International Journal of Computer Vision (2026) 134:275
https://doi.org/10.1007/s11263-026-02883-3

CORRECTION

Correction: Scene Prior Filtering for Depth Super-Resolution

Zhengxue Wang[1] · Zhiqiang Yan[2] · Ming-Hsuan Yang[3,4] · Jinshan Pan[1] · Guangwei Gao[1] · Ying Tai[5] · Jian Yang[1,5]

Correction to:
International Journal of Computer Vision (2026) 134:251
https://doi.org/10.1007/s11263-026-02829-9

In this article, in addition to Zhengxue Wang, Guangwei Gao and Jian Yang should also have been denoted as a corresponding authors.

The original article can be found online at https://doi.org/10.1007/s11263-026-02829-9.

✉ Zhengxue Wang
zxwang@njust.edu.cn

✉ Guangwei Gao
csggao@gmail.com

✉ Jian Yang
csjyang@njust.edu.cn

Zhiqiang Yan
yanzq@nus.edu.sg

Ming-Hsuan Yang
mhyang@ucmerced.edu

Jinshan Pan
jspan@njust.edu.cn

Ying Tai
yingtai@nju.edu.cn

[1] PCA Lab, Nanjing University of Science and Technology, Nanjing, China

[2] National University of Singapore, Singapore, Singapore

[3] University of California, Berkeley, USA

[4] Yonsei University, Seoul, South Korea

[5] PCA Lab, Nanjing University, Nanjing, China

Published online: 18 May 2026

International Journal of Computer Vision (2026) 134:251
https://doi.org/10.1007/s11263-026-02829-9

Scene Prior Filtering for Depth Super-Resolution

Zhengxue Wang[1] · Zhiqiang Yan[2] · Ming-Hsuan Yang[3,4] · Jinshan Pan[1] · Guangwei Gao[1] · Ying Tai[5] · Jian Yang[1,5]

Received: 1 December 2025 / Accepted: 21 March 2026

Abstract

Multi-modal fusion serves as a cornerstone for successful depth map super-resolution. However, commonly used fusion strategies, such as addition and concatenation, fall short of effectively bridging the modal gap. As a result, guided image filtering methods have been introduced to mitigate this issue. Nevertheless, it is observed that their filter kernels usually encounter significant texture interference and edge inaccuracy. To tackle these two challenges, we introduce a Scene Prior Filtering network, SPFNet, which utilizes the priors' surface normal and semantic map from large-scale models. Specifically, we propose an All-in-one Prior Propagation that computes similarity between multi-modal scene priors, *i.e.*, RGB, normal, semantic, and depth, to reduce the texture interference. Besides, we design a One-to-one Prior Embedding that continuously embeds every single modal prior into depth using Mutual Guided Filtering, further alleviating texture interference while enhancing edge representations. Our SPFNet has been extensively evaluated on both real-world and synthetic datasets, achieving state-of-the-art performance. Project page: https://yanzq95.github.io/projectpage/SPFNet/index.html.

Keywords Depth super-resolution · Scene prior filtering · Texture interference · Large-scale model

1 Introduction

Advances in sensor technology have led to the extensive application of depth cues in various fields, such as autonomous driving (Qiao et al., 2024; Sun et al., 2021; Yan et al., 2023), 3D reconstruction (Song et al., 2020; Yan et al., 2023; Yang et al., 2022), and virtual reality (Yuan et al., 2023; Zhou et al., 2023). However, depth measurements are typically low resolution (LR) due to sensor limitations and the complexity of imaging environments. Recently, a number of guided image filtering approaches (Kim et al., 2021; Li et al., 2019; Pan et al., 2019; Zhong et al., 2023) have been proposed to facilitate depth super-resolution (DSR). Nevertheless, the filter kernels, which are constructed directly from RGB images, often suffer from significant texture interference, and the clarity near edges is typically compromised. For instance, as depicted in Fig. 1(e) and (f) (yellow boxes), the filter kernels of DKN (Kim et al., 2021) and DAGF (Zhong et al., 2023) contain a substantial amount of textures. Fig. 1(i) shows that these two kernels display too many abrupt changes, while the ground-truth depth is considerably smoother. These observations demonstrate that the texture interference is not conducive to depth recovery. Moreover, within the white boxes, the edges and their neighboring pixels

Communicated by Yuchao Dai.

Zhengxue Wang and Zhiqiang Yan have equally contributed to this work.

✉ Zhengxue Wang
zxwang@njust.edu.cn

Zhiqiang Yan
yanzq@nus.edu.sg

Ming-Hsuan Yang
mhyang@ucmerced.edu

Jinshan Pan
jspan@njust.edu.cn

Guangwei Gao
csggao@gmail.com

Ying Tai
yingtai@nju.edu.cn

Jian Yang
csjyang@njust.edu.cn

1 PCA Lab, Nanjing University of Science and Technology, Nanjing, China

2 National University of Singapore, Singapore, Singapore

3 University of California, Berkeley, USA

4 Yonsei University, Seoul, South Korea

5 PCA Lab, Nanjing University, Nanjing, China

Published online: 29 April 2026

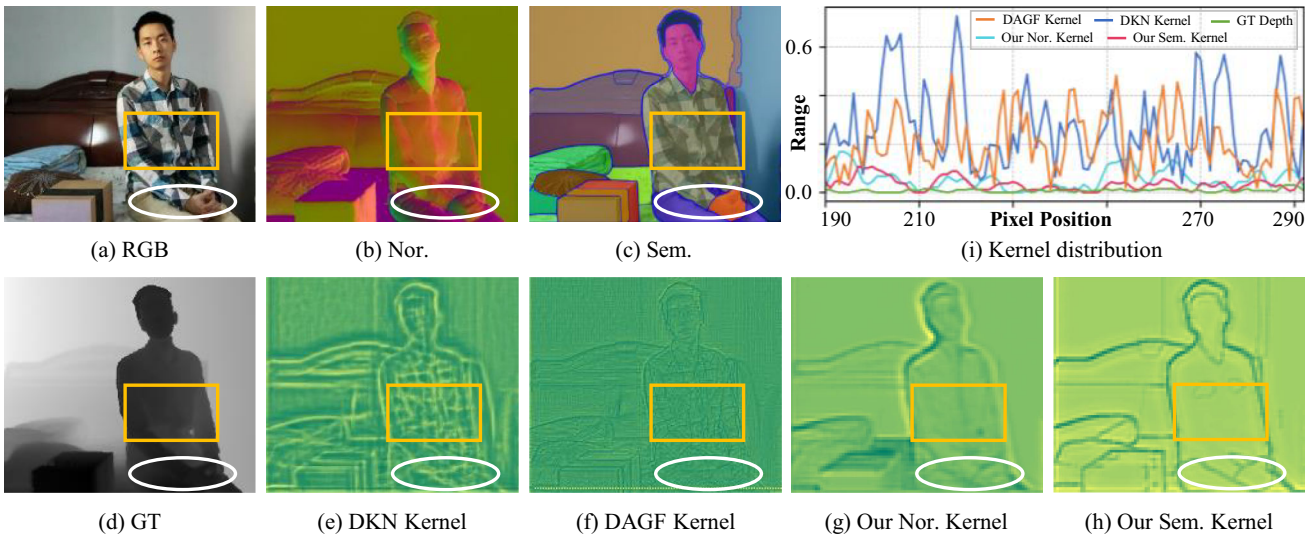


Fig. 1 Visualizations of scene priors and filter kernels. The normal (Nor.) (**b**) and semantic (Sem.) (**c**) are produced from (**a**) using Omnidata (Eftekhar et al., 2021) and SAM (Kirillov et al., 2023), respectively. (**e**) and (**f**) are the filter kernels derived from DKN (Kim et al., 2021) and DAGF (Zhong et al., 2023), while (**g**) and (**h**) are from our SPFNet. (**d**) is the ground-truth (GT) depth, and (**i**) is the normalized kernel distribution

show high similarity, leading to an insufficient contrast. Compared to the RGB input, Fig. 1(b) indicates that the surface normal is largely devoid of texture interference. On the other hand, Fig. 1(c) shows that the semantic map displays clear edges between different categories. These inherent characteristics are highly advantageous for constructing filter kernels with less interference and more distinct edges.

Drawing from the observations above, we design a novel scene prior filtering network (SPFNet) to reduce texture interference and enhance edge accuracy. Specifically, we first employ large-scale models (Eftekhar et al., 2021; Kirillov et al., 2023) to generate normal and semantic priors from RGB input. An all-in-one prior propagation (APP) is introduced, which computes the similarity between multi-modal scene priors, *i.e.*, RGB, normal, semantic, and depth, to weaken the interference. In addition, we present a one-to-one prior embedding (OPE) that sequentially incorporates each single-modal prior into depth using mutual guided filtering (MGF), further diminishing the interference and enhancing edges via normal and semantic. The MGF comprises a bidirectional path, that is, prior-to-depth filtering and depth-to-prior filtering. The prior-to-depth filtering transfers the accurate structural components from scene priors to depth. Conversely, the depth-to-prior filtering leverages depth knowledge to accentuate edges of these scene priors while downplaying the undesired textures.

As a result, Fig. 1(g) and (h) show that the kernels produced by our SPFNet are largely resistant to interference and exhibit precise edges. Furthermore, Fig. 1(i) shows that the distribution of our normal and semantic filter kernels aligns more closely with the GT, as compared to DKN and DAGF.

The main contributions of this work are:

- To address the issues of texture interference and edge inaccuracy in DSR, we are pioneering the incorporation of scene priors from large-scale models.
- We propose SPFNet, which recursively implements the novel all-in-one prior propagation, one-to-one prior embedding, and mutual guided filtering to further diminish texture interference and enhance edges.
- Extensive experiments on both real-world and synthetic datasets demonstrate that our SPFNet achieves superior performance, reaching the state-of-the-art.

2 Related Work

2.1 Multi-Modal Fusion Based DSR

Much progress in guided DSR (Gu et al., 2017; Yan et al., 2022; Zhong et al., 2021) based on deep learning has been made in recent years. Deng and Dragotti (2020) utilize multi-modal convolutional sparse coding to extract the common features between RGB and depth. DCTNet (Zhao et al., 2023) develops spherical space feature decomposition to separate shared and private features. In Zhao et al. (2023), project both RGB and depth features into a spherical space to separate their private features and align the shared ones. Similarly, (Yuan et al., 2023) introduce a deep contrast network to

split the depth into high-frequency and low-frequency maps. Most recently, a few methods (Deng et al., 2023; Wang et al., 2025, ?) exploit depth structure recovery. For instance, DADA (Metzger et al., 2023) combines anisotropic diffusion and a convolutional network to improve the edge transferring property of diffusion. In Wang et al. (2024), SGNet designs a structure-guided network that employs the gradient and frequency domains for structure enhancement. To enhance the accuracy of DSR in real-world scenarios,(He et al., 2021) establish a real-world RGB-D benchmark and develop a baseline method for real-world DSR based on octave convolution. DORNet (Wang et al., 2025) introduces a degradation-oriented and regularized framework, which selectively aggregates RGB and depth features by modeling the degradation representations between LR and HR depths. More recently, some methods have attempted to introduce foundation models to enhance depth quality. For example, (Yan et al., 2025) utilize a depth estimation foundation model to provide dual constraints for HR depth restoration. (Wang et al., 2025) introduce a powerful depth-aware model on large-scale datasets, which directly predicts accurate depth from RGB. (Viola et al., 2025) use sparse depth as a condition for a pre-trained depth generation model to restore dense depth from RGB and sparse depth. Unlike previous methods, which directly transform RGB features to depth, we focus more on leveraging normal and semantic priors to attenuate texture interference and improve accuracy near edges.

2.2 Guided Image Filtering based DSR

To transfer the structure information from guidance to target, numerous guided image filtering methods (He et al., 2012; Shen et al., 2015; Zhong et al., 2023) have been proposed in recent years. Li et al. (2019) develop joint image filtering based on deep convolutional networks to selectively transfer the structure from RGB to depth, and DKN (Kim et al., 2021) presents a deformable kernel network that explicitly generates a spatially-variant filter kernel and outputs sets of neighborhoods for each pixel. In Wang et al. (2023), utilize the hybrid side window filtering to propagate multi-scale structure knowledge from RGB to depth. In contrast, our method focuses more on minimizing texture interference within filter kernels and improving edge accuracy by exploiting scene priors and similarity maps.

2.3 Scene Prior Awareness

Surface normal and semantic priors contain rich geometry and boundary information. Recently, several methods (Kirillov et al., 2023; Qiu et al., 2019; Shao et al., 2024; Wu et al., 2023; Yang et al., 2018) have been designed to explore the use of these priors for facilitating downstream tasks. Xu et al. (2019) learn the geometric constraints between depth and surface normal in a diffusion module to improve the performance of depth completion. Qiu et al. (2019) integrate guided image and surface normal to enhance the depth accuracy, and Fan et al. (2020) present a data-fusion CNN method that fuses inferred surface normal and image for accurate free space detection. In addition, SKF (Wu et al., 2023) develops a semantic-aware knowledge-guided model to embed semantic prior in feature representation space for low-light image enhancement. In Jung et al. (2021), incorporate semantics into geometric representation to improve self-supervised monocular depth estimation.

Additionally, some methods attempt to leverage other prior to enhance the image. For instance, Xiao et al. (2025) optimize neuronal membrane potential and regulate spiking activity by designing a spiking attention block, while jointly modeling spatiotemporal features and mining non-local semantic priors to achieve efficient remote sensing image reconstruction. In Wang et al. (2025), propose a coarse-to-fine fusion strategy that integrates accurate but incomplete metric depth with complete yet relative geometric structure priors, enabling zero-shot generalization for depth restoration.

In this work, we leverage large-scale models to generate accurate surface normal and semantic priors from HR RGB. Both are utilized as prompts to alleviate texture interference and facilitate the quality of depth restoration.

2.4 Bidirectional Guidance

The bidirectional guidance mechanism is widely applied in many computer vision tasks, achieving sufficient feature representation through information interaction. For example, Dong et al. (2022) design a cross-domain adaptive filter to achieve mutual modulation of multi-modal inputs, and DAGF (Zhong et al., 2023) combines filter kernels from the guidance and target to model pixel-wise dependency between the two input images. In addition, Jiang et al. (2025) propose a direction-aware attention wavelet network based on mutual representation, which models the directionality of rain streaks via vector decomposition to achieve precise removal of heterogeneous rain streaks. Zhang et al. (2024) introduce a dual-task collaborative mutual promotion framework, which achieves joint optimization of image dehazing and depth estimation through an alternating difference perception. Different from these approaches that use bidirectional guidance to fully fuse guidance and target features, our method focuses more on leveraging the structural correlation between scene priors and depth to mitigate texture interference.

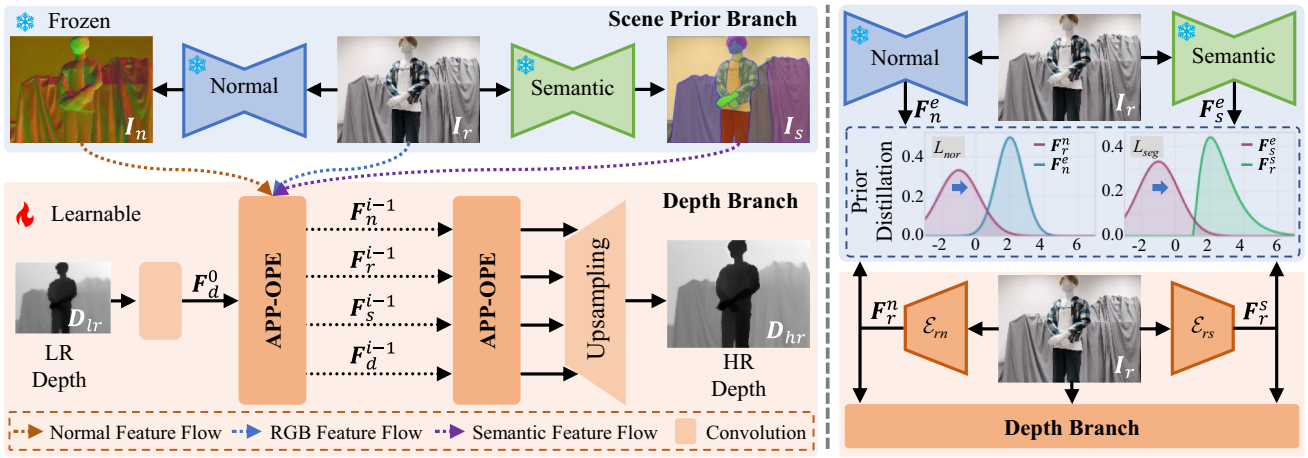


Fig. 2 Overview of SPFNet (left) and its distilled variant SPFNet-D (right). For SPFNet, large-scale models are first employed to generate the normal prior $\boldsymbol{I}_n$ and semantic prior $\boldsymbol{I}_s$ from RGB $\boldsymbol{I}_r$. Then, $\boldsymbol{I}_n$ and $\boldsymbol{I}_s$ are fed together with $\boldsymbol{I}_r$ into multiple iteratively executed APP-OPE modules (APP and OPE are connected in series) to transfer scene prior knowledge into depth features, thereby reconstructing HR depth $\boldsymbol{D}_{hr}$. For SPFNet-D, the encoded features of normal and semantic models are separately extracted, yielding $\boldsymbol{F}_n^e$ and $\boldsymbol{F}_s^e$. Additionally, a prior distillation regularization (consisting of normal term L_{nor} and semantic term L_{sem}) is introduced to distill normal and semantic features into RGB features $\boldsymbol{F}_r^n$ and $\boldsymbol{F}_r^s$, effectively reducing the computational overhead of large-scale models during inference. The distilled $\boldsymbol{F}_r^n$ and $\boldsymbol{F}_r^s$ replace the normal and semantic feature flows as input to the depth branch, whose architecture is kept identical to that of the original SPFNet

2.5 Image Super-Resolution Quality Assessment

Recently, numerous image quality assessment methods have been proposed to automatically evaluate the quality of reconstructed images. Zhou and Wang (2022) introduce a super-resolution (SR) image fidelity index that adaptively combines deterministic fidelity and statistical fidelity through an uncertainty weighting scheme. Li et al. (2024) design a full-reference bi-directional attention network for SR image quality assessment, enabling dynamic bidirectional interactions between the SR image generation process and the quality assessment process. In Zhou et al. (2020), develop a no-reference SR image quality assessment approach that adopts a two-stream convolutional network architecture to extract features related to structural degradation and texture distribution changes in SR images, thereby simulating the human visual system's perception of image distortions. More recently, Li et al. (2025) employ a bi-directional attention mechanism to model the bidirectional interaction between SR image generation and evaluation, integrating grouped multi-scale deformable convolution and sub-information excitation convolution to achieve adaptive assessment.

3 Scene Prior Filtering Network

In this section, we begin by introducing the overall architecture of our SPFNet and its distilled variant, termed SPFNet-D. We then provide a detailed description of the proposed all-in-one prior propagation (APP), one-to-one prior embedding (OPE), and mutual guided filtering (MGF). Finally, we present the loss function used for training.

3.1 Network Architecture

A naive guided DSR architecture generally incorporates an RGB guidance branch and a depth recovery branch. In our method, we introduce additional normal and semantic guidance branches, collectively forming the scene prior branch. Subsequently, in the depth recovery branch, the HR depth is progressively restored through multiple stages, under the guidance of the scene prior branch.

As shown in Fig. 2, our method comprises two variants: SPFNet (left) and SPFNet-D (right). Specifically, we first propose the original SPFNet, which leverages surface normal and semantic priors predicted by large-scale models as prompts to mitigate texture interference in RGB input while enhancing the representation of depth edges. To reduce the computational burden introduced by large-scale models during inference, we further introduce SPFNet-D. This variant distills the encoded features from surface normal and semantic models into RGB features to replace the direct use of estimated surface normal and semantic maps. Consequently, SPFNet-D effectively eliminates the additional computational overhead while attenuating texture interference in RGB features and refining depth edges.

SPFNet. As illustrated in Fig. 2 (left), SPFNet mainly consists of two components: the scene prior branch and depth branch. Given RGB $\boldsymbol{I}_r \in R^{sh \times sw \times 3}$ as input, the scene prior branch separately execute pre-trained large-scale models for

surface normal and semantic prediction, obtaining the normal prior $\boldsymbol{I}_n \in R^{sh\times sw\times 3}$ and semantic prior $\boldsymbol{I}_s \in R^{sh\times sw\times 1}$. h and w are the height and width of LR depth, and s is the upsampling factor. Subsequently, $\boldsymbol{I}_r$, $\boldsymbol{I}_n$, and $\boldsymbol{I}_s$ are each mapped into the feature space through 3×3 convolutional layers, generating the RGB feature flow $\boldsymbol{F}_r^0$, the normal feature flow $\boldsymbol{F}_n^0$, and the semantic feature flow $\boldsymbol{F}_s^0$. In the depth branch, the LR depth $\boldsymbol{D}_{lr} \in R^{h\times w\times 1}$ is first encoded through convolutional layers into the initial depth features $\boldsymbol{F}_d^0$. These features, along with the RGB, normal, and semantic feature flows from the scene prior branch, are then fed into multiple APP-OPE modules (where APP and OPE are cascaded). Specifically, APP computes the structural similarity between all scene priors and the depth to filter out structural information in the scene priors that is highly correlated with the depth, thereby preliminarily reducing irrelevant texture interference. Building on this, OPE iteratively takes the scene priors refined by APP, together with the similarity weights, as input to generate filtering kernels without texture interference. This enables the transfer of knowledge from the large model that aligns with the similarity weights to the depth features. As a result, APP-OPE effectively suppress interference from RGB textures while enhancing depth representations, producing enhanced depth features $\boldsymbol{F}_d^i$, normal features $\boldsymbol{F}_n^i$, RGB features $\boldsymbol{F}_r^i$, and semantic features $\boldsymbol{F}_s^i$. Finally, an upsampling module (composed of a transposed convolution layer and a 3×3 convolution layer) is used to aggregate these prior and depth features, thereby reconstructing HR depth $\boldsymbol{D}_{hr} \in R^{sh\times sw\times 1}$:

$$\boldsymbol{D}_{hr} = f_{up}(\sqcup(\boldsymbol{F}_n^i, \boldsymbol{F}_r^i, \boldsymbol{F}_s^i, \boldsymbol{F}_d^i)), \tag{1}$$

where $\sqcup(\cdot)$ represents concatenation operation.

To achieve a better balance between model complexity and reconstruction quality, we introduce a lightweight SPFNet-T. This model reduces the number of channels in all convolutional layers to one-seventh of those in SPFNet while keeping the original network architecture unchanged. As a result, the trainable parameter count of SPFNet-T is approximately 2.09% of the original model.

SPFNet-D. Additionally, we further design a scene prior distillation variant, SPFNet-D, which effectively eliminates additional computational overhead from large-scale models. As shown on the right side of Fig. 2, our SPFNet-D first extracts encoded features from the normal model and the semantic model separately, obtaining $\boldsymbol{F}_n^e$ and $\boldsymbol{F}_s^e$. Then, we use encoders ε_{rn} and ε_{rs} to map the RGB input to features $\boldsymbol{F}_r^n$ and $\boldsymbol{F}_r^s$, where the shapes of $\boldsymbol{F}_r^n$ and $\boldsymbol{F}_r^s$ are consistent with $\boldsymbol{F}_n^e$ and $\boldsymbol{F}_s^e$, respectively. ε_{rn} and ε_{rs} first extract initial RGB features through multiple stacked 3×3 convolutional layers and ReLU activation layers. Then, multiple transposed convolutional layers and convolutional layers with a stride of 2 are composed to ensure the feature resolution is consistent with that of the large model encoder output.

Next, we introduce a prior distillation regularization to transfer normal and semantic knowledge from large-scale models into the RGB features, including a normal term L_{nor} and a semantic term L_{sem}:

$$\begin{aligned} L_{nor} &= \sum_{k=1}^{C} \varphi\left(\frac{\boldsymbol{F}_N^n}{\tau_1}\right) \odot \left[\Gamma\left(\varphi\left(\frac{\boldsymbol{F}_N^n}{\tau_1}\right)\right) - \Gamma\left(\psi\left(\frac{\boldsymbol{F}_N^{nr}}{\tau_1}\right)\right)\right], \\ L_{sem} &= \sum_{k=1}^{C} \varphi\left(\frac{\boldsymbol{F}_N^s}{\tau_2}\right) \odot \left[\Gamma\left(\varphi\left(\frac{\boldsymbol{F}_N^s}{\tau_2}\right)\right) - \Gamma\left(\psi\left(\frac{\boldsymbol{F}_N^{sr}}{\tau_2}\right)\right)\right], \end{aligned} \tag{2}$$

where $\boldsymbol{F}_N^n = f_n(\boldsymbol{F}_n^e(k))$, $\boldsymbol{F}_N^{nr} = f_n(\boldsymbol{F}_r^n(k))$, $\boldsymbol{F}_N^s = f_n(\boldsymbol{F}_s^e(k))$, and $\boldsymbol{F}_N^{sr} = f_n(\boldsymbol{F}_r^s(k))$. C is the number of RGB feature channels, and f_n represents normalization. $\Gamma(\cdot)$ denotes log function. $\varphi(\cdot)$ and $\psi(\cdot)$ are softmax and log softmax functions, respectively. τ_1 and τ_2 are temperature parameters, while $\odot$ refers to element-wise multiplication.

Finally, the distilled RGB features $\boldsymbol{F}_r^n$ and $\boldsymbol{F}_r^s$ are fed into the depth branch to replace the original normal and semantic feature flows of SPFNet. This replacement mitigates texture interference and enhances depth edge representations without introducing additional computational overhead. Furthermore, the depth branch in SPFNet-D is maintained identically to that of the original SPFNet.

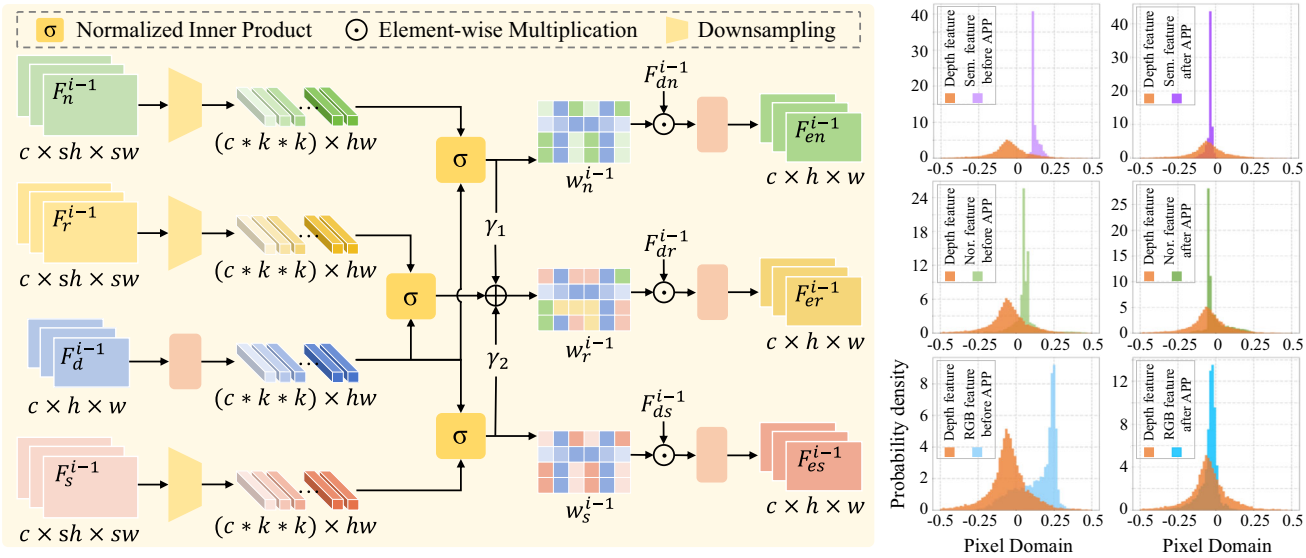
3.2 All-in-one Prior Propagation

The proposed APP weakens the texture interference by using the inherent characteristics of scene priors from large-scale models. It computes the similarity between the multi-modal scene priors (RGB, normal, semantic, and depth), which is then used to mitigate the interference and promote the generation of prior filter kernels. As shown in Fig. 3(a), we first downsample prior features $\boldsymbol{F}_n^{i-1}$, $\boldsymbol{F}_r^{i-1}$, and $\boldsymbol{F}_s^{i-1}$ to match the size of the depth features $\boldsymbol{F}_d^{i-1}$, producing downsampled features $\boldsymbol{F}_{dn}^{i-1}$, $\boldsymbol{F}_{dr}^{i-1}$, and $\boldsymbol{F}_{ds}^{i-1}$.

Next, APP unfolds (by 3×3 kernel) depth features and downsampled scene prior features into patches, denoted as n_j^{i-1}, r_j^{i-1}, s_j^{i-1}, and d_j^{i-1}, where $j \in [1, h\times w]$. Subsequently, we compute the patch similarity σ_j^p between the j-th prior patch p_j^{i-1} (*e.g.*, r_j^{i-1}, n_j^{i-1}, s_j^{i-1}) and depth patch d_j^{i-1} using normalized inner product:

$$\sigma_j^p = \left\langle \frac{p_j^{i-1}}{\left\| p_j^{i-1} \right\|}, \frac{d_j^{i-1}}{\left\| d_j^{i-1} \right\|} \right\rangle. \tag{3}$$

The overall similarity between the scene prior and depth features can be derived by calculating the σ_j^p of all patches,



(a) All-in-one Prior Propagation (b) Histogram comparison

Fig. 3 Scheme of (a) All-in-one Prior Propagation (APP), and (b) histogram comparison of scene prior features. Features $\boldsymbol{F}_{dn}^{i-1}$, $\boldsymbol{F}_{dr}^{i-1}$, and $\boldsymbol{F}_{ds}^{i-1}$ are downsampled from $\boldsymbol{F}_{n}^{i-1}$, $\boldsymbol{F}_{r}^{i-1}$, and $\boldsymbol{F}_{s}^{i-1}$, matching the size of depth features $\boldsymbol{F}_{d}^{i-1}$. The downsampling module consists of strided convolutions and ReLU activation functions

generating normal $\boldsymbol{w}_n^i$, semantic $\boldsymbol{w}_s^i$, and RGB $\boldsymbol{w}_r^i$ weights:

$$\boldsymbol{w}_r^i = \gamma_1 \circ \boldsymbol{w}_n^i + \gamma_2 \circ \boldsymbol{w}_s^i + f_r(\sqcup_{j=1}^{h\times w} \sigma_j^r), \tag{4}$$

where $\boldsymbol{w}_n^i = f_r(\sqcup_{j=1}^{h\times w} \sigma_j^n)$, $\boldsymbol{w}_s^i = f_r(\sqcup_{j=1}^{h\times w} \sigma_j^s)$. $\sqcup(\cdot)$ and $\circ$ represent concatenation and scalar multiplication. γ_1 and γ_2 are learnable constant parameters initialized to zero, which are used to adaptively adjust the contribution of normal and semantic priors similarity to RGB. f_r is a reshape function that transforms input to the dimensions of $\boldsymbol{F}_{dr}^{i-1}$.

Finally, the APP utilizes the similarity weights to attenuate interference in the scene prior features, generating enhanced normal features $\boldsymbol{F}_{en}^{i-1}$, RGB features $\boldsymbol{F}_{er}^{i-1}$, and semantic features $\boldsymbol{F}_{es}^{i-1}$. Fig. 3(b) illustrates a comparison of the distributions between the prior features and depth features before and after using the APP module. It can be found that the output prior features are closer to the distribution of depth features, demonstrating that our APP succeeds in calibrating prior features and reducing interference.

3.3 One-to-One Prior Embedding

The OPE module continuously integrates each single-modal prior (normal, semantic, and RGB) into depth to further reduce interference and enhance edges. Fig. 4(a) shows the main steps of the OPE module. First, given the enhanced scene prior features ($\boldsymbol{F}_{en}^{i-1}$, $\boldsymbol{F}_{er}^{i-1}$, $\boldsymbol{F}_{es}^{i-1}$), similarity weights ($\boldsymbol{w}_n^{i-1}$, $\boldsymbol{w}_r^{i-1}$, $\boldsymbol{w}_s^{i-1}$), and depth features $\boldsymbol{F}_d^{i-1}$, OPE conducts MGF, denoted as f_m, to successively embed each single-modal prior into depth features:

$$\begin{aligned}
\boldsymbol{F}_n^i, \boldsymbol{F}_{nd}^i &= f_m(f_c(\boldsymbol{F}_d^{i-1}), \boldsymbol{F}_{en}^{i-1}, \boldsymbol{w}_n^{i-1}),\\
\boldsymbol{F}_s^i, \boldsymbol{F}_{sd}^i &= f_m(\boldsymbol{F}_{id1}^{i-1}, \boldsymbol{F}_{es}^{i-1}, \boldsymbol{w}_s^{i-1}),\\
\boldsymbol{F}_r^i, \boldsymbol{F}_{rd}^i &= f_m(\boldsymbol{F}_{id2}^{i-1}, \boldsymbol{F}_{er}^{i-1}, \boldsymbol{w}_r^{i-1}),
\end{aligned} \tag{5}$$

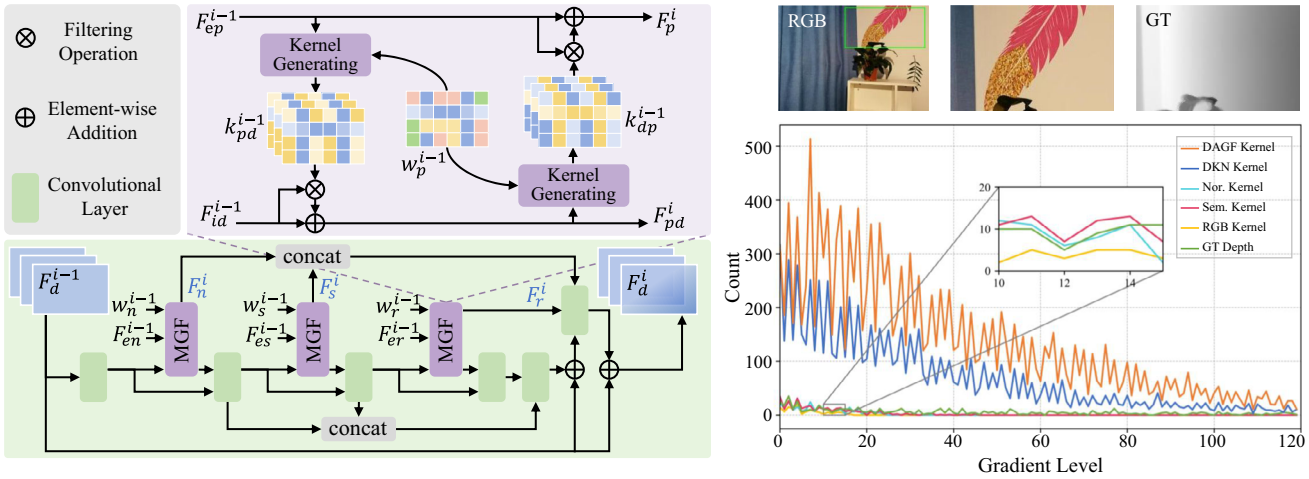
where $\boldsymbol{F}_{id1}^{i-1} = f_c(\sqcup(f_c(\boldsymbol{F}_d^{i-1}), \boldsymbol{F}_{nd}^i))$, $\boldsymbol{F}_{id2}^i = f_c(\sqcup(F_{id1}^i, \boldsymbol{F}_{ns}^i))$. $\boldsymbol{F}_n^i$, $\boldsymbol{F}_s^i$, and $\boldsymbol{F}_r^i$ indicate the filtered normal, semantic, and RGB features, respectively. In addition, $\boldsymbol{F}_{nd}^i$, $\boldsymbol{F}_{sd}^i$, and $\boldsymbol{F}_{rd}^i$ correspond to the filtered depth features of normal-to-depth, semantic-to-depth, and RGB-to-depth, respectively. f_c denotes convolutional layer. Then, OPE aggregates the filtered scene prior and depth features:

$$\boldsymbol{F}_d^i = f_c(\sqcup(\boldsymbol{F}_{dn}^i, \boldsymbol{F}_{ds}^i, \boldsymbol{F}_{dr}^i, \boldsymbol{F}_{fd}^i)) + \boldsymbol{F}_d^{i-1}, \tag{6}$$

where $\boldsymbol{F}_d^i$ represents the enhanced depth features. $\boldsymbol{F}_{fd}^i = f_c(\sqcup(\boldsymbol{F}_{if1}^i, \boldsymbol{F}_{if2}^i, f_c(\boldsymbol{F}_{rd}^i, \boldsymbol{F}_{if2}^i))) + \boldsymbol{F}_d^{i-1}$.

Mutual Guided Filtering. In contrast to existing guided filtering methods (Dong et al., 2022; Kim et al., 2021; Zhang & Wu, 2023; Zhong et al., 2023), our MGF focuses more on employing scene priors and similarity weights to reduce texture interference within the filter kernel and enhance the accuracy of the edge representations. As depicted in the purple part of Fig. 4(a), it includes both prior-to-depth filtering (P2D) and depth-to-prior filtering (D2P).

For P2D, given the APP-enhanced scene prior features $\boldsymbol{F}_{ep}^{i-1}$ (*e.g.*, $\boldsymbol{F}_{en}^{i-1}$, $\boldsymbol{F}_{er}^{i-1}$, $\boldsymbol{F}_{es}^{i-1}$), similarity weights $\boldsymbol{w}_p^{i-1}$ (*e.g.*,



(a) One-to-one Prior Embedding (b) Gradient histogram

Fig. 4 Scheme of (a) One-to-one Prior Embedding (OPE), and (b) gradient histogram of filter kernels in the texture area (green box). The surface normal, semantic, and RGB kernels are generated by our Mutual Guided Filtering (MGF)

$\boldsymbol{w}_n^{i-1}$, $\boldsymbol{w}_r^{i-1}$, $\boldsymbol{w}_s^{i-1}$), and depth features $\boldsymbol{F}_{id}^{i-1}$ as inputs, MGF first executes the kernel generator to construct a scene prior filtering kernel, denoted as $\boldsymbol{k}_{pd}^{i-1}$. The kernel generator consists of a 1×1 convolutional layer, a ReLU activation function layer, and a normalization operation. Furthermore, the prior-to-depth Filter Kernel Generator and the depth-to-prior Filter Kernel Generator share an identical network.

These kernels are subsequently applied to filter the depth features $\boldsymbol{F}_{id}^{i}$, thereby enabling the transfer of high-frequency components from the scene priors to the depth. The filtered depth features $\boldsymbol{F}_{pd}^{i}$ is described as:

$$\boldsymbol{F}_{pd}^{i} = \boldsymbol{F}_{id}^{i-1} \otimes \boldsymbol{k}_{pd}^{i-1} + \boldsymbol{F}_{id}^{i-1}, \tag{7}$$

where $\boldsymbol{k}_{pd}^{i-1} = f_{kg}(\boldsymbol{F}_{ep}^{i-1}, \boldsymbol{w}_p^{i-1})$. f_{kg} stands for kernel generating module, consisting of a 1×1 convolution and an activation function. $\otimes$ is the filtering operation.

Similarly, for depth-to-prior filtering, $\boldsymbol{w}_p^{i-1}$ and the filtered depth features $\boldsymbol{F}_{pd}^{i}$ are first fed into f_{kg}, generating the depth filter kernel $\boldsymbol{k}_{dp}^{i-1}$. Then, MGF filters prior features $\boldsymbol{F}_{ep}^{i-1}$ to preserve the structure required for the depth and further attenuate interference. The filtered prior features $\boldsymbol{F}_{p}^{i}$ is defined as:

$$\boldsymbol{F}_{p}^{i} = \boldsymbol{F}_{ep}^{i-1} \otimes \boldsymbol{k}_{dp}^{i-1} + \boldsymbol{F}_{ep}^{i-1}, \tag{8}$$

where depth filter kernel $\boldsymbol{k}_{dp}^{i-1} = f_{kg}(\boldsymbol{F}_{pd}^{i}, \boldsymbol{w}_p^{i-1})$.

As illustrated in Fig. 4(b), we present the gradient histogram comparisons of filter kernels. Notably, the surface normal, semantic, and RGB kernels predicted by our method exhibit less gradient variations and are closer to the ground truth depth than other methods. These results further demonstrate that our MGF can efficiently mitigate texture interference and enhance edge representations.

3.4 Loss Functions

Given the predicted depth $\boldsymbol{D}_{hr}$ and ground-truth depth $\boldsymbol{D}_{gt}$ as input, we introduce a common reconstruction loss to optimize our SPFNet:

$$L_{total1} = \sum_{q\in Q} \left\| \boldsymbol{D}_{gt}^{q} - \boldsymbol{D}_{hr}^{q} \right\|_1, \tag{9}$$

where Q is the set of valid pixels of $\boldsymbol{D}_{gt}$. $\|\cdot\|_1$ is L_1 norm.

Furthermore, to train the distillation variant SPFNet-D, we incorporate the reconstruction loss from original SPFNet with prior distillation regularization (as defined in Eq. (2)):

$$L_{total2} = \sum_{q\in Q} \left\| \boldsymbol{D}_{gt}^{q} - \boldsymbol{D}_{hr}^{q} \right\|_1 + \lambda_1 L_{nor} + \lambda_2 L_{sem}, \tag{10}$$

where λ_1 and λ_2 are hyperparameters, both set to 0.001.

4 Experimental Results

In this section, we conduct extensive experiments to evaluate the performance of our proposed SPFNet. Specifically, we first elaborate on the experimental setup and implementation details. Then, we present the comparison results with previous state-of-the-art methods across several experiments: synthetic DSR, real-world DSR, model complexity analysis, and joint DSR with denoising. To further demonstrate the

Table 1 Quantitative comparisons on synthetic DSR benchmarks. Following prior methods (He et al., 2021; Zhao et al., 2022), RMSE in centimeters is used as the evaluation metric, with lower values indicating better performance. The **best** and second-best results are marked

Methods	NYU-v2			RGB-D-D			Lu			Middlebury			Venue
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	
Bicubic downsampling													
DJF Li et al. (2016)	2.80	5.33	9.46	3.41	5.57	8.15	1.65	3.96	6.75	1.68	3.24	5.62	ECCV 2016
DSRNet Guo et al. (2018)	3.00	5.16	8.41	-	-	-	1.77	3.10	5.11	1.77	3.05	4.96	TIP 2018
DJFR Li et al. (2019)	2.38	4.94	9.18	3.35	5.57	7.99	1.15	3.57	6.77	1.32	3.19	5.57	PAMI 2019
PAC Su et al. (2019)	1.89	3.33	6.78	1.25	1.98	3.49	1.20	2.33	5.19	1.32	2.62	4.58	CVPR 2019
CUNet Deng and Dragotti (2020)	1.92	3.70	6.78	1.18	1.95	3.45	0.91	2.23	4.99	1.10	2.17	4.33	PAMI 2020
DKN Kim et al. (2021)	1.62	3.26	6.51	1.30	1.96	3.42	0.96	2.16	5.11	1.23	2.12	4.24	IJCV 2021
FDKN Kim et al. (2021)	1.86	3.58	6.96	1.18	1.91	3.41	0.82	2.10	5.05	1.08	2.17	4.50	IJCV 2021
FDSR He et al. (2021)	1.61	3.18	5.86	1.16	1.82	3.06	1.29	2.19	5.00	1.13	2.08	4.39	CVPR 2021
GraphSR De Lutio et al. (2022)	1.79	3.17	6.02	1.30	1.83	3.12	0.92	2.05	5.15	1.11	2.12	4.43	CVPR 2022
SUFT Shi et al. (2022)	1.12	2.51	4.86	1.10	1.69	2.71	1.10	1.74	3.92	1.07	1.75	3.18	MM 2022
DCTNet Zhao et al. (2022)	1.59	3.16	5.84	1.08	1.74	3.05	0.88	1.85	4.39	1.10	2.05	4.19	CVPR 2022
DAGF Zhong et al. (2023)	1.36	2.87	6.06	1.14	1.76	2.82	0.83	1.93	4.80	1.15	1.80	3.70	TNNLS 2023
RSAG Yuan et al. (2023)	1.23	2.51	5.27	1.14	1.75	2.96	**0.79**	1.67	4.30	1.13	2.74	3.55	AAAI 2023
DADA Metzger et al. (2023)	1.54	2.74	4.80	1.20	1.83	2.80	0.96	1.87	4.01	1.20	2.03	4.18	CVPR 2023
SSDNet Zhao et al. (2023)	1.60	3.14	5.86	**1.04**	1.72	2.92	0.80	1.82	4.77	**1.02**	1.91	4.02	ICCV 2023
SGNet Wang et al. (2024)	1.10	2.44	4.77	1.10	**1.64**	2.55	1.03	1.61	3.55	1.15	1.64	2.95	AAAI 2024
DORNet Wang et al. (2025)	1.19	2.70	5.60	1.15	1.80	2.97	0.92	1.75	4.41	1.05	1.76	3.48	CVPR 2025
SPFNet-T	1.52	3.03	5.71	1.16	1.77	2.90	0.80	1.69	4.31	1.04	1.77	3.36	-
SPFNet-D	**1.09**	2.39	4.67	1.12	1.71	2.56	0.92	1.57	3.22	1.06	1.59	2.87	-
SPFNet	**1.09**	**2.36**	**4.55**	1.13	1.71	**2.53**	0.90	**1.56**	**3.20**	1.05	**1.57**	**2.79**	-
Nearest-neighbor downsampling													
DJF Li et al. (2016)	3.54	6.20	10.21	2.14	3.32	4.92	2.54	4.71	7.66	2.14	3.77	6.12	ECCV 2016
DSRNet Guo et al. (2018)	3.49	5.70	9.76	-	-	-	2.57	4.46	6.45	2.08	3.26	5.78	TIP 2018
DJFR Li et al. (2019)	3.38	5.86	10.11	1.90	3.11	4.89	2.22	4.54	7.48	1.98	3.61	6.07	PAMI 2019
PAC Su et al. (2019)	2.82	5.01	8.64	-	-	-	2.48	4.37	6.60	1.91	3.20	5.60	CVPR 2019
CUNet Deng and Dragotti (2020)	4.09	6.25	10.23	1.85	3.07	5.01	2.15	4.33	7.72	2.06	3.97	6.36	PAMI 2020
DKN Kim et al. (2021)	2.46	4.76	8.50	1.92	2.91	4.46	2.35	4.16	6.33	1.93	3.17	5.49	IJCV 2021
FDKN Kim et al. (2021)	2.62	4.99	8.67	1.84	2.93	4.76	2.64	4.55	7.20	2.21	3.64	6.15	IJCV 2021
FDSR He et al. (2021)	2.50	4.62	7.77	1.83	2.77	4.07	2.17	3.97	6.51	1.85	2.97	5.31	CVPR 2021
DCTNet Zhao et al. (2022)	2.56	4.89	9.11	1.86	2.86	4.28	2.31	4.17	6.69	1.81	2.96	5.39	CVPR 2022
SUFT Shi et al. (2022)	2.05	4.11	7.26	1.85	2.79	3.95	2.07	3.63	6.16	1.76	2.76	5.16	MM 2022
DAGF Zhong et al. (2023)	2.35	4.62	7.81	1.78	**2.65**	3.95	1.96	3.81	6.16	1.78	2.73	4.75	TNNLS 2023
RSAG Yuan et al. (2023)	2.67	4.73	7.66	1.80	2.79	4.01	2.11	3.91	6.32	1.73	2.98	5.05	AAAI 2023
SGNet Wang et al. (2024)	1.94	4.04	7.31	1.81	2.66	**3.74**	1.98	3.58	5.82	1.72	2.76	4.43	AAAI 2024
DORNet Wang et al. (2025)	2.02	4.10	7.55	1.89	2.85	4.03	2.08	3.72	6.14	1.73	2.72	4.68	CVPR 2025
SPFNet-T	2.53	4.83	8.24	1.88	2.90	4.17	2.35	4.26	6.66	1.77	2.96	4.91	-
SPFNet-D	**1.92**	**3.96**	7.10	1.84	2.74	3.83	2.07	3.65	5.78	1.71	2.64	**4.39**	-
SPFNet	2.24	**3.96**	**6.74**	**1.76**	2.76	**3.74**	**1.94**	**3.46**	**5.48**	**1.65**	**2.62**	**4.39**	-

generalization capability of our approach, we also conduct experiments on other multi-modal image restoration tasks. Finally, we perform multiple ablation studies to comprehensively evaluate the effectiveness of our approach.

4.1 Experimental Setups and Implementation Details

We carry out extensive experiments on five DSR datasets, including NYU-v2 (Silberman et al., 2012), Lu (Lu et

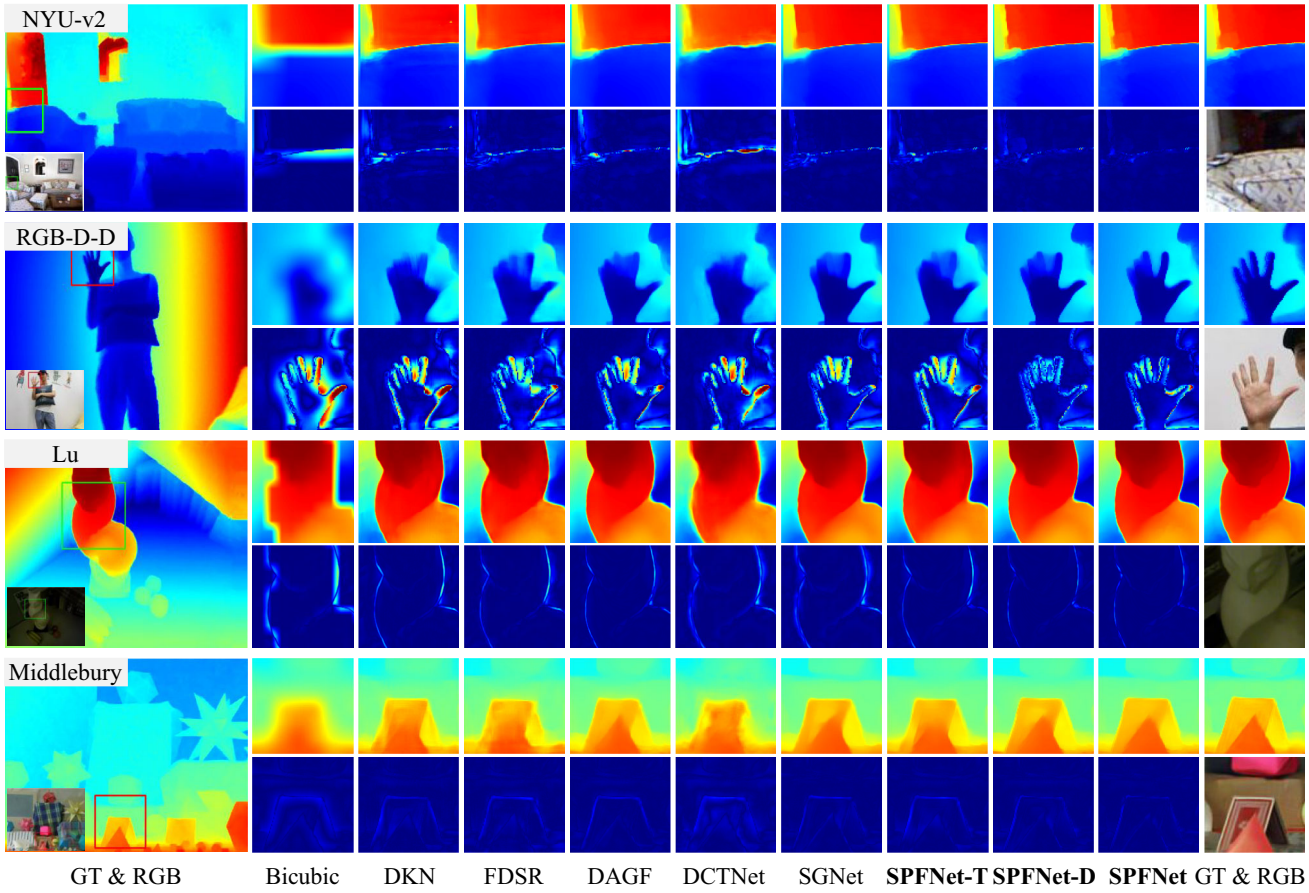


Fig. 5 Visual results and error maps on four synthetic datasets (×16). Brighter colors in error maps indicate larger errors

al., 2014), Middlebury (Hirschmuller & Scharstein, 2007; Scharstein & Pal, 2007), RGB-D-D (He et al., 2021), and TOFDSR (Yan et al., 2025). We utilize the Root Mean Square Error (RMSE) metric in centimeters to evaluate the DSR methods (Kim et al., 2021; Shi et al., 2022; Zhao et al., 2023). During training, the scene priors and GT depth are cropped to 256 × 256. We employ the Adam (Kingma & Ba, 2014) optimizer with an initial learning rate of 1×10^{-4} to train our method. The proposed model is implemented in PyTorch with a single NVIDIA RTX 4090.

4.2 Comparison with the State-of-the-Art DSR methods

We compare our method with state-of-the-art approaches on ×4, ×8, ×16, and ×32 DSR, including DJF (Li et al., 2016), DSRNet (Guo et al., 2018), DJFR (Li et al., 2019), PAC (Su et al., 2019), CUNet (Deng & Dragotti, 2020), DKN (Kim et al., 2021), FDKN (Kim et al., 2021), FDSR (He et al., 2021), SUFT (Shi et al., 2022), DCTNet (Zhao et al., 2022), GraphSR (De Lutio et al., 2022), DAGF (Zhong et al., 2023), RSAG (Yuan et al., 2023), DADA (Metzger et al., 2023), SSDNet (Zhao et al., 2023), SGNet (Wang et al., 2024), and DORNet (Wang et al., 2025).

Experimental Results on the Synthetic DSR. To assess the performance of our method on synthetic datasets, we conduct extensive experiments on the NYU-v2, Lu, Middlebury, and RGB-D-D datasets. Similar to previous methods Kim et al. (2021); Zhao et al. (2022); Zhong et al. (2023); Zhao et al. (2023), the LR depth is produced by downsampling the GT depth using bicubic interpolation and nearest-neighbor interpolation, respectively. Our model is trained on the 1, 000 RGB-D pairs of NYU-v2, with the remaining 449 pairs for testing. Furthermore, the model pre-trained on the NYU-v2 dataset is applied directly to evaluate the Lu (6 RGB-D pairs), Middlebury (30 pairs), and RGB-D-D (405 pairs) datasets without any fine-tuning.

Tab. 1 demonstrates that our SPFNet achieves state-of-the-art performance across multiple synthetic datasets under different downsampling (bicubic and nearest-neighbor). Specifically, as shown in Tab. 1, SPFNet surpasses most competing methods on four benchmark datasets under various downsampling. For example, for bicubic downsampling, our method decreases the RMSE (×16) by 0.22cm on NYU-v2 and by 0.35cm on Lu compared to the suboptimal approach.

Table 2 Quantitative comparisons at large-scale factor (×32) using bicubic downsampling

Datasets	DJFR Li et al. (2019)	CUNet Deng and Dragotti (2020)	DKN Kim et al. (2021)	FDSR He et al. (2021)	DCTNet Zhao et al. (2022)	SGNet Wang et al. (2024)	DORNet Wang et al. (2025)	**SPFNet-T**	**SPFNet-D**	**SPFNet**
Lu	9.94	10.97	8.98	8.62	9.28	7.23	8.28	7.32	6.42	**6.39**
NYU-v2	14.12	15.95	12.46	14.19	13.33	11.24	10.97	10.07	8.69	**8.06**
RGB-D-D	6.48	7.75	5.97	5.08	5.99	5.15	5.05	4.80	4.13	**3.97**
Middlebury	8.57	9.23	7.76	7.26	8.22	6.79	6.87	6.09	**5.59**	5.90

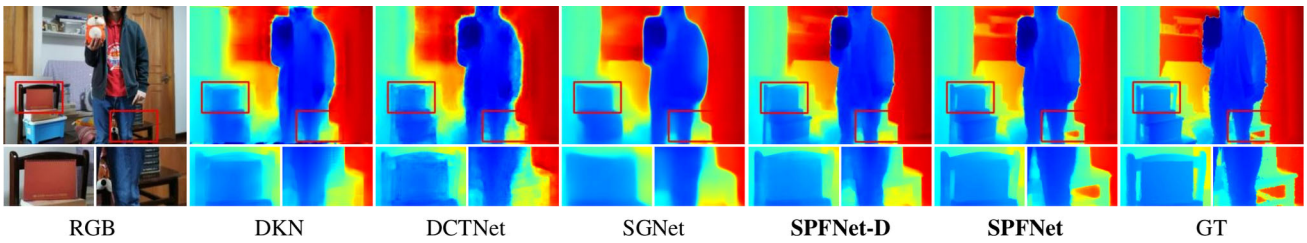


Fig. 6 Visual results on the synthetic RGB-D-D dataset (×32).

Table 3 Quantitative comparisons with existing state-of-the-art methods on the real-world RGB-D-D and TOFDSR datasets

Datasets	DJFR Li et al. (2019)	CUNet Deng and Dragotti (2020)	DKN Kim et al. (2021)	FDSR He et al. (2021)	DCTNet Zhao et al. (2022)	SUFT Shi et al. (2022)	SSDNet Zhao et al. (2023)	SGNet Wang et al. (2024)	**SPFNet-T**	**SPFNet-D**	**SPFNet**
RGB-D-D	5.52	5.84	5.08	5.49	5.43	5.41	5.38	5.32	4.68	**3.71**	4.21
TOFDSR	5.72	6.04	5.50	5.03	5.16	4.37	-	**4.33**	5.13	4.51	4.58

For nearest-neighbor downsampling, SPFNet outperforms the second-best method by 0.52*cm* on ×16 NYU-v2 and 0.34*cm* on ×16 Lu. Additionally, our SPFNet-D also delivers competitive accuracy while maintaining satisfactory efficiency. Specifically, SPFNet-D exceeds the second-best approach with RMSE reductions of 0.10*cm* under bicubic downsampling and 0.16*cm* under nearest-neighbor downsampling on the ×16 NYU-v2 dataset.

Fig. 5 presents the visual results on synthetic datasets with bicubic downsampling. It is evident that our method predicts depth with superior edge accuracy. For example, the edges of the human hand and the sculpture are more distinct than others, and the error maps display smaller errors.

Furthermore, Tab. 2 lists the quantitative comparisons on ×32 DSR, showing that both our SPFNet and SPFNet-D achieve the best performance at this large scale factor. For

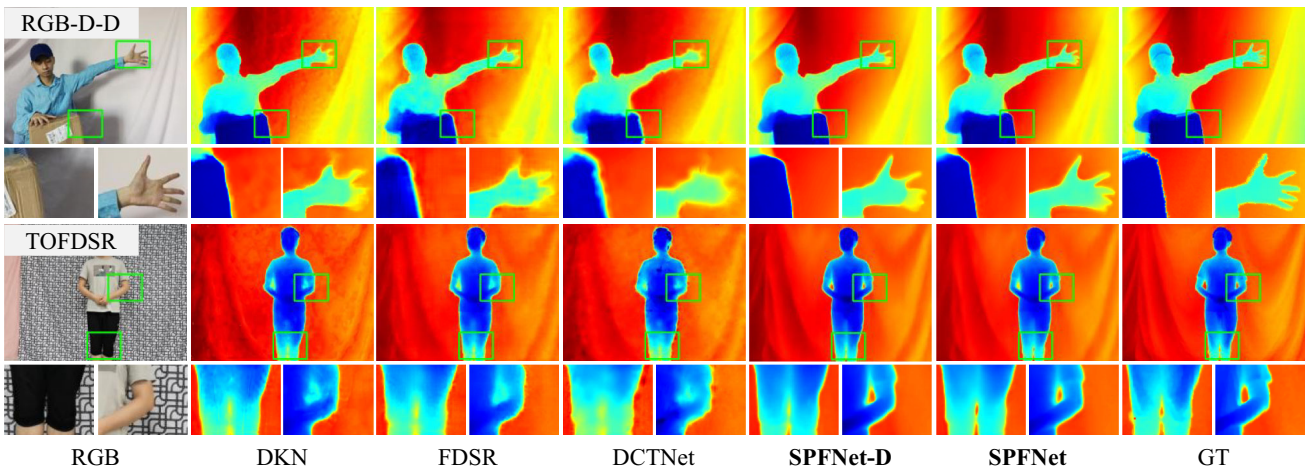


Fig. 7 Visual results on the real-world RGB-D-D and TOFDSR datasets

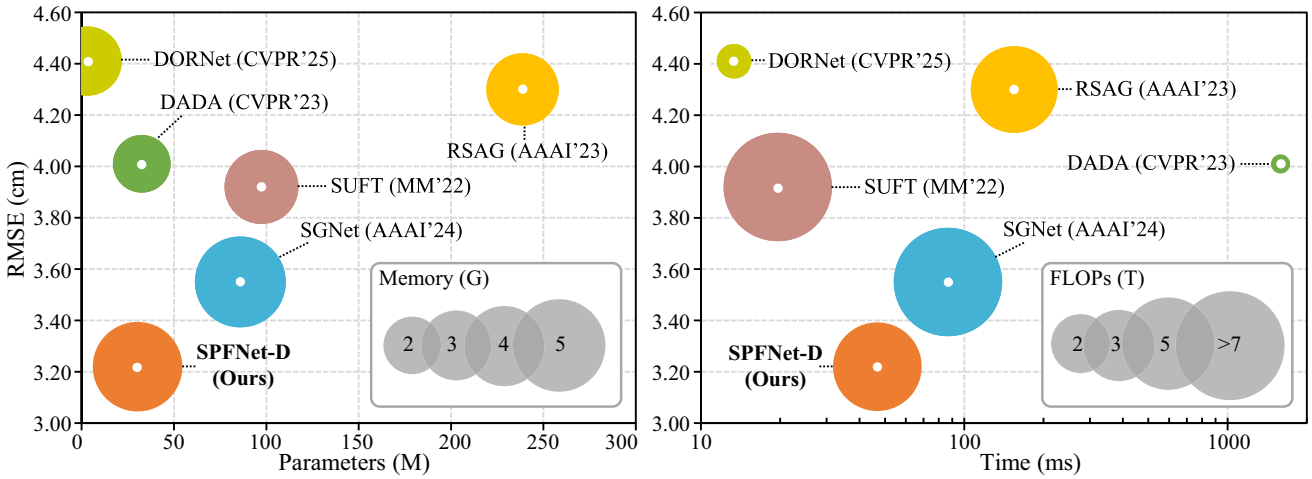


Fig. 8 Model complexity comparisons on the ×16 Lu dataset (bicubic downsampling), where inference time is tested on a single 4090 GPU. Left: Parameters and Memory comparison. Right: Inference time and FLOPs comparison

Table 4 Quantitative comparisons of joint DSR and denoising on four benchmark datasets

Methods	NYU-v2			RGB-D-D			Lu			Middlebury		
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16
Gaussian noise to LR depth												
DJF Li et al. (2016)	8.53	12.38	17.86	3.06	4.37	6.44	4.87	7.43	11.25	5.31	7.69	10.63
DJFR Li et al. (2019)	7.99	11.65	17.06	2.86	4.16	6.11	4.40	6.98	10.65	5.01	7.28	10.36
CUNet Deng and Dragotti (2020)	7.80	10.72	15.52	2.74	3.98	5.76	4.07	6.54	9.84	4.77	6.81	9.80
DKN Kim et al. (2021)	7.06	10.26	15.40	2.72	3.92	5.89	4.21	6.36	9.81	4.33	6.67	9.77
FDKN Kim et al. (2021)	7.41	10.69	15.72	2.73	4.03	6.09	4.12	6.54	11.63	4.48	6.66	9.86
FDSR He et al. (2021)	5.99	8.68	13.28	2.44	3.48	4.99	3.77	6.24	9.87	-	5.73	8.66
SUFT Shi et al. (2022)	5.46	7.94	12.23	2.46	3.42	4.85	3.59	6.01	9.90	3.51	5.61	8.70
DCTNet Zhao et al. (2022)	6.41	9.81	14.81	2.55	3.68	5.30	3.91	6.27	9.56	3.87	6.06	8.89
SGNet Wang et al. (2024)	5.33	7.95	12.19	2.41	3.30	4.61	**3.29**	5.45	8.27	3.33	5.11	7.88
DORNet Wang et al. (2025)	5.69	8.65	13.70	2.53	3.55	5.21	3.69	6.08	9.34	3.55	5.10	7.72
SPFNet-T	5.94	8.48	12.56	2.47	3.43	4.79	3.60	5.69	8.43	3.49	5.26	7.68
SPFNet-D	5.26	7.86	12.32	2.37	3.32	4.71	3.39	5.39	8.38	3.30	4.99	7.69
SPFNet	**5.14**	**7.44**	**10.96**	**2.36**	**3.29**	**4.47**	3.34	**5.30**	**7.93**	**3.23**	**4.85**	**7.20**
Gaussian noise to both LR depth and RGB image												
DJF Li et al. (2016)	8.91	13.14	18.72	3.13	4.56	6.81	4.98	7.49	11.69	5.43	7.82	10.89
DJFR Li et al. (2019)	8.47	12.48	18.11	2.90	4.32	6.56	4.46	7.07	10.97	5.14	7.50	10.64
CUNet Deng and Dragotti (2020)	8.32	11.49	16.60	2.90	4.04	6.07	4.39	6.43	10.11	5.12	6.97	9.90
DKN Kim et al. (2021)	7.54	11.04	16.77	2.74	4.05	6.34	4.09	6.27	10.39	4.48	6.88	9.99
FDKN Kim et al. (2021)	7.54	11.23	16.92	2.76	4.12	6.37	4.09	6.53	10.59	4.50	6.94	10.12
FDSR He et al. (2021)	6.88	10.30	16.00	2.47	3.70	5.79	3.69	6.20	10.21	-	6.30	9.33
SUFT Shi et al. (2022)	6.65	9.93	15.55	2.39	3.59	5.64	3.49	5.90	10.08	3.75	6.12	9.23
DCTNet Zhao et al. (2022)	7.38	10.99	16.90	2.56	3.87	6.05	3.81	6.18	10.16	4.22	6.65	9.75
SGNet Wang et al. (2024)	6.22	9.92	15.50	2.37	3.55	5.58	3.44	5.58	9.41	3.61	5.92	8.99
DORNet Wang et al. (2025)	6.03	8.69	13.32	2.47	3.49	5.09	3.68	6.01	9.41	3.70	5.47	7.82
SPFNet-T	6.04	8.60	12.67	2.42	3.44	4.77	3.57	5.69	8.76	3.54	5.28	7.93
SPFNet-D	6.59	9.79	15.45	2.40	3.61	5.67	3.33	5.64	9.34	3.66	5.94	8.90
SPFNet	**5.34**	**7.66**	**11.31**	**2.34**	**3.28**	**4.48**	**3.31**	**5.50**	**8.22**	**3.29**	**5.05**	**7.73**

instance, compared to the second-best approach, our SPFNet reduces the RMSE by 2.01cm on the NYU-v2 dataset and by 1.08cm on the RGB-D-D dataset.

Furthermore, Fig. 6 presents a visual comparison on $\times 32$ RGB-D-D dataset, demonstrating that our method maintains satisfactory performance even at large scale factors. For instance, the structure and edges of the chair predicted by our SPFNet and SPFNet-D in Fig. 6 are more accurate and closer to the ground truth compared to other methods.

Experimental Results on the Real-world DSR. We perform experiments on real-world RGB-D-D and TOFDSR datasets to evaluate the performance of DSR methods in real environments, where LR and ground-truth depth are captured in real-world scenes. Specifically, RGB-D-D consists of 2, 215 RGB-D pairs for training and 405 pairs for testing. Additionally, we employ the colorization approach (Levin et al., 2004) to fill in the raw LR depth in TOFDC (Yan et al., 2024) as new LR depth input, resulting in the TOFDSR dataset, which includes 10K RGB-D pairs in the training set and 560 pairs in the test set.

As shown in Tab. 3, our method achieves the best performance on the RGB-D-D dataset and delivers competitive results on the TOFDSR dataset. For example, SPFNet-D reduces RMSE by 30.26% compared to the second-best SGNet on RGB-D-D. Unlike synthetic datasets, LR depth captured in real-world environments often exhibits severe structural distortion. Fig. 7 shows the visual results on the real-world RGB-D-D and TOFDSR, demonstrating that our approach recovers more accurate edges compared to others. For instance, the edges of the hand (RGB-D-D) and the arm (TOFDSR) predicted by SPFNet in Fig. 7 are more closely aligned with the GT depth. Overall, these quantitative and visual results indicate that our method can effectively improve DSR performance in real-world scenarios.

Model Complexity Analysis. To fairly validate the effectiveness of our method, Fig. 8 presents a comparative analysis between SPFNet-D and previous state-of-the-art methods under identical experimental settings, evaluating performance, parameters, memory usage, inference time, and FLOPs. It can be clearly observed that our method achieves optimal performance while maintaining competitive computational costs. For example, compared with the second-best SGNet, our method achieves a 9.30% reduction in RMSE while maintaining comparable memory usage, and significantly decreases the number of parameters by 64.75%, inference time by 46.06%, and FLOPs by 68.59%.

Joint DSR and Denoising. Tab. 4 demonstrates that our method outperforms other approaches in joint DSR and denoising on four benchmark datasets using bicubic downsampling. Depth obtained in real-world environments is often noisy, which poses a challenge to HR depth restoration. Similar to the existing methods (Kim et al., 2021; Shin et al., 2023; Zhong et al., 2023), we first add Gaussian noise (mean 0 and standard deviation 0.07) to the LR depth as a new input. As evidenced by Tab. 4 (top), compared to the second-best method, SPFNet reduces the RMSE ($\times 16$) by 1.23cm on the NYU-v2 dataset and by 0.68cm on the Middlebury dataset. Fig. 9 (Noisy depth) presents the visual comparison on the $\times 8$ Middlebury dataset. Notably, both our SPFNet and SPFNet-D demonstrate strong noise robustness, enabling them to recover highly precise and sharp depth from noisy environments. For instance, the edges of the toy and the cord restored by our method are clearer compared to other approaches.

Following the previous method (Shin et al., 2023), we further introduce Gaussian noise to both the LR depth and RGB images as new inputs to simulate challenging real-world environments. As shown in Tab. 4 (bottom), our SPFNet surpasses the suboptimal approach ($\times 16$) by 4.19cm RMSE on the NYU-v2 dataset and by 1.19cm RMSE on the Lu dataset. In addition, although our SPFNet-D exhibits a performance drop compared to the original SPFNet, it still achieves results comparable to those of recent advanced methods (*e.g.*, SUFT (Shi et al., 2022), DCTNet (Zhao et al., 2022), and SGNet (Wang et al., 2024)) without incurring the additional computational cost of large-scale models. As illustrated in Fig. 9 (Noisy RGB-D), the edges of the bowling pin reconstructed by our method are visibly more accurate and closer to the GT depth. In summary, these results thoroughly confirm the robustness of our method.

4.3 Generalization on Other Restoration Tasks

To thoroughly validate the generalization capability of our method, we further conduct experiments on other guided image restoration tasks, including pan-sharpening, saliency map super-resolution, and depth completion, while keeping the overall architecture of our method unchanged.

Pan-Sharpening. Tab. 5 reports the superiority of SPFNet over advanced Pan-Sharpening methods on WorldView III and GaoFen2 datasets, including GFPCA (Liao et al., 2015), PanNet (Yang et al., 2017), MSDCNN (Yuan et al., 2018), SRPPNN (Cai & Huang, 2020), GPPNN (Xu et al., 2021), MutInf (Zhou et al., 2022), and PanFlow (Yang et al., 2023). Since the ground truth itself is not available, we follow previous methods (Yang et al., 2023; Yuan et al., 2018; Zhou et al., 2022) to generate synthetic datasets by using the Wald protocol tool (Wald et al., 1997). In the training and test sets, PAN images and LR multi-spectral images are cropped to sizes of 128×128 and 32×32, respectively. Additionally, PSNR, SSIM, SAM (Yuhas et al., 1992), and ERGAS (Alparone et al., 2007) are employed as image quality assessment metrics to evaluate our experimental results. Compared to the suboptimal method, we can clearly see that our SPFNet increases the PSNR by 0.1879dB on the WorldView III dataset and 0.1776dB on the WGaoFen2 dataset.

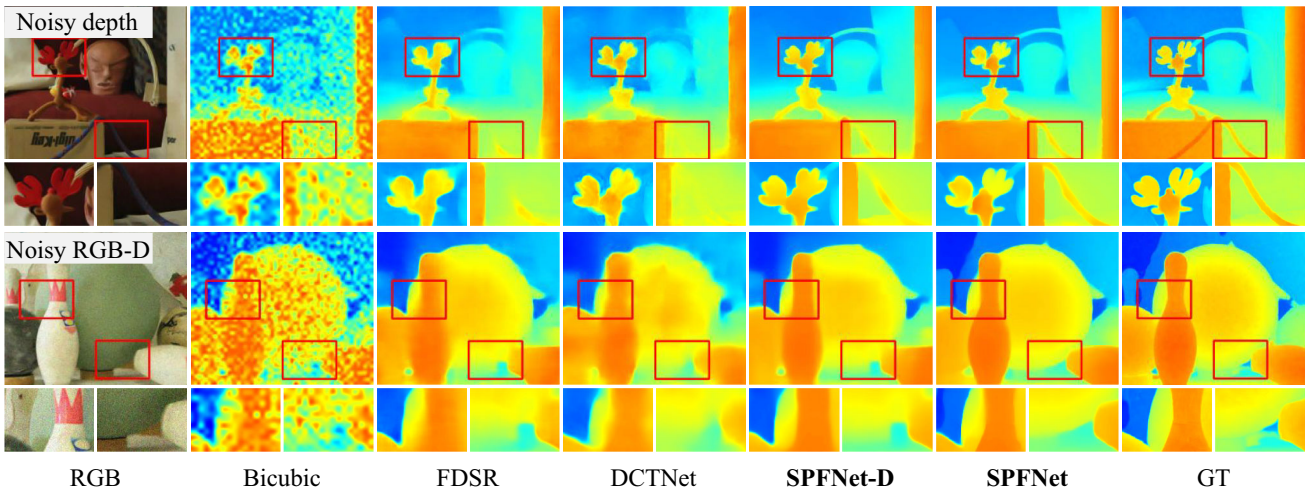


Fig. 9 Visual results of joint DSR and denoising on the ×8 Middlebury dataset

Table 5 Quantitative comparisons of Pan-Sharpening on WorldView III and GaoFen2 datasets

Methods	WorldView III				GaoFen2			
	PSNR↑	SSIM↑	SAMYuhas et al. (1992)↓	ERGAS Alparone et al. (2007)↓	PSNR↑	SSIM↑	SAMYuhas et al. (1992)↓	ERGAS Alparone et al. (2007)↓
GFPCA Liao et al. (2015)	22.3344	0.4826	0.1294	8.3964	37.9443	0.9204	0.0314	1.5604
PanNet Yang et al. (2017)	29.6840	0.9072	0.0851	3.4263	43.0659	0.9685	0.0178	0.8577
MSDCNN Yuan et al. (2018)	30.3038	0.9184	0.0782	3.1884	45.6874	0.9827	0.0135	0.6389
SRPPNN Cai and Huang (2020)	30.4346	0.9202	0.0770	3.1553	47.1998	0.9877	0.0106	0.5586
GPPNN Xu et al. (2021)	30.1785	0.9175	0.0776	3.2593	44.2145	0.9815	0.0137	0.7361
MutInf Zhou et al. (2022)	30.4907	0.9223	**0.0749**	3.1125	47.3042	0.9892	0.0102	0.5481
PanFlow Yang et al. (2023)	30.4873	0.9221	0.0751	3.1142	47.2533	0.9884	0.0103	0.5512
SPFNet	**30.6786**	**0.9244**	0.0769	**3.0447**	**47.4818**	**0.9895**	**0.0101**	**0.5441**

Table 6 Quantitative comparisons of saliency map super-resolution on DUT-OMRON (Yang et al., 2023). The metric is Fscore

Fscore	DJFR Li et al. (2019)	CUNet Deng and Dragotti (2020)	DKN Kim et al. (2021)	FDKN Kim et al. (2021)	FDSR He et al. (2021)	SUFT Shi et al. (2022)	DCTNet Zhao et al. (2022)	SGNet Wang et al. (2024)	SPFNet
×4	0.9858	0.9863	0.9917	0.9931	-	0.9948	0.9874	0.9951	**0.9967**
×8	0.9599	0.9497	0.9389	0.9789	0.9819	0.9825	0.9619	0.9829	**0.9844**
×16	0.8975	0.8972	0.9566	0.9543	0.9549	**0.9584**	0.9003	0.9483	0.9475

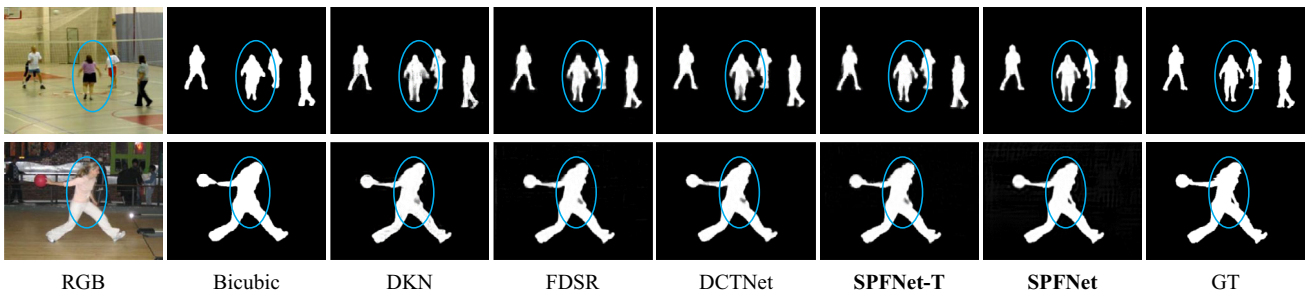
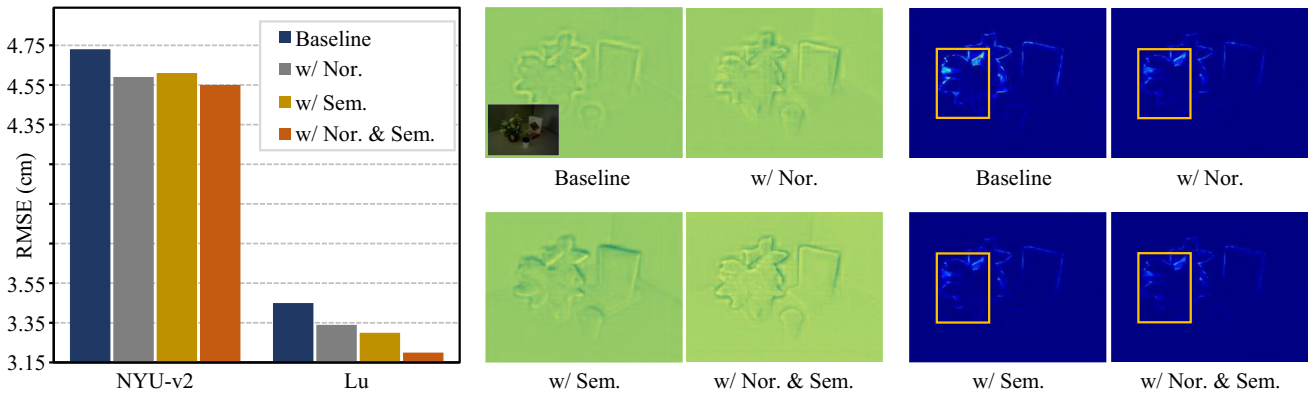


Fig. 10 Visual results of saliency map super-resolution on the DUT-OMRON dataset (×8)

Table 7 Quantitative comparisons of depth completion on the real-world TOFDC dataset

Metrics	CSPN Cheng et al. (2018)	FusionNet Van Gansbeke et al. (2019)	GuideNet Tang et al. (2020)	NLSPN Park et al. (2020)	CFormer Zhang et al. (2023)	RigNet Yan et al. (2022)	PointDC Yu et al. (2023)	TPVD Yan et al. (2025)	**SPFNet-D**	**SPFNet**
RMSE ↓	22.4	11.6	14.6	17.4	11.3	13.3	10.9	9.2	8.1	**8.0**
REL ↓	0.042	0.024	0.030	0.029	0.029	0.025	0.021	0.014	**0.013**	0.014
$\delta_{1.25}$ ↑	94.5	98.3	97.6	96.4	99.1	97.6	98.5	99.1	**99.2**	**99.2**
$\delta_{1.25^2}$ ↑	95.3	99.4	98.9	97.9	99.6	99.1	99.2	99.6	99.6	**99.7**
$\delta_{1.25^3}$ ↑	96.5	99.7	99.5	98.9	**99.9**	99.7	99.6	**99.9**	99.8	**99.9**



(a) Quantitative comparisons (b) Intermediate features (c) Error maps

Fig. 11 Ablation study of surface normal and semantic on ×16 DSR. The features and error maps are derived from Lu

Table 8 Complexity comparisons on NYU-v2 and Lu (×16), where inference time is tested on NYU-v2

Methods	w/o Nor. & Sem.				w/ Nor. & Sem.			
	NYU-v2	Lu	Params (M)	Time (ms)	NYU-v2	Lu	Params (M)	Time (ms)
FDSR He et al. (2021)	5.86	5.00	0.60	**14.13**	5.76 (−0.10)	4.59 (−0.41)	0.67 (+0.07)	**14.57** (+0.44)
SUFT Shi et al. (2022)	4.86	3.92	94.36	15.65	4.81 (−0.05)	4.07 (+0.15)	99.25 (+4.89)	28.58 (+12.93)
SGNet Wang et al. (2024)	4.77	3.55	85.94	88.81	4.70 (−0.07)	3.28 (−0.27)	86.89 (+0.95)	102.35 (+13.54)
SPFNet-T	5.80	4.44	**0.57**	26.47	5.71 (−0.09)	4.31 (−0.13)	**0.65** (+0.08)	29.57 (+3.10)
SPFNet	**4.73**	**3.45**	26.99	34.02	**4.55** (−0.18)	**3.20** (−0.25)	31.10 (+4.11)	51.59 (+17.57)

Saliency Map Super-resolution. Tab. 6 performs ×4, ×8, and ×16 experiments on the DUT-OMRON (Yang et al., 2023) dataset, which comprises 5, 168 pairs of RGB and saliency maps, where the LR saliency maps are generated by downsampling the GT saliency maps using bicubic interpolation. Following (Kim et al., 2021; Zhong et al., 2023), the pre-trained model on the NYU-v2 dataset is directly tested on DUT-OMRON without any fine-tuning. Besides, we employ the F-measure as an evaluation metric to maintain consistency with previous methods. Tab. 6 shows that our method achieves excellent performance. For example, our SPFNet surpasses the suboptimal approach by 0.0016 in the ×8 result. Furthermore, Fig. 10 presents the visual results, showing that our method can predict high-quality HR saliency maps with clearer edges than others.

Depth Completion. Tab. 7 compares our method with previous state-of-the-art depth completion (DC) approaches on the real-world TOFDC (Yan et al., 2024) dataset, including CSPN (Cheng et al., 2018), FusionNet (Van Gansbeke et al., 2019), GuideNet (Tang et al., 2020), NLSPN (Park et al., 2020), CFormer (Zhang et al., 2023), RigNet (Yan et al., 2022), PointDC (Yu et al., 2023), and TPVD (Yan et al., 2025). Consistent with these DC methods, we select RMSE, REL, and $\delta_{1.25^x}(x = 1, 2, 3)$ as evaluation metrics. It can be observed that both our SPFNet and SPFNet-D achieve satisfactory performance across all evaluation metrics, surpassing methods specifically designed for the DC task. For example, compared with the suboptimal TPVD (Yan et al., 2025) and PointDC (Yu et al., 2023), our SPFNet reduces RMSE by $1.2cm$ and $2.9cm$, respectively.

Table 9 Ablation study of SPFNet with different normal and semantic models on $\times 16$ DSR

Methods	Semantic models		Normal models		Datasets	
	MobileSAM Zhang et al. (2023)	SAM Kirillov et al. (2023)	Metric3Dv2 Hu et al. (2024)	Omnidata Eftekhar et al. (2021)	NYU-v2	Lu
(a)	✓		✓		4.56	3.30
(b)	✓			✓	4.61	3.33
(c)		✓	✓		4.56	3.25
(d)		✓		✓	**4.55**	**3.20**

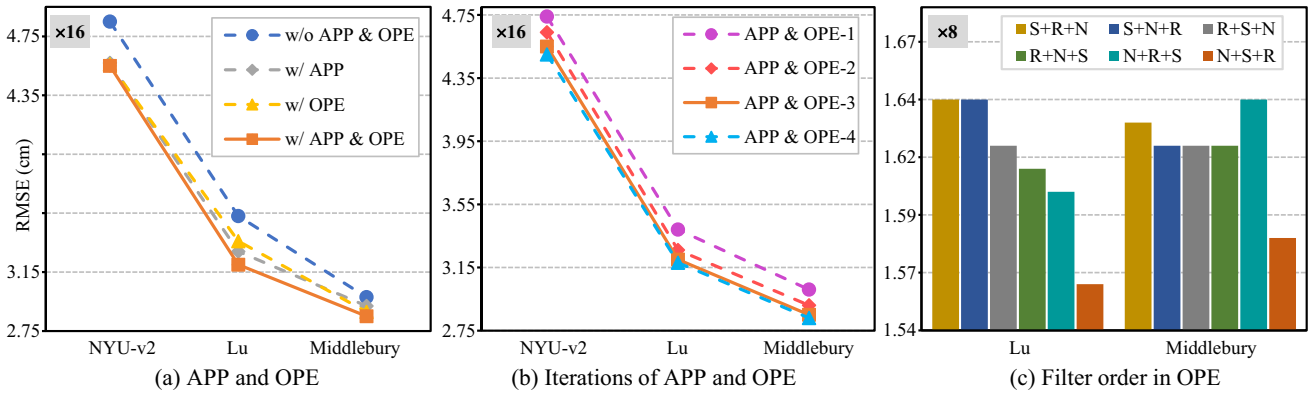


Fig. 12 Ablation study of APP and OPE. S+R+N refers to the filter in the order of semantic (S), RGB (R), and normal (N)

4.4 Ablation Studies

In this section, all ablation studies are conducted based on SPFNet on the synthetic dataset with bicubic downsampling.

Surface Normal and Semantic priors. Fig. 11 shows the ablation results on surface normal and semantic priors. For the baseline, similar to (He et al., 2021; Wang et al., 2024; Zhao et al., 2022), only RGB is utilized as guidance. From Fig. 11(a), we find that both surface normal and semantic priors contribute to a decrease in RMSE. When both are employed together, SPFNet achieves the best performance. For example, compared to the baseline, surface normal and semantic priors reduce the RMSE by $0.11cm$ and $0.15cm$ on the Lu dataset, respectively. Finally, SPFNet outperforms the baseline by $0.25cm$ on the Lu dataset.

Fig. 11(b) and (c) illustrate the visual results of intermediate depth features and error maps. We observe that both surface normal and semantic priors are capable of producing more distinctive edges and fewer errors than the baseline. Furthermore, when combining them, our SPFNet generates a much clearer structure and the much lesser errors. These results suggest that, by leveraging scene priors, our SPFNet can effectively enhance depth edges and minimize errors.

Tab. 8 provides the comparison of incorporating surface normal and semantic priors into previous advanced methods, further validating the effectiveness of our SPFNet. Specifically, for previous approaches, we add additional branches for both surface normal and semantic priors. These branches share the same architecture as the RGB branch present in their original networks. For one thing, as demonstrated by the second column (w/o Nor. & Sem.) in Tab. 8, our SPFNet achieves the best performance and competitive complexity even without relying on surface normal and semantic priors. For example, SPFNet is $0.1cm$ lower compared to the sub-optimal methods on the Lu dataset. For another thing, when compared to the second column, the third column (w/ Nor. & Sem.) in Tab. 8 indicates that nearly all DSR approaches benefit from integrating surface normal and semantic priors. Notably, our method still outperforms others. For instance, compared to the second-best method (SGNet), our SPFNet not only exhibits lower trainable parameters and inference time, but it also decreases the RMSE by $0.15cm$ on the NYU-2 dataset. In short, whether surface normal and semantic priors are employed or not, our method consistently delivers superior performance.

Different Large-Scale Models. Tab. 9 reveals the ablation study on different combinations of large-scale models. The results demonstrate the strong compatibility of our SPFNet with various foundation models, as all four configurations achieve satisfactory performance. Specifically, with the semantic model fixed, using Omnidata as the normal model yields a slight performance improvement over Metric3Dv2. Conversely, when the normal model is fixed, SAM exhibits certain advantages over MobileSAM. Finally, we

Table 10 Ablation study of MGF on NYU-v2 and Lu datasets (×8). D2P: depth-to-prior filtering, P2D: prior-to-depth filtering. D2P→P2D means that the D2P step precedes the P2D step

Methods	D2P	P2D	D2P→D2P	P2D→P2D	D2P→P2D	P2D→D2P	Similarity	Params (M)	NYU-v2	Lu
(a)								29.76 (±0.00)	2.59 (±0.00)	1.75 (±0.00)
(b)	✓							30.22 (+0.46)	2.51 (−0.08)	1.70 (−0.05)
(c)		✓						30.22 (+0.46)	2.48 (−0.11)	1.64 (−0.11)
(d)	✓		✓					30.68 (+0.92)	2.41 (−0.18)	1.60 (−0.15)
(e)		✓		✓				30.68 (+0.92)	2.39 (−0.20)	1.64 (−0.11)
(f)	✓	✓			✓			30.68 (+0.92)	2.46 (−0.13)	1.61 (−0.14)
(g)	✓	✓				✓		30.68 (+0.92)	2.40 (−0.19)	1.59 (−0.16)
(h)	✓	✓				✓	✓	30.68 (+0.92)	**2.36** (−0.23)	**1.56** (−0.19)

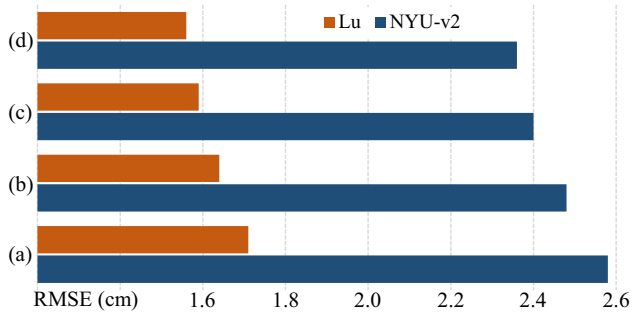


Fig. 13 Ablation study on the effectiveness of similarity computation on the NYU-v2 and Lu datasets (×8)

adopt configuration (d) as the default setup for SPFNet, which surpasses configurations (a)-(c) on the Lu dataset by RMSE margins of 0.10*cm*, 0.13*cm*, and 0.05*cm*, respectively.

APP and OPE. Fig. 12(a) presents the ablation on APP and OPE. The baseline (blue dotted line) removes all APP and OPE modules from SPFNet. It can be discovered that both our APP and OPE contribute to performance improvement. When they are utilized in conjunction, our method (solid orange line) achieves the best performance. For instance, our SPFNet surpasses the baseline by 0.3*cm* on the NYU-v2 dataset and 0.33*cm* on the Lu dataset, respectively.

In addition, Fig. 12(b) shows the ablation on different iterations of APP and OPE modules. It is evident that the performance incrementally improves as the number of APP and OPE increases. When APP & OPE-4 (4 iterations of APP and OPE) is utilized, the RMSE consistently reduces on the NYU and Lu datasets, but only slightly on the Middlebury dataset. To better balance complexity and performance, we select APP & OPE-3 (orange line) as the setting for SPFNet.

Finally, Fig. 12(c) presents an ablation study investigating the ordering of single-modal prior filters within the OPE module on the Lu and Middlebury datasets. These results clearly demonstrate the superiority of the 'N+S+R' configuration, which achieves the lowest RMSE.

MGF. Tab. 10 reports the ablation study of MGF on ×8 DSR. For the baseline (a), we remove all of the guided image filtering blocks in SPFNet. (b) and (c) employ solely depth-to-prior filtering (D2P) or prior-to-depth filtering (P2D), respectively, both of which enhance DSR performance compared to (a). (d)-(g) explore different combination orders of D2P and P2D, and the results show that these combinations achieve lower RMSE than using D2P or P2D alone. Based on (g), (h) further leverages the similarity weights to guide the generation of filter kernels, contributing to the best performance, *i.e.*, surpassing the baseline (a) by 0.23*cm* on the NYU-v2 dataset and 0.19*cm* on the Lu dataset, respectively.

Effectiveness of Similarity Computation. Fig. 13 illustrates the ablation study on multimodal similarity computation. (a) is the baseline, where the entire APP and similarity weights input in OPE are removed. (b) removes similarity computation in APP and similarity weights in OPE. (c) retains full APP but removes similarity weights in OPE. (d) uses full APP and OPE, *i.e.*, our default setting. These results demonstrate that similarity computation yields satisfactory performance gains, surpassing the baseline method (a) by 0.22*cm* in RMSE on the NYU-v2 dataset.

5 Conclusion

In this paper, we propose SPFNet, a novel DSR solution that utilizes the surface normal and semantic priors from large-scale models to effectively weaken the texture interference and improve the edge accuracy. Specifically, we design an all-in-one prior propagation that mitigates interference by calculating the similarity weights between multi-modal scene priors. Moreover, we develop the one-to-one prior embedding that continuously aggregates each single-modal prior into the depth using mutual guidance filtering, further reducing interference and enhancing the edge. Comprehensive experiments demonstrate that our SPFNet performs favorably against state-of-the-art approaches.

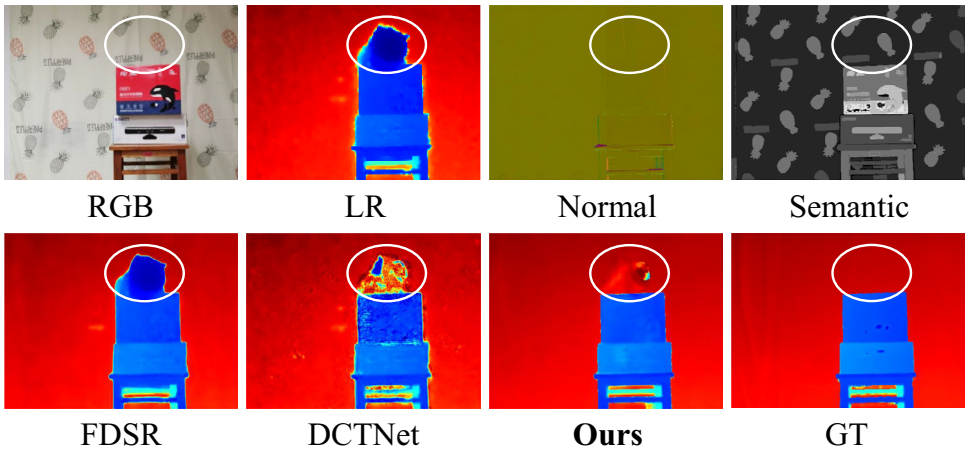


Fig. 14 Failure cases resulting from an inaccurate prior.

6 Discussion

Limitation. Our method utilizes high-quality prior knowledge from large models to reduce texture interference in RGB and enhance edges, thereby significantly improving the model's representation capability. However, when the input data falls outside the distribution of the large model's training samples, the priors may introduce noise. Such errors can potentially interfere with subsequent similarity computation and mutual filtering, degrading performance.

Failure Case. As depicted in Fig. 14, when the input RGB falls outside the distribution of the large model's training samples, the generated priors may not be sufficiently accurate (*e.g.*, blurred normal structures and depth-independent semantic information). In such cases, although our method achieves higher recovery quality than other approaches, it still struggles to precisely restore regions in real scenes that exhibit severe structural distortions and artifacts. For example, the artifacts shown in the white areas in Fig. 14.

The primary cause of this failure lies in the severe degradation inherent in real-world depth maps, such as structural distortion, artifacts, and holes. When the predicted scene prior is also inaccurate, the compounded errors lead to significant deviations in the computed similarity. These deviations cause the subsequent mutual guidance filter to transfer incorrect structural priors into the depth, thereby interfering with the depth reconstruction.

A potential solution is to build a collaborative optimization framework that jointly fine-tunes normal estimation, semantic segmentation, and DSR to improve the accuracy of scene priors. Besides, the large model could serve as an intermediate supervision signal rather than directly incorporating its outputs. This strategy will mitigate the adverse effects of low-quality priors in the inference stage.

Acknowledgements This work was supported by National Natural Science Foundation of China under Grant Nos. U24A20330, 62361166670 and 62406135, and Natural Science Foundation of Jiangsu Province under Grant No. BK20241198.

Data Availability Information on access to the datasets supporting the conclusions of this article is included therein.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

References

Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., & Bruce, L. M. (2007). Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing, 45*(10), 3012–3021.

Cai, J., & Huang, B. (2020). Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing, 59*(6), 5206–5220.

Cheng, X., Wang, P.& Yang, R. (2018) Learning depth with convolutional spatial propagation network. In: Proceedings of the European Conference on Computer Vision, 103–119

De Lutio, R., Becker, A., D'Aronco, S., Russo, S., Wegner, J.D. & Schindler, K. (2022) Learning graph regularisation for guided super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1979–1988

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

Deng, X., Xu, J., Gao, F., Sun, X., & Xu, M. (2023). Deepm 2 cdl: Deep multi-scale multi-modal convolutional dictionary learning network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(5), 2770–2787.

Dong, X., Yokoya, N., Wang, L. & Uezato, T. (2022) Learning mutual modulation for self-supervised cross-modal super-resolution. In: Proceedings of the European Conference on Computer Vision, 1–18. Springer

Eftekhar, A., Sax, A., Malik, J. & Zamir, A. (2021) Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10786–10796

Fan, R., Wang, H., Cai, P. & Liu, M. (2020) Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In: Proceedings of the European Conference on Computer Vision, 340–356. Springer

Gu, S., Zuo, W., Guo, S., Chen, Y., Chen, C.& Zhang, L. (2017) Learning dynamic guidance for depth image enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3769–3778

Guo, C., Li, C., Guo, J., Cong, R., Fu, H., & Han, P. (2018). Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing, 28*(5), 2545–2557.

He, K., Sun, J., & Tang, X. (2012). Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(6), 1397–1409.

He, L., Zhu, H., Li, F., Bai, H., Cong, R., Zhang, C., Lin, C., Liu, M. & Zhao, Y. (2021) 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, 9229–9238

Hirschmuller, H. & Scharstein, D. (2007) Evaluation of cost functions for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8. IEEE

Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., & Shen, S. (2024). Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 10579–10596.

Jiang, K., Jiang, J., Wang, Z., Geng, Z. & Liu, X. (2025) Dawn+: Wavelet-based image deraining meets direction-aware attention and mutual representation. IEEE Transactions on Neural Networks and Learning Systems

Jung, H., Park, E. & Yoo, S. (2021) Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 12642–12652

Kim, B., Ponce, J., & Ham, B. (2021). Deformable kernel networks for joint image filtering. *International Journal of Computer Vision, 129*(2), 579–600.

Kingma, D.P. & Ba, J. (2014) Adam: A method for stochastic optimization. arXiv:1412.6980

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al. (2023) Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015–4026

Levin, A., Lischinski, D., & Weiss, Y. (2004). Colorization using optimization. *ACM Transactions on Graphics, 23*(3), 689–694.

Li, Y., Huang, J.B., Ahuja, N.& Yang, M.H. (2016) Deep joint image filtering. In: Proceedings of the European Conference on Computer Vision, 154–169. Springer

Li, Y., Huang, J. B., Ahuja, N., & Yang, M. H. (2019). Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1909–1923.

Li, Y., Yang, X., Fu, J., Yue, G. & Zhou, W. (2024) Deep bi-directional attention network for image super-resolution quality assessment. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 1–6. IEEE

Li, Y., Yang, X., Yue, G., Fu, J., Jiang, Q., Jia, X., Rosin, P. L., Liu, H., & Zhou, W. (2025). Perception-oriented bidirectional attention network for image super-resolution quality assessment. *IEEE Transactions on Image Processing, 34*, 7728–7743.

Liao, W., Huang, X., Van Coillie, F., Thoonen, G., Pižurica, A., Scheunders, P. & Philips, W. (2015) Two-stage fusion of thermal hyperspectral and visible rgb image by pca and guided filter. In: Proceedings of Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 1–4

Lu, S., Ren, X. & Liu, F. (2014) Depth enhancement via low-rank matrix completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3390–3397

Metzger, N., Daudt, R.C. & Schindler, K. (2023) Guided depth super-resolution by deep anisotropic diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18237–18246

Pan, J., Dong, J., Ren, J.S., Lin, L., Tang, J. & Yang, M.H. (2019) Spatially variant linear representation models for joint filtering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1702–1711

Park, J., Joo, K., Hu, Z., Liu, C.K. & So Kweon, I. (2020) Non-local spatial propagation network for depth completion. In: Proceedings of the European Conference on Computer Vision, 120–136. Springer

Qiao, X., Poggi, M., Deng, P., Wei, H., Ge, C., & Mattoccia, S. (2024). Rgb guided tof imaging system: A survey of deep learning-based methods. *International Journal of Computer Vision, 132*(11), 4954–4991.

Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B. & Pollefeys, M. (2019) Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3313–3322

Scharstein, D. & Pal, C. (2007) Learning conditional random fields for stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8

Shao, S., Pei, Z., Chen, W., Chen, P. C., & Li, Z. (2024). Nddepth: Normal-distance assisted monocular depth estimation and completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 8883–8899.

Shen, X., Zhou, C., Xu, L. & Jia, J. (2015) Mutual-structure for joint filtering. In: Proceedings of the IEEE/CVF international conference on computer vision, 3406–3414

Shi, W., Ye, M. & Du, B. (2022) Symmetric uncertainty-aware feature transmission for depth super-resolution. In: Proceedings of the ACM International Conference on Multimedia, 3867–3876

Shin, J., Shin, S. & Jeon, H.G. (2023) Task-specific scene structure representations. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2272–2281

Silberman, N., Hoiem, D., Kohli, P. & Fergus, R. (2012) Indoor segmentation and support inference from rgbd images. In: Proceedings of the European Conference on Computer Vision, 746–760. Springer

Song, X., Dai, Y., Zhou, D., Liu, L., Li, W., Li, H. & Yang, R. (2020) Channel attention based iterative residual learning for depth map super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5631–5640

Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E. & Kautz, J. (2019) Pixel-adaptive convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11166–11175

Sun, B., Ye, X., Li, B., Li, H., Wang, Z. & Xu, R. (2021) Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7792–7801

Tang, J., Tian, F. P., Feng, W., Li, J., & Tan, P. (2020). Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing, 30*, 1116–1129.

Van Gansbeke, W., Neven, D., De Brabandere, B. & Van Gool, L. (2019) Sparse and noisy lidar completion with rgb guidance and uncertainty. In: Proceedings of the International Conference on Machine Vision Applications, 1–6. IEEE

Viola, M., Qu, K., Metzger, N., Ke, B., Becker, A., Schindler, K. & Obukhov, A. (2025) Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5359–5370

Wald, L., Ranchin, T., & Mangolini, M. (1997). Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing, 63*(6), 691–699.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C. & Novotny, D. (2025) Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference, 5294–5306

Wang, K., Zhao, L., Zhang, J., Zhang, J., Wang, A., & Bai, H. (2023). Joint depth map super-resolution method via deep hybrid-cross guidance filter. *Pattern Recognition, 136*, Article 109260.

Wang, Z., Chen, S., Yang, L., Wang, J., Zhang, Z., Zhao, H. & Zhao, Z. (2025) Depth anything with any prior. arXiv:2505.10565

Wang, Z., Wu, Y., Li, X., Yan, Z. & Yang, J. (2025) Spatiotemporal difference network for video depth super-resolution. arXiv:2508.01259

Wang, Z., Yan, Z., Pan, J., Gao, G., Zhang, K.& Yang, J. (2025) 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, 15813–15822

Wang, Z., Yan, Z., Wu, Y., Gao, G., Li, X. & Yang, J. (2025) Multi-order matching network for alignment-free depth super-resolution. arXiv:2511.16361

Wang, Z., Yan, Z. & Yang, J. (2024) Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, 5823–5831

Wu, Y., Pan, C., Wang, G., Yang, Y., Wei, J., Li, C. & Shen, H.T. (2023) Learning semantic-aware knowledge guidance for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1662–1671

Xiao, Y., Yuan, Q., Jiang, K., Huang, W., Zhang, Q., Zheng, T., Lin, C.W. & Zhang, L. (2025) Spiking meets attention: Efficient remote sensing image super-resolution with attention spiking neural networks. arXiv:2503.04223

Xu, S., Zhang, J., Zhao, Z., Sun, K., Liu, J. & Zhang, C. (2021) Deep gradient projection networks for pan-sharpening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1366–1375

Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H.& Li, H. (2019) Depth completion from sparse lidar data with depth-normal constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2811–2820

Yan, Z., Li, X., Wang, K., Chen, S., Li, J. & Yang, J. (2023) Distortion and uncertainty aware loss for panoramic depth completion. In: Proceedings of the International Conference on Machine Learning, 39099–39109. PMLR

Yan, Z., Lin, Y., Wang, K., Zheng, Y., Wang, Y., Zhang, Z., Li, J. & Yang, J. (2024) Tri-perspective view decomposition for geometry-aware depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4874–4884

Yan, Z., Wang, K., Li, X., Gao, G., Li, J. & Yang, J. (2025) Tri-perspective view decomposition for geometry aware depth completion and super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence

Yan, Z., Wang, K., Li, X., Zhang, Z., Li, G., Li, J., & Yang, J. (2022). Learning complementary correlations for depth super-resolution with incomplete data in real world. *IEEE Transactions on Neural Networks and Learning Systems, 35*(4), 5616–5626.

Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J. & Yang, J. (2022) Rignet: Repetitive image guided network for depth completion. In: Proceedings of the European Conference on Computer Vision, 214–230. Springer

Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J. & Yang, J. (2023) Desnet: Decomposed scale-consistent network for unsupervised depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence, 3109–3117

Yan, Z., Wang, Z., Dong, H., Li, J., Yang, J. & Lee, G.H. (2025) Ducos: Duality constrained depth super-resolution via foundation model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8361–8371

Yang, C., Zhang, L., Lu, H., Ruan, X. & Yang, M.H. (2013) Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3166–3173

Yang, G., Cao, X., Xiao, W., Zhou, M., Liu, A., Chen, X. & Meng, D. (2023) Panflownet: A flow-based deep network for pan-sharpening. In: Proceedings of the IEEE/CVF international conference on computer vision, 16857–16867

Yang, G., Zhao, H., Shi, J., Deng, Z. & Jia, J. (2018) Segstereo: Exploiting semantic information for disparity estimation. In: Proceedings of the European Conference on Computer Vision, 636–651

Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X. & Paisley, J. (2017) Pannet: A deep network architecture for pan-sharpening. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5449–5457

Yang, Y., Cao, Q., Zhang, J., & Tao, D. (2022). Codon: on orchestrating cross-domain attentions for depth super-resolution. *International Journal of Computer Vision, 130*(2), 267–284.

Yu, Z., Sheng, Z., Zhou, Z., Luo, L., Cao, S.Y., Gu, H., Zhang, H. & Shen, H.L. (2023) Aggregating feature point cloud for depth completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8732–8743

Yuan, J., Jiang, H., Li, X., Qian, J., Li, J. & Yang, J. (2023) Recurrent structure attention guidance for depth super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, 3331–3339

Yuan, Q., Wei, Y., Meng, X., Shen, H., & Zhang, L. (2018). A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(3), 978–989.

Yuhas, R.H., Goetz, A.F. & Boardman, J.W. (1992) Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In: JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. **1**: AVIRIS Workshop

Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S. & Hong, C.S. (2023) Faster segment anything: Towards lightweight sam for mobile applications. arXiv:2306.14289

Zhang, R., & Wu, J. (2023). A bidirectional guided filter used for rgb-d maps. *IEEE Transactions on Instrumentation and Measurement, 72*, 1–14.

Zhang, Y., Guo, X., Poggi, M., Zhu, Z., Huang, G. & Mattoccia, S. (2023) Completionformer: Depth completion with convolutions and vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18527–18536

Zhang, Y., Zhou, S. & Li, H. (2024) Depth information assisted collaborative mutual promotion network for single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2846–2855

Zhao, Z., Zhang, J., Gu, X., Tan, C., Xu, S., Zhang, Y., Timofte, R. & Van Gool, L. (2023) Spherical space feature decomposition for guided depth map super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 12547–12558

Zhao, Z., Zhang, J., Xu, S., Lin, Z. & Pfister, H. (2022) Discrete cosine transform network for guided depth map super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5697–5707

Zhong, Z., Liu, X., Jiang, J., Zhao, D., Chen, Z., & Ji, X. (2021). High-resolution depth maps imaging via attention-based hierarchical

multi-modal fusion. *IEEE Transactions on Image Processing, 31*, 648–663.

Zhong, Z., Liu, X., Jiang, J., Zhao, D., & Ji, X. (2023). Deep attentional guided image filtering. *IEEE Transactions on Neural Networks and Learning Systems, 35*(4), 12236–12250.

Zhou, M., Yan, K., Huang, J., Yang, Z., Fu, X. & Zhao, F. (2022) Mutual information-driven pan-sharpening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1798–1808

Zhou, M., Yan, K., Pan, J., Ren, W., Xie, Q., & Cao, X. (2023). Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision, 131*(1), 215–242.

Zhou, W., Jiang, Q., Wang, Y., Chen, Z., & Li, W. (2020). Blind quality assessment for image superresolution using deep two-stream convolutional networks. *Information Sciences,528*, 205–218.

Zhou, W. & Wang, Z. (2022) Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity. In: Proceedings of the ACM international conference on multimedia, 934–942