DORNet: A Degradation Oriented and Regularized Network for Blind Depth Super-Resolution

Zhengxue Wang[1]*, Zhiqiang Yan[1*†], Jinshan Pan[1], Guangwei Gao[2], Kai Zhang[3], and Jian Yang[1†]
[1]PCA Lab‡, Nanjing University of Science and Technology
[2]Nanjing University of Posts and Telecommunications [3]Nanjing University
{zxwang, yanzq, jspan, csjyang}@njust.edu.cn, csggao@gmail.com, kaizhang@nju.edu.cn

Abstract

Recent RGB-guided depth super-resolution methods have achieved impressive performance under the assumption of fixed and known degradation (e.g., bicubic downsampling). However, in real-world scenarios, captured depth data often suffer from unconventional and unknown degradation due to sensor limitations and complex imaging environments (e.g., low reflective surfaces, varying illumination). Consequently, the performance of these methods significantly declines when real-world degradation deviate from their assumptions. In this paper, we propose the Degradation Oriented and Regularized Network (DORNet), a novel framework designed to adaptively address unknown degradation in real-world scenes through implicit degradation representations. Our approach begins with the development of a self-supervised degradation learning strategy, which models the degradation representations of low-resolution depth data using routing selection-based degradation regularization. To facilitate effective RGB-D fusion, we further introduce a degradation-oriented feature transformation module that selectively propagates RGB content into the depth data based on the learned degradation priors. Extensive experimental results on both real and synthetic datasets demonstrate the superiority of our DORNet in handling unknown degradation, outperforming existing methods.

1. Introduction

Blind depth super-resolution (DSR) aims to recover precise high-resolution (HR) depth from low-resolution (LR) depth with unknown degradation, which has been widely applied in many fields, such as virtual reality [17, 33, 35, 44], augmented reality [6, 31, 41, 43, 48], and 3D reconstruction [3, 4, 34, 46, 49]. As shown in Fig. 1(a), recent RGB-guided DSR methods [2, 26, 45, 52, 58] have been proposed that integrate RGB features aligned with input depth based on the assumption of known and fixed degradation, achieving excellent performance on synthetic data.

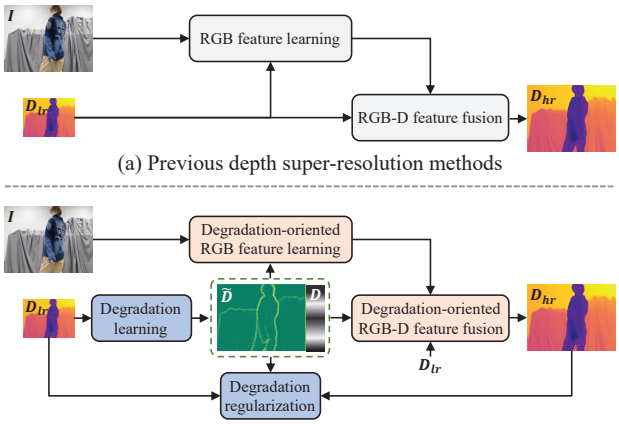


Figure 1. Previous methods (a) directly fuse the RGB information aligned with the LR depth, while our method (b) focuses more on modeling the degradation representation of the LR depth to provide targeted guidance for HR depth recovery.

However, due to limitations in sensor technology and imaging environments, depth data captured from real-world scenes often suffer from unconventional and unknown degradation [47] (*e.g.*, structural distortion and blur). Such real-world degradation results in structure inconsistency between depth and RGB, impairing the performance of previous methods. Moreover, real-world degradation labels are unavailable, preventing us from explicitly estimating the degradation between LR depth and HR depth.

As illustrated in Fig. 2(b) and Fig. 2(c), the LR depth synthesized using bicubic downsampling exhibits accurate depth structures, while the real-world LR depth experiences more severe structural distortion. Furthermore, Fig. 2(i)

*Equal contribution
†Corresponding authors
‡PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology.

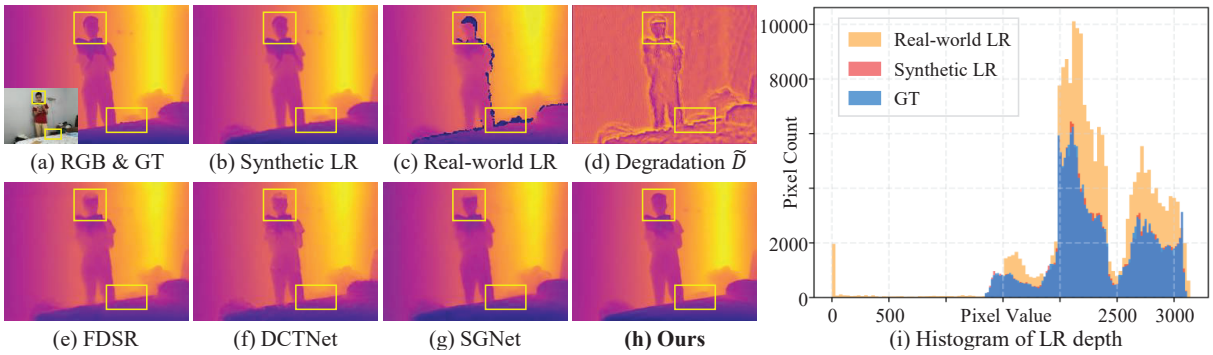


Figure 2. Visual results of LR depth, HR depth, and degradation representation. (b) and (c) are the synthetic and the real-world LR depth, respectively. (d) is the learned degradation representation $\tilde{\boldsymbol{D}}$. (e)-(g) are the HR depth predicted by FDSR [9], DCTNet [56], and SGNet [37], while (h) is produced by our DORNet. (i) is the histogram of real-world LR, synthetic LR, and ground-truth (GT) depth.

indicates that the distribution of the real-world LR depth shows a greater difference from the ground-truth depth compared to the synthetic LR depth. This makes it more challenging for DSR to restore accurate HR depth from LR depth with unknown degradation.

To address these issues, as shown in Fig. 1(b), we propose a degradation oriented and regularized network (DORNet). The DORNet utilizes degradation representations to guide the restoration of HR depth from real-world scenarios with unknown degradation. To this end, we present a self-supervised degradation learning strategy to estimate the implicit degradation representations between LR and HR depth. In this strategy, a router mechanism is first introduced to dynamically control the generation of degradation kernels with varying scales. We then design degradation regularization that leverages these kernels to deteriorate the predicted HR depth, yielding a new degraded depth. Consequently, the entire degradation process is learned by narrowing the distance between the new degraded depth and the LR depth, without using degradation labels. Furthermore, we observe that RGB can provide sharp and complete details for the degradation areas of the LR depth. Therefore, we propose utilizing the estimated degradation to adaptively select RGB features to guide and facilitate the RGB-D fusion. Concretely, we develop a degradation-oriented fusion scheme, deploying a degradation-oriented feature transformation module (DOFT). The DOFT produces filter kernels from learned degradation and then filters the RGB features, offering complementary contents for the depth features.

Owing to these designs, Fig. 2(d) demonstrates that the real-world degradation learned by DORNet accurately characterizes the degradation areas of the LR depth, thereby providing precise guidance for RGB-D fusion. Moreover, compared to previous approaches [9, 37, 56], Fig. 2(h) reveals that our method can effectively restore HR depth with more accurate and clearer structures.

In short, our contributions are as follows:

- We introduce a novel DSR framework termed DORNet, which utilizes degradation representations to adaptively address unknown degradation in real-world scenes.
- We design a self-supervised degradation learning strategy to model degradation representations of LR depth using routing selection-based degradation regularization.
- We propose a degradation-oriented fusion scheme that selectively transfers RGB content into depth by performing DOFT based on learned degradation priors.
- Extensive experiments demonstrate that our DORNet achieves state-of-the-art performance.

2. Related Work

2.1. Depth Map Super-Resolution

Synthetic Depth Super-Resolution. Many DSR methods [8, 23, 32, 38] have made significant progress on synthetic data with known degradation. For example, Hui et al. [11] develop a multi-scale guidance network to enhance the boundary clarity of depth. In [50], Ye et al. utilize the progressive multi-branch fusion network to restore HR depth with sharp boundaries. Recently, a few guided image filtering methods [12, 19, 59] have been proposed for transferring guidance information to the target. For instance, Li et al. [18] design a learning-based joint filtering method that propagates salient structures from guidance into target. Kim et al. [12] apply the deformable kernel network to learn sparse and spatially-variant filter kernels. Additionally, to extract common features from different modality inputs, Deng et al. [5] present a common and unique information splitting network based on multi-modal convolutional sparse coding. Similarly, Zhao et al. build the discrete cosine transform network [56] and the spherical spatial feature decomposition network [57] to separate the private and shared features between RGB and depth. Unlike these approaches, we focus on utilizing the degradation representations of LR depth to adaptively address unconventional and unknown degradation in real-world scenarios.

Real-world Depth Super-Resolution. Recently, real-

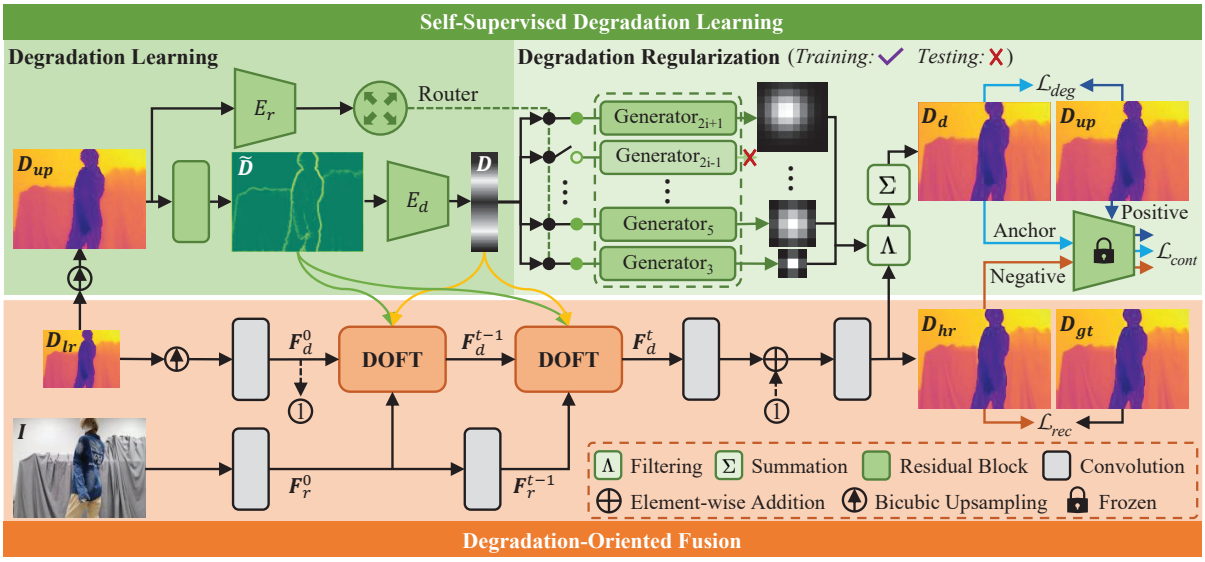


Figure 3. Overview of DORNet. Given $\boldsymbol{D}_{up}$ as input, the degradation learning first encodes it to produce degradation representations $\tilde{\boldsymbol{D}}$ and $\boldsymbol{D}$. Then, $\tilde{\boldsymbol{D}}$, $\boldsymbol{D}$, $\boldsymbol{D}_{lr}$, and $\boldsymbol{I}$ are fed into multiple degradation-oriented feature transformation (DOFT) modules, generating the HR depth $\boldsymbol{D}_{hr}$. Finally, $\boldsymbol{D}$ and $\boldsymbol{D}_{hr}$ are sent to the degradation regularization to obtain $\boldsymbol{D}_d$, which is used as input for the degradation loss $\mathcal{L}_{deg}$ and the contrastive loss $\mathcal{L}_{cont}$. The degradation regularization only applies during training and adds no extra overhead in testing.

world DSR [7, 9, 22, 29] targeting unknown degradation has attracted broad attention. For instance, Liu et al. [21] propose a robust optimization framework to address the issues of inconsistency in RGB edges and discontinuity in depth. Song et al. [29] employ both non-linear degradation with noise and interval down-sampling degradation to simulate LR depth for real-world DSR. Besides, He et al. [9] construct a real-world RGB-D dataset, and design a fast DSR network based on octave convolution. More recently, Yan et al. [42] introduce an auxiliary depth completion branch to recover dense HR depth from incomplete LR depth. Yuan et al. [53] develop a structure flow-guided model for real-world DSR, which learns a cross-modal flow map to guide the transfer of RGB structural information. Different from previous researches, we pay more attention to modeling the implicit degradation representations of LR depth, and selectively propagating RGB information into depth data based on the estimated degradation priors.

2.2. Degradation Representation Learning

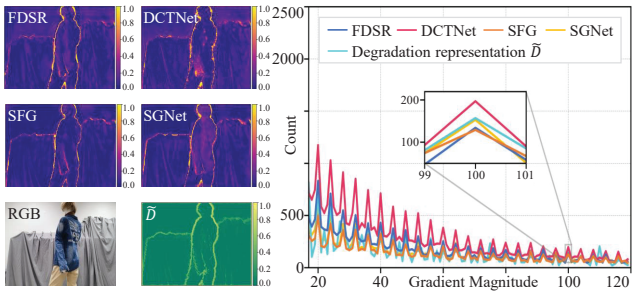
Degradation representations have been widely applied in several single-modal image restoration tasks [20, 36, 54]. For example, Wang et al. [36] learn degradation representations for blind image super-resolution by assuming that the degradation of different patches within each image is the same. Similarly, Xia et al. [40] develop a degradation estimator based on knowledge distillation to model the degradation representations. Li et al. [16] introduce a multi-scale degradation injection network to jointly optimize reblurring and deblurring. Additionally, some approaches [15, 51, 54] explore solutions that can be applied to various degradation in a single model. For instance, Li et al. [15] design an all-in-one image restoration framework, which can recover images with different degradation in one network. Inspired by them, we develop a self-supervised degradation learning strategy to estimate the degradation representations of LR depth using routing selection-based degradation regularization. The learned degradation priors are employed to guide the feature transformation between multi-modal inputs.

3. Method

3.1. Network Architecture

Given LR depth $\boldsymbol{D}_{lr} \in R^{h\times w\times 1}$ with unknown degradation and RGB $\boldsymbol{I} \in R^{sh\times sw\times 3}$ as inputs, our method aims to recover accurate HR depth $\boldsymbol{D}_{hr} \in R^{sh\times sw\times 1}$ by learning the degradation representations. h, w, and s represent the height, width, and upsampling factor, respectively.

As shown in Fig. 3, our DORNet mainly consists of a self-supervised degradation learning strategy (green part) and a degradation-oriented fusion scheme (orange part). Specifically, the upsampled LR depth $\boldsymbol{D}_{up} \in R^{sh\times sw\times 1}$ is first input into the degradation learning, producing both the router and the degradation representations, $\tilde{\boldsymbol{D}}$ and $\boldsymbol{D}$. Then, $\tilde{\boldsymbol{D}}$, $\boldsymbol{D}$, $\boldsymbol{D}_{lr}$, and $\boldsymbol{I}$ are sent to multiple degradation-oriented feature transformation modules (DOFT), which selectively propagate RGB information into the depth features, resulting in HR depth $\boldsymbol{D}_{hr}$. Next, the degradation regularization takes $\boldsymbol{D}$ as input and utilizes routing selection to adaptively generate degradation kernels with varying scales, all



(a) Error maps & $\tilde{D}$ (b) Gradient histogram

Figure 4. Visualization of error maps and degradation representation $\tilde{\boldsymbol{D}}$ (a), and their gradient histograms (b).

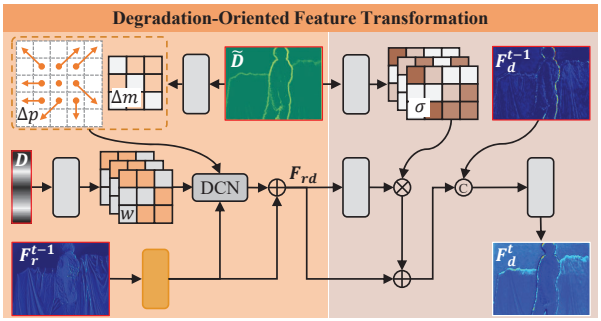


Figure 5. Details of DOFT. $\otimes$ is element-wise multiplication while © is concatenation. Orange rectangular box: residual group [55].

of which are sent into the filtering and summation modules together with $\boldsymbol{D}_{hr}$, obtaining the new degraded depth $\boldsymbol{D}_d$. Finally, $\boldsymbol{D}_d$ is employed as input for the degradation loss $\mathcal{L}_{deg}$ and the contrastive loss [39] $\mathcal{L}_{cont}$, further promoting the learning of degradation representations.

Furthermore, to balance computational complexity and performance, we present a more lightweight DSR model, DORNet-T, which is achieved by reducing all convolutional channels to $\frac{3}{8}$ of those in DORNet, while maintaining the entire network architecture unchanged.

3.2. Self-Supervised Degradation Learning

Degradation Learning. As illustrated in Fig. 3 (upper left), given $\boldsymbol{D}_{lr}$ as input, bicubic upsampling is first utilized to generate the upsampled depth $\boldsymbol{D}_{up}$. Then, we employ the residual block f_{rb} and the degradation encoder E_d to encode $\boldsymbol{D}_{up}$ into degradation representations $\tilde{\boldsymbol{D}}$ and $\boldsymbol{D}$, where $\tilde{\boldsymbol{D}} = f_{rb}(\boldsymbol{D}_{up})$ and $\boldsymbol{D} = E_d(\tilde{\boldsymbol{D}})$.

Next, inspired by the Mixture-of-Experts [1, 10, 24], we construct a router to dynamically allocate the degradation representation $\boldsymbol{D}$ to degradation regularization, thereby adaptively selecting degradation kernel generators of different scales. The learned router $\boldsymbol{\mathcal{R}}$ is formulated as:

$$\boldsymbol{\mathcal{R}} = \sigma(topK(E_r(\boldsymbol{D}_{up}))), \tag{1}$$

where σ and E_r are the softmax function and the routing encoder, respectively. $topK$ indicates the adaptive allocation of $\boldsymbol{D}$ to the top k degradation kernel generators from g candidate generators based on their scores.

Degradation Regularization. As depicted in Fig. 3 (upper right), given $\boldsymbol{D}$ as input, we first select k degradation kernel generators of different scales under the assignment of router $\boldsymbol{\mathcal{R}}$, adaptively producing a multi-scale degradation kernel set $\mathbb{S}$. As an example, the degradation kernel $\boldsymbol{s}_{2i+1}$ of size $(2i+1)\times(2i+1)$ in $\mathbb{S}$ is represented as:

$$\boldsymbol{s}_{2i+1} = f_g^{2i+1}(\boldsymbol{\mathcal{R}}, \boldsymbol{D}), i \geq 1, \tag{2}$$

where f_g^{2i+1} refers to the degradation kernel generator with a size of $(2i+1)\times(2i+1)$, consisting of MLP.

Then, the filtering and summation modules take the degradation kernel set $\mathbb{S}$ and the predicted HR depth $\boldsymbol{D}_{hr}$ as inputs to synthesize the degraded depth $\boldsymbol{D}_d$, which is used to supervise the learning of $\tilde{\boldsymbol{D}}$ and $\boldsymbol{D}$. Specifically, each degradation kernel in $\mathbb{S}$ is employed as a convolution kernel to individually convolve with $\boldsymbol{D}_{hr}$. The resulting convolution outputs are summed to generate $\boldsymbol{D}_d$:

$$\boldsymbol{D}_d = \textstyle\sum_{j=1}^{k} \Lambda(\mathbb{S}_j, \boldsymbol{D}_{hr}), \tag{3}$$

where Λ represents the convolution operation.

Next, we introduce a pre-trained VGG19 [28] to map $\boldsymbol{D}_{hr}$, $\boldsymbol{D}_d$, and $\boldsymbol{D}_{up}$ to the latent space, yielding negative sample $\boldsymbol{F}_n$, anchor sample $\boldsymbol{F}_a$, and positive sample $\boldsymbol{F}_p$, respectively. These samples are used as inputs for the contrastive loss $\mathcal{L}_{cont}$, pulling the degraded depth $\boldsymbol{D}_d$ closer to the LR depth $\boldsymbol{D}_{up}$ and pushing it away from the HR depth $\boldsymbol{D}_{hr}$, thereby facilitating the learning of degradation representations:

$$\mathcal{L}_{cont} = \textstyle\sum_{z=1}^{m} \alpha_z \cdot \frac{f_{l1}(\boldsymbol{F}_p^z - \boldsymbol{F}_a^z)}{f_{l1}(\boldsymbol{F}_n^z - \boldsymbol{F}_a^z)}, \tag{4}$$

where m denotes the number of latent space features, and α is a weight vector. f_{l1} refers to the L_1 distance.

Additionally, a degradation loss $\mathcal{L}_{deg}$ is employed to further optimize the degradation learning:

$$\mathcal{L}_{deg} = \frac{1}{Q} \textstyle\sum_{q=1}^{Q} \|\boldsymbol{D}_{up}^q - \boldsymbol{D}_d^q\|_1, \tag{5}$$

where Q refers to the number of training samples. $\|\cdot\|_1$ represents the L_1 loss function.

Fig. 4 presents a visual comparison of the learned degradation representation $\tilde{\boldsymbol{D}}$ with the error maps of previous methods, as well as a comparison of their gradient histograms. The visualizations and gradient distributions demonstrate that $\tilde{\boldsymbol{D}}$ successfully learns the degraded depth structures that is challenging for previous approaches to recover, thereby providing targeted guidance for enhancing these severely degraded depth features.

RMSE	DJF [18]	DJFR [19]	CUNet [5]	DKN [12]	FDKN [12]	FDSR [9]	DCTNet [56]	SUFT [25]	SSDNet [57]	SFG [53]	SGNet [37]	**DORNet-T**	**DORNet**
Params. (M)	0.08	0.08	0.21	1.16	0.69	0.60	0.48	22.01	-	63.53	8.97	0.46	3.05
RGB-D-D	5.54	5.52	5.84	5.08	5.37	5.49	5.43	5.41	5.38	3.88	5.32	3.84	**3.42**
TOFDSR	5.84	5.72	6.04	5.50	5.77	5.03	5.16	4.37	-	4.52	4.33	4.87	**4.21**

Table 1. Quantitative comparison with existing state-of-the-art methods on the real-world RGB-D-D and TOFDSR datasets.

RMSE	DJF [18]	DJFR [19]	CUNet [5]	DKN [12]	FDKN [12]	FDSR [9]	DCTNet [56]	SUFT [25]	SFG [53]	SGNet [37]	**DORNet-T**	**DORNet**
RGB-D-D	5.83	5.78	5.96	5.52	5.69	5.66	5.61	5.53	4.08	5.44	4.24	**3.68**
TOFDSR	8.21	7.03	8.64	5.96	6.86	5.58	5.46	5.08	5.46	5.11	5.07	**4.47**

Table 2. Quantitative comparison of joint DSR and denoising on the real-world RGB-D-D and TOFDSR datasets.

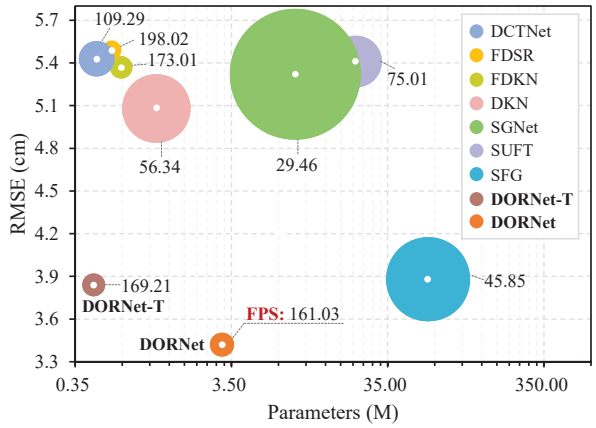


Figure 6. Complexity on RGB-D-D (w/o Noisy) tested by a 4090 GPU. A larger circle diameter indicates a higher inference time.

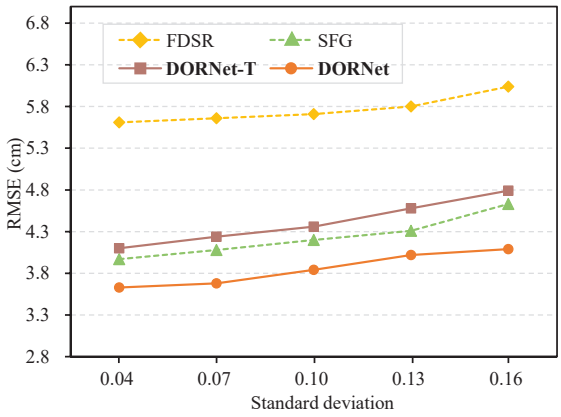


Figure 7. Robustness with different noises on RGB-D-D.

More importantly, degradation regularization is only applied in the training to facilitate the learning of degradation representations, and it does not introduce any additional computational overhead during testing.

3.3. Degradation-Oriented Fusion

As shown in the orange part of Fig. 3, $\boldsymbol{D}_{lr}$ is first input into bicubic upsampling. Then, the upsampled LR depth and $\boldsymbol{I}$ are mapped to $\boldsymbol{F}_d^0$ and $\boldsymbol{F}_r^0$, respectively.

Next, we take $\boldsymbol{F}_d^0$, $\boldsymbol{F}_r^0$, $\tilde{\boldsymbol{D}}$, and $\boldsymbol{D}$ as inputs and recursively conduct multiple DOFT to selectively propagate RGB content into the depth features, generating the enhanced depth feature $\boldsymbol{F}_d^t$:

$$\boldsymbol{F}_d^t = f_{do}^t(\tilde{\boldsymbol{D}}, \boldsymbol{D}, \boldsymbol{F}_d^{t-1}, \boldsymbol{F}_r^{t-1}), \tag{6}$$

where f_{do}^t refers to t-th DOFT.

Finally, the HR depth $\boldsymbol{D}_{hr}$ is predicted by fusing depth features $\boldsymbol{F}_d^0$ and $\boldsymbol{F}_d^t$:

$$\boldsymbol{D}_{hr} = f_c(\boldsymbol{F}_d^0 + f_c(\boldsymbol{F}_d^t)), \tag{7}$$

where f_c refers to the convolutional layer, indicated by the gray rectangular box in Fig. 3 and 5.

Degradation-Oriented Feature Transformation. Fig. 5 shows that DOFT includes degradation-oriented RGB feature learning (left part) and RGB-D feature fusion (right part). Specifically, DOFT first maps $\tilde{\boldsymbol{D}}$ to the offset $\triangle p$ and modulation scalar $\triangle m$, both of which are utilized to dynamically adjust the receptive field of the deformable convolution (DCN) [60] f_d. Then, we generate the weights w of the DCN using $\boldsymbol{D}$ to focus its attention on RGB features that match the degraded depth structures.

Next, given RGB feature $\boldsymbol{F}_r^{t-1}$ as input, $\triangle p$, $\triangle m$, and w are together used to adaptively learn the RGB feature $\boldsymbol{F}_{rd}$ aligned with the degradation representations:

$$\boldsymbol{F}_{rd} = f_d(f_{rg}(\boldsymbol{F}_r^{t-1}), \triangle p, \triangle m, w) + f_{rg}(\boldsymbol{F}_r^{t-1}), \tag{8}$$

where f_{rg} is the residual group [55], a feature extraction unit consisting of residual block and channel attention.

Finally, we encode $\tilde{\boldsymbol{D}}$ as an affinity coefficient σ for the selective transfer of learned RGB feature $\boldsymbol{F}_{rd}$ to the depth, resulting in the enhanced depth feature $\boldsymbol{F}_d^t$:

$$\boldsymbol{F}_d^t = f_c([\boldsymbol{F}_d^{t-1}, \sigma \otimes f_c(\boldsymbol{F}_{rd}) + \boldsymbol{F}_{rd}]), \tag{9}$$

where $\boldsymbol{F}_d^{t-1}$ is the input depth feature of DOFT. $[\cdot]$ denotes concatenation. $\otimes$ refers to element-wise multiplication.

3.4. Loss Function

Given the predicted HR depth $\boldsymbol{D}_{hr}$ and the ground-truth depth $\boldsymbol{D}_{gt}$, we first introduce the reconstruction loss $\mathcal{L}_{rec}$ to optimize our DORNet:

$$\mathcal{L}_{rec} = \frac{1}{Q}\sum_{q=1}^{Q} \|\boldsymbol{D}_{gt}^q - \boldsymbol{D}_{hr}^q\|_1. \tag{10}$$

Figure 8. Visual results (left) and error maps (right) on the real-world RGB-D-D dataset (w/o Noise).

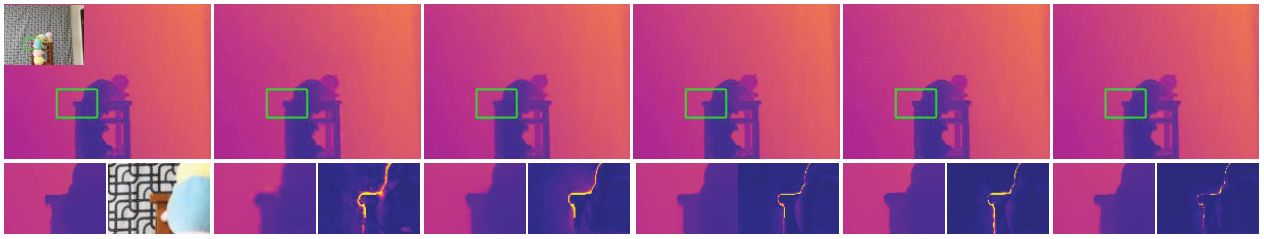


Figure 9. Visual results (left) and error maps (right) on the real-world TOFDSR dataset (w/o Noise).

Then, combining Eqs. (4) and (5), the total training loss $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{deg} + \lambda_2 \mathcal{L}_{cont}, \tag{11}$$

where λ_1 and λ_2 are hyper-parameters.

4. Experiments

4.1. Experimental Setups

Datasets. We conduct extensive experiments on both real-world RGB-D-D [9], TOFDSR [46], and synthetic NYU-v2 [27] datasets. Specifically, for the RGB-D-D, the training set comprises $2,215$ RGB-D pairs, while the test set contains 405 pairs. Additionally, the colorization method [14] is used to fill in the raw LR depth of the TOFDC [46], obtaining the TOFDSR that includes $10K$ RGB-D pairs for training and 560 pairs for testing. In the real-world scenarios, the LR depth is obtained using the ToF camera of the Huawei P30 Pro. Following [12, 37, 56], the synthetic NYU-v2 consists of $1,000$ RGB-D pairs for training and 449 pairs for testing, with the LR depth generated by bicubic downsampling from the GT depth.

To weaken the interference of erroneous depth in the TOFDSR dataset, all methods calculate the loss and RMSE only for valid pixels where the GT depth is within the range of $0.1m$ to $5m$. For the RGB-D-D and NYU-v2 datasets, we maintain the same settings as in previous methods [9, 9, 56].

Implementation Details. We employ the root mean square error (RMSE) in centimeter as the evaluation metric to be consistent with previous DSR methods [9, 31, 53, 57]. The Adam [13] optimizer with an initial learning rate of 1×10^{-4} is used to train our DORNet. Besides, we implement our model in PyTorch using the NVIDIA GeForce RTX 4090. The hyper-parameters are set as $\lambda_1 = \lambda_2 = 0.1$.

4.2. Comparison with the State-of-the-Art

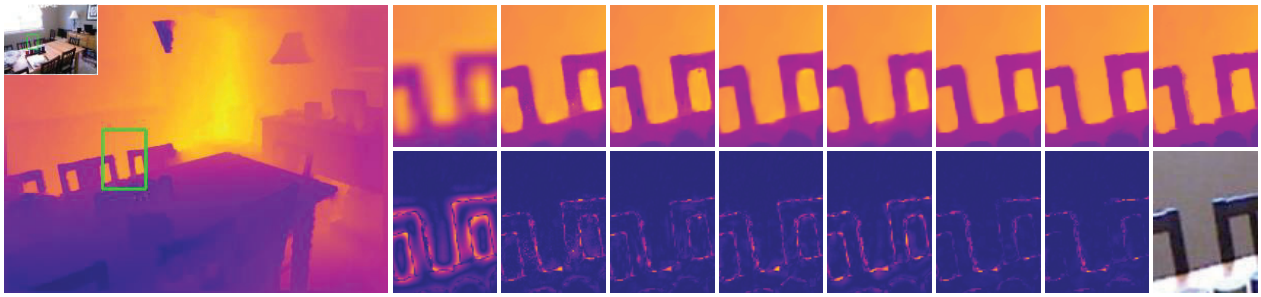
We compare DORNet with popular methods, *i.e.*, DJF [18], DJFR [19], PAC [30], CUNet [5], DKN [12], FDKN [12], FDSR [9], GraphSR [4], DCTNet [56], SUFT [25], DADA [23], SSDNet [57], SFG [53], and SGNet [37]. To ensure a fair comparison, we directly cite the data from their papers for methods with existing experimental results. For other approaches, we utilize their released code to retrain and test under the same settings.

Comparison on Real-World Dataset. Tab. 1 indicates that our DORNet outperforms other advanced methods on the real-world RGB-D-D and TOFDSR datasets. From the first two rows of Tab. 1, it can be seen that DORNet surpasses SFG [53] by $0.46cm$ on RGB-D-D while also significantly reducing the number of parameters. Moreover, the third row demonstrates that our method decreases the RMSE by $0.12cm$ on TOFDSR compared to SGNet [37].

Furthermore, Figs. 8 and 9 present the visual results on the RGB-D-D and TOFDSR. In the error maps, a brighter color means a larger error. Obviously, for severely degraded LR depth, our method succeeds in recovering accurate depth structures. For instance, the handbag in Fig. 8 predicted by our method is more precise than others. Additionally, the error maps in Fig. 9 show that DORNet reconstructs HR depth with fewer errors.

Fig. 6 illustrates that our method achieves a satisfactory balance among parameters, inference time, FPS, and performance. For example, compared to lightweight DCTNet ($0.48M$), our DORNet-T ($0.46M$) reduces RMSE by 29% and inference time by 35%. Moreover, DORNet surpasses the second-best approach by 11% while significantly decreasing both parameters and inference time.

Robustness to Noise. Tab. 2 demonstrates that our method



GT & RGB Bicubic FDKN DKN FDSR DCTNet **DORNet-T** **DORNet** GT & RGB

Figure 10. Visual results (top) and error maps (bottom) on the synthetic NYU-v2 dataset ($\times 8$).

RMSE	PAC [30]	CUNet [5]	DKN [12]	FDSR [9]	GraphSR [4]	DCTNet [56]	SUFT [25]	DADA [23]	SSDNet [57]	SFG [53]	SGNet [37]	**DORNet-T**	**DORNet**
Params. (M)	-	0.21	1.16	0.60	32.53	0.48	22.01	32.53	-	63.53	35.42	0.46	3.05
$\times 4$	1.89	1.92	1.62	1.61	1.79	1.59	1.12	1.54	1.60	1.45	**1.10**	1.33	1.19
$\times 8$	3.33	3.70	3.26	3.18	3.17	3.16	2.51	2.74	3.14	2.84	**2.44**	2.90	2.70
$\times 16$	6.78	6.78	6.51	5.86	6.02	5.84	4.86	4.80	5.86	5.56	**4.77**	5.95	5.60

Table 3. Quantitative comparison with existing state-of-the-art methods on the synthetic NYU-v2 dataset.

exhibits robustness in noisy environments. Similar to previous approaches [12, 53], we add Gaussian noise (mean 0 and standard deviation 0.07) and Gaussian blur (standard deviation 3.6) to upsampled LR depth as new input. We can see that DORNet outperforms SFG [53] by $0.40cm$ in RMSE on the RGB-D-D. For experiments on adding noise before LR depth pre-upsampling, please see our appendix.

Fig. 7 shows the comparison across different noise levels, with the standard deviation of Gaussian noise ranging from 0.04 to 0.16, while the Gaussian blur remains unchanged. We can observe that as the noise levels increase, the performance of all methods gradually declines. However, our DORNet consistently outperforms other approaches at each noise level. For instance, our method reduces RMSE by $0.36cm$ (standard deviation 0.10) and by $0.29cm$ (standard deviation 0.13) compared to SFG [53].

Comparison on Synthetic Dataset. Tab. 3 shows that our method achieves comparable performance on the NYU-v2 dataset. The first row lists the model parameters with a scale factor of 4. For example, compared to the SGNet [37], our DORNet significantly reduces the parameters by 91%, while only increasing the RMSE by 8% ($\times 4$). Furthermore, for lightweight DSR, our DORNet-T outperforms DCTNet by 16% and FDSR by 17% in RMSE ($\times 4$). Fig. 10 reveals that the depth structures predicted by our method is more closely aligned with the ground-truth depth. For instance, the edges of chair exhibit less error than others.

In summary, all of these quantitative comparisons and visual results demonstrate that our method effectively enhances the performance of real-world DSR.

4.3. Generalization Ability

To further evaluate the generalization ability of our method, we implement it on **pan-sharpening** and **depth completion** tasks. Please see our appendix for the details.

Methods	Params (M)	w/o Noisy	w/ Noisy
baseline + DASR [36]	3.44	3.86	4.06
baseline + KDSR [40]	3.73	3.65	3.86
baseline + DL & DR (**Ours**)	**3.05**	**3.42**	**3.69**

Table 4. Comparison of different degradation learning methods on the real-world RGB-D-D dataset. DL indicates Degradation Learning, while DR refers to Degradation Regularization.

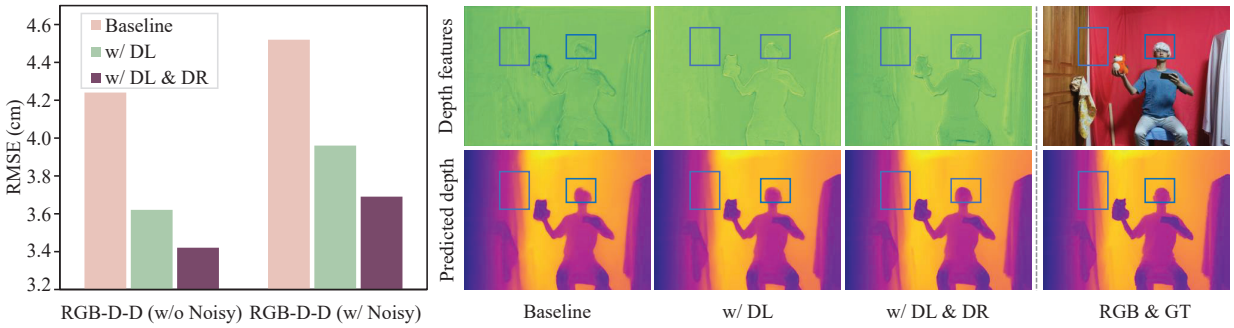
4.4. Ablation Studies

Degradation Learning and Regularization. Fig. 11 and Tab. 4 present the ablation study of degradation learning (DL) and degradation regularization (DR). For the baseline, we first remove the entire DL and DR in DORNet. Then, we utilize concatenation to replace all DOFT. Additionally, only the reconstruction loss is used during the training.

Fig. 11(a) reveals that DL significantly reduces RMSE by modeling the degradation representations. When DR is combined, our method achieves the best performance. For example, DORNet outperforms the baseline by $0.82cm$ (w/o Noisy) and $0.83cm$ (w/ Noisy). Fig. 11(b) presents the visual results of the depth features and predicted depth. Compared to the baseline, DL contributes to generating clearer structures. When DR is employed together with DL, our approach produces more accurate depth.

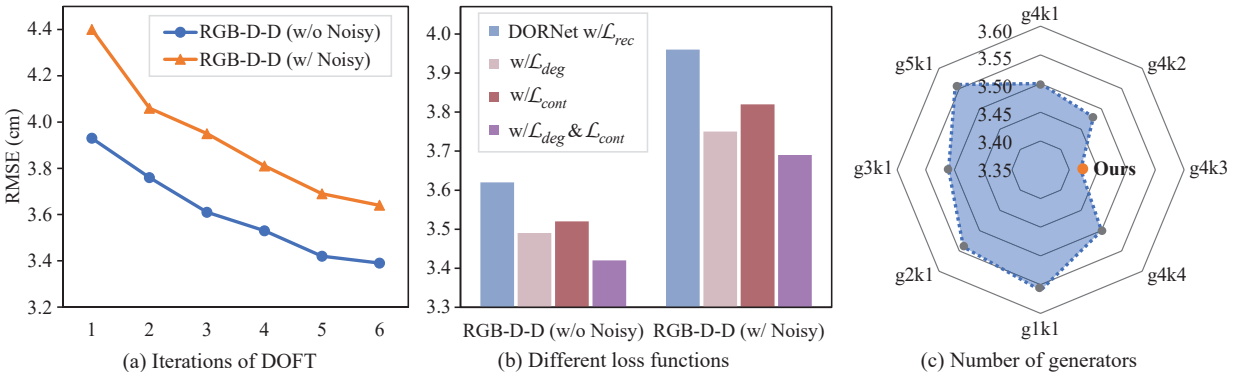
Furthermore, Tab. 4 lists the comparison results of DL and DR with previous degradation learning methods. Specifically, we replace the entire DL and DR with the degradation learning modules from DASR [36] and KDSR [40], respectively. It can be observed that our approach surpasses DASR by $0.44cm$ and KDSR by $0.23cm$ in RMSE (w/ Noisy). These results further demonstrate that our DL and DR can learn more accurate degradation representations and effectively enhance DSR performance.

Different Recursion Numbers of DOFT. Fig. 12(a) de-



(a) Quantitative comparison (b) Visual comparison

Figure 11. Ablation study of degradation learning (DL) and degradation regularization (DR) on the RGB-D-D dataset.



(a) Iterations of DOFT (b) Different loss functions (c) Number of generators

Figure 12. Ablation study of DORNet with (a) different numbers of DOFT, (b) different loss functions, and (c) different numbers of degradation kernel generators. 'g4k3': DR selects 3 (k) out of 4 generators (g) of size $(2i+1) \times (2i+1)$, $1 \leq i \leq 4$, based on the router.

picts the ablation study of different iterations of DOFT. The baseline is the entire DORNet with all loss functions. It is evident that performance incrementally improves as the number of DOFT iterations increases. When the number of iterations reaches 6, the reduction in RMSE begins to slow down. To better trade-off between the model complexity and performance, our DORNet iterates 5 DOFT.

Different Loss Functions. Fig. 12(b) presents the ablation study of different loss functions. The baseline is the entire DORNet using only the reconstruction loss $\mathcal{L}_{rec}$. Obviously, we can see that both the degradation loss $\mathcal{L}_{deg}$ and contrastive loss $\mathcal{L}_{cont}$ contribute to performance improvement. When $\mathcal{L}_{deg}$ and $\mathcal{L}_{cont}$ are deployed together, our method achieves the lowest RMSE. For example, compared to the baseline, our DORNet decreases the RMSE by $0.20cm$ (w/o Noisy) and $0.27cm$ (w/ Noisy) on RGB-D-D.

Number of Generators. Fig. 12(c) shows the ablation study of different numbers of degradation kernel generators on the RGB-D-D dataset (w/o Noisy). The baseline is the entire DORNet with $\mathcal{L}_{rec}$, $\mathcal{L}_{deg}$, and $\mathcal{L}_{cont}$. We conduct experiments with 8 sets of different generator selection settings. As an example, 'g4k3' indicates that DR adaptively selects 3 out of 4 different-scale degradation kernel generators based on the router $\mathcal{R}$, producing 3 degradation kernels of different scales. Firstly, we observe that the RMSE of 'g4k1' is lower than that of 'g1k1', 'g2k1', 'g3k1', and 'g5k1', indicating that more generators may not necessarily result in better performance. Secondly, 'g4k3' achieves better DSR performance than 'g4k1', 'g4k2', and 'g4k4'. Therefore, we select 'g4k3' as the setting for DORNet.

5. Conclusion

In this paper, we proposed the degradation oriented and regularized network, a novel real-world DSR solution that learns degradation representations of low-resolution depth to provide targeted guidance. Specifically, we designed a self-supervised degradation learning strategy to model the degradation representations using routing selection-based degradation regularization. This enables label-free implicit degradation learning that adaptively addresses unknown degradation in real-world scenes. Furthermore, we developed a degradation-oriented feature transformation module to perform effective RGB-D fusion. Based on the learned degradation priors, the module selectively propagates RGB content into depth, thereby restoring accurate high-resolution depth. Extensive experiments demonstrate the effectiveness and superiority of our method.

Acknowledgements

This work was supported by the National Science Fund of China under Grant Nos. U24A20330 and 62361166670.

References

[1] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *ICCV*, pages 23555–23564, 2023. 4

[2] Xuanhong Chen, Hang Wang, Jialiang Chen, Kairui Feng, Jinfan Liu, Xiaohang Wang, Weimin Zhang, and Bingbing Ni. Intrinsic phase-preserving networks for depth super resolution. In *AAAI*, pages 1210–1218, 2024. 1

[3] Jaesung Choe, Sunghoon Im, Francois Rameau, Minjun Kang, and In So Kweon. Volumefusion: Deep depth fusion for 3d scene reconstruction. In *ICCV*, pages 16086–16095, 2021. 1

[4] Riccardo De Lutio, Alexander Becker, Stefano D'Aronco, Stefania Russo, Jan D Wegner, and Konrad Schindler. Learning graph regularisation for guided super-resolution. In *CVPR*, pages 1979–1988, 2022. 1, 6, 7

[5] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2020. 2, 5, 6, 7

[6] Junkai Fan, Kun Wang, Zhiqiang Yan, Xiang Chen, Shangbing Gao, Jun Li, and Jian Yang. Depth-centric dehazing and depth-estimation from real-world hazy driving video. *arXiv preprint arXiv:2412.11395*, 2024. 1

[7] Xiao Gu, Yao Guo, Fani Deligianni, and Guang-Zhong Yang. Coupled real-synthetic domain adaptation for real-world deep depth enhancement. *IEEE Transactions on Image Processing*, 29:6343–6356, 2020. 3

[8] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing*, 28(5):2545–2557, 2018. 2

[9] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *CVPR*, pages 9229–9238, 2021. 2, 3, 5, 6, 7

[10] Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Frequency-adaptive pan-sharpening with mixture of experts. In *AAAI*, pages 2121–2129, 2024. 4

[11] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *ECCV*, pages 353–369, 2016. 2

[12] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021. 2, 5, 6, 7

[13] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *SIGGRAPH*, pages 689–694. 2004. 6

[15] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17452–17462, 2022. 3

[16] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *ECCV*, pages 736–753, 2022. 3

[17] Ling Li, Xiaojian Li, Shanlin Yang, Shuai Ding, Alireza Jolfaei, and Xi Zheng. Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. *IEEE Transactions on Industrial Informatics*, 17(6):3920–3928, 2020. 1

[18] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, pages 154–169, 2016. 2, 5, 6

[19] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1909–1923, 2019. 2, 5, 6

[20] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, pages 574–591, 2022. 3

[21] Wei Liu, Xiaogang Chen, Jie Yang, and Qiang Wu. Robust color guided depth map restoration. *IEEE Transactions on Image Processing*, 26(1):315–327, 2016. 3

[22] Xianming Liu, Deming Zhai, Rong Chen, Xiangyang Ji, Debin Zhao, and Wen Gao. Depth restoration from rgb-d data via joint adaptive regularization and thresholding on manifolds. *IEEE Transactions on Image Processing*, 28(3):1068–1079, 2018. 3

[23] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *CVPR*, pages 18237–18246, 2023. 2, 6, 7

[24] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 4

[25] Wuxuan Shi, Mang Ye, and Bo Du. Symmetric uncertainty-aware feature transmission for depth super-resolution. In *ACMMM*, pages 3867–3876, 2022. 5, 6, 7

[26] Jisu Shin, Seunghyun Shin, and Hae-Gon Jeon. Task-specific scene structure representations. In *AAAI*, pages 2272–2281, 2023. 1

[27] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 6

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[29] Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang. Channel attention based iterative residual learning for depth map super-resolution. In *CVPR*, pages 5631–5640, 2020. 3

[30] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, pages 11166–11175, 2019. 6, 7

[31] Baoli Sun, Xinchen Ye, Baopu Li, Haojie Li, Zhihui Wang, and Rui Xu. Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *CVPR*, pages 7792–7801, 2021. 1, 6

[32] Qi Tang, Runmin Cong, Ronghui Sheng, Lingzhi He, Dan Zhang, Yao Zhao, and Sam Kwong. Bridgenet: A joint learn-

ing network of depth map super-resolution and monocular depth estimation. In *ACMMM*, pages 2148–2157, 2021. 2

[33] Haotian Wang, Meng Yang, Ce Zhu, and Nanning Zheng. Rgb-guided depth map recovery by two-stage coarse-to-fine dense crf models. *IEEE Transactions on Image Processing*, 32:1315–1328, 2023. 1

[34] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *ICCV*, pages 16055–16064, 2021. 1

[35] Kun Wang, Zhiqiang Yan, Junkai Fan, Wanlu Zhu, Xiang Li, Jun Li, and Jian Yang. Dcdepth: Progressive monocular depth estimation in discrete cosine domain. In *NeurIPS*, pages 64629–64648, 2024. 1

[36] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, pages 10581–10590, 2021. 3, 7

[37] Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *AAAI*, pages 5823–5831, 2024. 2, 5, 6, 7

[38] Zhengxue Wang, Zhiqiang Yan, Ming-Hsuan Yang, Jinshan Pan, Jian Yang, Ying Tai, and Guangwei Gao. Scene prior filtering for depth map super-resolution. *arXiv preprint arXiv:2402.13876*, 2024. 2

[39] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021. 4

[40] Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Knowledge distillation based degradation estimation for blind super-resolution. In *ICLR*, 2023. 3, 7

[41] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *ECCV*, pages 378–395, 2022. 1

[42] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Guangyu Li, Jun Li, and Jian Yang. 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, 2022. 3

[43] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *ECCV*, pages 214–230, 2022. 1

[44] Zhiqiang Yan, Xiang Li, Kun Wang, Shuo Chen, Jun Li, and Jian Yang. Distortion and uncertainty aware loss for panoramic depth completion. In *ICML*, pages 39099–39109, 2023. 1

[45] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Desnet: Decomposed scale-consistent network for unsupervised depth completion. In *AAAI*, pages 3109–3117, 2023. 1

[46] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *CVPR*, pages 4874–4884, 2024. 1, 6

[47] Zhiqiang Yan, Zhengxue Wang, Kun Wang, Jun Li, and Jian Yang. Completion as enhancement: A degradation-aware selective image guided network for depth completion. *arXiv preprint arXiv:2412.19225*, 2024. 1

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 1

[49] Yuxiang Yang, Qi Cao, Jing Zhang, and Dacheng Tao. Codon: On orchestrating cross-domain attentions for depth super-resolution. *International Journal of Computer Vision*, 130(2):267–284, 2022. 1

[50] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing*, 29:7427–7442, 2020. 2

[51] Guanghao Yin, Wei Wang, Zehuan Yuan, Wei Ji, Dongdong Yu, Shouqian Sun, Tat-Seng Chua, and Changhu Wang. Conditional hyper-network for blind super-resolution with multiple degradations. *IEEE Transactions on Image Processing*, 31:3949–3960, 2022. 3

[52] Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Recurrent structure attention guidance for depth super-resolution. In *AAAI*, pages 3331–3339, 2023. 1

[53] Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Structure flow-guided network for real depth super-resolution. In *AAAI*, pages 3340–3348, 2023. 3, 5, 6, 7

[54] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *CVPR*, pages 5825–5835, 2023. 3

[55] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 4, 5

[56] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5697–5707, 2022. 2, 5, 6, 7

[57] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. In *ICCV*, pages 12547–12558, 2023. 2, 5, 6, 7

[58] Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, Zhiwen Chen, and Xiangyang Ji. High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. *IEEE Transactions on Image Processing*, 31:648–663, 2021. 1

[59] Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, and Xiangyang Ji. Deep attentional guided image filtering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[60] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 5