Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19200-0

CA-DBMNet: a channel attention based dual branch multi-scale network for depth map super-resolution

Yongwei Miao[1] · **Xinjie Zhang**[2] · **Yuliang Sun**[3] · **Jinrong Wang**[1]

Received: 28 July 2023 / Revised: 18 February 2024 / Accepted: 7 April 2024

Abstract
Scene depth information plays a fundamental role and can be beneficial to various computer vision or visual robotics applications. The scene color image acquired by consumer depth sensors usually has a high resolution, whilst its depth map counterpart often performs low resolution or man-made artifacts. Due to its strong similarity in terms of scene structures between RGB-D pairs, taking the color image as prior information, this paper proposes a Dual Branch Multi-scale Network (CA-DBMNet) based on the channel attention mechanism which can effectively guide the task of depth map super-resolution (SR). The network consists of two branches–color image feature extraction branch and depth map super-resolution branch. The first branch adopts the feature pyramid structure to extract the color image features, capturing image features and structures at different scales. The second branch is composed of three modules: 1) A dense residual feature fusion (DRFF) module to integrate the extracted features from two branches with dense connection and residual learning; 2) A channel multi-scale (CMS) module to exploit multi-scale features from depth feature maps; 3) A channel attention (CA) module to effectively enhance the channel proportion of high-frequency components in the depth feature maps. Extensive experiments demonstrate that CA-DBMNet can effectively reconstruct the high-resolution depth map with complete scene structures and sharp edges.

Keywords Channel attention · Depth map · Super-resolution (SR) · Dual branch · Multi-scale

✉ Yongwei Miao
ywmiao@hznu.edu.cn

Yuliang Sun
sunyuliang@zjsru.edu.cn

[1] School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China

[2] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China

[3] School of Information Science and Technology, Zhejiang Shuren University, Hangzhou 310015, China

Published online: 16 April 2024

1 Introduction

Scene depth map super-resolution (SR) is an important issue in the literature of computer vision and visual robotics. The depth information of 3d scenes acquired by consumer depth sensors always plays a fundamental role in various real-world applications of augmented reality [1], scene segmentation [2], autonomous driving [3], 3d reconstruction [4], robotics navigation[5], etc. However, due to the limitations of the resolution of depth sensors or photosensitive devices, the low-resolution depth maps significantly affect the performance of these downstream applications. Therefore, it is imperative to devise effective depth map SR algorithms to enhance the utilization of depth information.

In general, the difficulty of depth map SR is that it always leads to some unavoidable loss of scene structures or fine details [6], which becomes even more severe as the down-sampling factor is increased. To tackle this issue, many researchers always adopt color images to guide the SR task of the corresponding low-resolution depth map, such as optimized-based methods [7–9], dictionary learning-based [10, 11] and deep learning-based schemes [12, 13]. The intuition behind these methods is that color images and their corresponding depth maps represent the photometric color and geometric depth of the same scene from the same perspective, and they always have strong structural similarities. Although the guidance from high-resolution color images within the depth map SR reconstruction can effectively alleviate the issue of edge and detail blurring, the texture information such as flash-reflection patterns on the object surface in the color image may lead to man-made artifacts on the reconstructed depth map. The key to color-guided schemes is to excavate the rich structural information in color images and suppress texture copy artifacts.

Recently, some methods have exploited multi-scale information for better feature extraction. Hui et al. [14] presented a multi-scale guided convolutional neural network (CNN) to realize a depth map SR task. This network learns the rich feature information of scene depth maps at different scales so that it can better adapt to the fine structures of depth maps and be suitable for the task of large-scale depth upsampling. Similarly, Zuo et al. [15] introduced a residual network structure at each scale in the upsampling step to effectively recover the high-frequency details of depth maps. These methods always adopt one or two convolutional networks at each scale while extracting the scene features of high-resolution color images, but it is still difficult to achieve a desirable reconstruction of the fine structures in depth maps. Different from the aforementioned works, our network learns multi-scale information on both color images and depth maps. In addition, the proposed network integrates a channel attention mechanism for better structure reconstruction and artifact suppression.

With the high-resolution scene color image as prior information, here we propose a Channel Attention based Dual Branch Multi-scale Network (CA-DBMNet) to effectively guide depth map SR. The proposed network consists of two branches, i.e., the color image feature extraction branch, and the depth map SR branch, as shown in Fig. 1. The color image feature extraction branch uses the feature pyramid structure, which consists of a down-sampling operation for shallow feature extraction, an upsampling procedure for deep feature extraction, and a skip-connection process for feature fusion. The structural information at multiple scales can be learned by such down-sampling and upsampling operations. The depth map SR branch comprises a dense residual feature fusion (DRFF) module, a channel multi-scale (CMS) module, and a channel attention (CA) module. To realize the high-resolution color image guidance, the DRFF module adopts dense connection and residual connection to fuse the feature maps from two branches. CMS module can effectively learn the structural information of depth maps at different scales by average group convolution to enlarge multi-level

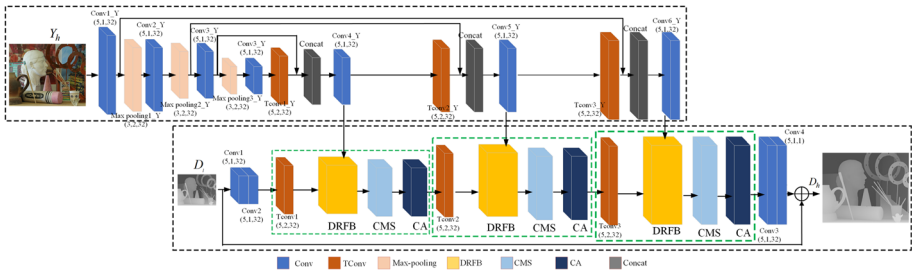

Fig. 1 The network architecture of our proposed CA-DBMNet. This network consists of the color image feature extraction branch and the depth map SR branch. The top branch takes a high-resolution color image as input, uses a feature pyramid to extract multi-scale features, and guides depth map SR. The bottom branch takes a low-resolution depth map as input and employs a dense residual feature fusion (DRFF) module, a channel multi-scale (CMS) module, and a channel attention (CA) module to effectively reconstruct the corresponding high-resolution depth map.

receptive fields. CA module assigns different weights to the channels of the scene feature map, which can effectively reflect the differences between channels. The adaptive channel weights are used to reduce the influence of high-resolution color image artifacts on the depth map SR. The proposed network has achieved gratifying performance in various cases and possesses significant potential to improve the performance of real-world downstream tasks, such as object detection, scene reconstruction, and semantic segmentation. The main technical contributions can be summarized as follows:

- A feature extraction branch of the scene color image based on the feature pyramid structure is proposed, effectively extracting the shallow and deep features of the high-resolution color image. It combines features by skip connection to extract rich features of the color image and provide better guidance for depth map SR.
- A channel attention module for scene depth feature map based on attention mechanism is presented. The module can adaptively learn the weights of different channels of the depth feature map, and effectively suppress the artifacts caused by high-resolution color images, thus realizing the effective reconstruction of high-frequency structures such as edge information of the depth map.
- A dual branch multi-scale network is introduced for the task of depth map SR. This network can simultaneously extract multi-scale features of color images and depth maps, and also fuse the image structural information at each scale to guide the depth map SR.

The rest of this paper is organized as follows. In Section 2, the related works of depth map SR and learning-based multi-scale feature extraction are reviewed. The details of the proposed network CA-DBMNet are given in Section 3, including the overall framework and key modules. Experimental results are presented in Section 4. Section 5 concludes this paper and gives future improvements.

2 Related works

2.1 Single image super-resolution

Deep learning methods have achieved dramatic performance in single image SR tasks. Liu et al. [16] proposed a multi-scale encoder-decoder network with the guidance of a phase

congruency edge map. Aiming at small-scale pedestrian detection, Pang et al. [17] proposed a unified framework that integrates SR and classification sub-networks. Similarly, Liu et al. [18] proposed a joint SR and deblurring network with decoupled cooperative learning. To capture long-range feature similarities, Mei et al. [19] integrated a Cross-Scale Non-Local Attention module into a recurrent neural network. To tackle the case of unpaired image sets, Maeda [20] designed a generative adversarial framework that produced pseudo-clean low-resolution images for SR network training. Wang et al. [21] proposed a lightweight network DDistill-SR that captured and reused important information through a plug-in reparameterized dynamic distillation unit. Lee et al. [22] proposed an optimization method in the network training process by removing the inherent noise. Based on the assumption that underlying image distribution is scale-invariant, Scanvic et al. [23] introduced a new self-supervised loss for SR network training. More recently, Wu et al. [24] proposed a self-attention-free network CFSR that utilizes large kernel convolution for lightweight feature extraction.

Most single image SR networks are CNN-based, while recent Transformer-based SR models showcase better performance. Due to the differences in resolution and structure between depth maps and color images, these methods applicable to RGB images cannot be directly used for depth map SR tasks.

2.2 Depth map super-resolution

Among different methods for depth map SR, color-guided approaches are most commonly-seen. These works take low-resolution depth maps and the corresponding high-resolution color images as joint inputs, where RGB information plays as guidance. The intuition behind these methods is that color images contain rich structural information and accurately represent the visual characteristics of the captured scenes. Existing studies can be classified into optimized-based methods, dictionary learning-based, and deep learning-based schemes.

The optimized-based schemes usually adjust color image guidance by manually modeling depth smoothness. Zuo et al. [7] proposed an explicit evaluation model to quantitatively measure the inconsistency between the depth edge map and the color edge map. Khoddami et al. [8] introduced a structure-preserving guided filter for depth map SR which can overcome the defects of depth maps, such as texture-copying artifacts, halo artifacts, and blurring edges. Wang et al. [9] presented a depth map enhancement method based on a dual normal-depth regularization with a re-weighted graph Laplacian prior, which constrains edge consistency between the surface normal map and depth map.

Benefiting from the dictionary learning strategy, Liu et al. [10] presented a depth map SR method which employed a joint dictionary learning method with both low- and high-resolution depth maps and thus built a sparse vector classification scheme that can be used in depth map SR. Li et al. [11] presented a scheme for depth image SR based on multi-dictionary learning with an edge regularization model, which can learn three dictionaries of three parts based on the assumption that the low-resolution, high-resolution, and edge-depth images share the same sparse representation.

The deep-learning based approaches always learn from scene datasets and recover depth maps through deep neural networks. Owing to a joint bilateral filter, Li et al. [25] applied a CNN to the task of depth upsampling. This network jointly extracts the structural features from both color and depth images and concatenates feature maps through another sub-network. Ye et al. [26] first learned a binary map from a low-resolution depth map and the corresponding color image, and thus reconstructed a high-quality depth map through an edge-guided interpolation. Zhu et al. [27] designed CNNs for concurrent edge detection and

depth map interpolation operation. Using a color-guided strategy, Wen et al. [12] introduced a progressive method to recover the high-frequency details of depth maps while alleviating man-made artifacts. Lutio et al. [28] regarded the transformation from color image to corresponding depth map as a pixel-wise translation during depth map enhancement.

Although the structural information of color images can guide the task of depth map SR, depth-color inconsistency may occur due to color image texture copying. The aforementioned works used various strategies to alleviate such artifacts. Different from these methods, we design a dual-branch multi-scale network integrated with a channel attention mechanism to suppress texture-copy artifacts.

2.3 Multi-scale feature extraction

Multi-scale feature extraction in CNNs plays a vital role in various computer vision and visual robotics tasks such as object detection [29], semantic segmentation [30, 31], and SR enhancement [15]. To effectively extract the multi-scale features of scene images, some works adopt multi-scale input, skip-connection, or recursive models. He et al. [29] presented a spatial pooling network for object detection, which can feed images at different scales to enhance multi-scale feature representation. Long et al. [30] introduced a fully convolutional multi-scale representation network, exploiting feature maps at different scales by convolution operation to perform better semantic segmentation results. Owing to the convolutional network architecture, Ronneberger et al. [31] utilized a U-net structure to extract multi-scale information for image segmentation by down-sampling and upsampling operations. Apart from these methods, another multi-scale strategy is to integrate dilated convolution into the backbone network and get scene features at multiple scaling factors for extracting context information. Song et al. [32] treated depth map SR as a series of novel view synthesis sub-tasks, and used a multi-scale fusion strategy to effectively exploit the feature maps at different scales. To deal with the issue of depth map SR with large scaling factors, Zuo et al. [15] adopted multi-scale frequency synthesis with local residual learning to extract rich features. This network can maintain spatial information and extract fine structures on multi-scale depth feature maps.

Different from the existing methods, we combine a multi-scale scheme not only in the depth map SR branch but also in the color image feature extraction branch. Furthermore, by incorporating channel-wise attention, we also assign different weights for multi-scale feature maps to reconstruct the high-frequency information of depth maps.

3 The proposed CA-DBMNet

Owing to the channel attention mechanism, our proposed network CA-DBMNet adopts the high-resolution scene color images as prior and reconstructs corresponding depth maps. The structural information extracted from different depth map channels may be inconsistent, which could lead to artifacts during the reconstruction of the depth map. Our CA-DBMNet utilizes channel-wise attention to adaptively recalibrate the weights of each channel and excavate high-frequency features. The conventional color-guided approach [27] only considers color-depth fusion at a single scale and ignores fine details under different scaling factors. Owing to a dual branch multi-scale framework, this paper applies a feature pyramid structure to both the color image feature extraction branch and the depth map SR branch, as shown in Fig.1. In the network branch of color image feature extraction, we adopt the feature pyra-

mid structure that consists of top-down and subsequent bottom-up convolutional operations to learn the structural information. This feature pyramid can effectively extract multi-scale structural information of color images and also provide sound guidance for depth map SR. In the network branch of depth SR, we employ an image pyramid structure to progressively upscale the low-resolution depth map and perform the SR operation in a coarse-to-fine manner. This branch adopts a two-layer convolution to extract features and then applies the Bicubic interpolation to restore low-frequency information of the depth map. For each level of the pyramid structure, the dense connection, channel multi-scale, and channel attention are employed to effectively reconstruct high-frequency features of the depth map.

3.1 Network branch of color image feature extraction

In general, since the scene color image and its corresponding depth map represent the color and geometric depth of the same scene from the same perspective, there is a strong structural similarity between RGB-D pairs. The low-resolution depth map usually contains less high-frequency detail information, while the corresponding high-resolution color image of the same scene contains rich high-frequency texture information. Thus, we can employ the color image to assist depth map SR for a better reconstruction. The function of the color image feature extraction branch is mainly to provide structural information as a priori for the depth map SR branch. Under the guidance of high-resolution color images, unlike most enhancement methods [12] that only use a single-scale convolution layer for feature extraction, this paper adopts a multi-scale scheme to effectively extract shallow and deep structural information. In addition, the shallow features obtained after down-sampling is further fused with the deep features obtained after upsampling through skip connections, extracting rich structural information to guide depth map SR.

The proposed color image feature extraction branch can be divided into three parts. The first part is the down-sampling operation. In this part, the input high-resolution color image is gradually down-sampled through the convolution layer and max-pooling layer so that the down-sampled color image matches the same resolution as the input depth map. Specifically, as shown in Fig. 1, the upsampling factor of depth map SR is 8. The color map feature extraction branch undergoes three down-samplings, where the kernel of all convolution layers is 3×3 with the stride 1 and the channel number 32. The kernel, stride, and channel number of max-pooling layers are 3×3, 2, and 32, respectively. The second part is the upsampling operation. In this part, the low-resolution color feature map is upsampled by transposed convolution, reaching the same resolution as the depth feature map, which can guide the depth feature map SR at different scales. The kernel, stride, and channel number of these transposed convolution layers are 5×5, 2, and 32, respectively. The third part is the fusion operation. In this part, the initial color feature map and the upsampled color feature map are concatenated by skip connection. Then the convolution operation is carried out with kernel 3×3, 2 stride, and 32 channels. The above operations in the branch of color image feature extraction can be formulated as follows,

$$\begin{aligned} F_1^Y &= \sigma(W_1^Y * Y_h + b_1^Y), \\ F_i^Y &= Maxpool(F_{i-1}^Y), i \in \{2, 4, 6\} \\ F_j^Y &= \sigma(W_j^Y * F_{j-1}^Y + b_j^Y), j \in \{3, 5, 7\} \\ F_k^Y &= \sigma(W_k^Y \bullet F_{k-1}^Y + b_k^Y), k \in \{8, 11, 14\} \\ F_{k+2}^Y &= \sigma(W_{k+2}[F_j^Y, F_{k+2}^Y] + b_{k+2}^Y) \end{aligned} \tag{1}$$

where σ represents the activation function PRelu, $*$ and $\bullet$ are convolution and deconvolution respectively, W and b are the weights and the biases of convolution respectively. Y_h means the high-resolution color image, F_i^Y is the feature map after the max-pooling operation at i-th layer, F_j^Y represents the feature map extracted by j-th convolution layer, F_k^Y denotes the feature map after transposed convolution at k-th layer, and F_{k+2}^Y represents the feature map fused from shallow and deep features at the same scale.

3.2 Network branch of depth map super-resolution

The network branch of depth map SR adopts a low-resolution scene depth map as input and employs an end-to-end multi-scale upsampling scheme. To obtain the high-resolution depth map, this branch consists of a shallow feature extraction module, an upsampling module, a DRFF module, a CMS module, a CA module, and a depth map reconstruction module.

Depth map shallow feature extraction The scene features with different frequencies in the depth maps often require different reconstruction and processing strategies. For example, the traditional Bicubic interpolation has an excellent performance in dealing with low-frequency features such as smooth surfaces [14]. However, this interpolation scheme is not suitable for recovering high-frequency features such as sharp edges. In our network branch, the low-frequency features are first interpolated by Bicubic interpolation before feature extraction. Then the low-frequency components of the interpolated depth map are kept and later combined with the output of the network. This branch focuses on extracting the high-frequency components, which is beneficial to restoring the high-frequency structural information and reducing the computational cost. The input of this branch is the low-frequency depth map D_l, the features extracted from two-layer convolution are F_1 and F_2, as follows,

$$F_1 = \sigma(W_1 * D_l + b_1), F_2 = \sigma(W_2 * F_1 + b_2) \tag{2}$$

Depth map upsampling In the color image feature extraction branch, the high-resolution color image will be continuously down-sampled until it reaches the same resolution as the scene depth map. To maintain the same resolution between feature maps from both branches in the fusion stage, the low-resolution scene depth map is also upsampled $2\times$ in an end-to-end manner. The upsampling operation is realized by transposed convolution, which has 5×5 kernel size with stride and channel numbers set to 2 and 32, respectively. This operation can be formulated as follows,

$$F_t = \sigma(W_t * F_{t-1} + b_t), t \in 3, 7, 11 \tag{3}$$

where F_t and F_{t-1} are the feature maps before and after transposed convolution at t-th layer.

Dense residual feature fusion (*DRFF*) As mentioned before, the high-resolution color image can be down-sampled and thus obtain a color feature map. In this network branch, the low-resolution depth map is converted to the same size as the color image, producing a corresponding depth feature map. The introduced DRFF module fuses the color feature map and depth feature map at the same scale to fully realize the color-guided depth SR. As shown in Fig.2, the DRFF module consists of dense connections and residual connections. The dense connection takes the outputs of each preceding layer as the input of the following layers, which can strengthen feature propagation by connecting low-level and high-level features. To further alleviate the issue of vanishing-gradient, the local residual structure is introduced to effectively optimize the network performance.

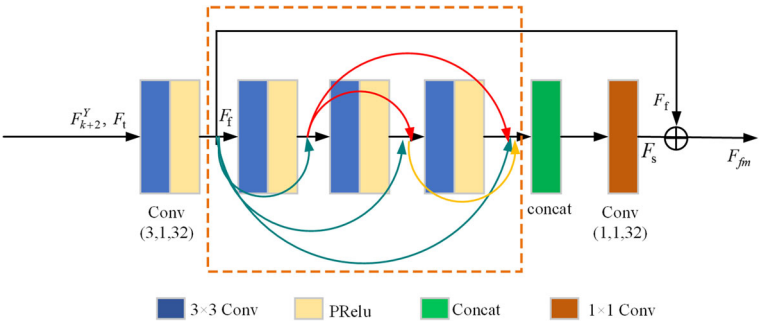

Fig. 2 The dense residual feature fusion (DRFF) module. This module employs dense connections to strengthen feature propagation and uses local residual structure to alleviate the vanishing-gradient problem

Specifically, the inputs of the DRFF module are color feature map F^Y_{k+2} and depth feature map F_t. These two feature maps are concatenated and produce a fused feature map F_f through a 3×3 convolution as,

$$F_f = \sigma(W_d[F_t, F^Y_{k+2}]) \tag{4}$$

Taking F_f as input, the dense connection can be calculated as,

$$F_{d,l} = \sigma(W_{d,l}[F_f, F_{d,0}, F_{d,1}, \cdots, F_{d,l-1}]) \tag{5}$$

where $F_{d,l}$ means the output of dense connection at layer l, $[F_f, F_{d,0}, F_{d,1}, \cdots, F_{d,l-1}]$ represents feature map concatenation, $W_{d,l}$ are the weights of convolution at layer l. Assume there are L convolutional layers in total, and each layer has G channels; the number of output feature maps is $(L+1) \times G$. To reduce the dimension of feature maps and the computational complexity of the network, a 1×1 convolution kernel is applied here. Furthermore, to make the dense connection fully recover the fine structures of scene depth, we adopt the residual connection structure to link the feature map F_f to the last layer. The final output of the DRFF module can be represented as follows (see Fig.2),

$$F_{fm} = F_s + F_f \tag{6}$$

where F_{fm} is the output feature map. The last convolution kernel is 1×1, the rest convolutions have 3×3 kernel with stride and channel numbers set to 1 and 32, respectively. The number of dense connection layers L is 5.

Depth map multi-scale optimization In the task of depth map SR, the multi-scale information is usually excavated from the scene depth image, which is the key to achieving high-precision reconstruction. In our network, the depth map SR branch uses the pyramid structure to extract multi-scale information. At each pyramid level, the multi-scale feature is extracted and optimized by channel grouping. Specifically, our depth SR network adopts a channel multi-scale (CMS) strategy to further exploit multi-scale information from depth feature maps. As shown in Fig. 3, to extract the detailed multi-scale features, group convolution is employed on depth feature maps, and thus residual blocks are stacked to enlarge the receptive field.

The input of the CMS module is the feature map F_{fm} from the output of DRFF. A 1×1 convolution is applied first to get a feature map with 64 channels, which is further divided

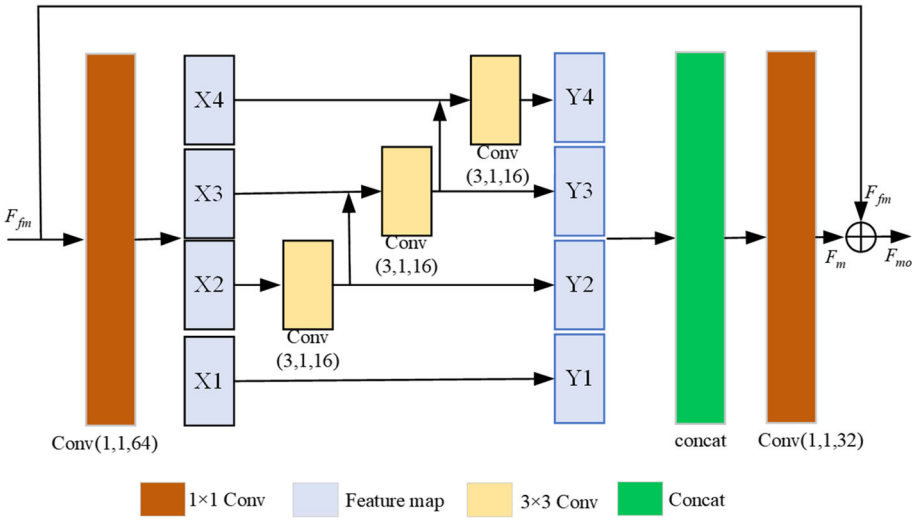

Fig. 3 Multi-scale optimization. A grouping and concatenation strategy is exploited to extract depth map multi-scale information by enlarging receptive fields of feature maps

into 4 groups. Each group x_i has its feature map subset with a corresponding convolution kernel K_i, producing an output feature map y_i $(i = 1, 2, 3, 4)$. The third and fourth groups add their feature map subsets x_i with y_{i-1}. This multi-scale grouping can be formulated as follows,

$$y_i = \begin{cases} x_i & i = 1 \\ K_i(x_i) & i = 2 \\ K_i(x_i + y_{i-1}) & i = 3, 4 \end{cases} \tag{7}$$

The outputs from these four groups are concatenated and filtered by a 1×1 convolution, producing a feature map F_m. This grouping and concatenation strategy makes convolution operation more effective for extracting features by enlarging receptive fields of the feature map and getting better multi-scale representation. The final output of the CMS module F_{mo} is the sum of feature map F_m and F_{fm} as follows,

$$F_{mo} = F_m + F_{fm} \tag{8}$$

Channel attention (*CA*) The depth map features obtained by different channels in the depth feature map are different, and each channel has a different effect on the task of depth SR. To focus on the informative high-frequency features, we adopt the CA module to generate different attention for the channel-wise feature as follows,

$$F_{ca} = f_{CA}(F_{mo}) \tag{9}$$

where F_{mo} is the output of CMS module, F_{ca} is the output of CA module. The next subsection will detail the channel attention operation f_{CA}.

Depth map reconstruction The goal of this module is to generate a high-resolution depth map by adaptively combining the feature maps. The feature map output by the CA module is processed by another convolution operation and combined with the low-frequency feature map obtained by Bicubic interpolation thus reconstructing the final high-resolution depth map.

3.3 Channel attention mechanism

In the area of computer vision and computer graphics, the attention mechanism has become an important component of various neural networks [33, 34], which can improve the feature selection ability of the entire network by setting higher weights to those channels containing high-frequency information whilst setting lower weights to those of low- frequency information. Generally speaking, channel attention can learn the weight distribution of image features according to each channel dimension [35]. The learned weights can thus be applied to the original feature channels so that it is concentrated on crucial feature channels whilst ignoring the unimportant ones. The depth map contains both the low-frequency structure of smooth regions and also the high-frequency depth boundaries which could make a more significant impact on the depth SR. In the network branch of depth map SR, the channel attention mechanism can adaptively aggregate different feature channels and assign larger weights to those channels containing high-frequency structures.

As shown in Fig. 4, the input of the CA module is the output from the CMS module, which is further encoded by V, P, and Q operations in parallel. Here, V represents the data preprocessing, producing component F_{mov} with tensor size $C \times H \times W$ via one-layer convolution; P and Q represent the shaping operation, producing component P_{moP} with tensor size $C \times (H * W)$ and component P_{moQ} with tensor size $(H * W) \times C$. After encoding, the weight of channel attention θ is obtained by *Softmax* operation on the dot product between P_{moP} and P_{moQ}. The output of the CA module F_{ca} is the dot product between θ and F_{moV}. The above process can be formulated as follows,

$$\begin{aligned} F_{moV} &= V(F_{mo}), \\ F_{moP} &= P(F_{mo}), \\ F_{moQ} &= Q(F_{mo}), \\ \theta &= Softmax(F_{moP} \cdot F_{moQ}), \\ F_{ca} &= F_{moV} \cdot \theta \end{aligned} \tag{10}$$

where $\cdot$ is the dot product. The kernel size in V, P, and Q is 3×3, the stride and channel numbers are 1 and 32, respectively.

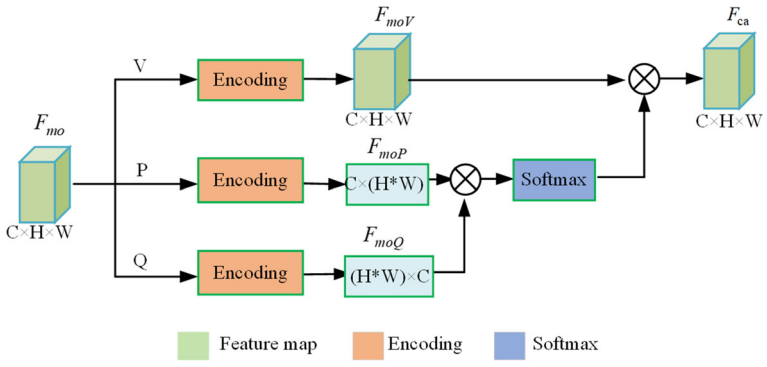

Fig. 4 Channel attention. The depth feature map is encoded by V, P, and Q, where V represents the data preprocessing, P and Q represent shaping operation. The weights of channel attention are generated according to the outputs of P and Q operations

3.4 Loss function

The proposed CA-DBMNet takes low-resolution depth maps as input and finally reconstructs high-resolution depth maps. Let F represents the model of our CA-DBMNet and K is the number of training samples. The loss function can be formulated as follows,

$$L(\theta) = \frac{1}{K}\sum_{i=1}^{K} \| F(D_{l(i)}, Y_{h(i)}; \theta) - D_{h(i)} \|^2 \tag{11}$$

where θ is the learning parameters. For each training sample i, $Y_{h(i)}$ is a high-resolution color image, $D_{l(i)}$ and $D_{h(i)}$ mean low-resolution depth map and corresponding restored high-resolution depth map, respectively.

4 Experimental results and discussion

4.1 Implementation details

In this paper, our proposed CA-DBMNet adopts the pyramid structure to extract multi-scale information from the color image and combines the shallow and deep features of the same scale with skip connections. Thus, the multi-scale structure information of the color feature map is effectively exploited in the depth map SR. Based on the channel attention mechanism, this network can adaptively assign a higher weight to the channel with a more

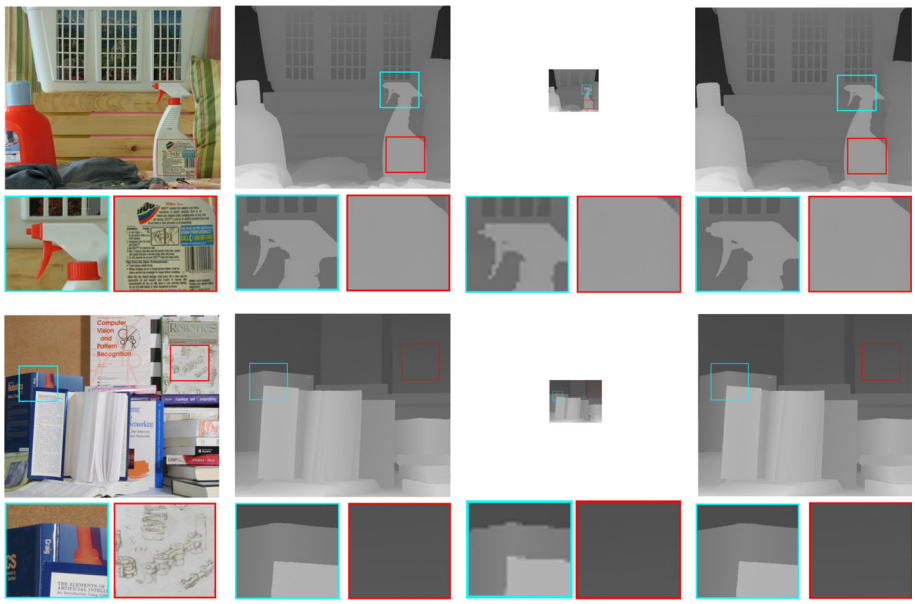

Fig. 5 Visual performance of depth map SR reconstruction on Middlebury [38] testing data: (a) input high-resolution color images, (b) the ground truth depth maps, (c) input low-resolution depth maps, (d) the results using the proposed method

significant contribution to the task of depth map SR. Thus, it effectively suppresses man-made artifacts such as shadow texture copying and focuses on high-frequency structures such as depth boundaries. The proposed network is implemented with the Tensorflow framework and trained on an NVIDIA Tesla V100 with 16G GPU memory. It is optimized using Adam optimizer [36] with $\beta_1 = 0.9$, $\beta_2 = 0.99$. The initial learning rate is set to $1e-4$ and decreased to $1e-5$ after 30 epochs.

The high-resolution scene color images and corresponding low-resolution depth maps are fed into the color image feature extraction branch and the depth map SR branch, respectively. Our experiments use 58 RGB-D scene images from the MPI dataset [37] and 34 RGB-D scene images from the Middlebury dataset [38]. During network training, 82 RGB-D scene images are selected as the training candidates, and the other 10 scene images are used as the test set for experimental verification. The training candidates are rotated by 90 degrees and normalized to the range [0, 1] to augment the training samples. The upscaling factors $2\times$,$4\times$, $8\times$, and $16\times$ are employed for generating depth map SR results. Unlike [39] uses large-scale images for training, we partition the scene depth map into regular and small overlapping patches to reduce the network training time and also maintain its performance. Then the low-resolution depth maps are generated by Bicubic interpolation from high-resolution maps. The size of the depth map patch is set according to upsampling factor. When the upsampling factor is set to $\{2\times, 4\times, 8\times, 16\times\}$, the size of low-resolution depth map patch is $\{24 \times 24, 16 \times 16, 12 \times 12, 8 \times 8\}$, and the size of color image patch is $\{48 \times 48, 64 \times 64, 96 \times 96, 128 \times 128\}$, whilst the size of the output depth map is also $\{48 \times 48, 64 \times 64, 96 \times 96, 128 \times 128\}$ respectively.

4.2 Visual performance

To verify the effectiveness of our proposed depth map SR network CA-DBMNet, we analyze the visual results of depth map SR in terms of global and local aspects. From the global aspect, the network achieves two goals of depth map SR, such as image amplification and image clarity. Figure 5 shows the depth SR results in $8\times$ upsampling case on the Middlebury [38] testing data. Figures 5(a) and 5(c) present input high-resolution color images and corresponding depth maps respectively. Figure 5(d) shows the enhancement results of different depth maps, which successfully upsample the low-resolution inputs into the specified high-resolution depth maps. By comparing the depth maps of different scenes in Fig. 5(c) and Fig. 5(d), it can be seen that our SR results are sharper and cleaner than the original depth map input. These results are consistent with the corresponding ground truth in Fig. 5(b), and the overall structures can be well maintained. From the local aspect, the blue boxes given in Fig. 5 identify the depth map high-frequency edge structures, such as water spout and book edge in Fig. 5(a), which are the fine structures that are difficult to recover in the depth map SR task. The red boxes identify the detailed texture structures in color images, such as the production instruction text of the watering can and the pattern on the book in Fig. 5(a), which may interfere with the depth map SR. As seen from the red boxes in Fig. 5(d), the structures recovered from the depth map SR are not affected by the texture information from color images. The artifacts caused by texture copying do not appear in the recovered high-resolution depth maps.

Figure 6 shows the reconstruction results of depth map SR by using our proposed network in $8\times$ upsampling case on the MPI [37] testing data. From the global view, the proposed network produces high-resolution depth maps in Fig. 6(d) from low-resolution depth maps of various scenes in Fig. 6(c). As shown in the red boxes of Fig. 6(d), texture information

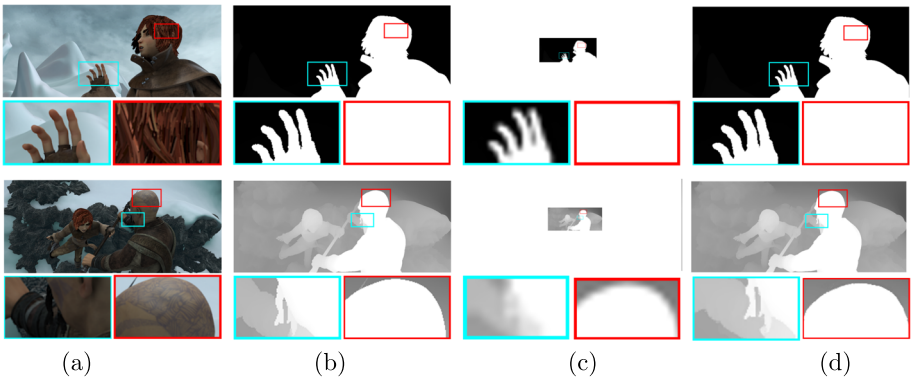

Fig. 6 Visual performance of depth map SR reconstruction on MPI [37] testing data: (a) input high-resolution color images, (b) the ground truth depth maps, (c) input low-resolution depth maps, (d) the results using the proposed method

in the color images, such as hair texture and scalp tattoo, do not cause man-made artifacts in the depth map SR. The proposed network also restores high-resolution depth maps with precise details, such as the edges of fingers and beards in the blue boxes of Fig. 6(d).

4.3 Comparisons

To demonstrate the effectiveness of the proposed network CA-DBMNet, we use Root Mean Squared Error (RMSE) as the evaluation metric for comparison with other existing methods. We evaluate the reconstruction results on Middlebury RGB-D scene dataset [38] and compare with 6 traditional methods (Bicubic, GF [40], RMRF [41], TGV [42], JID [43], AR [44]) and 5 learning-based approaches (SRCNN [6], MSG [14], MFR [15], PMBA [13], LAP [45]). The Middlebury dataset is divided into Groups A, B, and C in our experiments. The experimental results of CA-DBMNet and other methods are analyzed and compared under upsampling factors $2\times$, $4\times$, $8\times$, and $16\times$ for Groups A and B. For Group C data, this paper only performs $2\times$, $4\times$, and $8\times$ since the resolution of the input depth map is too low to reconstruct under $16\times$.

Tables 1, 2, and 3 respectively show the RMSE of the depth map SR results in testing on these three data groups. The smallest RMSE value represents the optimal results of all methods, which are marked in boldface. The sub-optimal results are underlined. It can be seen from Table 1 that when testing on Group A data, the average RMSE error of our CA-DBMNet is only 1.30, which is reduced by 56.67%, 55.17%, 62.53%, and 53.07% respectively, compared with bicubic, GF [40], JID [43] and AR [44]. Experimental results demonstrate that the effectiveness of CA-DBMNet is superior to traditional methods. It can be seen from Table 2 that when testing on Group B data, the average RMSE error of our CA-DBMNet is only 1.35, which is reduced by 44.90%, 12.90%, and 21.05% respectively, compared with learning-based methods SRCNN [6], MFR [15] and PMBA [13]. SRCNN [6] does not consider the multi-scale information of depth maps. On the contrary, MFR [15] and PMBA [13] use the multi-scale information of depth maps while ignoring the multi-scale information of color images. Experimental results can verify the effectiveness of our CA-DBMNet in terms of multi-scale feature extraction from both high-resolution color images and low-resolution depth maps. Table 3 shows the comparisons of reconstruction results

Table 1 Quantitative depth map SR results (in RMSE) on group A of Middlebury Dataset [38]

Mehods	Art				Book				Moebius			
	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×
Bicubic	2.66	3.85	5.52	8.37	1.03	1.58	2.27	3.36	0.93	1.40	2.06	2.98
GF [40]	2.93	3.79	4.97	7.88	1.16	1.57	2.09	3.18	1.09	1.44	1.88	2.86
RMRF [41]	2.31	3.26	4.31	6.68	1.14	1.53	2.18	2.92	0.97	1.44	2.21	2.79
TGV [42]	3.03	3.79	4.79	7.11	1.29	1.61	1.99	2.94	1.12	1.46	1.91	2.63
JID [43]	1.16	1.84	2.77	10.86	0.59	0.93	1.14	9.15	0.61	0.87	1.37	10.38
AR [44]	3.07	3.99	4.68	6.87	1.38	1.94	2.05	2.84	0.98	1.23	1.73	2.56
SRCNN [6]	0.98	2.29	4.75	7.81	0.39	0.94	2.15	3.24	0.45	0.97	2.01	2.82
LAP [45]	0.88	1.79	2.73	6.31	0.78	0.94	1.29	2.35	0.77	0.95	1.33	2.37
MSG [14]	0.73	1.65	3.01	5.76	0.41	0.69	1.48	2.96	0.44	0.76	1.44	2.91
MFR [15]	0.71	1.54	2.71	4.35	0.42	0.63	1.05	1.78	0.42	0.72	1.10	1.73
PMBA [13]	0.61	2.04	3.63	5.38	0.41	0.92	1.68	2.55	0.39	0.84	1.41	2.09
CA-DBMNet	**0.46**	**1.42**	**2.57**	**4.15**	**0.31**	**0.54**	**0.98**	**1.63**	**0.35**	**0.59**	**0.96**	**1.63**

* **Boldface** indicates the best value for each evaluation, while the underline indicates the second best

between our proposed network and other methods when testing on Group C data. As seen from Table 3, the average RMSE error of CA-DBMNet is only 1.35, which is reduced by 17.09%, 22.97%, and 5.00% respectively, compared with LAP [45], MS [14], and MSG [14]. Here, LAP [45] and MS [14] lack the guidance of color images, and MSG [14] does not fully fuse the feature map. On the contrary, CA-DBMNet effectively fuses the color feature map and depth feature map under the guidance of the scene color images at multiple scales.

It can be seen from Tables 1, 2, and 3 that the reconstruction performance of our method is superior to other methods in most cases of the Middlebury dataset [38]. The reasons are as follows. Firstly, the color map feature extraction branch in our network can extract rich

Table 2 Quantitative depth map SR results (in RMSE) on group B of Middlebury Dataset [38]

Mehods	Dolls				Laundry				Reindeer			
	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×
Bicubic	0.91	1.31	1.96	2.63	1.61	2.41	3.45	5.09	1.94	2.81	3.99	5.82
GF [40]	1.25	1.31	1.86	3.62	2.21	2.54	3.42	4.56	2.68	3.05	4.06	5.32
RMRF [41]	1.14	1.49	1.94	2.45	1.47	2.06	2.87	4.22	1.82	2.58	3.24	4.91
TGV [42]	1.17	1.42	2.05	4.44	1.84	2.21	3.92	6.75	2.41	2.67	4.29	8.80
JID [43]	0.73	0.96	1.26	2.06	0.72	1.19	1.77	3.47	0.91	1.47	2.19	4.15
AR [44]	1.01	1.23	1.65	2.23	2.39	2.43	3.01	4.47	2.99	3.09	4.33	4.99
SRCNN [6]	0.63	1.11	1.92	2.61	0.81	1.87	3.87	5.63	0.67	1.74	3.45	5.04
LAP [45]	0.77	0.98	1.42	2.28	0.78	1.12	1.67	3.79	0.81	1.31	1.92	4.56
MSG [14]	0.61	0.92	1.47	3.29	0.51	1.12	2.09	4.26	0.62	1.32	2.43	4.97
MFR [15]	0.61	0.89	1.22	1.74	0.61	1.11	1.75	3.01	0.65	1.23	2.06	3.74
PMBA [13]	**0.36**	0.95	1.47	2.03	0.38	1.14	2.19	3.31	**0.41**	1.39	2.74	4.12
CA-DBMNet	0.43	**0.79**	**1.09**	**1.55**	**0.36**	**0.89**	**1.49**	**2.74**	0.45	**1.09**	**1.89**	**3.39**

* **Boldface** indicates the best value for each evaluation, while the underline indicates the second best

Table 3 Quantitative depth map SR results (in RMSE) on group C of Middlebury Dataset [38]

Mehods	Tsukuba			Venus			Teddy			Cones		
	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
Bicubic	5.81	8.56	12.30	1.32	1.91	2.76	1.99	2.90	4.07	2.45	3.61	5.30
GF[40]	8.12	9.41	12.51	1.63	1.93	2.69	2.49	2.93	3.98	3.33	3.87	5.29
TGV[42]	7.21	10.31	17.51	2.15	2.52	4.04	2.71	3.31	5.39	3.51	4.45	7.14
JID[43]	3.48	5.95	10.91	0.81	1.17	1.76	1.28	2.94	2.76	1.69	4.17	5.11
SRCNN[6]	5.47	8.11	11.80	1.27	1.85	2.67	1.88	2.77	3.95	2.34	3.43	5.15
LAP[45]	1.72	5.34	8.94	0.72	0.77	1.34	0.97	1.68	2.96	0.98	2.85	4.67
MS[14]	2.21	5.21	10.25	0.59	0.78	1.18	0.99	1.78	3.18	1.13	2.95	5.23
MSG[14]	1.85	4.29	8.42	0.14	0.35	1.04	0.71	1.49	2.76	0.91	2.61	4.23
CA-DBMNet	**1.71**	**4.12**	**8.36**	**0.13**	**0.32**	**1.01**	**0.68**	**1.45**	**2.35**	**0.85**	**2.38**	**4.05**

* **Boldface** indicates the best value for each evaluation, while the underline indicates the second best

structural information and pass it to the depth map SR branch, which provides effective guidance for depth information recovery. Secondly, unlike the existing methods that employ the multi-scale structure only in one branch, our CA-DBMNet adopts the multi-scale structure in both branches to fully extract their multi-scale information; Finally, our network introduces the channel attention mechanism into depth map SR, which can adaptively assign different weights to the feature map channels, and thus help for depth map enhancement.

To further demonstrate the effeciveness of our proposed method in real-world scenarios, we also implement our network on NYU V2 indoor RGBD dataset[46]. This data is captured by consumer-level scanning devices. Following the common splitting, we use 1000 images as training data and the rest 449 images for testing. We evaluate the reconstruction results and compare with other state-of-the-art methods, including DCTNet [47], AHMF [48], DSR-Diff [49], SSDNet [50]. It should be mentioned that the DSR-Diff [49] is a diffusion model based guided superresolution method which has a time-consuming nature of diffusion model. As seen from Table 4, our CA-DBMNet produces satisfactory depth reconstruction from real-world data. This demonstrates that our method has competitive performance compared with other the-state-of-the-arts.

Figure 7 illustrates the reconstruction effect of CA-DBMNet and other methods on the test image "reindeer" scene under upsampling factor 8×. Figure 7(a) is the high-resolution color image, and Fig. 7(b) gives the ground truth depth map. Figure 7(c) shows the effect of depth map SR results using RMRF [41], while the fine structure and the edge information are blurred. Figure 7(d) gives the depth map SR results using GF [40], while the fine structure and edges have noticeable artifacts. Figure 7(e) shows the depth map SR results using TGV [42], while the depth map is disturbed by color textures such as background elements

Table 4 Quantitative depth map SR results (in RMSE) on NYU V2 Dataset [46]

Mehods	DCTNet [47]	SSDNet [50]	AHMF [48]	DSR-Diff [49]	CA-DBMNet
4×	1.59	1.60	1.40	**1.25**	1.43
8×	3.16	3.14	2.89	**2.57**	2.87
16×	5.84	5.86	5.64	**4.91**	5.69

* **Boldface** indicates the best value for each evaluation, while the underline indicates the second best

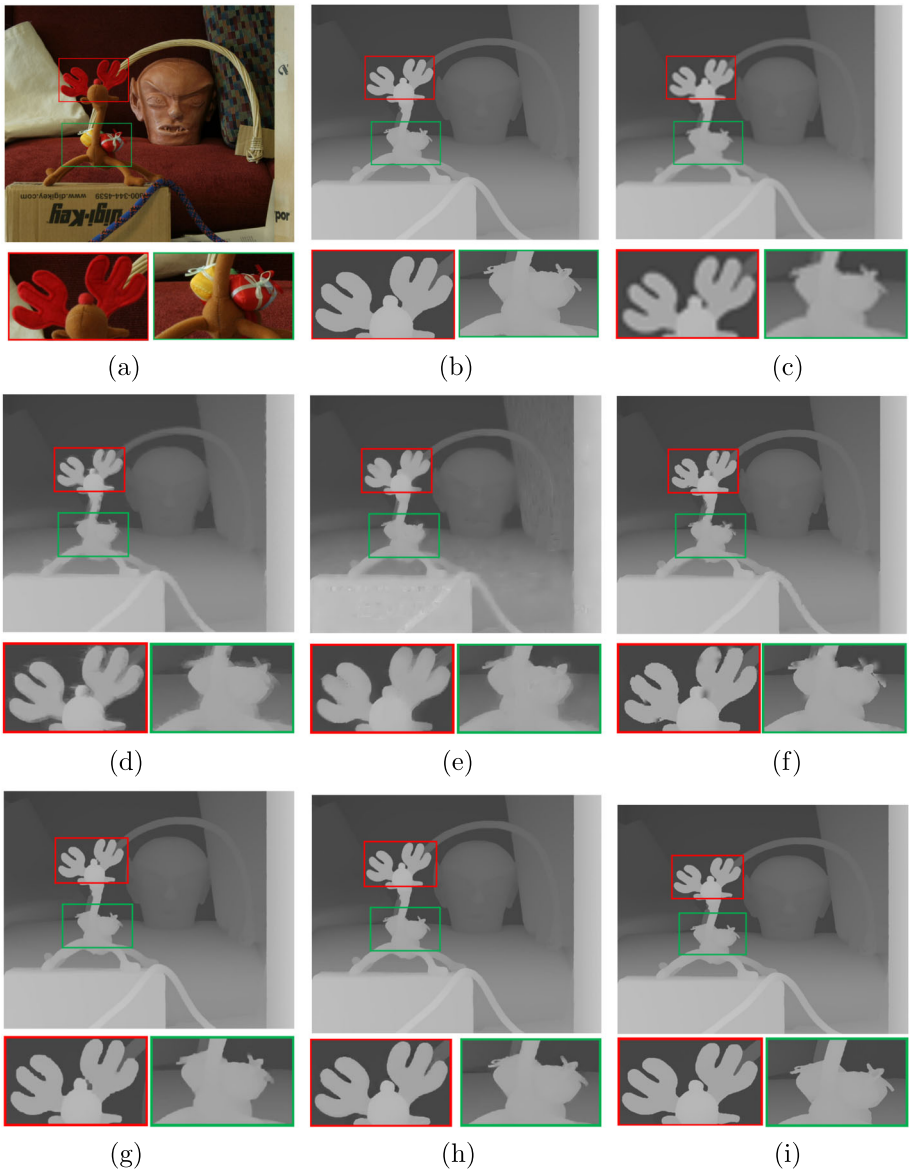

Fig. 7 Visual quality comparisons for depth map SR reconstruction on the test image "reindeer" scene under upsampling factor 8×: (a) high-resolution color image, (b) the ground truth depth map, (c) RMRF [41], (d) GF [40], (e) TGV [42], (f) SRCNN [6], (g) PMBA [13], (h) MFR [15], (i) our proposed network

appear. Fig.7(f) gives the result using SRCNN [6], which still has artifacts or blurring in local structures. Figures 7(g) and 7(h) are the depth map SR results of PMBA [13] and MFR[15] respectively. These two methods can recover the main structures of the underlying depth map but perform poorly in several local structures. For the cropped zoomed regions of the reconstruction results via PMBA [13], local structures within the red box are blurred, and the structures in the upper-left corner and upper-right corner within the blue box are not clear.

For the cropped zoomed regions of the reconstruction results via MFR [15], edges within both the red and small boxes show jagged blurs. Figure 7(i) shows the results using our network, which performs well for the depth map global structure and local details. The structural blurs caused by upsampling and the artifacts caused by texture copying can be effectively avoided. It demonstrates that our CA-DBMNet can effectively recover the fine structure and edge information of depth maps by exploiting multi-scale features extracted from both network branches and focusing on those high-frequency features through the channel attention mechanism.

4.4 Ablation studies

To explore the role of each module in our proposed CA-DBMNet, the ablation studies of the color image guidance module, DRFF module, CMS module, and CA module are carried out. The color image guidance represents the multi-scale feature extraction and guidance in the color image feature extraction branch. The DRFF module fuses color feature maps and depth feature maps. The CMS module adopts channel grouping to achieve multi-scale feature extraction, whilst the CA module can assign the weights of depth feature map channels in an adaptive manner.

Table 5 shows the experimental results of ablation studies under upsampling factor $8\times$. The first line is the baseline network, which does not include those four modules. These modules are added to the baseline network in turn for analysis. The full CA-DBMNet (the last row of Table 5) achieves the best performance, demonstrating that the dual branch parallel multi-scale feature extraction, dense fusion strategy, and channel attention mechanism can contribute to the depth map SR. Notably, the color image guidance module can improve the reconstruction results because the extracted multi-scale color feature map will provide rich structural information for depth map enhancement. The CA module improves depth map SR performance, presumably because this module assigns the appropriate weights to feature map channels at each scale.

5 Conclusions

In this work, we propose a dual branch network CA-DBMNet for scene depth map SR. The network adopts the multi-scale mechanism in both the color image feature extraction branch and the depth map SR branch, which fully extracts the scene structural information at various scales and provides good guidance for depth map SR. The proposed network

Table 5 Ablation studies on modules

Image Guided	DRFF	CMS	CA	RMSE
×	×	×	×	2.8007
×	×	✓	×	2.7738
✓	×	✓	×	2.7560
×	✓	✓	×	2.5949
×	✓	✓	✓	2.6715
✓	✓	✓	✓	2.5074

employs the feature pyramid structure in the color image feature extraction branch. This structure combines shallow and deep features, realizes the effective fusion of multi-level information, and extracts the structural information from color images. In the depth map SR branch, this network extracts the features through dense connection, channel multi-scale, and channel attention. The channel attention mechanism can better restore the fine structure of depth maps by allocating channel weights. The experimental results show that our CA-DBMNet can effectively reconstruct the high-frequency structure of depth maps and suppress the texture artifacts which is reported better performance compared with state-of-the-art approaches.

In the future, we will consider the extraction of high-frequency structures from high-resolution color images and further avoid the issue of texture copying using advanced network modules. We could also consider adding deep supervision to further solve the gradient vanishing problem while the network layers are deep.

Appendix A

The following abbreviations are used in this paper.

Table 6 The abbreviations and corresponding description

Name	Description
CA	channel attention module
CA-DBMNet	channel attention based dual branch multi-scale network
CMS	channel multi-scale module
CNN	convolution neural network
DRFF	dense residual feature fusion module
RMSE	root mean squared error
SR	super-resolution

Acknowledgements This work was partially supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ23F020002, and the National Natural Science Foundation of China under Grant No. 61972458. The authors would like to thank the anonymous reviewers for their helpful and valuable comments and suggestions.

Availability of data Publicly available datasets were analyzed in this study. This data can be found here: MPI dataset [http://sintel.is.tue.mpg.de] and Middlebury RGB-D scene dataset [https://vision.middlebury.edu/stereo/data/]. Other data will be made available on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kwon O, Park J, Oh S (2023) Renderable neural radiance map for visual navigation. In: Proc of the IEEE/cvf conf comput vis pattern recognit (CVPR), pp 9099–9108
2. Fooladgar F, Kasaei S (2020) A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks. Multimed Tools Appl 79:4499–4524
3. Prakash A, Chitta K, Geiger A (2021) Multi-modal fusion transformer for end-to-end autonomous driving. In: Proc of the IEEE/CVF conf comput vis pattern recognit (CVPR), pp 7077-7087
4. Li J, Gao W, Wu Y, Liu Y, Shen Y (2022) High-quality indoor scene 3d reconstruction with RGB-D cameras: a brief review. Comp Visual Media 8(3):369–393
5. Seichter D, Lewandowski B, Höchemer D, Wengefeld T, Gross H M (2020) Multi-task deep learning for depth-based person perception in mobile robotics. In: Proc of the IEEE int conf intell robot syst (IROS), pp 10497–10504
6. Chao D, Chen C L, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: Proc of the Eur conf on comput vis (ECCV), pp 184–199
7. Zuo Y, Wu Q, Zhang J, An P (2018) Explicit edge inconsistency evaluation model for color-guided depth map enhancement. IEEE Trans Circuits Syst Video Technol 28(2):439–453
8. Khoddami AA, Moallem P, Kazemi M (2022) Depth map super resolution using structure-preserving guided filtering. IEEE Sensors J 22(13):13144–13152
9. Wang J, Sun L, Xiong R, Shi Y, Zhu Q, Yin B (2022) Depth map super-resolution based on dual normal-depth regularization and graph Laplacian prior. IEEE Trans Circuits Syst Video Technol 32(6):3304–3318
10. Liu LW, Wang LH, Zhang M (2015) Depth map super-resolution based on joint dictionary learning. Multimed Tools Appl 74:467–477
11. Li S, Wang A, Hong S, Wu Y, Li D, Wu Y, Liang J (2020) Super resolution of single depth image based on multi-dictionary learning with edge feature regularization. Multimed Tools Appl 79:34813–34834
12. Wen Y, Sheng B, Li P, Lin W, Feng DD (2019) Deep color guided coarse-to-fine convolutional network cascade for depth image superresolution. IEEE Trans Image Process 28(2):994–1006
13. Ye X, Sun B, Wang Z, Yang J, Xu R, Li H, Li B (2020) PMBANet: progressive multi-branch aggregation network for scene depth super-resolution. IEEE Trans. on Image Process. 29:7427–7442
14. Hui T W, Chen C L, Tang X (2016) Depth map super-resolution by deep multi-scale guidance. In: Proc of the Eur conf on comput vis (ECCV), pp 353–369
15. Zuo Y, Wu Q, Fang Y, An P, Huang L, Chen Z (2020) Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. IEEE Trans Circuits Syst Video Technol 30(2):297–306
16. Liu H, Fu Z, Han J, Shao L, Hou S, Chu Y (2019) Single image super-resolution using multi-scale deep encoder-decoder with phase congruency edge map guidance. Inf Sci 473:44–58
17. Pang Y, Cao J, Wang J, Hang J (2019) JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. IEEE Trans Inf Forensics Secur 14(12):3322–3331
18. Liu H, Qin J, Fu Z, Li X, Han J (2020) Fast simultaneous image super-resolution and motion deblurring with decoupled cooperative learning. J Real-time Image PR 17:1787–1800
19. Mei Y, Fan Y, Zhou Y, Huang L, Huang T, Shi H (2020) Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proc of the IEEE/CVF conf comput vis pattern recognit (CVPR), pp 5690-5699
20. Maeda S (2020) Unpaired image super-resolution using pseudo-supervision. In: Proc of the IEEE/cvf conf comput vis pattern recognit (CVPR), pp 291–300
21. Wang Y, Su T, Li Y, Cao J, Wang G, Liu X (2022) DDistill-SR: Reparameterized dynamic distillation network for lightweight image super-resolution. IEEE Trans Multimedia 25:7222–7234
22. Lee MK, Heo J-P (2023) Noise-free optimization in early training steps for image super-resolution. arXiv preprint arXiv:2312.17526
23. ScanVic J, Davies M, Abry P, Tachella J (2023) Self-supervised learning for image super-resolution and deblurring. arXiv preprint arXiv:2312.11232
24. Wu G, Jiang J, Jiang J, Liu X (2024) Transforming image super-resolution: a ConvFormer-based efficient approach. arXiv preprint arXiv:2401.05633
25. Li Y, Huang J B, Ahuja N, Yang M H (2016) Deep joint image filtering. In: Proc of the Eur conf on comput vis (ECCV), pp 154-169
26. Ye X, Duan X, Li H (2018) Depth super-resolution with deep edge inference network and edge-guided depth filling. In: Proc of the IEEE int conf acoust, speech signal process (ICASSP), pp 1398–1402
27. Zhu J, Zhai W, Cao Y, Zha Z J (2018) Co-occurrent structural edge detection for color-guided depth map super-resolution. In: Proc of the inter conf on multimedia modeling, pp 93–105

28. Lutio RD, D'aronco S, Wegner JD, Schindler K (2019) Guided super-resolution as pixel-to-pixel transformation. In: Proc of the IEEE/CVF int conf comput vis (ICCV), pp 8828–8836
29. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
30. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proc of the IEEE/CVF conf comput vis pattern recognit (CVPR), pp 3431–3440
31. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Proc of the int conf medical image comput and compter-assisted intervention (MICCAI), vol 9351, pp 234–241
32. Song X, Dai Y, Qin X (2016) Deep depth super-resolution: learning depth super-resolution using deep convolutional neural network. In: Proc of the Asian conf comput vis (ACCV), pp 360–376
33. Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: Convolutional block attention module. In: Proc of the Eur conf on comput vis (ECCV), pp 3–19
34. Dai T, Zha H, Jiang Y, Xia S T (2019) Image super-resolution via residual block attention networks. In: Proc of the IEEE/CVF int conf comput vis workshop, pp 3879–3886
35. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Learning a discriminative feature network for semantic segmentation. In: Proc of the IEEE/CVF conf comput vis pattern recognit (CVPR), pp 1857–1866
36. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Proc of the Int Conf on Learning Representations (ICLR), pp 13
37. Butler DJ, Wulff J, Stanley GB, Black MJ (2012) A naturalistic open source movie for optical flow evaluation. In: Proc of the Eur conf on comput vis (ECCV), pp 611–625
38. Pal CJ, Weinman JJ, Tran LC, Scharstein D (2012) On learning conditional random fields for stereo. Inter J Comput Vis 99(3):319–337
39. Riegler G, Ferstl D, Rüther M, Bischof H (2016) A deep primal-dual network for guided depth super-resolution. In: Proc of the British mach vis conf (BMVC), Article no. 7
40. He K, Jian S, Tang X (2013) Guided image filtering. IEEE Trans on Pattern Anal and Mach Intell 35(6):1397–1409
41. Liu W, Chen X, Yang J, Wu Q (2017) Robust color guided depth map restoration. IEEE Trans on Image Process 26(1): 315–327
42. Ferstl D, Reinbacher C, Ranftl R, Ruther M, Bischof H (2013) Image guided depth upsampling using anisotropic total generalized variation. In: Proc of the IEEE int conf on comput vis (ICCV), pp 993–1000
43. Kiechle M, Hawe S, Kleinsteuber M (2013) A joint intensity and depth co-sparse analysis model for depth map super-resolution. In: Proc of the IEEE inter conf on comput vis (ICCV), pp 1545–1552
44. Yang J, Ye X, Li K, Hou C, Wang Y (2014) Color-guided depth recovery from RGB-D data using an adaptive autoregressive model. IEEE Trans on Image Processing 23(8):3443–3458
45. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proc of the IEEE/CVF conf comput vis pattern recognit (CVPR), pp 624–632
46. Silberman N, Hoiem D, Kohli P, Rob F (2012) Indoor segmentation and support inference from RGBD images. In: Proc of the Eur Conf on Comput Vis (ECCV), pp 746–760
47. Zhao Z, Zhang J, Xu S, Lin Z, Pfister H (2022) Discrete cosine transform network for guided depth map super-resolution. In: Proc of the IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 5687–5697
48. Zhong Z, Liu X, Jiang J, Zhao D, Chen Z, Ji X (2021) High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. IEEE Trans Image Process 21:648–663
49. Shi Y, Xia B, Zhu R, Liao Q, Yang W (2023) DSR-Diff: Depth map super-resolution with diffusion model. arXiv preprint arXiv:2311.09919
50. Zhao Z, Zhang J, Gu X, Tan C, Xu S, Zhang Y, Timofte R, Van Gool L (2023) Spherical space feature decomposition for guided depth map super-resolution. In: Proc of the IEEE/CVF int conf comput vis (ICCV), pp 12547–12558