Structure Flow-Guided Network for Real Depth Super-resolution

Jiayi Yuan*, Haobo Jiang*, Xiang Li, Jianjun Qian, Jun Li†, Jian Yang†

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education
Jiangsu Key Lab of Image and Video Understanding for Social Security
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{jiayiyuan, jiang.hao.bo, xiang.li.implus, csjqian, junli, csjyang}@njust.edu.cn

Abstract

Real depth super-resolution (DSR), unlike synthetic settings, is a challenging task due to the structural distortion and the edge noise caused by the natural degradation in real-world low-resolution (LR) depth maps. These defeats result in significant structure inconsistency between the depth map and the RGB guidance, which potentially confuses the RGB-structure guidance and thereby degrades the DSR quality. In this paper, we propose a novel structure flow-guided DSR framework, where a cross-modality flow map is learned to guide the RGB-structure information transferring for precise depth upsampling. Specifically, our framework consists of a cross-modality flow-guided upsampling network (CFUNet) and a flow-enhanced pyramid edge attention network (PEANet). CFUNet contains a trilateral self-attention module combining both the geometric and semantic correlations for reliable cross-modality flow learning. Then, the learned flow maps are combined with the grid-sampling mechanism for coarse high-resolution (HR) depth prediction. PEANet targets at integrating the learned flow map as the edge attention into a pyramid network to hierarchically learn the edge-focused guidance feature for depth edge refinement. Extensive experiments on real and synthetic DSR datasets verify that our approach achieves excellent performance compared to state-of-the-art methods. Our code is available at: https://github.com/Yuanjiayii/DSR_SFG.

Introduction

With the fast development of cheap RGB-D sensors, depth maps have played a much more important role in a variety of computer vision applications, such as object recognition (Blum et al. 2012; Eitel et al. 2015), 3D reconstruction (Newcombe et al. 2011), and virtual reality (Meuleman et al. 2020)). However, the defects (e.g., low resolution and structural distortion) lying in the cheap RGB-D sensors (e.g., Microsoft Kinect and HuaweiP30Pro), still hinder their more extensive applications in the real world. Also, although the popular DSR methods (Song et al. 2020; Kim, Ponce, and Ham 2021; Sun et al. 2021) have achieved excellent DSR accuracy on synthetic LR depth maps, the significant domain gap between the real and synthetic data largely degrades their DSR precision on the real data.

This domain gap is mainly caused by different generation mechanisms of the LR depth map. The synthetic LR depth map is usually generated via artificial degradation (e.g., down-sampling operation), while the real one is from natural degradation (e.g., noise, blur, and distortion). Different from the synthetic data, there are two challenges of the real-data DSR as below. The first one is the severe structural distortion (see Fig. 1 (c)), especially for the low-reflection glass surface or the infrared-absorbing surface. The second one is the edge noise even the holes (see Fig. 1 (d)), caused by the physical limitations or the low processing power of the depth sensors. Both of the challenges above present a significant difference between the real and the synthetic data, which inherently degrades the generalization precision of the synthetic DSR methods to the real data.

In this paper, we develop a novel structure flow-guided DSR framework to handle the above challenges. For the structural distortion, we propose a cross-modality flow-guided upsampling network (CFUNet) that learns a structured flow between the depth map and the RGB image to guide their structure alignment for the recovery of the distorted depth structure. It includes two key components: a trilateral self-attention module and a cross-modality cross-attention module. In detail, the former leverages the geometric and semantic correlations (i.e., coordinate distance, pixel difference, and feature difference) to guide the relevant depth-feature aggregation into each depth feature to supplement the missing depth-structure information. The latter utilizes the enhanced depth feature and the RGB feature as the input for their sufficient message passing and flow-map generation. Finally, we combine the flow map with the grid-sampling mechanism for the coarse HR depth prediction.

For the edge noise, we present a flow-enhanced pyramid edge attention network (PEANet) that integrates the learned structure flow map as the edge attention into a pyramid network to learn the edge-focused guidance feature for the edge refinement of the coarse HR depth map predicted above. Considering the structure clue (i.e., edge region tends to own significant flow-value fluctuations) lying in the learned flow map, we combine the flow map with the RGB feature to form the flow-enhanced RGB feature for highlighting the RGB-structure region. Then, we feed the flow-enhanced

[*] These authors contributed equally.

[†] corresponding authors

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

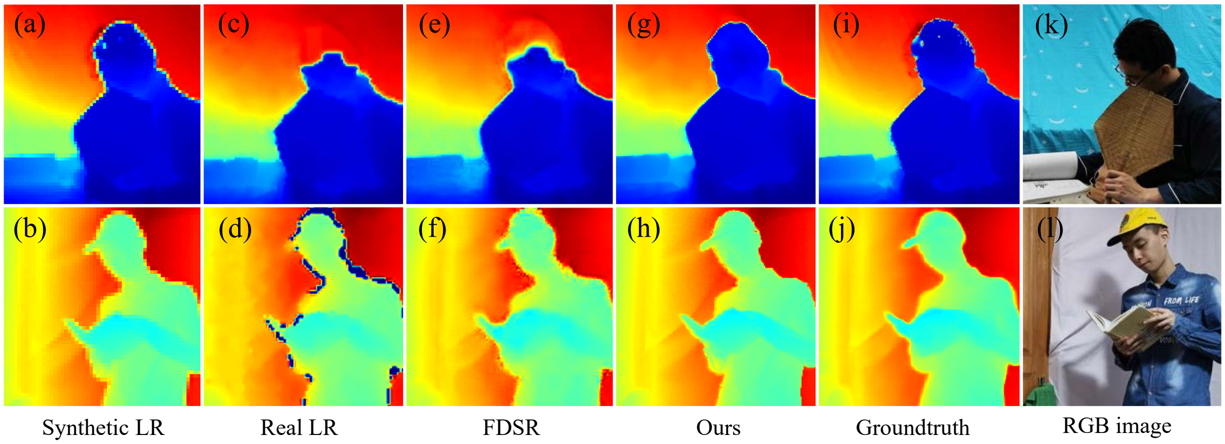


Figure 1: In this paper, we propose a novel structure flow-guided method for real-world DSR. Our method obtains better depth edge recovery (g-h), compared to (e) and (f) using the SOTA method, FDSR (He et al. 2021). (a-b) Synthetic LR depth maps; (c) Real LR depth map with the structural distortion; (d) Real LR depth map with the edge noise (e.g., holes); (i-j) Ground-truth HR depth maps; (k-l) RGB image guidance.

RGB feature into an iterative pyramid network for its edge-focused guidance feature learning. The low-level guidance features effectively filter the RGB-texture noise (guided by the flow map), while the high-level guidance features exploit the rich context information for more precise edge-feature capture. Finally, we pass the learned guidance feature and the depth feature into a decoder network to predict the edge-refined HR depth map. Extensive experiments on challenging real-world datasets verify the effectiveness of our proposed method (see examples in Fig. 1(g-h)). In summary, our contributions are as follows:

- We propose an effective cross-modality flow-guided upsampling network (CFUNet), where a structure flow map is learned to guide the structure alignment between the depth map and the RGB image for the recovery of the distorted depth edge.
- We present a flow-enhanced pyramid edge attention network (PEANet) that integrates the flow map as edge attention into a pyramid network to hierarchically learn the edge-focused guidance feature for edge refinement.
- Extensive experiments on the real and synthetic datasets verify the effectiveness of the proposed framework, and we achieve state-of-the-art restoration performance on multiple DSR dataset benchmarks.

Related Work

Synthetic Depth Super-resolution

The synthetic depth super-resolution (DSR) architectures can be divided into the pre-upsampling methods and the progressive upsampling methods (Wang, Chen, and Hoi 2020). The pre-upsampling DSR methods first upsample the input depth with interpolation algorithms (e.g., bicubic) from LR to HR, and then feed it into depth recovery network layers. (Li et al. 2016) introduced the first pre-upsampling network architecture. As this method handles arbitrary scaling factor depth, more and more similar approaches have been presented to further facilitate DSR task (Li et al. 2019; Lutio et al. 2019; Zhu et al. 2018; Chen and Jung 2018; Hao et al. 2019; Su et al. 2019). However, upsampling in one step is not suitable for large scaling factors simply because it usually leads to losing much detailed information. To tackle these issues, a progressive upsampling structure is designed in MSG-net(Tak-Wai, Loy, and Tang 2016), which gradually upsamples the LR depth map by transposed convolution layers at different scale levels. Since then, various progressive upsample-based methods have been proposed that greatly promote the development of this domain(Guo et al. 2019; He et al. 2021; Zuo et al. 2019). Recently, the joint-task learning framework achieves impressive performance, such as DSR & completion (Yan et al. 2022), depth estimation & enhancement (Wang et al. 2021) and DSR & depth estimation (Tang et al. 2021; Sun et al. 2021). Inspired by these joint-task methods, we combine the alignment task with the super-resolution task to distill the cross-modality knowledge for robust depth upsampling.

Real-World Depth Super-resolution

In recent years, the super-resolution for real-world images has been under the spotlight, which involves upsampling, denoising, and hole-filling. Early traditional depth enhancement methods (Yang et al. 2014; Liu et al. 2016, 2018) are based on complex and time-consuming optimization. For fast CNN-based DSR, AIR (Song et al. 2020) simulates the real LR depth map by combining the interval degradation and the bicubic degradation, and proposes a channel attention-based network for real DSR. PAC (Su et al. 2019) and DKN (Kim, Ponce, and Ham 2021) utilize the adaptive kernels calculated by the neighborhood pixels in RGB image for robust DSR. FDSR(He et al. 2021) proposes the octave convolution for frequency domain separation, which achieves outstanding performance in real datasets. Although

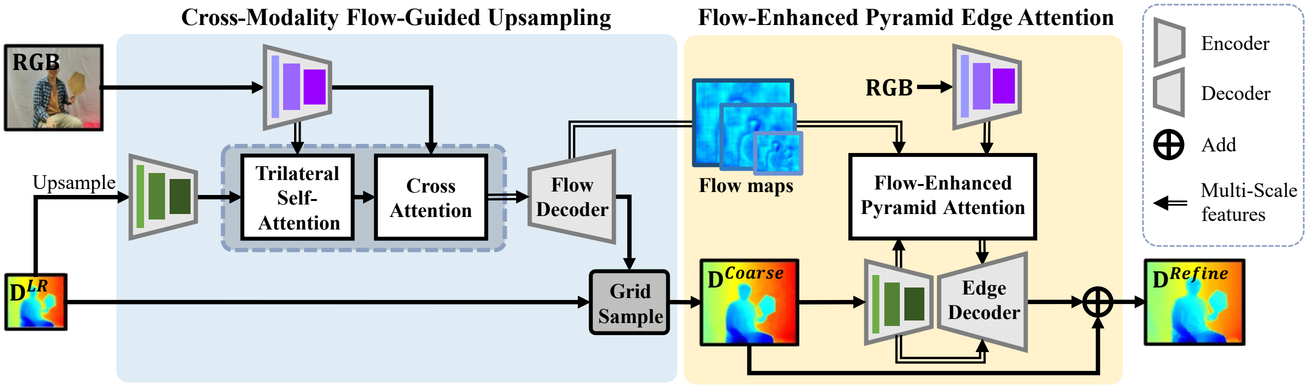


Figure 2: The pipeline of our structure flow-guided DSR framework. Given the LR depth map and the RGB image, the left block (blue) first generates the flow maps through a trilateral self-attention module and a cross-attention module, and predicts the coarse depth map $\mathbf{D}^{coarse}$ with the flow-based grid-sampling. Then, the right block (yellow) integrates the RGB/depth features and the flow map (as edge attention) to learn the edge-focused guidance feature for edge refinement ($\mathbf{D}^{refine}$).

these methods handle the large modality gap between the guidance image and depth map, the structure misalignment between the depth map and the RGB image still leads them to suffer from serious errors around the edge regions. Different from the general paradigms, we introduce a novel structure flow-guided framework, which exploits the cross-modality flow map to guide the RGB-structure information transferring for real DSR.

Approach

In the following, we introduce our structure flow-guided DSR framework for robust real-world DSR. As shown in Fig. 2, our framework consists of two modules: a cross-modality flow-guided upsampling network (CFUNet) and a flow-enhanced pyramid edge attention network (PEANet). Given an LR depth map $\mathbf{D}^{LR} \in \mathbb{R}^{H_0 \times W_0}$ and its corresponding HR RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ $(H/H_0 = W/W_0 = s$ and s is the scale factor), CFUNet first learns the cross-modality flow to guide the structure alignment between depth and RGB for coarse HR depth prediction. Then, PEANet exploits the structure flow as edge attention to learn the edge-focused guidance feature for edge refinement.

Cross-Modality Flow-Guided Upsampling Network

As demonstrated in Fig. 1 (c), the structural distortion of the real LR depth map leads to the significant structure misalignment between the RGB image and the depth map, which potentially damages the structure guidance of RGB images for depth edge recovery. To handle it, our solution is to learn an effective cross-modality flow map between the depth and the RGB to identify their structure relationship. Then, guided by the learned flow map, we align the structure of the depth map to the RGB image for the recovery of the distorted depth edge. Next, we will describe our network in terms of the feature extraction, the trilateral attention-based flow generation, and the flow-guided depth upsampling.

Feature Extraction. To achieve the consistent input size, we first upsample the LR depth map $\mathbf{D}^{LR}$ to a resolution map $\mathbf{D}^{Bic} \in \mathbb{R}^{H \times W}$ with the bicubic interpolation.

Then, we feed the upsampled depth map and the RGB image into an encoder for their feature extraction: $\{\mathbf{F}_l \in \mathbb{R}^{H \times W \times D}\}_{l=1}^{L}$ and $\{\mathbf{G}_l \in \mathbb{R}^{H \times W \times D}\}_{l=1}^{L}$ where the subscript l denotes the feature output in l-th layer of the encoder.

Trilateral Attention-Based Flow Generation. The key to generating a reliable cross-modality flow map is to model a robust relationship between the RGB and the depth map. Nevertheless, the serious structural distortion caused by the natural degradation potentially increases the modality gap between the depth and the RGB. Thereby, it's difficult to directly exploit a general attention mechanism to model such a relationship. To mitigate it, we target at enhancing the depth feature through a proposed trilateral self-attention block so that the distorted depth-structure information can be largely complemented for relationship modeling. As shown in Fig. 3, our trilateral self-attention block fuses the geometric-level correlation and the semantic-level correlation to jointly guide the depth feature enhancement. It's noted that we just enhance the depth feature $\mathbf{F}_L$ in the last layer (L-th layer):

$$\bar{\mathbf{F}}_L^{(i)} = \sum_j \boldsymbol{\alpha}_{i,j} \cdot (\boldsymbol{\beta}_{i,j}^{low} + \boldsymbol{\beta}_{i,j}^{high}) \cdot \boldsymbol{\gamma}_{i,j} \cdot \mathbf{F}_L^{(j)} + \mathbf{F}_L^{(j)}, \qquad (1)$$

where $\mathbf{F}_L^{(j)}$ $(1 \leq j \leq H \times W)$ denotes the j-th depth-pixel feature and $\bar{\mathbf{F}}_L^{(i)}$ denotes the i-th enhanced depth feature $(1 \leq i \leq H \times W)$. The geometric-level correlation contains a spatial kernel $\boldsymbol{\alpha} \in \mathbb{R}^{(H \times W) \times (H \times W)}$ and a low-level color kernel $\boldsymbol{\beta}^{low} \in \mathbb{R}^{(H \times W) \times (H \times W)}$, while the semantic-level correlation contains a high-level color semantic kernel $\boldsymbol{\beta}^{high} \in \mathbb{R}^{(H \times W) \times (H \times W)}$ and a depth semantic kernel $\boldsymbol{\gamma} \in \mathbb{R}^{(H \times W) \times (H \times W)}$. In detail, we formulate the spatial kernel as a coordinate distance-aware Gaussian kernel:

$$\boldsymbol{\alpha}_{i,j} = \text{Gaussian}(\| \text{Coor}(i) - \text{Coor}(j) \|_2, \sigma_s), \qquad (2)$$

where $\text{Gaussian}(x, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ is the Gaussian function. $\text{Coor}(i) \in \mathbb{R}^2$ denotes the row-column coordi-

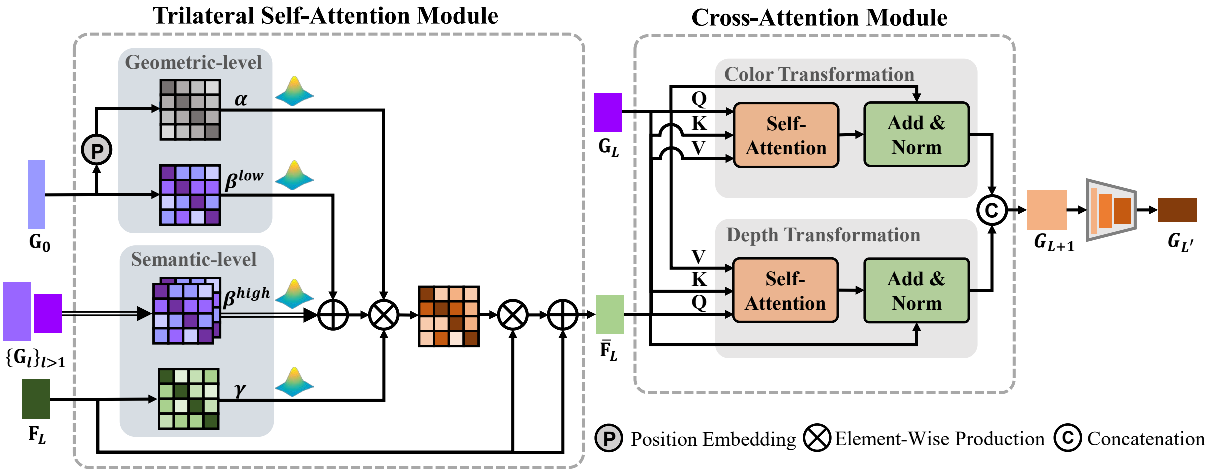


Figure 3: The architecture of the trilateral self-attention module and the cross-attention module.

nates of pixel i at the depth map and σ_s is the kernel variance. The low-level and high-level color kernels are defined by the Gaussian kernels with the low-level and the semantic-level RGB feature similarity, whose kernel sum is:

$$\beta_{i,j}^{low} + \beta_{i,j}^{high} = \sum_{l=0}^{L} \text{Gaussian}(\|\mathbf{G}_l^{(i)} - \mathbf{G}_l^{(j)}\|_2, \sigma_c). \quad (3)$$

The depth semantic kernel is designed based on the depth feature similarity in the L-th layer:

$$\gamma_{i,j} = \text{Gaussian}(\|\mathbf{F}_L^{(i)} - \mathbf{F}_L^{(j)}\|_2, \sigma_d). \quad (4)$$

Guided by the geometric and semantic kernels above, the correlated depth information can be effectively aggregated into each depth feature through Eq.1 for depth feature completion and enhancement.

Then, we feed the enhanced depth feature $\bar{\mathbf{F}}_L$ and the RGB feature $\mathbf{G}_L$ into the cross-attention block for their efficient cross-modality feature intersection:

$$\tilde{\mathbf{F}}_L^{(i)} = \bar{\mathbf{F}}_L^{(i)} + \text{MLP}(\sum_j \text{SM}_j(\phi_q(\bar{\mathbf{F}}_L^{(i)})^\top \phi_k(\mathbf{G}_L^{(j)}))\phi_v(\mathbf{G}_L^{(j)})),$$
$$\tilde{\mathbf{G}}_L^{(i)} = \mathbf{G}_L^{(i)} + \text{MLP}(\sum_j \text{SM}_j(\phi_q(\mathbf{G}_L^{(i)})^\top \phi_k(\bar{\mathbf{F}}_L^{(j)})\phi_v(\bar{\mathbf{F}}_L^{(j)})),$$
$$(5)$$

where SM indicates the softmax function. ϕ_q, ϕ_k, and ϕ_v are the projection functions of the query, the key, and the value in our nonlocal-style cross-attention module. With the query-key similarity, the value can be retrieved for feature enhancement. Then, we concatenate the enhanced depth feature $\tilde{\mathbf{F}}_L$ and RGB feature $\tilde{\mathbf{G}}_L$ and pass them into a multi-layer convolutional network to obtain their correlated feature at each layer $\{\mathbf{G}_l\}_{l=L+1}^{L'}$. Finally, following (Dosovitskiy et al. 2015), based on the previously extracted features $\{\mathbf{G}_l\}_{l=1}^{L}$ and the correlated features $\{\mathbf{G}_l\}_{l=L+1}^{L'}$, we exploit a decoder network to generate the multi-layer flow maps $\{\boldsymbol{\Delta}_l\}_{l=1}^{L'}$, where the flow generation in layer l can be formulated as:

$$\mathbf{G}_l^{flow}, \boldsymbol{\Delta}_l = \text{deconv}(\text{Cat}[\mathbf{G}_{l-1}^{flow}, \boldsymbol{\Delta}_{l-1}, \mathbf{G}_{L'-l+1}]), \quad (6)$$

where $\mathbf{G}_l^{flow}$ denotes the intermediate flow feature and deconv consisting of a deconvolution operation and a convolutional block ($\mathbf{G}_1^{flow}, \boldsymbol{\Delta}_1 = \text{deconv}(\mathbf{G}_{L'})$).

Flow-Guided Depth Upsampling. With the learned flow map $\boldsymbol{\Delta}_{L'}$ in the last layer, we combine it with the grid-sampling strategy for the HR depth map prediction. In detail, the value of the HR depth map is the bilinear interpolation of the neighborhood pixels in LR depth map $\mathbf{D}^{LR}$, where the neighborhoods are defined according to the learned flow field, which can be formulated as:

$$\mathbf{D}^{coarse} = \text{Grid-Sample}(\mathbf{D}^{LR}, \boldsymbol{\Delta}_{L'}), \quad (7)$$

where Grid-Sample denotes the upsampling operation computing the output using pixel values from neighborhood pixels and pixel locations from the grid (Li et al. 2020).

Flow-Enhanced Pyramid Edge Attention Network

In order to further improve our DSR precision in the case of the edge noise problem, we propose a flow-enhanced pyramid network, where the learned structure flow is served as the edge attention to hierarchically mine edge-focused guidance feature from the RGB image for the edge refinement of $\mathbf{D}^{coarse}$. Specifically, we first feed the previously predicted HR depth map $\mathbf{D}^{coarse}$ and the RGB image into an encoder network to extract their features: $\{\mathbf{F}_t^{coarse}\}_{t=1}^{T}$ and $\{\mathbf{G}_t\}_{t=1}^{T}$, where subscript t indicates the extracted feature at the t-th layer. Then, we propose the flow-enhanced pyramid attention module and the edge decoder module as follows for refined HR depth prediction.

Flow-Enhanced Pyramid Attention Module. In this module, we target at combining the RGB feature and the flow map to learn the edge-focused guidance feature $\{\mathbf{G}_t^{guide}\}$ at each layer. In detail, for the t-th layer, with the RGB feature $\mathbf{G}_t$ and its corresponding flow map $\boldsymbol{\Delta}_{L'-t}$, we first fuse the flow information into the RGB feature to form the flow-enhanced RGB feature,

$$\mathbf{G}_t^{flow} = \boldsymbol{\Delta}_{L'-t} \cdot \mathbf{G}_t + \mathbf{G}_t, \quad (8)$$

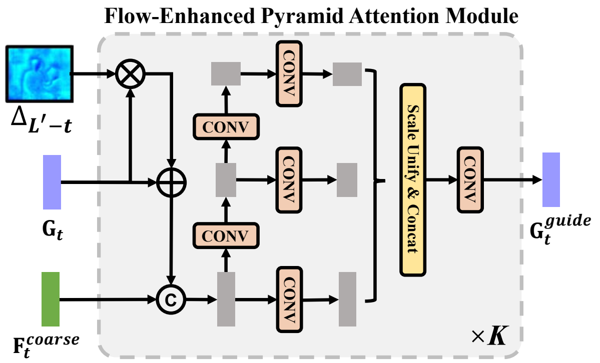


Figure 4: The architecture of the pyramid attention module. The subscript t denotes the feature output in the t-th layer of the encoder ($1 \leq t \leq T$). '$\times K$' indicates the iteration times of the guidance feature updating.

where $\mathbf{\Delta}_{L'-t} \cdot \mathbf{G}_t$ is expected to exploit the significant flow-value fluctuations at the edge region in $\mathbf{\Delta}_{L'-t}$ to better highlight the structure region of the RGB feature. To further smooth the texture feature in $\mathbf{G}_t^{flow}$, we concatenate it with the texture-less depth feature $\mathbf{F}_t^{coarse}$ to obtain the texture-degraded RGB feature $\tilde{\mathbf{G}}_t^{flow}$. Then, we feed $\tilde{\mathbf{G}}_t^{flow}$ into a pyramid network to extract its edge-focused guidance features $\{\tilde{\mathbf{G}}_{t,k}^{flow}\}_{k=1}^{K}$ at different scales. The low-level guidance feature is to filter the texture noise (guided by the flow map), while the high-level feature is to exploit the rich context information for edge-feature capture. After that, we unify the scales of the hierarchical feature $\{\tilde{\mathbf{G}}_{t,k}^{flow}\}_{k=1}^{K}$ using the bicubic interpolation and pass the concatenated feature into a convolutional block to generate the flow-enhanced RGB guidance feature $\mathbf{G}_t^{guide}$ at the t-th layer. Notably, we design an iterative architecture to progressively refine the RGB guidance feature as illustrated in Fig. 4.

Edge Decoder. Guided by the flow-based guidance features $\{\mathbf{G}_t^{guide}\}_{t=1}^{T}$ learned at each layer, we progressively decode the depth feature in an iterative manner:

$$\mathbf{F}_{t+1}^{edge} = \mathrm{FU}(\mathrm{Cat}(\mathbf{F}_t^{edge}, \mathbf{G}_{T-t+1}^{guide}, \mathbf{F}_{T-t+1}^{coarse})), \tag{9}$$

where FU function indicates the fusion and upsampling operation following (Guo et al. 2020) and the initial feature $\mathbf{F}_1^{edge}$ is obtained by the convolutional operation on $\mathbf{F}_T^{coarse}$. Finally, we pass $\mathbf{F}_{T+1}^{edge}$ into a convolutional block to obtain the edge-refined HR depth map $\mathbf{D}^{refine}$.

Loss Function

We train our model by minimizing the smooth-L1 loss between the ground-truth depth map $\mathbf{D}^{gt}$ and the network output of each sub-network, including the coarse depth prediction $\mathbf{D}^{coarse}$ and the refined one $\mathbf{D}^{refine}$:

$$\mathcal{L}_{dsr} = \sum_{i=1}^{H \times W} \ell\left(\mathbf{D}_i^{coarse} - \mathbf{D}_i^{gt}\right) + \ell\left(\mathbf{D}_i^{refine} - \mathbf{D}_i^{gt}\right), \tag{10}$$

where the subscript i denote the pixel index and the smooth-L1 loss function is defined as:

$$\ell(u) = \begin{cases} 0.5u^2, & \text{if } |u| \leq 1 \\ (|u| - 0.5)\,, & \text{otherwise.} \end{cases} \tag{11}$$

Experiments

Experimental Setting

To evaluate the performance of our method, we perform extensive experiments on real-world RGB-D-D dataset (He et al. 2021), ToFMark dataset (Ferstl et al. 2013) and synthetic NYU-v2 dataset (Silberman et al. 2012). We implement our model with PyTorch and conduct all experiments on a server containing an Intel i5 2.2 GHz CPU and a TITAN RTX GPU with almost 24 GB. During training, we randomly crop patches of resolution 256×256 as groundtruth and the training and testing data are normalized to the range $[0, 1]$. In order to balance the training time and network performance, the parameters L, L', K, T are set to 3, 6, 3, 2 in this paper. We quantitatively and visually compare our method with 13 state-of-the-art (SOTA) methods: TGV (Ferstl et al. 2013), FBS (Barron and Poole 2016), MSG (Tak-Wai, Loy, and Tang 2016), DJF (Li et al. 2016), DJFR (Li et al. 2019), GbFT (AlBahar and Huang 2019), PAC (Su et al. 2019), CUNet (Deng and Dragotti 2020), FDKN (Kim, Ponce, and Ham 2021), DKN (Kim, Ponce, and Ham 2021), FDSR (He et al. 2021), CTKT (Sun et al. 2021) and DCTNet (Zhao et al. 2022). For simplicity, we name our **S**tructure **F**low-**G**uided method as **SFG**.

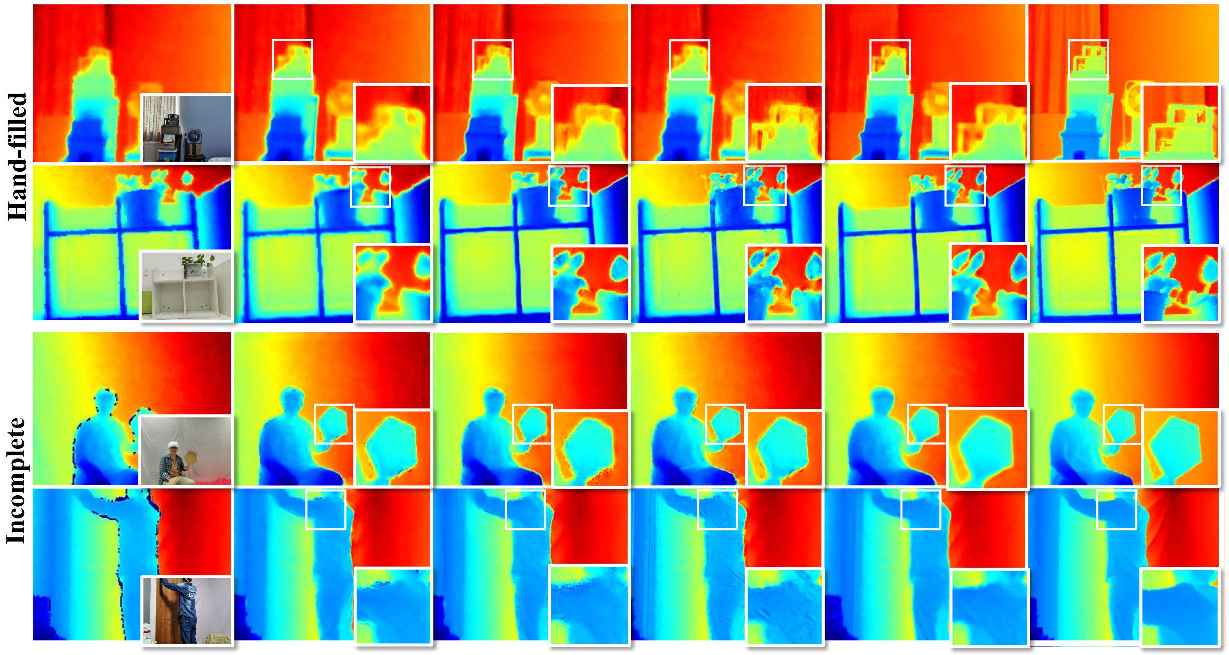
Experiments on Real Datasets

Depth maps captured by cheap depth sensors usually suffer from structural distortion and edge noise. To verify the efficiency and robustness of our proposed method, we employ our method on two challenging benchmarks: RGB-D-D dataset and ToFMark dataset.

Evaluation on Hand-Filled RGB-D-D. To evaluate the performance of our method on real LR depth maps, we conduct experiments on RGB-D-D datasets captured by two RGB-D sensors: Huawei P30 Pro (captures RGB images and LR depth maps) and Helios TOF camera (captures HR depth maps). The LR inputs are shown in Fig. 5, which suffer from the low resolution (LR size is 192×144 and target size is 512×384) and random structural missing in the edge region. Following FDSR (He et al. 2021), we first use 2215 hand-filled RGB/D pairs for training and 405 RGB/D pairs for testing. As listed in the first row of Table 1, the proposed model outperforms SOTA methods by a significant margin.

The first two rows in Fig. 5 show the visual DSR comparisons on hand-filled RGB-D-D dataset. We can see that edges in the results of DKN (Kim, Ponce, and Ham 2021) and DCTNet (Zhao et al. 2022) are over-smoothed and the artifacts are visible in the FDSR results. In contrast, our results show more accurate structures without texture copying.

Evaluation on Incomplete RGB-D-D. To further verify the DSR performance of our method in the case of edge noise (e.g., edge holes), instead of the hole completion

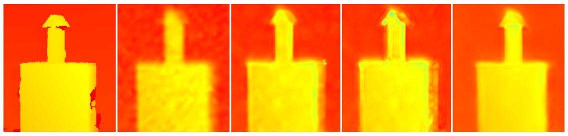


(a) LR depth (b) DKN (c) FDSR (d) DCTNet (e) SFG (ours) (f) Groundtruth

Figure 5: Visual comparison on RGB-D-D dataset. The first (last) two rows show DSR results of hand-filled (incomplete) LR.

RMSE	Bicubic	MSG	DJF	DJFR	CUNet	DKN	FDKN	FDSR	DCTNet	SFG (ours)
Hand-filled	7.17	5.50	5.54	5.52	5.84	5.08	5.37	5.34	5.28	**3.88**
Incomplete	-	7.90	5.70	5.52	6.54	5.43	5.87	5.59	5.49	**4.79**
Noisy	11.57	10.36	5.62	5.71	6.13	5.16	5.54	5.63	5.16	**4.45**

Table 1: Quantitative comparison on RGB-D-D dataset.



(a) Groundtruth (b) Bicubic (c) DKN (d) DCTNet (e) SFG (ours)

Figure 6: Visual comparison on ToFMark dataset.

	DJFR	DKN	FDKN	FDSR	DCTNet	SFG (ours)
RMSE	0.27	0.26	0.28	0.28	0.27	**0.25**

Table 2: Quantitative comparison on ToFMark dataset.

above, we directly test SFG on unfilled RGB-D dataset and achieve the lowest RMSE as shown in the second row of Table 1. Moreover, as shown in the last two rows in Fig. 5, the edges recovered by our method are sharper with fewer artifacts and visually closest to the ground-truth map. It's mainly attributed to the edge-focused guidance learning with our flow-enhanced pyramid edge attention network.

Evaluation on Noisy RGB-D-D and ToFMark. We evaluate the denoising and generalization ability of our method on ToFMark dataset consisting of three RGB-D pairs. The LR inputs have irregular noise and limited resolution (LR depth is 120 × 160 and target size is 610 × 810). To simulate the similar degradation for training, we add the Gaussian noise (mean 0 and standard deviation 0.07) and the Gaussian blur (kernel size 5) on the 2215 RGB-D pairs from RGB-D-D dataset to generate the noisy training dataset. Testing dataset consists of 405 RGB-D pairs from noisy RGB-D-D dataset and 3 RGB-D pairs from ToFMark dataset. As shown in the last row of Table 1 and Table 2, our method achieves the lowest RMSE in noisy RGB-D-D dataset and the lowest RMSE in ToFMark dataset, which proves its ability for noise removing. As shown in Fig. 6, it is observed that DKN (Kim, Ponce, and Ham 2021) and DCTNet (Zhao et al. 2022) introduce some texture artifacts and noise in the low-frequency region, while SFG recovers clean surface owing to the edge attention network with effective texture removing.

Experiments on Synthetic Datasets

Since most popular methods are designed for synthetic datasets, we further evaluate our method on NYU-v2 datasets for a more comprehensive comparison. Following

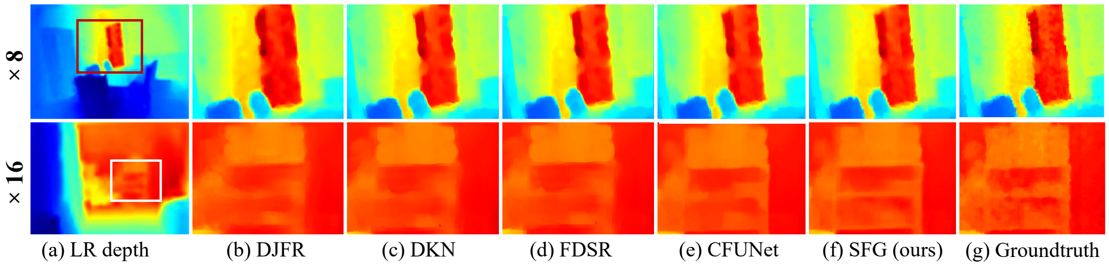


Figure 7: Visual comparison of ×8 and ×16 DSR results on NYU-v2 dataset.

RMSE	TGV	FBS	DJFR	GbFT	PAC	CUNet	FDKN	DKN	FDSR	DCTNet	CTKT	SFG (ours)
×4	4.98	4.29	2.38	3.35	2.39	1.89	1.86	1.62	1.61	1.59	<u>1.49</u>	**1.45**
×8	11.23	8.94	4.94	5.73	4.59	3.58	3.33	3.26	3.18	3.16	**2.73**	<u>2.84</u>
×16	28.13	14.59	9.18	9.01	8.09	6.96	6.78	6.51	5.86	5.84	**5.11**	<u>5.56</u>

Table 3: Quantitative comparison on NYU-v2 dataset in terms of average RMSE (cm).

Model	RMSE
CFUNet	4.22
CFUNet *w/o* TriSA	4.34
CFUNet *w/o* cross-attention	4.57

Table 4: Ablation study of CFUNet on RGB-D-D dataset.

Datasets	SFG	SFG *w/o* PEANet
RGB-D-D	3.88	4.22
NYU-v2 (×4)	1.45	1.82
NYU-v2 (×8)	2.84	3.76
NYU-v2 (×16)	5.55	5.90

Table 5: Ablation study (in RMSE) of PEANet.

the widely used data splitting criterion, we sample 1000 RGB-D pairs for training and the rest 449 RGB-D pairs for testing. As shown in Table 3, the proposed method still achieves comparable results with the SOTA methods on all upsampling cases (×4, ×8, ×16). In addition, Fig. 7 presents that our ×8 and ×16 upsampled depth maps own higher accuracy and more convincing results. It verifies that our method not only performs DSR well in low-quality maps with noise and missing structure, but also achieves high-quality precision in the case of large-scale upsampling.

Ablation Analysis

Ablation Study on CFUNet. As shown in the first row of Table 4, we still achieve the lowest RMSE criterion just with the single CFUNet (SFG w/o PEANet) on RGB-D-D dataset when compared with SOTA methods. It proves the effectiveness of the learned structure flow map for real DSR. Table 4 also shows that removing the trilateral self-attention (TriSA) and cross-attention modules in CFUNet causes performance degradation on RGB-D-D datasets, which verifies the necessity of the depth enhancement for reliable flow map.

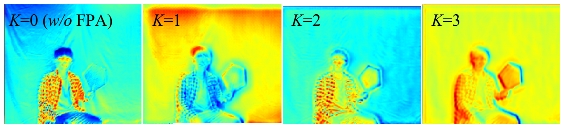


Figure 8: Visual comparison of guidance features using FPA with different iteration times *K*, *i.e.*, from 0 (*w/o* FPA) to 3.

Ablation Study on PEANet. To analyze the effectiveness of PEANet, we train the network with and without PEANet on the synthetic dataset (NYU-v2) and the real-world dataset (RGB-D-D). As shown in Table 5, PEANet consistently brings the RMSE gain under both real and synthetic dataset settings. It's mainly due to our edge-focused guidance feature learning for robust edge refinement. In addition, Fig. 8 shows the guidance features under varying iteration times in FPA (**F**low-Enhanced **P**yramid **A**ttention) module from 0 (*w/o* FPA) to 3. Visually, as the number of iterations increases, the edge regions tend to receive more attention.

Conclusion

In this paper, we proposed a novel structure flow-guided DSR framework for real-world depth super-resolution, which deals with issues of structural distortion and edge noise. For the structural distortion, a cross-modality flow-guided upsampling network was presented to learn a reliable cross-modality flow between depth and the corresponding RGB guidance for the reconstruction of the distorted depth edge, where a trilateral self-attention combines the geometric and semantic correlations for structure flow learning. For the edge noise, a flow-enhanced pyramid edge attention network was introduced to produce edge attention based on the learned flow map and learn the edge-focused guidance feature for depth edge refinement with a pyramid network. Extensive experiments on both real-world and synthetic datasets demonstrated the superiority of our method.

Acknowledgements

This work was supported by the National Science Fund of China under Grant Nos. U1713208 and 62072242.

References

AlBahar, B.; and Huang, J.-B. 2019. Guided image-to-image translation with bi-directional feature transformation. In *ICCV*, 9016–9025.

Barron, J. T.; and Poole, B. 2016. The fast bilateral solver. In *ECCV*, 617–632.

Blum, M.; Springenberg, J. T.; Wülfing, J.; and Riedmiller, M. 2012. A learned feature descriptor for object recognition in RGB-D data. In *ICRA*, 1298–1303.

Chen, B.; and Jung, C. 2018. Single depth image super-resolution using convolutional neural networks. In *ICASSP*, 1473–1477.

Deng, X.; and Dragotti, P. L. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, PP(99): 1–1.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow With Convolutional Networks. In *ICCV*, 2758–2766.

Eitel, A.; Springenberg, J. T.; Spinello, L.; Riedmiller, M.; and Burgard, W. 2015. Multimodal deep learning for robust RGB-D object recognition. In *IROS*, 681–687.

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.

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 Transactions on Image Processing*, 2545–2557.

Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; and Tan, M. 2020. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, 5407–5416.

Hao, X.; Lu, T.; Zhang, Y.; Wang, Z.; and Chen, H. 2019. Multi-Source Deep Residual Fusion Network for Depth Image Super-resolution. In *RSVT*, 62–67.

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 Baseline and. In *CVPR*, 9229–9238.

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

Li, X.; You, A.; Zhu, Z.; Zhao, H.; Yang, M.; Yang, K.; Tan, S.; and Tong, Y. 2020. Semantic flow for fast and accurate scene parsing. In *ECCV*, 775–793. Springer.

Li, Y.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2016. Deep joint image filtering. In *ECCV*, 154–169.

Li, Y.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2019. Joint image filtering with deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1909–1923.

Liu, W.; Chen, X.; Yang, J.; and Wu, Q. 2016. Robust color guided depth map restoration. *IEEE Transactions on Image Processing*, 26(1): 315–327.

Liu, X.; Zhai, D.; Chen, R.; Ji, X.; Zhao, D.; and Gao, W. 2018. Depth restoration from RGB-D data via joint adaptive regularization and thresholding on manifolds. *IEEE Transactions on Image Processing*, 28(3): 1068–1079.

Lutio, R. d.; D'aronco, S.; Wegner, J. D.; and Schindler, K. 2019. Guided super-resolution as pixel-to-pixel transformation. In *ICCV*, 8829–8837.

Meuleman, A.; Baek, S.-H.; Heide, F.; and Kim, M. H. 2020. Single-shot monocular rgb-d imaging using uneven double refraction. In *CVPR*, 2465–2474.

Newcombe, R. A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A. J.; Kohi, P.; Shotton, J.; Hodges, S.; and Fitzgibbon, A. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 127–136. Ieee.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*, 746–760.

Song, X.; Dai, Y.; Zhou, D.; Liu, L.; Li, W.; Li, H.; and Yang, R. 2020. Channel attention based iterative residual learning for depth map super-resolution. In *CVPR*, 5631–5640.

Su, H.; Jampani, V.; Sun, D.; Gallo, O.; Learned-Miller, E.; 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.

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

Tang, Q.; Cong, R.; Sheng, R.; He, L.; Zhang, D.; Zhao, Y.; and Kwong, S. 2021. BridgeNet: A Joint Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation. In *ACMMM*, 2148–2157.

Wang, K.; Zhang, Z.; Yan, Z.; Li, X.; Xu, B.; Li, J.; and Yang, J. 2021. Regularizing Nighttime Weirdness: Efficient Self-Supervised Monocular Depth Estimation in the Dark. In *ICCV*, 16055–16064.

Wang, Z.; Chen, J.; and Hoi, S. C. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3365–3387.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, G.; Li, J.; and Yang, J. 2022. Learning Complementary Correlations for Depth Super-Resolution With Incomplete Data in Real World. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, J.; Ye, X.; Li, K.; Hou, C.; and Wang, Y. 2014. Color-guided depth recovery from RGB-D data using an adaptive autoregressive model. *IEEE transactions on image processing*, 23(8): 3443–3458.

Zhao, Z.; Zhang, J.; Xu, S.; Lin, Z.; and Pfister, H. 2022. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, 5697–5707.

Zhu, J.; Zhai, W.; Cao, Y.; and Zha, Z.-J. 2018. Co-occurrent structural edge detection for color-guided depth map super-resolution. In *MMM*, 93–105.

Zuo, Y.; Wu, Q.; Fang, Y.; An, P.; Huang, L.; and Chen, Z. 2019. Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2): 297–306.