IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 July 2024, accepted 21 July 2024, date of publication 29 July 2024, date of current version 27 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435504

RESEARCH ARTICLE

Latent Edge Guided Depth Super-Resolution Using Attention-Based Hierarchical Multi-Modal Fusion

HUI LAN AND CHEOLKON JUNG, (Member, IEEE)
School of Electronic Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Cheolkon Jung (zhengzk@xidian.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62111540272.

ABSTRACT Color guided depth super-resolution (SR) aims to reconstruct a high-resolution (HR) depth image from a low-resolution (LR) one guided by its paired HR color image. However, when the sampling factor is large, color guided depth SR suffers from reconstructing accurate depth edges due to the severe loss of high frequency (HF) components. In this paper, we propose a latent edge guided depth SR network using attention-based hierarchical multi-modal fusion, named LEDSRNet. We extract the hierarchical multi-modal features from HR color and LR depth images, and perform selective fusion to estimate the residual map for depth SR. Firstly, we perform gradient map estimation to generate accurate depth edges from the input HR color image and the interpolated LR depth image, and filter out unnecessary edges in the HR color image while preventing texture copying artifacts in depth SR. Then, we perform depth upsampling to get depth edges from the input LR depth image and refine them guided by gradient features in the latent space. Moreover, we fuse the features extracted from gradient map estimation and depth upsampling to obtain the residual map for depth SR. Finally, we reconstruct SR depth image by adding the residual map to the interpolated LR depth image. We design an attention based multi-level residual block (AMRB) as the basic block for LEDSRNet to extract both shallow and deep features in color and depth images for hierarchical multi-modal fusion. In the loss function, we use a binarized gradient map from the ground truth depth image, i.e. mask map, to calculate the loss for edge and smooth areas separately, preventing excessive smoothing of edge regions in the reconstructed SR depth image. Extensive experiments show that LEDSRNet reconstructs accurate depth edges even in the large sampling factor and achieves the best performance in RMSE with low running time and small model parameters. They indicate that LEDSRNet outperforms state-of-the-art methods in terms of both visual quality and quantitative measurements.

INDEX TERMS Depth super-resolution, attention, gradient estimation, latent edge, mask map, multi-modal fusion.

I. INTRODUCTION

Depth image has been widely used in scene reconstruction [1], robotics [2], and autonomous driving [3]. However, the common depth cameras such as Microsoft kinect and Lidar cannot obtain high quality and high resolution (HR) depth images (e.g., the resolution of depth images acquired by Kinect 2.0 is only 512 × 424) [4]. It is required to reconstruct super-resolution (SR) depth images from low resolution (LR) depth images. The simplest method for depth SR is image interpolation, such as bicubic, bilinear and joint bilateral upsampling (JBU) [5]. However, the depth images obtained by such methods are usually too smooth, and it is difficult to recover high-quality and HR images, especially when the sampling factor is high. To solve this problem, some traditional methods have achieved good performance by constructing hand-crafted filters or objective functions. However, this kind of methods are usually useful for the images of specific scenes, and it is difficult to be widely used for the depth images of real scenes. Color and depth

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

represent different attributes in the same scene, and the HR color image has strong structural similarity to the LR depth image. Therefore, color guided depth SR is proposed and has achieved outstanding results.

Owing to the rapid advancements of deep learning, it has been gradually applied to the field of image SR including depth SR [6], [7], medical image SR [8], remote sensing image SR [9], thermal image SR [10], and face SR [11]. Due to its powerful feature extraction and representation ability, deep learning methods have achieved a significant advantage in improving the quality of reconstructed SR depth images. For upsampling of a single depth image, the deep learning method can estimate the corresponding SR depth image from a single LR depth image by learning the mapping relationship. Dong et al. [12] proposed a single image SR reconstruction method, called a super-resolution convolutional neural network (SRCNN), which used only three convolutional layers to map the LR feature space to the HR feature space. Since SRCNN had a relatively simple structure with a small receptive fields, its learning ability of features was limited. However, it was the first application of deep learning method to the image SR. In the color guided depth SR such as Hui et al.'s method [13], the features were extracted from the HR color image and the LR depth image, while the depth image was upsampled and reconstructed under the guidance of the HR color image features. However, not all the features in the HR color image are beneficial to the depth SR because the color image also contains its complex textures. If the useful and useless textures in the color image cannot be effectively distinguished, it is easy to cause texture copying artifacts in the color guided depth SR.

The existing methods have three main problems that need to be further treated: 1) Existing methods usually use the LR depth images or the interpolated LR depth images as input of the proposed network, ignoring that both LR depth image and interpolated LR depth image contribute positively to depth SR. 2) There is no target solution to the problem of texture copying artifacts, resulting in the inability to effectively filter the useless edge information of the color image when the sampling factor is large. 3) Although the edge information for depth SR is mostly from the color image, most methods suffer from selecting valid depth edges from the color image.

In this paper, we propose a latent edge guided depth SR network using attention-based hierarchical multi-modal fusion, named LEDSRNet. Different from the existing methods, the proposed method fully extracts the edge features in the latent space from HR color image and the interpolated LR depth image to estimate a fine gradient map. The LR depth image is upsampled with the guidance of the latent edge features to further refine the depth edges and generate the SR depth image. LEDSRNet consists of three subnetworks: gradient estimation, LR depth upsampling and fusion. Specifically, we first convert the HR color image to the gray scale, and concatenate the interpolated LR depth image as the input of the gradient estimation subnetwork. We use an encoder-decoder structure [14] to extract multi-scale texture features to estimate an accurate edge map. The interpolated LR depth image is to provide depth structure information and filter out unwanted edge details in the color image. The LR depth upsampling subnetwork is guided by the decoder of the gradient estimation subnetwork. During LR depth upsampling, the high frequency (HF) details for depth SR are further refined. Then, we use the fusion subnetwork to fully fuse the multi-modal features extracted from gradient estimation and LR depth upsampling to obtain the residual map between the interpolated LR depth image and the corresponding HR one. Finally, we reconstruct the SR depth image by adding the learned residual map to the interpolated LR depth image. Experimental results demonstrate that LEDSRNet outperforms the state-of-the-art methods for depth SR in terms of root mean square error (RMSE), peak signal to noise ratio (PSNR), and mean absolute difference (MAD). Fig. 1 shows the whole network architecture of the proposed LEDSRNet.

In our previous work [15], we proposed a depth SR network guided by blurry depth and clear intensity edges, named DSRNet. DSRNet distinguished effective edges from a number of HR edges with the guidance of blurry depth and clear intensity edges, thus successfully reconstructing depth edges in SR results. The key idea of DSRNet is to extract and fuse the color and depth features for the SR depth reconstruction. However, DSRNet mainly focused on extracting the features of depth image by taking the features of color image as supplementary information. It could not effectively deal with the redundant edges in HR color image, thus causing the texture copying artifacts in depth SR. According to careful observations and analysis, the edge information provided by depth image is very limited, especially when the sampling factor is large, and most clear edge information comes from HR color image. Thus, we take the gradient estimation subnetwork as the backbone to extract a depth edge map from HR color image and interpolated LR depth image. We use the interpolated LR depth image to remove the redundant edges in the HR color image. Guided by the edge features, the LR depth image is gradually upsampled and finally the SR depth image is reconstructed. Fig. 2 illustrates the feature maps extracted at different scales when the sampling factor is 4. It can be observed that the feature maps obtained from the encoder of the gradient estimation subnetwork contain a large amount of redundant edges and structures. After the first convolution layer and attention-based multi-level residual block (AMRB), a large number of edges are extracted from the HR color image that contains redundant ones. As the network deepens, the redundant edge details are filtered out. On the decoder side, the features gradually approach the content of the gradient map. In the LR depth upsampling subnetwork, only shallow depth structure information is extracted at the first part. Under the guidance of the gradient estimation subnetwork, the initial structural features are upsampled step

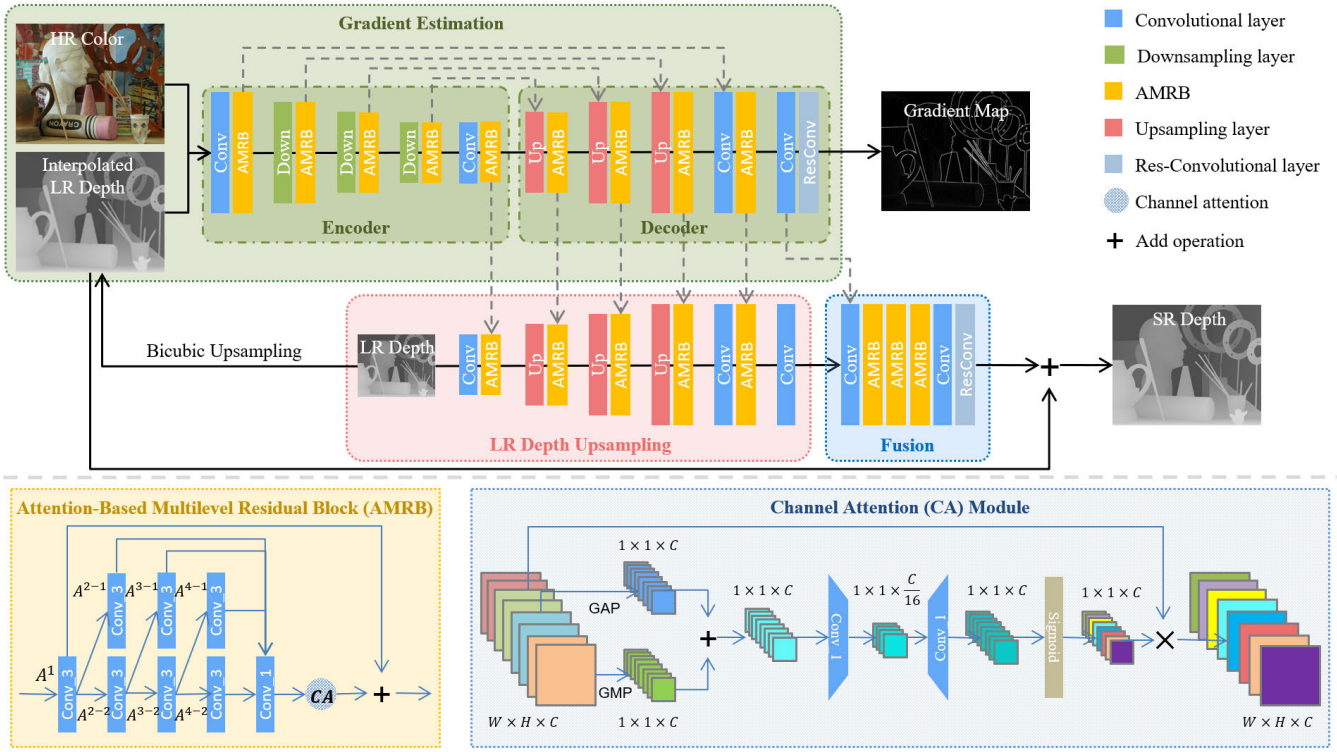

FIGURE 1. Whole architecture of the proposed latent edge guided depth super-resolution network (LEDSRNet). For ease to representation, we illustrate LEDSRNet with sampling factor ×8. LEDSRNet consists of three subnetworks: gradient estimation, LR depth upsampling and fusion. The hierarchical multi-modal features extracted from color and depth images are concatenated, while the output depth image is obtained by adding the residual map and interpolated LR depth image. The mask map is used to preserve depth edges in the loss calculation.

by step and the accurate edge information is generated. After the fusion subnetwork, the edge features are converted into residual features required for depth SR. Therefore, LEDSRNet processes the edge information of the depth image in multiple steps, filters out unnecessary details, and further reconstructs an accurate SR depth image.

Compared with existing methods, the main contributions of LEDSRNet are as follows:

- We propose a latent edge guided depth SR network using attention-based hierarchical multi-modal fusion (LEDSRNet). We extract the hierarchical multi-modal features from HR color and LR depth images, and perform selective fusion in the latent space to estimate the residual map for depth SR. Based on the selective fusion through an attention module, the residual connection in each block enables LEDSRNet to learn the HF details for depth SR while ignoring the smooth region.
- We present a simple yet effective attention-based multi-level residual block (AMRB) as the basic block. AMRB takes the advantage of residual connection, feature reuse, and parallel multi-layer convolution. In AMRB, the shallow convolutional layer extracts texture and edge information, while the deep convolutional layer extracts deep semantic features in a larger receptive field. Thus, through the multi-layer parallel convolution, AMRB can extract more multilevel features than residual block with feature reuse, which is less complex than dense blocks. Since channel attention integrates features of different receptive fields, assign weights adaptively, and retain effective features, AMRB filters out useless edges by learning effective features, thus predicting fine edge details for depth SR.
- We gradually filter out redundant edge information by gradient estimation, LR depth upsampling, and fusion, thus effectively preventing texture copying artifacts in depth SR.
- We introduce a mask map, i.e. binarized gradient map from the ground truth depth image, in the loss function to calculate the loss for edge and smooth areas separately. The mask map enables LEDSRNet to preserve depth edges in the loss calculation, thus preventing edge smoothing in depth SR and generating accurate edge map and SR depth image.

The rest of this paper is organized as follows. Section II reviews the related work on depth SR. Section III describes LEDSRNet in detail including the network architecture and implementation details. The experimental results are provided in Section IV, and finally Section V concludes this paper.

II. RELATED WORK

According to the additional input, we divide related work into two categories: Single depth SR and color guided depth SR.

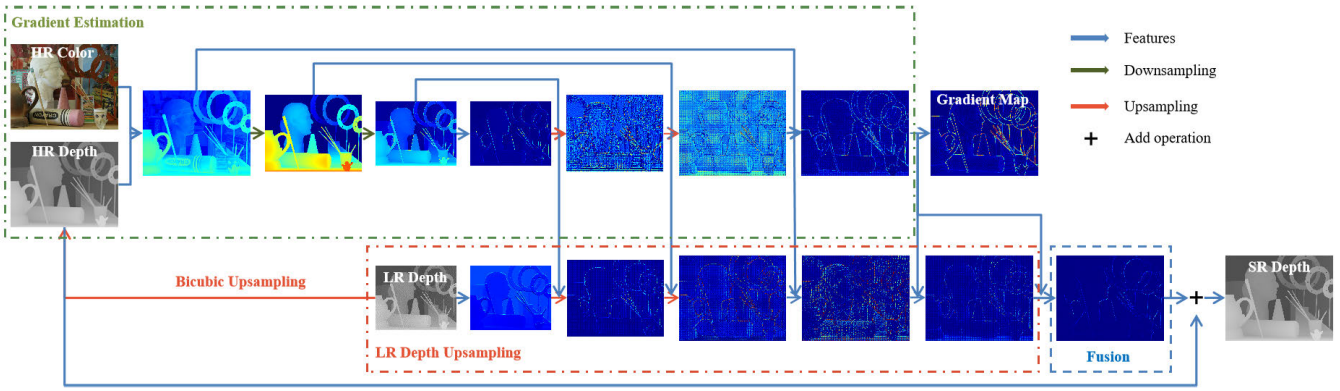

FIGURE 2. Illustration of feature maps at different scales in LEDSRNet when the sampling factor is 4. The feature maps of the encoder side in the gradient estimation subnetwork contain a large amount of redundant edges and structures, while those of the decoder side gradually extracts the precise depth edge by removing unnecessary ones. Guided by the edge features, the LR depth upsampling subnetwork gradually recovers residual features.

Since LEDSRNet is based on deep learning, we review them focusing on deep learning-based depth SR.

A. SINGLE DEPTH SR

Single depth SR is similar to single image SR (SISR) since it only extracts feature details from LR image and restores the corresponding HR image. However, due to the feature difference between depth image and color image, it may not be possible to get excellent by applying SISR methods directly to the depth image. Xie et al. [16] used a Markov Random Field (MRF) to construct an HR edge map from the LR depth edge map. With the guidance of the HR edge map, the LR depth image was upsampled via a joint bilateral filter. However, traditional filter-based methods often failed to recover complex edges. Dong et al. [17] proposed a deep convolutional network for image SR that consisted of three convolutional layers. Huang et al. [18] presented pyramid-structured depth SR based on residual dense blocks that used dense connection layers and residual learning to model the mapping between high-frequency (HF) residuals and LR depth image. Jiang et al. [19] proposed a hierarchical dense block (HDB) for image SR to estimate residual feature maps in a coarse-to-fine manner. They extracted texture features through multiple parallel interlacing dense blocks, and generated the final SR image by fusing multiple HDB features. Fang et al. [20] firstly estimated the soft edges of the HR image and the rough HR image at the same time, and then used a fusion network, which cascaded with multiple residual blocks to fully fuse the soft edges and rough HR images obtained in the first stage to generate clear SR image. To fully explore the mapping relationship between LR and HR features, Wu et al. [21] used iterative upsampling and downsampling operations to construct a deep feedback mechanism by projecting HR representation to the LR spatial domain and then back-projecting to the HR spatial domain. The deep feedback block imitates the process of image degradation and reconstruction iteratively. The attention mechanism brings an opportunity for selective fusion of increasingly complex features. Zamir et al. [22] proposed attention based multi-resolution feature aggregation that received complementary contextual information from LR and HR representations, and gradually refined edge details in HR image. Li et al. [23] extracted multi-scale features by different receptive field filters, and incorporated channel shuffle into the attention mechanism to get the relationship between the feature channels and improve the feature selection capacity. Chai et al. [24] proposed transformer branch and convolution branch to extract long-range and short-range dependencies, respectively, thus extracting rich and heterogeneous features from two branches. Liu et al. [25] proposed a blind image SR method that combined CNN and transformer. The contrast learning was incorporated into the transformer network to learn the degeneration representation of an image with unknown noise. Shi et al. [26] constructed a multi-scale parallel face reconstruction network that combined local pixel attention and global transformer attention. Ye et al. [27] proposed a slice-based single depth SR network to realize arbitrary sampling factors. Specifically, the depth image was divided into several slices according to the depth of the scene, and each slice was refined by the depth features of different scales. Finally, each slice was adaptively weighted by the distance-aware weighting network to obtain the final output. Zamir et al. [28] further presented a multi-stage network architecture that progressively learned restoration functions for the degraded inputs and divided the restoration process into manageable steps. In the multi-stage architecture, a key component is the information exchange among different stages. Jiang et al. [29] leveraged wavelet transformation to decompose the features into LF and HF components and then employed two different branches to separately process them and reconstruct LR and HF components. Then, they recover SR image by the inverse discrete wavelet transformation (IDWT).

Such methods are suitable for depth SR without HR color guidance, which can usually achieve good performance when the sampling factor is small. However, when the sampling factor is large, it is difficult to recover high-quality SR depth images only from LR depth images.

B. COLOR GUIDED DEPTH SR

Compared with single depth SR, color guided depth SR (CDSR) can obtain more edge information from HR color image. However, it is worth noting that not all the edges of color image are useful for depth SR. Xu et al. [30] proposed a depth SR method using multi-directional dictionary learning joint local gradient and nonlocal structural regularization. They classified depth patches according to the geometrical directions and learned a compact online dictionary for depth SR. Li et al. [31] proposed a multi-scale guidance feature extraction branch and a depth estimation branch for depth SR. They used a correlation-controlled color guidance block to fuse multi-modal features in each scale. Zhao et al. [32] proposed a simultaneous color-depth SR method in 3D videos to generate high-quality color-depth images from the low-quality ones. Hui et al. [13] proposed a residual based multi-scale guidance network for depth SR. They extracted the multi-scale features from LR depth image and HR color image, respectively. Guided by the features of color image, the depth features are upsampled in each scale. The multi-level feature fusion has inspired and affected many subsequent works. Guo et al. [33] used an encoder-decoder structure to extract multi-level features from the interpolated upsampled depth image and estimate the residual map. At each scale of the encoder part, the features of the depth image downsampled by the pooling layer were fused separately to supplement the depth details. At the decoder part, the edge details were supplemented by fusing the multi-level features of the color image. Wang et al. [34] proposed deep edge-aware learning for depth SR. They first fused the features extracted from HR color and LR depth images to produce a clear edge map. Then, they introduced traditional and learning based restoration modules to solve the information loss when the sampling factor is large. Liu et al. [35] proposed a progressive depth reconstruction method to improve the accuracy of SR depth image. The feature representation of each sampling factor was obtained by fully recombining the extracted depth feature and color image feature with the adaptive feature recombination model and the joint attention mechanism. Under the supervision of multi-scale loss function, the multi-scale SR depth image was generated. Wang et al. [36] proposed a multi-scale feedback network (MSF-Net) for guided depth SR to effectively extract and refine multi-scale features by multi-scale feedback learning. Ye et al. [37] proposed a progressive multi-branch aggregation network. They designed attention-based error feed-forward/-back modules to iteratively estimate the lost high-frequency information and refine the depth image. Meanwhile, HR color image was used to complement the texture details required. Metzger et al. [38] combined guided anisotropic diffusion with a deep convolutional network for guided depth SR. Anisotropic diffusion is a form of iterative edge aware filtering, which achieves good performance for a large sampling factor. Mehri et al. [39] introduced transformer into CDSR to generate local attention weights from LR depth and HR color images, and used a multi-level fusion module to generate global attention map to highlight the edge information for depth upsampling. They implemented the depth upsampling at arbitrary sampling factors. Ariav and Cohen [40] proposed a two branch fully transformer-based depth SR network. Color branch generates HR edge weights to be a global guidance for depth SR, while depth branch learns edge details by a cross-attention guidance module with the help of color features. Chai et al. [41] proposed a binocular image reconstruction network that used a dynamic convolution pyramid to extract local features of binocular images and capture global context information by cross-view transformer. Zhao et al. [42] and Metzger et al. [38] distinguished valid edges from useless ones in color image based on spherical space mapping and anisotropic diffusion.

III. PROPOSED METHOD

A. PROBLEM FORMULATION

LEDSRNet aims at achieving the LR depth image upsampling by estimating a depth edge map from the HR color image and the interpolated LR depth image. We combine edge features in gradient estimation and depth features in LR depth upsampling to generate an accurate residual map and add it with the interpolated LR depth image for SR reconstruction. Given an HR color image $C^H \in R^{(\rho h \times \rho w \times 3)}$, we convert it into grayscale, which is denoted as $G^H \in R^{(\rho h \times \rho w \times 1)}$. Denote HR depth image (ground truth) as $D_{GT} \in R^{(\rho h \times \rho w \times 1)}$, LR depth image as $D^L \in R^{(h \times w \times 1)}$, and the interpolated LR depth image as $D^{IL} \in R^{(\rho h \times \rho w \times 1)}$. We obtain D^L from D_{GT} by bicubic downsampling, where $\rho \geq 1$ is the upscaling factor (e.g., 2, 4, 8 and 16). We denote the generated residual map as $R^H \in R^{(\rho h \times \rho w \times 1)}$ and the final SR depth image as $D_{SR} \in R^{(\rho h \times \rho w \times 1)}$. The ground truth of gradient estimation subnetwork is the gradient map of D_{GT} which is denoted as $E_{GT} \in R^{(\rho h \times \rho w \times 1)}$. Since Sobel operator takes advantages of fast computation speed and anti-interference ability to noise, we use it to get the gradient map as follows:

$$E_{GT} = Sobel(D_{GT}) \tag{1}$$

where $Sobel(\cdot)$ denotes Sobel operation. LEDSRNet is based on the residual learning to learn the lost high frequency (HF) component in bicubic interpolation upsampling.

B. NETWORK ARCHITECTURE

As shown in Fig. 1, LEDSRNet reconstructs SR depth image with clear edges through three collaborative stages: Gradient estimation, LR depth upsampling and fusion.

1) ATTENTION-BASED MULTI-LEVEL RESIDUAL BLOCK

We use an attention-based multi-level residual block (AMRB) as basic block of LEDSRNet. The shallow features of convolutional neural networks (CNNs) usually contain local features such as textures and edges, while the deep features contain mostly semantic information. We take the advantages of dense block and residual block to construct AMRB. Under the limited parameters, AMRB reuses the deep and shallow

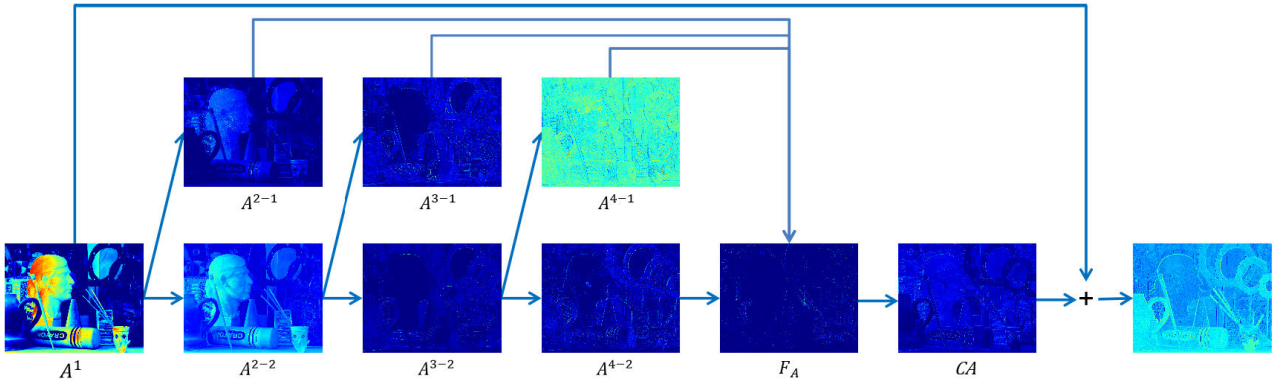

FIGURE 3. Example of the feature layers extracted by the attention-based multi-level residual block (AMRB). AMRB effectively extracts and fuses shallow structural information and deep semantic information for edge estimation. The network structure of AMRB is shown in Fig. 1.

features well, while effectively dealing with the problem of gradient disappearance. The attention module can assign a larger weight to important regions and a smaller weight to unimportant ones. As shown in Fig. 1, AMRB contains five parts with a channel attention module. The first part of AMRB is a convolutional layer to extract the initial features. The next three parts contain two convolutional layers for deeper features extraction. We use the output of one convolutional layer as input to the next part, while another convolutional layer is used to preserve the features of the current deep. The last part is a 1×1 convolutional layer, which is used to fuse the output features of all convolutional layers from the second part to the fourth part, realize the feature reuse of different receptive fields, and compress 256 channels to 64. The channel attention module is used to preserve more important features during channel compression. Finally, we use a global residual connection to learn the HF information and ignore the smoothing information that already exists. Fig. 3 shows the feature layers extracted by AMRB. The shallow convolutional layers extract local structural features, while the deep convolutional layers extract global semantic features. After fusing the shallow and deep features, the channel attention module can effectively select valid edge information for depth SR.

AMRB is expressed as follows:

$$A^1 = Conv(Input) \tag{2}$$
$$A^{2-1} = Conv(A^1) \tag{3}$$
$$A^{2-2} = Conv(A^1) \tag{4}$$
$$A^{3-1} = Conv(A^{2-2}) \tag{5}$$
$$A^{3-2} = Conv(A^{2-2}) \tag{6}$$
$$A^{4-1} = Conv(A^{3-2}) \tag{7}$$
$$A^{4-2} = Conv(A^{3-2}) \tag{8}$$
$$Output = c(A^{4-2}, A^{4-1}, A^{3-1}, A^{2-1}) \tag{9}$$
$$Output = Ca(Conv(Output)) + A^1 \tag{10}$$

where *Input* and *Output* are the input and output features of AMRB; A^1 is the initial feature extracted by the first part of AMRB; A^{i-1} and A^{i-2}, $i \in 2, 3, 4$ are the features obtained by the second part to the fourth part of AMRB; and $Ca(\cdot)$ denoted as the channel attention module.

2) GRADIENT ESTIMATION

We design a U-Net based structure with skip connections to extract a set of hierarchical gradient features from HR color image and interpolated LR depth image, and generate a clear gradient map. We provide the gradient maps generated by the gradient estimation subnetwork in Fig. 4. It can be observed that the gradient estimation subnetwork generates clear edges even in $\times 8$, i.e. a large sampling factor. C^H contain a lot of clear but redundant edge information, and D^{IL} can provide rough depth edge reference to prevent texture copying artifacts. We convert C^H into intensity scales G^H to remove unnecessary color information and concatenated with D^{IL} as input of this subnetwork. As shown in Fig. 1, the encoder branch contains five parts to extract features of different receptive fields. The first part has one convolutional layer and one AMRB, which is used to extract initial features with the original resolution. The next three parts have the same structure that consists of one downsampling layer and one AMRB to extract multi-scale semantic features. In this work, we use a convolutional layer with stride 2 for downsampling. The last layer consists of one convolutional layer with stride 1 and one AMRB that further integrates the LR features, which is consistent with the feature extraction of the LR depth upsampling subnetwork. The encoder branch can be expressed as follows:

$$Conv(\cdot) = \sigma(W * Input + b) \tag{11}$$
$$F_g e^1 = AMRB(Conv(c(G^H, D^{IL})) \tag{12}$$
$$F_g e^{i+1} = AMRB(Downsampling(F_g e^i)) \tag{13}$$
$$F_g e^5 = AMRB(Conv(F_g e^4)) \tag{14}$$

where W and b stand for the weight and bias in the first convolutional layer, respectively; $*$ represents convolution operation; σ is the element-wise rectified linear unit (ReLU) activation function; $Conv(\cdot)$ represents the convolutional layer; $c(\cdot)$ means the concatenation operation; $AMRB(\cdot)$ is

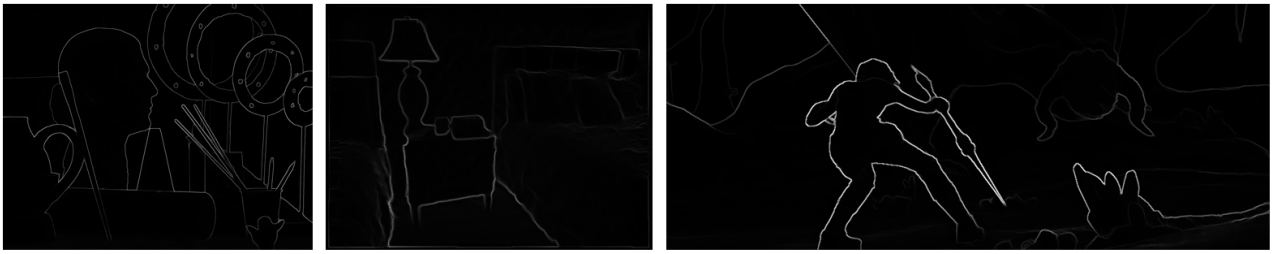

FIGURE 4. Gradient maps generated by the gradient estimation subnetwork for ×8. (a) Middlebury dataset. (b) NYU dataset. (c) MPI Sintel dataset.

the attention-based multi-level residual block (AMRB); and *Downsampling*(·) means the downsampling layer, which is a convolutional layer with kernel size 3 × 3 and stride 2. $F_g e^1$ is the features extracted from input G^H and D^{IL} and $F_g e^{i+1}$ is the features extracted from $F_g e^i$ by other layers of encoder part, in which $i \in 1, 2, 3$ when the sampling factor is ×8.

The structure of the decoder branch corresponds to the that of encoder branch and contains five parts when the sampling factor is ×8. When upsampling and fusing the multi-scale features from the encoder branch to prevent the information loss, the useless edge information is further removed to generate an accurate gradient map. The first three parts of the decoder branch consists of one upsampling layer and one AMRB. Here, we use the sub-pixel convolutional layer for upsampling [43]. After the first three parts, the edge features are upsampled to the original resolution (i.e. the same resolution as the depth ground truth). The fourth part contains one convolutional layer and one AMRB for integrating HR edge features. Then, an accurate edge map is generated through the fifth part which consists of two convolutional layers. The kernel size of all convolutional layers are 3 × 3, followed by one Relu layer except the last one. The convolutional layer without Relu operation to generate the output image is named as the residual convolutional layer. To reduce the parameters of LEDSRNet, we set the number of channels per layer to 64. However, the output channels of residual convolutional layer is determined by the output image. The encoder branch is expressed as follows:

$$F_g d^1 = AMRB(Upsampling(c(F_g e^4, F_g e^5)) \quad (15)$$
$$F_g d^2 = AMRB(Upsampling(c(F_g d^1, F_g e^3))) \quad (16)$$
$$F_g d^3 = AMRB(Upsampling(c(F_g d^2, F_g e^2))) \quad (17)$$
$$F_g d^4 = AMRB(Conv(c(F_g d^3, F_g e^1))) \quad (18)$$
$$F_g d^5 = Conv(F_g d^4) \quad (19)$$
$$E = ResConv(F_g d^5) \quad (20)$$

where *Upsampling*(·) means the upsampling layer, and $F_g d^i$, $i \in 1, 2, 3, 4, 5$ are the features obtained by the each layer in decoder branch; *ResConv*(·) is the residual convolutional layer; and E is the output gradient map of the gradient estimation subnetwork. The gradient estimation subnetwork can effectively distinguish useful edge information and reconstruct a depth edge map.

3) LR DEPTH UPSAMPLING

We have obtained a clear edge map by the gradient estimation subnetwork, but the edge map contains redundant edges that are not required for depth SR. The LR depth image is of low resolution, but it contains clear depth edge information, which plays a key role in preventing the texture copying artifacts and further removing unnecessary edges of color image. Therefore, the LR depth upsampling subnetwork extracts multi-scale depth features and estimates the residual information guided by the gradient information. As shown in Fig. 1, the structure of the LR depth upsampling subnetwork is similar to the decoder branch of gradient estimation subnetwork. In the figure, the input of the LR depth upsampling subnetwork is the LR depth image D^L. We first use a convolutional layer and an AMRB to extract initial LR depth features. Then, the multi-scale depth features are extracted from the initial LR depth by three upsampling layers, and the edge information extracted from the color image by the decoder branch of the gradient estimation subnetwork is adaptively fused at each scale. Note that the sub-pixel convolutional layer [43] is used for upsampling and each upsampling layer is followed by one AMRB to get more complex features. After the upsampling layers, we use a convolutional layer with one AMRB to further extract the SR depth feature and fuse it with the corresponding edge features. Finally, two convolutional layers are used to integrate all types of HR features and generate the final SR depth features. The kernel size of all convolutional layers are 3 × 3, and the channels of each layer is 64. In the feature extraction and fusion, the LR depth upsampling subnetwork filters out unwanted edge information in a hierarchical manner to prevent texture copying artifacts. The LR depth upsampling subnetwork is expressed as follows:

$$F_d^1 = AMRB(Conv(D^L)) \quad (21)$$
$$F_d^2 = AMRB(Upsampling(c(F_g e^5, F_d^1))) \quad (22)$$
$$F_d^3 = AMRB(Upsampling(c(F_g d^1, F_d^2))) \quad (23)$$
$$F_d^4 = AMRB(Upsampling(c(F_g d^2, F_d^3))) \quad (24)$$
$$F_d^5 = AMRB(Conv(c(F_g d^3, F_d^4))) \quad (25)$$
$$F_d^6 = Conv(Conv(F_d^5)) \quad (26)$$

where F_d^i, $i \in 1, 2, 3, 4, 5, 6$ are the features obtained by the each layer in the LR depth upsampling subnetwork.

(a) (b)

FIGURE 5. (a) Mask map of *Art* without binarization. (b) Mask map of *Art* with binarization. The binarization operation highlights the edge region of the depth image. Specifically, the mask value of the edge region is 1, otherwise the value is 0 so that the loss calculation can be performed on the edge and smooth region separately to prevent excessive smoothing of edge regions in the reconstructed SR depth image.

4) FUSION

The gradient estimation and LR depth upsampling subnetworks are used to fully extract features from G^H, D^{IL} and D^L, respectively, and generate HR edge maps and SR depth features. Then, we use a simple fusion subnetwork to combine useful information from HR edge features and depth features to generate an accurate residual map R^H. The fusion subnetwork consists of one convolutional layer whose kernel size is 1×1 to compress the 128 channels to 64, three AMRBs and one residual convolutional layer whose kernel size is 3×3. F_d^6 and $F_g d^5$ are the input of the fusion subnetwork and the output is R^H. R^H and D^H are then added to generate the final SR depth image D_{SR}. The fusion subnetwork is expressed as follows:

$$R^H = ResConv(AMRB_3(Conv(c(F_d^6, F_g d^5)))) \tag{27}$$

$$D_{SR} = R^H + D^H \tag{28}$$

where $AMRB_3(\cdot)$ means 3 consecutive AMRBs.

5) LOSS FUNCTION

L_1 and L_2 losses have been commonly used for depth SR. However, both of them average the difference between the prediction result and its ground truth in a whole image, which is not effective in considering HF components such as details and boundaries in depth. Moreover, L_2 loss is sensitive to outliers and cannot rapidly converge in early training. To solve these problems, we propose to combine the mask map from the ground truth depth with L_1 and L_2 losses in the loss function, denoted as ML_1 and ML_2, respectively. The main idea of the proposed mask loss is to use the depth edge map E_{GT} of the ground truth depth D_{GT} as mask M to constrain L_1 and L_2 losses so that the losses can be calculated separately for edge and smooth regions. Since the edge map is a vector type and its proportion is relatively small in information, it is less helpful for loss functions. Thus, we perform binarization on E_{GT} and magnify the proportion of the edges so as to increase the constraint on the image edges. The difference is shown in Fig. 5, and binarized mask map highlights important edge regions.

The original L_1 and L_2 losses are expressed as:

$$L_1(x, y) = |x - y|^1 \tag{29}$$

$$L_2(x, y) = |x - y|^2 \tag{30}$$

where the x and y are the ground truth and SR result, respectively.

The proposed ML_1 and ML_2 are expressed as follows:

$$\begin{aligned} ML_1(x, y) &= L_1(M \times x, M \times y) \\ &\quad + L_1((1 - M) \times x, \quad (1 - M) \times y) \\ ML_2(x, y) &= L_2(M \times x, M \times y) \\ &\quad + L_2((1 - M) \times x, \quad (1 - M) \times y) \end{aligned} \tag{31}$$

We use ML_1 for gradient estimation. For depth SR, ML_1 is used to speed up the convergence until epoch is 120 (i.e. $1 \leq epoch \leq 120$), then use ML_2 to generate reconstruction results until training is finished (i.e. $120 < epoch \leq 200$) as follows:

$$Loss_1 = ML_1(E_{GT}, E) \tag{32}$$

$$if\,(epoch \leq 120) \Longrightarrow Loss_2 = ML_1(D_{GT}, D_{SR}) \tag{33}$$

$$else \Longrightarrow Loss_2 = ML_2(D_{GT}, D_{SR}) \tag{34}$$

At the same time, *SSIM* loss L_{SSIM} is used to constrain the structure information of the output depth image. L_{SSIM} estimates luminance, contrast and structure simultaneously as follows:

$$L_{SSIM}(x, y) = [l(x, y)]^{\alpha}[c(x, y)]^{\beta}[s(x, y)]^{\gamma} \tag{35}$$

Luminance part:

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{36}$$

Contrast part:

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{37}$$

Structure part:

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{38}$$

For image reconstruction, higher *SSIM* is better. Thus, we define $Loss_3$ as:

$$Loss_3 = 1 - L_{SSIM}(D_{GT}, D_{SR}) \tag{39}$$

our total loss is defined as:

$$Loss = Loss_1 + Loss_2 + w \times Loss_3 \tag{40}$$

where μ_x and μ_y are means of x and y, respectively; σ_x^2 and σ_y^2 are variances of x and y, respectively; σ_{xy} is the covariance of x and y; L is the range of pixel values; $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ are the constants, and $c_3 = c_2/2$; and w is the weight for *SSIM* loss. We set $k_1 = 0.01$, $k_2 = 0.03$, and $w = 0.1$.

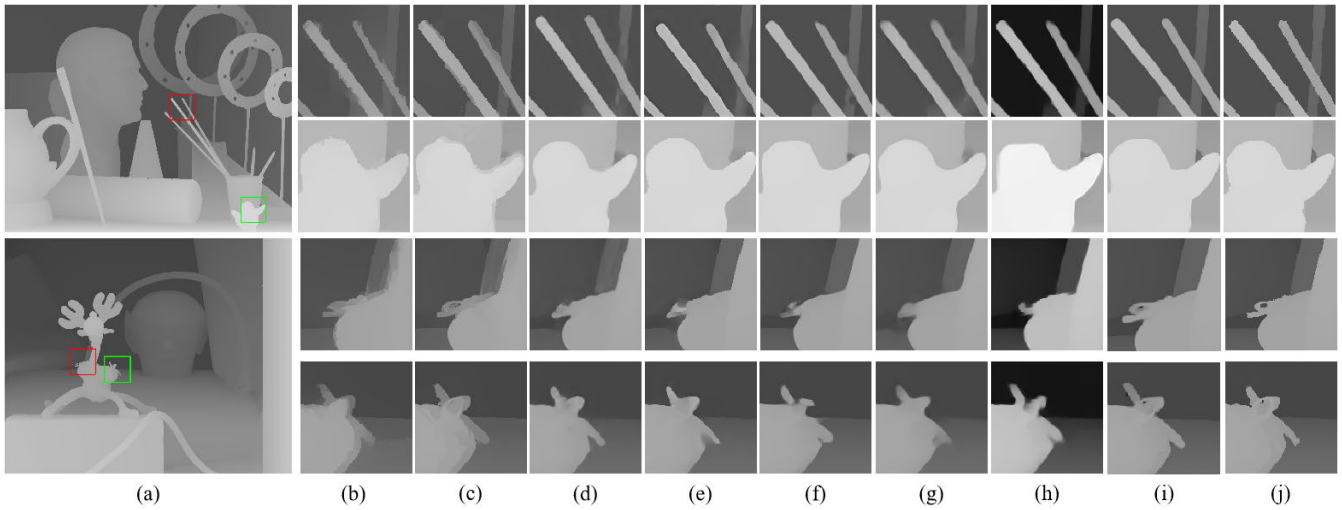

FIGURE 6. Visual comparison for ×8 on synthetic Middlebury dataset (*Art* and *Reindeer*): (a) Ground truth depth image, (b) FGI [44], (c) RGDR [45], (d) MSG [13], (e) CGN [46], (f) PMBANet [37], (g) DSR [34], (h) PDRNet [35], (i) Proposed LEDSRNet, (j) Ground truth. Depth patches were enlarged to enhance contrast for clear visualization (Top: Red box. Bottom: Green box).

TABLE 1. Quantitative depth upsampling results (in PSNR/MAD) for ×4 on synthetic Middlebury dataset. Higher PSNR and lower MAD indicate better performance. Best performance is shown in bold font.

Methods	Art	Books	Dolls	Laundry	Moebius	Reindeer	Average
Bicubic	35.14/1.82	42.83/0.97	44.55/0.86	39.10/1.12	44.10/0.95	38.00/1.24	40.62/1.16
FGI [44]	35.73/1.17	42.46/0.80	42.83/0.880	39.60/0.89	43.27/0.97	38.39/0.91	40.38/0.92
RGDR [45]	36.31/1.06	43.07/0.78	46.05/0.87	40.21/0.77	45.09/0.76	38.68/0.80	41.57/0.84
SRCNN [12]	39.92/0.63	46.82/0.27	47.12/0.29	43.86/0.40	47.16/0.28	42.72/0.35	44.60/0.37
MIRNet [22]	38.85/0.32	46.55/0.17	47.54/0.20	43.12/0.22	48.31/0.16	40.56/0.25	44.16/0.22
MPRNet [28]	39.67/0.28	46.69/0.15	47.87/0.18	43.78/0.19	48.86/0.14	40.90/0.22	44.63/0.19
MSG [13]	38.39/0.46	45.33/0.15	46.69/0.25	40.71/0.28	45.40/0.21	39.50/0.31	42.67/0.28
CGN [46]	41.52/0.29	47.38/0.19	48.58/0.23	46.33/0.20	49.24/0.19	40.93/0.22	45.66/0.22
PMBANet [37]	41.94/0.26	48.95/0.15	48.67/0.19	47.07/0.17	49.75/0.16	45.27/0.17	46.94/0.18
DSR [34]	41.56/0.28	47.96/0.11	48.40/0.14	46.19/0.18	48.58/0.17	40.86/0.21	45.59/0.18
PDRNet [35]	39.60/0.40	46.40/0.17	47.46/0.25	43.63/0.24	47.54/0.21	41.73/0.21	44.39/0.25
DAGF [47]	43.11/0.33	**50.25**/0.19	49.40/0.24	47.84/0.22	50.24/0.19	45.96/0.25	47.80/0.24
DSRNet [15]	42.45/0.17	47.97/0.11	49.24/0.13	47.57/0.11	49.82/0.11	46.62/0.12	47.28/0.12
Ours	**43.22/0.16**	48.47/**0.10**	**49.66/0.12**	**48.81/0.10**	**50.35/0.10**	**47.55/0.10**	**48.01/0.11**

TABLE 2. Quantitative depth upsampling results (in PSNR/MAD) for ×8 on synthetic Middlebury dataset. Higher PSNR and lower MAD indicate better performance. Best performance is shown in bold font.

Methods	Art	Books	Dolls	Laundry	Moebius	Reindeer	Average
Bicubic	31.87/3.67	39.60/2.51	41.48/2.52	35.56/2.83	40.64/2.67	34.47/2.74	37.27/2.82
FGI [44]	33.34/2.01	40.03/1.13	40.38/1.15	37.15/1.36	40.86/1.49	36.24/1.31	38.00/1.41
RGDR [45]	33.74/1.72	40.75/1.13	41.81/1.21	37.32/1.12	41.09/1.15	36.38/1.14	38.52/1.25
MIRNet [22]	34.49/0.85	42.24/0.45	42.93/0.52	38.97/0.65	43.27/0.43	36.22/0.52	39.69/0.57
MPRNet [28]	35.40/0.72	42.65/0.35	43.62/0.46	39.10/0.54	43.32/0.44	36.96/0.54	40.16/0.50
MSG [13]	35.62/0.69	42.79/0.36	43.47/0.46	38.19/0.65	43.32/0.43	37.48/0.52	40.15/0.51
CGN [46]	35.94/0.56	42.83/0.31	44.21/0.37	39.90/0.36	44.21/0.30	38.19/0.38	40.88/0.38
PMBANet [37]	36.15/0.61	42.88/0.26	44.61/0.32	40.31/0.34	44.49/0.26	38.47/0.34	41.15/0.36
DSR [34]	36.07/0.61	42.74/0.28	43.99/0.42	40.07/0.43	44.27/0.29	38.22/0.35	40.89/0.40
PDRNet [35]	36.77/0.60	42.88/0.27	44.67/0.30	40.82/0.32	45.33/0.23	39.19/0.30	41.61/0.34
DAGF [47]	35.99/0.60	43.59/0.30	44.74/0.39	40.15/0.39	44.61/0.30	38.83/0.37	41.32/0.39
DSRNet [15]	37.54/**0.38**	**44.06/0.22**	45.60/**0.25**	42.67/**0.25**	46.11/**0.21**	40.91/**0.24**	42.82/**0.26**
Ours	**38.08**/0.39	43.27/0.24	**46.09/0.25**	**42.91**/0.26	**46.24/0.21**	**40.81**/0.27	**42.90**/0.27

TABLE 3. Quantitative depth upsampling results (in PSNR/MAD) for ×16 on synthetic Middlebury dataset. Higher PSNR and lower MAD indicate better performance. Best performance is shown in bold font.

Methods	Art	Books	Dolls	Laundry	Moebius	Reindeer	Average
Bicubic	28.15/5.72	35.81/3.82	38.30/3.8	33.40/4.45	37.30/4.35	31.56/3.72	34.09/4.31
FGI [44]	29.60/3.65	36.74/1.75	37.92/1.71	33.88/2.37	37.84/2.43	32.88/1.95	34.81/2.31
RGDR [45]	29.94/3.43	36.84/1.68	39.57/1.73	33.83/1.98	38.33/1.71	31.94/1.81	35.08/2.06
MIRNet [22]	29.26/3.16	37.00/1.12	38.94/1.15	32.73/1.54	38.05/1.16	30.84/1.82	34.47/1.65
MPRNet [28]	29.93/2.14	38.14/0.85	39.90/0.91	33.55/1.32	38.70/0.91	32.82/1.36	35.51/1.25
MSG [13]	30.95/1.53	37.95/0.76	39.34/0.84	33.56/1.12	38.85/0.76	33.66/0.99	35.72/1.00
CGN [46]	32.38/1.27	39.28/0.71	41.56/0.64	36.93/0.75	40.14/0.69	34.86/0.75	37.53/0.80
PMBANet [37]	32.67/1.22	40.17/0.59	41.73/0.59	38.11/0.71	40.35/0.67	34.62/0.74	37.94/0.75
DSR [34]	32.60/1.28	39.80/0.69	41.69/0.73	37.03/1.04	40.24/0.67	34.84/0.92	37.70/0.89
PDRNet [35]	33.12/1.12	39.28/0.62	41.65/0.61	36.98/0.72	40.60/0.61	36.11/0.64	37.96/0.72
DAGF [47]	**34.16**/1.14	**42.20**/0.48	**42.95**/0.61	38.57/0.69	42.71/0.52	37.08/0.68	**39.61**/0.69
DSRNet [15]	33.55/**0.86**	41.61/0.44	42.63/**0.48**	**39.66/0.56**	**42.86/0.40**	36.89/0.49	39.53/**0.54**
Ours	33.84/**0.86**	41.47/**0.42**	42.91/**0.48**	38.77/0.59	42.66/0.41	**37.96/0.46**	39.60/**0.54**

TABLE 4. Quantitative depth upsampling results (in RMSE) on NYU-Depth-V2 dataset. Lower RMSE indicates better performance. Best performance is shown in bold font.

Factors	Bicubic	FGI [44]	RGDR [45]	SRCNN [12]	MIRNet [22]	MPRNet [28]	MSG [13]	CGN [46]	PMBANet [37]	DSR [34]	PDRNet [35]	DAGF [47]	DSRNet [15]	Ours
×4	3.31	3.16	3.26	1.64	2.23	1.62	1.30	1.20	**1.06**	1.18	1.30	1.36	1.11	**1.06**
×8	5.56	5.56	5.48	-	3.91	3.22	3.73	2.52	2.28	2.44	2.27	2.87	2.17	**2.13**
×16	10.49	9.37	8.82	-	7.15	6.23	6.53	5.18	4.98	5.15	4.42	6.06	**4.00**	4.25

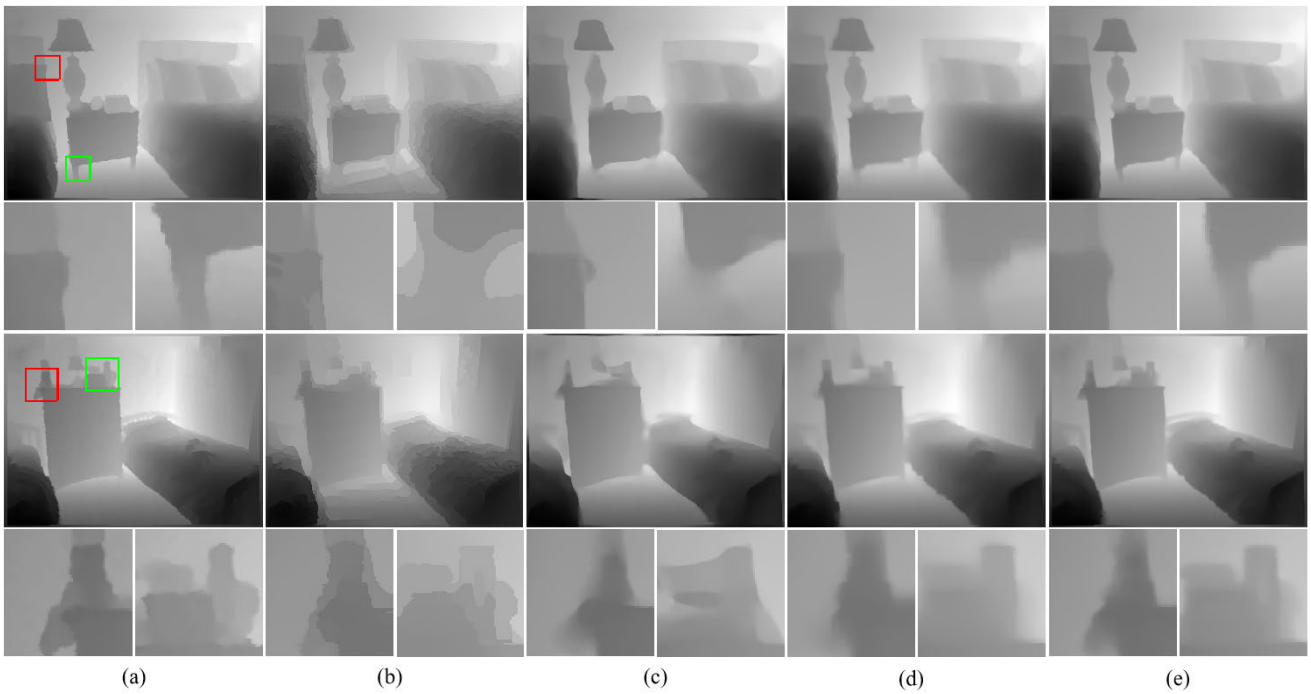

FIGURE 7. Visual comparison for ×16 on NYU-Depth-V2 dataset. (a) Ground truth depth image, (b) RGDR [45]. (c) PMBANet [37]. (d) PDRNet [35]. (e) Proposed LEDSRNet. Depth patches are enlarged for clear visualization (Top: Red box. Bottom: Green box).

IV. EXPERIMENTAL RESULTS

A. DATASETS

We use the same training dataset as MSG [13] that include 58 RGB-D image pairs from MPI Sintel depth dataset [48] and 34 RGB-D image pairs from Middlebury dataset (6, 10, 18 images are from 2001 [49], 2006 [50] and 2014 [51] datasets, respectively). To evaluate the performance of LEDSRNet, we test on 6 standard RGB-D image pairs from Middlebury 2005 [50]: Art, Books, Moebius, Dolls, Laundry and Reindeer. Similar to the previous work [35], HR color images and depth images were cropped into 128×128 patches with stride 32 for sampling factors 4, 8 and 16. The LR

TABLE 5. Quantitative depth upsampling results (in RMSE) on MPI Sintel dataset. Lower RMSE indicates better performance. Best performance is shown in bold font.

	Alley-1-32			Ambush-5-37			Cave-2-46			Market-5-27			Shanman-2-22			Temple-3-48			
Methods	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	Average
Bicubic	6.83	9.31	13.08	6.79	9.58	13.52	4.69	6.56	9.30	5.65	7.87	11.79	6.55	8.39	11.01	6.39	8.78	12.07	8.79
RGDR [45]	6.39	9.51	13.82	6.49	10.87	16.61	4.30	7.01	11.70	5.08	8.16	13.08	6.37	8.53	11.99	5.91	9.03	13.13	9.33
SRCNN [12]	4.83	-	-	4.23	-	-	3.29	-	-	3.73	-	-	5.43	-	-	4.62	-	-	-
MIRNet [22]	5.24	9.18	14.49	4.52	8.33	15.26	3.55	6.10	10.56	4.34	7.59	13.73	6.55	9.25	12.62	4.84	8.01	11.69	8.66
MPRNet [28]	4.98	8.87	12.61	4.34	8.25	14.31	3.41	6.02	9.04	4.04	7.32	12.72	6.13	8.72	10.75	4.75	7.83	11.20	8.07
MSG [13]	5.09	7.46	9.74	4.46	6.65	8.83	3.54	4.93	7.93	4.41	6.01	10.96	6.10	8.85	10.19	4.29	5.92	7.94	6.85
CGN [46]	4.86	6.96	9.43	3.83	6.31	8.15	3.42	4.86	7.43	4.26	5.81	10.83	5.75	8.21	9.72	4.14	5.71	6.71	6.47
PMBANet [37]	5.60	6.84	9.10	4.23	5.83	**7.01**	3.88	4.71	6.96	5.02	5.58	10.82	6.50	8.16	9.76	6.65	5.12	6.54	6.57
DSR [34]	4.78	6.86	9.23	3.92	5.93	7.18	3.91	4.22	7.31	4.21	5.63	10.84	5.86	8.26	9.78	4.26	5.32	6.65	6.34
PDRNet [35]	4.91	**5.25**	9.19	4.48	**4.17**	7.11	3.49	**3.86**	7.16	5.06	5.03	10.80	6.09	6.81	9.26	5.02	**4.73**	**6.01**	6.02
DAGF [47]	3.95	5.84	8.05	3.42	5.70	8.38	2.67	4.16	5.85	2.33	4.54	7.81	4.96	**6.48**	**8.46**	3.91	5.70	7.83	5.56
DSRNet [15]	3.30	5.62	7.91	2.79	6.02	11.65	2.10	4.15	**5.48**	1.86	4.22	**6.17**	5.13	7.47	9.00	2.93	6.64	10.19	5.70
Ours	**2.73**	5.63	**7.44**	**2.35**	5.77	10.51	**1.87**	4.04	5.75	**1.51**	**4.01**	6.29	**4.54**	7.46	9.20	**2.40**	5.53	10.64	**5.42**

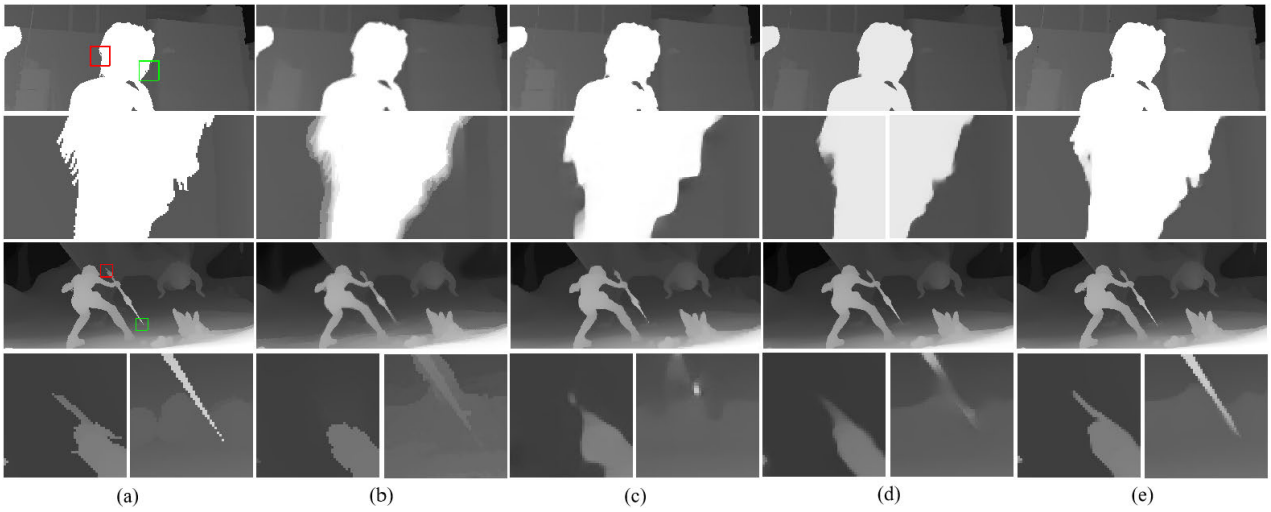

FIGURE 8. Visual comparison for ×8 on MPI Sintel dataset (*Alley* − 1 − 32 and *Cave* − 2 − 46). (a) Ground truth depth image, (b) RGDR [45]. (c) PMBANet [37]. (d) PDRNet [35]. (e) Proposed LEDSRNet. Depth patches are enlarged for clear visualization (Top: Red box. Bottom: Green box).

depth images are obtained by bicubic downsampling for a given sampling factor. We set the mini-batch size to 32. For data augmentation, we randomly perform horizontal flip and 90 degree rotation. We also perform experiments on the NYU-Depth-V2 dataset [52] captured by Kinect. Following the common splitting method [37], we use the first 1000 RGB-D image pairs as training set and the remaining 449 RGB-D image pairs as testing set. MPI Sintel dataset [48] is employed for validation to evaluate generalization of LEDSRNet.

B. IMPLEMENTATION DETAILS

LEDSRNet is implemented with PyTorch framework [53] on Nvidia GTX 3090. We use the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for network optimization. The initial learning rate is 0.0002 and decreased by 0.5 for every 50 epochs, where the total epoch is 200. To evaluate the performance, we compare LEDSRNet with three types of methods: 1) Traditional methods of bicubic interpolation, FGI [44] and RGDR [45]; 2) Three CNN-based single image SR methods of SRCNN [12], MIRNet [22] and MPRNet [28]; 3) Seven CNN-based color guided depth SR methods of MSG [13], CGN [46], PMBANet [37], DSR [34], PDRNet [35], DAGF [47] and DSRNet [15]. The results of them are generated by their official codes which are tuned to generate the best results. As evaluation metrics, peak signal to noise ratio (PSNR), root mean squared error (RMSE), and mean absolute difference (MAD) are used. We do not use SSIM for evaluation because most of the methods have similar values and are not discriminating. In depth SR, more attention is paid to the boundary recovery of an image, but SSIM is ineffective in measuring it. Instead, we use MAD for evaluation that measures the error of predicting the depth map at the pixel level. SRCNN only provides the trained models for ×2 and ×4, but does not provide the models for other factors. Thus, we do not provide the results when the sampling factor is ×8 or ×16 in Tables 4 and 5.

C. PERFORMANCE COMPARISON

1) MIDDLEBURY DATASET

Tables 1, 2 and 3 show quantitative measurements in sampling factors ×4, ×8 and ×16 on Middlebury dataset, respectively. Both single image SR and color guided depth SR based on CNN generally perform better than the traditional

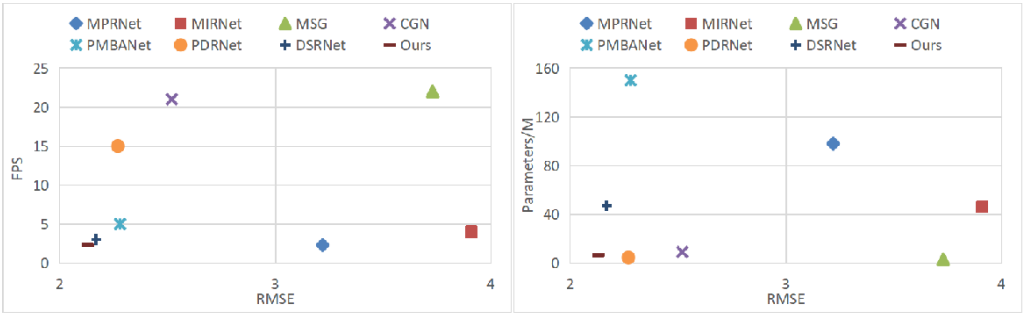

FIGURE 9. RMSE comparison with respect to (a) frames per second (FPS) and (b) model parameters among recent CNN-based methods with the sampling factor ×8. Higher FPS and lower parameters indicate better performance.

methods, while the color guided depth SR methods are significantly superior to the single image SR methods, especially when the sampling factor is large. Since the information extracted from a single depth image is limited, the restoration of edge details needs richer and finer features when the sampling factor is large, while the corresponding HR color image can provide clear edge features. LEDSRNet obviously outperforms DAGF [47] and DSRNet [15]. As the sampling factor increases, the performance of DAGF [47] and DSRNet [15] is gradually close to that of LEDSRNet, which the comprehensive performance of DAGF [47] in PSNR and MAD is slightly worse than LEDSRNet. The performance of DSRNet [15] is nearly similar to LEDSRNet, but its number of parameters is much higher than the proposed method as shown in Fig. 9. However, in the NYU and MPI datasets, the results of LEDSRNet are much higher than the others. Compared with all methods, the proposed LEDSRNet selects accurate depth edges from HR color and LR depth images by gradient estimation, and successfully reconstructs an accurate SR depth image with the help of the mask map. Thus, LEDSRNet achieves the best performance in different sampling factors and significantly outperforms the others in terms of both PSNR and MAD metrics. Fig. 6 further demonstrates the visual performance of LEDSRNet under the sampling factor ×8 on synthetic Middlebury dataset. It is obvious that LEDSRNet obtains the most similar reconstruction results to the ground truth in terms of details and boundaries. The results of traditional methods contain significant texture copying artifacts because they cannot successfully distinguish valid and redundant color edges. However, due to the lack of supplementary information from the color image, the single depth image SR methods cannot accurately recover edge details in depth. Other color-guided depth SR methods pay more attention to extracting the depth features and fail to estimate accurate edge details from the color image so that they cannot reconstruct depth details.

2) NYU-DEPTH-V2 DATASET

NYU-Depth-V2 dataset [52] captured by the real depth sensor, which is consist of 1449 densely labeled pairs of aligned RGB-D images with the size of 640 × 480. Due to the inaccurate sampling by depth cameras, the boundaries of actual depth images are often severely affected by noise. The selected methods are trained and tested on NYU dataset with the same training-testing method for a fair comparison. Table 4 shows quantitative measurements on NYU-Depth-V2 dataset. LEDSRNet also achieves outstanding performance in real depth images, while successfully dealing sensor noise and achieving the best performance in depth SR in different sampling factors (×4, ×8 and ×16). DSRNet achieves higher RMSE results on factor ×16 than LEDSRNet, but its number of parameters is much higher than LEDSRNet as shown in Fig. 9. As shown in Fig. 7, LEDSRNet preserves more depth boundaries and geometric details than other methods even in factor ×16.

3) MPI SINTEL DATASET

MPI Sintel dataset contains synthetic data and we choose six RGBD images (*Alley* − 1 − 32, *Ambush* − 5 − 37, *Cave* 2 − 46, *Market* − 5 − 27, *Shanman* − 2 − 22, and *Temple* − 3 − 48) from MPI Sintel dataset [48] for evaluation. For all methods, we use the models trained on Middlebury dataset to perform the evaluation on RGB-D image pairs. Table 5 compares LEDSRNet with other methods. LEDSRNet achieves best performance in all test image pairs in the sampling factor ×4, while LEDSRNet achieves comparable performance to the others in ×8 and ×16. However, LEDSRNet performs significantly better than the others in average. Fig. 8 shows the visual comparison on factor ×8 among them. LEDSRNet achieves better geometric details and depth boundaries that the others. For *Alley* − 1 − 32 (Top), the edges of our results are sharper and more detailed than the others. For *Cave* − 2 − 46 (Bottom), LEDSRNet is more accurate than them in estimating fine edge structures. Thus, LEDSRNet shows outstanding performance with high robustness on different test datasets.

4) MODEL COMPLEXITY COMPARISON

Running time and model parameters are used to evaluate the complexity of CNN-based methods. Fig. 9 shows the running time, model parameters and performance on NYU-Depth-V2

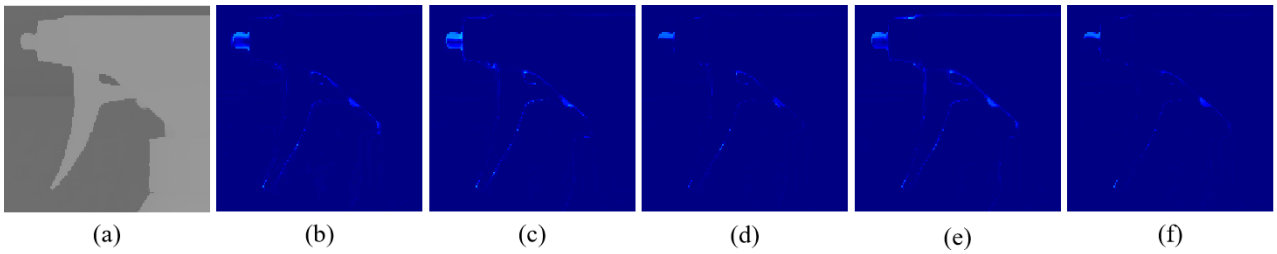

FIGURE 10. Ablation study on the loss function in synthetic Middlebury dataset (Laundry). (a) Ground truth depth image. (b) W/O gradient mask. (c) W/O depth mask. (d) W/O mask. (e) W/O SSIM loss. (f) LEDSRNet. We provide the difference map between the ground truth and the reconstruction result.

TABLE 6. Effectiveness of loss functions in LEDSRNet. Higher PSNR and lower MAD indicate better performance. Best performance is shown in bold font.

Methods	Middlebury	NYU	MPI
W/O gradient mask	42.77/0.28	**41.72**/0.72	33.58/0.54
W/O depth mask	42.22/0.28	41.68/0.72	33.43/**0.52**
W/O mask	42.50/0.28	41.68/0.73	33.58/0.53
W/O SSIM loss	42.43/0.30	41.71/0.73	**33.88**/0.60
LEDSRNet	**42.90/0.27**	**41.72/0.69**	33.65/0.53

TABLE 7. Effectiveness of LEDSRNet in different AMRB configurations. Higher PSNR and lower MAD indicate better performance. Best performance is shown in bold font.

Methods	Parameters	Middlebury	NYU	MPI
3 layers	5208454	42.77/0.27	41.68/0.72	33.56/0.54
5 layers	7858822	42.46/0.27	41.67/0.72	33.17/0.57
W/O CA	6523778	42.63/0.27	41.66/0.72	**33.68**/0.54
7 convs	6244162	37.46/1.06	35.42/1.66	28.41/2.22
LEDSRNet	6533638	**42.90/0.27**	**41.72/0.69**	33.65/**0.53**

dataset (×8) for several CNN-based methods. For running time, we compare average frames per second (FPS) among different methods on 449 test image pairs with 640 × 480. FPS and model parameters of them are obtained by running their open-source codes on the same condition or their papers. LEDSRNet achieves an optimal trade-off between running time and performance. PDRNet [35], CGN [46] and MSG [13] achieve higher FPS but much lower performance than LEDSRNet. Although LEDSRNet are slightly more than PDRNet [35] and MSG [13] in model parameters, LEDSRNet achieves a good balance between parameters and performance. As shown in Fig. 9(a), the RMSE value of LEDSRNet is close to that of DSRNet. The frames per second (FPS) of LEDSRNet is around 2.3, while FPS of DSRNet [15] is around 3. Fig. 9(b) shows the relationship between the number of parameters and performance. The number of parameters in DSRNet [15] is about 47, while the number of parameters in LEDSRNet is about 6.5. Although the number of parameters in PDRNet [35], CGN [46] and MSG [13] is also small, the RMSE performance is significantly worse than LEDSRNet.

D. ABLATION STUDY

To justify the network structure, we perform the ablation study on LEDSRNet with sampling factor ×8. We perform the experiments on Middlebury dataset. First, we explore the effectiveness of the loss function as follows:

- W/O gradient mask: We use traditional L_1 to replace masked $Loss_1$.
- W/O depth mask: We use traditional L_1 and L_2 to replace masked $Loss_2$.
- W/O mask: We use traditional L_1 and L_2 without mask.
- W/O SSIM loss: We train LEDSRNet without SSIM loss.
- LEDSRNet: Full model.

Table 6 and Fig. 10 show quantitative measurements, which indicates: (1) Compared with the traditional L_1 and L_2 loss, the proposed mask map has a positive impact on the depth SR. The mask map can effectively produce edge details and smooth regions in the SR depth image. (2) The SSIM loss can supervise SR depth reconstruction in luminance, contrast and structure simultaneously.

Then, we evaluate the AMRB structure by changing the configuration in AMRB and test its effectiveness by testing LEDSRNet with and without AMRB. The number of layers refers to the number of 3 × 3 convolution layers before the last 1 × 1 one as follows:

- 3 layers: As shown in Fig. 1, we removed the last two 3 × 3 convolution layers.
- 5 layers: As shown in Fig. 1, we add two 3 × 3 convolution layers before the final 1 × 1 convolution layer.
- W/O CA: AMRB without channel attention module.
- 7 conv layers: AMRB consists of 7 continuous convolution layers with kernel size 3 × 3.
- LEDSRNet: We use AMRB with 4 layers as our basic block.

Table 7 and Figs. 11(b)-(e) show the effectiveness of AMRB, which indicates: (1) AMRB with 4 layers can obtain more deep and shallow features, but its performance decreases with the increase of the layers. In addition, the increase of the layers also lead to a significant increase in the number of parameters, and thus we choose the 4-layer AMRB as the basic block of LEDSRNet. (2) AMRB is able to extract much richer features than a simple 4 convolutional layers.

In LEDSRNet, we perform binarization on E_{GT} to get the mask map M. In addition, we add the interpolated LR depth image D^{IL} as the input of LEDSRNet as follows.

- W/O binarization: Without binarization on the mask map M.

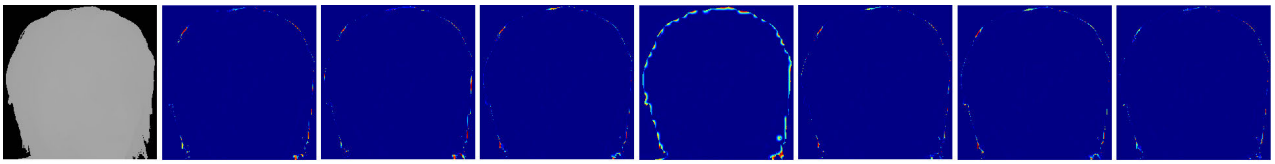

FIGURE 11. Ablation study on AMRB, mask binarization, and input of LEDSRNet in synthetic Middlebury dataset (Laundry). (a) Ground truth depth image. (b) 3 layers. (c) 5 layers. (d) W/O CA. (e) 7 convolutional layers. (f) W/O binarization. (g) W/O HR depth image. (h) LEDSRNet. We provide the difference map between the ground truth and the reconstruction result.

TABLE 8. Effectiveness of binarization and input of LEDSRNet on the performance. Higher PSNR and lower MAD indicate better performance. Best performance is shown in bold font.

Methods	Middlebury	NYU	MPI
W/O binarization	42.70/0.28	41.59/0.71	33.38/0.55
W/O HR depth image	42.40/0.30	41.61/0.73	33.55/0.58
LEDSRNet	**42.90/0.27**	**41.72/0.69**	**33.65/0.53**

- W/O HR depth image: HR color image alone as the input for the gradient estimation subnetwork.
- LEDSRNet: Full model.

Table 8 and Figs. 11(f)(g) show effectiveness of the binarization operation and the input of LEDSRNet, which indicates: (1) The mask map can significantly improve the performance of LEDSRNet. (2) Using interpolated LR depth image as input is very helpful for depth SR.

V. CONCLUSION

In this paper, we have proposed a latent edge guided depth SR network using attention-based hierarchical multi-modal fusion, named LEDSRNet. We have adopted the attention-based hierarchical multi-modal fusion for depth SR to extract and fuse the multi-modal and multi-scale features from HR color and LR depth images. Moreover, we have utilized a binarized mask map to calculate L_1 and L_2 losses for edge and smooth areas separately, thus preventing edge smoothing in depth SR. LEDSRNet consists of three main subnetworks: gradient estimation, LR depth upsampling and fusion. The gradient estimation subnetwork is used to extract rich edge information from HR color image and interpolated LR depth image, and filter out useless HR edges to generate a depth edge map. Edge features from the decoder branch in the gradient estimation subnetwork guide the LR depth image upsampling to refine depth edges in the latent space. By taking the advantages of dense block and residual block, AMRB is used to effectively fuse shallow and deep features. Various experiments on Middlebury, NYU-Depth-V2 dataset, and MPI Sintel datasets demonstrate that LEDSRNet outperforms state-of-the-art methods in terms of both visual quality and quantitative measurements. Ablation studies verify the effectiveness of AMRB and mask map, which enable LEDSRNet to contain accurate depth information.

Our future work includes introducing LEDSRNet into the 3D video compression to save bits using multi-sensor collaboration.

REFERENCES

[1] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3D for autonomous agents: A survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2019.

[2] D. Ball, P. Ross, A. English, P. Milani, D. Richards, A. Bate, B. Upcroft, G. Wyeth, and P. Corke, "Farm workers of the future: Vision-based robotics for broad-acre agriculture," *IEEE Robot. Autom. Mag.*, vol. 24, no. 3, pp. 97–107, Sep. 2017.

[3] J. Zhang, Q. Su, C. Wang, and H. Gu, "Monocular 3D vehicle detection with multi-instance depth and geometry reasoning for autonomous driving," *Neurocomputing*, vol. 403, pp. 182–192, Aug. 2020.

[4] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight cameras in computer graphics," *Comput. Graph. Forum*, vol. 29, no. 1, pp. 141–159, Mar. 2010.

[5] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 99, p. 96, Jul. 2007.

[6] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 154–169.

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

[8] M.-I. Georgescu, R. T. Ionescu, A.-I. Miron, O. Savencu, N.-C. Ristea, N. Verga, and F. S. Khan, "Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2194–2204.

[9] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.

[10] X. Chen, G. Zhai, J. Wang, C. Hu, and Y. Chen, "Color guided thermal image super resolution," in *Proc. Vis. Commun. Image Process. (VCIP)*, 2016, pp. 1–4.

[11] K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, "Dual-path deep fusion network for face image hallucination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 378–391, Jan. 2022.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[13] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 353–369.

[14] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 161–169.

[15] H. Lan and C. Jung, "DSRNet: Depth super-resolution network guided by blurry depth and clear intensity edges," *Signal Process., Image Commun.*, vol. 121, Feb. 2024, Art. no. 117064.

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

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 184–199.

[18] 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, Dec. 2019.

[19] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical dense recursive network for image super-resolution," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107475.
[20] F. Fang, J. Li, and T. Zeng, "Soft-edge assisted network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 4656–4668, 2020.
[21] G. Wu, Y. Wang, and S. Li, "Single depth map super-resolution via a deep feedback network," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 19, no. 2, Mar. 2021, Art. no. 2050072.
[22] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 492–511.
[23] W. Li, J. Li, J. Li, Z. Huang, and D. Zhou, "A lightweight multi-scale channel attention network for image super-resolution," *Neurocomputing*, vol. 456, pp. 327–337, Oct. 2021.
[24] X. Chai, F. Shao, Q. Jiang, and H. Ying, "TCCL-net: Transformer-convolution collaborative learning network for omnidirectional image super-resolution," *Knowl.-Based Syst.*, vol. 274, Aug. 2023, Art. no. 110625.
[25] Q. Liu, P. Gao, K. Han, N. Liu, and W. Xiang, "Degradation-aware self-attention based transformer for blind image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 7516–7528, 2024.
[26] J. Shi, Y. Wang, Z. Yu, G. Li, X. Hong, F. Wang, and Y. Gong, "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-CNN structure for face super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 2608–2620, 2024.
[27] 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 Conf. Multimedia*, 2020, pp. 1809–1818.
[28] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14816–14826.
[29] X. Jiang, N. Wang, J. Xin, K. Li, X. Yang, J. Li, X. Wang, and X. Gao, "FABNet: Frequency-aware binarized network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 6234–6247, 2023.
[30] W. Xu, Q. Zhu, and N. Qi, "Depth map super-resolution via joint local gradient and nonlocal structural regularizations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8297–8311, Dec. 2022.
[31] T. Li, H. Lin, X. Dong, and X. Zhang, "Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107513.
[32] 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, Apr. 2019.
[33] 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, May 2019.
[34] Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li, "Depth upsampling based on deep edge-aware learning," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107274.
[35] 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, Mar. 2022.
[36] J. Wang, C. Li, Y. Shi, D. Wang, M.-E. Wu, N. Ling, and B. Yin, "MSF-net: Multi-scale feedback reconstruction for guided depth map super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 709–723, Nov. 2024.
[37] 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.
[38] N. Metzger, R. C. Daudt, and K. Schindler, "Guided depth super-resolution by deep anisotropic diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18237–18246.
[39] A. Mehri, P. Behjati, and A. D. Sappa, "TnTViT-G: Transformer in transformer network for guidance super resolution," *IEEE Access*, vol. 11, pp. 11529–11540, 2023.
[40] I. Ariav and I. Cohen, "Fully cross-attention transformer for guided depth super-resolution," *Sensors*, vol. 23, no. 5, p. 2723, Mar. 2023.
[41] X. Chai, F. Shao, H. Chen, B. Mu, and Y.-S. Ho, "Super-resolution reconstruction for stereoscopic omnidirectional display systems via dynamic convolutions and cross-view transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.
[42] Z. Zhao, J. Zhang, X. Gu, C. Tan, S. Xu, Y. Zhang, R. Timofte, and L. Van Gool, "Spherical space feature decomposition for guided depth map super-resolution," 2023, *arXiv:2303.08942*.
[43] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[44] Y. Li, D. Min, M. N. Do, and J. Lu, "Fast guided global interpolation for depth and motion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 717–733.
[45] W. Liu, X. Chen, J. Yang, and Q. Wu, "Robust color guided depth map restoration," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 315–327, Jan. 2017.
[46] Y. Zuo, Y. Fang, P. An, X. Shang, and J. Yang, "Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensity-guided refinement," *IEEE Trans. Multimedia*, vol. 23, pp. 772–783, 2021.
[47] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Deep attentional guided image filtering," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2023.
[48] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2012, pp. 611–625.
[49] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proc. IEEE Workshop Stereo Multi-Baseline Vis. (SMBV)*, Dec. 2001, pp. 131–140.
[50] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[51] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, Germany. Cham, Switzerland: Springer, 2014, pp. 31–42.
[52] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2012, pp. 746–760.
[53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in Pytorch," in *Proc. Adv. Neural Inf. Process. Syst. Workshop Autodiff*, 2017, pp. 1–4.

HUI LAN received the B.S. degree in electronic information science and technology from Northwest University, China, in 2016. She is currently pursuing the Ph.D. degree with Xidian University, China. Her research interests include depth super-resolution and video compression.

CHEOLKON JUNG (Member, IEEE) is a Born Again Christian. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, Republic of Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with Samsung Advanced Institute of Technology, Samsung Electronics, Republic of Korea, from 2002 to 2007. He was also a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of Xidian Media Laboratory. His main research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3DTV.