Intrinsic Phase-Preserving Networks for Depth Super Resolution

Xuanhong Chen[1,2*], Hang Wang[3*], Jialiang Chen[1,2], Kairui Feng[4], Jinfan Liu[1], Xiaohang Wang[1], Weimin Zhang[1,2], Bingbing Ni[1,2 †]

[1]Shanghai Jiao Tong University, Shanghai 200240, China
[2]USC-SJTU Institute of Cultural and Creative Industry
[3]Huawei
[4]National Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University
{chen19910528, nibingbing}@sjtu.edu.cn

Abstract

Depth map super-resolution (DSR) plays an indispensable role in 3D vision. We discover an non-trivial spectral phenomenon: the components of high-resolution (HR) and low-resolution (LR) depth maps manifest the same intrinsic phase, and the spectral phase of RGB is a superset of them, which suggests that a phase-aware filter can assist in the precise use of RGB cues. Motivated by this, we propose an intrinsic phase-preserving DSR paradigm, named IPPNet, to fully exploit inter-modality collaboration in a mutually guided way. In a nutshell, a novel Phase-Preserving Filtering Module (PPFM) is developed to generate dynamic phase-aware filters according to the LR depth flow to filter out erroneous noisy components contained in RGB and then conduct depth enhancement via the modulation of the phase-preserved RGB signal. By stacking multiple PPFM blocks, the proposed IPPNet is capable of reaching a highly competitive restoration performance. Extensive experiments on various benchmark datasets, e.g., NYU v2, RGB-D-D, reach SOTA performance and also well demonstrate the validity of the proposed phase-preserving scheme. Code: https://github.com/neuralchen/IPPNet/.

Introduction

As a necessary means of environmental perception, depth sensors are widely equipped in edge devices (e.g., mobile phones, VR/AR glasses), enabling new applications such as visual SLAM and gesture recognition. Constrained by cost and sensor capability, the acquired depth map is usually of low-resolution (LR) (e.g., Apple: 24×24, Huawei: 240×180), seriously hindering its application. As such, DSR has been extensively studied (Li et al. 2018; Xie, Feris, and Sun 2016; Riegler, Rüther, and Bischof 2016; Song, Dai, and Qin 2019; Sun et al. 2021) as one of the promising solutions. Generally, the depth map is usually associated with a high-resolution (HR) color image (typically, 1920×1080) with nearly-aligned content, motivating the usage of RGB images for recovering the HR depth map, i.e., RGB guided DSR (Li et al. 2019; Kim, Ponce, and Ham 2021; Tang, Chen, and Zeng 2021; He et al. 2021; Wang et al. 2023b) .

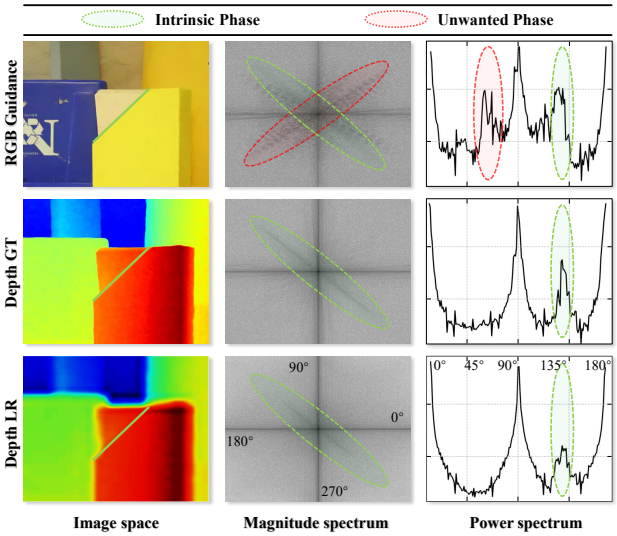

Figure 1: Observing the spectrum maps, we find that depth maps discard specific-phase signals in RGB images that correspond to texture patterns with specific gradients. Thus, direct utilization of RGB information to guide the depth map upsampling will bring undesirable artifacts. Inspired by this, we propose a phase preserving filtering scheme to selectively preserve desired geometric patterns while pruning erroneous components.

From the perspective of information utilization, existing DSR frameworks can be classified into two categories: single depth map super-resolution (SDSR) and RGB-guided depth map super-resolution (GDSR). Insufficient available information (i.e., depth map only) makes SDSR lag behind in performance, which results in SDSR being rarely adopted. In contrast, GDSR methods focus on exploring the full utilization of the corresponding RGB images, including two types of cooperative guidance schemes, i.e., joint filtering (Su et al. 2019; Kim, Ponce, and Ham 2021), and progressive feature aggregation (He et al. 2021; Zhong et al. 2022; Ye et al. 2020). These approaches usually use RGB information to strengthen the features of depth map unidirectionally and obtain promising performance gain; however, unrestricted RGB information often brings over-texture arti-

*Equal contribution.
†Corresponding author: Bingbing Ni.

facts, resulting in undesirable performance degradation.

For a long time, various works (Xie, Feris, and Sun 2016; Kim, Ponce, and Ham 2021; He et al. 2021) have been devoted to better exploration of edge information in RGB images. However, utilization of edge/texture presents a "cut-both-ways" effect. Concretely, edges in RGB images are extremely diverse and complex (e.g., character, logo), and most of them are detrimental to DSR compared to depth map (i.e., only object outlines), often leading to over-textured predictions. To reveal this, we delve into spectrum analysis of data (e.g., Discrete Fourier Transform (DFT)) and find that depth maps and RGB images show significant differences in frequency domain. As shown in Fig. 1, we can observe that spectral components are mainly clustered on the spokes corresponding to the principal geometric patterns (i.e., lines with certain-direction gradient) in intensity space. In this paper, we call the phase of these spokes as the intrinsic phase. Spectral components located in those spokes actually correspond to principal texture patterns of original depth map and RGB image. Comparing the two figures, there are fewer spokes in the depth map than in the RGB, illustrating that the depth map discards specific-phase components in corresponding RGB image. The vanishing components are considered as noise for depth map. This spectral property suggests that a phase-aware filter should be designed for noisy texture pruning during the process of guiding DSR.

Inspired by the above guidelines and spectral property, we present a brand-new phase-preserving network, named *IPPNet*, to realize efficient and high-performance cooperation of RGB and depth map in GDSR. In the heart of IPPNet, a novel *Phase-Preserving Filtering Module* (PPFM) is proposed for cross-content guided collaborative filtering between depth map features and color image features. In detail, PPFM contains two information streams, namely a phase-aware filtering stream and a modulation stream. The former generates a set of phase-specific filters/masks based on intrinsic phase of depth map features to mask out unwanted components in RGB features, and the latter is in charge of strengthening depth map features with phase-aligned RGB features. Despite its simplicity, PPFM can greatly improve spectral properties, achieving compelling performance while maintaining the desired model size.

We extensively experiment with the proposed framework in terms of both qualitative and quantitative evaluations on mainstream benchmark datasets (Silberman et al. 2012; Scharstein and Pal 2007; Lu, Ren, and Liu 2014; Park et al. 2011) and real-world dataset (He et al. 2021). It is demonstrated that our IPPNet achieves state-of-the-art performance (e.g., NYU v2 (Silberman et al. 2012) $16\times$: 4.68, Middlebury (Scharstein and Pal 2007) $8\times$: 1.51, Lu (Lu, Ren, and Liu 2014) $16\times$: 2.88). In particular, our IPPNet outperforms all existing methods on real-world branch of RGB-D-D dataset. Comprehensive ablation experiments and visualizations validate that our proposed phase-preserving filtering scheme outperforms with existing unidirectional schemes.

Related Work

Depth Map Super-Resolution (DSR). SDSR methods (Ferstl et al. 2013; Xie, Feris, and Sun 2016; Riegler, Rüther, and Bischof 2016; Song, Dai, and Qin 2019; Sun et al. 2021) leverage only LR depth map to recover accurate HR depth map, often resulting in limited performance. GDSR methods utilize RGB modality as guidance to reconstruct the degraded LR depth map and has become the mainstream. Early works (Tomasi and Manduchi 1998; Kopf et al. 2007; Yang et al. 2007) based on heuristic rules to transfer fine-grained structures from the guidance image. Later, data-driven approaches like DJF (Li et al. 2016) and DG (Gu et al. 2017) use CNNs to extract features and directly regress the target image. DKN (Kim, Ponce, and Ham 2021) proposes to learn spatially-variant kernels explicitly for each pixel to obtain upsampling results as a weighted average. JIIF (Tang, Chen, and Zeng 2021) represents target image with local latent codes from both input image and guide image and learns interpolation weights via graph attention mechanism. However, former works (Su et al. 2019; Kim, Ponce, and Ham 2021; He et al. 2021) utilize RGB information to strengthen depth map unidirectionally, while such a guidance scheme brings over-texture artifacts. In this work, we design PPFM to strengthen depth features with phase-aligned RGB features, obtaining more pleasant fusion quality.

Guided Depth Completion. Recently, several methods (Albahar and Huang 2019; Hu et al. 2021; Tang et al. 2021; Zhao et al. 2021; Rho, Ha, and Kim 2022; Lee et al. 2022) utilize dual-branch framework to separately extract features of sparse depth and color images, and fuse them for final depth predictions. Although these works (Tang et al. 2021; Zhao et al. 2021; Rho, Ha, and Kim 2022) share a similar framework structure to our model, we aim an essentially different target and working mechanism. Instead of heuristic design, we conduct in-depth theoretical analysis and find the spectrum phenomenon, suggesting that a phase-preserving filter mechanism should be designed to selectively exploit RGB features. Furthermore, we design the phase-preserving filtering module to gain precise fusion according to the homogeneous spectral properties of LR and HR depth maps. The above features make our framework significantly different from previous designs. we are the first to conduct an in-depth theoretical analysis and design in GDSR. We emphasize that RGB features should be so as to filter out unwanted components and conduct depth enhancement via the modulation of the phase-preserved RGB signal.

Methodology

Problem Definition

In GDSR task, there is a LR depth image $D_L \in \mathbb{R}^{H\times W\times 1}$ and its corresponding HR RGB guidance $G \in \mathbb{R}^{sH\times sW\times 3}$, where s is the scaling factor. The goal is to up-sample the LR depth image to recover the HR depth target $D_H \in \mathbb{R}^{sH\times sW\times 1}$ under the guidance of HR RGB image, where H and W are the width and height of the input depth image. Specifically, bicubic interpolation is first employed on the LR depth image D_L to get an up-sampled high-resolution initialization $D_\uparrow \in \mathbb{R}^{sH\times sW\times 1}$. Then the obtained RGBD pair $(D_\uparrow, G)$ is feed into the neural network $\mathcal{F}$ to recover residual details between D_H and $D_\uparrow$. HR depth target D_H is

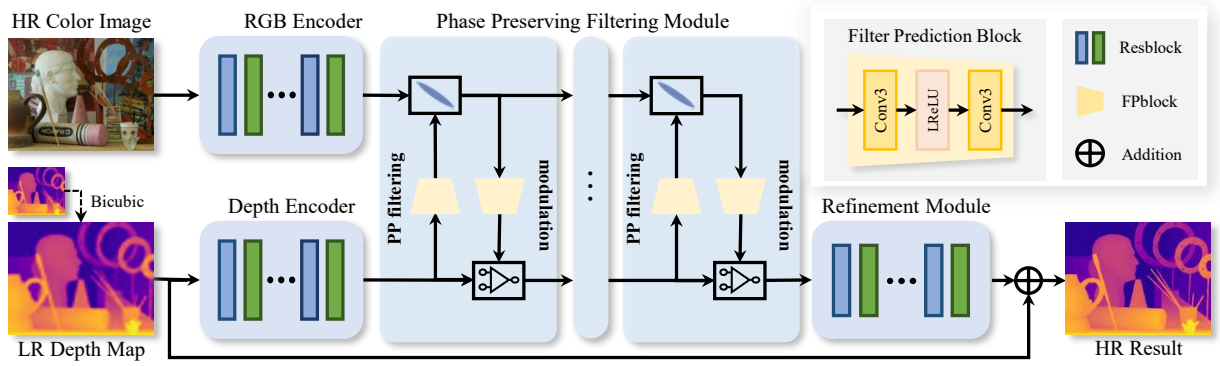

Figure 2: IPPNet framework overview. IPPNet consists of RGB encoder, depth encoder, PPFM, and refinement module.

calculated as $D_H = D_\uparrow + \mathbf{\Phi}(D_\uparrow, G)$, where $\mathbf{\Phi}$ is the neural network to learn the residual details.

The Phase Power Spectrum

In order to analyze the spectral distribution of RGB images and depth maps, we rely on the DFT and the Fourier power spectrum. We compute the spectral representation for depth map and RGB image from the discrete Fourier Transform $\mathcal{F}$ of 2D data I of size $H \times W$,

$$\mathcal{F}(I)(k,\ell) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} e^{-2\pi j \cdot \frac{wk}{W}} e^{-2\pi j \cdot \frac{h\ell}{H}} \cdot I(h,w), \quad \text{for} \quad k = 0,\dots,H-1, \quad \ell = 0,\dots,W-1, \tag{1}$$

via phase integration over set of spectra with phase φ:

$$AI(\varphi) = \int_{\Omega_\varphi} |\mathcal{F}(\omega)|\, e^{j\varphi} \mathrm{d}\omega, \tag{2}$$

where Ω_φ represents the set of spectra with phase φ. $\omega \in \Omega_\varphi$ denotes a spectra frequency with phase φ. In this paper, we employ phase power spectrum to count the intensities of components of different phases. It is worth pointing out that the phase power spectrum calculates results along the phase dimension rather than the frequency direction, which enables us to measure the power of components with specific phases. We use this tool to quantify the ability of our framework to filter for specific phase components, thereby validating our motivation.

The Phase-Preserving Module

Previous GDSR methods (Su et al. 2019; Kim, Ponce, and Ham 2021; He et al. 2021) devote efforts to design heuristic joint filtering schemes, which directly transfer the RGB guidance to depth flow via hand-craft or model-induced fusion modules. As revealed in spectrum properties, RGB contains a lot of extra undesired texture patterns for depth map. These extra textures correspond to surface details rather than real object shapes, which are erroneous noise for depth maps, and naively using them introduces unwanted artifacts, e.g., over-texture, and halo artifacts. Therefore, it is crucial to eliminate the noise patterns to avoid propagating erroneous structure details, which requires precise filters tailored to the input. Due to the homology of LR and HR depth map (i.e., LR is produced by HR via a specific degradation kernel), they exhibit highly similar spectral properties, e.g., most of the spectral components of the two maps overlap in phase. Whiteness this, we argue that the tailored anisotropic filters can be derived from the LR depth map to suppress chaos in RGB. The guided RGB image G can be represented as a combination of two components:

$$G = \hat{G} + n, \tag{3}$$

where $\hat{G}$ represents the component that is consistent with the spectral phase of the depth signal. n denotes the textured details. However, these detailed textures are undesirable noise for the depth map.

Based on above analysis, we propose a novel cross-modality guided *Phase-Preserving Filtering Module* to achieve phase-aware filtering between depth map features and color image features. Specifically, the proposed paradigm consists of two information streams, namely the phase-aware filtering stream and the modulation stream, where the former stream is designed to dynamically generate phase-aware filters derived from depth map features to eliminate unwanted texture components in the RGB features, and the latter aims to strengthen the depth map features by modulating with purified RGB features. Given the feature $x_D \in \mathbb{R}^{C\times H\times W}$ of depth map and the feature $x_G \in \mathbb{R}^{C\times H\times W}$ of RGB guidance (C is the number of channels), PPFM can be formulated as,

$$\begin{cases} x_G^i = x_G^i * \mathbf{\Psi}_D^i(x_D^i), \\ x_D^i = x_D^i + \lambda\, x_D^i * \mathbf{\Psi}_G^i(x_G^i), \end{cases} \quad i \in \{0,\cdots,N-1\}, \tag{4}$$

where $\mathbf{\Psi}_D^i$ and $\mathbf{\Psi}_G^i$ represent the spatial kernel generator. In our implementation, we use a simple *Filter Prediction Block* (FPblock), which is a two layers block as shown in Fig. 2. $*$ denotes convolution, and λ is the modulation ratio. In our implementation, λ is fixed as 1. $\mathbf{\Psi}_D^i(x_D^i)$ and $\mathbf{\Psi}_G^i(x_G^i)$ are the generated phase-aware filters explicitly based on depth features and filtered RGB features, respectively. $\mathbf{\Psi}_D^i(x_D^i)$ is designed to filter noise RGB features, and $\mathbf{\Psi}_G^i(x_G^i)$ is used to strengthen depth map with cleaned RGB information, enabling the multi-modality information fusion in a collaborative manner.

Analysis of the Phase-Preserving Filtering Module with DFT

We use DFT to analyze the rationality of above filtering process theoretically. Actually, $\mathbf{\Psi}_D^i$ is designed to adjust the

gain of filters, so we abstract it as an amplification factor $\mathcal{K}$. Then, we can rewrite Eq. 4 as

$$x_G^i * (\mathcal{K} \cdot x_D^i) = \hat{x}_G^i * (\mathcal{K} \cdot x_D^i) + n^i * (\mathcal{K} \cdot x_D^i). \tag{5}$$

We decompose it using DFT:

$$\begin{aligned} &\mathcal{F}\left(x_G^i * \left(\mathcal{K} \cdot x_D^i\right)\right) \\ &= \mathcal{F}\left(\hat{x}_G^i\right) \cdot \mathcal{F}\left(\mathcal{K} \cdot x_D^i\right) + \mathcal{F}\left(n^i\right) \cdot \mathcal{F}\left(\mathcal{K} \cdot x_D^i\right) \\ &= \int_0^{2\pi} \left|\mathcal{F}\left(\hat{x}_G^i\right)\right| e^{j\varphi} \cdot \left|\mathcal{F}\left(\mathcal{K} \cdot x_D^i\right)\right| e^{j\varphi} \mathrm{d}\varphi \\ &+ \int_0^{2\pi} \left|\mathcal{F}\left(n^i\right)\right| e^{j\varphi} \cdot \left|\mathcal{F}\left(\mathcal{K} \cdot x_D^i\right)\right| e^{j\varphi} \mathrm{d}\varphi. \end{aligned} \tag{6}$$

We finally express the DFT results in polar form, where φ is the phase and $|\cdot|$ is the magnitude. According to above spectrum properties, when φ falls within the intrinsic phase of depth map, the first term in Eq. 6 will remain and the second term will approach 0. Otherwise, both will be close to 0. In this way, our module can achieve precise denoise for RGB. Additionally, we show a toy-level experiment in Fig. 1 of *supplementary material*, and it can be seen that the spectrum of filtered RGB features is very close to the depth.

Considering that the feature embedding of a specific position has integrated the information of its neighborhoods after feature extraction, i.e., it has encoded the features of a patch region in the input space (Isola et al. 2017; Chen et al. 2020), we further reduce the kernel size of filters $\mathbf{\Psi}_D^i(x_D^i)$ and $\mathbf{\Psi}_G^i(x_G^i)$ in Eq. 4 to 1 to make this operator more flexible and efficient. It can be readily implemented by pixel-wise multiplication (i.e., modulation (Hu et al. 2022; Chen et al. 2023; Chen, Wang, and Ni 2021) or gating (Chen et al. 2022)), enabling a more simplified structure and lower computational overhead, as

$$x_G^i = x_G^i \odot \mathbf{\Psi}_D^i(x_D^i), \tag{7}$$

where $i \in \{0, \cdots, N-1\}$. $\odot$ denotes pixel-wise multiplication. Note that the difference between Eqs. 4 and 7 is that the way of information fusion is simplified from convolution to element-wise multiplication. In fact, the pixel-to-pixel multiplication in feature space is equivalent to the patch-to-path convolution in input space.

Network Architecture

Based on the proposed PPFM, we further propose a simple and efficient framework, named IPPNet, to achieve high-performance DSR. It contains four components, i.e., RGB encoder, depth encoder, PPFM, and refinement module. Two feature encoders convert RGB image and depth map into embedding space, respectively. Then, RGB features and depth features are integrated in a mutually guided and impulsed way by PPFM. Finally, the refinement module further enriched the deep features to produce more accurate results. The whole framework pipeline is shown in Fig. 2.

Feature Extraction. The feature embeddings of RGB image and depth map are generated by RGB encoder and depth encoder, respectively. Following EDSR (Lim et al. 2017), both encoders are composed of ResBlocks (He et al. 2016) without BN (Ioffe and Szegedy 2015). Such a simple modification brings more range flexibility to the network and thus contributes to better performance.

Image Reconstruction. The processed depth feature is further enriched by the refinement module, which also consists of several ResBlocks without normalization layers. Finally, a convolution layer is applied to the refined features to generate the residual image $R \in \mathbb{R}^{H \times W \times 1}$, and the restored depth map is obtained by the sum of residual image and the input depth image, $\widehat{D}_H = D_\uparrow + R$.

Optimization Objective. Following previous image restoration works (Kim, Ponce, and Ham 2021; Tang, Chen, and Zeng 2021; Zamir et al. 2022; Wang et al. 2023a,c), we train the whole model by minimizing a standard $\mathcal{L}_1$ constraint between the predicted depth map $\widehat{D}_H$ and the ground truth depth map D_H^{gt}, which can be written as

$$\mathcal{L} = ||D_H^{gt} - \widehat{D}_H||_1. \tag{8}$$

Experiments

Guided Depth Map Super-Resolution

Implementation Details. Following JIIF (Tang, Chen, and Zeng 2021), HR image is randomly cropped into (256, 256) patches during training. LR input depth map is generated from HR ground truth using bicubic downsampling at different ratios ($4\times, 8\times, 16\times$). Depth encoder, RGB encoder and refinement module consist of 4 Resblocks (He et al. 2016) without batch normalization layer. The number N of phase-preserving filtering blocks is set to 3. Adam(Kingma and Ba 2015) optimizer is used to train our model. The learning rate is initially set to 1×10^{-4} and then halved every 8K iterations. Our method is implemented with PyTorch (Paszke et al. 2017), and one NVIDIA Tesla V100 GPU is used for training. Detailed implementation of phase-preserving filter is provided in supplementary material. The source code will be released for reproducibility.

Evaluation Metric. Root mean squared error (RMSE) serves as the evaluation metric to measure the difference between prediction and ground truth. The RMSE is first calculated for each depth map and finally averaged.

Experiments on NYU v2. NYU v2 (Silberman et al. 2012) contains 1449 RGB/D pairs collected by Kinect, of which 1000 RGB/D pairs are used for training, and 449 RGB/D pairs are used for testing. Our model is also tested on Middlebury (Scharstein and Pal 2007; Hirschmüller and Scharstein 2007) and Lu (Lu, Ren, and Liu 2014).

Results on NYU v2. Table 1 compares the RMSE results for $4\times, 8\times, 16\times$ upsampling of existing approaches on three benchmarks. It can be observed that our method achieves the best performance on six of nine tasks and obtains the best performance in terms of average RMSE, even under challenging scaling factors of $8\times$ and $16\times$. In particular, on the LU dataset, IPPNet surpasses the current state-of-the-art approaches with a notable margin (i.e., $4\times : 0.11$, $8\times : 0.11$, $16\times : 0.89$). These empirical results verify the superiority of the cross-modality phase-preserving filtering over depth images and RGB images, in which the depth information is first used to filter the erroneous noise contained in the RGB

Methods	NYU v2			Middlebury			LU			Average		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	4.28	7.14	11.58	2.28	3.98	6.37	2.42	4.54	7.38	3.00	5.22	8.44
GF (He, Sun, and Tang 2013)	5.84	7.86	12.41	3.24	4.36	6.79	4.18	5.34	8.02	4.42	5.85	9.07
DG (Gu et al. 2017)	3.68	5.78	10.08	1.97	4.16	5.27	2.06	4.19	6.90	2.57	4.71	7.41
DGF (Wu et al. 2018)	3.21	5.92	10.45	1.94	3.36	5.81	2.45	4.42	7.26	2.53	4.57	7.84
DJF (Li et al. 2016)	2.80	5.33	9.46	1.68	3.24	5.62	1.65	3.96	6.75	2.04	4.18	7.28
DMSG (Hui, Loy, and Tang 2016)	3.02	5.38	9.17	1.88	3.45	6.28	2.30	4.17	7.22	2.40	4.33	7.57
DJF (Li et al. 2019)	2.38	4.94	9.18	1.32	3.19	5.57	1.15	3.57	6.77	1.62	3.90	7.17
DSRNet (Guo et al. 2019)	3.00	5.16	8.41	1.77	3.05	4.96	1.77	3.10	5.11	2.18	3.77	6.16
PAC (Su et al. 2019)	1.89	3.33	6.78	1.32	2.62	4.58	1.20	2.33	5.19	1.47	2.76	5.52
FDKN (Kim, Ponce, and Ham 2021)	1.86	3.58	6.96	1.08	2.17	4.50	*0.82*	2.10	5.05	1.25	2.62	5.50
DKN (Kim, Ponce, and Ham 2021)	1.62	3.26	6.51	1.23	2.12	4.24	0.96	2.16	5.11	1.27	2.51	5.29
FDSR (He et al. 2021)	1.61	3.18	5.86	1.13	2.08	4.39	1.29	2.19	5.00	1.34	2.48	5.08
DCTNet (Zhao et al. 2022)	1.59	3.08	5.80	*1.05*	2.00	4.20	0.87	1.87	4.34	1.17	2.32	4.78
JIIF (Tang, Chen, and Zeng 2021)	1.37	2.76	5.27	1.09	1.82	*3.31*	0.85	1.73	4.16	1.10	2.10	4.25
AHMF (Zhong et al. 2022)	1.40	2.89	5.64	1.07	*1.63*	**3.14**	0.88	1.66	*3.71*	1.12	2.06	4.16
PMBANet (Ye et al. 2020)	**1.06**	**2.28**	*4.98*	1.23	2.02	3.36	0.90	*1.62*	3.82	*1.06*	*1.97*	*4.05*
IPPNet (ours)	*1.20*	*2.45*	**4.69**	**0.98**	**1.51**	*3.31*	**0.79**	**1.48**	**2.88**	**0.99**	**1.81**	**3.63**

Table 1: Quantitative depth map upsampling results on three benchmark datasets. The best performance is marked with Bold (1st best), Italics (2nd best). For NYU v2 dataset we calculate in centimeters, for Middlebury dataset and Lu dataset we calculate RMSE with depth value scaled to the range [0, 255] (Lower RMSE values, better performance).

Methods	RMSE	Methods	RMSE
SVLRM (2019)	8.05	DJF (2016)	7.90
DJFR (2019)	8.01	DKN (2021)	7.38
FDKN (2021)	7.50	FDSR (2021)	7.50
DCTNet (2022)	*7.37*	**IPPNet** (ours)	**7.15**
FDSR* (2021)	5.49	DCTNet* (2022)	*5.43*
IPPNet* (ours)	**5.32**		

Table 2: Quantitative results on real-world branch of RGB-D-D dataset. Note that FDSR*, DCTNet* and IPPNet* report the results directly trained on real-world branch data.

Methods	RMSE			Edge Errors		
	×4	×8	×16	×4	×8	×16
SVLRM	3.39	5.59	8.28	5.08	15.18	34.30
DJF	3.41	5.57	8.15	5.65	17.07	35.32
DJFR	3.35	5.57	7.99	5.26	15.66	34.54
FDKN	1.18	1.91	3.41	1.39	3.41	11.73
DKN	1.30	1.96	3.42	2.11	3.55	12.93
FDSR	1.13	1.82	3.06	1.38	3.09	12.47
DCTNet	*1.08*	*1.74*	*3.05*	*1.36*	*3.07*	*12.42*
IPPNet	**1.05**	**1.67**	**2.61**	**1.29**	**2.95**	**12.04**

Table 3: RMSE and edge errors results on RGB-D-D. Note that all models are trained on NYU v2 without finetuning.

features, and the pre-filtered RGB information is then utilized to enhance the depth features via modulation.

Fig. 3 provides several qualitative visual comparisons of different approaches. In these examples, our proposed IPPNet produces better fine-grained depth details, while other methods may generate blurred edges or artifacts in complicated areas. For instance, in the 2-nd row, IPPNet produces

Methods	Art			Books		
	4×	8×	16×	4×	8×	16×
Bicubic	6.07	7.27	9.59	5.15	5.45	5.97
DMSG (2016)	6.19	7.26	9.53	5.38	5.18	5.20
PDN (2016)	3.11	4.48	7.35	1.56	2.24	3.46
DG (2017)	2.96	4.41	7.06	1.64	2.35	3.50
DJF (2019)	4.25	6.43	9.05	2.20	3.35	4.94
PAC (2019)	5.34	7.69	10.66	2.11	3.12	4.60
FDKN (2021)	3.14	4.47	7.61	1.49	2.13	3.40
DKN (2021)	3.01	4.14	*7.01*	1.44	2.10	3.09
FDSR (2021)	2.93	4.01	7.23	1.32	1.93	2.95
JIIF (2021)	*2.79*	*3.87*	7.14	*1.30*	*1.75*	*2.47*
IPPNet (ours)	**2.52**	**3.62**	**6.94**	**1.28**	**1.73**	**2.39**

Table 4: RMSE on noisy depth map upsampling.

fewer artifacts around the chair, while other methods generate blurred boundaries. Observing the example in the 4-th row, our method obtains a better result with more accurate edges of leaves and more pleasant visual effects.

Experiments on RGB-D-D. RGB-D-D (He et al. 2021) is a large-scale real-world DSR dataset, which contains 4811 pairs of RGB/D images. Each image pair contains HR RGB image from a mobile phone, real-world LR depth map captured by ToF camera on mobile phone, and an HR depth map captured by industrial Helios camera. Following FDSR (He et al. 2021), we introduce two standards for a fair comparison.

Testing Without Retraining on RGB-D-D. To validate the generalizability of the model, we directly test the model trained on NYU v2 without retraining. The quantitative comparison is shown in Table 3 in terms of RMSE and edge errors introduced by (He et al. 2021). Edge errors are designed to measure the local accuracy, we report the percentage of errors over 1.2% in the edge area. Please refer to

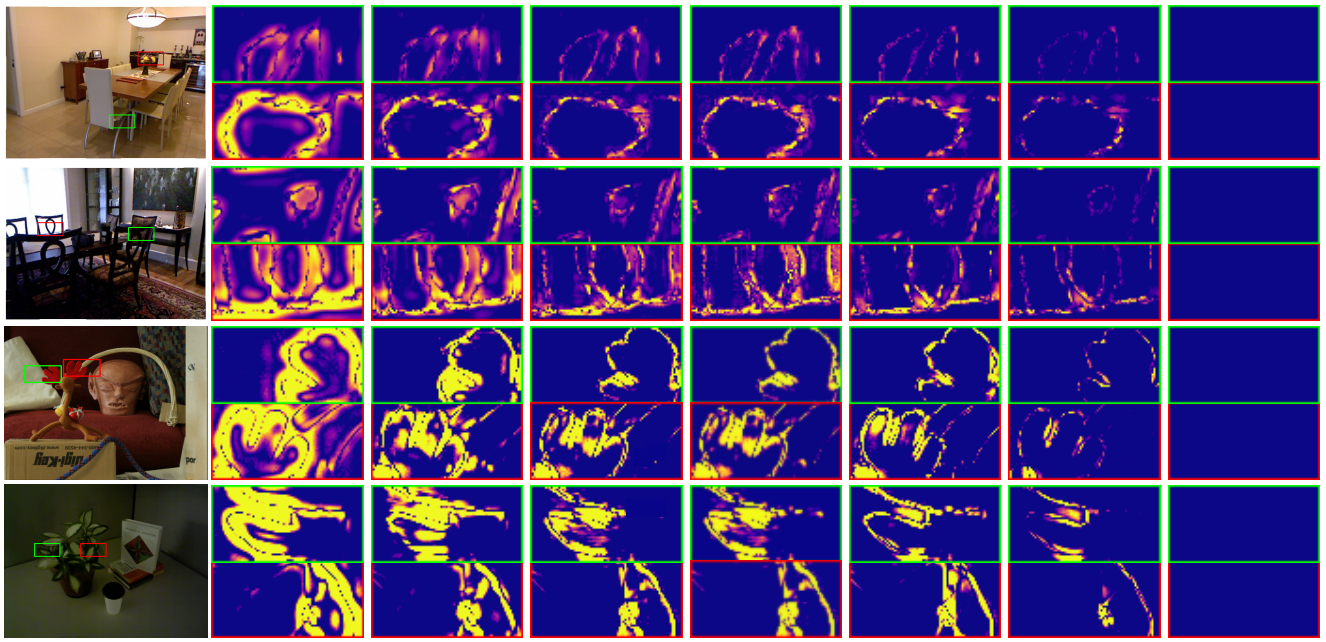

Figure 3: Visual comparisons of $8\times$ GDSR. For better comparison, we show the error map between the results and ground truth. First two rows show the results on NYU v2, and last two rows show the results on Middlebury and Lu, respectively.

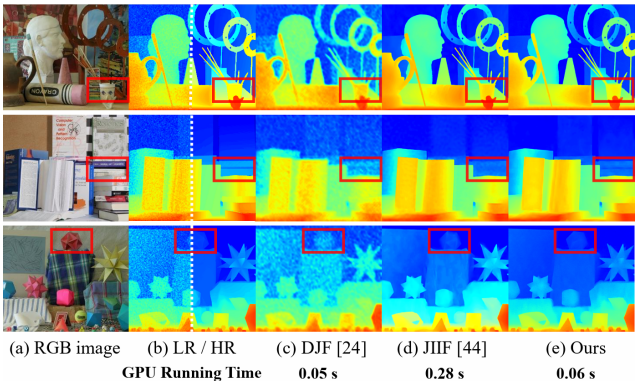

Figure 4: Visual comparisons of $8\times$ guided noisy upsampling results on Noisy Middlebury dataset.

FDSR for more details about the edge errors. We can observe that the proposed IPPNet has best performance, which means our method produces more accurate results.

Training in Real-World Manner on RGB-D-D. Under this setting, we train the model on the real-world branch of RGB-D-D, where LR depth maps are captured by ToF camera of mobile phone in real scenes instead of downsampling from HR depth maps. The size of the LR depth map is 192×144 and the target resolution is 512×384. Other training settings remain the same as in Sec. *Implementation Details*. We also evaluate the existing $\times 4$ models (trained on NYU v2) and report the results in Table 2 for comparison. We can observe that IPPNet achieves state-of-the-art performance over various settings. Specifically, when directly trained on RGB-D-D, IPPNet surpasses current advanced approaches with a notable margin, further demonstrating its superiority and generalization ability. Our results show finer boundaries and more visual pleasant details.

Guided Noisy Depth Map Super-Resolution

Datasets and Training Details. Noisy Middlebury (Park et al. 2011) is used for testing, including 3 RGB/D pairs, *i.e.*, *Art*, *Books*. Following (Kim, Ponce, and Ham 2021), we simulate noisy LR input depth map by adding multiplicative Gaussian noise $\eta(x) = \mathcal{N}(0, \tau x)$, where τ is the magnitude of the noise and x is proportional to the depth value. We train the model using a noisy LR depth map with clean HR color guidance from NYU v2 and do not fine-tune the model on Noisy Middlebury. The other settings remain the same as in Sec. *Implementation Details*.

Results. Table 4 presents the RMSE results for $4\times$, $8\times$, $16\times$ upsampling on Noisy Middlebury. It can be observed that IPPNet achieves state-of-the-art results on all guided noisy depth map super-resolution tasks, demonstrating that our method is capable of handling noisy data. This phenomenon is in line with the better empirical performance of our method on previous tasks, which verifies the superiority of the proposed phase-preserving filtering scheme. Fig. 4 shows the $8\times$ upsampling results of different methods. These visualizations clearly show IPPNet obtains more accurate depth results compared to other algorithms. Inference time for image size of 1376×1088 on a NVIDIA TITAN XP GPU is also tested. It can be observed that our IPPNet achieves high efficiency whilst with the best performance.

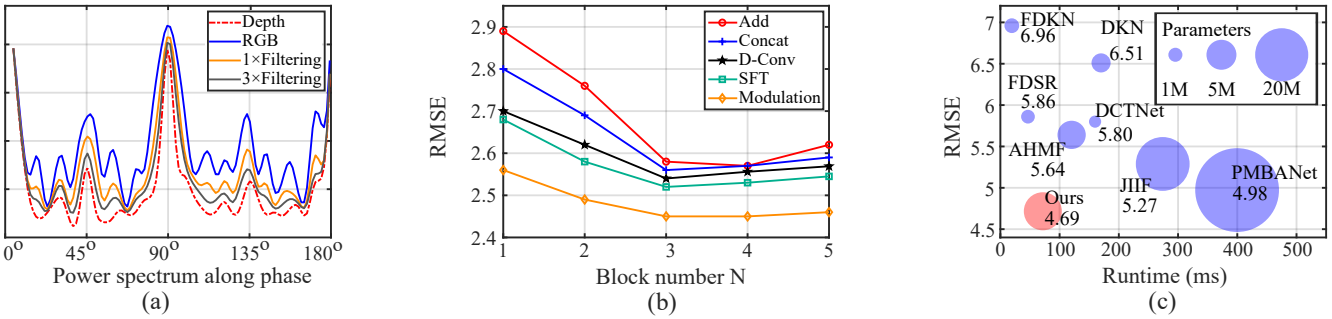

Figure 5: (a) Phase power spectrum of RGB features before and after PPFM. (b) Sensitivity of PPFM's number with different fusion strategies. RMSE in (a) and (b) are reported for 8× SR on NYU v2. (c) Inference time vs. RMSE of different methods for 16× SR scale on NYU v2. It can be observed that our model achieves high efficiency whilst with best performance.

	Configurations	×4	×8	×16
I	w/o RGB guidance	1.92	3.91	7.62
II	w/o Filtering Stream	1.37	2.84	5.61
III	fuseD RGB and depth	1.32	2.76	5.18
IV	w/o Refinement Module	1.36	2.82	5.45
V	place PPFM after RM	1.29	2.64	4.92
	Ours	**1.20**	**2.45**	**4.69**

Table 5: Results of ablation experiments on the NYU v2.

Ablation Study

Spectrum Analysis of PPFM. Fig. 5(a) compares the phase power spectrum of the RGB features before and after PPFM. The curve data is the average result on the whole NYU v2. We also provide the phase power spectrum of LR depth map. After the filtering process of intrinsic phase-preserving filter generated by depth features, the noise patterns contained in RGB features are effectively suppressed. With 3 × PPFMs, most of the undesired textures in RGB images are greatly attenuated, which verifies the effectiveness of PPFM in selectively preserving desired geometric patterns.

Effect of Different Cross-Modality Fusion Strategies. In Fig. 5(b), we evaluate the performance of different fusion strategy. Several widely used fusion methods, i.e., addition, concatenation, deformable convolution (Kim, Ponce, and Ham 2021) and SFT (Wang et al. 2021) are introduced for comparison. The latter two methods are specially designed for image restoration. It can be observed that the setting of *Modulation* delivers the best performance, which demonstrates its effectiveness. Therefore, we adopt this simple strategy as the fusion solution in all experiments.

Model Complexity Comparison with SOTA. Fig. 5 (c) compares the model complexity of current advanced methods for ×16 SR scale. The running time is tested at size of 640 × 480 on one NVIDIA TITAN XP GPU over 10 independent runs. The size of the circle represents the parameters of the corresponding model. From the figure, we can see that our IPPNet achieves high efficiency whilst with best performance, demonstrating the superiority of the phase-preserving filtering scheme for the guided DSR task.

Effect of Different Design Choices. To evaluate the effectiveness of the major components and justify several design choices, we conduct ablation study experiments on NYU v2. The comparison results are reported in Table 5. Due to limited space, we refer readers to the supplementary material for more network structure details. Exp. I only take depth maps as input without information fusion. It is difficult to recover accurate high-resolution depth map without the guidance of HR color images, showing that RGB image plays a vital role in the DSR task. Exp. II removes the filtering stream from the PPFM and only modulation stream is reserved. Since RGB images contain erroneous structure details (i.e., unwanted phase for depth maps) which are harmful to depth maps, such a modification leads to obvious performance degradation. Exp. III fuse RGB and depth simultaneously as previous works (Tang et al. 2021; Zhao et al. 2021; Rho, Ha, and Kim 2022), i.e., depth filters RGB and RGB filters depth in the meantime. Such order-insensitive fusion scheme would risk introducing unwanted textures in RGB. The obtained result is worse than our model. In our scheme, depth is first used to filter RGB to perform phase-preserving filtering to remove over-textured noise. Exp. IV removes RM. The result shows the refinement of depth features is able to further boost model's performance. Exp. V places PPFM after RM, and under such configuration, undesirable performance degradation occurs, which indicates that the cross-modality information fusion should be conducted in the middle stage of the whole pipeline.

Conclusion

In this paper, we propose a phase-preserving filtering DSR framework, named *IPPNet*, to realize more effective cooperation of RGB image and depth map. According to the spectrum phenomenon, we find that RGB guidance contains a lot of extra undesired texture patterns for depth map. Therefore, we design a novel *Phase Preserving Filtering Module* to achieve cross-content guided collaborative filtering between depth and RGB features, and build the IPPNet framework. Extensive experimental results and analytical studies verify its prominent performance under various guided depth map super-resolution settings.

Acknowledgements

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

Albahar, B.; and Huang, J. 2019. Guided Image-to-Image Translation With Bi-Directional Feature Transformation. In *ICCV*, 9015–9024.

Chen, X.; Ni, B.; Liu, N.; Liu, Z.; Jiang, Y.; Truong, L.; and Tian, Q. 2020. CooGAN: A Memory-Efficient Framework for High-Resolution Facial Attribute Editing. In *ECCV*.

Chen, X.; Ni, B.; Liu, Y.; Liu, N.; Zeng, Z.; and Wang, H. 2023. SimSwap++: Towards Faster and High-Quality Identity Swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, X.; Wang, H.; and Ni, B. 2021. X-volution: On the unification of convolution and self-attention. *CoRR*, abs/2106.02253.

Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; and Liu, Z. 2022. Simple Baselines for Image Restoration. In *ECCV*.

Ferstl, D.; Reinbacher, C.; Ranftl, R.; Rüther, M.; and Bischof, H. 2013. Image Guided Depth Upsampling Using Anisotropic Total Generalized Variation. In *ICCV*, 993–1000.

Gu, S.; Zuo, W.; Guo, S.; Chen, Y.; Chen, C.; and Zhang, L. 2017. Learning Dynamic Guidance for Depth Image Enhancement. In *CVPR*, 712–721.

Guo, C.; Li, C.; Guo, J.; Cong, R.; Fu, H.; and Han, P. 2019. Hierarchical Features Driven Residual Learning for Depth Map Super-Resolution. *IEEE TIP*, 28(5): 2545–2557.

He, K.; Sun, J.; and Tang, X. 2013. Guided Image Filtering. *IEEE TPAMI*, 35(6): 1397–1409.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

He, L.; Zhu, H.; Li, F.; Bai, H.; Cong, R.; Zhang, C.; Lin, C.; Liu, M.; and Zhao, Y. 2021. Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline. In *CVPR*.

Hirschmüller, H.; and Scharstein, D. 2007. Evaluation of Cost Functions for Stereo Matching. In *CVPR*.

Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; and Gong, X. 2021. PENet: Towards Precise and Efficient Image Guided Depth Completion. In *ICRA*.

Hu, X.; Chen, X.; Ni, B.; Li, T.; and Liu, Y. 2022. Bivolution: A Static and Dynamic Coupled Filter. In *AAAI*.

Hui, T.; Loy, C. C.; and Tang, X. 2016. Depth Map Super-Resolution by Deep Multi-Scale Guidance. In *ECCV*, 353–369.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 448–456.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 5967–5976.

Kim, B.; Ponce, J.; and Ham, B. 2021. Deformable Kernel Networks for Joint Image Filtering. *IJCV*, 129(2): 579–600.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kopf, J.; Cohen, M. F.; Lischinski, D.; and Uyttendaele, M. 2007. Joint bilateral upsampling. *ACM Trans. Graph.*, 26(3): 96.

Lee, Y.; Park, S.; Kang, B.; and Park, H. 2022. Multimodal Characteristic Guided Depth Completion Network. In *ACCV*.

Li, J.; Fang, F.; Mei, K.; and Zhang, G. 2018. Multi-scale Residual Network for Image Super-Resolution. In *ECCV*, volume 11212, 527–542.

Li, Y.; Huang, J.; Ahuja, N.; and Yang, M. 2016. Deep Joint Image Filtering. In *ECCV*.

Li, Y.; Huang, J.; Ahuja, N.; and Yang, M. 2019. Joint Image Filtering with Deep Convolutional Networks. *Trans. PAMI*, 41(8): 1909–1923.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *CVPRW*.

Lu, S.; Ren, X.; and Liu, F. 2014. Depth Enhancement via Low-Rank Matrix Completion. In *CVPR*.

Pan, J.; Dong, J.; Ren, J. S. J.; Lin, L.; Tang, J.; and Yang, M. 2019. Spatially Variant Linear Representation Models for Joint Filtering. In *CVPR*.

Park, J.; Kim, H.; Tai, Y.; Brown, M. S.; and Kweon, I. 2011. High quality depth map upsampling for 3D-TOF cameras. In *ICCV*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic Differentiation in PyTorch. In *NeurIPS Workshop*.

Rho, K.; Ha, J.; and Kim, Y. 2022. GuideFormer: Transformers for Image Guided Depth Completion. In *CVPR*.

Riegler, G.; Ferstl, D.; Rüther, M.; and Bischof, H. 2016. A Deep Primal-Dual Network for Guided Depth Super-Resolution. In *BMVC*.

Riegler, G.; Rüther, M.; and Bischof, H. 2016. ATGV-Net: Accurate Depth Super-Resolution. In *ECCV*, 268–284.

Scharstein, D.; and Pal, C. 2007. Learning Conditional Random Fields for Stereo. In *CVPR*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Song, X.; Dai, Y.; and Qin, X. 2019. Deeply Supervised Depth Map Super-Resolution as Novel View Synthesis. *IEEE TCSVT*, 29(8): 2323–2336.

Su, H.; Jampani, V.; Sun, D.; Gallo, O.; Learned-Miller, E. G.; and Kautz, J. 2019. Pixel-Adaptive Convolutional Neural Networks. In *CVPR*, 11166–11175.

Sun, B.; Ye, X.; Li, B.; Li, H.; Wang, Z.; and Xu, R. 2021. Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution. In *CVPR*, 7792–7801.

Tang, J.; Chen, X.; and Zeng, G. 2021. Joint Implicit Image Function for Guided Depth Super-Resolution. In *ACM MM*.

Tang, J.; Tian, F.; Feng, W.; Li, J.; and Tan, P. 2021. Learning Guided Convolutional Network for Depth Completion. *IEEE Trans. Image Process.*, 30: 1116–1129.

Tomasi, C.; and Manduchi, R. 1998. Bilateral Filtering for Gray and Color Images. In *ICCV*, 839–846.

Wang, H.; Chen, X.; Ni, B.; Liu, Y.; and Liu, J. 2023a. Omni Aggregation Networks for Lightweight Image Super-Resolution.

Wang, X.; Chen, X.; Ni, B.; Tong, Z.; and Wang, H. 2023b. Learning Continuous Depth Representation via Geometric Spatial Aggregator. In *AAAI*.

Wang, X.; Chen, X.; Ni, B.; Wang, H.; Tong, Z.; and Liu, Y. 2023c. Deep Arbitrary-Scale Image Super-Resolution via Scale-Equivariance Pursuit. In *CVPR*.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*.

Wu, H.; Zheng, S.; Zhang, J.; and Huang, K. 2018. Fast End-to-End Trainable Guided Filter. In *CVPR*, 1838–1847.

Xie, J.; Feris, R. S.; and Sun, M.-T. 2016. Edge-Guided Single Depth Image Super Resolution. *IEEE TIP*, 25(1): 428–438.

Yang, Q.; Yang, R.; Davis, J.; and Nistér, D. 2007. Spatial-Depth Super Resolution for Range Images. In *CVPR*.

Ye, X.; Sun, B.; Wang, Z.; Yang, J.; Xu, R.; Li, H.; and Li, B. 2020. PMBANet: Progressive Multi-Branch Aggregation Network for Scene Depth Super-Resolution. *IEEE TIP*, 29: 7427–7442.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zhao, S.; Gong, M.; Fu, H.; and Tao, D. 2021. Adaptive Context-Aware Multi-Modal Network for Depth Completion. *IEEE Trans. Image Process.*, 30: 5264–5276.

Zhao, Z.; Zhang, J.; Xu, S.; Lin, Z.; and Pfister, H. 2022. Discrete Cosine Transform Network for Guided Depth Map Super-Resolution. In *CVPR*.

Zhong, Z.; Liu, X.; Jiang, J.; Zhao, D.; Chen, Z.; and Ji, X. 2022. High-Resolution Depth Maps Imaging via Attention-Based Hierarchical Multi-Modal Fusion. *IEEE TIP*, 31: 648–663.