}_0^{(b)}\|_2. \tag{19}$$

Since T is fixed and $\lambda_{\max} > 0$, the factor $e^{\lambda_{\max} T} > 1$. Thus, the local structural variations (edges, fine details) are exponentially sensitive to the initial noise. Even an infinitesimal change in the random seed (i.e., $\|\delta\mathbf{Z}_0\|_2 \to 0$) results in significantly different local depth patterns.

Unified Formulation and Conclusion. Combining both theorems, the simulated depth map is a function of the hyperparameters:

$$\hat{\mathbf{D}}_r^T = \mathcal{D}\left(\text{ODESolve}\left(\mathbf{f}_\theta, \epsilon_\theta;\, \mathbf{Z}_0^{(\mathbf{D}_r)}, t = T, \mathbf{Z}^{(\mathbf{C})} = \mathcal{E}(\mathbf{C})\right)\right). \tag{20}$$

- **Global Scale (via T)**: The error decays as $\mathcal{O}(1/T)$. To achieve accurate metric depth, one must use a sufficiently large number of steps.

Table 9: The analysis of the randomness of the diffusion model. The precision that is improved is in bold.

type	Mid.2014		NYUv2		KITTI	
	AbsRel	RMSE	AbsRel	RMSE	AbsRel	RMSE
MS1	0.052	0.207	0.063	0.246	0.114	2.66
MS1+Ours	**0.038**	**0.140**	**0.047**	**0.178**	**0.096**	**1.87**
MS2	0.053	0.203	0.062	0.242	0.113	2.59
MS2+Ours	**0.038**	**0.139**	**0.047**	**0.178**	**0.095**	**1.86**
MS3	0.050	0.191	0.063	0.247	0.116	2.60
MS3+Ours	**0.038**	**0.139**	**0.047**	**0.178**	**0.096**	**1.87**

Table 10: The cost of the proposed method.

type	AbsRel	para/M	memo/MB	infe/image$\cdot s^{-1}$
DAV2-large	-	335.3	1887	18.0
small	0.044	**53.0**	**214**	**28.8**
base	0.039	93.8	375	28.1
large	**0.038**	210	844	23.1

- **Local Scale (via $\mathbf{Z}_0$)**: The variability grows exponentially due to the nature of the reverse process. This property is fundamental to generative models, enabling the estimation of aleatoric uncertainty in ambiguous regions (e.g., textureless areas).

This theoretical framework provides a solid foundation for developing adaptive step-size schedulers and noise-ensemble strategies for robust raw depth simulation containing diverse global and local distortions.

Notably, recent Stable-Sim2Real Xu et al. (2025) overlaps with the proposed data-simulation on the core idea of simulating depth distortion using diffusion models. The proposed method simulates global and local scale distortion by controlling the diffusion steps and initial noise, while Stable-Sim2Real generates realistic depth data with real sensor noise through two-stage diffusion and 3D discriminator, focusing on simulating the noise patterns of real data rather than scale distortion.

A.2 More Ablation Studies

The randomness of the diffusion model. The diffusion model may generate different possible results with different initial noise. In this test, we set the random seed of 1,2,3 to fix the estimated results of Marigold Ke et al. (2024), denoted as "MS1, MS2, MS3" in Table 9, respectively. Then we recover these results by the proposed model. Results in Table 9 show that the diffusion model differs in results by a small margin with different initial noise, and the proposed method effectively recovers different results estimated by diffusion models. Notice that all results in this paper are calculated once because the only factor that can change the results is the random seed. While the random seed is fixed, the results will never be changed.

Parameters and Speed. We adopt ConvNeXt Woo et al. (2023) as our base model, which provides choices of "small", "base", and "large" models for different needs. We adopt the AbsRel of Middlebury 2014 in Table 1 (in the paper) for showing the precision, and the number of parameters (denoted as "para"), memory cost (denoted as "memo"), and inference speed (denoted as "infe") are used to evaluate the cost of different models. Note that all models are tested on a RTX 3090 GPU with a resolution of 320×320, and the static results of DAV2 is applied for testing costs. Results in Table 10 show the cost of different models, while ConvNeXt-large achieves the best accuracy, ConvNeXt-small achieves the lowest cost, and acceptable accuracy. Additionally, as a further application of the DAV2 model, our approach achieves acceptable results in terms of model cost.

The settings of hyperparameters. Two hyperparameters are verified in this part. The first one is the probability of 0.3 in implementation details. We tried the probability from 0 to 0.5 in steps of

Table 11: Ablation study of the setting of hyperparameter P.

P	Mid.2014		NYUv2	
	AbsRel	RMSE	AbsRel	RMSE
0	0.042	0.152	0.049	0.180
0.1	0.041	0.149	0.048	0.177
0.2	0.039	0.148	**0.046**	0.176
0.3	**0.038**	**0.140**	0.047	0.178
0.4	0.039	0.145	0.047	**0.175**
0.5	0.041	0.150	0.048	0.177

Table 12: Ablation study of the setting of hyperparameter s.

s	Mid.2014		NYUv2	
	AbsRel	RMSE	AbsRel	RMSE
0	0.041	0.153	0.049	0.180
1	**0.038**	**0.140**	**0.047**	0.178
2	0.040	0.150	0.048	**0.177**
3	0.042	0.152	0.050	0.181

0.1, and found 0.3 is the best choice in Table 11. The second one is the number s of views $\tilde{\mathbf{V}}$ in the 3D loss. s is set from 0 to 3, while $s = 1$ is verified as the best choice in Table 12.

The effectiveness of weighted median filter. A simple weighted median filter is used for dealing with flying pixels in output depth. Table 13 demonstrates the effectiveness of the weighted median filter, which mainly helps the metrics of RMSE.

A.3 Recovering Heavy Structure Misalignments.

The heavy structure misalignments remain prevalent in real-world depth maps Wang et al. (2023a), which also affects the scene scale in 3D scenes. Thus we verify our method on recovering the structure misalignments are generated by SGBM Hirschmuller (2007) on Middlebury 2014 dataset. Quantitative results in Table 14 and visual results in Fig. 3 demonstrate the generalization capability of our method, with the SOTA performance of recovering structure misalignments in 3D scenes.

A.4 More 3D metrics

Table 15 evaluates Chamfer Distance (CD) and Mean Angle Error (MAE) recovering Marigold, verifying the effectiveness and progressiveness of the proposed method in 3D space.

A.5 More Visual Results

Due to space limitations, visual results are provided here. Figs. 4, 5, 6 prove the effectiveness of the proposed method on recovering different 3D distortions on 3 datasets. Notably, DAV2 is the DFM we adopted and still suffered from 3D distortions, the proposed method effectively recovers the shape of bed in Fig. 6. Additionally, Fig. 7 shows that the proposed method can also recover the 3D distortions caused by low-cost sensors in RGBDD dataset. It further proves the generalization ability of the proposed method.

A.6 Limitations

There are two main limitations to the proposed method. The first aspect is that in the process of utilizing DFM, although other metrics have gained advantages, there is a phenomenon that the iRMSE metric is not as good as the DFM results. We believe that this is due to still approximating the scale problem as a global variable of the image, lacking region level or even pixel level perception for close range, which will become our future research problem. Secondly, using DFM only during the testing phase can achieve good results, which comes from treating DFM results as raw depth

Table 13: Ablation study of the weighted median filter (WMF).

method	Mid.2014		NYUv2		KITTI	
	AbsRel	RMSE	AbsRel	RMSE	AbsRel	RMSE
w/o WMF	**0.038**	0.148	**0.046**	0.188	0.097	1.97
with WMF	**0.038**	**0.140**	0.047	**0.178**	**0.096**	**1.87**

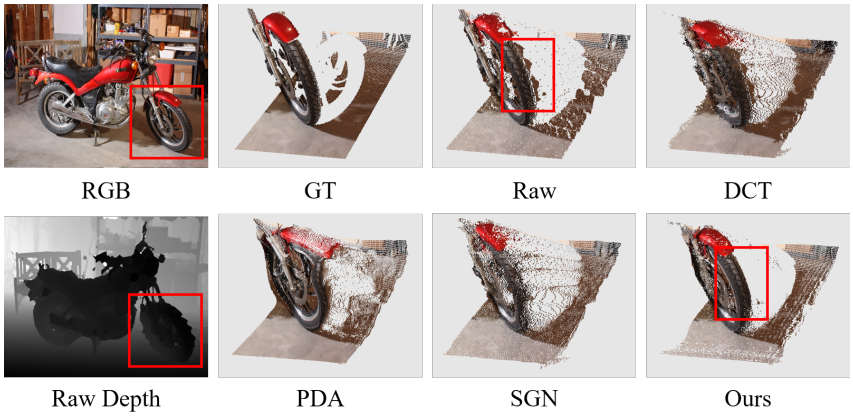


Figure 3: Visual results of recovering heavy structure misalignments on the Middlebury 2014 dataset.

inputs to the network. Although this proves that our approach can generalize the results of DFM (DAV2), it to some extent weakens the innovation of our weakly supervised design.

REFERENCES

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pp. 75602–75637, 2025.

Yaqiao Dai, Renjiao Yi, Chenyang Zhu, Hongjun He, and Kai Xu. Multi-resolution monocular depth map fusion by self-supervised gradient-based composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 488–496, 2023.

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

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4340–4349, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361. IEEE, 2012.

Chuchen Guo, Weijie Zhou, Zheng Liu, and Ying He. You should learn to stop denoising on point clouds in advance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3212–3219, 2025.

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 9229–9238, 2021.

Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

Table 14: Quantitative results of recovering heavy structure misalignments on the Middlebury 2014 dataset.

method	AbsRel↓	RMSE↓	iRMSE↓	$\delta_{1.25}$ ↑
DKN	0.027	0.248	34.9	96.5
DCT	0.043	0.252	34.0	96.6
LeRes	0.195	0.618	112	67.6
SGN	0.032	0.237	30.9	97.0
G2	**0.024**	0.227	30.8	96.9
ASDN	0.101	0.488	68.1	84.2
PDA	0.066	0.294	39.2	95.9
Ours	0.034	**0.187**	**27.4**	**97.9**

Table 15: Quantitative results of recovering Marigold results on Middlebury, NYUv2, KITTI.

	Mid. 2014		NYUv2		KITTI	
	CD↓	MAE↓	CD↓	MAE↓	CD↓	MAE↓
DKN(2021)	0.076	36.2	0.192	59.6	13.8	44.9
DCT(2022)	0.078	44.3	0.239	68.0	**11.4**	42.2
SGN(2024)	0.079	41.2	0.247	63.5	13.1	45.9
G2(2024)	0.079	29.9	0.251	56.7	13.9	46.8
ASDN(2025)	0.088	44.2	0.268	69.1	14.3	49.8
PDA(2025)	0.094	37.8	0.332	61.8	12.3	51.9
Ours	**0.058**	**24.1**	**0.175**	**39.8**	**11.4**	**41.7**

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Junjie Hu, Chenyu Bao, Mete Ozay, Chenyou Fan, Qing Gao, Honghai Liu, and Tin Lun Lam. Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8244–8264, 2022.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 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*, 2025.

Chao Jia and Brian L. Evans. Constrained 3d rotation smoothing via global manifold regression for video stabilization. *IEEE Trans. Signal Process.*, 62(13):3293–3304, 2014.

Jiahui Kang, Qing Cai, Runqing Tan, Yimei Liu, and Zhi Liu. C2pd: Continuity-constrained pixel-wise deformation for guided depth super-resolution. *arXiv preprint arXiv:2501.07688*, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.

Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1477–1485, 2023.

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



Figure 4: Visual results of recovering MiDaS on the Middlebury 2014 dataset.

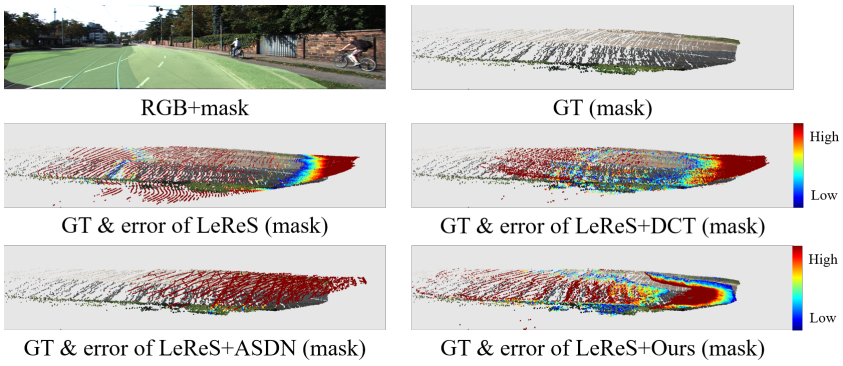


Figure 5: Visual results of recovering LeReS on the KITTI dataset.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17070–17080, 2025b.

Wei Liu, Pingping Zhang, Yinjie Lei, Xiaolin Huang, Jie Yang, and Michael Ng. A generalized framework for edge-preserving and structure-preserving image smoothing. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6631–6648, 2021.

Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *Proceedings of the IEEE international conference on computer vision*, pp. 49–56, 2013.

S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9685–9694, 2021.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 120–136. Springer, 2020.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

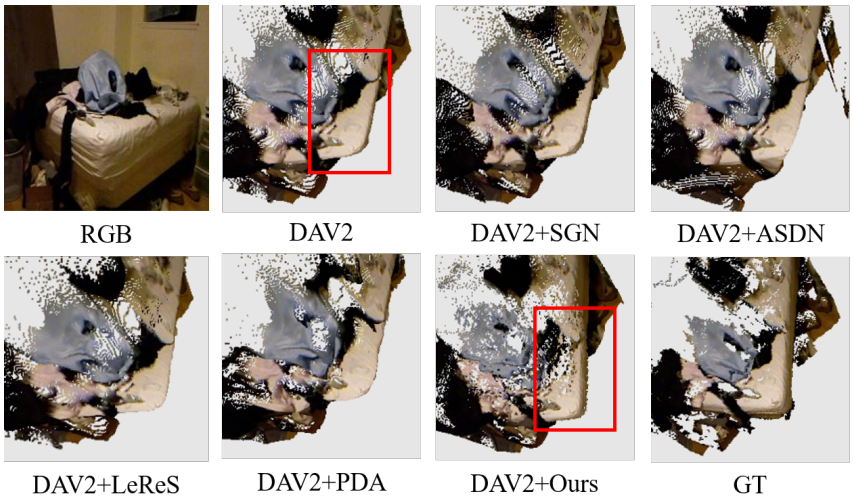


Figure 6: Visual results of recovering DAV2 on the NYUv2 dataset.

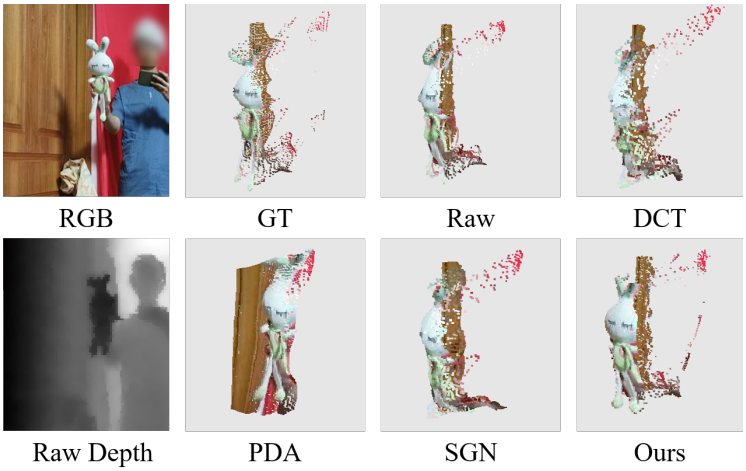


Figure 7: Visual results of recovering the real-world low-resolution depth maps on the RGBDD dataset.

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

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proceedings of the 36th German Conference on Pattern Recognition, Münster, Germany, September 2-5, 2014*, pp. 31–42. Springer, 2014.

Yunxiao Shi, Manish Kumar Singh, Hong Cai, and Fatih Porikli. Decotr: Enhancing depth completion with 2d and 3d attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10736–10746, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 746–760. Springer, 2012.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th acm international conference on multimedia*, pp. 4390–4399, 2021.

Haotian Wang, Meng Yang, Ce Zhu, and Nanning Zheng. Rgb-guided depth map recovery by two-stage coarse-to-fine dense crf models. *IEEE Transactions on Image Processing*, 32:1315–1328, 2023a.

Haotian Wang, Meng Yang, Ce Zhu, and Nanning Zheng. Rgb-guided depth map recovery by two-stage coarse-to-fine dense crf models. *IEEE transactions on image processing*, 32:1315–1328, 2023b.

Haotian Wang, Meng Yang, and Nanning Zheng. G2-monodepth: A general framework of generalized depth inference from monocular rgb+x data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3753–3771, 2024a.

Haotian Wang, Meng Yang, Xinhu Zheng, and Gang Hua. Scale propagation network for generalizable depth completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 1908–1922, 2025a.

Jin Wang, Longhua Sun, Ruiqin Xiong, Yunhui Shi, Qing Zhu, and Baocai Yin. Depth map super-resolution based on dual normal-depth regularization and graph laplacian prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3304–3318, 2021.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 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*, pp. 5261–5271, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.

Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. 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*, pp. 8445–8453, 2019.

Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5823–5831, 2024b.

Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5823–5831, 2024c.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5249–5260, 2025.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 611–620, 2020.

Mutian Xu, Chongjie Ye, Haolin Liu, Yushuang Wu, Jiahao Chang, and Xiaoguang Han. Stablesim2real: Exploring simulation of real-captured 3d data with two-stage depth diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2609–2619, 2025.

Wei Xu, Qing Zhu, and Na Qi. Depth map super-resolution via joint local gradient and nonlocal structural regularizations. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (12):8297–8311, 2022.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. volume 37, pp. 21875–21911, 2024.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2021.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6480–6494, 2023.

Shangshu Yu, Meiqing Wu, and Siew-Kei Lam. Vfm-depth: Leveraging vision foundation model for self-supervised monocular depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Structure flow-guided network for real depth super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3340–3348, 2023.

Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5697–5707, 2022.

Zhiyuan Zhou, Ruofeng Liu, Taichi Liu, Weijian Zuo, Shanshan Wang, Zhiqing Hong, and Desheng Zhang. Any to full: Prompting depth anything for depth completion in one stage. *arXiv preprint arXiv:2603.05711*, 2026.

Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13116–13125, 2020.