Dual Graph Regularized Deep Unfolding Network for Guided Depth Map Super-resolution
–Supplementary Material–

This document provides additional details to complement the main paper:
- **Section A:** More related work.
- **Section B:** Proof for Lemma 1.
- **Section C:** Additional implementation details and experimental settings.
- **Section D:** Extended experimental results and visual comparisons.
- **Section E:** Convergence Discussion.

A. Related Work

A.1. Guided Depth Map Super-resolution

Guided depth map super-resolution (GDSR) refers to the process of restoring fine spatial details in a low-resolution (LR) depth map using a high-resolution (HR) color image as a reference. Depending on the algorithmic design paradigm, existing methods can be broadly grouped into filtering-based, optimization-based, and learning-based approaches.

Filtering-based methods enhance the resolution of LR images by using a weighted averaging technique, where the weights are derived from various strategies. For instance, Eisemann *et al.* [1] propose an edge-preserving filter that considers both spatial distance and intensity difference. Based on the local linear assumption, He *et al.* [2] introduce a translation variant image filter. These approaches are attractive for their simplicity and low computational cost. Nevertheless, their dependence on the local linear assumption restricts their capability to capture complex cross-modal correlations.

Optimization-based methods address the GDSR problem from a global optimization perspective. The objective function of these methods typically comprises two components: data fidelity term and regularization term. The former maintains data consistency, while the latter enhances structural alignment with the guidance image. Representative regularization functions include the Markov random field (MRF) [3], Auto-Regress model (AR) [4] and conditional random field (CRF) [5]. Although these methods overcome certain limitations of filtering-based approaches, they rely on manually designed objective functions, which may not fully capture complex image priors.

Recently, learning-based methods have emerged as the dominant paradigm in various computer vision tasks, outperforming traditional hand-designed methods with their ability to autonomously learn complex features. To this end, more and more advanced architectures are proposed and applied in the field of GDSR [6–12]. For example, Li *et al.* [6] present a data-driven filter capable of adaptively identifying and injecting critical structural cues from the guidance image into the target image. Kim *et al.* [13] develop deformable kernel networks for joint image filtering, where both sampling locations and kernel weights are learned in a content-adaptive manner. Tang *et al.* [14] introduce a dual-branch framework for jointly learning depth map SR and monocular depth estimation, aiming to bridge the information gap between the two tasks. Yuan *et al.* [10] address noise and distortion in real-world low-resolution depth maps by estimating structure flow from guidance RGB images. Tang *et al.* [15] develop a joint implicit image function framework that represents the high-resolution depth map as a continuous function of spatial coordinates guided by RGB input. This implicit formulation enables accurate depth prediction without explicit upsampling and effectively preserves structural details. Zhao *et al.* [7] propose a discrete cosine transform network with semi-coupled blocks to better extract informative features. Zuo *et al.* [16] develop a guided implicit function framework that models depth as a continuous function of spatial coordinates and introduces a scale-aware fusion module for effective multi-scale feature aggregation. Zhou *et al.* [17] design a spatial frequency information integration network that fuses cross-modality features in both the spatial and frequency domains. Wang *et al.* [11] propose a degradation-oriented blind depth super-resolution method that jointly learns degradation representation and reconstruction. Yan *et al.* [12] leverage a frozen foundation model to extract semantic-aligned global features and introduce a duality-constrained optimization

framework that enforces consistency between the degradation and reconstruction processes. Although these learning-based methods have achieved superior performance by leveraging powerful feature extraction capabilities, they are typically designed as black-box models with limited interpretability, and performance improvements often come at the cost of increased model complexity and parameters.

A.2. Deep Unfolding Networks

In recent years, Deep Unfolding Networks (DUNs) have gained wide attention for their ability to combine the interpretability of model-based optimization with the learning flexibility of deep neural networks. By unfolding iterative solvers into structured layers, DUNs offer transparent and controllable architectures for various image restoration tasks. A representative work, LISTA [18], unfolds sparse coding into a trainable network, inspiring a range of DUN-based models in denoising [19, 20], deblurring [21, 22], and super-resolution [23, 24]. In the context of GDSR, several studies have adopted DUN-based designs. Riegler *et al.* [25] propose a deep primal-dual network that mimics optimization dynamics, using dual gradients to enforce RGB-depth consistency. Marivani *et al.* [26] propose a multimodal unfolding framework that fuses RGB and depth priors at each stage to progressively refine structural details. Zhao *et al.* [27] propose a discontinuity-aware unfolding network that jointly refines depth and gradient for better structure preservation. Metzler *et al.* [28] introduce a deep anisotropic diffusion model that simulates diffusion iterations with learned directional filters. Dai et al. [29] proposed an indoor depth recovery framework based on deep unfolding with a non-local prior, where a non-local auto-regressive regularization term is introduced to exploit repetitive depth structures in the scene. More recently, De Lutio *et al.* propose LGR [30], a hybrid framework that incorporates graph-based regularization into convolutional neural networks for guided depth super-resolution. Specifically, LGR learns a data-driven similarity matrix to construct the Laplacian graph and embeds a differentiable regularization module into the end-to-end training pipeline. As the first work to explicitly combine deep learning with graph optimization in GDSR, LGR bridges an important research gap and achieves competitive performance. However, LGR still faces several notable limitations. Although sparse graph construction effectively reduces computational cost, it inherently restricts the receptive field to local neighborhoods, making it difficult to capture long-range structural dependencies. In addition, the fixed adjacency structure restricts the model to fixed input sizes, which reduces its ability to handle images in different resolutions. Finally, flattening the spatial features into one-dimensional vectors during graph processing breaks the natural two-dimensional layout of the depth map and may reduce geometric accuracy and spatial consistency.

B. Proof for Lemma 1

This proof is based on **Definition 1** in the main paper, which defines the dual-graph regularization term:

$$\min_{\boldsymbol{X}} \frac{1}{2}\sum_{i=1}^{H}\sum_{j=1}^{H}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|_2^2(\boldsymbol{S}_r)_{ij}+\frac{1}{2}\sum_{i=1}^{W}\sum_{j=1}^{W}\|(\boldsymbol{X}^\top)_i-(\boldsymbol{X}^\top)_j\|_2^2(\boldsymbol{S}_c)_{ij}. \tag{1}$$

Lemma 1: *By introducing the Laplacian matrices $\boldsymbol{L}_r \in \mathbb{R}^{H\times H}$ and $\boldsymbol{L}_c \in \mathbb{R}^{W\times W}$ in terms with $\boldsymbol{S}_r$ and $\boldsymbol{S}_c$, respectively, Eq. (1) is equivalent to the following expression,*

$$\min_{\boldsymbol{X}} tr(\boldsymbol{X}^\top\boldsymbol{L}_r\boldsymbol{X})+tr(\boldsymbol{X}\boldsymbol{L}_c\boldsymbol{X}^\top). \tag{2}$$

Proof: *As the two Laplacian matrices $\boldsymbol{L}_r$ and $\boldsymbol{L}_c$ are constructed by the affinity graphs $\boldsymbol{S}_r \in \mathbb{R}^{H\times H}$ and $\boldsymbol{S}_c \in \mathbb{R}^{W\times W}$, respectively, we can formulate that*

$$\boldsymbol{L}_r=\boldsymbol{U}_r-\boldsymbol{S}_r, \tag{3}$$

$$\boldsymbol{L}_c=\boldsymbol{U}_c-\boldsymbol{S}_c, \tag{4}$$

where $\boldsymbol{U}_r \in \mathbb{R}^{H\times H}$ and $\boldsymbol{U}_c \in \mathbb{R}^{W\times W}$ are two degree matrices, i.e., $(\boldsymbol{U}_r)_{ii}=\sum_j(\boldsymbol{S}_r)_{ij}$ and $(\boldsymbol{U}_c)_{ii}=\sum_j(\boldsymbol{S}_c)_{ij}$. According to the definition of vector norm, we get

$$\frac{1}{2}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|_2^2=\frac{1}{2}(\|\boldsymbol{X}_i\|_2^2+\|\boldsymbol{X}_j\|_2^2-2\langle\boldsymbol{X}_i,\boldsymbol{X}_j\rangle), \tag{5}$$

By combining Eq. (1) and Eq. (5), we can rewrite Eq. (1) as

$$\min_{\boldsymbol{X}} \frac{1}{2}\sum_{i=1}^{H}\sum_{j=1}^{H}(\|\boldsymbol{X}_i\|_2^2+\|\boldsymbol{X}_j\|_2^2-2\langle\boldsymbol{X}_i,\boldsymbol{X}_j\rangle)(\boldsymbol{S}_r)_{ij}+\frac{1}{2}\sum_{i=1}^{W}\sum_{j=1}^{W}(\|(\boldsymbol{X}^\top)_i\|_2^2+\|(\boldsymbol{X}^\top)_j\|_2^2-2\langle(\boldsymbol{X}^\top)_i,(\boldsymbol{X}^\top)_j\rangle)(\boldsymbol{S}_c)_{ij}. \tag{6}$$

Then, Eq. (6) can be reformulated as

$$\begin{aligned}
&\min_{\boldsymbol{X}} \sum_{i=1}^{H}\sum_{j=1}^{H}(\|\boldsymbol{X}_i\|_2^2 - 2\langle \boldsymbol{X}_i, \boldsymbol{X}_j\rangle)(\boldsymbol{S}_r)_{ij} + \sum_{i=1}^{W}\sum_{j=1}^{W}(\|(\boldsymbol{X}^\top)_i\|_2^2 - 2\langle (\boldsymbol{X}^\top)_i (\boldsymbol{X}^\top)_j\rangle)(\boldsymbol{S}_c)_{ij} \\
=&\min_{\boldsymbol{X}} \sum_{i=1}^{H}((\boldsymbol{X}^\top \boldsymbol{U}_r \boldsymbol{X})_{ii} - (\boldsymbol{X}^\top \boldsymbol{S}_r \boldsymbol{X})_{ii}) + \sum_{i=1}^{W}((\boldsymbol{X}\boldsymbol{U}_c\boldsymbol{X}^\top)_{ii} - (\boldsymbol{X}\boldsymbol{S}_c\boldsymbol{X}^\top)_{ii}) \\
=&\min_{\boldsymbol{X}}\ tr(\boldsymbol{X}^\top \boldsymbol{L}_r \boldsymbol{X}) + tr(\boldsymbol{X}\boldsymbol{L}_c\boldsymbol{X}^\top).
\end{aligned} \tag{7}$$

The proof is completed.

C. Experimental Details

C.1. Dataset

Following the experimental protocols of existing guided depth super-resolution methods [17, 31–33], we evaluate our method on two benchmark datasets: NYU v2 [34] and RGB-D-D [35].

For the NYU v2 dataset [34], the first 1,000 RGB-D image pairs are used for training, while the remaining 449 pairs are used for testing. To further evaluate the generalization ability of our model, we also perform cross-dataset testing on five representative datasets: (1) 1,064 RGB-D images from Sintel [36], (2) the test set of DIDOE [37], (3) 500 image pairs from the SUN RGB-D test set [38], (4) the official test set of RGB-D-D [35], and (5) the test set of DIML Indoor [39]. Low-resolution (LR) depth maps are generated by downsampling the high-resolution (HR) ground-truth depth maps using bicubic interpolation with scaling factors of $4\times$, $8\times$, and $16\times$. The corresponding HR RGB images are used as guidance.

The RGB-D-D dataset [35] is a real-world benchmark dataset, where the HR depth maps are captured using a Helios ToF camera, and the LR depth maps are obtained with a Huawei P30 Pro. To ensure fair and consistent evaluation, we strictly follow the official training and testing splits provided by the dataset.

C.2. Implementation Details

Our framework is implemented in PyTorch and trained on a workstation equipped with two NVIDIA RTX 5090 GPUs. We use image patches of size 320×320 and a batch size of 16. Training is conducted for 200 epochs using the Adam optimizer [40] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The learning rate is initialized to 1×10^{-4} and reduced by half after the 100th epoch to ensure stable convergence.

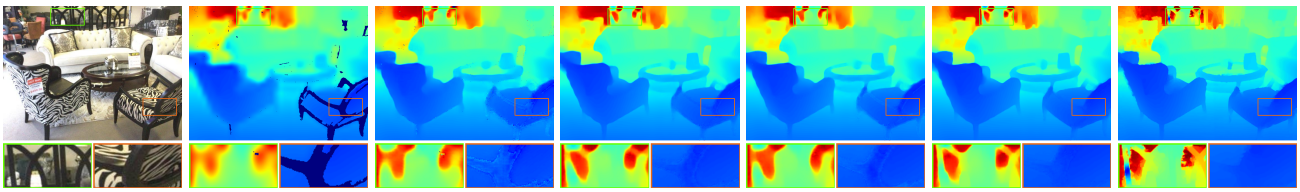
The proximal net is implemented as a lightweight U-Net with two downsampling layers and two upsampling layers. At each resolution level, a residual block is used as the basic building unit. The downsampling path aggregates multi-scale structural information, and the upsampling path restores fine details with skip connections. The iterative parameters α, μ, β, and λ are initialized to 1 and jointly optimized during training. The model performs three iterative refinement stages to progressively enhance reconstruction quality. The $\mathcal{L}_1$ loss is used for supervision. Following prior works [7, 17, 33], we apply standard data augmentation techniques, including random flipping and rotation, to improve the model's generalization ability. For fair comparison, all learning-based methods are trained and evaluated using the same datasets and experimental protocols. The root mean square error (RMSE) is used to measure the difference between the predicted depth values and the ground-truth depth values:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i - \hat{D}_i)^2}, \tag{8}$$

where D_i and $\hat{D}_i$ denote the ground-truth and predicted depth values at pixel i, respectively, and N is the total number of pixels. A lower RMSE indicates better reconstruction performance.

D. Experiments

In this section, we present more visualization results as well as additional subjective and objective comparisons to further demonstrate the visual quality and quantitative performance of our method. In addition, we conduct a new experiment on RGB-guided joint depth map completion and super-resolution to further verify the effectiveness and generalization capability of the proposed method.



(a) Guidance (b) Input (c) DJFR [6] (d) LGR [30] (e) DORNet [11] (f) **LapNet** (g) GT

Figure 1. Qualitative comparison for joint depth super-resolution and completion on the SUN RGB-D dataset [38].

D.1. Joint Depth Map Completion and Super-resolution

Experimental Settings. To evaluate the robustness of our method under realistic conditions, we simulate Kinect-like degradation, which is commonly observed in structured light sensors. This degradation includes two types of artifacts: structured holes near object boundaries and random dropouts in flat or reflective regions. To replicate this, we first downsample the HR depth maps using bicubic interpolation. Then, we apply a binary degradation mask that combines: i) structured missing regions, generated by detecting and dilating depth edges to mimic boundary-related holes, and ii) random missing pixels, introduced by sampling a Bernoulli distribution with a predefined missing probability. The final degraded depth map is obtained by masking the downsampled depth map through element-wise multiplication. We conduct experiments on the SUN RGB-D [38], and allocate the first 1,500 image pairs for training, with the remaining 355 pairs reserved for evaluation.

Results. We compare our method with several representative baselines, as shown in Table 1. LapNet achieves the best performance at all three scaling factors ($4\times$, $8\times$, and $16\times$), highlighting its strong robustness in practical degradation settings, particularly in cases involving simultaneous depth value absence and spatial resolution loss. Visual comparisons are presented in Fig. 1. As illustrated, DJFR [6] fails to recover missing regions, while LGR [30] and DORNet [11] can inpaint the holes but often result in incomplete or distorted structures, as seen in the zoomed-in areas. In contrast, our method not only fills in the missing depth values but also reconstructs more complete and accurate structures. These results highlight the effectiveness of our dual-prior design in addressing complex degradations and preserving structural integrity under challenging conditions.

Table 1. RMSE comparison on SUN RGB [41] dataset for joint depth map completion and super-resolution.

Scale	Bicubic	DJFR [6]	DKN [13]	DCTNet [7]	MMNet [42]	AHMF [43]	LGR [30]	SFINet [17]	DORNet [11]	**LapNet**
$4\times$	22.69	3.59	2.73	2.57	3.17	<u>2.45</u>	2.78	2.73	2.46	**1.92**
$8\times$	22.83	3.39	3.06	2.66	2.69	2.52	2.87	3.04	<u>2.49</u>	**2.05**
$16\times$	23.15	3.87	3.42	3.05	3.59	2.82	3.22	3.40	<u>2.74</u>	**2.31**

D.2. Comparison with More Methods

We further compare the proposed method with two diffusion-based methods and one deep unfolding-based methods. The corresponding quantitative results are presented in Table 2. As can be seen, our method achieves the best performance among all the compared methods.

Table 2. RMSE/MAE comparison on NYU v2 [34] dataset for guided depth map super-resolution.

Scale	StableSR [44]	TVT [45]	DeepSN-Net [46]	**LapNet**
$4\times$	1.86/0.71	1.54/0.60	1.42/0.52	**1.05/0.38**
$8\times$	4.04/1.62	3.79/1.55	3.66/1.48	**2.33/0.91**
$16\times$	6.92/2.84	6.41/2.63	5.98/2.41	**4.55/1.86**

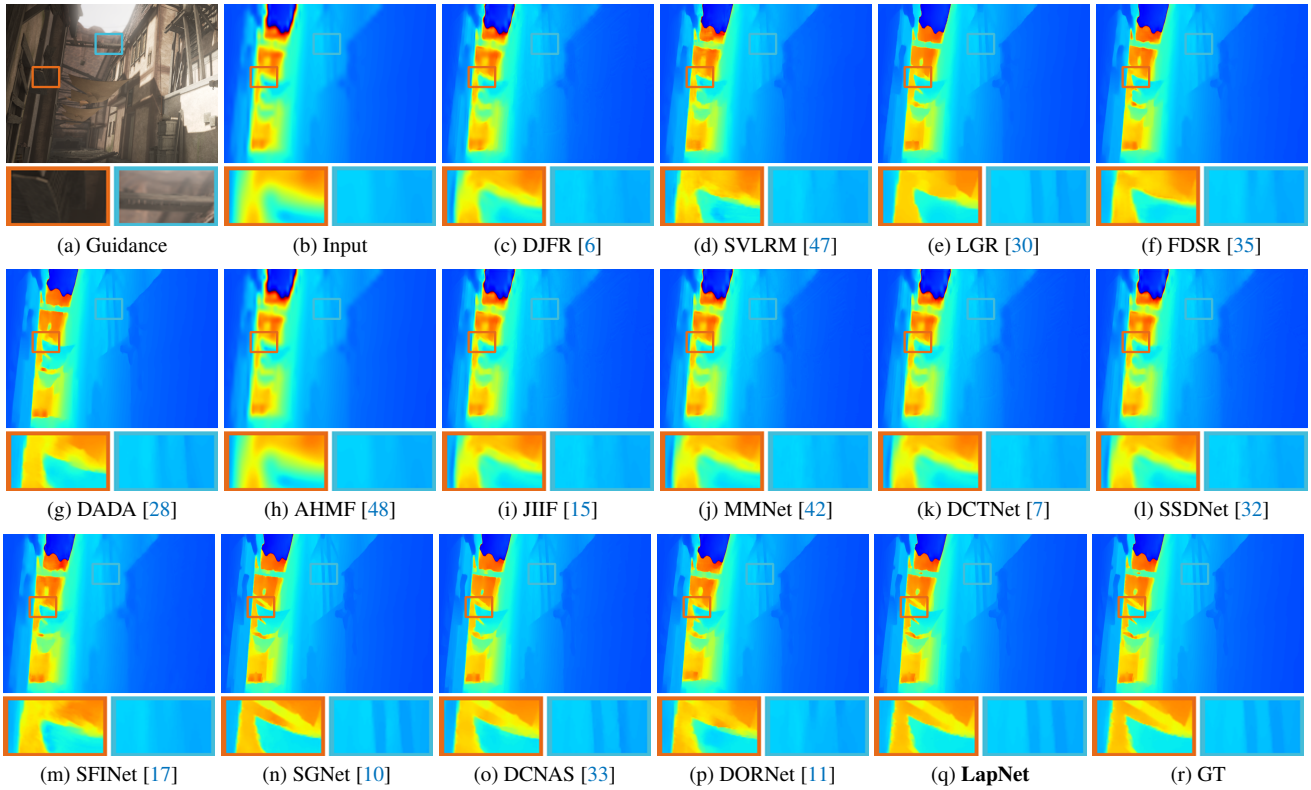
D.3. More ablation experiments

Effect of Skip Connections in the Proximal Network. To prevent the loss of important structural information during cross-stage propagation, we design a skip connection strategy in the proximal network, enabling the model to reuse features across iterations. Specifically, both the reconstructed depth $\boldsymbol{X}_k$ and the intermediate decoder features $\boldsymbol{F}_{k-1}$ from the previous stage are passed to the encoder of the current stage. To validate the effectiveness of this strategy, we construct a variant model, `Model8`, in which all skip connections are removed and only current-stage inputs are used. As presented in Table 3,

eliminating skip connections leads to consistent performance degradation on all three benchmark datasets, indicating that they are beneficial for reducing information loss and improving reconstruction quality.

Table 3. **Ablation Study.** RMSE comparison of proximal network design strategies for 8× GDSR.

Method	NYU v2 [34]	Sintel [36]	DIDOE [37]
`Model8`	2.52	5.26	5.84
LapNet	**2.33**	**5.05**	**5.51**



(a) Guidance (b) Input (c) DJFR [6] (d) SVLRM [47] (e) LGR [30] (f) FDSR [35]

(g) DADA [28] (h) AHMF [48] (i) JIIF [15] (j) MMNet [42] (k) DCTNet [7] (l) SSDNet [32]

(m) SFINet [17] (n) SGNet [10] (o) DCNAS [33] (p) DORNet [11] (q) **LapNet** (r) GT

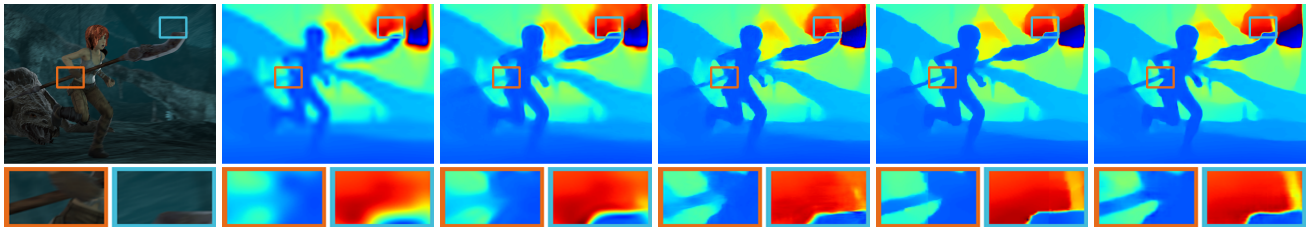
Figure 2. Qualitative comparison for 16× depth map super-resolution (*Sintel dataset [36] Image-85*).

D.4. More Visualization Results

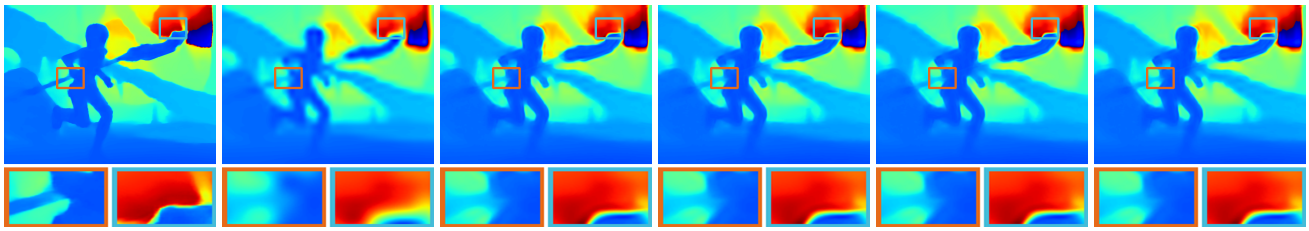
In this subsection, we present additional qualitative comparison results (see Fig. 2, Fig. 3, Fig. 4, Fig. 5). As shown in the figures, our method reconstructs more complete structural details and effectively suppresses artifacts caused by incorrect texture transfer.

E. Convergence Discussion

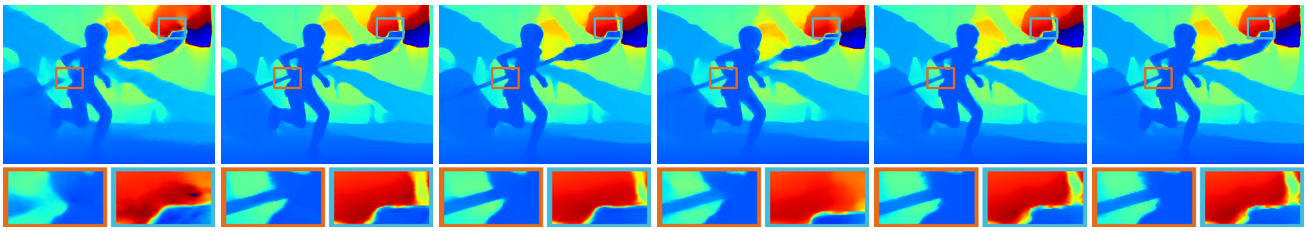
The proposed method is derived by unfolding a fixed number of ADMM iterations, and thus each stage of the network corresponds to one step of the underlying optimization algorithm. This gives the model a clear optimization interpretation and preserves the key structure of ADMM. While learnable modules are introduced to enhance representation power, the overall framework remains optimization-driven and retains strong interpretability. Following the common practice in deep unfolding literature, we do not pursue a strict theoretical proof for the convergence of the learned network, since the inclusion of neural parameterization generally makes such analysis difficult. Instead, we highlight that the proposed model demonstrates stable empirical convergence and consistent reconstruction improvement in practice across different experimental settings.



(a) Guidance (b) Input (c) DJFR [6] (d) SVLRM [47] (e) LGR [30] (f) FDSR [35]

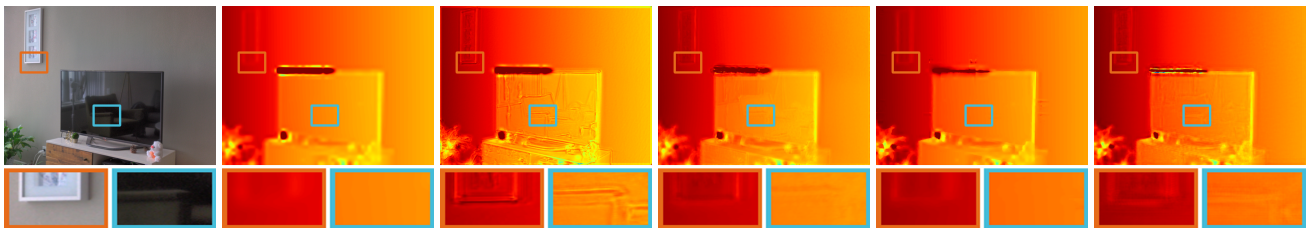


(g) DADA [28] (h) AHMF [48] (i) JIIF [15] (j) MMNet [42] (k) DCTNet [7] (l) SSDNet [32]

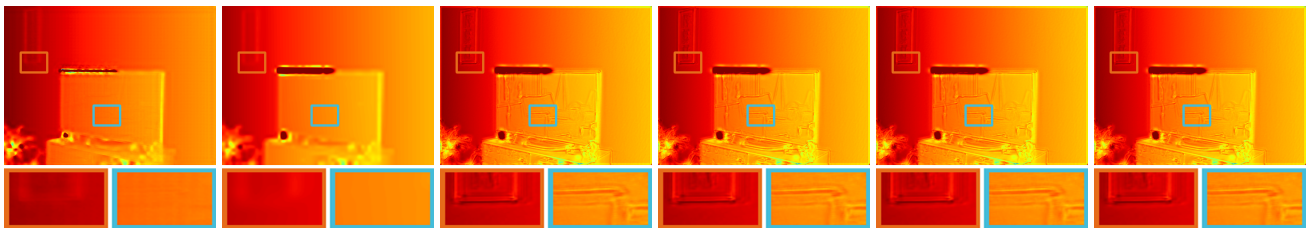


(m) SFINet [17] (n) SGNet [10] (o) DCNAS [33] (p) DORNet [11] (q) **LapNet** (r) GT

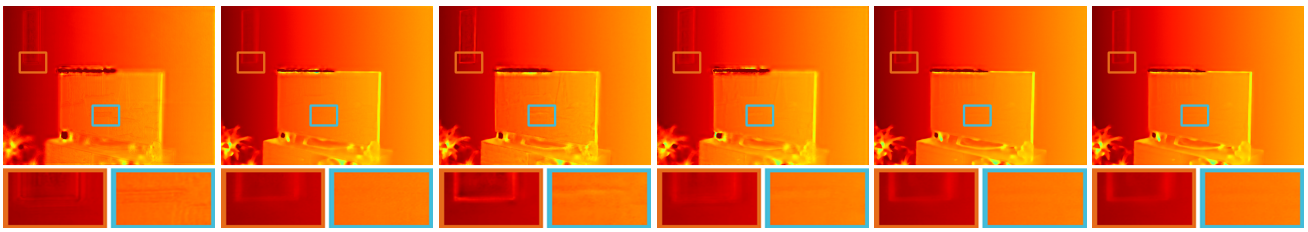
Figure 3. Qualitative comparison for 16× depth map super-resolution (*Sintel dataset [36] Image-483*).



(a) Guidance (b) Input (c) DJFR [6] (d) SVLRM [47] (e) LGR [30] (f) FDSR [35]

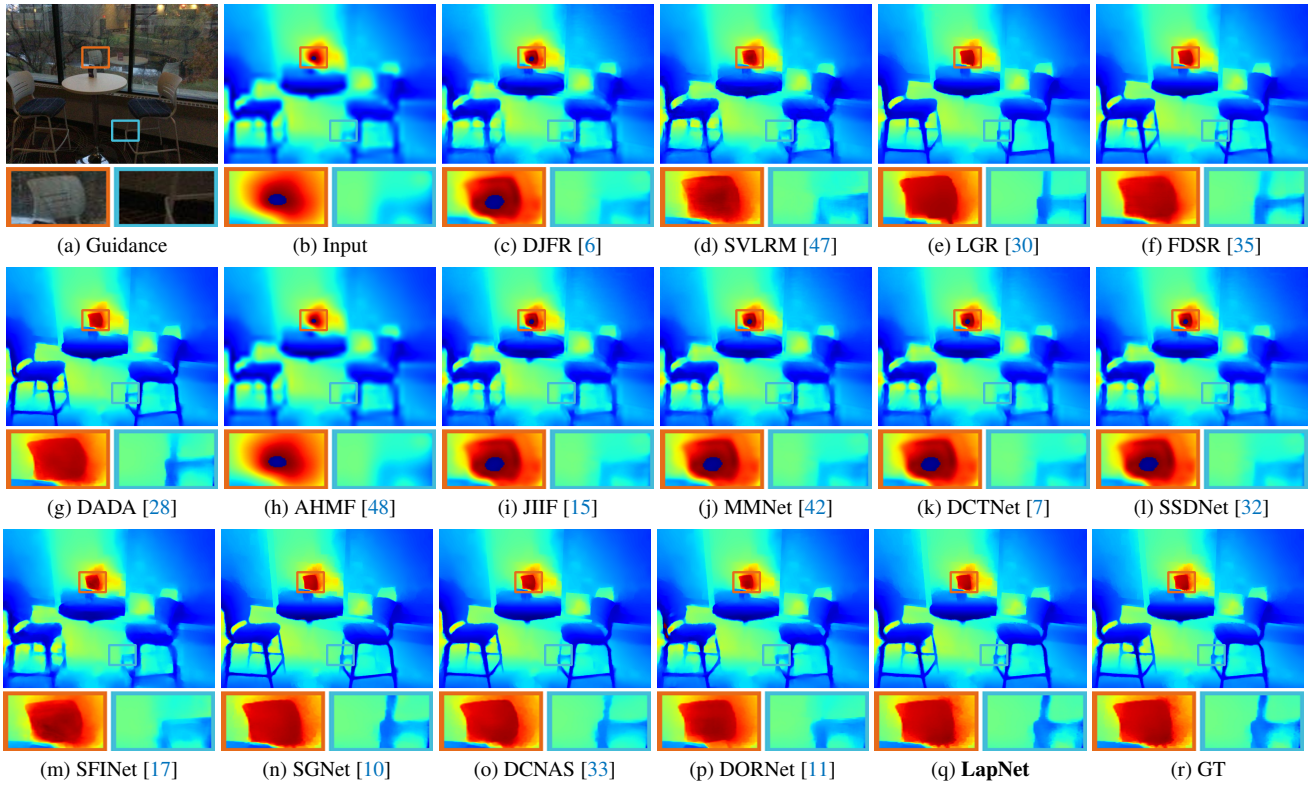


(g) DADA [28] (h) AHMF [48] (i) JIIF [15] (j) MMNet [42] (k) DCTNet [7] (l) SSDNet [32]



(m) SFINet [17] (n) SGNet [10] (o) DCNAS [33] (p) DORNet [11] (q) **LapNet** (r) GT

Figure 4. Qualitative comparison for 16× depth map super-resolution (*DIDOE dataset [37] Image-110*).



(a) Guidance (b) Input (c) DJFR [6] (d) SVLRM [47] (e) LGR [30] (f) FDSR [35]

(g) DADA [28] (h) AHMF [48] (i) JIIF [15] (j) MMNet [42] (k) DCTNet [7] (l) SSDNet [32]

(m) SFINet [17] (n) SGNet [10] (o) DCNAS [33] (p) DORNet [11] (q) **LapNet** (r) GT

Figure 5. Qualitative comparison for 16× depth map super-resolution (*SUN RGB-D dataset [38] Image-393*).

References

[1] E. Eisemann and F. Durand, "Flash photography enhancement via intrinsic relighting," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 673–678, 2004. 1

[2] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013. 1

[3] Y. Zuo, Q. Wu, J. Zhang, and P. An, "Minimum spanning forest with embedded edge inconsistency measurement model for guided depth map enhancement," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4145–4159, 2018. 1

[4] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from rgb-d data using an adaptive autoregressive model," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3443–3458, 2014. 1

[5] H. Wang, M. Yang, C. Zhu, and N. Zheng, "Rgb-guided depth map recovery by two-stage coarse-to-fine dense crf models," *IEEE Transactions on Image Processing*, vol. 32, pp. 1315–1328, 2023. 1

[6] Y. Li, J. B. Huang, N. Ahuja, and M. H. Yang, "Joint image filtering with deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1909–1923, 2019. 1, 4, 5, 6, 7

[7] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. 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. 1, 3, 4, 5, 6, 7

[8] W. Shi, M. Ye, and B. Du, "Symmetric uncertainty-aware feature transmission for depth super-resolution," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3867–3876, 2022.

[9] X. Wang, X. Chen, B. Ni, Z. Tong, and H. Wang, "Learning continuous depth representation via geometric spatial aggregator," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 2698–2706, 2023.

[10] Z. Wang, Z. Yan, and J. Yang, "Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 5823–5831, 2024. 1, 5, 6, 7

[11] Z. Wang, Z. Yan, J. Pan, G. Gao, K. Zhang, and J. Yang, "Dornet: A degradation oriented and regularized network for blind depth super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15813–15822, 2025. 1, 4, 5, 6, 7

[12] Z. Yan, Z. Wang, H. Dong, , J. Li, J. Yang, and G. H. Lee, "Ducos: Duality constrained depth super-resolution via foundation model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1

[13] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, pp. 1–22, 2021. 1, 4

[14] Q. Tang, R. Cong, R. Sheng, L. He, D. Zhang, Y. Zhao, and S. Kwong, "Bridgenet: A joint learning network of depth map super-resolution and monocular depth estimation," in *Proceedings of ACM International Conference on Multimedia*, p. 2148–2157, 2021. 1

[15] J. Tang, X. Chen, and G. Zeng, "Joint implicit image function for guided depth super-resolution," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4390–4399, 2021. 1, 5, 6, 7

[16] Y. Zuo, Y. Hu, Y. Xu, Z. Wang, Y. Fang, J. Yan, W. Jiang, Y. Peng, and Y. Huang, "Learning guided implicit depth function with scale-aware feature fusion," *IEEE Transactions on Image Processing*, vol. 34, pp. 3309–3322, 2025. 1

[17] M. Zhou, J. Huang, K. Yan, D. Hong, X. Jia, J. Chanussot, and C. Li, "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3, 4, 5, 6, 7

[18] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 399–406, 2010. 2

[19] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3587–3596, 2017. 2

[20] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1256–1272, 2017. 2

[21] H. Wang, T. Zhang, M. Yu, J. Sun, W. Ye, C. Wang, and S. Zhang, "Stacking networks dynamically for image restoration based on the plug-and-play framework," in *Proceedings of European Conference on Computer Vision*, pp. 446–462, Springer, 2020. 2

[22] C. Mou, Q. Wang, and J. Zhang, "Deep generalized unfolding networks for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17378–17389, 2022. 2

[23] Q. Ning, W. Dong, G. Shi, L. Li, and X. Li, "Accurate and lightweight image super-resolution with model-guided deep unfolding network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 240–252, 2021. 2

[24] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3214–3223, 2020. 2

[25] G. Riegler, D. Ferstl, M. Rüther, and B. Horst, "A deep primal-dual network for guided depth super-resolution," in *BMVC*, 2016. 2

[26] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multimodal deep unfolding for guided image super-resolution," *IEEE Transaction on Image Processing*, vol. 29, pp. 8443–8456, 2020. 2

[27] L. Zhao, J. Zhang, J. Zhang, H. Bai, and A. Wang, "Joint discontinuity-aware depth map super-resolution via dual-tasks driven unfolding network," *TIM*, vol. 73, pp. 1–14, 2024. 2

[28] N. Metzger, R. C. Daudt, and K. Schindler, "Guided depth super-resolution by deep anisotropic diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18237–18246, June 2023. 2, 5, 6, 7

[29] Y. Dai, J. Zhang, F. Fang, and G. Zhang, "Indoor depth recovery based on deep unfolding with non-local prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12355–12364, 2023. 2

[30] R. de Lutio, A. Becker, S. D'Aronco, S. Russo, J. D. Wegner, and K. Schindler, "Learning graph regularisation for guided super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4, 5, 6, 7

[31] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Guided depth map super-resolution: A survey," *ACM Computing Surveys*, vol. 55, no. 14, pp. 1–36, 2023. 3

[32] Z. Zhao, J. Zhang, X. Gu, C. Tan, S. Xu, Y. Zhang, R. Timofte, and L. Van Gool, "Spherical space feature decomposition for guided depth map super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12547–12558, 2023. 5, 6, 7

[33] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and S. Wang, "Dual-level cross-modality neural architecture search for guided image super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 9, pp. 8249–8267, 2025. 3, 5, 6, 7

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the European Conference on Computer Vision*, pp. 746–760, 2012. 3, 4, 5

[35] L. He, H. Zhu, F. Li, H. Bai, R. Cong, C. Zhang, C. Lin, M. Liu, and Y. 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. 3, 5, 6, 7

[36] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proceedings of the European Conference on Computer Vision*, pp. 611–625, 2012. 3, 5, 6

[37] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, *et al.*, "Diode: A dense indoor and outdoor depth dataset," *arXiv preprint arXiv:1908.00463*, 2019. 3, 5, 6

[38] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015. 3, 4, 7

[39] J. Cho, D. Min, Y. Kim, and K. Sohn, "Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset," *Expert Systems with Applications*, vol. 178, p. 114877, 2021. 3

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv: Learning*, 2014. 3

[41] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proceedings of European Conference on Computer Vision*, pp. 611–625, 2012. 4

[42] M. Zhou, K. Yan, J. Pan, W. Ren, Q. Xie, and X. Cao, "Memory-augmented deep unfolding network for guided image super-resolution," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 215–242, 2023. 4, 5, 6, 7

[43] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Deep attentional guided image filtering," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023. 4

[44] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5929–5949, 2024. 4

[45] Q. Yi, S. Li, R. Wu, L. Sun, Y. Wu, and L. Zhang, "Fine-structure preserved real-world image super-resolution via transfer vae training," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12415–12426, 2025. 4

[46] X. Deng, C. Zhang, L. Jiang, J. Xia, and M. Xu, "Deepsn-net: Deep semi-smooth newton driven network for blind image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2632–2646, 2025. 4

[47] J. Dong, J. Pan, J. S. Ren, L. Lin, J. Tang, and M.-H. Yang, "Learning spatially variant linear representation models for joint filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8355–8370, 2022. 5, 6, 7

[48] Z. Zhong, X. Liu, J. Jiang, D. Zhao, Z. Chen, and X. Ji, "High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion," *IEEE Trans. Image Process.*, vol. 31, pp. 648–663, 2022. 5, 6, 7