Learning Hierarchical Color Guidance for Depth Map Super-Resolution

Runmin Cong, *Senior Member, IEEE,* Ronghui Sheng, Hao Wu, Yulan Guo, Yunchao Wei, Wangmeng Zuo, Yao Zhao, *Fellow, IEEE,* and Sam Kwong, *Fellow, IEEE*

***Abstract*—Color information is the most commonly used prior knowledge for depth map super-resolution (DSR), which can provide high-frequency boundary guidance for detail restoration. However, its role and functionality in DSR have not been fully developed. In this paper, we rethink the utilization of color information and propose a hierarchical color guidance network to achieve DSR. On the one hand, the low-level detail embedding module is designed to supplement high-frequency color information of depth features in a residual mask manner at the low-level stages. On the other hand, the high-level abstract guidance module is proposed to maintain semantic consistency in the reconstruction process by using a semantic mask that encodes the global guidance information. The color information of these two dimensions plays a role in the front and back ends of the attention-based feature projection (AFP) module in a more comprehensive form. Simultaneously, the AFP module integrates the multi-scale content enhancement block and adaptive attention projection block to make full use of multi-scale information and adaptively project critical restoration information in an attention manner for DSR. Compared with the state-of-the-art methods on four benchmark datasets, our method achieves more competitive performance both qualitatively and quantitatively. The code and results can be found from the link of https://rmcong.github.io/HCGNet_TIM2024.**

***Index Terms*—Depth map, Super-resolution, Hierarchical color guidance, Residual mask, Semantic mask, Adaptive projection.**

I. Introduction

DEPTH maps describe the distance relationship of the scene including the occlusion and overlap of objects, which is essential for the 3D understanding tasks, such as autonomous driving [1], 3D reconstruction [2], [3], object recognition [4], [5], and salient object detection [6]–[15], *etc*. However, due to the limitations of existing depth acquisition devices, the resolution of the acquired depth maps is relatively low, especially for the low-power depth sensors equipped on smartphones. The low-resolution (LR) depth map cannot match the high-resolution (HR) color image in resolution, thereby hindering the further expansion of depth-oriented applications. Therefore, the super-resolution reconstruction technology for depth maps came into being, which has practical research value and industrial application value.

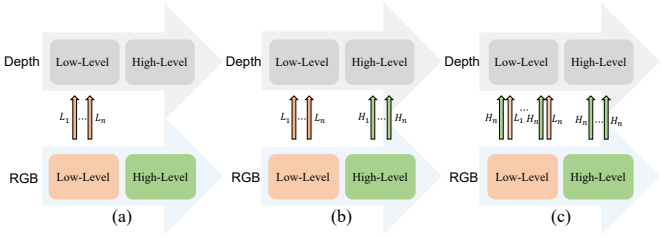


Fig. 1: Illustration of the color guidance in DSR. Mode (a) only utilizes the low-level color information to guide the reconstruction of detail information; Mode (b) treats different levels of color information indiscriminately; Mode (c) represents our guidance model, which divides color information into two parts, *i.e*., low-level and high-level information, and allows them to play different roles.

Depth map super-resolution (DSR) is a challenging task that aims to reconstruct the LR depth map into an HR depth map. This task is inherently an ill-posed inverse problem due to the absence of a unique mapping between LR and HR depth maps. Furthermore, it is particularly difficult to recover fine details, such as sharp boundaries, especially when dealing with large upsampling factors [16]–[18]. Because of the structural similarity with depth maps and readily accessible, HR color images can naturally provide comprehensive guidance information for the DSR task, and numerous color-guided DSR approaches have been proposed. However, what kind of color information is to be utilized for guidance and how the implementation is to be conducted still remain open topics in color-guided DSR. For example, current DSR techniques utilize the color boundary information, either explicitly or implicitly, to enhance the reconstruction of details [19]–[22]. But such structural congruence is not universally applicable. The RGB image contains not only the object boundary but also the texture boundary inside the object, while the depth map only has the object boundary. In other words, as for the color-guided DSR, a critical issue is the effective exploitation of color guidance information to enhance depth details while

Runmin Cong is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, also with the School of Control Science and Engineering, Shandong University, Jinan 250061, China, and also with the Key Laboratory of Machine Intelligence and System Control, Ministry of Education, Jinan 250061, Shandong, China (e-mail: rmcong@sdu.edu.cn).

Ronghui Sheng, Yunchao Wei, and Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: ronghuisheng@bjtu.edu.cn; wychao1987@gmail.com; yzhao@bjtu.edu.cn).

Hao wu is with the Artificial Intelligence Institute, Beijing Normal University, Beijing 100875, China (email: wuhao@bnu.edu.cn).

Yulan Guo is with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha 410073, China, and also with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail:guoyulan@sysu.edu.cn).

Wangmeng Zuo is with School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China (e-mail: wmzuo@hit.edu.cn).

Sam Kwong is with the Lingnan University, Hong Kong SAR, China (e-mail: samkwong@ln.edu.hk).

mitigating the texture-copying artifacts introduced by the color image. Furthermore, in the form of guidance, some modes and strategies are designed, such as using low-level color features as detailed guidance, usually directly concatenating color features with depth features [23], [24] (as shown in Fig. 1(a)), or treating different levels of color features equally as guidance [25], [26] (as shown in Fig. 1(b)), *etc*. However, these methods do not fully consider the roles and diversity of different color information in the guidance phase, indicating a necessity for more in-depth and comprehensive exploration to leverage the full spectrum of color guidance information effectively.

Motivated by the above analysis, the core theoretical contribution of our work lies in rethinking the utilization of color information in the DSR task and distinguishing the roles of low-level and high-level color information, thereby making them guide the depth branch in a divide-and-conquer manner, as shown in Fig. 1(c). On the one hand, the low-level color features contain fine-grained detailed information (*e.g*., boundaries) [24], [27] that DSR needs to pay attention to, which is helpful for the detail recovery of the depth map. However, the representation of these features is too specific, with a lot of interference. And simply transferring color features may introduce unnecessary interfering boundaries, resulting in texture replication. To address this, we learn a residual mask in the designed low-level detail embedding (LDE) module to highlight the spatial locations of color features that are most consistent with depth features, thereby adaptively guiding the information transmission from color features to depth features. On the other hand, the high-level color features contain global abstract information, which describes scene content more comprehensively and preserves semantic outlines. The existing approaches do not specifically consider high-level color abstract information, but discard it [28] or treat it the same as low-level detail information [25]. Considering that the semantic consistency of the scene may be shifted or blurred during the depth reconstruction, we design a high-level abstract guidance (HAG) module to modify the initial reconstruction features by using a semantic mask that encodes the global abstract guidance information. It is worth mentioning that the LDE and HAG modules we designed also have good portability, which can be transplanted into existing color-guided DSR methods to improve their performance (see the validation experiments in Section IV-C).

In addition, to achieve better recovery, we need to map the low-resolution features to the desired high-resolution reconstruction features. The existing methods, such as DBPN [25], directly map features between LR and HR domains instead of selecting the reconstruction region, which greatly increases the complexity of the model and introduces additional errors. In fact, the focus of the DSR task is not to generate content from scratch but to supplement, refine, and enhance the details such as boundaries. From this point of view, blindly and indiscriminately performing super-resolution reconstruction on all regions is a sub-optimal way, which also difficult to achieve the purpose of optimizing important regions with more severe degradation. To this end, we design the attention-based feature projection (AFP) module, including a multi-scale content enhancement (MCE) block and multiple adaptive attention projection (AAP) blocks. The core contribution of the AFP module lies in the designed AAP block, which reinforces the important restoration regions in an attention manner, thereby suppressing interference and improving the reconstruction performance. The whole reconstruction process is a restoration pipeline from coarse to fine, focusing on using different levels of color information for guided reconstruction. All modules cooperate with each other to hierarchically reconstruct the depth features, thereby obtaining the final depth map at the target resolution.

To summarize, the contributions of this work are as follows:

- We reexplore the role of the color information in DSR and propose a hierarchical color guidance network (HCGNet). Comprehensive experiments on four benchmark datasets show that the proposed method achieves more competitive performance both qualitatively and quantitatively.
- The LDE and HAG modules work together to achieve hierarchical color guidance in DSR task. Concretely, the LDE module distinguishes between similar as well as interfering regions in the form of residual masks, thereby effectively utilizing high-frequency complementary guidance of color features. And the HAG module extracts complete semantic outlines from high-level abstract features in the form of semantic masks, thereby alleviating semantic shifts and ambiguities for global reconstruction.
- We design an AAP block to reinforce the key restoration regions in the attention domain, thereby suppressing the valueless redundancy and improving the reconstruction performance with optimized computation.

II. Related Work

A. Non Color-Guided DSR

Non color-guided DSR directly reconstructs a HR depth map from a LR depth map without any external guidance information. Earlier works proposed some local filtering-based methods, which mostly use high-pass filters to recover the boundary information of the depth map. For example, Yang *et al*. [29] proposed a post-processing method to improve the spatial resolution and accuracy of depth images by iterative bilateral filtering. The filtering-based method has lower operational complexity, but its ability to recover depth details is unsatisfactory, and the over-smooth and blurred boundaries are prone to appear in the reconstructed depth map. In recent years, the research focus has gradually shifted to deep learning solutions for SR, and many high-performance algorithms have emerged, such as DBPN [30], DEWRN [31], SRN [32], SADN [33], DTSR [34], [35], [36], *etc*. Taking into account the particularity of the depth map, deep learning-based DSR methods usually needs to design a specific network structure to improve the reconstruction performance. Riegler *et al*. [37] incorporated the total generalized variational constraint at the back-end of DCNN to form an end-to-end ATGV-Net. Song *et al*. [38] proposed to reconstruct the depth map in a way that a series of view synthesis sub-tasks can be learned in parallel. Sun *et al*. [39] proposed a depth-controlled slicing network that learns a set of slicing branches in a divide-and-conquer

manner and is parameterized by a distance-aware weighting scheme to adaptively integrate the different depths in the set.

B. Color-Guided DSR

As mentioned earlier, it is effortless for some depth cameras (such as Kinect) to acquire HR color images while acquiring depth maps. Therefore, the color-guided DSR model has received widespread attention in recent years and has become the mainstream model. The color-guided DSR is based on the similar structural information between the depth map and the aligned color image, *i.e*., the depth boundaries have a strong symbiotic relationship with the luminance boundaries. Filter-based approaches consider coeval structural relationships in addition to depth-neighborhood relationships when designing filters. For example, Kopf *et al*. [40] proposed a joint bilateral upsampling filter model by combining a Gaussian function based on the depth image neighborhood position relationship. He *et al*. [41] developed a local linear model of the filtered image and the bootstrap image model, and then proposed a bootstrap filter. Wang *et al*. [42] proposed a dual normal-depth regularization term to constrain the edge consistency between the normal map and the depth map. Recently, learning-based approaches have successfully applied DCNN to the field of color-guided DSR. Wen *et al*. [28] used a coarse-to-fine DCNN network to learn different filters with different kernel sizes, thus enabling data-driven training to replace the manually designed filters. Huang *et al*. [43] proposed a deep dense residual network with a pyramidal structure that leverages multi-scale features to predict high-frequency residuals through dense connectivity and residual learning. Guo *et al*. [25] designed a residual U-Net structure for the deep reconstruction task and introduced hierarchical feature-driven residual learning. Zuo *et al*. [21] proposed a data-driven super-resolution network based on global and local residual learning. Sun *et al*. [24] proposed a progressive multi-branch aggregation network that utilizes multi-scale information and high-frequency features to fully reconstruct the depth map in a progressive way. They also demonstrate that the low-level color information is only suitable for early feature fusion and does not help much for DSR at $\times 2$ and $\times 4$ cases.

III. Methodology

A. Overview

The overview of the proposed network is shown in Fig. 2, which is a dual-stream hierarchical reconstruction architecture. Given a LR depth map $D_{LR} \in \mathbb{R}^{h\times w\times 1}$ and the corresponding HR color image $I_{HR} \in \mathbb{R}^{H\times W\times 3}$ as inputs, the goal of our task is to reconstruct and generate a SR version of depth map $D_{SR} \in \mathbb{R}^{H\times W\times 1}$ with the same resolution of the color image. To be concise, we first extract the multi-level features of RGB and depth features via five progressive convolution blocks (green blocks in Fig. 2), where each block includes two 3×3 convolution layers and a 1×1 convolution layer. The obtained RGB and depth features are denoted as F_c^i and F_d^i ($i = \{1, 2, 3, 4, 5\}$), respectively.

Then, we achieve color-guided depth feature learning and detail restoration under the cooperation of the AFP, LDE, and HAG modules. It is worth noting that there are three inputs (if any) are sent to the AFP module: (1) The depth backbone features F_d^i in the corresponding level; (2) The low-level detail features F_{LDE}^i generated by the LDE module, which is used for detailed restoration in the low-level reconstruction stage; (3) The dense transfer features $(\mathrm{F}_{tr}^{i+1}, \mathrm{F}_{tr}^{i+2}, \cdots, \mathrm{F}_{tr}^{5})$ from all the completed reconstruction levels. At different reconstruction levels, the input features of the AFP module are different, which are specifically formulated as:

$$\mathrm{F}_{in}^i = \begin{cases} \mathrm{F}_d^i, & i = 5 \\ Concat(\mathrm{F}_d^i, \mathrm{F}_{tr}^k), & i = \{3, 4\} \\ Concat(\mathrm{F}_{LDE}^i, \mathrm{F}_{tr}^k), & i = \{1, 2\} \end{cases}, \tag{1}$$

where $Concat$ denotes the concatenation operation along the channel dimension, F_d^i represent the depth backbone features of the i-th level, F_{LDE}^i are the low-level detail features generated by the i-th LDE module, and F_{tr}^k denote the transfer features from the k-th completed reconstruction level, which can be calculated by:

$$\mathrm{F}_{tr}^k = (\mathrm{F}_{HAG}^k)\downarrow + \mathrm{F}_d^k, \tag{2}$$

where F_{HAG}^k represent the output features of the k-th HAG module, $\downarrow$ denotes the downsampling operation, and $k = \{i + 1, i + 2, \cdots, 5\}$. It should be noted that the inputs of the LDE module include the depth features F_d^i and color features F_c^i of the corresponding layer.

After that, the high-level abstract features F_c^5 and the depth backbone features F_d^i are fed into the HAG module to modify the output features F_{AFP}^i of AFP module and generate the reconstruction features F_{HAG}^i. Finally, the pixel-shuffle and convolution operations are performed on the features F_d^1 and F_{HAG}^1 to obtain the final upsampled depth map D_{SR}.

Note that, our model is trained by minimizing L_1 loss, which can be formulated as:

$$Loss = \|D_{SR} - D_{HR}\|_1, \tag{3}$$

where D_{SR} and D_{HR} denote the predicted depth SR result and ground truth, respectively, and $\|\cdot\|_1$ is the L_1 norm function. In the following subsections, we will provide the technical details of the AFP, LDE, and HAG modules one by one.

B. Attention-based Feature Projection Module

In order to achieve the depth map super-resolution, we need to map the low-resolution features to the desired high-resolution reconstruction features. Specifically, there are two issues that need to be paid attention to: (1) In order to recover more severely degenerated local details (such as depth boundaries and fine objects), simply increasing the depth of the network is insufficient and unwise. Therefore, we introduce a Multi-scale Content Enhancement (MCE) block to enhance the depth features before projection, using different receptive fields to recover detailed features at different scales as much as possible. (2) The information between the LR and HR domains is not absolutely one-to-one correspondence in the projection process, and the interference of excessive interfering information is likely to introduce additional errors, thereby impairing the reconstruction accuracy. To this end, we propose

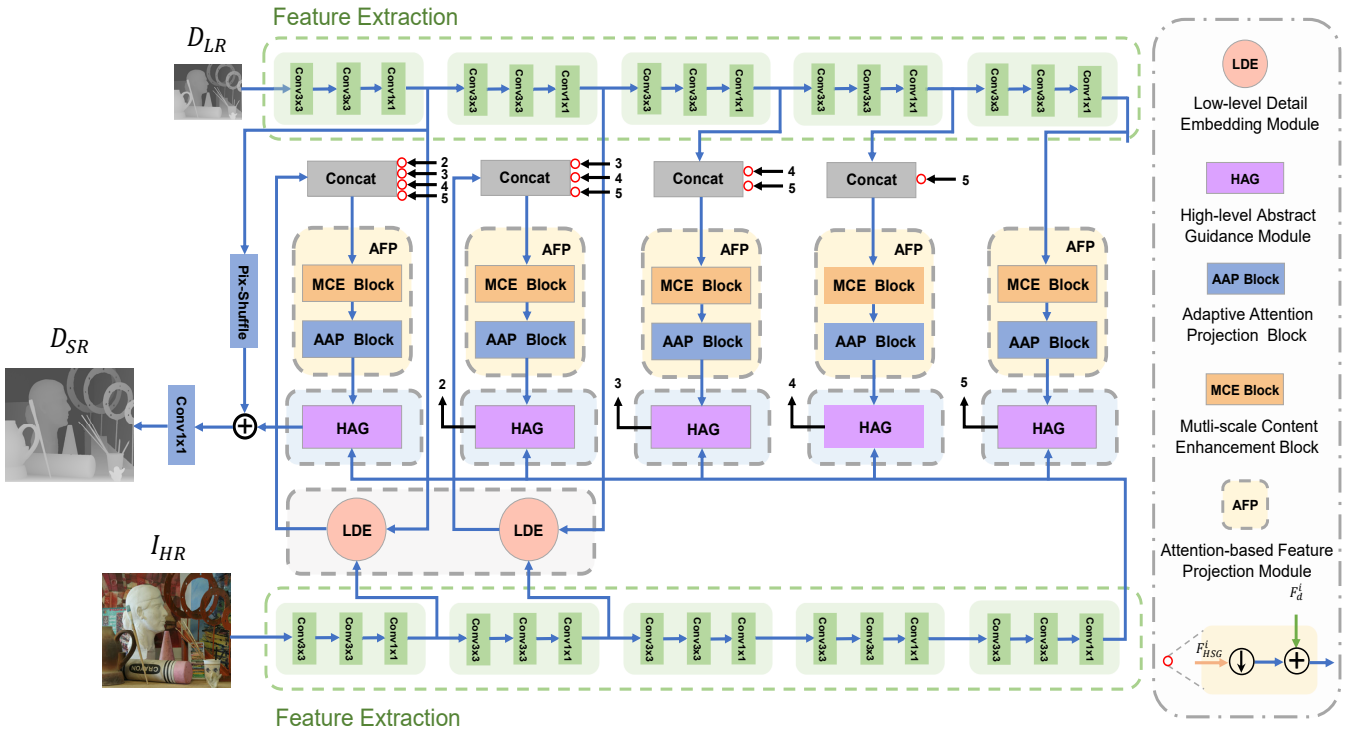


Fig. 2: The architecture of HCGNet. The LR depth map and HR color image are first embedded into the feature extraction unit to extract multi-level features. Then, the Attention-based Feature Projection (AFP) module, Low-level Detail Embedding (LDE) module, and High-level Abstract Guidance (HAG) module work together to gradually recover details in LR depth features and generate the HR depth map. The use of color information is manifested in two aspects. On the one hand, the low-level color features are used in the low-level reconstruction stage to restore details through the LDE module. On the other hand, the high-level abstract features are used at the end of the AFP module to provide semantic guidance through the HAG module.

an Adaptive Attention Projection (AAP) block to project valid information in the attention domain, guaranteeing the effectiveness and compactness of the projected features. Note that, four cascaded AAP blocks are used in the AFP module to achieve better performance. In summary, as shown in Fig. 3, the MCE block and AAP blocks together form the AFP module to achieve depth feature reconstruction.

1) Multi-scale Content Enhancement block: Multi-scale information can effectively perceive and model different details, which is of great significance for detail restoration in DSR. As shown in the lower left of Fig. 3, the MCE block contains a stack of four dilated convolution layers with different dilation rates, which are applied to capture more details with different scales of receptive fields [44], [45]. Moreover, we employ dense connections to obtain full information from all previous layers, which are formulated as:

$$\mathrm{F}_m^i = \begin{cases} MD(\mathrm{F}_{in}^i), & m = 1 \\ MD(Concat(\mathrm{F}_{in}^i, \mathrm{F}_1^i, \cdots, \mathrm{F}_{m-1}^i)), & m = \{2, 3, 4\} \end{cases} \tag{4}$$

where MD denotes the multi-scale dilated convolution operation with different dilation rates of 1, 2, 3, and 4, and F_m^i is the output of each multi-scale dilated convolution. Finally, all the multi-scale dilation features are concatenated and fused by a 1×1 convolution layer:

$$\mathrm{F}_{MCE}^i = DeConv(Conv_{1\times1}(Concat(\mathrm{F}_{in}^i, \mathrm{F}_1^i, \cdots, \mathrm{F}_4^i))), \tag{5}$$

where $Conv_{1\times1}$ denotes the convolution layer with the kernel size of 1×1, F_{MCE}^i is the output of MCE block, which perceives content information of different scales, and $DeConv$ denotes the upsampling operation implemented by the deconvolution layer.

2) Adaptive Attention Projection block: The super-resolution process of depth maps needs to bridge the huge gap between the LR domain and the HR domain. In fact, the focus of the DSR task is not to generate content from scratch but to supplement, refine, and enhance the details such as boundaries. From the perspective of the frequency domain, low-frequency information is usually included in the smooth regions while high-frequency regions contain more boundary information. Therefore, to extract clear color boundaries and suppress their interfering textures, we need to correct the error information progressively while extracting the high-frequency features of the image. Moreover, blindly and indiscriminately performing super-resolution reconstruction on all regions is a sub-optimal way, which is also difficult to achieve the purpose of optimizing important regions with more severe degradation. In other words, in the process of restoring information from the LR domain to the HR domain (which is also called the projection process), interference may be introduced without

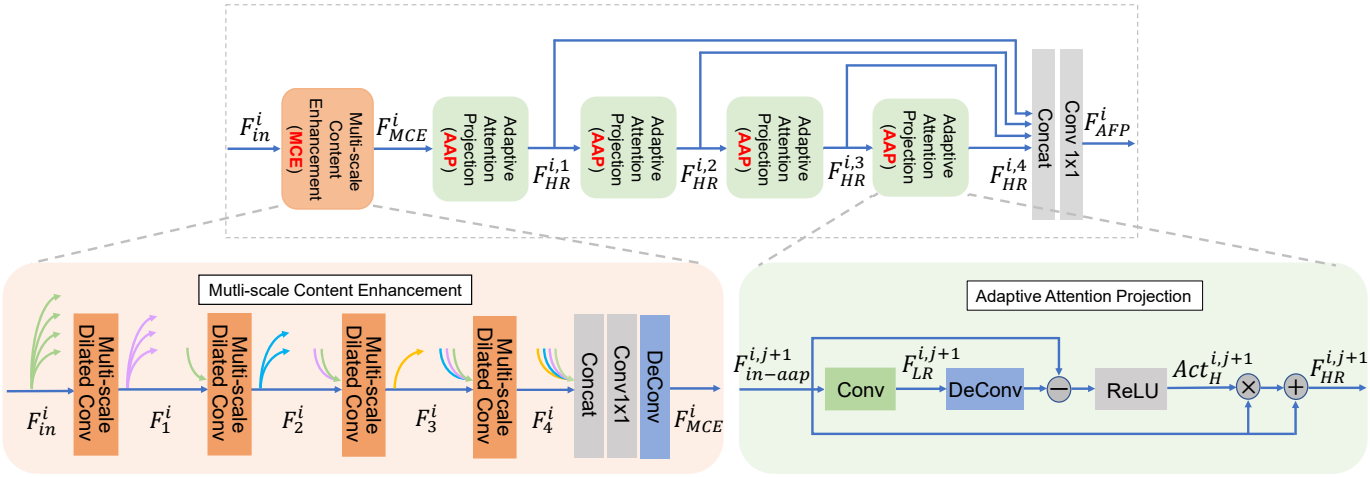


Fig. 3: The whole architecture of AFP module and details of sub-blocks, *i.e.*, MCE block and AAP block.

filtering, thereby introducing additional errors and affecting the reconstruction accuracy. Hence, we design the AAP block to reinforce the key restoration regions in an attention manner, thereby suppressing the interference and improving the reconstruction performance, as shown in the bottom flowchart of Fig. 3.

For the AAP blocks, the input of the first AAP block is the output features of the MCE block, while the input of the other blocks is the output of the previous AAP block. As such, the input of the AAP block can be uniformly formulated as:

$$\mathrm{F}_{in-aap}^{i,j+1} = \begin{cases} \mathrm{F}_{MCE}^{i}, & j = 0 \\ \mathrm{F}_{HR}^{i,j}, & j = \{1,2,3\} \end{cases} \tag{6}$$

where $\mathrm{F}_{HR}^{i,j}$ is the HR output of j-th AAP block in the i-th level (will be formulated further below).

For the algorithm of AAP, we simulate the DSR process by using both down-projection and up-projection, thereby obtaining the reconstructed HR feature map under worse conditions. These projection blocks can be interpreted as self-correcting processes that provide projection errors to the sampling layer, and thus progressively generate better solutions:

$$\mathrm{F}_{HR_{rough}}^{i,j+1} = DeConv(Conv(\mathrm{F}_{in-aap}^{i,j+1})), \tag{7}$$

where $Conv$ is a convlution layer for down-projection, and $DeConv$ is a deconvolution layer for up-projection.

Then, we subtract the reconstructed HR features from the original HR features to generate the residual features and extract the high-frequency features of the image, which encode the content information that needs to be recovered during reconstruction. The projected attention map is calculated by:

$$\mathrm{Act}_{H}^{i,j+1} = ReLU(\mathrm{F}_{in-aap}^{i,j+1} - \mathrm{F}_{HR_{rough}}^{i,j+1}), \tag{8}$$

where $ReLU$ denotes the rectified linear unit. The projected attention map will correct errors in reconstruction and avoid the degradation caused by feature projection between the LR and HR domains.

Finally, the residual map is activated as a projected attention map and used to adaptively refine the original HR features:

$$\mathrm{F}_{HR}^{i,j+1} = \mathrm{Act}_{H}^{i,j+1} \otimes \mathrm{F}_{in-aap}^{i,j+1} + \mathrm{F}_{in-aap}^{i,j+1}, \tag{9}$$

where $\otimes$ denotes the element-wise multiplication. With four serial AAP blocks, four HR reconstruction features from coarse to fine are generated. Combining them, we can obtain the final output of the AFP module:

$$\mathrm{F}_{AFP}^{i} = Conv_{1\times1}(Concat(\mathrm{F}_{HR}^{i,1}, \mathrm{F}_{HR}^{i,2}, \mathrm{F}_{HR}^{i,3}, \mathrm{F}_{HR}^{i,4})), \tag{10}$$

where F_{AFP}^{i} denote the initial reconstruction depth features.

C. Low-level Detail Embedding Module

As is well-known, high-resolution color images are readily available and contain much useful information, such as boundaries, textures, and semantic information, *etc*. Therefore, introducing color guidance into the DSR model has become the mainstream idea in this field. However, there is no complete consensus on which color information to use and how to use it. Considering the different roles of color features at different levels, we provide a differentiated solution of color guidance strategy in this paper. Concretely, we design a Low-level Detail Embedding (LDE) module at the low-level reconstruction stage to leverage low-level color features for enhancing the high-frequency details of depth features, such as boundaries. In addition, we design a High-level Abstract Guidance (HAG) module, where the high-level abstract color features are used to perform content correction on the original reconstruction features, preventing content shifts during the depth reconstruction. We will introduce the LDE module in this subsection, and provide the details of the HAG module in the next subsection.

For depth map super-resolution, accurate and sharp boundary reconstruction has always been the focus of researchers' unremitting efforts. It just so happens that what the low-level layer of the color branch learns is detailed information such as texture, boundary, *etc*. Therefore, we introduce the color features at lower levels (*i.e.*, the first two layers) of the HR color branch through the LDE module and take the output as one of the inputs of the AFP module. However, the depth boundaries are not absolutely consistent with the RGB boundaries. In fact, the boundaries in the depth map are mainly the object boundaries, while the color image

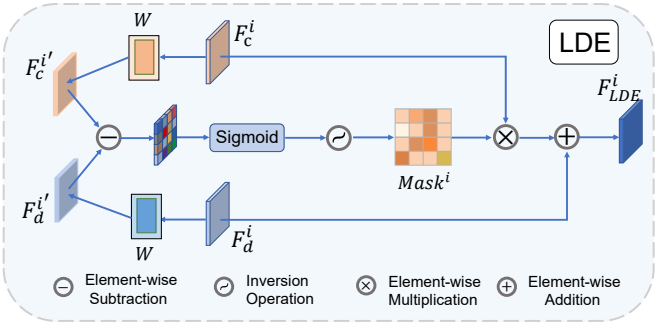


Fig. 4: The architecture of Low-level Detail Embedding Module.

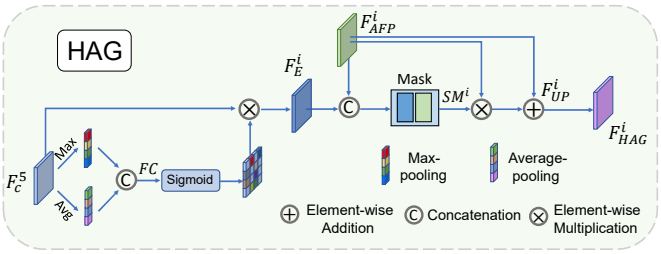


Fig. 5: The architecture of High-level Abstract Guidance Module.

includes rich texture boundaries inside the object in addition to the object boundaries. Obviously, the texture boundaries are interferences for DSR. However, it is difficult to determine the delineation of texture and boundaries very clearly in network reconstruction, so completely discarding texture details is not the best option. Therefore, instead of removing all texture information with absolute gradient boundaries, our proposed LDE module suppresses the interfering information of RGB by learning residual masks, which is shown in Fig. 4.

Concretely, we first make a domain mapping of the RGB-D features using point-wise convolution and depth-wise convolution. The subsequent key step of subtraction actually serves to distinguish between similar as well as interfering regions in the RGB-D features. Since explicitly filtering out the appropriate complementary regions is difficult, we obtain the regions with large variances in the RGB-D modality, *i.e.*, the redundant regions, by subtractive operations. The reverse residual mask, unlike the binarized mask, will increase the weight of complementary information and reduce the weight of information in redundant regions, thus differentiating the guiding role of low-level color features:

$$\mathrm{RM}^i = 1 - Sigmoid(Conv_{1\times1}(W_c \cdot \mathrm{F}_c^i - W_d \cdot \mathrm{F}_d^i)), \quad (11)$$

where RM^i denotes the residual mask, W_c and W_d represent the mapping matrix for the color features and depth features, $Sigmoid$ is the normalization operation, and i indexes the lower level here, which is equal to 1 or 2.

In this way, the residual mask highlights the most relevant part of the color and depth information, so we multiply it with the initial color features to obtain the effective color features that can be used for depth reconstruction guidance. Moreover, we believe that the feature representation of the color information filtered by the residual mask in the lower levels is in the same domain as the depth information, so the final fusion adopts a direct summation scheme, which is also consistent with objective rules. Therefore, the final output of the LDE module can be formulated as:

$$\mathrm{F}_{LDE}^i = \mathrm{RM}^i \otimes \mathrm{F}_c^i + \mathrm{F}_d^i, \quad (12)$$

where $\otimes$ denotes the element-wise multiplication.

D. High-level Abstract Guidance Module

As analyzed earlier, the existing methods mainly focus on extracting color features to supplement the details for depth reconstruction, just like the functions implemented by our LDE module. However, let us rethink the role of color-guided features: Is this detailed guidance strategy sufficient? In fact, high-level color features are very important for many tasks, which contains abstract global information and preserves semantic outlines. In the DSR task, the existing methods ignore one issue, *i.e.*, the global abstract information preserving ability of the reconstructed features. As the reconstruction process progresses, there is a possibility that semantic consistency may be shifted or blurred, which is very unfavorable for the subsequent depth-oriented application tasks. This is mainly caused by the lack of global guidance in the reconstruction. Fittingly, our color branch can provide high-resolution, offset-free color guidance information. This also benefits from the fact that the semantic information extraction for the color branch does not rely on the process of depth reconstruction, but extracts shallow texture features and high-level abstract features through multi-layer convolution deepening. Since no resolution change is involved, the high-resolution color information is not shifted during the extraction process. Inspired by these, we design a HAG module to maintain the content attribute during the depth reconstruction, which is equipped after each AFP module. Specifically, the high-level color information at the top layer of the color branch is utilized to generate a mask that encodes the global content guidance information, and is further used to modify the initial reconstruction features F_{AFP}^i (*i.e.*, the output of the AFP module).

As shown in Fig. 5, we first enhance the top-layer color features F_c^5 at the spatial level [46], thereby generating the reweighted color features F_E^i that highlight the important locations:

$$\mathrm{Act}_{SL}^i = Sigmoid(FC(Concat(Max(\mathrm{F}_c^5), Avg(\mathrm{F}_c^5)))), \quad (13)$$

$$\mathrm{F}_E^i = \mathrm{F}_c^5 \otimes \mathrm{Act}_{SL}^i, \quad (14)$$

where FC denotes the fully connected layers, Max and Avg are the max-pooling operation along channel dimension and global average pooling, respectively, and Act_{SL}^i is the spatial-level attention.

Considering the auxiliary role of semantic features, we still take the depth reconstruction features as the dominant one in the guiding process. Thus, we concatenate the enhanced

color features with the original reconstruction features, and then generate a semantic mask:

$$\mathrm{SM}^i = Conv_{1\times1}(PReLU(Conv_{3\times3}(Concat(\mathrm{F}^i_{AFP}, \mathrm{F}^i_E)))), \tag{15}$$

where $PReLU$ is the parametric rectified linear unit, and $Conv_{n\times n}$ denotes the convolution layer with the kernel size of $n \times n$.

With the semantic mask, the original depth reconstruction features can be refined by:

$$\mathrm{F}^i_{HAG} = \mathrm{SM}^i \otimes \mathrm{F}^i_{AFP} + \mathrm{F}^i_{AFP}, \tag{16}$$

where F^i_{HAG} are the output of the corresponding HAG module. It should be noted that F^i_{HAG} are the final reconstruction feature at each stage, where the features of the last layer F^1_{HAG} will be directly used to generate the upsampled depth map, and the reconstruction features of other layers will be sent to the AFP module through the dense transmission to realize the progressive learning of the entire network.

IV. Experiments

A. Datasets and Implementation Details

To demonstrate the effectiveness of the proposed method, we conduct comprehensive experiments on the Middlebury dataset, NYU v2 dataset [57], real-world RGB-D-D dataset [58], and Lu dataset [59]. All these datasets are produced with the alignment of color and depth images, and all the ground-truth values of the upsampled depth map are owned by the dataset. From the perspective of dataset construction, for depth cameras such as Kinect, most of the scenes are RGB-D aligned with default parameters. With tools such as Matlab, the accuracy can be further improved by co-calibration. In addition, the device usually includes an SDK that enables users to interpolate the depth output to match the resolution of RGB images. However, the quality of the interpolated depth map can not be guaranteed. Thus, the NYU v2 dataset undergoes additional processing such as depth completion and calibration for improved quality and accuracy.

- We collect 36 RGB-D images from Middlebury dataset (6, 21, 9 images from 2001 [60], 2006 [61], and 2014 [62] datasets, respectively) for training, and 6 standard depth maps from Middlebury 2005 dataset [63] for testing. The resolution of the image is mostly around 1000×1000.
- As for the NYU v2 dataset, it is a real-scene dataset captured by the Kinect camera. We use the first 1000 pairs with the resolution of 640×480 for training and the remaining 449 pairs for testing. The model trained on the NYU v2 dataset is also directly used to test on the Lu dataset and RGB-D-D dataset for generalization evaluation.
- The RGB-D-D dataset is a real-world indoor dataset captured by a Huawei P30 pro cellphone equipped with the color camera and time of flight (TOF) camera, also with a resolution of 640×480. Following the setting in [58], 2215 pairs are used for training and 405 pairs for evaluation.
- The Lu dataset only consists of 6 RGB-D pairs acquired by the ASUS Xtion Pro camera, all of which are used for testing.

For quantitative evaluation, the metrics of Mean Absolute Difference (MAD) and Root Mean Square Error (RMSE) are introduced [22], [69], [70]. Following PMBANet [24], we augment our training samples by cropping the HR pairs into around 15000 HR patches with the squared size of 64, 128, and 256 according to the upsampling factors of $\times4$, $\times8$, and $\times16$. Meanwhile, we perform horizontal and vertical flipping of the training samples with a probability of 0.5. The LR depth patches are downsampled into a fixed size of 16×16 from HR depth patches via Bicubic interpolation. With respect to the structural details of our HCPNet, in the feature extraction phase, the network maps the RGB-D features to each of the 64-channel dimensions and maintains the number of channels constant throughout the feature extraction phase. For the LDE, HAG, and MCE modules, the input channel dimension changes depending on the number of input features, and the output is uniformly 64 channels. Our HCGNet is implemented by PyTorch with an NVIDIA 3090 GPU. As a hyperparameter, the batch size is set to 8 at $\times8$ and $\times4$ factor cases and set to 4 for the $\times16$ factor case. During training, we use Adam optimizer with momentum of 0.9, and the network optimization parameters of β_1, β_2, and ϵ are set to 0.9, 0.99, and $1e^{-8}$, respectively. The initial learning rate is set to $1e^{-4}$, and it is decreased by multiplying by 0.1 for the first 100 epochs and the next 50 epochs.

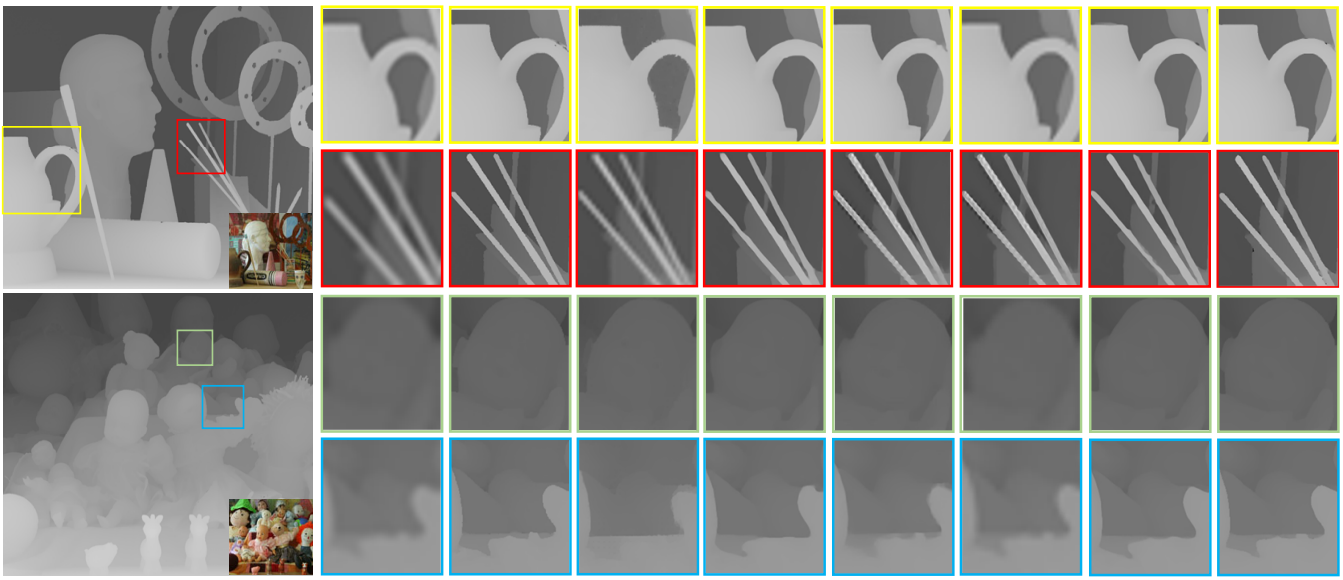
B. Performance Comparison

1) Middlebury Dataset: We compare our method with some state-of-the-art DSR methods under different upsampling factors ($\times4$, $\times8$, and $\times16$), including four traditional depth SR methods (*i.e.*, TGV [47], EG [48], JGF [49], and CDLLC [50]) and nine deep-learning based methods (*i.e.*, GSRPT [51], MDDL [52], DEIN [53], DJF [54], CCFN [28], CTKT [55], BridgeNet [22], PMBANet [24] and MIGNet [56]).

The quantitative comparisons of the MAD score are reported in Table I. We can observe that the traditional DSR models often fail to achieve satisfactory performance, especially when dealing with more complex scenes (*e.g.*, Art) or larger upsampling factors (*e.g.*, $\times16$). In contrast, the deep-learning based DSR models show more competitive performance, in which the proposed HCGNet outperforms other SOTA methods among different scenarios and achieves the best overall average MAD performance under different upsampling factors. Moreover, our method achieves obvious performance improvement under a challenging large sampling factor (such as $\times16$). Compared with the **second best** method, the MAD value of the Mobius scene is decreased from 0.54 to 0.45, with a percentage gain of 16.7%, and the MAD percentage gain of the Books scene reaches 13.7%. Fig. 6 demonstrates the visual comparisons of different methods under the factor of $\times8$. As visible, our method can recover more complete and accurate depth details. In the first image, compared with the results of other methods, the boundaries around sticks are sharper and smoother with less jagged noise, and the structure of the

TABLE I: Quantitative DSR results (in MAD ↓) on the Middlebury 2005 dataset. The best performance is marked in bold, and the second-best performance is underlined.

Methods	Art			Books			Dolls			Laundry			Mobius			Reindeer			Average		
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16
TGV [47]	0.65	1.17	2.30	0.27	0.42	0.82	0.33	0.70	2.20	0.55	1.22	3.37	0.29	0.49	0.90	0.49	1.03	3.05	0.43	0.84	2.11
M EG [48]	0.48	0.71	1.35	0.15	0.36	0.70	0.27	0.49	0.74	0.28	0.45	0.92	0.23	0.42	0.75	0.36	0.51	0.95	0.30	0.49	0.90
JGF [49]	0.47	0.78	1.54	0.24	0.43	0.81	0.33	0.59	1.06	0.36	0.64	1.20	0.25	0.46	0.80	0.38	0.64	1.09	0.34	0.59	1.08
CDLLC [50]	0.53	0.76	1.41	0.19	0.46	0.75	0.31	0.53	0.79	0.30	0.48	0.96	0.27	0.46	0.79	0.43	0.55	0.98	0.34	0.54	0.95
GSRPT [51]	0.48	0.74	1.48	0.21	0.38	0.76	0.28	0.48	0.79	0.33	0.56	1.24	0.24	0.49	0.80	0.31	0.67	1.07	0.31	0.55	1.02
MDDL [52]	0.46	0.62	1.87	0.24	0.37	0.73	0.29	0.51	0.79	0.32	0.53	1.11	0.19	0.37	0.74	0.41	0.51	0.95	0.32	0.49	1.03
DEIN [53]	0.40	0.64	1.34	0.22	0.37	0.78	0.22	0.38	0.73	0.23	0.36	0.81	0.20	0.35	0.73	0.26	0.40	0.80	0.26	0.42	0.87
DJF [54]	0.40	1.07	2.05	0.16	0.45	1.00	0.20	0.49	0.99	0.28	0.71	1.67	0.18	0.46	1.02	0.23	0.60	1.32	0.24	0.63	1.46
CCFN [28]	0.43	0.72	1.50	0.17	0.36	0.69	0.25	0.46	0.75	0.24	0.41	**0.71**	0.23	0.39	0.73	0.29	0.46	0.95	0.27	0.47	0.89
CTKT [55]	0.25	0.56	1.44	**0.12**	0.28	0.67	0.18	0.39	0.65	0.16	0.40	0.76	**0.14**	0.29	0.69	0.18	0.41	0.77	0.17	0.39	0.83
BridgeNet [22]	0.30	0.58	1.49	0.14	0.24	0.51	0.19	0.34	0.64	0.17	0.34	**0.71**	0.15	0.26	0.54	0.19	0.31	0.70	0.19	0.35	0.77
PMBANet [24]	0.26	0.51	1.22	0.15	0.26	0.59	0.19	0.32	0.59	0.17	0.31	**0.71**	0.16	0.26	0.67	0.17	0.33	0.74	0.18	0.33	0.75
MIGNet [56]	**0.21**	0.47	**1.08**	0.15	0.25	0.53	0.21	0.35	0.71	0.19	0.35	0.81	0.15	0.26	0.55	0.22	0.36	0.82	0.19	0.34	0.75
HCGNet (Ours)	**0.21**	**0.41**	1.38	0.13	**0.23**	**0.44**	**0.17**	**0.31**	**0.57**	**0.14**	**0.28**	0.78	**0.14**	**0.24**	**0.45**	**0.15**	**0.26**	**0.64**	**0.16**	**0.29**	**0.71**



(a) GT and Color images (b) LR (c) GT (d) TGV (e) CTKT (f) GSRPT (g) Bicubic (h) PMBANet (i) Ours

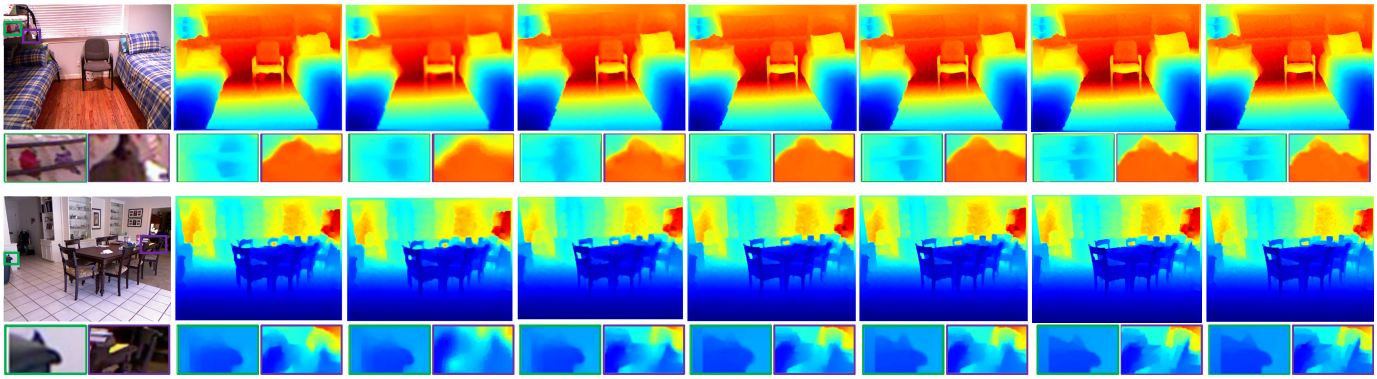
Fig. 6: Visual comparisons of ×8 upsampling results on the Middlebury 2005 dataset. (a) HR depth maps and color images. (b)-(c) LR depth maps and the HR GT. (d)-(i) The reconstructed HR depth maps by the TGV, CTKT, GSRPT, Bicubic, PMBANet, and Our method, respectively.

TABLE II: Quantitative comparisons with state-of-art methods (in RMSE ↓) on the NYU v2 dataset. The best performance is marked in bold, and the second-best performance is underlined. Note that, the depth values are measured in centimeters

	Bicubic	TGV [47]	DJFR [64]	SDF [65]	SVLRM [66]	DKN [19]	FDKN [19]	FDSR [58]	PMBANet [24]	JIIF [67]	DCT [68]	HCGNet (Ours)
×4	8.16	6.98	3.38	3.04	1.74	1.62	1.86	1.61	1.30	1.37	1.59	**1.22**
×8	14.22	11.23	5.86	5.67	5.59	3.26	3.58	3.18	2.75	2.76	3.16	**2.53**
×16	22.32	28.13	10.11	9.97	7.23	6.51	6.96	5.86	5.48	5.27	5.84	**4.85**

distant object is more complete and continuous. For the Dolls image, our method can restore the clear and sharp boundaries of the reconstructed objects. For example, the left ear of the doll in the third row is restored with complete contour and the surrounding values are artifact-free, and the sleeves of the doll (*e.g.*, the raised part in the middle) shown in the last row are also recovered more satisfactorily than other methods. In general, the comparison methods can more or less introduce additional shape distortion and noise, resulting in blurring, discontinuities, and changed depth values in the reconstructed images. In contrast, our method is able to recover objects more completely while taking into account the restoration of depth details without introducing destructive contamination.

2) NYU v2 Dataset: We evaluate our method on the NYU v2 dataset and compare it with other SOTA methods, including Bicubic, TGV [47], DJFR [64], SDF [65], SVLRM [66], DKN [19], FDKN [19], FDSR [58], PMBANet [24], JIIF [67] and DCT [68].

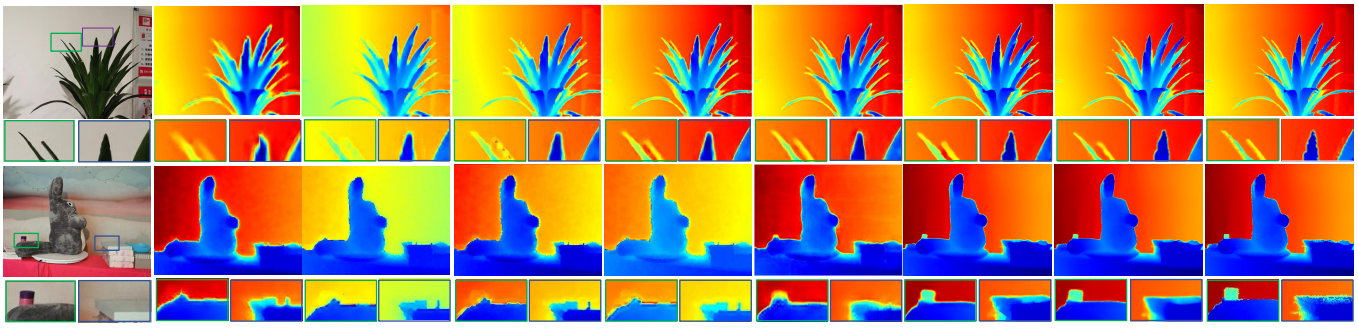


(a) Color images (b) SVLRM (c) SDF (d) DJFR (e) DKN (f) FDSR (g) Ours (h) GT

Fig. 7: Visual comparisons of different methods under ×8 upsampling on the NYU v2 dataset. (a) Color images. (b) SVLRM. (c) SDF. (d) DJFR. (e) DKN. (f) FDSR. (g) Ours. (h) GT.

TABLE III: Quantitative comparisons with state-of-art methods (in RMSE ↓) on the RGB-D-D dataset. The best performance is marked in bold, and the second-best performance is underlined. The depth values are measured in centimeters. FDSR$^+$ and HCGNet$^+$ are the models retrained on the RGB-D-D dataset.

	SDF [65]	DJFR [64]	FDKN [19]	DKN [19]	FDSR [58]	FDSR$^+$	DCT [68]	JIIF [67]	HCGNet (Ours)	HCGNet$^+$ (Ours)
×4	2.00	3.35	1.18	1.30	1.16	1.11	<u>1.08</u>	1.17	1.13	**0.99**
×8	3.23	5.57	1.91	1.96	1.82	<u>1.71</u>	1.74	1.80	1.77	**1.49**
×16	5.16	8.15	3.41	3.42	3.06	3.01	3.05	2.84	<u>2.70</u>	**2.00**



(a) Color images (b) DJFR (c) FDKN (d) DKN (e) FDSR (f) FDSR$^+$ (g) HCGNet (h) HCGNet$^+$ (i) GT

Fig. 8: Visual comparisons of different methods under ×8 upsampling on the RGB-D-D dataset. (a) Color images. (b) DJFR. (c) FDKN. (d) DKN. (e) FDSR. (f) FDSR$^+$. (g) HCGNet (Ours). (h) HCGNet$^+$ (Ours). (i) GT.

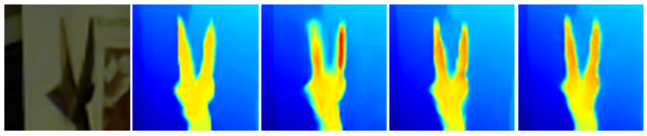
The quantitative results are provided in Table II. We can see that our method outperforms all the SOTA methods under different upsampling factors. In the most challenging $\times 16$ upsampling situation, compared with the **second best** method, the RMSE of our method reaches 4.85, with a percentage gain of 7.4%. Fig. 7 shows some visual comparisons on the NYU v2 dataset under the $\times 8$ factor case. It can be seen that our method has obvious advantages in boundary reconstruction and depth preservation. For example, in the overlapping regions of the bed sheet and bed frame marked in purple in the first image, the depth map reconstructed by our method has sharper boundaries and more accurate depth values. In the second image, our method shows a stronger ability to portray details, such as the object above the trash can in the left image and the oblique border area of the chair in the right image.

3) RGB-D-D Dataset: On this dataset, we evaluate our method (*i.e.*, HCGNet and HCGNet$^+$) and other SOTA methods, including SDF [65], DJFR [64], DKN [19], FDKN [19], FDSR [58], FDSR$^+$, DCT [68], and JIIF [67]. The superscript '$+$' indicates the corresponding model was retrained on the RGB-D-D dataset, and no superscript mark indicates the model was only trained on the NYU v2 dataset [57] without retraining or fine-tuning.

The quantitative results are reported in Table III. Compared to other models without retraining, our method (*i.e.*, HCGNet) achieves the best performance under large factor cases and even outperforms the retrained FDSR$^+$ method under the most challenging case of $\times 16$, which also demonstrates the generalizability of our model. Concretely, compared with the **second best** method without retraining, the average RMSE value drops from 3.05 to 2.70 under $\times 16$ case with a percentage gain of 11.5%, and the percentage gain under $\times 8$ factor reaches 5.5%. At the same time, we can see a further improvement in

TABLE IV: Quantitative comparisons with state-of-art methods (in RMSE ↓) on the Lu dataset. The best performance is marked in bold, and the second-best performance is underlined.

	Bicubic	FDKN [19]	DKN [19]	FDSR [58]	DCT [68]	JIIF [67]	Ours
×4	2.42	**0.82**	0.96	1.29	0.88	<u>0.85</u>	1.02
×8	4.54	2.10	2.16	2.19	1.85	<u>1.73</u>	**1.68**
×16	7.38	5.05	5.11	5.00	4.39	<u>4.16</u>	**3.75**



(a) Color image (b) GT (c) DKN (d) JIIF (e) Ours

Fig. 9: Visual comparisons of different methods under ×8 upsampling on the Lu dataset. (a) Color image. (b) GT. (c) DKN. (d) JIIF. (e) Ours.

the performance of the model after retraining. Our retrained version (*i.e.*, HCGNet$^+$) achieves the best performance under all factor cases and the percentage gain reaches 33.6% under $\times 16$ case compared with the retained FDSR$^+$ method. For the visual comparison, Fig. 8 demonstrates the details on the RGB-D-D dataset under $\times 8$ case. In the first image, our method restores the leaves with clear and sharp boundaries, and even the sawtooth shape of the leaves in the lower right corner is partially restored. In the more challenging second image, our method recovers not only the accurate values both around and inside the cube in the lower left corner but also the complete boundaries of the books.

4) Lu Dataset: We also evaluate our method on the Lu Dataset and compare it with other SOTA methods, including Bicubic, FDKN [19], DKN [19], FDSR [58], DCT [68] and JIIF [67]. The Lu dataset was captured in low light and a relatively complex environment, which challenges the generalization ability of the method. The scale of the dataset is relatively small, so all methods are not retrained on this dataset. As shown in Table IV, our method achieves competitive performance compared with other methods, especially at the large factor cases (*i.e.*, $\times 8$, $\times 16$). For example, compared with the **second best** method, the average RMSE value is decreased from 4.16 to 3.75 under $\times 16$ factor case, with a percentage gain of 9.9%. Fig. 9 shows a visual example of different methods on the Lu dataset. As visible, the proposed method punches above its weight in terms of detail reconstruction, such as the sharper borders around the statue and more accurate depth values (*e.g.*, the ears) inside the statue.

C. Ablation Study

1) Validation of modules. Ablation studies are conducted to verify the effectiveness of each key component designed in the proposed HCGNet on both the Middlebury 2005 dataset and the NYU v2 dataset. We remove the color branch and simplify the AFP module as the baseline model. Specifically, we replace the MCE block with a few simple convolution layers and replace the AAP block with the residual block accordingly. Then, we sequentially add the HAG, LDE, and AFP modules to the baseline to verify the effectiveness of each module. For the validation of MCE and AAP blocks in AFP, we replace the MCE with the similar RDB [71] block and the AAP module with the similar DBPN [30] block. The quantitative results under the $\times 8$ case are reported in Table V, and some visual examples are shown in Fig. 10.

TABLE V: Ablation studies of our HCGNet in terms of average MAD (↓) values on the Middlebury 2005 dataset and RMSE (↓) values on the NYU v2 dataset ($\times 8$ case).

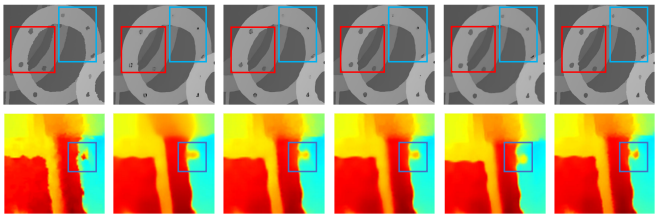
Model	Baseline	HAG	LDE	RDB+DBPN	AFP	MAD	RMSE
1	✓					0.42	3.80
2	✓	✓				0.36	2.88
3	✓	✓	✓			0.34	2.74
4	✓	✓	✓	✓		0.32	2.61
5	✓	✓	✓		✓	**0.29**	**2.53**

TABLE VI: Quantitative evaluation in RMSE (↓) with different number of AFP and AAP modules on the NYU v2 dataset ($\times 8$ case).

Model	RMSE	Model	RMSE
5AFPs (Ours)	**2.53**	4AAPs (Ours)	**2.53**
4AFPs	2.63	3AAPs	2.64
3AFPs	2.68	5AAPs	2.61

First, we verify the role of color guidance in DSR. After introducing the HAG module into the baseline to provide high-level abstract guidance, we can see that the MAD value on the Middlebury dataset is improved from 0.42 to 0.36 with the percentage gain of 14.3%, and RMSE values on the NYU v2 dataset decreased from 3.80 to 2.88 with the percentage gain of 24.2%. As visible, compared with the second and third columns in the first image of Fig. 10, the indistinguishable hole regions in the baseline model (marked in red box) have been obviously restored. Then, we embed the color detail information into the depth reconstruction branch through the LDE module. As reported in model 3 of Table V, the performance is further improved, where the percentage gains of the MAD value on the Middlebury 2005 dataset and RMSE value on the NYU v2 dataset reach 5.6% and 4.9% compared with model 2, respectively. At the same time, as shown in Fig. 10(d), the holes in the visualization result are also reconstructed more clearly. In summary, all these results demonstrate that our proposed HAG and LDE modules provide effective guidance information for depth map super-resolution reconstruction from two perspectives, and improve the reconstruction accuracy.

Furthermore, we verify the effectiveness of the proposed AFP module. On the basis of our DSR framework for color guidance, we introduce the AFP module to achieve multi-scale content enhancement and adaptive attention projection. Compared with model 3 in Table V, the MAD score is further improved from 0.34 to 0.29 with a percentage gain of 14.7%, and the RMSE score is decreased from 2.74 to 2.53 with a percentage gain of 7.7%. In terms of the visualization results shown in Fig. 10(f), the final full model (model 5) can not only maintain the details of the holes in the red-marked area but also effectively update the incorrectly reconstructed holes in the lower right corner of model 3 in the first row. To further illustrate the superiority of the AFP module design over



(a) (b) (c) (d) (e) (f)

Fig. 10: Visual comparisons of different ablation models (×8 case). (a) Ground truth. (b) Baseline (model 1). (c) Baseline + HAG (model 2). (d) Baseline + HAG + LDE (model 3). (e) Baseline + HAG + LDE + RDB + DBPN (model 4). (f) Baseline + HAG + LDE + AFP (model 5).

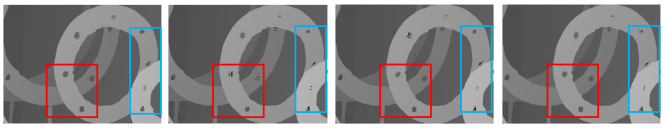
similar methods, we add the ablation experiments of model 4. Comparing model 4 and model 5, we can see that after replacing MCE and AAP with RDB and DBPN respectively, the model performance drops obviously. On the Middlebury 2005 dataset, the MAD score increases from 0.29 to 0.32, with a performance drop of 10.3%. While on the NYU v2 dataset, the RMSE score increases from 2.53 to 2.61, with a performance drop of 3.2%. The experimental results also prove the effectiveness of our proposed AFP module in-depth reconstruction.

Finally, we conduct an ablation study to compare the performance of different numbers of modules within the AFP and AAP components, the results of which are illustrated in Table VI. Specifically, we conduct experiments on the AFP modules by sequentially removing them from level 5 and level 4 of the network, while still retaining the HAG's global guidance. The results indicate that the removal of one AFP module (reducing the number from five to four) leads to an increase in RMSE from 2.53 to 2.63, corresponding to a performance decrease of 4%. Since the proposed HCGNet is a hierarchical progressive reconstruction, the quality of the reconstruction at each layer has an essential impact on subsequent layers. For the ablation of AAP blocks, compared with the four AAP blocks, the RMSE value of three blocks increases from 2.53 to 2.64, with a performance drop of 4.3%. Meanwhile, observation reveals that too many AAP blocks can also lead to performance degradation. This is because, at the current stage, the quality of the depth features is limited, and the correction capability of the AAP module is similarly constrained. Excessively deep stacking may, in fact, hinder the restoration of the depth features. Based on the above experimental results, we finally set the number of AAP to 4.

2) Portability of modules. To further validate the effectiveness of the proposed LDE and HAG modules, we conduct validation experiments by adding LDE and HAG modules to the PMBANet [24], which is a network considering only low-level color information (Mode (a) in Fig. 1). Specifically, we use the depth features of the corresponding layer, which is modified by the LDE module, as the input of the current MBA block, while the global HAG module is used to refine the depth features after each reconstruction branch. As shown in Table VII, it can be seen that after adding these two modules to the PMBANet, the performance is improved on both two

TABLE VII: Quantitative evaluation of adding our modules to PMBANet, including the MAD (↓) on the Middlebury 2005 dataset and the RMSE (↓) on the NYU v2 dataset (×8 case). PMBANet* denotes the retrained version.

Methods	MAD	RMSE	Running Time
PMBANet* [24]	0.33	2.75	**13.6ms**
PMBANet*+LDE	0.31	2.72	18.8ms
PMBANet*+LDE+HAG	**0.30**	**2.58**	21.1ms



(a) (b) (c) (d)

Fig. 11: Visual examples of adding our modules to PMBANet. (a) GT. (b) PMBANet*. (c) PMBANet*+LDE. (d) PMBANet*+LDE+HAG.

datasets. For example, on the Middlebury 2005 dataset, with both LDE and HAG modules, compared with the original PMBANet, the average MAD value is improved from 0.33 to 0.30. On the NYU v2 dataset, introducing both the LDE and HAG modules reduces the RMSE of PMBANet from 2.75 to 2.58 with a percentage gain of 6.2%. Some visual examples of adding our modules to PMBANet are shown in Fig. 11. We can see that, compared with the original result of PMBANet, the PMBANet model with our LDE and HAG modules improves the accuracy and details of the reconstruction, such as the hole regions. These experiments all again demonstrate the effectiveness and portability of our proposed hierarchical color guidance mechanism.

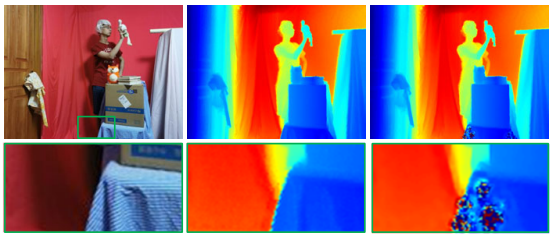
D. Discussion

For the training time, it takes about 24 hours to obtain the final ×8 model for 200 epochs with a batch size of 8. And our ×4 model only requires about 10 hours for training with a batch size of 8. This suggests that the training time is correlated with both our specific task and batch size settings, thus existing methods focus more on the running time of the test. The running time comparisons on the NYU v2 dataset (640 × 480) are shown in Table VIII. It can be seen that our method processes ×8 DSR task in less than 20ms per image on the NYU v2 dataset (640 × 480) while achieving optimal performance compared to other SOTA methods. In other words, our model achieves a good balance in terms of performance and efficiency.

In terms of performance, the advantages are more pronounced when dealing with more complex scenes or larger upsampling factors. But as reported in Table III and Table IV, our method has no obvious advantages when dealing with low-scale (×4) DSR on the Lu dataset and real-world RGB-D-D dataset. Some failure cases under the ×4 factor on the RGB-D-D dataset are also provided in Fig. 12. It has been observed that our algorithm may suffer from reconstruction errors when confronted with regions of sudden changes in brightness inside objects in the low-depth range of the scene.

TABLE VIII: Average running time and RMSE (↓) of different methods on the NYU v2 dataset ($\times 8$ case).

Methods	RMSE	Running Time
PMBANet* (TIP'20) [24]	2.75	**13.6ms**
DKN (IJCV'21) [19]	3.26	81.1ms
JIIF (ACMMM'21) [67]	2.76	132.0ms
DCT (CVPR'22) [68]	3.21	46.5ms
Ours	**2.53**	19.5ms



(a) Color images (b) GT (c) Ours

Fig. 12: Illustration of the failure cases under the $\times 4$ factor on the RGB-D-D dataset.

In these regions, the brightness boundary of the color image varies greatly, but the depth variation is small, so the inconsistency between the color brightness and depth boundaries may cause reconstruction errors in small regions. Therefore, the robustness of the model to such challenging scenarios needs to be further optimized in the future.

V. Conclusion

In this paper, we rethink the utilization of color information in DSR and propose a novel framework HCGNet. On the one hand, to supplement the high-frequency color information for the depth features, we embed the low-level detail features through the LDE module; On the other hand, to maintain semantic consistency in the reconstruction process, we encode the global abstract guidance information in the HAG module. In addition, we design the AFP module to make full use of multi-scale information while projecting effective information for reconstruction in an attention manner. Experiments on four benchmark datasets demonstrate that the proposed network outperforms other state-of-the-art methods both qualitatively and quantitatively.

References

[1] H. Hu, T. Zhao, Q. Wang, F. Gao, L. He, and Z. Gao, "Monocular 3-D vehicle detection using a cascade network for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[2] Z. Pan, J. Hou, and L. Yu, "Optimization RGB-D 3D reconstruction algorithm based on dynamic SLAM," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[3] C. Gu, Y. Cong, and G. Sun, "Three birds, One stone: Unified laser-based 3D reconstruction across different media," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[4] S. Liu, R. Wu, J. Qu, and Y. Li, "HDA-Net: Hybrid convolutional neural networks for small objects recognization at airports," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[5] A. Kosuge, S. Suehiro, M. Hamada, and T. Kuroda, "mmWave-YOLO: A mmwave imaging radar-based real-time multiclass object recognition system for ADAS applications," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022.

[6] R. Cong, Q. Qin, C. Zhang, Q. Jiang, S. Wang, Y. Zhao, and S. Kwong, "A weakly supervised learning framework for salient object detection via hybrid labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 534–548, 2023.

[7] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6800–6815, 2022.

[8] H. Wen, C. Yan, X. Zhou, R. Cong, Y. Sun, B. Zheng, J. Zhang, Y. Bao, and G. Ding, "Dynamic selective network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 9179–9192, 2021.

[9] C. Zhang, R. Cong, Q. Lin, L. Ma, F. Li, Y. Zhao, and S. Kwong, "Cross-modality discrepant interaction network for RGB-D salient object detection," in *Proc. ACM Int. Conf. Multim. (ACM MM)*, 2021, pp. 2094–2102.

[10] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, and S. Kwong, "Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection," in *Proc. ACM Int. Conf. Multim. (ACM MM)*, 2023, pp. 406–416.

[11] Y. Mao, Q. Jiang, R. Cong, W. Gao, F. Shao, and S. Kwong, "Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2435–2448, 2022.

[12] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "RRNet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–11, 2022.

[13] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, and S. Kwong, "Does Thermal really always matter for RGB-T salient object detection?" *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2022.3216476.

[14] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 88–100, 2021.

[15] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. European Conf. Comput. Vis. (ECCV)*, 2020, pp. 225–241.

[16] X. Song, Y. Dai, D. Zhou, L. Liu, W. Li, H. Li, and R. Yang, "Channel attention based iterative residual learning for depth map super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5630–5639.

[17] X. Song, D. Zhou, W. Li, Y. Dai, L. Liu, H. Li, R. Yang, and L. Zhang, "WAFP-Net: Weighted attention fusion based progressive residual learning for depth map super-resolution," *IEEE Trans. Multim.*, vol. 24, pp. 4113–4127, 2022.

[18] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, and X. Fan, "Color-guided depth map super resolution using convolutional neural network," *IEEE Access*, vol. 5, pp. 26 666–26 672, 2017.

[19] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 579–600, 2021.

[20] P. Liu, Z. Zhang, Z. Meng, N. Gao, and C. Wang, "PDR-Net: Progressive depth reconstruction network for color guided depth map super-resolution," *Neurocomputing*, vol. 479, pp. 75–88, 2022.

[21] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 297–306, 2020.

[22] Q. Tang, R. Cong, R. Sheng, L. He, D. Zhang, Y. Zhao, and S. Kwong, "BridgeNet: A joint learning network of depth map super-resolution and monocular depth estimation," in *Proc. ACM Int. Conf. Multim. (ACM MM)*, 2021, pp. 2148–2157.

[23] L. Zhao, H. Bai, J. Liang, B. Zeng, A. Wang, and Y. Zhao, "Simultaneous color-depth super-resolution with conditional generative adversarial networks," *Pattern Recognit.*, vol. 88, pp. 356–369, 2019.

[24] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, "PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 7427–7442, 2020.

[25] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, 2019.

[26] X. Cao, Y. Luo, X. Zhu, L. Zhang, Y. Xu, H. Shen, T. Wang, and Q. Feng, "DAEANet: Dual auto-encoder attention network for depth map super-resolution," *Neurocomputing*, vol. 454, pp. 350–360, 2021.

[27] D. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 275–292.

[28] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng, "Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 994–1006, 2019.

[29] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, pp. 1–8.

[30] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1664–1673.

[31] W. Hsu and P. Jian, "Detail-enhanced wavelet residual network for single image super-resolution," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[32] S. Ahmadi, L. Kästner, J. C. Hauffen, P. Jung, and M. Ziegler, "Photothermal-SR-Net: A customized deep unfolding neural network for photothermal super resolution imaging," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[33] W. Shi, F. Tao, and Y. Wen, "Structure-aware deep networks and pixel-level generative adversarial training for single image super-resolution," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[34] Y. Zhu, S. Wang, Y. Zhang, Z. He, and Q. Wang, "A dual transformer super-resolution network for improving the definition of vibration image," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[35] J. Wu, R. Cong, L. Fang, C. Guo, B. Zhang, and P. Ghamisi, "Unpaired remote sensing image super-resolution with content-preserving weak supervision neural network," *Sci. China Inf. Sci.*, vol. 66, no. 1, 2023.

[36] F. Li, Y. Wu, H. Bai, W. Lin, R. Cong, and Y. Zhao, "Learning detail-structure alternative optimization for blind super-resolution," *IEEE Trans. Multim.*, vol. 25, pp. 2825–2838, 2023.

[37] G. Riegler, M. Rüther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 268–284.

[38] X. Song, Y. Dai, and X. Qin, "Deeply supervised depth map super-resolution as novel view synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2323–2336, 2019.

[39] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, "Depth super-resolution via deep controllable slicing network," in *Proc. ACM Int. Conf. Multim. (ACM MM)*, 2020, pp. 1809–1818.

[40] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, 2007.

[41] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, 2013.

[42] J. Wang, L. Sun, R. Xiong, Y. Shi, Q. Zhu, and B. Yin, "Depth map super-resolution based on dual normal-depth regularization and graph laplacian prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3304–3318, 2022.

[43] L. Huang, J. Zhang, Y. Zuo, and Q. Wu, "Pyramid-structured depth MAP super-resolution based on deep dense-residual network," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1723–1727, 2019.

[44] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 636–644.

[45] Y. Huang, F. Zheng, R. Cong, W. Huang, M. R. Scott, and L. Shao, "MCMT-GAN: multi-task coherent modality transferable GAN for 3D brain image synthesis," *IEEE Trans. Image Process.*, vol. 29, pp. 8187–8198, 2020.

[46] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[47] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 993–1000.

[48] J. Xie, R. S. Feris, and M. Sun, "Edge-guided single depth image super resolution," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 428–438, 2016.

[49] M. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 169–176.

[50] J. Xie, C. Chou, R. S. Feris, and M. Sun, "Single depth image super resolution and denoising via coupled dictionary learning with local constraints and shock filtering," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2014, pp. 1–6.

[51] R. de Lutio, S. D'Aronco, J. D. Wegner, and K. Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8828–8836.

[52] J. Wang, W. Xu, J. Cai, Q. Zhu, Y. Shi, and B. Yin, "Multi-direction dictionary learning based depth map super-resolution with autoregressive modeling," *IEEE Trans. Multim.*, vol. 22, no. 6, pp. 1470–1484, 2020.

[53] X. Ye, X. Duan, and H. Li, "Depth super-resolution with deep edge-inference network and edge-guided depth filling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 1398–1402.

[54] Y. Li, J. Huang, N. Ahuja, and M. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 154–169.

[55] B. Sun, X. Ye, B. Li, H. Li, Z. Wang, and R. Xu, "Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7792–7801.

[56] Y. Zuo, H. Wang, Y. Fang, X. Huang, X. Shang, and Q. Wu, "MIG-Net: Multi-scale network alternatively guided by intensity and gradient features for depth map super-resolution," *IEEE Trans. Multim.*, vol. 24, pp. 3506–3519, 2022.

[57] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 746–760.

[58] L. He, H. Zhu, F. Li, H. Bai, R. Cong, C. Zhang, C. Lin, M. Liu, and Y. Zhao, "Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9229–9238.

[59] S. Lu, X. Ren, and F. Liu, "Depth enhancement via low-rank matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3390–3397.

[60] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1-3, pp. 7–42, 2002.

[61] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, pp. 1–8.

[62] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. 36th German Conf. Pattern Recognit.*, 2014, pp. 31–42.

[63] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, pp. 1–8.

[64] Y. Li, J. Huang, N. Ahuja, and M. Yang, "Joint image filtering with deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1909–1923, 2019.

[65] B. Ham, M. Cho, and J. Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 192–207, 2018.

[66] J. Pan, J. Dong, J. S. J. Ren, L. Lin, J. Tang, and M. Yang, "Spatially variant linear representation models for joint filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1702–1711.

[67] J. Tang, X. Chen, and G. Zeng, "Joint implicit image function for guided depth super-resolution," in *Proc. ACM Int. Conf. Multim. (ACM MM)*, 2021, pp. 4390–4399.

[68] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister, "Discrete cosine transform network for guided depth map super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 5697–5707.

[69] C. Li, J. Guo, B. Wang, R. Cong, Y. Zhang, and J. Wang, "Single underwater image enhancement based on color cast removal and visibility restoration," *J. Electronic Imaging*, vol. 25, no. 3, p. 033012, 2016.

[70] J. Hu, Q. Jiang, R. Cong, W. Gao, and F. Shao, "Two-branch deep neural network for underwater image enhancement in HSV color space," *IEEE Signal Process. Lett.*, vol. 28, pp. 2152–2156, 2021.

[71] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2472–2481.