IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 10, 2020, accepted July 1, 2020, date of publication July 7, 2020, date of current version July 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007667

Guided Depth Map Super-Resolution Using Recumbent Y Network

TAO LI[1], (Member, IEEE), XIUCHENG DONG[1], AND HONGWEI LIN[2], (Member, IEEE)
[1]School of Electrical Engineering and Electronic Information, Xihua University, Chengdu 610039, China
[2]College of Electrical Engineering, Northwest Minzu University, Lanzhou 730000, China

Corresponding author: Tao Li (lucia634@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61901392.

ABSTRACT Low spatial resolution is a well-known problem for depth maps captured by low-cost consumer depth cameras. Depth map super-resolution (SR) can be used to enhance the resolution and improve the quality of depth maps. In this paper, we propose a recumbent Y network (RYNet) to integrate the depth information and intensity information for depth map SR. Specifically, we introduce two weight-shared encoders to respectively learn multi-scale depth and intensity features, and a single decoder to gradually fuse depth information and intensity information for reconstruction. We also design a residual channel attention based atrous spatial pyramid pooling structure to further enrich the feature's scale diversity and exploit the correlations between multi-scale feature channels. Furthermore, the violations of co-occurrence assumption between depth discontinuities and intensity edges will generate texture-transfer and depth-bleeding artifacts. Thus, we propose a spatial attention mechanism to mitigate the artifacts by adaptively learning the spatial relevance between intensity features and depth features and reweighting the intensity features before fusion. Experimental results demonstrate the superiority of the proposed RYNet over several state-of-the-art depth map SR methods.

INDEX TERMS Depth map super-resolution, convolutional neural network, UNet network, atrous spatial pyramid pooling, attention mechanism.

I. INTRODUCTION

Depth map has many applications in practice, such as autonomous driving, virtual reality, 3D reconstruction. Recent consumer depth cameras have provided a convenient way to acquire depth maps. However, the depth maps captured by these cameras usually suffer from low spatial resolution. For instance, the resolution of the depth map taken by Kinect v2 is only 512×424. In order to solve this issue, depth map super-resolution (SR) is developed to enhance the resolution of depth maps.

Depth map SR aims to reconstruct a high-resolution (HR) depth map from its low-resolution (LR) counterpart. It has inherent ill-posedness, since there may exist multiple HR depth maps that can produce an identical LR depth map after degradation. Numerous depth map SR methods have been proposed to alleviate the ill-posedness, including filter-based methods, optimization-based methods, and learning-based methods. Considering recent RGB-D cameras can simultaneously capture an LR depth map and an HR color image, some of the depth map SR methods employ the HR color image or intensity image as guidance to enhance the quality of depth map, based on the co-occurrence assumption between depth discontinuities and intensity edges.

The idea of the filter-based methods is to use the neighboring LR depth values to estimate the HR pixel value in a filtering manner. The most representative work is joint bilateral up-sampling (JBU) [1] where bilateral weights are determined based on the guidance color image. The main advantages of the filter-based methods are simple and low computational complexity. However, they often tend to yield blur artifacts along depth discontinuities.

The key idea of optimization-based methods is to formulate depth map SR as an objective function minimization which generally consists of a data fidelity term and powerful depth map priors. For instance, Dong *et al.* [2] incorporate joint local and nonlocal regularization priors into a unified depth SR framework. Although such optimization-based methods are able to produce high quality HR depth maps, they often involve a time-consuming minimization process. Moreover,

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

their performance may degrade quickly when the test depth statistics are diverse from the adopted depth prior.

Another research direction is the development of learning-based methods, which learn a nonlinear relationship between LR and HR depth maps. A pioneer work in [3] is to co-train three dictionaries (namely, LR depth dictionary, HR depth dictionary, and HR color dictionary) to imply the nonlinear relationship. In recent years, due to the outstanding performance of deep convolutional neural network (DCNN) in various computer vision tasks, a great deal of attention has been shifted to deep learning-based depth map SR methods. For example, Hui *et al.* [4] propose a deep MSG-Net which learns HR intensity features to complement the LR depth features.

Although considerable progress has been achieved in deep learning-based depth map SR methods, there still exist some aspects which need to be further considered. (1) It is well known that multi-scale network structure can extract and take full advantage of different-level information, including low-level local fine information and high-level global semantic information. However, to design an effective multi-scale network for depth map SR task is still an open problem. (2) Existing networks usually directly use the intensity guidance but ignoring the depth-intensity spatial relevance. It would result in texture-transfer and depth-bleeding artifacts when the co-occurrence assumption between depth discontinuities and intensity edges does not hold.

To address the aforementioned problems, we propose a recumbent Y network (RYNet) for depth map SR. Specifically, our RYNet adopts an encoders-decoder architecture endowed with an effective multi-scale structure. Given an interpolated LR depth map and an HR intensity image as inputs, two weight-shared depth encoder and intensity encoder are used to generate multi-scale depth features and intensity features, respectively. Then, we employ a junction module to connect the encoders with the unified decoder. In junction module, we design residual channel attention based atrous spatial pyramid pooling (RCA-ASPP) blocks to further extract multi-scale atrous convolved depth features and intensity features for subsequent fusion. In the decoder part, we apply spatial attention mechanism to learn the spatial relevance between the depth encoder features and intensity encoder features passed from the same hierarchical encoder level by skip connection. The intensity encoder features are then adaptively reweighted to improve the accuracy of intensity guidance. Next, at each scale level of the decoder, we concatenate for fusion the corresponding depth encoder features, the reweighted intensity encoder features, and the decoder features from the preceding scale level, and then perform reconstruction.

In summary, our main contributions are listed as follows:

(1) We propose a RYNet for depth map SR by using a multi-scale encoders-decoder architecture.

(2) We develop an effective RCA-ASPP block to further increase the feature's scale diversity and explore multi-scale feature correlations.

(3) We introduce a spatial attention mechanism with the ability of learning depth-intensity spatial relevance to handle depth-intensity feature fusion. The proposed spatial attention based feature fusion block can effectively suppress texture-transfer and depth-bleeding artifacts.

Experimental results demonstrate the superior performance of the proposed RYNet.

II. RELATED WORK

A. DEEP LEARNING-BASED SINGLE COLOR IMAGE SUPER-RESOLUTION

Since Dong *et al.* firstly introduce SRCNN [5] which demonstrates that DCNN can be used to learn the mapping function between LR and HR spaces, deep learning has been successfully applied to single color image SR tasks. The deep learning-based single color image SR methods benefit from their novel network architectures and appropriate learning principles. For example, residual connections [6] and dense connections [7] are applied to enhance the network learning capability for color image SR. The adoption of adversarial loss in SRGAN [8] and ESRGAN [9] bring significant gains in perceptual SR quality. The attention mechanisms, such as channel attention [10], [11], non-local attention [12], and color attention [13], are used to exploit the inherent feature correlations in color image SR networks. The feedback mechanism in DBPN [14] and SRFBN [15] work in a top-down manner, feeding high-level information back to refine low-level information. When retrained with depth map datasets, these deep learning-based methods can be generalized to deal with depth map SR problem.

B. DEPTH MAP SUPER-RESOLUTION

1) SINGLE DEPTH MAP SUPER-RESOLUTION

Plenty of single depth map SR methods have been developed to obtain an HR depth map from a single LR depth map. Single depth map SR is challenging due to its severely undetermined nature. Many filter-based methods have been proposed to explore structural characteristics of depth maps. For example, Hornacek *et al.* [16] exploit 3D patch-wise self-similarity across depth according to rigid body transformation. Lei *et al.* [17] design a depth map up-sampling filter by considering depth smoothness, texture similarity and view synthesis quality.

Optimization-based methods attempt to alleviate the ill-posedness of single depth map SR by employing effective optimization models. Aodha *et al.* [18] search in the range image database to find appropriate HR candidate patches for each LR input depth patch, and then formulate the selection of right candidate as a Markov Random Field (MRF) labeling problem. Xie *et al.* [19] propose an edge-guided single depth map SR method, which constructs the HR edge map from the LR depth edges through an MRF optimization.

The leaning-based single depth map SR methods usually fall into two categories: sparse representation-based and deep learning-based. In the former category, Ferstl *et al.* [20]

estimate edge priors by sparse coding with an external dictionary, and then merge the edge priors into a variational energy minimization using a Total Generalized Variation (TGV) regularization. Mandal *et al.* [21] construct sub-dictionaries of exemplars to restore HR depth map. In the latter category, Riegler *et al.* [22] integrate an anisotropic TGV regularization term on top of a deep network to construct an end-to-end ATGV-Net for depth SR. Song *et al.* [23] represent the single depth map SR task as a series of novel view synthesis sub-tasks, each of which can be solved by end-to-end deep learning. Huang *et al.* [24] propose a pyramid-structured DCNN with dense-residual connection to progressively generate depth maps of various levels.

Due to the limitation of insufficient available information, single depth map SR methods can only perform well in the case of small scaling factors. Their performance will deteriorate as the scaling factor increases.

2) COLOR GUIDED DEPTH MAP SUPER-RESOLUTION

A large number of color guided depth map SR methods have emerged in the literature. Methods in this category are generally based on the co-occurrence assumption between depth discontinuities and the corresponding intensity edges. Compared to single depth map SR, color guided depth map SR is more reliable and robust by introducing additional guidance information from the aligned HR color image.

The filter-based methods design local filters whose weights are determined by the affinity measure based on RGB-D image pairs. For instance, Liu *et al.* [25] use geodesic distances inferred from an aligned HR color image to upsample an LR depth map. He *et al.* [26] propose a local linear edge-preserving filter called guided image filter to perform guided upsampling. Lo *et al.* [27] present a joint trilateral filtering algorithm that exploits spatial distance, color difference, and local depth gradient for depth map SR.

The optimization-based methods exploit various optimization models for color guided SR, including MRF [28], [29], auto-regressive (AR) model [30], [31], weighted least squares (WLS) [32]–[34], total variation (TV) [35], [36], and graph signal model [37]. Specifically, Diebel *et al.* [28] design a MRF framework, which contains a pairwise depth measurement potential and an image guided depth smoothness prior potential. Zhou *et al.* [34] propose alternately guided depth SR using WLS and zero-order reverse filtering. Jiang *et al.* [36] propose a unified depth SR model with transform domain regularization and spatial multi-directional TV prior. Liu *et al.* [37] design a depth SR optimization framework by combining both internal graph-signal smoothness prior and external depth-color gradient consistency. Yu *et al.* [38] propose color guided depth up-sampling based on edge sparsity and super-weighted L_0 gradient minimization.

At an early stage, the leaning-based methods mainly benefit from the adoption of sparse representation. Kiechle *et al.* [39] train a bimodal co-sparse analysis model to capture the interdependency across the RGB-D pair. Tosic *et al.* [40] present a method for learning joint depth and intensity sparse generative models and use joint basic pursuit to infer sparse coefficients. Motivated by the success of DCNN, deep learning-based methods [41]–[47] are also developed. The complete deep primal-dual network in [41] contains a fully convolutional network for an initial HR depth estimation, and a non-local variational primal-dual network for HR depth refinement. A joint image filter based on convolutional neural network (CNN) is introduced in [43] to selectively transfer salient structures from guidance image to the target depth. A cascade coarse-to-fine CNN is proposed in [44] to learn data-driven filters for color guided depth SR problem. The local and global residual learning is adopted in [45] to learn the frequency-dependent mapping function for coarse-to-fine depth reconstruction. A residual UNet deep network named DepthSR-Net is presented in [46] to infer HR depth map by hierarchical features driven residual learning. A deep edge-aware learning framework in [47] is used to estimate depth edges as reconstruction cues and then two depth restoration modules are used to recover HR depth map. The proposed method in this paper also belongs to the deep learning-based SR category with color guidance.

C. UNet AND YNet

The UNet network [48] and its variations, which are representative encoder-decoder architectures, have attracted extensive interest in many computer vision tasks, including image segmentation [49]–[51], image SR [52], [53], image generation [54], [55], and object detection [56]. Typically, the UNet network is composed of a contracting path in encoder, an expanding path in decoder, and some skip connections between two paths. The contracting path gradually reduces the spatial dimensions of feature maps and captures higher-level contextual information. The expanding path gradually recovers the spatial details. And the skip connections are added between layers at the same hierarchical level in encoder and decoder, merging lower layer features with higher layer features to enhance networks' learning ability.

The most relevant work to ours is YNet in [57], which generalizes UNet for joint segmentation and classification tasks. It has one encoder and two decoders, where one decoder branch works for segmentation and the other for classification. In contrast, to solve depth map SR task, we elaborately design a RYNet structure which has two encoders (depth encoder and intensity encoder) and one decoder, resembling a recumbent Y. The RYNet can make full use of the depth information and intensity guidance information with the two encoders, and combine them in the decoder to recover a high-quality HR depth map. In addition, we introduce residual channel attention based atrous spatial pyramid pooling (RCA-ASPP) in the central junction of RYNet to further extract expressive multi-scale features and achieve efficient feature fusion. Meanwhile, the feature fusion from each skip connection pair is guided by a spatial attention mechanism,

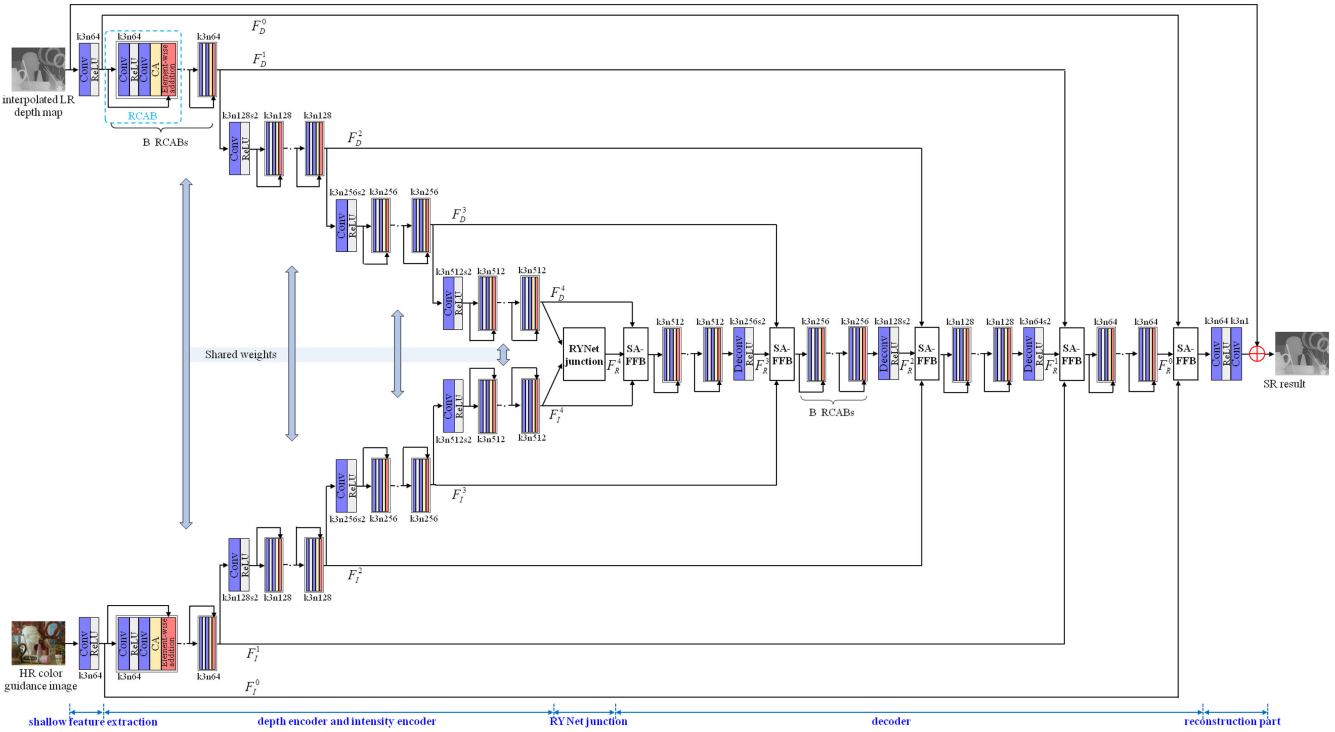

FIGURE 1. Framework of the proposed Recumbent Y Network (RYNet). Kernel size (k), number of output feature maps (n), and stride (s) are indicated for each convolutional layer. For simplicity, the default s = 1 is omitted.

designed to exploit spatial correlations between depth features and intensity features,

III. PROPOSED METHOD

A. RYNet ARCHITECTURE

As shown in Fig. 1 (please refer to the electronic version for better visualization), the overall architecture of our RYNet network consists of six parts: the shallow feature extraction, the depth encoder, the intensity encoder, the junction module, the decoder, and finally the reconstruction part. Instead of directly taking LR depth map as input, we upscale it to the desired solution using Bicubic interpolation in advance. D_{LR}, I_{HR}, and D_{SR} denote the input interpolated LR depth map, the input HR color guidance image, and the reconstructed HR depth output of RYNet, respectively. We use one convolutional (Conv) layer to extract the shallow features F_D^0 and F_I^0 from D_{LR} and I_{HR}, respectively, which can be formulated as

$$F_D^0 = f_{SF\_D}(D_{LR}), \quad F_I^0 = f_{SF\_I}(I_{HR}), \tag{1}$$

where f_{SF_D} denotes the shallow depth feature extraction function, and f_{SF_I} the shallow intensity feature extraction function. Note that the ReLU activation functions are omitted for clarity unless specified.

F_D^0 and F_I^0 are fed into the depth encoder and the intensity encoder as inputs, respectively. Both encoders contain L different scale levels (L = 4 in Fig. 1). In every scale level except the 1st one, down-sampling operation is first performed by using a Conv layer with stride 2 to enlarge receptive field and deal with features in higher semantic level. Meanwhile the number of feature channels is doubled. Then, inspired by the success of RCAN in [10], we stack B residual channel attention blocks (RCAB for short) after each down-sampling Conv layer. As shown in the top-left corner of Fig. 1, an RCAB contains a residual block and a channel attention (CA) block, which focuses on more informative components and possesses discriminative learning ability. A detailed structure of CA block will be elaborated in Section III-B. To ensure the consistency between depth features and intensity features to enhance the subsequent feature fusion efficiency, the depth encoder and intensity encoder share the same weights. Two encoder branches are formulated as:

$$F_D^l = f_{Enc}^l(F_D^{l-1}), \quad F_I^l = f_{Enc}^l(F_I^{l-1}), \tag{2}$$

respectively, where f_{Enc}^l denotes the encoding function at the l-th scale level integrating down-sampling and B RCABs, $l \in \{1, \ldots, L\}$. F_D^l and F_I^l refer to the output of l-th depth scale level and l-th intensity scale level, respectively. In this way, two encoders gradually reduce the spatial resolution of features maps and capture multi-level semantic features.

The RYNet junction module is used to connect the encoders with the decoder. In RYNet junction, residual channel attention based atrous spatial pyramid pooling (RCA-ASPP) block is designed for further multi-scale feature extraction. High-level depth features and intensity features are also fused to be fed into the decoder. More details on the RYNet junction are provided in the Section III-B.

The decoder also consists of L scale levels, each of which has the same spatial resolution as the corresponding encoder scale level. We add skip connections between corresponding scale levels to transfer lower level information from the encoders to the decoder. The fusion of 3 feature maps is then implemented using a spatial attention based feature fusion block (SA-FFB) located at the head of each decoder scale level. The SA-FFB in the l-th decoder scale level is formulated as:

$$\tilde{F}_R^l = f_{SA-FFB}^l(F_D^l, F_I^l, F_R^l), \tag{3}$$

where f_{SA-FFB}^l is the function of SA-FFB to generate the output $\tilde{F}_R^l$. The detailed network structure of SA-FFB is introduced in Section III-C. The following B cascaded RCABs are used to recover details. At the end of each decoder scale level except the last one, a deconvolutional (Deconv) layer with stride 2 is employed to handle spatial up-sampling and feature channel compression. At last, an extra SA-FFB is inserted to fuse F_D^0, F_I^0, and F_R^0. The full decoder branch is formulated as:

$$\begin{aligned} F_R^{l-1} &= f_{Dec}^l(F_D^l, F_I^l, F_R^l), \\ and\ \ \tilde{F}_R^0 &= f_{SA-FFB}^0(F_D^0, F_I^0, F_R^0), \end{aligned} \tag{4}$$

where f_{Dec}^l refers to the decoding function at the l-th scale level which produces F_R^{l-1}. And $\tilde{F}_R^0$ represents the output of the whole decoder branch. The decoder gradually increases the spatial resolution of features and recovers more depth details.

Global residual learning is used for stable and fast training. In the reconstruction part, we stack two Conv layers as well as a ReLU activation layer between them to reconstruct the residual depth map. And then we apply the element-wise summation of the residual depth map and the interpolated LR depth map D_{LR} to obtain the final output D_{SR}, which can be expressed as

$$D_{SR} = D_{LR} + f_{REC}(\tilde{F}_R^0) \tag{5}$$

where f_{REC} refers to the reconstruction function consisting two Conv layers and one ReLU layer in between. The overall process of our RYNet is formulated as

$$D_{SR} = f_{RYNet}(D_{LR}, I_{HR}) \tag{6}$$

where f_{RYNet} denotes the function of RYNet.

Then RYNet is optimized with a loss function. We choose $L2$ loss as most previous works do. Given a training set $\{D_{LR}^i, I_{HR}^i, D_{HR}^i\}_{i=1}^N$, which contains N interpolated LR depth map, N HR color guidance images, and N HR ground truth depth maps. The loss function of RYNet can be formulated as:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|f_{RYNet}(D_{LR}^i, I_{HR}^i) - D_{HR}^i\|^2, \tag{7}$$

where Θ denotes the parameters of the network. The minimization of (7) is achieved by using stochastic gradient descent.

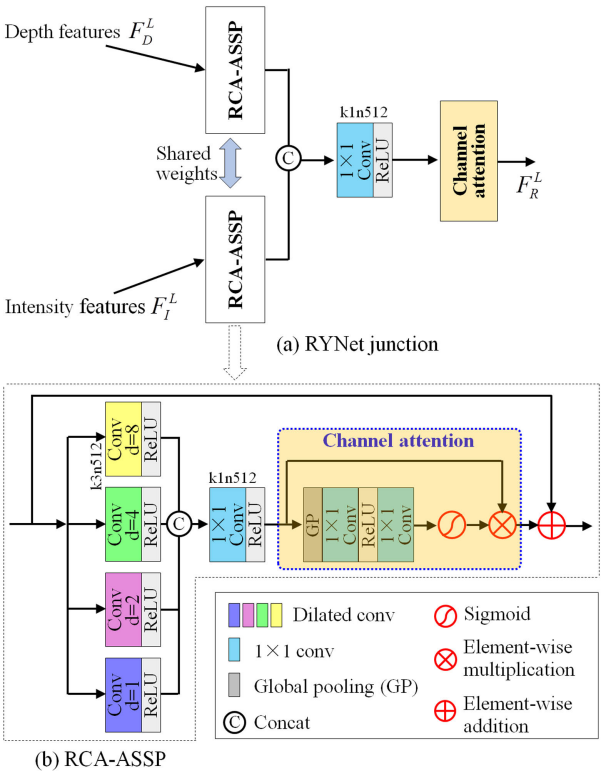

FIGURE 2. The structure of RYNet junction.

B. RYNet JUNCTION

The network architecture of the RYNet junction is illustrated in Fig. 2 (a). The high-level intensity features F_I^L from intensity encoder and the high-level depth features F_D^L from depth encoder are sent to the junction module for feature fusion. We first apply two RCA-ASPP blocks to further extract the multi-scale features of F_I^L and F_D^L, respectively. These two RCA-ASPP blocks share the same weights. Then we concatenate the features out of two RCA-ASPP blocks together and feed it into a 1×1 Conv layer to fuse information. Finally, we use a CA block to focus on more informative fusion features. Details of RCA-ASPP block and CA block are shown in Fig. 2 (b). Specifically, the output F_R^L of RYNet junction can be formulated as:

$$\begin{aligned} F_R^L &= f_{CA}(f_F(\left[f_{RCA-ASPP}(F_I^L), f_{RCA-ASPP}(F_D^L)\right])) \\ &= f_J(F_I^L, F_D^L), \end{aligned} \tag{8}$$

where $f_{RCA-ASPP}$, f_F, and f_{CA} represent the function of RCA-ASPP block, feature fusion Conv, and CA block, respectively. $\left[\cdot,\cdot\right]$ refers to the feature concatenation. And f_J denotes the overall function of RYNet junction.

Multi-scale context information would help resolve the misalignment between depth maps and color guidance image, and result in more efficient and robust feature fusion. To this end, we propose the RCA-ASPP block as shown in Fig. 2 (b) to further increase the feature's scale diversity.

We first combine four parallel atrous convolutional layers with different dilation rates of 1, 2, 4, and 8, and then concatenate multiple atrous convolved features into one feature using

a 1×1 Conv layer. This forms the basic atrous spatial pyramid pooling (ASPP) structure. Moreover, to adaptively pay attention to different channel-wise features, we take the CA mechanism [10]. In a CA block, a global pooling is used to aggregate the multi-scale features across spatial dimension to generate a squeezed channel descriptor. A Conv/ReLU/Conv combination followed by a sigmoid gating unit is then applied to learn nonlinear interactions between channels. Finally, an element-wise multiplication operation is performed to channel-wisely reweight the multi-scale features according to weight values out of the sigmoid activation. Our final RCA-ASPP block is constructed by cascading the basic ASPP and the CA block with residual connection. Thus the proposed RCA-ASPP block not only enriches scale diversity of features, but also adaptively explores multi-scale feature correlations, resulting in more expressive feature learning for the subsequent depth-intensity feature fusion.

C. SPATIAL ATTENTION BASED FEATURE FUSION BLOCK

Most of the color guided depth map SR methods rely on the co-occurrence assumption between depth discontinuities and aligned intensity edges. However, the co-occurrence assumption does not always hold. The assumption might be violated when object surface has continuous depth but with complicated texture, or adjacent object surfaces have different depths but with similar color or texture, therefore, leading to texture-transfer and depth-bleeding artifacts. In order to mitigate the artifacts, we develop a spatial attention (SA) mechanism as shown in Fig. 3, to exploit the spatial relevance between intensity encoder features and depth encoder features. The SA module adaptively rescale the intensity encoder features F_I^l in spatial domain before its fusion with F_D^l and F_R^l.

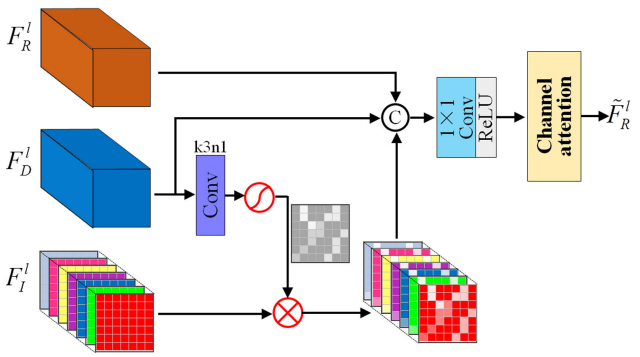

FIGURE 3. Spatial attention based feature fusion block (SA-FFB).

Let $F_I^l = [\mu_1, \mu_2, \ldots, \mu_{H\times W}]$ be one input of SA, which has C feature maps with size of $H \times W$. And $\mu_i \in \mathbb{R}^{C\times 1\times 1}$ is the i-th spatial features of F_I^l. Here we first take a Conv layer ($k3n1$) and a sigmoid activation to generate different attention weights from F_D^l for each spatial feature in F_I^l, as

$$w = \delta(f_{SA\_W}(F_D^l)) \tag{9}$$

where f_{SA_W} denotes the function of the Conv layer which learns the spatial attention. δ denotes the sigmoid gating function. Then we use the spatial attention weight map $w = [w_1, w_2, \ldots, w_{H\times W}] \in \mathbb{R}^{1\times H\times W}$ to reweight the intensity features F_I^l by

$$\tilde{F}_I^l = f_{SA}(F_I^l, w), \tag{10}$$

where $\tilde{F}_I^l = [v_1, v_2, \ldots, v_{H\times W}]$ are the reweighted intensity features, and $v_i \in \mathbb{R}^{C\times 1\times 1}$. f_{SA} denotes the space-wise multiplication operation, which implies $v_i = w_i \cdot \mu_i$. We employ the SA mechanism to learn the distinct importances of different intensity spatial features and then re-calibrate them. The relevant spatial features in F_I^l are strengthened and the irrelevant ones are suppressed. After concatenating $\tilde{F}_I^l$, F_D^l and F_R^l, we finally adopt a 1×1 Conv layer and a CA block to effectively fuse them.

D. IMPLEMENTATION DETAILS

Now we specify the implementation details of our RYNet. By comprehensively considering the network complexity and sufficient feature's scale diversity, we set $L = 4$ scale levels in encoders and decoder. We set the number of RCABs cascaded in each scale level as $B = 4$. We set 3×3 as the size of all Conv and Deconv layers except where otherwise stated. For Conv and Deconv layers with kernel size 3×3, we apply zero-padding strategy to keep the size fixed. The output_padding of Deconv layers is set to 1. The number of filters in each weight layer is set according to the n value indicated in Fig. 1, Fig. 2, and Fig. 3.

IV. EXPERIMENTAL RESULTS

A. SETTINGS

1) TRAINING DATA

The training dataset consisting of 82 RGB-D images is provided by [4]. To make full use of data, we adopt data augmentation (flipping and rotation) as [46] do. The HR RGB color images are transferred to corresponding HR intensity guidance images beforehand. Then the HR depth maps and HR intensity images are cropped into 96×96 image patches for scaling factors $2\times$ and $4\times$, and 128×128 image patches for scaling factors $8\times$ and $16\times$, respectively. The adjacent patches have 24-pixels overlap. To synthesize LR depth patches, the corresponding HR depth patches are further down-sampled using Bicubic with the given scaling factor. All the training patches are normalized into [0, 1].

2) TRAINING SETTINGS

Our network models are implemented with Pytorch framework. The networks are optimized using Adam [58] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 64. In the network initialization, the filter weight parameters are initialized using the "MSRA" method in [59], and the bias parameters are set to 0. The initial learning rate is set to 10^{-4} and decreased to half after every 40 epochs. The training is stopped after 80 epochs since validation accuracy stops further improving. We train a specific network for each scaling factor ($2\times$, $4\times$, $8\times$, and $16\times$), respectively.

3) TEST DATA

To keep consistency with previous works [45], [46], we use 10 depth maps from the hole-filled Middlebury dataset [60] for test under various scaling factors, including Art, Book, Dolls, Laundry, Moebius, Reindeer, Cones, Teddy, Tsukuba, and Venus. None of the test depth maps occur in the training set. We conduct noise-free and noisy experiments to demonstrate the robustness of our network. In the noisy experiment, Gaussian noise with variance of 25 is added to the LR depth maps to simulate the ToF-like degradation. In addition, we evaluate our network on real data [36] captured by Kinect. The experiment results on several standard benchmark datasets (Middlebury [60], ToFMark [35], and NYU Depth v2 [61]) are also provided for a statistical evaluation.

4) BASELINE METHODS

We compare our method with the baseline Bicubic interpolation method and the following state-of-art methods:

1) Deep learning-based single color image SR methods, *i.e.*, RCAN [10] and SAN [11].

2) Single depth map SR methods, such as, Aodha *et al.* [18], and ATGV-Net [22].

3) Color guided depth map SR methods, including DJFR [43], JID [39], TSDR [36], MSG-Net [4], MFR-SR [45], and DepthSR-Net [46].

We briefly explain the experimental details of the deep learning-based comparison methods. For RCAN [10] and SAN [11], we use their source codes and retrain them with our augmented training dataset. The authors of ATGV-Net [22], DJFR [43], MSG-Net [4], and DepthSR-Net [46] have provided their source codes for training and testing. We directly use their publicly available models for some experimental cases. For other cases where their models are inaccessible, we retrain them using their training codes. We only perform numerical comparison with the reported results of TSDR [36] and MFR-SR [45] in the noise-free and noisy experiments.

5) EVALUATION METRICS

The quantitative performance is reported in terms of RMSE and PSNR (dB). The best result for each evaluation is highlighted in bold, whereas the second best one is underlined.

B. EXPERIMENT ON NOISE-FREE DATA

In this subsection, we evaluate the performance of RYNet on noise-free depth maps. The quantitative results for scaling factors of $2\times$, $4\times$, $8\times$, and $16\times$ are reported in Table 1. It can be observed that all the deep learning-based methods achieve significant performance gains over other methods for all the scaling factors, reflecting that DCNN can facilitate learning more accurate LR-HR mapping relationships. In the case of small scaling factors such as $2\times$ and $4\times$, the performances of RCAN [10] and SAN [11] without color guidance are comparable to, or even better than those of color guided methods. While for large scaling factors such as $8\times$ and $16\times$, the increasing feature ambiguity hinders the performance of methods without color guidance. The results demonstrate that color guidance can help alleviating feature ambiguity and resulting in SR performance improvement, especially on large scaling factors. Benefiting from the increase of scale diversity by multi-scale encoders-decoder structure and RCA-ASPP structure, and the discriminative feature learning by spatial attention mechanism, the proposed RYNet outperforms other competing methods for all scaling factors. Specifically, compared with the second best RMSE results, our average reductions are up to 0.12 over SAN [11] for scaling factor of $2\times$, and up to 0.20, 0.18, and 0.50 over DepthSR-Net [46] for scaling factors of $4\times$, $8\times$, and $16\times$, respectively. And compared with the second best PSNR results, our average gains are up to 3.61dB and 2.13dB over MSG-Net [4] for scaling factors of $2\times$ and $4\times$, and up to 1.22dB and 1.17dB over DepthSR-Net [46] for scaling factors of $8\times$ and $16\times$, respectively.

In order to compare visual quality, the reconstructed results produced by different methods on depth map Cones for $4\times$ SR and Venus for $8\times$ SR are shown in Fig. 4 and Fig. 5, respectively. The absolute error images between reconstructed results and ground truth images are also provided in terms of jet colormap to give an explicit comparison. Specifically, the reconstructed depth maps by Bicubic interpolation, DJFR [43], and JID [39] are over-smoothed to some extent. The patch-based method [18] suffers from severe blocking artifacts. ATGV-Net [22] produces obvious reconstruction artifacts, especially in the case of $4\times$ SR. Although deep learning-based methods RCAN [10] and SAN [11] generate competitive results for $4\times$ SR, they inevitably introduce blurring artifacts for $8\times$ SR, due to the difficulty in learning accurate mapping relationship without color guidance information. Recent MSG-Net [4] and DepthSR-Net [46] recover depth edges well, but moderate reconstruction errors still present in their smooth regions. By contrast, the proposed RYNet shows superiority in reducing reconstruction errors as well as recovering sharp edges and accurate depth details. Our reconstructed depth maps are more visually appealing and closer to the ground truth, which can be observed from the highlighted cropped regions.

C. EXPERIMENT ON NOISY DATA

To compare the robustness of comparison methods, we evaluate the performance of different methods on the noisy data in this subsection. We add Gaussian noise with mean 0 and variance 25 to the LR depth maps in our training dataset, and retrain the RYNet model for $8\times$ SR. Then, the trained model is evaluated on the test data including 10 noisy depth maps. Due to limited space, seven representative or state-of-the-art algorithms from Section IV-B are selected as comparison baselines of noisy experiments. The quantitative results are presented in Table 2. As illustrated in Table 2, although without color guidance, RCAN [10] and SAN [11] yield better results than MSG-Net [4]. And the proposed RYNet, MFR-SR [45], and DepthSR-Net [46] perform the 1st, 2nd, and 3rd best in terms of average

TABLE 1. Quantitative comparison (in RMSE / PSNR (dB)) on noise-free data.

Method	Art	Book	Dolls	Laundry	Moebius	Reindeer	Cones	Teddy	Tsukuba	Venus	Avg.
					2×						
Bicubic	2.63 / 39.74	1.04 / 47.77	0.91 / 48.96	1.60 / 44.03	0.87 / 49.32	1.92 / 42.45	2.51 / 40.14	1.93 / 42.40	5.83 / 32.82	1.30 / 45.84	2.05 / 43.35
RCAN	0.44 / 55.35	0.36 / 57.11	0.45 / 55.01	0.40 / 56.10	0.40 / 55.99	0.41 / 55.94	0.69 / 51.31	0.70 / 51.25	0.67 / 51.58	0.41 / 55.92	0.49 / 54.56
SAN	0.45 / 55.10	0.35 / 57.19	0.45 / 55.00	0.40 / 56.07	0.41 / 55.89	0.41 / 55.90	0.70 / 51.19	0.71 / 51.09	**0.53 / 53.66**	0.44 / 55.32	0.49 / 54.64
Aodha *et al.*	2.05 / 41.89	0.97 / 48.38	1.27 / 46.03	1.89 / 42.62	1.18 / 46.70	1.54 / 44.39	2.70 / 39.52	2.40 / 40.54	5.33 / 33.59	0.93 / 48.73	2.03 / 43.24
ATGV-Net	0.50 / 54.11	0.25 / 60.29	0.36 / 57.04	0.40 / 56.10	0.30 / 58.46	0.42 / 55.59	1.17 / 46.77	0.81 / 49.97	2.65 / 39.65	1.22 / 46.41	0.81 / 52.44
DJFR	-	-	-	-	-	-	-	-	-	-	-
JID	1.27 / 46.06	0.67 / 51.62	0.75 / 50.58	0.77 / 50.36	0.62 / 52.30	1.07 / 47.55	1.57 / 44.21	1.20 / 46.56	3.51 / 37.22	0.63 / 52.19	1.21 / 47.87
TSDR	-	-	-	-	-	-	-	-	-	-	-
MSG-Net	0.56 / 53.16	0.26 / 59.94	0.37 / 56.70	0.37 / 56.75	0.31 / 58.17	0.42 / 55.59	0.91 / 49.00	0.71 / 51.07	1.85 / 42.80	0.14 / 65.08	0.59 / 54.83
MFR-SR	0.71 / 51.11	0.42 / 55.67	0.60 / 52.57	0.61 / 52.42	0.42 / 55.67	0.65 / 51.87	-	-	-	-	-
DepthSR-Net	0.53 / 53.64	0.43 / 55.54	0.49 / 54.27	0.44 / 55.23	0.44 / 55.29	0.52 / 53.86	0.79 / 50.15	0.83 / 49.78	1.36 / 45.46	0.51 / 54.06	0.63 / 52.73
RYNet	**0.26 / 59.88**	**0.18 / 63.09**	**0.27 / 59.45**	**0.22 / 61.23**	**0.24 / 60.66**	**0.25 / 60.12**	**0.50 / 54.10**	**0.58 / 52.84**	1.08 / 47.49	**0.13 / 65.58**	**0.37 / 58.44**
					4×						
Bicubic	3.87 / 36.38	1.60 / 44.05	1.31 / 45.80	2.40 / 40.52	1.33 / 45.63	2.80 / 39.19	3.84 / 36.45	2.85 / 39.04	8.57 / 29.47	1.91 / 42.52	3.05 / 39.90
RCAN	1.40 / 45.19	0.58 / 52.80	0.87 / 49.35	0.86 / 49.45	0.72 / 50.93	1.01 / 48.00	2.12 / 41.61	1.59 / 44.09	3.54 / 37.15	0.70 / 51.28	1.34 / 46.99
SAN	1.42 / 45.06	0.58 / 52.90	0.83 / 49.70	0.86 / 49.47	0.70 / 51.25	1.03 / 47.91	2.13 / 41.57	1.54 / 44.38	3.69 / 36.79	0.71 / 51.06	1.35 / 47.01
Aodha *et al.*	3.93 / 36.23	1.73 / 43.39	2.00 / 42.12	3.06 / 38.41	1.87 / 42.70	2.84 / 39.06	5.22 / 33.78	3.79 / 36.56	9.40 / 28.67	1.83 / 42.87	3.57 / 38.38
ATGV-Net	1.52 / 44.52	0.47 / 54.63	0.78 / 50.30	0.95 / 48.61	0.62 / 52.26	1.07 / 47.56	2.91 / 38.84	1.48 / 44.70	6.67 / 31.65	0.32 / 58.05	1.68 / 47.11
DJFR	3.72 / 36.73	1.56 / 44.27	1.37 / 45.38	2.27 / 40.99	1.27 / 46.07	2.58 / 39.91	3.55 / 37.11	2.79 / 39.22	8.36 / 29.68	1.74 / 43.32	2.92 / 40.27
JID	1.96 / 42.29	0.83 / 49.73	0.95 / 48.56	1.30 / 45.85	0.85 / 49.51	1.45 / 44.93	2.85 / 39.03	1.77 / 43.18	5.94 / 32.66	0.87 / 49.38	1.88 / 44.51
TSDR	1.57 / 44.19	1.05 / 47.69	-	0.98 / 48.34	-	1.19 / 46.64	-	1.46 / 44.88	4.79 / 34.53	-	-
MSG-Net	1.40 / 45.23	0.46 / 54.94	0.73 / 50.85	0.79 / 50.21	0.58 / 52.86	0.98 / 48.27	2.60 / 39.85	1.49 / 44.69	4.29 / 35.48	0.35 / 57.34	1.37 / 47.97
MFR-SR	1.54 / 44.38	0.63 / 52.14	0.89 / 49.14	1.11 / 47.22	0.72 / 50.98	1.23 / 46.33	-	-	-	-	-
DepthSR-Net	1.22 / 46.42	0.61 / 52.35	0.81 / 49.97	0.79 / 50.22	0.68 / 51.54	0.97 / 48.40	2.33 / 40.77	1.37 / 45.37	**3.29 / 37.78**	0.64 / 52.01	1.27 / 47.48
RYNet	**0.98 / 48.29**	**0.36 / 57.11**	**0.59 / 52.66**	**0.64 / 52.06**	**0.50 / 54.18**	**0.74 / 50.70**	**1.92 / 42.46**	**1.36 / 45.48**	3.31 / 37.73	**0.25 / 60.28**	**1.07 / 50.10**
					8×						
Bicubic	5.46 / 33.39	2.33 / 40.79	1.86 / 42.75	3.45 / 37.38	1.97 / 42.23	3.98 / 36.14	5.52 / 33.29	4.04 / 35.99	12.33 / 26.31	2.75 / 39.35	4.37 / 36.76
RCAN	2.97 / 38.67	1.19 / 46.59	1.30 / 45.86	1.81 / 42.98	1.16 / 46.88	2.17 / 41.40	4.34 / 35.38	2.94 / 38.77	7.79 / 30.30	1.50 / 44.61	2.72 / 41.14
SAN	2.79 / 39.21	0.98 / 48.26	1.32 / 45.71	1.57 / 44.23	1.13 / 47.08	1.94 / 42.36	4.24 / 35.58	2.86 / 39.01	7.54 / 30.59	1.38 / 45.31	2.58 / 41.73
Aodha *et al.*	7.05 / 31.17	2.81 / 39.17	3.05 / 38.46	4.47 / 35.12	3.11 / 38.28	4.78 / 34.54	7.94 / 30.14	6.62 / 31.72	15.76 / 24.18	3.21 / 37.99	5.88 / 34.08
ATGV-Net	3.16 / 38.15	1.08 / 47.50	1.33 / 45.65	1.87 / 42.71	1.12 / 47.18	2.06 / 41.86	5.15 / 33.90	2.86 / 39.01	11.57 / 26.87	1.10 / 47.27	3.13 / 41.01
DJFR	4.84 / 34.43	2.15 / 41.48	1.75 / 43.27	3.05 / 38.45	1.90 / 42.54	3.55 / 37.12	6.12 / 32.39	3.49 / 37.27	14.16 / 25.11	2.22 / 41.22	4.32 / 37.33
JID	2.77 / 39.27	1.04 / 47.81	1.25 / 46.16	1.86 / 42.74	1.19 / 46.61	2.23 / 41.16	4.51 / 35.04	2.69 / 39.53	9.68 / 28.42	1.23 / 46.35	2.85 / 41.31
TSDR	2.30 / 40.86	1.06 / 47.62	-	1.58 / 44.15	-	1.75 / 43.30	-	1.99 / 42.15	9.39 / 28.67	-	-
MSG-Net	2.42 / 40.45	0.88 / 49.24	1.10 / 47.34	1.51 / 44.53	0.94 / 48.70	1.76 / 43.24	4.23 / 35.61	2.76 / 39.31	8.43 / 29.62	1.04 / 47.79	2.51 / 42.58
MFR-SR	2.71 / 39.47	1.05 / 47.71	1.22 / 46.40	1.75 / 43.27	1.10 / 47.30	2.06 / 41.85	-	-	-	-	-
DepthSR-Net	2.27 / 41.02	0.90 / 49.03	1.11 / 47.21	1.31 / 45.81	0.96 / 48.47	1.57 / 44.23	3.46 / 37.36	2.50 / 40.16	6.95 / 31.30	1.10 / 47.26	2.21 / 43.19
RYNet	**2.04 / 41.94**	**0.73 / 50.88**	**0.97 / 48.40**	**1.21 / 46.47**	**0.81 / 49.96**	**1.41 / 45.13**	**3.33 / 37.68**	**2.28 / 40.97**	**6.84 / 31.44**	**0.70 / 51.26**	**2.03 / 44.41**
					16×						
Bicubic	8.16 / 29.89	3.34 / 37.64	2.63 / 39.74	5.10 / 33.97	2.86 / 39.00	5.85 / 32.78	7.67 / 30.43	6.09 / 32.44	16.57 / 23.75	3.89 / 36.32	6.22 / 33.60
RCAN	4.52 / 35.03	1.71 / 43.49	1.86 / 42.72	2.70 / 39.49	1.73 / 43.36	3.18 / 38.09	6.64 / 31.69	4.52 / 35.02	13.44 / 25.56	1.95 / 42.31	4.22 / 37.68
SAN	4.33 / 35.41	1.67 / 43.68	1.87 / 42.69	2.63 / 39.75	1.76 / 43.21	3.16 / 38.12	6.60 / 31.74	4.52 / 35.02	14.39 / 24.97	1.81 / 42.97	4.27 / 37.76
Aodha *et al.*	10.74 / 27.51	4.54 / 35.00	4.76 / 34.57	6.43 / 31.96	4.80 / 34.50	8.07 / 30.00	11.52 / 26.90	8.56 / 29.48	20.75 / 21.79	6.05 / 32.49	8.62 / 30.42
ATGV-Net	8.70 / 29.34	3.51 / 37.22	2.80 / 39.18	5.36 / 33.55	3.06 / 38.42	6.21 / 32.26	8.00 / 30.07	6.44 / 31.95	17.10 / 23.47	4.14 / 35.79	6.53 / 33.13
DJFR	7.00 / 31.23	3.47 / 37.31	2.47 / 40.28	4.72 / 34.66	2.88 / 38.96	4.94 / 34.26	8.65 / 29.39	5.69 / 33.03	17.04 / 23.50	3.26 / 37.86	6.01 / 34.05
JID	10.30 / 27.87	8.64 / 29.40	6.70 / 31.60	10.01 / 28.13	9.17 / 28.88	5.69 / 33.03	14.30 / 25.03	6.61 / 31.72	16.53 / 23.77	5.50 / 33.33	9.35 / 29.28
TSDR	4.30 / 35.45	1.59 / 44.10	-	2.19 / 41.26	-	3.12 / 38.22	-	**3.18 / 38.18**	14.13 / 25.13	-	-
MSG-Net	4.17 / 35.72	1.70 / 43.54	1.63 / 43.89	2.63 / 39.74	1.69 / 43.59	2.92 / 38.83	6.19 / 32.29	3.95 / 36.20	13.83 / 25.31	1.74 / 43.30	4.05 / 38.24
MFR-SR	4.35 / 35.36	1.78 / 43.12	1.74 / 43.32	3.01 / 38.56	1.73 / 43.37	3.74 / 36.67	-	-	-	-	-
DepthSR-Net	3.91 / 36.29	1.54 / 44.36	1.54 / 44.39	2.34 / 40.73	1.56 / 44.29	2.44 / 40.37	5.28 / 33.67	3.72 / 36.71	13.02 / 25.84	1.48 / 44.75	3.68 / 39.14
RYNet	**3.37 / 37.57**	**1.37 / 45.37**	**1.37 / 45.42**	**2.01 / 42.08**	**1.37 / 45.43**	**2.22 / 41.20**	**4.40 / 35.27**	3.67 / 36.84	**10.76 / 27.50**	**1.22 / 46.41**	**3.18 / 40.31**

TABLE 2. Quantitative comparison (in RMSE / PSNR (dB)) on noisy data (8×).

Method	Art	Book	Dolls	Laundry	Moebius	Reindeer	Cones	Teddy	Tsukuba	Venus	Avg.
Bicubic	6.85 / 31.42	4.77 / 34.57	4.55 / 34.96	5.38 / 33.52	4.58 / 34.92	5.75 / 32.93	6.89 / 31.36	5.73 / 32.97	13.00 / 25.85	5.08 / 34.01	6.26 / 32.65
RCAN	4.16 / 35.75	1.90 / 42.54	2.39 / 40.55	2.82 / 39.12	2.13 / 41.55	2.96 / 38.71	6.12 / 32.40	4.07 / 35.94	10.16 / 27.99	2.24 / 41.12	3.90 / 37.57
SAN	4.20 / 35.67	1.89 / 42.60	2.30 / 40.90	2.88 / 38.94	2.10 / 41.67	3.07 / 38.38	6.45 / 31.94	4.09 / 35.91	9.98 / 28.15	2.12 / 41.59	3.91 / 37.58
ATGV-Net	4.53 / 35.01	2.35 / 40.70	2.49 / 40.19	3.23 / 37.96	2.45 / 40.35	3.27 / 37.84	5.85 / 32.78	4.12 / 35.84	11.80 / 26.70	2.68 / 39.57	4.28 / 36.69
MSG-Net	4.30 / 35.46	2.30 / 40.89	2.48 / 40.24	3.10 / 38.29	2.60 / 39.84	3.25 / 37.89	5.79 / 32.88	4.03 / 36.02	11.19 / 27.15	2.55 / 40.02	4.16 / 36.87
MFR-SR	3.97 / 36.16	2.13 / 41.56	2.13 / 41.56	3.01 / 38.56	2.82 / 39.13	**2.25 / 41.09**	5.01 / 34.13	3.79 / 36.56	9.95 / 28.17	**1.99 / 42.15**	3.71 / 37.91
DepthSR-Net	3.96 / 36.17	2.08 / 41.77	2.25 / 41.07	2.75 / 39.33	2.18 / 41.35	2.95 / 38.72	5.37 / 33.53	3.85 / 36.43	10.76 / 27.50	2.13 / 41.58	3.83 / 37.75
RYNet	**3.47 / 37.33**	**1.88 / 42.66**	**1.97 / 42.23**	**2.47 / 40.29**	**1.87 / 42.67**	2.68 / 39.55	**4.74 / 34.61**	**3.45 / 37.37**	**9.76 / 28.34**	2.18 / 41.37	**3.45 / 38.64**

RMSE and average PSNR values. Our average RMSE reductions and average PSNR gains over MFR-SR [45] and DepthSR-Net [46] are 0.26/0.73dB and 0.38/0.89dB, respectively.

The visual comparison of the noisy depth map Art for 8× SR is provided in Fig. 6. It can be observed from the highlighted regions that although most of the learning-based methods without color guidance successfully suppress noise,

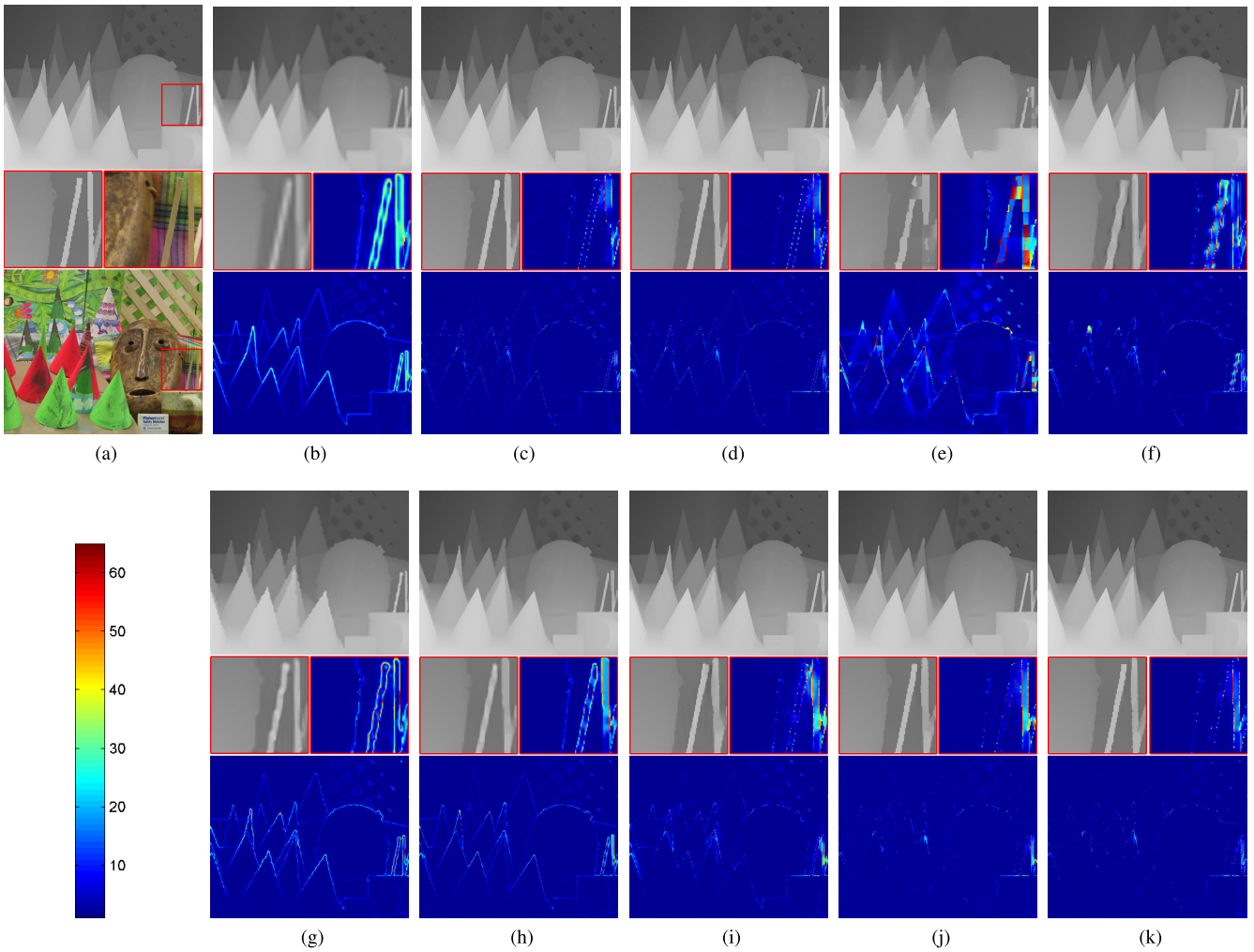

FIGURE 4. Visual comparison of depth map Cones for 4× SR. (a) Original depth map and its aligned color image (RMSE/PSNR(dB)), (b) Bicubic (3.84/36.45), (c) RCAN (2.12/41.61), (d) SAN (2.13/41.57), (e) Aodha *et al.* (5.22/33.78), (f) ATGV-Net (2.91/38.84), (g) DJFR (3.55/37.11), (h) JID (2.85/39.03), (i) MSG-Net (2.60/39.85), (j) DepthSR-Net(2.33/40.77), (k) RYNet (1.92/42.46).

they produce over-smooth boundaries and inaccurate depth details, such as the incorrect lip shapes in Fig. 6 (d)-(f). The result of MSG-Net [4] is somewhat blurred. DepthSR-Net [46] preserves the edge sharpness well, but still fails to recover some depth details, e.g. the disappearing vertical stick in the bottom-right highlighted region of Fig. 6 (h). By contrast, the proposed RYNet performs well in suppressing noise, and recovering sharp depth edges and accurate depth details. The quantitative and qualitative comparisons verify the robustness of RYNet to noise.

D. EXPERIMENT ON REAL DATA

In this subsection, we perform experiment on real data from [36] captured by Microsoft Kinect camera. Kinect simultaneously returns a 512 × 424 depth map and a 1920 × 1080 color image. Besides the low resolution defect, Kinect depth maps also frequently contain depth holes, such as structural missing values along depth edges and random missing values in flat areas. As the preliminary work on the depth map SR of real data, we first calibrate the depth maps and the color images under the resolution of 480 × 270 according to the calibration parameters, and fill the depth holes by interpolation. The hole-filled depth maps are then up-sampled to 1920 × 1080 by Bicubic interpolation and used as the depth input of RYNet. Due to the limited space, we only provide the visual comparison with TSDR [36] and DepthSR-Net [46] on depth map Yoga in Fig. 7. It can be seen that TSDR [36] produces smooth reconstruction results. Both DepthSR-Net [46] and the proposed RYNet perform well in restoring depth edge sharpness. But the reconstruction of DepthSR-Net [46] appears somewhat distorted, such as in the railing regions. In contrast, the proposed RYNet makes an accurate structural reconstruction.

E. STATISTICAL EVALUATION ON DEPTH MAP DATASET

To evaluate the statistical performance of the proposed RYNet, we carry out experiments on three public benchmark datasets: Middlebury [60] (60 RGB-D image pairs),

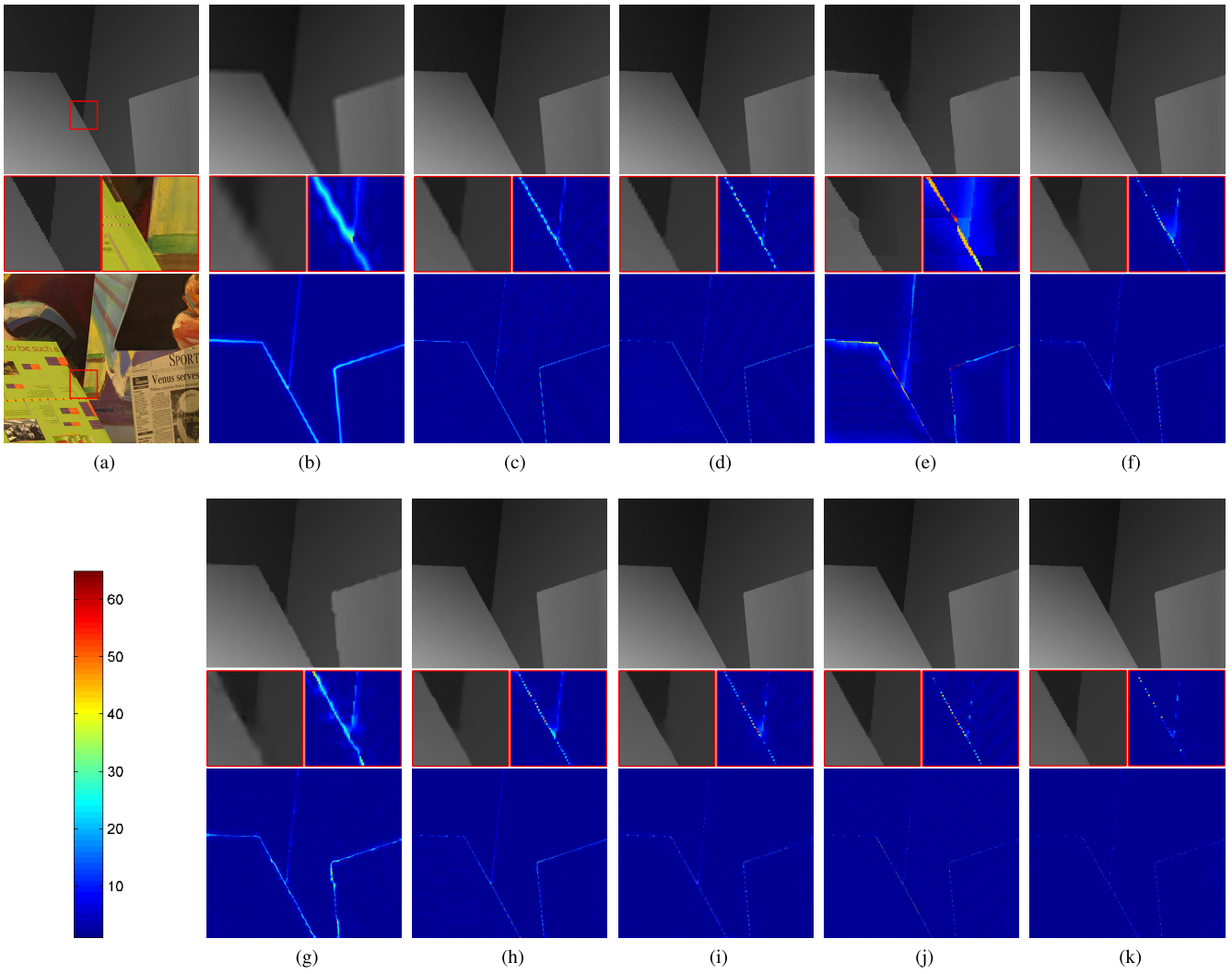

FIGURE 5. Visual comparison of depth map Venus for 8× SR. (a) Original depth map and its aligned color image (RMSE/PSNR(dB)), (b) Bicubic (2.75/39.35), (c) RCAN (1.50/44.61), (d) SAN (1.38/45.31), (e) Aodha *et al.* (3.21/37.99), (f) ATGV-Net (1.10/47.27), (g) DJFR (2.22/41.22), (h) JID (1.23/46.35), (i) MSG-Net (1.04/47.79), (j) DepthSR-Net(1.10/47.26), (k) RYNet (0.70/51.26).

TABLE 3. Average RMSE / PSNR (dB) on three depth datasets (4×).

Method	Middlebury	ToFMark	NYU_Depth
Bicubic	6.52 / 37.27	2.62 / 40.00	2.36 / 41.44
RCAN	1.29 / 47.34	1.40 / 45.43	1.34 / 45.84
SAN	1.29 / 47.39	1.37 / 45.58	1.30 / 46.10
ATGV-Net	1.49 / 47.34	1.45 / 45.23	1.28 / 46.31
MSG-Net	1.28 / 48.07	1.41 / 45.47	1.31 / 46.18
DepthSR-Net	1.19 / 47.84	1.37 / 45.59	1.34 / 45.88
RYNet	**0.98 / 50.34**	**1.21 / 46.75**	**1.06 / 48.02**

ToFMark [35] (3 RGB-D image pairs of real scenes), and NYU Depth v2 [61] (1449 RGB-D image pairs of real scenes).

For the limited space, six representative or state-of-the-art methods from Section IV-B are selected as baselines, including Bicubic interpolation, RCAN [10], SAN [11], ATGV-Net [22], MSG-Net [4], and DepthSR-Net [46]. The average RMSE and PSNR values for 4× depth map SR achieved by these methods are shown in Table 3. The proposed method achieves the lowest RMSE and the best PSNR on all the datasets. It is followed by ATGV-Net [22], MSG-Net [4], SAN [11], and DepthSR-Net [46] according to the general results. More exactly, our average RMSE reductions and PSNR gains of all the 1512 RGB-D image pairs in three datasets over ATGV-Net [22], MSG-Net [4], SAN [11], and DepthSR-Net [46] are 0.24/1.76dB, 0.25/1.85dB, 0.25/1.96dB, and 0.27/2.16dB, respectively. As shown in Fig. 8, we also provide the RMSE/PSNR improvement probability distributions over the four baselines to better illustrate the statistical performance of the proposed method. The RMSE reduction distributions and PSNR gain distributions are all positively biased, proving that the proposed method is statistically superior to the competing baselines.

The NYU Depth v2 dataset is composed of 464 video scenes taken with a Kinect. There are unavoidable local structural misalignment between hole-filled Kinect depth maps and color images, which may deteriorate the SR performance

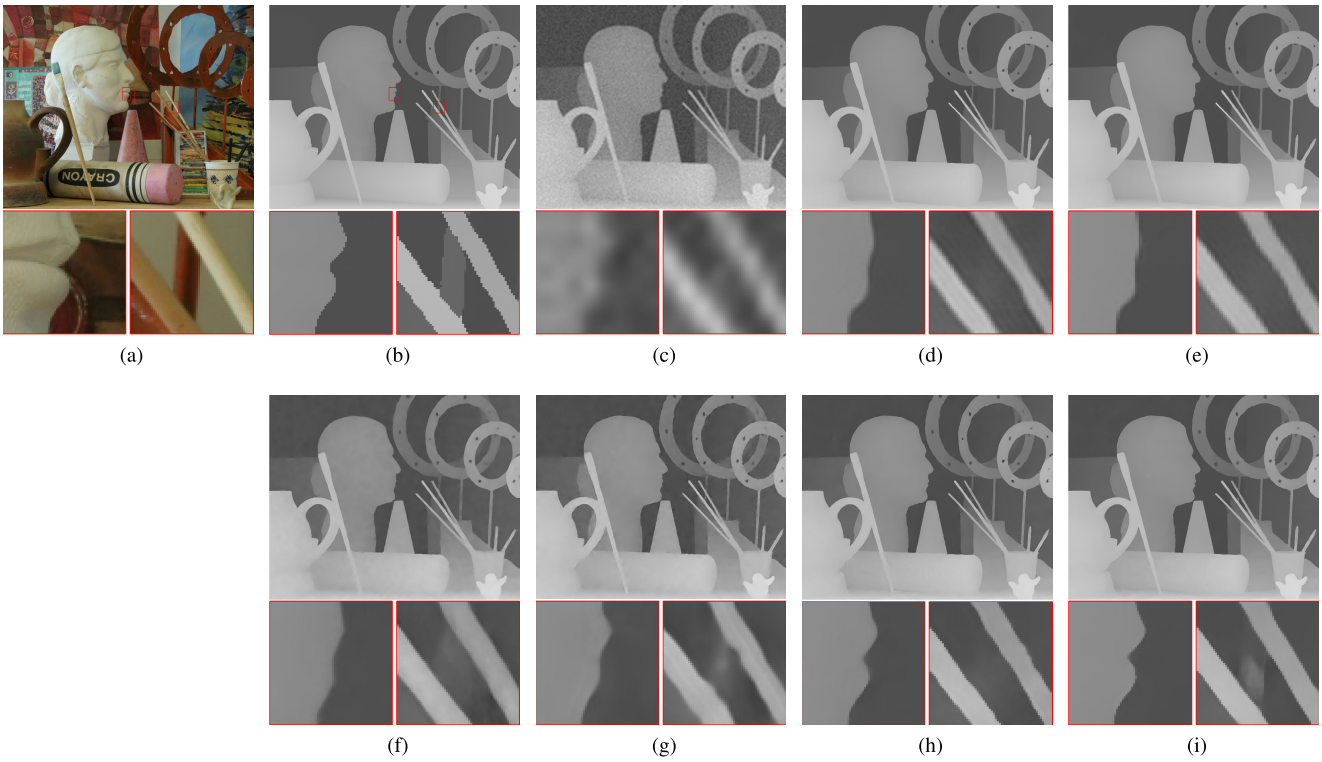

FIGURE 6. Visual comparison of noisy depth map Art for 8× SR. (a) Color guidance image, (b) original depth map (RMSE/PSNR(dB)) (c) Bicubic (6.85/31.42), (d) RCAN (4.16/35.75), (e) SAN (4.20/35.67), (f) ATGV-Net (4.53/35.01), (g) MSG-Net (4.30/35.46), (h) DepthSR-Net(3.96/36.17), (i) RYNet (3.47/37.33).

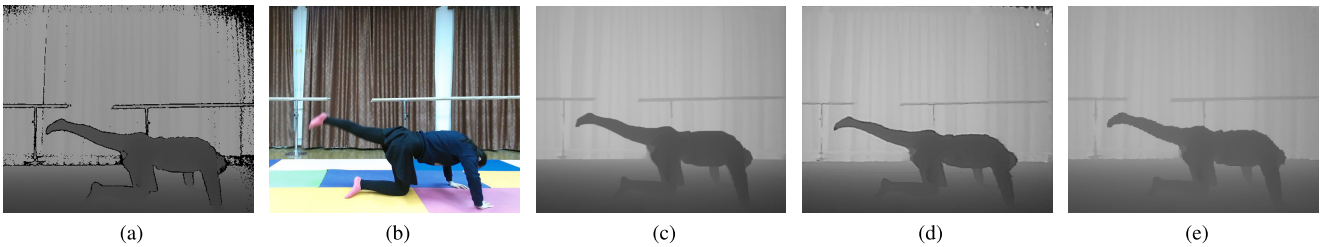

FIGURE 7. Visual comparison of real data Yoga captured by Kinect. From left to right, (a) Raw LR depth maps from [36], (b) the HR color image, and the reconstruction results of (c) TSDR, (d) DepthSR-Net, and (e) RYNet.

of most depth map SR methods with color guidance. This explains why the single depth map SR methods such as SAN [11] and ATGV-Net [22] outperform MSG-Net [4] and DepthSR-Net [46] on the NYU Depth v2 dataset. However, the proposed method still exhibits robustness to the local depth-color structural misalignment due to the increase of the feature's scale diversity and the adoption of spatial attention. As shown in Fig. 9, the fake edge artifacts caused by misalignment are obvious in DepthSR-Net [46], especially along the edges in the bottom-left highlighted regions. The proposed method not only effectively suppresses the fake edge artifacts, but also performs well in recovering depth details.

F. ABLATION ANALYSIS

In this subsection, we perform ablation experiments to verify our network choices, including multi-scale encoders-decoder

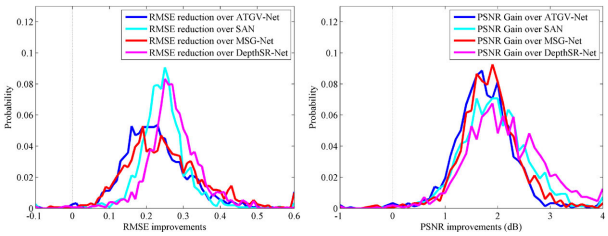

FIGURE 8. The probability distributions of RMSE reductions and PSNR gains of the proposed RYNet over four state-of-the-art baselines.

(MSED) architecture, RCA-ASPP block, spatial attention (SA), and RCAB. Due to the limited space, instead of enumerating all possible combinations, we select six representative network models (from M-1 to M-6) with different combinations of MSED, RCA-ASPP, SA, and RCAB as

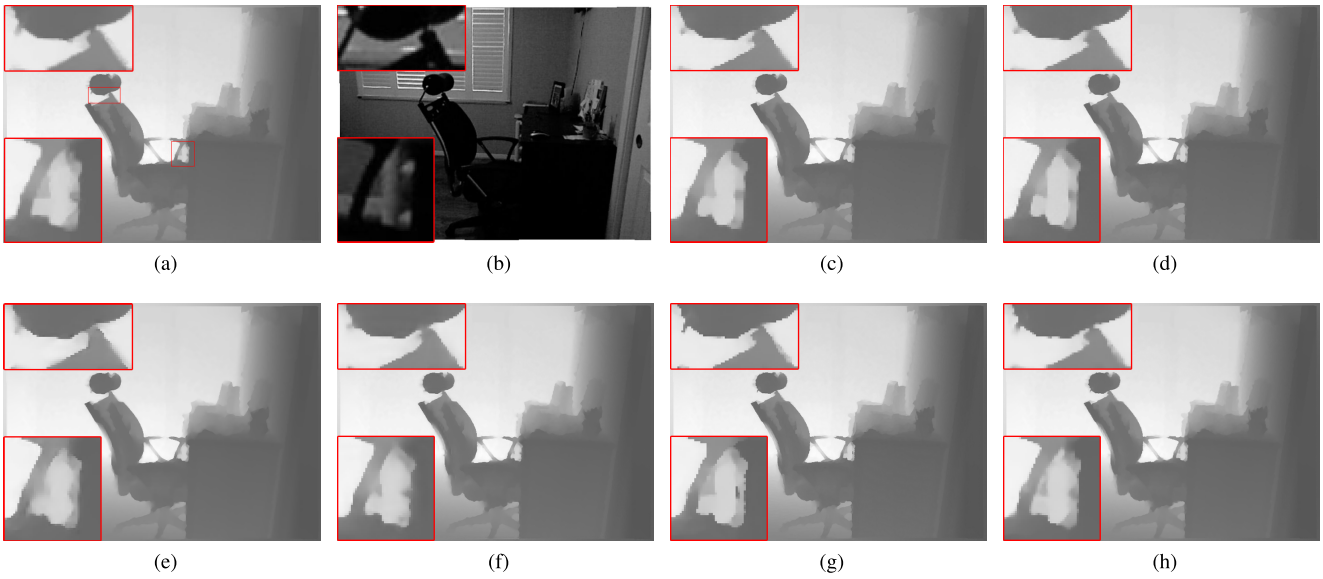

FIGURE 9. An example of the SR results on NYU Depth v2 dataset. (a) Original depth map (RMSE/PSNR(dB)), (b) the aligned color image, (c) RCAN (1.74/43.30), (d) SAN (1.66/43.72), (e) ATGV-Net (1.77/43.18), (f) MSG-Net (1.84/42.83), (g) DepthSR-Net (1.91/42.51), (h) RYNet (1.42/45.08).

TABLE 4. Ablation investigation of RYNet (8×).

Different combinations of MSED, RCA-ASPP, SA, and RCAB						
Structure name	M-1	M-2	M-3	M-4	M-5	M-6
MSED	×	×	√	√	√	√
RCA-ASPP	×	√	×	√	√	√
SA	×	√	√	×	√	√
RCAB	×	√	√	√	×	√
Average RMSE	2.64	2.38	2.11	2.06	2.09	2.03
Average PSNR (dB)	41.82	42.62	43.95	44.07	43.98	44.41

shown in Table 4. The model M-2 is actually a single-scale encoders-decoder architecture with $L = 1$. In model M-3, we directly concatenate F_D^L and F_I^L in the junction module. In model M-4, the intensity encoder features F_I^l without reweighting are directly fused with the depth encoder features and the decoder features from the preceding scale level. The model M-5 substitutes RCAB with a residual block. M-6 is exactly the complete RYNet model used in previous experiment sections.

As listed in Table 4, compared to M-6, network models M-2 to M-5 suffer an increase of 0.35, 0.08, 0.03, and 0.06 in average RMSE, and a decrease of 1.79dB, 0.46dB, 0.34dB, and 0.43dB in average PSNR, respectively. It proves that the adoption of MSED, RCA-ASPP, SA, and RCAB can separately improve the performance. In addition, the comparison between single-scale model M-2 with $L = 1$ and multi-scale model M-6 with $L = 4$ indicates the more scale diversity, the better. The considerable performance gap between the basic model M-1 and the complete model M-6 clearly demonstrates that all our network choices work coherently and make positive complementary contributions to performance.

G. TRAINING CONVERGENCE

In this subsection, we visualize the convergence process of the proposed RYNet. The average PSNR curves on the ten test depth maps in Section IV-B and ToFMark dataset versus training epochs are shown in Fig. 10. The trends are similar for different scaling factors, so only the curve for 2× SR is provided. Clearly, with the growth of epochs, the PSNR curves increase monotonically and reach their convergent states after about 30 epochs. Hence, the maximal epoch number 80 seems appropriate.

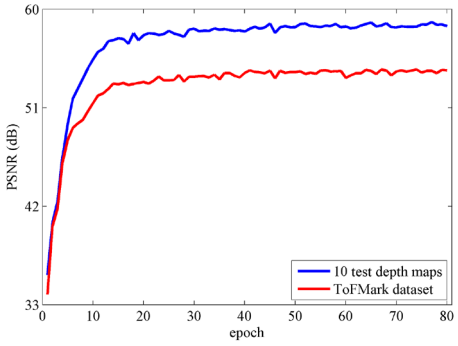

FIGURE 10. The PSNR (dB) values of RYNet on ten test depth maps in Section IV-B and ToFMark dataset with different training epochs.

H. RUNNING TIME

We evaluate the running time on the same desktop computer with a NVIDIA RTX 2080Ti GPU and Intel Core i9

TABLE 5. Average running time (sec) on 1080 × 1320 test depth maps.

Bicubic	RCAN	SAN	Aodha *et al.*	ATGV-Net
-	0.45	4.99	743.45	0.39
DJFR	**JID**	**MSG-Net**	**DepthSR-Net**	**RYNet**
5.62	869.60	0.37	1.87	0.71

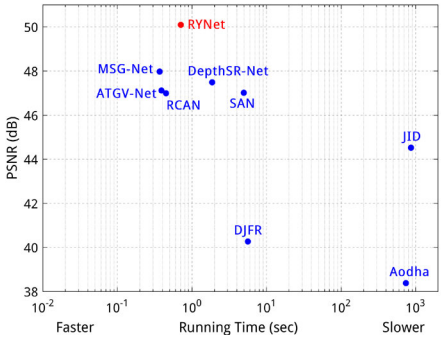

FIGURE 11. Running time and accuracy trade-off for different depth map SR methods.

3.5GHz GHz CPU. Table 5 lists the average running time of different methods for 4× SR of 1080 × 1320 test depth maps. The running time of Bicubic interpolation is negligible. Generally, deep learning-based methods are computationally more efficient than optimization-based methods and traditional learning-based methods. The speed of the proposed RYNet is faster than all the competing methods except RCAN [10], ATGV-Net [22], and MSG-Net [4]. Fig. 11 presents the trade-offs between the running time and PSNR performance on the ten test depth maps in Section IV-B. The results show that the proposed RYNet is an appropriate choice for depth map SR with comprehensive consideration of computational complexity and performance.

V. CONCLUSION

In this paper, we propose a recumbent Y network (RYNet) using an encoders-decoder architecture to fuse depth information and intensity information for the depth map SR task. Our RYNet introduces a residual channel attention based atrous spatial pyramid pooling structure to enrich feature's scale diversity and adapt the network to focus on feature channels with more informative scales. We also introduce a spatial attention mechanism to increase the spatial correlation between depth features and intensity guidance features to mitigate the texture-transfer and depth-bleeding artifacts. Comparison to recent depth map SR methods has shown that our RYNet achieves the state-of-the-art performance for different scaling factors. In future, we will extend this work to other computer vision tasks with multiple inputs and single output.

REFERENCES

[1] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.
[2] W. Dong, G. Shi, X. Li, K. Peng, J. Wu, and Z. Guo, "Color-guided depth recovery via joint local structural and nonlocal low-rank regularization," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 293–301, Feb. 2017.
[3] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 152–157.
[4] T. W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 353–369.
[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comp. Vis.*, 2014, pp. 184–199.
[6] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.
[7] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[8] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[9] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3637–3641.
[10] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[11] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.
[12] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *Proc. ICLR*, 2019, pp. 11065–11074.
[13] X. Xu and X. Li, "SCAN: Spatial color attention networks for real single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.
[14] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.
[15] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3867–3876.
[16] M. Hornacek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1123–1130.
[17] J. Lei, L. Li, H. Yue, F. Wu, N. Ling, and C. Hou, "Depth map super-resolution considering view synthesis quality," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1732–1745, Apr. 2017.
[18] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 71–84.
[19] 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.
[20] D. Ferstl, M. Ruther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 513–521.
[21] S. Mandal, A. Bhavsar, and A. K. Sao, "Depth map restoration from undersampled data," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 119–134, Jan. 2017.
[22] G. Riegler, M. Ruther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 268–284.
[23] 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, Aug. 2019.
[24] 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.
[25] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 169–176.
[26] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.
[27] K.-H. Lo, Y.-C.-F. Wang, and K.-L. Hua, "Edge-preserving depth map upsampling by joint trilateral filter," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 371–384, Jan. 2018.

[28] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. NIPS*, 2005, pp. 291–298.
[29] J. Li, Z. Lu, G. Zeng, R. Gan, and H. Zha, "Similarity-aware patchwork assembly for depth image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3374–3381.
[30] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 158–171.
[31] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3443–3458, Aug. 2014.
[32] Y. Li, D. Min, M. N. Do, and J. Lu, "Fast guided global interpolation for depth and motion," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 717–733.
[33] 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.
[34] K. Zhou, S. Yu, and C. Jung, "Alternately guided depth super-resolution using weighted least squares and zero-order reverse filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1847–1851.
[35] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
[36] Z. Jiang, Y. Hou, H. Yue, J. Yang, and C. Hou, "Depth super-resolution from RGB-D pairs with transform and spatial domain regularization," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2587–2602, May 2018.
[37] X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao, "Depth super-resolution via joint color-guided internal and external regularizations," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1636–1645, Apr. 2019.
[38] S. Yu, H. Lan, and C. Jung, "Intensity guided depth upsampling using edge sparsity and super-weighted L_0 gradient minimization," *IEEE Access*, vol. 7, pp. 140553–140565, 2019.
[39] M. Kiechle, S. Hawe, and M. Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1545–1552.
[40] I. Tosic and S. Drewes, "Learning joint intensity-depth sparse representations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2122–2132, May 2014.
[41] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof, "A deep primal-dual network for guided depth super-resolution," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–14.
[42] W. Zhou, X. Li, and D. Reynolds, "Guided deep network for depth map super-resolution: How much can color help?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1457–1461.
[43] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Joint image filtering with deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1909–1923, Aug. 2019.
[44] 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, Feb. 2019.
[45] 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, Feb. 2020.
[46] 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.
[47] 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.
[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[49] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[51] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019.
[52] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782.
[53] X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth, "RUNet: A robust UNet architecture for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–3.
[54] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. NIPS*, 2017, pp. 406–416.
[55] P. Esser and E. Sutter, "A variational U-net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8857–8866.
[56] L. Han, X. Li, and Y. Dong, "Convolutional edge constraint-based U-net for salient object detection," *IEEE Access*, vol. 7, pp. 48890–48900, 2019.
[57] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images," in *Proc. MICCAI*, 2018, pp. 893–901.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[60] *The Middlebury Stereo Datasets*. Accessed: Jun. 2015. [Online]. Available: http://vision.middlebury.edu/stereo/data/
[61] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

TAO LI (Member, IEEE) received the B.S. degree in electronics and information engineering and the Ph.D. degree in communication and information system from Sichuan University, Chengdu, China, in 2005 and 2017, respectively. She joined the School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, in 2017. Her research interests include image/video compression and restoration, image/video super-resolution, and computer vision.

XIUCHENG DONG received the B.S. and M.S. degrees from Chongqing University, China, in 1985 and 1990, respectively. He is currently a Professor with the School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, China. He is also the Dean of the Electrical Engineering and Electronic Information, Xihua University. His research interests include modern control theory, adaptive control, and robotics.

HONGWEI LIN (Member, IEEE) received the M.S. degree in communication and information System from Xidian University, China, in 2011, and the Ph.D. degree in communication and information system from Sichuan University, China, in 2019. He is currently a Lecturer with the Electrical Engineering, Northwest Minzu University. His research interests include image processing and video compression and communication.