FOUNDDSR: A GENERALIZABLE FOUNDATION MODEL WITH GUIDED 2D GAUSSIAN SPLATTING FOR DEPTH SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce FoundDSR, a generalizable foundation model for robust depth reconstruction across unseen data distributions using RGB-D pairs. FoundDSR begins with a guided 2D Gaussian Splatting strategy to model depth representations with Gaussian primitives. This strategy employs high-resolution RGB as prompts to optimize the Gaussian parameters, thereby encouraging each Gaussian primitive to anisotropically deform along high-frequency structural directions. The resulting Gaussian-upsampled representations are then mapped to high-resolution depth through an effective depth reconstruction branch. Furthermore, to mitigate training instability and bias toward dominant sources caused by distribution gaps in large-scale heterogeneous data, we introduce heterogeneous federated learning that allocates each data source to an independent client for local optimization and global aggregation. This design effectively endows FoundDSR with stable scalability to diverse and large-scale training data. Extensive zero-shot evaluations on synthetic, real-world, arbitrary-scale, and noisy conditions demonstrate that FoundDSR consistently outperforms existing state-of-the-art approaches, confirming its strong robustness and generalization to unknown scenes.

1 INTRODUCTION

Depth maps capture rich geometric information, making them essential for various computer vision tasks, including augmented reality Zhao et al. (2023); Wang et al. (2023a), virtual reality Ye et al. (2025b), 3D reconstruction Yan et al. (2026); Sun et al. (2021), and robotic navigation Yuan et al. (2023); Song et al. (2020). Due to the limitations of sensor technology and imaging environments, consumer-grade devices typically acquire low-resolution (LR) depth corrupted by unknown degradations. To this end, numerous depth super-resolution (DSR) methods Yang et al. (2022); Chen et al. (2024); Zheng et al. (2025) have been proposed to restore accurate high-resolution (HR) depth from their LR inputs, with RGB employed as guidance to facilitate depth recovery.

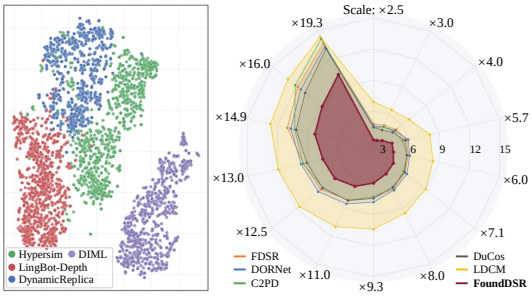
However, real-world scenarios are inherently variable, with acquisition environments and target objects often diverging significantly from training data. Moreover, different types of depth sensors often exhibit distinct noise distributions and degradation patterns during imaging. Most existing DSR methods Wang et al. (2023b; 2024) rely on small-scale and homogeneous data, resulting in suboptimal zero-shot generalization to domains outside the training distribution. Although recent works Yan et al. (2025b); Wang et al. (2026c;a) leverage geometric priors from pretrained vision foundation models to improve robustness, they are still fine-tuned on limited and less diverse data, failing to overcome the zero-shot generalization bottleneck in real-world scenes.

In this paper, we propose FoundDSR, a novel foundation model for DSR trained on million-scale data to significantly improve zero-shot generalization across unseen domains. To enable arbitrary-scale depth restoration, we develop a guided 2D Gaussian Splatting strategy that models depth features with continuous Gaussian primitives. Specifically, this strategy employs RGB gradients as prompts to adaptively modulate both the covariance and spatial position of each Gaussian primitive, driving them to anisotropically deform in orientation, shape, and position along high-frequency structural directions. This anisotropic design facilitates accurate depth reconstruction at arbitrary upsampling factors. Subsequently, we introduce a simple yet effective depth reconstruction branch.

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.

Eschewing complex module designs, it adopts a straightforward encoder-decoder architecture to restore accurate HR depth from the Gaussian-upsampled depth representations, while maintaining low inference latency.

Furthermore, Fig. 1(a) visualizes the embedding distributions across different data sources, revealing pronounced distribution gaps in large-scale heterogeneous data. During mixed-data training, such discrepancies can bias optimization toward dominant sources with larger proportions and destabilize training. To address this issue, we introduce a heterogeneous federated learning approach that assigns each data source to an independent client for local optimization. Subsequently, the server periodically aggregates and redistributes the client weights. This mechanism effectively mitigates cross-source optimization interference, significantly improving both scalability and training stability on large-scale heterogeneous datasets.



(a) t-SNE of different data (b) Zero-shot comparison

Figure 1: (a) t-SNE visualization of feature embeddings from different data sources. (b) Average RMSE across all benchmarks at different scales.

Fig. 1(b) compares the generalization performance of our FoundDSR with previous depth foundation models and DSR approaches across multiple scales. These results demonstrate that our method exhibits superior zero-shot generalization to unseen scenarios, achieving an average RMSE reduction of 20.66% over the second-best method across all scales.

In summary, our contributions are as follows:

- For the first time, we propose a novel DSR foundation model that is trained on large-scale heterogeneous data, substantially improving zero-shot generalization to unseen scenes.
- We present a guided 2D Gaussian Splatting that leverages high-resolution RGB as prompts to anisotropically deform Gaussian primitives along high-frequency structural directions, effectively modeling depth representations at arbitrary scales.
- We introduce heterogeneous federated learning to mitigate training instability and bias toward dominant data sources induced by distribution gaps across large-scale multi-source data, enabling stable scaling to larger and more diverse training data.
- Extensive experiments across various datasets demonstrate the superior zero-shot generalization and robustness of our FoundDSR over existing state-of-the-art approaches.

2 Related Work

2.1 Depth Super-Resolution

Recent deep learning-based DSR methods Zhong et al. (2025; 2026); Wang et al. (2026b) mainly focus on RGB-D feature fusion, exploring strategies such as shared and modality-specific feature decomposition Deng & Dragotti (2020), frequency decoupling He et al. (2021), and guided filtering Zhong et al. (2023); Wang et al. (2026c). For example, DCTNet Zhao et al. (2022) disentangles shared and private RGB-D features in Euclidean and spherical feature spaces to suppress depth-irrelevant cues during multi-modal feature fusion. SGNet Wang et al. (2024) proposes a structure-guided DSR network that simultaneously extracts and aggregates the high- and low-frequency information from multi-modal inputs across the spatial, gradient, and frequency domains. DKN Kim et al. (2021) incorporates a deformable mechanism into joint image filtering, adaptively transferring RGB features to depth features. More recently, several methods Wang et al. (2025); Yan et al. (2025b) have explored improving the generalization of DSR models. For instance, DuCos Yan et al. (2025b) presents a dual-constraint network that leverages a foundation model for depth estimation as a prompt to enhance generalization across different scenes. However, these methods still rely on fine-tuning on small-scale datasets, which limits their performance to unseen domains. To address this issue, we introduce a novel DSR foundation model trained on large-scale heterogeneous data, which demonstrates superior robustness and zero-shot generalization.

2.2 VISION FOUNDATION MODEL

Foundation models have achieved groundbreaking progress in many computer vision fields, such as image segmentation and depth estimation. Their core principle is to improve generalization by scaling up training data rather than introducing increasingly intricate module designs. For instance, SAM Kirillov et al. (2023) constructs a large-scale segmentation dataset and introduces the first segment anything model, achieving remarkable zero-shot segmentation performance. Recently, Depth Anything Yang et al. (2024a) introduces the first foundation model for monocular depth estimation, jointly trained on large-scale labeled and unlabeled data. This has inspired a subsequent line of foundation models Ke et al. (2024); Hu et al. (2024) for depth estimation. For instance, PromptDA Lin et al. (2025) employs low-cost LiDAR as a prompt for the decoder of a depth estimation foundation model, improving metric depth prediction. Unlike these foundation models, we introduce a novel DSR foundation model that leverages federated learning and guided 2D Gaussian Splatting, significantly improving zero-shot generalization across diverse scenes and arbitrary upsampling factors.

2.3 GAUSSIAN SPLATTING

Gaussian Splatting (GS) has shown strong potential for image representation. 3D GS Kerbl et al. (2023); Xu et al. (2025) combines explicit Gaussian representations with differentiable rasterization, enabling efficient 3D reconstruction and novel view synthesis. However, 3D GS typically depends on accurate camera intrinsics and extrinsics for projection and rendering, restricting its use when calibration data is unavailable. Recently, 2D GS Zeng et al. (2025); Chen et al. (2025) has been introduced, enabling direct modeling of geometric structures within the pixel coordinate system without requiring explicit 3D projection transformations. For example, GaussianImage Zhang et al. (2024) proposes a pioneering 2D GS paradigm for image representation and compression, utilizing 2D Gaussians instead of 3D ones to substantially reduce the storage requirements of Gaussian representation. Different from these single-modal approaches, we introduce a cross-modal guided 2D GS. This strategy leverages the first-order gradients of high-resolution RGB as structural prompts to optimize depth Gaussian primitives, thereby enhancing the accuracy of depth geometry.

3 METHOD

3.1 PRELIMINARIES: 2D GAUSSIAN SPLATTING

Given an image $\boldsymbol{Z}$, we model it as a continuous spatial representation using a set of 2D Gaussians Peng et al. (2026), where $\boldsymbol{Z}(x, y)$ at any continuous position (x, y) is obtained:

$$\boldsymbol{Z}(x,y) = \sum_{m=1}^{M} \mathcal{G}_m(x,y), \quad \mathcal{G}_m(x,y) = \boldsymbol{c}_{rgb,m} \cdot \frac{1}{2\pi|\boldsymbol{\Sigma}_m|} \exp\left(-\frac{1}{2}\boldsymbol{d}_m^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{d}_m\right), \tag{1}$$

where M represents the number of Gaussians. Each Gaussian kernel $\mathcal{G}_m$ is parameterized by its spatial position $\boldsymbol{\mu}_m = (\mu_{x,m}, \mu_{y,m})\top$, covariance matrix $\boldsymbol{\Sigma}_m$, and color coefficients $\boldsymbol{c}_{rgb,m} \in [0,1]^3$. $\boldsymbol{d}_m = (x - \mu_{x,m}, y - \mu_{y,m})\top$ denotes the displacement vector from the Gaussian center to the query point (x, y). The covariance matrix $\boldsymbol{\Sigma}_m$ can be further factorized into:

$$\boldsymbol{\Sigma}_m = \boldsymbol{R}_m \begin{bmatrix} \sigma_{x,m}^2 & 0 \\ 0 & \sigma_{y,m}^2 \end{bmatrix} \boldsymbol{R}_m^\top, \quad \boldsymbol{R}_m = \begin{bmatrix} \cos(\theta_m) & -\sin(\theta_m) \\ \sin(\theta_m) & \cos(\theta_m) \end{bmatrix}, \tag{2}$$

where $\sigma_{x,m}$ and $\sigma_{y,m}$ are the standard deviations that control the shape of each Gaussian primitive. $\boldsymbol{R}_m$ is the rotation matrix and θ_m is the rotation angle, which together determine the orientation.

3.2 OVERALL FRAMEWORK

As shown in Fig. 2, FoundDSR incorporates heterogeneous federated learning to collaboratively optimize large-scale RGB-D data, mitigating the interference caused by heterogeneous data. All clients share the same architecture, each comprising a guided 2D GS branch and a depth reconstruction branch. Given an LR depth $\boldsymbol{D}_{lr} \in R^{h \times w \times 1}$, where h and w denote its height and width, the guided 2D GS first employs an encoder $\mathcal{E}_{lr}$ to predict initial Gaussians $\mathcal{G}$. Then, RGB gradients $\boldsymbol{I}_x$, $\boldsymbol{I}_y$, and $\boldsymbol{M}_{x,y}$ are utilized to modulate the covariance and position of $\mathcal{G}$, adaptively deforming each

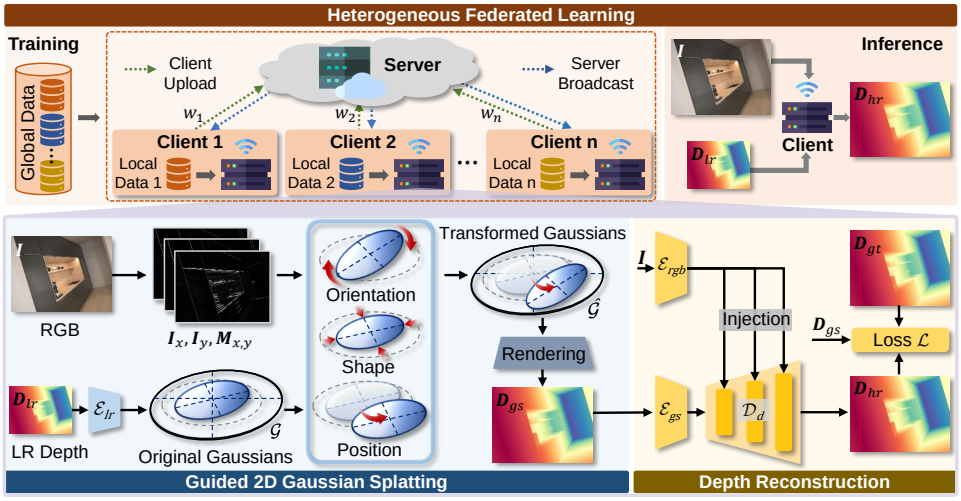


Figure 2: Pipeline of FoundDSR. In federated learning, global data are distributed to independent clients for local optimization, while the server aggregates and redistributes model weights. For each client, LR depth is encoded into initial Gaussian $\mathcal{G}$, which are modulated by RGB gradients for anisotropic structural deformation. The transformed Gaussian $\hat{\mathcal{G}}$ are then rendered into $\boldsymbol{D}_{gs}$. Finally, $\boldsymbol{D}_{gs}$ and RGB $\boldsymbol{I}$ are fed into a depth reconstruction to predict HR depth $\boldsymbol{D}_{HR}$.

Gaussian in orientation, shape, and position along high-frequency structural directions, yielding the transformed Gaussians $\hat{\mathcal{G}}$. Next, $\hat{\mathcal{G}}$ is rendered into an upsampled depth representation $\boldsymbol{D}_{gs}$. Finally, a depth reconstruction branch with dual encoders and a single decoder fuses the RGB $\boldsymbol{I} \in R^{sh \times sw \times 3}$ and $\boldsymbol{D}_{gs}$ to predict the HR depth $\boldsymbol{D}_{hr} \in R^{sh \times sw \times 1}$, where s is the upsampling factor.

3.3 Guided 2D Gaussian Splatting

As illustrated in the blue region of Fig. 2, our guided 2D GS leverages RGB structures as prompts for Gaussian prediction and modulation. For clarity, we omit the Gaussian index m in the following formulations. Specifically, the LR depth $\boldsymbol{D}_{lr}$ is first fed into the network $\mathcal{E}_{lr}$ to predict the original Gaussian primitives $\mathcal{G}$, as defined in Eq. 1, where $\mathcal{E}_{lr}$ consists of multiple MLP heads that separately estimate different Gaussian parameters. Then, we compute the RGB gradients along the x and y directions (denoted as $\boldsymbol{I}_x$ and $\boldsymbol{I}_y$) and their corresponding magnitudes $\boldsymbol{M}_{x,y}$:

$$\boldsymbol{I}_x = \frac{\partial \boldsymbol{I}}{\partial x}, \quad \boldsymbol{I}_y = \frac{\partial \boldsymbol{I}}{\partial y}, \quad \boldsymbol{M}_{x,y} = \sqrt{\boldsymbol{I}_x^2 + \boldsymbol{I}_y^2}. \tag{3}$$

Next, $\boldsymbol{I}_x$, $\boldsymbol{I}_y$, and $\boldsymbol{M}_{x,y}$ are jointly employed to modulate the position $\boldsymbol{\mu}$, rotation angle θ, and covariance matrix $\boldsymbol{\Sigma}$ of $\mathcal{G}$. This strategy enables the orientations, shapes, and positions of the 2D Gaussians to undergo adaptive anisotropic deformation along the gradient directions, facilitating the recovery of high-frequency geometric information in rendered depth. The modulated position is:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{bmatrix}, \begin{cases} \hat{\mu}_x = \mu_x + \frac{\boldsymbol{I}_x}{\boldsymbol{M}_{x,y}+\epsilon} \\ \hat{\mu}_y = \mu_y + \frac{\boldsymbol{I}_y}{\boldsymbol{M}_{x,y}+\epsilon} \end{cases}, \tag{4}$$

where ϵ is a small constant for numerical stability. The modulated angle $\hat{\theta}$ and matrix $\hat{\boldsymbol{R}}$ are:

$$\hat{\boldsymbol{R}} = \begin{bmatrix} \cos(\hat{\theta}) & -\sin(\hat{\theta}) \\ \sin(\hat{\theta}) & \cos(\hat{\theta}) \end{bmatrix}, \quad \begin{cases} \phi = \arctan2(\boldsymbol{I}_y, \boldsymbol{I}_x) + \frac{\pi}{2} \\ \hat{\theta} = \theta + \lambda(\phi - \theta) \end{cases}, \tag{5}$$

where λ is a learnable constant controlling the rotation toward the edge-tangent direction. Based on $\hat{\boldsymbol{R}}$, we obtain the covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{R}} \begin{bmatrix} \sigma_x^2(1+\alpha \boldsymbol{M}_{x,y}) & 0 \\ 0 & \frac{\sigma_y^2}{1+\beta \boldsymbol{M}_{x,y}} \end{bmatrix} \hat{\boldsymbol{R}}^\top, \tag{6}$$

where α and β are learnable parameters that scale the Gaussian along tangent and normal directions.

The transformed Gaussian $\hat{\mathcal{G}}$ are formulated as:

$$\hat{\mathcal{G}} = \boldsymbol{c}_d \cdot \frac{1}{2\pi|\hat{\boldsymbol{\Sigma}}|} \exp\left(-\frac{1}{2}\hat{\boldsymbol{d}}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{d}}\right), \quad (7)$$

where the displacement vector $\hat{\boldsymbol{d}} = (x - \hat{\mu}_x, y - \hat{\mu}_y)\top$. $\boldsymbol{c}_d$ denotes the depth coefficient from the prediction head in $\mathcal{E}_{lr}$. Finally, following existing 2D Gaussian rendering methods Ye et al. (2025a); Peng et al. (2026), $\hat{\mathcal{G}}$ is rendered into the upsampled depth representation $\boldsymbol{D}_{gs}$. Fig. 3 visualizes the Gaussian primitives predicted with and without our RGB-guided strategy. The results show that our method effectively modulates the positions, shapes, and orientations of the Gaussian primitives, enabling them to better align with the depth structures.

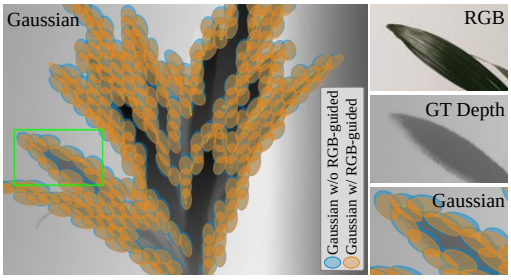


Figure 3: Visualization of Gaussian primitive modulation with and without RGB guidance.

3.4 Depth Reconstruction

As shown in the yellow region of Fig. 2, our depth reconstruction branch aligns with the prevailing paradigm of most depth foundation models Yang et al. (2024a;b); Lin et al. (2025), adopting a simple yet effective encoder–decoder network to reconstruct accurate HR depth $\boldsymbol{D}_{hr}$ from the RGB guidance $\boldsymbol{I}$ and the $\boldsymbol{D}_{gs}$ predicted by the guided 2D GS. Without relying on intricate module designs, we first employ two DINOv2 Oquab et al. (2023) encoders, $\mathcal{E}_{rgb}$ and $\mathcal{E}_{gs}$, to extract RGB features and depth features. Then, a DPT-based decoder Ranftl et al. (2021) $\mathcal{D}_d$ is introduced, which aggregates multi-scale RGB-D features from the encoders through a straightforward addition operation, progressively recovering spatial details and ultimately yielding $\boldsymbol{D}_{hr}$.

In this work, following prior foundation models Lin et al. (2025); Yu et al. (2026b), we adopt the ViT-L variant of DINOv2 as the default encoders for both $\mathcal{E}_{rgb}$ and $\mathcal{E}_{gs}$ in our FoundDSR. To reduce computational cost while preserving competitive generalization, we further introduce a lightweight FoundDSR-S, which replaces ViT-L with ViT-S and keeps the remaining architecture unchanged.

3.5 Heterogeneous Federated Learning

As depicted in the orange region of Fig. 2, we introduce a heterogeneous federated learning strategy to mitigate optimization interference caused by distribution discrepancies across diverse data sources. Specifically, the large-scale global training data are partitioned by source into multiple local subsets, each of which is used to independently optimize a client. All clients share the same DSR architecture, consisting of our proposed guided 2D GS and depth reconstruction network, and follow an identical training strategy to ensure consistent local optimization. In each communication round, the model parameters learned by individual clients are uploaded to the server for adaptive aggregation Shi et al. (2025). The aggregated parameters are then redistributed to all clients, enabling iterative refinement and collaborative training across heterogeneous data sources.

Algorithm 1: Model Parameter Aggregation

Input: Global model parameters ψ^t at round t, local model parameters $\{\hat{\psi}_i^t\}_{i=1}^n$ from n clients
Output: Updated global model parameters ψ^{t+1}

Step 1: Compute client updates
for each client $i = 1$ to n **do**
  $\Delta_i^t \leftarrow \hat{\psi}_i^t - \psi^t$
end for
Step 2: Compute average update direction
$\bar{\Delta}^t \leftarrow \frac{1}{n}\sum_{i=1}^n \Delta_i^t$
Step 3: Compute directional consistency and aggregation weights
for each client $i = 1$ to n **do**
  $c_i^t \leftarrow \frac{\Delta_i^t \cdot \bar{\Delta}^t}{\|\Delta_i^t\|\|\bar{\Delta}^t\|}$ {Direction consistency of client parameter updates}
  $w_i^t \leftarrow \frac{\exp(c_i^t/\tau)}{\sum_{j=1}^n \exp(c_j^t/\tau)}$ {Aggregation weight with temperature coefficient τ}
end for
Step 4: Aggregate global model
$\psi^{t+1} \leftarrow \sum_{i=1}^n w_i^t \hat{\psi}_i^t$
return ψ^{t+1}

Alg. 1 outlines the aggregation strategy for client model parameters. At round t, given the global model parameters ψ^t and the locally trained parameters $\{\hat{\psi}_i^t\}_{i=1}^n$ from n clients (where n equals the number of data sources, set to 4 by default), we first compute each client

update, yielding Δ_i^t. The average update $\bar{\Delta}^t$ is then estimated based on all Δ_i^t. To improve update stability, we further evaluate the consistency between each client's update direction and $\bar{\Delta}^t$, producing consistency scores c_i^t. These scores are subsequently used to assign aggregation weights w_i^t to each client, ensuring that clients whose update directions are better aligned with $\bar{\Delta}^t$ receive higher weights, thereby effectively improving aggregation stability. Finally, the aggregated parameters ψ^{t+1} are broadcast to all clients for the next communication round.

3.6 Loss Function

To optimize our FoundDSR, we introduce a training loss $\mathcal{L}$ composed of a reconstruction loss $\mathcal{L}_{rec}$ and a Gaussian loss $\mathcal{L}_{gs}$. The former supervises each client to recover the final HR depth $\boldsymbol{D}_{hr}$, while the latter constrains the guided 2D Gaussian Splatting process to preserve structural fidelity in the rendered depth representation $\boldsymbol{D}_{gs}$:

$$\mathcal{L} = \underbrace{\textstyle\sum_{q\in\mathbb{Q}} ||\boldsymbol{D}_{gt}^q - \boldsymbol{D}_{hr}^q||_1}_{\text{Reconstruction loss } \mathcal{L}_{rec}} + \underbrace{\rho\left(\textstyle\sum_{q\in\mathbb{Q}} ||\boldsymbol{D}_{gt}^q - \boldsymbol{D}_{gs}^q||_1 + \left(1 - \frac{\sigma_{gs,gt} + C}{\sigma_{gs}\sigma_{gt} + C}\right)\right)}_{\text{Gaussian loss } \mathcal{L}_{gs}}, \tag{8}$$

where $\boldsymbol{D}_{gt}$ represents the ground-truth (GT) depth, and $\mathbb{Q}$ is the valid pixel set of $\boldsymbol{D}_{gt}$. $||\cdot||_1$ denotes the L_1 norm. $\rho = 0.1$ is a balancing coefficient. σ_{gs} and σ_{gt} are the standard deviations of $\boldsymbol{D}_{gs}$ and $\boldsymbol{D}_{gt}$, respectively, while $\sigma_{gs,gt}$ denotes their covariance. $C = 10^{-6}$ is used for numerical stability.

4 Experiments

4.1 Experimental Setups

Datasets. To train FoundDSR, we construct a large-scale dataset of approximately 1.4 million RGB-D pairs from DIML Cho et al. (2021); Kim et al. (2016; 2017; 2018), Hypersim Roberts et al. (2021), DynamicReplica Karaev et al. (2023), and LingBot-Depth Tan et al. (2026). The trained model is directly evaluated on common benchmarks without any fine-tuning, including 449 pairs NYU-v2 Silberman et al. (2012), 6 pairs Lu Lu et al. (2014), 30 pairs Middlebury Hirschmuller & Scharstein (2007); Scharstein & Pal (2007), 100 pairs iBims-1 Koch et al. (2018), 405 pairs RGB-D-D He et al. (2021), and 560 pairs TOFDSR Yan et al. (2025a). For synthetic data, we follow the standard protocol in prior DSR works Zhao et al. (2022); Wang et al. (2025) that synthesize LR depth from GT depth using bicubic downsampling. For real-world scenarios, the RGB-D-D He et al. (2021) and TOFDSR Yan et al. (2025a) datasets also provide real-world LR depth captured by ToF cameras, which is subject to inherently more complex degradations than synthetic counterparts.

Implementation Details. Following prior works Kim et al. (2021); Zhao et al. (2022), we use root mean square error (RMSE) in centimeters as the default evaluation metric, *with additional metrics reported in the appendix*. During training, FoundDSR is optimized using AdamW Loshchilov & Hutter (2017), with an initial learning rate of 5×10^{-6} for the encoder and 5×10^{-5} for other layers. For data augmentation, we first apply random cropping, random horizontal and vertical flipping, and random 90-degree rotations. Then, we randomly sample scaling factors from $[3, 11]$, providing diverse scale supervision while leaving larger factors for evaluating extrapolation ability. Gaussian noise (with zero mean and a standard deviation of 0.07) is injected into a randomly selected 1% of the training samples. The entire training process is conducted on 4 NVIDIA RTX 4090 GPUs.

4.2 Comparison with the State-of-the-Art

We compare FoundDSR with advanced DSR methods, including FDSR He et al. (2021), DCT-Net Zhao et al. (2022), DORNet Wang et al. (2025), C2PD Kang et al. (2025), and DuCos Yan et al. (2025b), as well as depth foundation models using low-quality depth inputs, such as PromptDA Lin et al. (2025), Marigold-DC Viola et al. (2025), InfiniDepth Yu et al. (2026a), and LDCM Yu et al. (2026b). Following prior works Lin et al. (2025); Yu et al. (2026b), foundation models are evaluated using their released weights, while DSR methods are retrained on the same data as FoundDSR.

Comparisons on Synthetic Dataset. Tab. 1 presents zero-shot comparisons over 14 integer and non-integer scales. The results confirm that our method exhibits strong zero-shot generaliza-

Table 1: Zero-shot comparisons on 5 synthetic benchmarks under integer and non-integer scales. All foundation models use their large variants. The best and second-best results are highlighted.

Methods	Datasets	×2.5	×3.0	×4.0	×5.7	×6.0	×7.1	×8.0	×9.3	×11.0	×12.5	×13.0	×14.9	×16.0	×19.3
PromptDA	NYU-v2	17.34	17.06	17.64	17.57	17.61	17.58	18.44	18.01	18.36	18.56	18.66	18.49	19.24	19.00
Marigold-DC		9.12	9.15	9.26	9.48	9.53	9.80	10.01	10.30	10.90	11.38	11.55	12.13	12.53	13.79
InfiniDepth		8.91	8.73	8.70	8.68	8.71	9.02	9.15	9.54	10.15	10.46	10.86	11.20	11.65	13.56
LDCM		4.54	4.57	4.95	6.15	6.50	7.04	7.58	8.44	9.18	9.89	10.28	11.12	11.52	13.12
FDSR		3.90	3.64	4.03	3.85	4.08	4.28	4.99	5.13	5.60	6.30	6.82	9.00	9.98	12.29
DCTNet		3.33	3.41	3.54	4.05	4.17	4.64	5.01	5.65	6.57	7.56	7.93	9.22	9.98	11.97
DORNet		3.82	3.71	3.92	4.05	4.14	4.21	4.80	5.28	5.75	6.69	7.30	8.64	9.95	12.23
C2PD		3.51	3.43	3.73	3.67	3.99	4.19	4.76	5.02	5.67	6.09	7.03	7.96	10.24	12.62
DuCos		3.38	3.17	3.33	3.93	4.04	4.47	4.80	5.33	5.91	6.36	6.64	7.87	8.80	11.46
FoundDSR-S		2.23	2.19	2.31	2.84	2.95	3.32	3.60	4.03	4.57	5.08	5.33	6.21	6.85	8.79
FoundDSR		1.95	1.95	2.17	2.71	2.81	3.16	3.42	3.84	4.33	4.75	4.96	5.70	6.21	7.91
PromptDA	Lu	22.12	22.20	22.33	22.58	22.58	22.86	23.02	23.03	23.53	23.49	23.72	23.90	23.37	23.75
Marigold-DC		36.74	36.76	36.47	35.82	36.11	35.01	36.51	27.42	37.30	33.88	34.25	26.18	36.11	35.47
LDCM		2.49	2.18	2.73	3.92	4.19	4.64	5.14	5.80	6.21	6.67	6.94	7.59	7.65	8.68
FDSR		2.03	1.78	2.09	2.07	2.14	2.49	3.09	3.34	4.04	4.70	5.04	6.21	6.35	8.01
DCTNet		2.02	1.99	1.99	2.22	2.28	2.64	2.98	3.57	4.38	5.07	5.35	6.23	6.59	7.82
DORNet		2.39	2.39	2.41	2.57	2.65	2.94	3.16	3.65	4.40	4.98	5.19	5.88	6.35	7.76
C2PD		2.84	2.42	2.54	2.41	2.92	2.96	3.91	3.72	4.54	4.87	5.27	5.74	7.42	8.87
DuCos		2.24	2.15	2.17	2.32	2.40	2.56	2.76	3.15	3.92	4.32	4.57	5.46	5.99	7.71
FoundDSR-S		1.09	1.03	1.03	1.25	1.32	1.49	1.69	1.99	2.51	3.02	3.27	4.05	4.57	6.00
FoundDSR		0.94	0.90	0.93	1.17	1.22	1.46	1.61	1.92	2.40	2.85	3.05	3.72	4.09	5.37
PromptDA	Middlebury	14.57	14.70	14.61	15.06	14.98	15.24	15.16	15.23	15.59	15.79	15.85	15.49	16.13	16.22
Marigold-DC		5.08	5.56	6.01	5.97	5.72	5.92	6.10	6.17	6.48	6.97	6.47	7.12	7.44	8.22
LDCM		2.26	2.17	2.34	3.30	3.44	3.93	4.30	4.73	5.19	5.62	5.65	6.16	6.39	7.15
FDSR		1.61	1.64	1.87	2.01	2.09	2.24	2.54	2.78	3.15	3.77	3.93	4.98	5.53	6.66
DCTNet		1.79	1.82	1.90	2.22	2.27	2.56	2.82	3.22	3.72	4.36	4.48	5.21	5.61	6.63
DORNet		1.99	1.99	2.01	2.16	2.19	2.37	2.58	2.85	3.18	3.65	3.72	4.39	4.76	6.13
C2PD		1.78	1.86	1.97	2.01	2.10	2.21	2.41	2.64	3.04	3.30	3.69	4.35	5.39	6.81
DuCos		1.87	1.77	1.81	1.99	2.05	2.22	2.40	2.67	2.97	3.16	3.38	3.96	4.54	6.07
FoundDSR-S		1.26	1.25	1.31	1.50	1.52	1.70	1.85	2.12	2.37	2.63	2.76	3.18	3.57	4.68
FoundDSR		0.99	0.98	1.06	1.26	1.29	1.48	1.62	1.87	2.13	2.41	2.50	2.95	3.35	4.37
PromptDA	RGB-D-D	4.89	4.89	4.93	5.06	5.06	5.17	5.24	5.33	5.39	5.46	5.46	5.47	5.54	5.53
Marigold-DC		3.15	3.19	3.24	3.38	3.44	3.58	3.73	3.94	4.19	4.49	4.61	4.95	5.22	5.75
InfiniDepth		3.61	3.55	3.55	3.46	3.50	3.57	3.67	3.80	4.03	4.22	4.39	4.57	4.85	5.60
LDCM		2.14	2.08	2.40	2.73	2.82	3.14	3.54	3.84	4.05	4.41	4.51	4.94	5.28	5.79
FDSR		1.58	1.68	1.89	1.90	1.98	2.08	2.39	2.57	2.74	3.06	3.23	3.96	4.48	5.24
DCTNet		1.80	1.83	1.94	2.25	2.32	2.55	2.73	2.96	3.27	3.57	3.69	4.12	4.42	5.13
DORNet		1.82	1.85	1.93	2.19	2.25	2.47	2.66	2.92	3.25	3.50	3.59	3.94	4.27	5.05
C2PD		1.61	1.77	2.01	1.90	2.10	2.17	2.54	2.65	2.87	3.08	3.43	3.94	4.69	5.55
DuCos		1.54	1.51	1.62	1.89	1.94	2.13	2.29	2.53	2.88	3.14	3.23	3.57	3.90	4.85
FoundDSR-S		1.26	1.25	1.30	1.52	1.56	1.74	1.89	2.08	2.32	2.55	2.59	2.88	3.11	3.78
FoundDSR		1.08	1.08	1.17	1.41	1.46	1.63	1.78	1.98	2.21	2.43	2.46	2.71	2.90	3.43
PromptDA	iBims-1	37.84	38.21	38.40	38.93	39.39	38.70	40.08	40.95	41.85	42.65	43.06	43.12	44.17	45.10
Marigold-DC		39.83	40.27	37.83	37.04	37.08	37.12	36.74	35.83	37.45	38.33	38.27	38.17	39.65	41.46
LDCM		23.10	21.65	21.93	24.28	25.02	26.06	27.13	29.18	30.31	31.73	32.37	33.81	34.53	37.18
FDSR		14.82	15.63	16.67	19.15	19.41	20.79	21.83	23.20	25.06	27.23	28.16	31.07	32.91	36.10
DCTNet		13.80	14.26	16.15	19.36	19.86	21.62	22.67	24.37	26.25	28.11	28.80	30.95	32.32	35.64
DORNet		13.01	13.61	15.36	18.70	19.23	21.09	22.20	24.20	26.14	27.90	28.59	30.64	32.10	35.46
C2PD		12.38	14.02	14.87	16.93	18.05	19.54	20.80	22.22	24.77	26.42	27.97	29.57	33.11	36.94
DuCos		13.29	13.23	15.27	18.05	18.51	20.31	21.48	23.43	25.31	26.98	27.74	29.95	31.57	35.38
FoundDSR-S		12.80	12.76	13.47	15.76	16.18	17.72	18.84	20.72	22.94	24.81	25.51	27.28	28.55	31.77
FoundDSR		11.19	11.33	12.52	15.11	15.58	17.03	18.19	20.10	22.21	24.00	24.70	26.48	27.70	30.57

tion. Although recent foundation models also use low-quality depth as a prompt, they primarily regress depth from RGB rather than learning the LR-to-HR depth mapping. Consequently, FoundDSR achieves the lowest RMSE, outperforming suboptimal LDCM by 46.09%, 46.54%, 47.57%, 45.08%, and 19.78% on the NYU-v2, Lu, Middlebury, RGB-D-D, and iBims-1, respectively (×16.0). On the low-light Lu, Marigold-DC degrades substantially due to its strong reliance on RGB inherited from its depth estimation backbone, which become unreliable and deviate from the pretraining distribution under poor illumination. Moreover, compared with advanced DSR methods, FoundDSR shows increasing gains at larger scales, reducing the average RMSE over DCTNet, DORNet, C2PD, and DuCos by 29.38% and 36.26% on NYU-v2 at ×8.0 and ×16.0, respectively.

Fig. 4 shows the visual results, confirming that our method effectively restores degraded depth structures while preserving fine geometric information. Furthermore, Fig. 5 presents zero-shot comparisons across arbitrary upsampling factors. As the scale increases, all competing methods exhibit higher RMSE, whereas our FoundDSR consistently achieves the lowest error with the most gradual performance degradation, demonstrating strong generalization.

Comparisons on Real-World Dataset. Tab. 2 reports the comparisons on real-world RGB-D-D and TOFDSR datasets with unknown degradations. The results show that FoundDSR exhibits superior

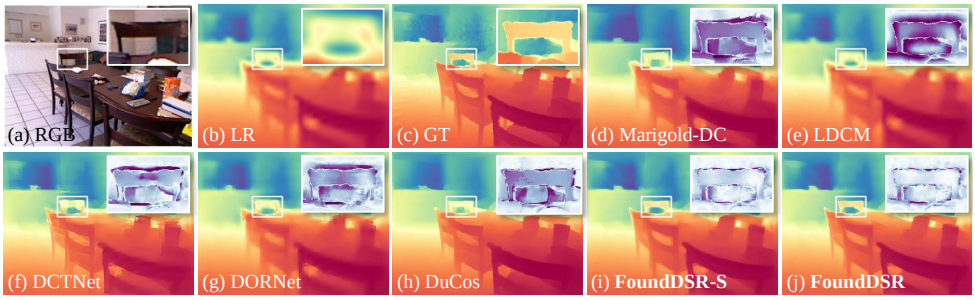


Figure 4: Visual comparisons on the $\times 16.0$ NYU-v2 dataset. The zoomed-in patches in (d)–(j) highlight the error maps within the white boxes, where darker purple shades indicate higher errors.

Table 2: Zero-shot comparisons on the real-world RGB-D-D and TOFDSR datasets.

RMSE	PromptDA	Marigold-DC	InfiniDepth	LDCM	FDSR	DCTNet	DORNet	C2PD	DuCos	**FoundDSR-S**	**FoundDSR**
RGB-D-D	7.06	6.85	6.87	7.65	7.41	7.82	7.49	7.26	7.44	6.93	6.68
TOFDSR	8.79	7.80	8.48	8.22	7.23	7.45	7.34	7.23	7.42	7.05	6.92

Table 3: Zero-shot comparisons on multiple $\times 4.0$ benchmarks with noise levels from 0.02 to 0.04.

Methods	NYU-v2			Lu			Middlebury			RGB-D-D			iBims-1		
	0.02	0.03	0.04	0.02	0.03	0.04	0.02	0.03	0.04	0.02	0.03	0.04	0.02	0.03	0.04
PromptDA	17.42	17.34	17.36	22.49	22.42	22.40	15.12	15.52	11.93	5.23	5.35	5.53	38.83	39.06	39.48
Marigold-DC	8.71	8.79	8.97	36.42	36.04	35.94	5.75	5.83	5.65	2.93	2.93	2.96	33.90	33.80	33.89
LDCM	6.00	7.40	9.24	4.64	6.24	7.99	4.33	5.88	7.46	3.00	3.62	4.34	23.76	26.85	30.44
FDSR	5.82	6.56	7.05	3.18	3.66	3.96	3.00	3.49	3.86	2.26	2.56	2.92	18.61	19.71	20.62
DCTNet	6.24	6.94	7.81	3.32	3.94	4.30	3.35	3.96	4.22	2.58	2.95	3.44	20.08	21.83	23.36
DORNet	5.79	6.81	7.96	3.95	4.69	5.45	3.49	4.13	4.79	2.65	3.01	3.36	19.56	21.67	23.79
C2PD	5.53	6.42	7.22	2.97	3.59	4.08	3.17	3.38	3.91	2.40	2.74	3.15	17.82	18.94	20.17
DuCos	5.37	6.61	6.98	2.88	3.34	4.12	3.06	3.33	3.84	2.17	2.53	2.81	17.89	19.57	20.10
FoundDSR-S	4.44	5.13	5.73	2.41	2.69	3.15	2.46	2.93	3.26	1.88	2.19	2.44	16.52	17.77	18.94
FoundDSR	4.21	4.91	5.48	2.12	2.59	3.07	2.28	2.79	3.20	1.85	2.15	2.40	16.26	17.58	18.63

zero-shot generalization over both depth foundation models and existing DSR methods, outperforming the second-best method by $0.17cm$ and $0.31cm$ in RMSE on RGB-D-D and TOFDSR, respectively. *Please see our appendix for visual results.*

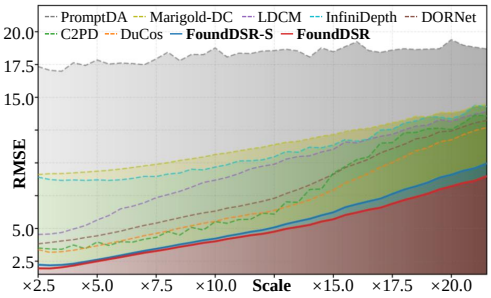


Figure 5: Zero-shot comparisons of continuous upsampling factors with prior foundation models and DSR methods on NYU-v2.

Robustness to Noise. Following prior works Kim et al. (2021); Wang et al. (2025), we additionally corrupt the LR depth with Gaussian noise to simulate more challenging degradation conditions. As reported in Tab. 3, we evaluate all methods on five benchmarks under zero-mean Gaussian noise with standard deviations of 0.02, 0.03, and 0.04. FoundDSR consistently shows strong robustness across all noise levels, reducing the average RMSE over the second-best DuCos by 14.82%, 15.15%, and 13.39% across all datasets at noise levels of 0.02, 0.03, and 0.04, respectively. *Qualitative comparisons are provided in the appendix.*

Complexity Analysis. Tab. 4 compares model complexity and computational efficiency. As a foundation model, FoundDSR inevitably involves a relatively large number of parameters. Nevertheless, our lightweight FoundDSR-S remains competitive with existing DSR methods such as C2PD and DuCos in parameter count and GPU memory usage, while requiring fewer FLOPs, lower inference latency, and substantially improving reconstruction accuracy. Compared with previous foundation models, FoundDSR achieves a more favorable trade-off between computational cost and reconstruction accuracy. Notably, Marigold-DC is built upon a latent diffusion model and requires a 50-step denoising process, resulting in significantly higher inference costs than other approaches. For consistency with other methods, we report only the FLOPs of a forward pass for Marigold-DC. Compared with the second-best foundation model LDCM, FoundDSR reduces inference time by 60.87% while improving reconstruction performance by 52.67%.

Table 4: Complexity on $\times 16.0$ NYU-v2, evaluated on a single NVIDIA RTX 4090 GPU.

Methods	Param. (M)	Memo. (G)	FLOPs (G)	Time (s)	RMSE
C2PD	62.05	1.75	414.75	0.061	8.80
DuCos	34.38	9.02	2927.49	0.025	10.24
Marigold-DC	949.58	9.83	94753.76	14.02	13.79
InfiniDepth	358.15	23.11	16446.19	0.081	13.56
LDCM	238.62	3.01	903.07	0.069	13.12
FoundDSR-S	51.85	1.08	90.76	0.016	6.85
FoundDSR	657.44	3.65	1248.86	0.027	6.21

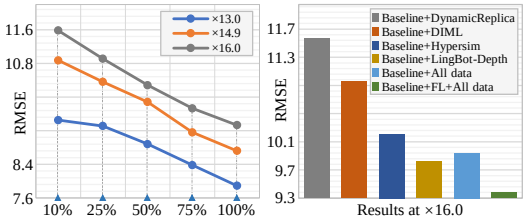
Table 5: Ablation study of position (Pos.) and covariance (Cov.) modulation in our guided 2D GS strategy on $\times 13.0$, $\times 14.9$, and $\times 16.0$ NYU-v2, with time measured at the $\times 16.0$ scale.

	2D GS	Pos.	Cov.	Param. (M)	Time (s)	$\times 13.0$	$\times 14.9$	$\times 16.0$
(a)				48.52	0.013	6.75	8.48	9.42
(b)	✓			51.04	0.015	5.63	6.56	7.26
(c)	✓	✓		51.85	0.016	5.60	6.50	7.18
(d)	✓		✓	51.85	0.016	5.52	6.39	7.02
(e)	✓	✓	✓	51.85	0.016	5.33	6.21	6.85

4.3 Ablation Study

Guided 2D Gaussian Splatting. Tab. 5 presents the ablation study of our guided 2D GS, where baseline (a) removes the entire guided 2D GS. *For efficiency, all ablations are conducted using our ViT-S variant.* Compared with (a), standard 2D GS (b) consistently improves performance across different scales, validating the effectiveness of continuous Gaussian representations for arbitrary-scale depth upsampling. Building upon (b), (c) and (d) further enhance reconstruction quality by incorporating our RGB-guided strategy to modulate the positions or covariances of Gaussian primitives, with only marginal computational overhead. When both are jointly modulated, our method (e) achieves the best performance, reducing RMSE over baseline (a) by $1.42cm$, $2.27cm$, and $2.57cm$ on NYU-v2 at $\times 13.0$, $\times 14.9$, and $\times 16.0$, respectively.

Training Data Scaling. To evaluate the scalability of our method, we construct training subsets containing 10% to 100% of the full dataset by random sampling, while maintaining the same sampling ratio across all data sources. As shown in Fig. 6(a), increasing the data consistently improves performance across different upsampling factors, demonstrating the strong data-scaling capability of FoundDSR. Compared with using only 10% of the data, training on the full dataset reduces RMSE by $1.56cm$, $2.15cm$, and $2.25cm$ at $\times 13.0$, $\times 14.9$, and $\times 16.0$, respectively.



(a) Training data scaling (b) Federated learning (FL)

Figure 6: Ablation study on (a) training data scaling and (b) heterogeneous federated learning. RMSE is averaged over five synthetic datasets.

Federated Learning. Fig. 6(b) presents the ablation study of federated learning, where the baseline removes all federated learning components. These results indicate that directly mixing all data sources fails to achieve proportional gains from the increased data scale and even underperforms single-source training on LingBot-Depth, highlighting the optimization interference caused by distribution discrepancies. With our heterogeneous federated learning strategy, our method effectively alleviates such interference and achieves the best performance. Specifically, it reduces the average RMSE by $0.56cm$ compared with direct mixed-data training. These quantitative comparisons demonstrate the scalability of our method to larger-scale and more diverse data sources.

5 Conclusion

In this paper, we propose FoundDSR, the first foundation model for depth super-resolution, trained on million-scale heterogeneous data to substantially enhance zero-shot generalization to out-of-distribution data and arbitrary upsampling factors. To facilitate continuous and arbitrary-scale depth upsampling, we introduce an RGB-guided 2D Gaussian Splatting that leverages high-resolution RGB prompts to modulate the position and covariance of each Gaussian primitive, enabling anisotropic deformation in position, shape, and orientation along structural directions. Furthermore, FoundDSR incorporates heterogeneous federated learning that effectively mitigates optimization instability and dominant-source bias caused by distribution gaps across large-scale non-uniform data sources, thereby improving scalability to larger and more diverse training data. Extensive zero-shot experiments under synthetic, real-world, arbitrary-scale, and noisy settings consistently demonstrate the strong generalization of FoundDSR to out-of-distribution data.

AI USE STATEMENT

In this work, we used generative AI tools for translation, language polishing, expression refinement, and the detection of typographical and grammatical errors. 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.

ETHICS STATEMENT

Our study focuses on depth super-resolution, a fundamental problem in computer vision. The training and evaluation are conducted using publicly available RGB-D datasets from diverse sources, without collecting or processing personally identifiable or sensitive information. This research does not involve human subjects and has been conducted in accordance with standard research ethics and responsible research practices.

REPRODUCIBILITY STATEMENT

To facilitate verification and extension of our work, we provide detailed descriptions of the model architecture, training settings, and evaluation protocols in the paper and supplementary materials. The complete source code and trained model weights will be made publicly available after publication to support reproducibility and further research.

REFERENCES

Du Chen, Liyi Chen, Zhengqiang Zhang, and Lei Zhang. Generalized and efficient 2d gaussian splatting for arbitrary-scale super-resolution. In *ICCV*, pp. 26435–26445, 2025.

Xuanhong Chen, Hang Wang, Jialiang Chen, Kairui Feng, Jinfan Liu, Xiaohang Wang, Weimin Zhang, and Bingbing Ni. Intrinsic phase-preserving networks for depth super resolution. In *AAAI*, pp. 1210–1218, 2024.

Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset. *Expert Systems with Applications*, 178: 114877, 2021.

Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (10):3333–3348, 2020.

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 *CVPR*, pp. 9229–9238, 2021.

Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, pp. 1–8. IEEE, 2007.

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*, 46(12):10579–10596, 2024.

Jiahui Kang, Qing Cai, Runqing Tan, Yimei Liu, and Zhi Liu. C2pd: Continuity-constrained pixel-wise deformation for guided depth super-resolution. In *AAAI*, pp. 4212–4220, 2025.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *CVPR*, pp. 13229–13239, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pp. 9492–9502, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

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

Sunok Kim, Dongbo Min, Bumsub Ham, Seungryong Kim, and Kwanghoon Sohn. Deep stereo confidence prediction for depth estimation. In *ICIP*, pp. 992–996, 2017.

Youngjung Kim, Bumsub Ham, Changjae Oh, and Kwanghoon Sohn. Structure selective depth superresolution for rgb-d cameras. *IEEE Transactions on Image Processing*, 25(11):5227–5238, 2016.

Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing*, 27(8):4131–4144, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *ECCV*, pp. 331–348, 2018.

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 *CVPR*, pp. 17070–17080, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, pp. 3390–3397, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Long Peng, Anran Wu, Wenbo Li, Xinjie Zhang, Xueyuan Dai, Xin Di, Haoze Sun, Renjing Pei, Yang Wang, Yang Cao, et al. Pixel to gaussian: Ultra-fast continuous super-resolution with 2d gaussian modeling. In *ICLR*, volume 2026, pp. 92172–92199, 2026.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12179–12188, 2021.

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 *ICCV*, pp. 10912–10922, 2021.

Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *CVPR*, pp. 1–8, 2007.

Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: adaptive optimization of aggregation weights in federated learning using client vectors. In *CVPR*, pp. 30651–30660, 2025.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pp. 746–760. Springer, 2012.

Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang. Channel attention based iterative residual learning for depth map super-resolution. In *CVPR*, pp. 5631–5640, 2020.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Baoli Sun, Xinchen Ye, Baopu Li, Haojie Li, Zhihui Wang, and Rui Xu. Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *CVPR*, pp. 7792–7801, 2021.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, et al. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In *ICCV*, pp. 5359–5370, 2025.

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.

Xiaohang Wang, Xuanhong Chen, Bingbing Ni, Zhengyan Tong, and Hang Wang. Learning continuous depth representation via geometric spatial aggregator. In *AAAI*, pp. 2698–2706, 2023b.

Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *AAAI*, pp. 5823–5831, 2024.

Zhengxue Wang, Zhiqiang Yan, Jinshan Pan, Guangwei Gao, Kai Zhang, and Jian Yang. Dornet: A degradation oriented and regularized network for blind depth super-resolution. In *CVPR*, pp. 15813–15822, 2025.

Zhengxue Wang, Lina Liu, Xiang Li, Zhiqiang Yan, and Jian Yang. Graph-aware prompting network for generalizable depth super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026a.

Zhengxue Wang, Yuan Wu, Xiang Li, Zhiqiang Yan, and Jian Yang. Spatiotemporal difference network for video depth super-resolution. In *AAAI*, pp. 10403–10411, 2026b.

Zhengxue Wang, Zhiqiang Yan, Ming-Hsuan Yang, Jinshan Pan, Guangwei Gao, Ying Tai, and Jian Yang. Scene prior filtering for depth super-resolution: Zhengxue wang et al. *International Journal of Computer Vision*, 134(5):251, 2026c.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, pp. 16453–16463, 2025.

Zhiqiang Yan, Kun Wang, Xiang Li, Guangwei Gao, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry aware depth completion and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Zhiqiang Yan, Zhengxue Wang, Haoye Dong, Jun Li, Jian Yang, and Gim Hee Lee. Ducos: Duality constrained depth super-resolution via foundation model. In *ICCV*, pp. 8361–8371, 2025b.

Zhiqiang Yan, Jianhao Jiao, Zhengxue Wang, and Gim Hee Lee. Event-driven dynamic scene depth completion. *NeurIPS*, 38:142971–142993, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pp. 10371–10381, 2024a.

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

Yuxiang Yang, Qi Cao, Jing Zhang, and Dacheng Tao. Codon: On orchestrating cross-domain attentions for depth super-resolution. *International Journal of Computer Vision*, 130(2):267–284, 2022.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025a.

Xinchen Ye, Aokai Zhang, and Rui Xu. Semantics-driven contrastive learning for real-world depth super resolution. In *ACMMM*, pp. 3085–3093, 2025b.

Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. Infinidepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields. *arXiv preprint arXiv:2601.03252*, 2026a.

Zhu Yu, Zhengyi Zhao, Runmin Zhang, Lingteng Qiu, Kejie Qiu, Yisheng He, Siyu Zhu, Zilong Dong, Si-Yuan Cao, and Hui-Liang Shen. Large depth completion model from sparse observations. *arXiv preprint arXiv:2605.30115*, 2026b.

Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Structure flow-guided network for real depth super-resolution. In *AAAI*, pp. 3340–3348, 2023.

Zhaojie Zeng, Yuesong Wang, Tao Guan, Chao Yang, and Lili Ju. Instant gaussianimage: A generalizable and self-adaptive image representation via 2d gaussian splatting. In *ICCV*, pp. 27896–27905, 2025.

Xinjie Zhang, Xingtong Ge, Tongda Xu, Dailan He, Yan Wang, Hongwei Qin, Guo Lu, Jing Geng, and Jun Zhang. Gaussianimage: 1000 fps image representation and compression by 2d gaussian splatting. In *ECCV*, pp. 327–345. Springer, 2024.

Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pp. 5697–5707, 2022.

Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. In *ICCV*, pp. 12547–12558, 2023.

Huan Zheng, Wencheng Han, and Jianbing Shen. Decoupling fine detail and global geometry for compressed depth map super-resolution. In *CVPR*, pp. 951–960, 2025.

Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, and Xiangyang Ji. Deep attentional guided image filtering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, and Shiqi Wang. Dual-level cross-modality neural architecture search for guided image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zhiwei Zhong, Peilin Chen, Qiangqiang Shen, Bo Li, and Shiqi Wang. Dual graph regularized deep unfolding network for guided depth map super-resolution. In *CVPR*, pp. 16322–16332, 2026.

A APPENDIX

A.1 DATASET DETAILS

As shown in Tab. 6, we collect 4 publicly available RGB-D datasets to train our FoundDSR, consisting of 1 real-world and 3 synthetic datasets, covering diverse scenes.

Table 6: Overview of the training datasets.

Dataset	Type	Resolution	Statistic
DIML	Real-World	1344×756	1.6K
Hypersim	Synthetic	1024×768	46.1K
DynamicReplica	Synthetic	1280×720	336.9K
LingBot-Depth	Synthetic	1280×960	1.0M
Total	-	-	1.4M

DIML Dataset. DIML Cho et al. (2021); Kim et al. (2016; 2017; 2018) is captured in real-world environments using Kinect v2 and ZED stereo cameras. We use only its indoor subset, which contains approximately 1.6K training samples spanning 18 indoor scenes.

Hypersim Dataset. Hypersim Roberts et al. (2021) is generated from a large repository of synthetic scenes created by professional artists, containing 77.4K samples across 461 indoor scenes. We remove RGB-D pairs with invalid depth values, leaving approximately 46.1K samples for training.

DynamicReplica Dataset. DynamicReplica Karaev et al. (2023) is built upon Replica Straub et al. (2019) by introducing moving humans and animals into scanned indoor environments, yielding richer and more diverse dynamic scenes. It contains 524 synthetic video sequences.

LingBot-Depth Dataset. LingBot-Depth Tan et al. (2026) contains large-scale real-world and synthetic RGB-D data. Since the real-world subset includes a considerable number of invalid depth samples, we only adopt the synthetic subset. It simulates real-world sensing imperfections and renders paired RGB-D data in Blender using self-hosted 3D assets, covering 442 indoor scenes.

A.2 METRICS

In the main paper, we follow common practice Zhao et al. (2022); Wang et al. (2025) and adopt root mean square error (RMSE) as the primary evaluation metric. For a more comprehensive evaluation, additional quantitative comparisons using mean absolute error (MAE) and $\delta_{1.05}$ are provided in the supplementary material. These metrics are defined as follows:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{|\mathbb{Q}|}\sum \left(\boldsymbol{D}_{gt}^{q} - \boldsymbol{D}_{hr}^{q}\right)^{2}}, \\ \text{MAE} &= \frac{1}{|\mathbb{Q}|}\sum \left|\boldsymbol{D}_{gt}^{q} - \boldsymbol{D}_{hr}^{q}\right|, \\ \delta_{1.05} &= \frac{|\mathbb{S}|}{|\mathbb{Q}|}, \mathbb{S} : max(\frac{\boldsymbol{D}_{gt}^{q}}{\boldsymbol{D}_{hr}^{q}}, \frac{\boldsymbol{D}_{hr}^{q}}{\boldsymbol{D}_{gt}^{q}}) < 1.05, \end{aligned} \tag{9}$$

where $\boldsymbol{D}_{gt}$ and $\boldsymbol{D}_{hr}$ denote the ground-truth depth and the predicted HR depth, respectively. $\mathbb{Q}$ is the valid pixel set of $\boldsymbol{D}_{gt}$, and q is one of the valid pixels.

A.3 COMPARISONS WITH MORE EVALUATION METRICS

Tab. 7 and Tab. 8 further present zero-shot comparisons in terms of MAE and $\delta_{1.05}$ metrics, respectively. The results demonstrate that our FoundDSR consistently exhibits stronger generalization capability than existing DSR methods and foundation models, particularly at large upsampling factors. Compared with the more challenging large-scale settings, smaller upsampling factors require less model capacity to achieve satisfactory performance. Consequently, our lightweight variant, FoundDSR-S, slightly outperforms FoundDSR in some low-scale settings. As the upsampling factor increases, the advantage of the larger FoundDSR becomes increasingly pronounced. For example, at $\times 16.0$, FoundDSR reduces the MAE of the second-best method by 26.54%, 31.16%, 25.81%, 31.58%, and 29.42% on NYU-v2, Lu, Middlebury, RGB-D-D, and iBims-1, respectively, while improving $\delta_{1.05}$ by 2.23, 2.84, 3.95, 1.61, and 3.26 percentage points.

Table 7: Quantitative comparisons on five synthetic benchmarks using the **MAE**↓ (cm) metric.

Methods	Datasets	×2.5	×3.0	×4.0	×5.7	×6.0	×7.1	×8.0	×9.3	×11.0	×12.5	×13.0	×14.9	×16.0	×19.3
PromptDA	NYU-v2	7.58	7.54	7.67	7.74	7.73	7.67	8.04	7.82	8.01	9.16	8.16	8.23	8.50	8.49
Marigold-DC		4.18	4.20	4.27	4.39	4.41	4.56	4.68	4.83	5.14	5.38	5.53	5.79	6.04	6.73
InfiniDepth		3.33	3.37	3.42	3.53	3.59	3.78	3.91	4.60	4.52	4.72	4.94	5.11	5.45	6.53
LDCM		1.57	1.63	1.82	2.32	2.47	2.64	3.01	3.40	3.86	4.25	4.51	4.95	5.24	6.19
FDSR		1.12	1.10	1.19	1.28	1.33	1.48	1.68	1.84	2.10	2.49	2.74	3.70	4.34	5.54
DCTNet		1.26	1.29	1.36	1.58	1.63	1.82	2.01	2.30	2.84	3.36	3.56	4.17	4.62	5.65
DORNet		1.07	1.09	1.11	1.19	1.23	1.33	1.54	1.75	1.86	2.42	2.75	3.47	4.11	5.38
C2PD		1.14	1.16	1.21	1.29	1.37	1.50	1.63	1.82	2.15	2.39	2.82	3.20	4.26	5.78
DuCos		0.86	0.83	0.93	1.19	1.24	1.43	1.57	1.79	2.08	2.36	2.51	3.18	3.73	5.17
FoundDSR-S		0.84	0.82	0.87	1.07	1.11	1.26	1.38	1.56	1.81	2.09	2.22	2.65	3.02	4.00
FoundDSR		0.81	0.79	0.84	1.04	1.08	1.23	1.34	1.51	1.73	1.96	2.07	2.44	2.74	3.59
Marigold-DC	Lu	11.59	11.69	11.53	11.51	11.55	11.28	11.73	8.62	11.78	11.34	11.74	8.90	12.23	12.65
LDCM		0.77	0.77	0.91	1.26	1.37	1.54	1.68	1.98	2.19	2.44	2.56	2.91	2.97	3.56
FDSR		0.60	0.55	0.63	0.70	0.72	0.80	0.91	1.03	1.21	1.45	1.58	2.10	2.37	3.11
DCTNet		0.77	0.79	0.79	0.88	0.90	1.01	1.10	1.31	1.60	1.89	1.98	2.41	2.58	3.22
DORNet		0.57	0.58	0.59	0.68	0.70	0.80	0.88	1.02	1.25	1.46	1.54	1.89	2.12	2.92
C2PD		0.73	0.65	0.74	0.71	0.79	0.84	0.98	1.05	1.27	1.40	1.55	1.76	2.35	3.41
DuCos		0.51	0.50	0.52	0.59	0.62	0.68	0.75	0.86	1.06	1.21	1.29	1.72	1.99	2.89
FoundDSR-S		0.45	0.43	0.44	0.49	0.51	0.56	0.61	0.71	0.84	0.99	1.05	1.34	1.53	2.18
FoundDSR		0.49	0.48	0.48	0.53	0.55	0.59	0.64	0.71	0.83	0.96	1.00	1.24	1.37	1.93
Marigold-DC	Middlebury	2.33	2.41	2.54	2.65	2.59	2.73	2.78	2.93	3.12	3.38	3.28	3.66	3.80	4.33
LDCM		0.93	0.95	1.08	1.45	1.51	1.74	1.93	2.17	2.47	2.75	2.80	3.15	3.31	3.86
FDSR		0.76	0.78	0.86	0.94	0.97	1.04	1.14	1.24	1.42	1.71	1.81	2.43	2.77	3.55
DCTNet		0.93	0.95	0.98	1.09	1.11	1.23	1.35	1.55	1.85	2.20	2.28	2.73	2.97	3.64
DORNet		0.80	0.81	0.83	0.91	0.93	1.01	1.09	1.21	1.38	1.61	1.68	2.11	2.38	3.26
C2PD		0.89	0.92	0.94	0.95	0.98	1.03	1.10	1.18	1.35	1.47	1.64	1.97	2.55	3.63
DuCos		0.73	0.71	0.74	0.84	0.86	0.94	1.01	1.12	1.25	1.37	1.46	1.84	2.17	3.14
FoundDSR-S		0.66	0.65	0.67	0.75	0.76	0.84	0.91	1.01	1.12	1.25	1.30	1.51	1.72	2.35
FoundDSR		0.61	0.61	0.63	0.72	0.73	0.80	0.87	0.96	1.07	1.18	1.23	1.43	1.61	2.18
PromptDA	RGB-D-D	2.75	2.72	2.70	2.71	2.72	2.75	2.80	2.86	2.92	2.97	2.99	3.02	3.05	3.10
Marigold-DC		1.26	1.28	1.31	1.38	1.42	1.49	1.57	1.68	1.84	2.01	2.05	2.27	2.41	2.77
InfiniDepth		1.08	1.09	1.11	1.16	1.21	1.30	1.35	1.46	1.65	1.78	1.90	2.02	2.18	2.72
LDCM		0.62	0.63	0.80	0.94	0.98	1.13	1.32	1.49	1.65	1.87	1.93	2.19	2.39	2.75
FDSR		0.49	0.51	0.58	0.60	0.62	0.66	0.75	0.84	0.94	1.10	1.18	1.56	1.83	2.39
DCTNet		0.65	0.66	0.69	0.77	0.79	0.86	0.94	1.06	1.24	1.43	1.50	1.77	1.94	2.41
DORNet		0.49	0.50	0.53	0.62	0.64	0.72	0.79	0.89	1.03	1.18	1.23	1.47	1.67	2.24
C2PD		0.49	0.52	0.56	0.57	0.62	0.67	0.74	0.82	0.95	1.04	1.21	1.35	1.76	2.46
DuCos		0.40	0.39	0.43	0.53	0.55	0.62	0.68	0.77	0.90	1.02	1.07	1.31	1.52	2.60
FoundDSR-S		0.39	0.38	0.40	0.47	0.48	0.54	0.59	0.67	0.76	0.85	0.88	1.03	1.15	1.54
FoundDSR		0.36	0.36	0.39	0.46	0.47	0.53	0.57	0.64	0.72	0.80	0.83	0.94	1.04	1.34
PromptDA	iBims-1	13.84	14.10	14.14	14.15	14.42	13.78	14.50	14.91	15.45	15.66	15.96	16.05	16.37	17.17
Marigold-DC		14.05	14.26	13.37	13.30	13.29	13.37	13.46	12.93	14.04	14.59	14.60	14.64	15.72	16.86
LDCM		4.78	4.62	5.11	6.45	6.79	7.40	8.01	9.05	9.81	10.64	11.05	11.96	12.36	14.17
FDSR		2.91	3.05	3.21	3.99	4.03	4.51	4.81	5.48	6.36	7.44	8.01	9.70	10.78	13.14
DCTNet		3.26	3.33	3.74	4.68	4.84	5.46	5.87	6.65	7.68	8.76	9.17	10.44	11.19	13.34
DORNet		2.10	2.21	2.65	3.66	3.81	4.44	4.84	5.60	6.55	7.46	7.90	9.17	10.04	12.54
C2PD		2.13	2.38	2.50	3.16	3.43	3.95	4.22	4.91	5.95	6.60	7.51	8.32	10.21	13.25
DuCos		2.18	2.17	2.35	3.08	3.22	3.80	4.18	4.92	5.74	6.57	7.02	8.44	9.45	12.28
FoundDSR-S		2.10	2.10	2.24	2.77	2.87	3.27	3.58	4.12	4.84	5.55	5.88	6.71	7.32	9.41
FoundDSR		2.07	2.10	2.27	2.77	2.88	3.22	3.47	3.96	4.57	5.19	5.47	6.16	6.67	8.38

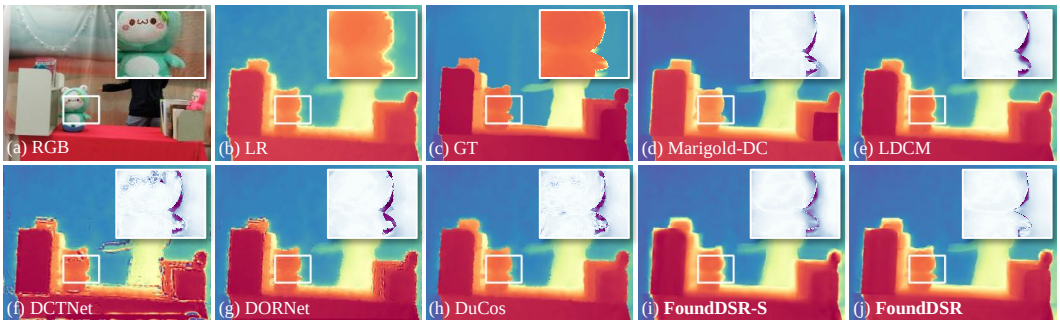


Figure 7: Visual comparisons on the real-world RGB-D-D dataset. The zoomed-in patches in (d)–(j) highlight the error maps within the white boxes, where darker purple shades indicate higher errors.

Table 8: Quantitative comparisons on five synthetic benchmarks using the $\delta_{1.05} \uparrow$ metric.

Methods	Datasets	×2.5	×3.0	×4.0	×5.7	×6.0	×7.1	×8.0	×9.3	×11.0	×12.5	×13.0	×14.9	×16.0	×19.3
PromptDA	NYU-v2	92.39	92.46	92.20	92.13	92.16	92.26	91.57	91.90	91.59	91.38	91.45	91.23	90.66	90.73
Marigold-DC		95.91	95.87	95.69	95.38	95.33	94.97	94.56	94.21	93.47	92.87	92.58	91.85	91.12	89.73
InfiniDepth		94.37	94.21	94.06	93.80	93.59	93.29	92.79	92.74	91.63	91.15	90.81	90.31	89.30	87.52
LDCM		98.88	98.78	98.39	97.43	97.23	96.55	95.94	95.24	94.23	93.39	93.09	92.09	91.28	89.74
FDSR		99.00	99.10	99.00	99.05	98.98	98.88	98.62	98.46	98.15	97.56	97.13	95.04	93.45	90.99
DCTNet		99.19	99.16	99.10	98.83	98.76	98.51	98.25	97.84	96.84	95.66	95.28	93.81	92.62	90.66
DORNet		99.12	99.13	99.09	99.07	99.02	99.04	98.75	98.53	98.22	97.71	97.23	95.10	94.34	91.72
C2PD		99.21	99.21	99.15	99.14	99.06	98.94	98.80	98.60	98.19	97.84	97.11	96.34	93.82	90.46
DuCos		99.26	99.33	99.30	99.13	99.09	98.93	98.80	98.54	98.19	97.81	97.61	96.27	94.95	91.87
FoundDSR-S		99.58	99.60	99.57	99.40	99.36	99.22	99.10	98.93	98.64	98.26	98.05	97.25	96.40	94.43
FoundDSR		99.69	99.69	99.63	99.46	99.43	99.30	99.21	99.04	98.80	98.54	98.39	97.79	97.18	95.51
Marigold-DC	Lu	74.42	74.19	74.74	73.47	73.66	73.83	73.28	78.51	73.28	72.67	73.00	76.63	71.61	69.80
LDCM		98.07	97.98	97.36	95.73	95.22	94.54	93.93	92.46	91.50	90.35	89.76	88.20	87.99	85.34
FDSR		98.06	98.19	97.90	97.80	97.73	97.48	96.99	96.57	95.76	94.66	94.11	91.68	90.49	86.92
DCTNet		97.91	97.88	97.74	97.29	97.17	96.72	96.23	95.32	93.96	92.57	92.08	90.07	89.36	86.20
DORNet		98.41	98.38	98.31	97.93	97.80	97.35	96.94	96.20	95.25	94.37	94.01	92.49	91.55	87.72
C2PD		97.94	98.11	97.86	97.97	97.72	97.52	97.11	96.65	95.67	95.05	94.43	93.45	91.17	86.04
DuCos		98.67	98.73	98.63	98.32	98.19	97.93	97.66	97.10	96.30	95.64	95.26	93.30	92.18	87.85
FoundDSR-S		99.13	99.18	99.17	98.91	98.84	98.58	98.36	97.91	97.33	96.56	96.21	94.86	94.24	91.16
FoundDSR		99.34	99.37	99.32	99.02	98.93	98.68	98.44	98.09	97.49	96.92	96.59	95.49	95.02	92.18
Marigold-DC	Middlebury	90.88	90.51	89.80	88.50	88.89	87.87	87.58	86.12	85.19	83.68	83.50	81.62	81.05	77.96
LDCM		97.62	97.49	96.56	94.11	93.73	92.28	91.11	89.55	87.65	85.90	85.52	83.30	82.59	79.29
FDSR		97.77	97.71	97.33	96.96	96.83	96.49	96.03	95.50	94.41	92.51	91.68	87.54	85.55	80.72
DCTNet		97.28	97.25	97.08	96.44	96.31	95.57	94.86	93.49	91.37	88.97	88.26	85.28	84.04	79.95
DORNet		97.56	97.56	97.50	97.20	97.17	96.82	96.34	95.72	94.61	93.22	92.66	89.70	88.07	82.47
C2PD		97.47	97.31	97.20	97.16	97.01	96.79	96.46	96.05	95.11	94.47	93.41	91.17	87.52	80.28
DuCos		97.86	98.00	97.90	97.53	97.45	97.13	96.78	96.29	95.66	94.90	94.35	91.50	89.58	83.42
FoundDSR-S		98.53	98.56	98.46	98.10	98.80	97.71	97.38	96.91	96.35	95.63	95.28	93.75	92.60	88.47
FoundDSR		98.98	99.01	98.86	98.47	98.43	98.03	97.72	97.29	96.74	96.12	95.80	94.52	93.53	89.85
PromptDA	RGB-D-D	97.13	97.11	97.02	96.78	96.76	96.59	96.46	96.31	96.32	96.05	96.11	96.05	95.91	95.75
Marigold-DC		98.05	97.99	97.87	97.57	97.46	97.14	96.83	96.39	95.80	95.14	94.87	94.08	93.48	92.06
InfiniDepth		97.64	97.60	97.49	97.31	97.12	96.77	96.57	96.12	95.48	94.94	94.54	94.03	93.35	91.29
LDCM		99.20	99.17	98.70	98.05	97.90	97.34	96.76	96.14	95.45	94.66	94.45	93.51	92.86	91.42
FDSR		99.31	99.22	99.02	99.03	98.97	98.89	98.62	98.43	98.22	97.81	97.61	96.35	95.33	93.07
DCTNet		99.13	99.10	98.99	98.70	98.64	98.43	98.27	98.03	97.58	97.01	96.78	95.77	95.08	93.17
DORNet		99.15	99.14	99.10	98.91	98.87	98.68	98.53	98.29	97.95	97.60	97.47	96.73	96.05	93.87
C2PD		99.28	99.14	98.95	99.05	98.89	98.82	98.55	98.39	98.14	97.90	97.53	97.09	95.75	92.97
DuCos		99.40	99.43	99.36	99.17	99.13	98.97	98.83	98.60	98.26	97.99	97.89	97.25	96.52	93.99
FoundDSR-S		99.62	99.62	99.59	99.44	99.40	99.27	99.15	98.99	98.77	98.54	98.49	98.06	97.70	96.37
FoundDSR		99.72	99.72	99.66	99.50	99.47	99.34	99.23	99.07	98.87	98.68	98.65	98.35	98.13	97.23
PromptDA	iBims-1	86.92	86.86	87.07	87.33	87.08	88.26	87.57	87.65	87.18	87.25	87.22	86.99	86.72	86.38
Marigold-DC		86.50	86.51	86.37	85.83	85.74	85.64	84.98	85.85	84.27	83.47	83.61	83.49	81.79	80.76
LDCM		93.68	93.70	92.89	91.32	91.09	90.19	89.56	88.60	87.51	86.52	86.19	85.13	84.50	82.60
FDSR		95.99	95.84	95.64	94.92	94.93	94.50	94.27	93.63	92.90	91.87	91.24	89.02	87.63	84.71
DCTNet		94.55	94.54	94.24	93.48	93.35	92.81	92.38	91.71	90.41	88.90	88.41	86.70	85.71	83.30
DORNet		96.34	96.22	95.83	95.15	95.03	94.55	94.15	93.63	92.78	92.01	91.63	90.05	88.93	85.62
C2PD		96.39	96.35	96.47	95.80	95.60	95.10	94.93	94.35	93.55	93.00	92.05	91.22	88.84	84.81
DuCos		96.70	96.49	97.16	96.37	96.12	95.69	95.30	94.66	93.98	93.30	92.94	91.11	89.79	86.15
FoundDSR-S		96.80	96.81	96.78	96.48	96.43	96.09	95.82	95.40	94.83	94.23	93.94	93.03	92.39	90.01
FoundDSR		97.12	97.05	96.78	96.29	96.22	95.79	95.61	95.20	94.73	94.28	94.15	93.55	93.05	91.40

A.4 More Visual Comparisons

Fig. 7 shows qualitative comparisons on real-world scenarios with unseen degradations. The results further confirm the strong generalization capability of our method in reconstructing high-quality depth under challenging real-world conditions. For example, the toy restored by FoundDSR exhibits more accurate depth structures than those produced by competing methods.

Figs. 8 and 9 present qualitative comparisons and corresponding error maps under integer and non-integer upsampling factors, respectively. FoundDSR consistently delivers superior reconstruction quality over existing DSR methods and depth foundation models across different scales, recovering more accurate depth structures with lower errors. For example, the cabinet at $\times 13.0$ in Fig. 8 and the arm at $\times 12.5$ in Fig. 9 reconstructed by our method exhibit more faithful geometric structures and are closer to the GT depth than those of competing methods.

Furthermore, Fig. 10 provides visual comparisons under different levels of zero-mean Gaussian noise, with standard deviations ranging from 0.02 to 0.04. FoundDSR remains robust under increasing noise levels, effectively suppressing noise while recovering accurate geometric details. For instance, the arm reconstructed by our method in Fig. 10 contains less residual noise and lower reconstruction errors than other approaches. These results further demonstrate the robustness and strong zero-shot generalization of our FoundDSR.

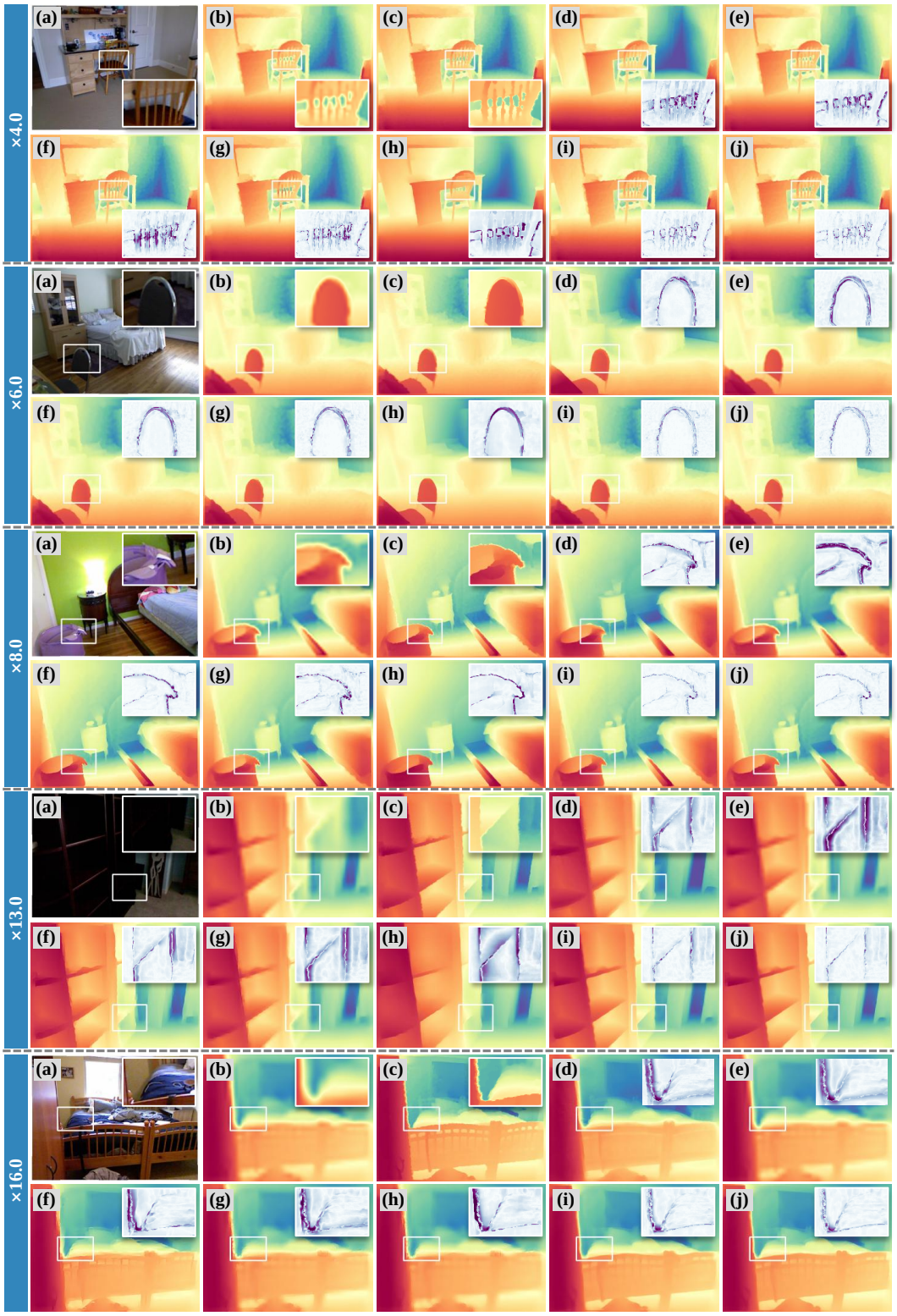


Figure 8: Visual comparisons on NYU-v2 across integer scales. (a)–(j) denote RGB, LR, GT, Marigold-DC, LDCM, DCTNet, DORNet, DuCos, FoundDSR-S, and FoundDSR, respectively.

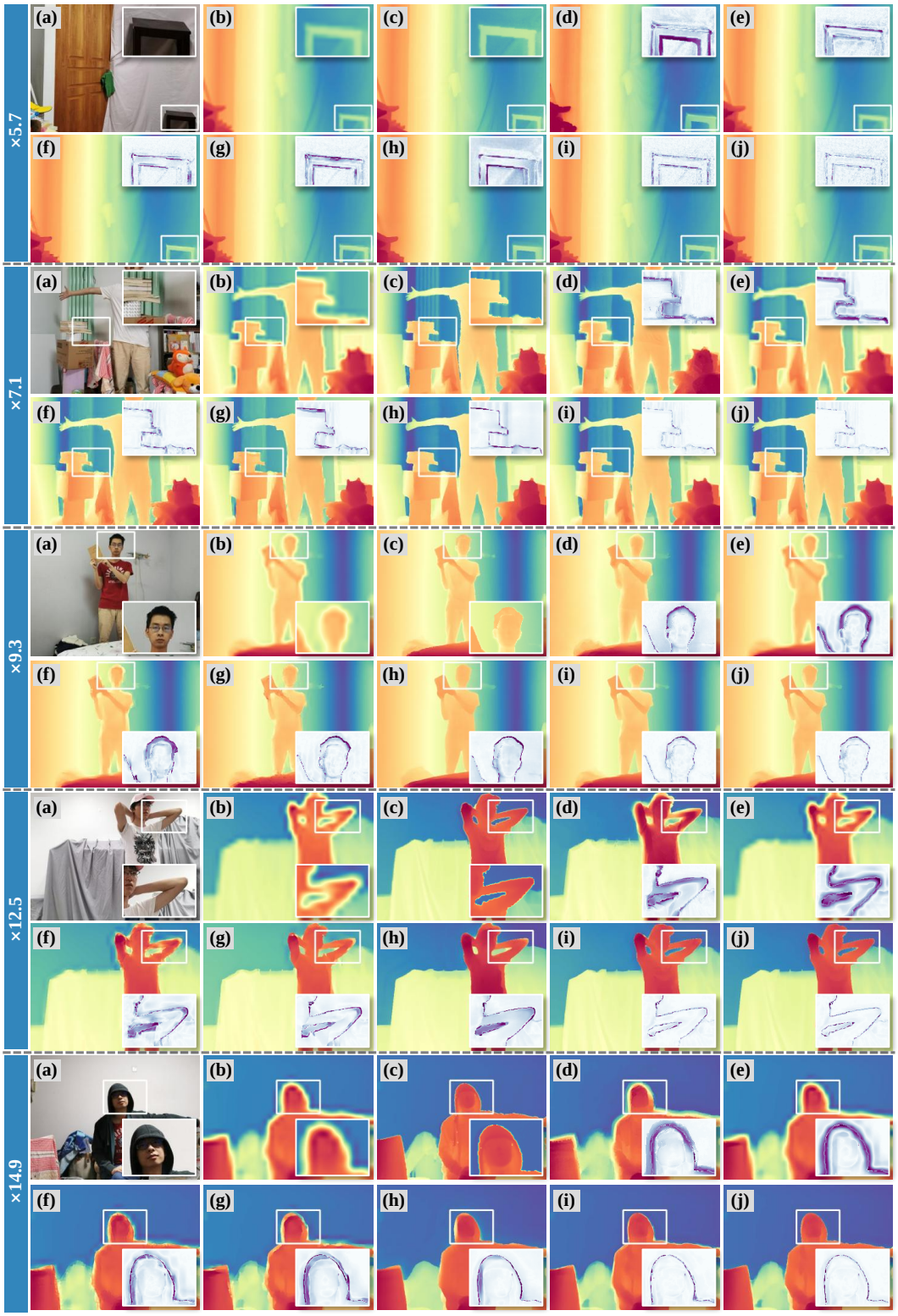


Figure 9: Visual comparisons on RGB-D-D across non-integer scales. (a)–(j) denote RGB, LR, GT, Marigold-DC, LDCM, DCTNet, DORNet, DuCos, FoundDSR-S, and FoundDSR, respectively.

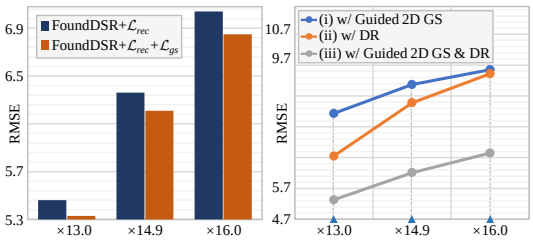
A.5 More Ablation Study

Different Loss Functions. Fig. 11(a) presents the ablation study on different loss functions, including the standard reconstruction loss $\mathcal{L}_{rec}$ and our Gaussian loss $\mathcal{L}_{gs}$. While $\mathcal{L}_{rec}$ alone already yields



Figure 10: Visual comparisons on $\times 4.0$ RGB-D-D under various noise levels. (a)–(j): RGB, LR, GT, Marigold-DC, LDCM, DCTNet, DORNet, DuCos, FoundDSR-S, and FoundDSR.

a strong baseline, incorporating $\mathcal{L}_{gs}$ consistently reduces the RMSE across all upsampling factors, with more pronounced improvements at higher scales. Specifically, adding $\mathcal{L}_{gs}$ decreases RMSE by $0.13cm$, $0.15cm$, and $0.19cm$ at upsampling scales of $\times 13.0$, $\times 14.9$, and $\times 16.0$, respectively. These results demonstrate that $\mathcal{L}_{gs}$ effectively improves the Gaussian-upsampled representation and facilitates more accurate depth reconstruction.



(a) Different loss functions (b) Guided 2D GS and DR

Figure 11: Ablation study of (a) different losses and (b) guided 2D GS and depth reconstruction (DR) on $\times 13.0$, $\times 14.9$, and $\times 16.0$ NYU-v2.

Guided 2D GS and Depth Reconstruction Branches. Fig. 11(b) analyzes the contributions of the guided 2D GS and depth reconstruction (DR) branches. The results show that guided 2D GS (i) stabilizes performance across different scale factors, reducing the performance gap between continuous scales. In contrast, the DR branch (ii) substantially improves reconstruction accuracy at every scale, but exhibits larger fluctuations across continuous scale factors. When both branches are combined, FoundDSR (iii) achieves the best performance across all evaluated scales, reducing RMSE by $2.80cm$ and $1.42cm$ compared with (i) and (ii), respectively, on $\times 16.0$ NYU-v2.