Color-Image Guided Depth Map Super-Resolution Based on Iterative Depth Feature Enhancement

Lijun Zhao[1*], Ke Wang[1], Jinjing Zhang[2], Jialong Zhang[1], and Anhong Wang[1]
[1] Taiyuan University of Science and Technology
Taiyuan, 030051, China
[e-mail: leejun@tyust.edu.cn]
[2] North University of China
Taiyuan, 030051, China
[e-mail: B1707007@st.nuc.edu.cn]
*Corresponding author: Lijun Zhao

Received March 16, 2023; accepted July 30, 2023; published August 31, 2023

Abstract

With the rapid development of deep learning, Depth Map Super-Resolution (DMSR) method has achieved more advanced performances. However, when the upsampling rate is very large, it is difficult to capture the structural consistency between color features and depth features by these DMSR methods. Therefore, we propose a color-image guided DMSR method based on iterative depth feature enhancement. Considering the feature difference between high-quality color features and low-quality depth features, we propose to decompose the depth features into High-Frequency (HF) and Low-Frequency (LF) components. Due to structural homogeneity of depth HF components and HF color features, only HF color features are used to enhance the depth HF features without using the LF color features. Before the HF and LF depth feature decomposition, the LF component of the previous depth decomposition and the updated HF component are combined together. After decomposing and reorganizing recursively-updated features, we combine all the depth LF features with the final updated depth HF features to obtain the enhanced-depth features. Next, the enhanced-depth features are input into the multi-stage depth map fusion reconstruction block, in which the cross enhancement module is introduced into the reconstruction block to fully mine the spatial correlation of depth map by interleaving various features between different convolution groups. Experimental results can show that the two objective assessments of root mean square error and mean absolute deviation of the proposed method are superior to those of many latest DMSR methods.

Keywords: Convolution neural network, depth map super-resolution, high-low frequency decomposition, joint image filtering.

http://doi.org/10.3837/tiis.2023.08.006 **ISSN : 1976-7277**

1. Introduction

People often use scene depth information as a kind of auxiliary information to guide computer vision tasks to better complete scene analysis and understanding include self-driving vehicles, virtual reality, etc. Consumer-grade depth cameras are limited by hardware sensors. Hence, it is challenging to obtain high-quality and High-Resolution (HR) depth maps directly. Low-resolution (LR) depth maps always suffer from serious structural loss along image discontinuous areas, which makes it impossible to accurately reconstruct an HR depth map from only one LR depth map. In contrast to single Depth Map super-resolution (DMSR), color image-guided depth map Super-Resolution (SR) methods always use HR color images as prior information to guide structural restoration of LR depth maps by exploring structural similarity.

Existing DMSR methods are approximately divided into three types: weighted filtering-based methods, prior-regularized methods, and neural network-based methods. Weighted filtering-based depth map SR methods usually need to locally calculate filtering weight based on the spatial similarity between different pixels within each image block [1-3]. As a result, it often takes a long time to achieve DMSR reconstruction after pixel-by-pixel filtering. Prior-regularized DMSR methods usually use data item and regularization items to construct objective functions and use the complex optimization algorithms for iterative reconstruction [4-6]. However, these kinds of approaches have high computational complexity, since these objective functions are always non-convex and it is hard to get the optimal solution for these DMSR approaches.

Recently, many scholars have devoted themselves to the research of DMSR methods based on deep neural networks [7-10]. For example, *Hui et al*. [11] proposed a progressive up-sampling network structure to extract HR color features and LR depth features and used the texture features rich in HR color features to eliminate the ambiguity of LR depth features. Unlike depth SR methods based on multi-scale feature extraction, *He et al*. [12, 13] used octave convolution to successively divide color features into two components at multiple stages, after which color features from various stages were aggregated with depth features from different layers respectively to progressively enhance the fine detail information of depth map features. However, due to the differences of illumination and inherent characteristics of different object surfaces in the depth maps and corresponding color images, it is difficult to transfer structural features. At the same time, the structures at the end of the reconstruction module in many depth maps SR networks [8, 10, 12, 14-16] are too simple, which greatly limits the accuracy improvement of depth map reconstruction.

Based on the above analysis, we proposes a color-image guided DMSR method based on iterative depth feature enhancement. Specifically, to eliminate the interference of color LF information and adaptive filter unwanted color features, we propose to leverage the decomposed depth features to guide color feature extraction for high-efficiency structural transfer. Meanwhile, we adopt recursive feature disassembly and reassemble to gradually enhance depth feature in depth high-frequency updating block. Inspired by MobileNet and ShuffleNet [17, 18], we introduce criss-cross enhancement modules at the end of the network to fully explore depth spatial dependence by interleaving various features from different groups of convolution operations.

The others of this article is arranged as below. Firstly, we review single DMSR methods and joint DMSR methods in Section 2. After that, we describe the implementation process of the proposed method in detail in Section 3, which is followed by the experimental results of objective quality contrast, visual quality contrast, and ablation research are presented in Section 4. At last, we give a conclusion in Section 5.

2. Related Work

2.1 Single Depth Map Super-Resolution

To obtain high-quality depth information, various single DMSR methods are studied by many researchers. For instance, *Wang et al.* [19] used local self-similarity of depth map to construct many paired image blocks, namely HR depth image block and LR depth edge block, and then they deduced high-resolution boundary map of depth map through Markov model. Finally, joint bilateral filtering was used to realize high-quality reconstruction of depth map. *Chen et al.* [20] used a convolution neural network to predict high-quality depth edge map and used it as the weight of the regularization term of the total change model to achieve better depth map reconstruction.

Different from traditional single DMSR methods, many researchers are working on single depth map SR based convolution neural networks. For instance, *Chen et al.* [21] proposed an image SR reconstruction method based on the attention mechanism on the feature map, which reconstructed the original low-resolution image into a multi-scale SR image. In addition, the existing CNN-based image super-resolution methods have too many parameters while maintaining high-quality reconstruction, which is difficult to be applied to the edge-devices with limited computing and memory resources. To solve the above problems, *Chen et al.* [22] proposed a progressive feature aggregation network to gradually extract and enhance the multi-scale information of low resolution images. *Du et al.* [23] used a single model adapted to any scale factor to solve the image SR problem. *Huang et al.* [24] proposed a pyramid dense residual network for DMSR. This method used dense jump connection to aggregate different levels of features and used residual learning to iterative generate HR depth maps. Similarly, *Song et al.* [25] proposed an iterative residual learning-based depth map SR framework, which learned the High-Frequency (HF) components of the depth map in a progressive way from coarse to fine and constrained the learning of depth refinement module through total generalized variation regularization and consistency loss. In addition, *Wu et al.* [26] effectively enhanced feature representation of depth map by using iterative up-sampling and down-sampling operations. *Ye et al.* [27] proposed a deep controllable slicing network with a group of slicing branches for accurate depth map recovery. To sum up, single DMSR methods cannot accurately reconstruct HR depth map with high quality, since there are not many dependable clues from a single LR depth map for depth map restoration, which lacks fine-details and edge-structures when an up-sampling factor is extremely large.

2.2 Joint Depth Map Super-Resolution

By contrast, joint DMSR methods can use high structural similarity between HR color image and LR depth map to enhance depth boundary information. For instance, *Ham et al.* [4] reformulated guided image filtering as a converged quickly non-convex optimization problem by using structural information from the guided and input images. *Barron et al.* [1] cast depth super-resolution problem as an optimization problem restricted by image-dependent bilateral-smooth term and data-fidelity term. *Ferstl et al.* [5] converted DMSR task as a convex optimization problem with the high-order regularization. *Yang et al.* [28] built a stereo-vision-assisted model by using three constraints: non-local and local prior constraints, as well as stereo-disparity prior constraint. However, these traditional joint DMSR methods general depend on high complexity optimization and consume a lot of computational time, which greatly restricts their wide applications and deployments.

Recently, CNN-based methods have achieved remarkable performance in the field of DMSR [29-31]. For instance, *Zuo et al*. [14] constructed a deep convolution neural network to extract multi-scale intensity features and used dense connection, local and global residual learning to recover HF details from coarse to fine. After that, *Ye et al*. [16] iteratively used up-sampling and down-sampling errors and applied an attention mechanism to gradually highlight depth boundary features. *Guo et al*. [32] used residual U-Net to combine the LR depth map and guide map features level by level at the decoding end through hierarchical feature-driven residual learning. Inspired by residual U-Net [32], *Cao et al*. [33] constructed a dual-branch auto-encoder attention network, which included guidance and target auto-encoder network. The first network was trained by both color and depth reconstruction loss, while the second network was only regularized by depth reconstruction loss. That is to say, a dual auto-encoder attention network was trained by multi-task loss. Similarly, *Tang et al.* [34] proposed a joint multi-task learning network to simultaneously implement depth map SR and monocular depth estimation, which was also optimized by multi-task loss. This network used a HF attention and content guidance module to make information interaction between the monocular depth estimation task and super-resolution reconstruction task of the depth map. Although these approaches can greatly improve the performance of DMSR task, it is still necessary to further study the DMSR topic, since higher-accuracy depth image can provide better geometric structure information for 3D reconstruction.

3. The Proposed Method

Although many DMSR methods can improve the resolution of low-quality LR depth map to a certain extent, while ensuring the clarity of the reconstructed depth map, feature differences and consistency between dual modalities of color and depth maps have not been fully mined by these methods. More importantly, the high-frequency information of color images has not been well leveraged to improve depth map quality. Our motivation comes from that the difference of color and depth high-frequency features is far smaller than that of color and depth features. Consequently, we propose a color-image guided DMSR method based on iterative depth feature enhancement. Given a LR depth map $D_{low} \epsilon R^{w*h*1}$ and corresponding HR color map $C_{high} \epsilon R^{8w*8h*3}$, HR depth map $D_H \epsilon R^{8w*8h*1}$ can be predicted by the proposed method, which consists of three blocks: sampling-based color HF prediction block, depth high-frequency updating block, and multi-stage depth reconstruction block, as shown in **Fig. 1**. In this section, the implementation of 8x depth map SR will be taken as an example to introduce the proposed network.

We first use Bicubic up-sampling to enlarge D_{low} to the same scale as C_{high} to obtain the initialized depth map $D_I \epsilon R^{8w*8h*1}$. Then, we use two cascaded convolution modules to extract shallow color features and depth features from the HR color map C_{high} and initialized depth map D_{low} respectively. This module includes a series of operations, that is, 3×3 convolution operation, Leaky ReLU (LReLU) activation function and 1×1 convolution operation. Next, shallow color feature $F_{c(0)}$ are sent into sampling-based color HF prediction block to estimate color HF features so as to eliminate the interference of color map low-frequency features. In this block, down-sampling convolution and up-sampling convolution are leveraged to extract color low-frequency features F_c^L from shallow color feature $F_{c(0)}$, after which we can obtain color high-frequency features F_c^H by subtracting F_c^L from $F_{c(0)}$.

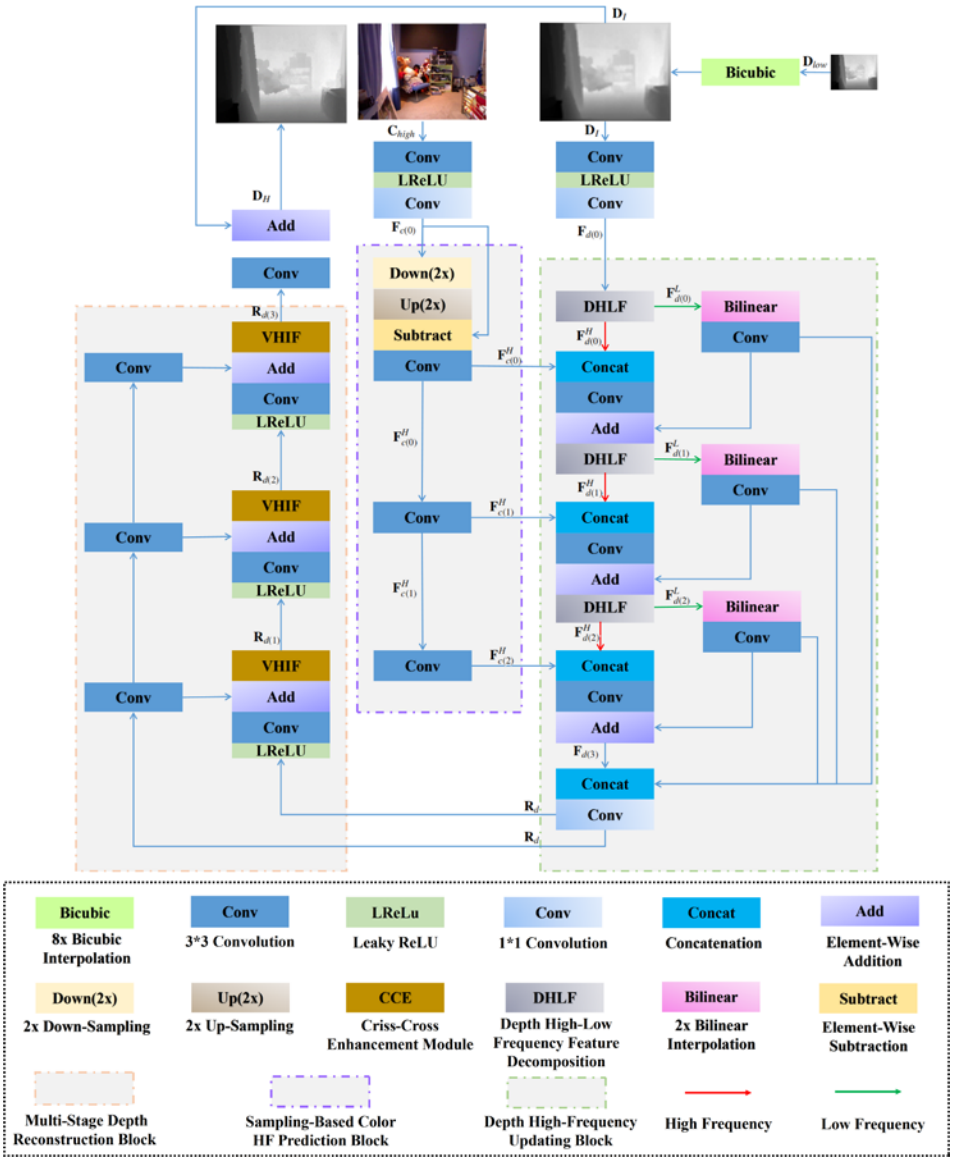

Fig. 1. The diagram of the proposed method.

In the depth high-frequency updating block, we adopt octave convolution [13] as an efficient way of high-low frequency feature decomposition operation to divide depth shallow feature $F_{d(0)}$ as the initial depth high-frequency features $F_{d(0)}^{H}$ and the initial depth low-frequency features $F_{d(0)}^{L}$. In the first stage of this block, the initial depth high-frequency features $F_{d(0)}^{H}$ and color high-frequency features $F_{c(0)}^{H}$ are concatenated along channel dimension, after which the 3×3 convolution operation is used to combine color and depth high-frequency features as the enhanced depth high-frequency features. At the same time, the bilinear interpolation operation with an up-sampling factor of 2 is used to restore the initial depth low-frequency features $F_{d(0)}^{L}$ to the same resolution as $F_{d(0)}$, which is followed by the 3×3 convolution operation to decrease the channel number of convolution features. Finally, the restored depth low-frequency features $F_{d(0)}^{L}$ and the updated depth high-frequency features

are added together element by element to obtain the updated depth features $F_{d(1)}$ in the first stage. Analogously, we can obtain depth features $F_{d(2)}$ and $F_{d(3)}$ in the second and third stages. Finally, at the end of depth high-frequency updating block, a sequential concatenation operation and 1×1 convolution operation are used to aggregate depth low-frequency features at each stage with $F_{d(3)}$ along the channel dimension to obtain R_d. The whole process of depth high-frequency updating block can be written as follows:

$$\begin{gathered}
F_{d(0)} = f_{c(1)}(f_{LReLU}(f_{c(3)}(D_I))), \\
\mathrm{F_{c(0)}} = \mathrm{f_{c(1)}}(\mathrm{f_{LReLU}}(\mathrm{f_{c(3)}}(\mathrm{C_{high}}))), \\
\mathrm{F^H_{c(i)}} = \mathrm{f_{c(3)}}(\mathrm{F_{c(0)}} - \mathrm{f_{2x\uparrow}}(\mathrm{f_{2x\downarrow}}(\mathrm{F_{c(0)}}))), \\
\mathrm{F^H_{d(i)}}, \mathrm{F^L_{d(i)}} = \mathrm{f_{DHLH}}(\mathrm{F_{d(i)}}), \\
\mathrm{F_{d(i+1)}} = \mathrm{f_{c(3)}}(\mathrm{C}(\mathrm{F^H_{d(i)}}, \mathrm{F^H_{c(i)}})) + \mathrm{f_{c(3)}}(\mathrm{f_{BILI}}(\mathrm{F^L_{d(i)}})), (\mathrm{i} = 0,1,2).
\end{gathered} \tag{1}$$

in which $f_{c(3)}$, f_{LReLU} and $f_{c(1)}$ denotes 3×3 convolution, Leaky ReLU activation function and 1×1 convolution respectively. f_{DHLH} and f_{BILI} denote Depth High-Low Frequency (DHLF) feature decomposition and bilinear interpolation operation respectively. C denotes concatenation operation along channel dimension. $f_{2x\uparrow}$ and $f_{2x\downarrow}$ denote the 2x up-sampling and 2x down-sampling operations. These two operations denote the convolution with a stride of 2 and transposed-convolution respectively.

In multi-stage depth reconstruction block, three sequential feature enhancement modules in turn are used to enhance the depth feature of R_d for multi-stage enhancement. The first sequential module uses Leaky ReLU and 3×3 convolution to extract non-linear features R_n from R_d. At the same time, we use a 3×3 convolution to extract linear features R_l. Next, we add the linear features R_l with nonlinear features R_n element by element, after which depth spatial correlation is fully mined by the criss-cross enhancement (CCE) module. This module interleaves various features to obtain the enhanced depth feature $R_{d(1)}$. By analogy, we can get the depth enhancement features $R_{d(2)}$ and $R_{d(3)}$ in the second and third stages. Finally, the depth enhancement feature $R_{d(3)}$ in the third stage is processed by an output convolution operation, and then its output is added element by element with D_I to obtain the final high-resolution reconstruction depth map D_H. The procedure of multi-stage depth reconstruction block can be written as:

$$\begin{gathered}
\mathrm{R_{d(1)}} = \mathrm{f_{CCE}}(\mathrm{f_{c3}}(\mathrm{f_{LReLU}}(\mathrm{R_d}))) + \mathrm{f_{c3}}(\mathrm{R_d}), \\
\mathrm{R_{d(2)}} = \mathrm{f_{CCE}}(\mathrm{f_{c3}}(\mathrm{f_{LReLU}}(\mathrm{R_{d(1)}}))) + \mathrm{f_{c3}}(\mathrm{f_{c3}}(\mathrm{R_d})), \\
\mathrm{R_{d(3)}} = \mathrm{f_{CCE}}(\mathrm{f_{c3}}(\mathrm{f_{LReLU}}(\mathrm{R_{d(2)}}))) + \mathrm{f_{c3}}(\mathrm{f_{c3}}(\mathrm{f_{c3}}(\mathrm{R_d}))), \\
\mathrm{D_H} = \mathrm{R_{d(3)}} + \mathrm{D_I}.
\end{gathered} \tag{2}$$

in which f_{CCE} denotes the criss-cross enhancement module.

3.1 Criss-Cross Enhancement Module

Generally, the standard convolution layer has higher computational complexity than a sequential of group convolution and point-wise convolution, which are widely applied in the lightweight network models such as MobileNet and ShuffleNet [17, 18]. Inspired by these models, we propose a lightweight Criss-Cross Enhancement (CCE) module to fully mine spatial and channel correlation between different features from different group convolutions, as displayed in **Fig. 2**. Next, we will introduce the CCE module in the multi-stage deep reconstruction block. The input of the first CCE module is $Z_u = R_{l(0)} + R_{n(0)}$. In this module, the group number of group convolution in the upper branch is 4 for group-feature extraction to obtain $Z_{G=4}$, and the features of the Leaky ReLU activated $Z_{G=4}$ are extracted along the

channel dimension by point-wise convolution, and then its output is added with Z_u by a residual connection to obtain Z_4. Then, we add Z_u with $Z_{G=4}$ together, and the resulting Z_m is sent to the middle branch with a group number of 8, and the features are extracted by convolution to obtain $Z_{G=8}$. The subsequent operations are the same as the upper branch, and then we can obtain the final output Z_8 of the middle branch. Similar to the middle branch, in the lower branch, we add Z_u with $Z_{G=8}$ together, and the resulting Z_d is successively processed through group convolution and point-wise convolution with a group number of 16, and the Z_d and Z_{16}^p are added together to obtain Z_{16}. Finally, a concatenation operation and an output convolution are respectively used to aggregate Z_4, Z_8 and Z_{16} and perform feature-channel shrinkage so as to obtain the enhanced depth feature $R_{d(1)}$.

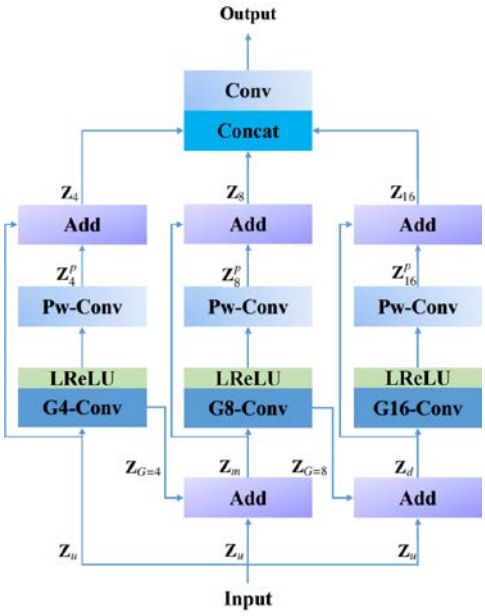

Fig. 2. The network structure of the proposed criss-cross enhancement (CCE) module.

3.2 Loss Function

For various image restoration tasks, mean square error loss and mean absolute error loss functions are widely used to constrain the network training. However, it has been proved that image restoration network trained with L_1 norm-regularized loss function obtains better performance than that with L_2 norm-regularized loss function. Therefore, the L_1 norm is used to constrain data loss L_{data} to supervise the learning of the proposed depth map SR network, which can be written as:

$$L_{data} = \left\| D_{high} - D_{GT} \right\|_1. \tag{3}$$

where $\|\cdot\|_1$ represents the L_1 norm, D_{high} is the depth map predicted by the proposed network, and D_{GT} is the corresponding Ground-Truth (GT) image.

4. Experimental Results and Analysis

In this section, we will first briefly describe the implementation details of the proposed method. Then, the proposed method is compared with a lot of traditional and CNN-based approaches to show the absolute advantages of the proposed method in terms of Root Mean Square Error

(RMSE) and Mean Absolute Deviation (MAD). At last, the irreplaceability of each module in the proposed network is verified by ablation studies.

4.1 Implementation Details

In the proposed method, two extensively-used RGB-D datasets are chosen for training and testing, that is, the Middlebury RGB-D dataset and NYU-v2 RGB-D dataset. Specifically, the top 1000 pairs of color and depth maps of NYU-v2 RGB-D dataset are used as the training dataset, while the remaining 449 pairs of color+depth maps are used to compare the performance of different depth map SR approaches. The Middlebury RGB-D dataset includes 36 pairs of RGB-D images, among which 6, 21, and 9 images come from the Middlebury (2001), (2006), and (2014) datasets respectively. These images are used as the training dataset. Middlebury testing dataset is composed of several color and depth maps (Art, Books, Moebius, Dollars, Laundry, and Reindeer) from the Middlebury (2005) dataset. The Bicubic interpolation method is used to down-sample the ground-truth depth maps to obtain corresponding LR depth maps. For training, the proposed network is optimized with the Adam optimizer. Additionally, we set the initial learning rate of the network to be 1e-4, and the learning rate of 100 epochs per iteration is multiplied by 0.1. Our method is implemented by using the deep learning framework of PyTorch on the NVIDIA TITAN RTX GPU.

4.2 Performance Comparison on NYU-v2 RGB-D Dataset

To prove the superiority of the our method, we compare it with many state-of-the-art DMSR methods in term of the objective quality at various up-scaling factors (4x, 8x, 16x). Traditional depth map SR methods have Bicubic interpolation method, JBU [3], TGV [5], MRF [6], GF [2], FBS [1], Park [35] and Ham [4], while deep learning-based depth map SR methods include DJF [15], DMSG [11], DJFR [9], FDSR [12], DKN [10], FDKN [10], Bridge [34], DSR [36], and DAEA [33]. **Table 1** lists the RMSE values of these methods under different scale factors (4x, 8x, 16x). As given in **Table 1**, the performances of CNN-based depth SR approaches are much better than those of the traditional depth SR methods. This comes from that traditional methods often depends on highly complex optimization models, which greatly limits their further application and deployment. Specifically, the greatest performance method in the traditional methods is JBU [3], and the RMSE values at different up-scaling factors (4x, 8x, 16x) are 4.07, 8.29, and 13.35 respectively. Among the deep learning-based depth map SR methods, DJF [15] has the worst performance. Compared with the traditional depth map SR method JBU [3], it can be clearly found that the performance of DJF [15] is much better than JBU [3] at different up-sampling factors (4x, 8x, 16x). From the above analysis, it can be found that the objective performance index of the deep learning-based methods is far better than that of the traditional ones.

Among deep learning-based methods, our method can achieve the best performance for depth map SR under different scale factors. Specifically, at 4x, compared with DSR [36], the RMSE of our method is decreased by 0.38 from 1.49 to 1.11, and the performance of our method is improved by 26%. At 8x, compared with the Bridge method [34], the RMSE value is decreased by 0.39 from 2.63 to 2.24, and our performance is improved by 15%. Different from depth SR at the up-scaling factor of 4x and 8x, to alleviate loss problem of detail and structure information under large-scale sampling for color guided depth map SR, when 16x depth map SR, we use two-level depth map SR to realize 16x joint depth map SR. Specifically, we apply the 4x depth map SR network structure twice to form a cascaded two-stage 16x depth map SR network. At 16x, compared with DAEA [33], the RMSE of the proposed method is decreased by 0.91 from 4.55 to 3.64, and our performance is improved by 20%. From the

above objective performance comparison, we can found that the proposed method is superior to many advanced depth map SR methods for 4x, 8x, and 16x depth SR.

Table 1. Objective performance comparison between classic methods and proposed method in term of RMSE on NYU-v2 dataset.

Method	**Bicubic**	**MRF**[6]	**GF**[2]	**JBU**[3]	**TGV**[5]	**FBS**[1]
4x	8.16	7.84	7.32	4.07	6.98	4.29
8x	14.22	13.98	13.62	8.29	11.23	8.94
16x	22.32	22.22	22.03	13.35	28.13	14.59
Method	**Park**[35]	**Ham**[4]	**DJF**[15]	**DMSG**[11]	**DJFR**[9]	**DKN**[10]
4x	5.20	5.27	3.54	3.02	3.38	1.62
8x	9.56	12.31	6.20	5.38	5.86	3.26
16x	18.10	19.24	10.21	9.17	10.11	6.51
Method	**FDKN**[10]	**FDSR**[12]	**DSR**[36]	**DAEA**[33]	**Bridge**[34]	**Our(32)**
4x	1.86	1.61	**1.49**	1.58	1.54	**1.11**
8x	3.58	3.18	2.73	2.79	**2.63**	**2.24**
16x	6.96	5.86	5.11	**4.55**	4.98	**3.64**

Table 2. Objective performance comparison between classic methods and proposed method in term of MAD on Middlebury 2005 dataset.

Method	**CLMF**[37]	**JGU**[38]	**PB**[39]	**TGV**[5]	**CDLLC**[40]
4x	0.447	0.338	0.557	0.430	0.338
8x	0.783	0.590	0.920	0.838	0.540
16x	1.515	1.083	1.557	2.107	0.947
Method	**EG**[41]	**DMSG**[11]	**DSR**[36]	**Bridge**[34]	**CGN**[42]
4x	0.295	0.280	**0.163**	0.190	0.198
8x	0.490	0.515	0.355	0.345	0.350
16x	0.902	1.005	0.830	0.765	0.762
Method	**MFR**[7]	**MIG**[29]	**RDN**[14]	**RMIG**[8]	**Ours(64)**
4x	0.253	0.188	0.203	0.212	**0.142**
8x	0.417	**0.340**	0.377	0.343	**0.317**
16x	0.782	0.748	**0.728**	0.833	**0.675**

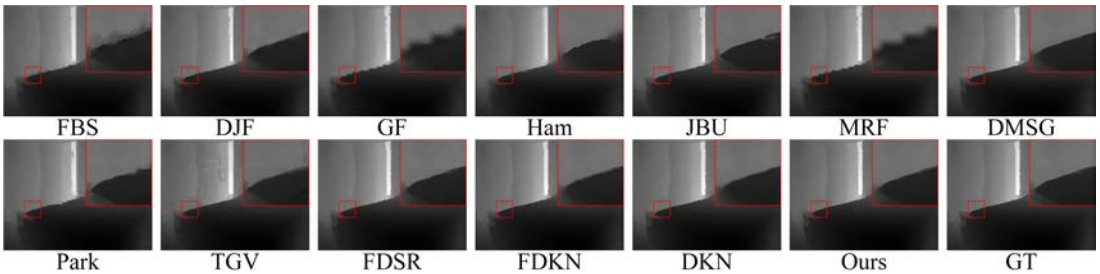

Fig. 3. The visual comparison of the 8x up-sampling results on 1002-th depth map from NYU-v2 RGB-D Dataset.

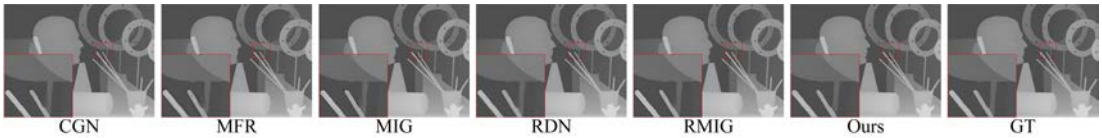

Fig. 4. The visual comparison of the 8x up-sampling results on the Art depth map.

To further exhibit the superiority of the proposed method in visual quality comparison, we compare different depth map SR methods. As shown in **Fig. 3**, it can be clearly seen that the predicted depth images by several depth SR methods such as FBS [1], GF [2] and MRF [6] have serious boundary distortion problems, when they are compared with GT images, whose boundary information is sharp and object surface has a characteristic of piece-wise smoothness. The boundary regions of these predicted depth images have large protrusions and blurring outcomes. Meanwhile, the depth maps predicted by Ham [4], JBU [3] and DMSG [11] have dense yet small serration. In contrast, the results predicted by DJF [15], Park [35], TGV [5], FDSR [12], FDKN [10] and DKN [10] are more clear, but there are still subtle variations, as compared with the GT images. From the above visual contrast, it can be clearly found that the proposed method has an advantage in the detail restoration of the reconstructed depth map.

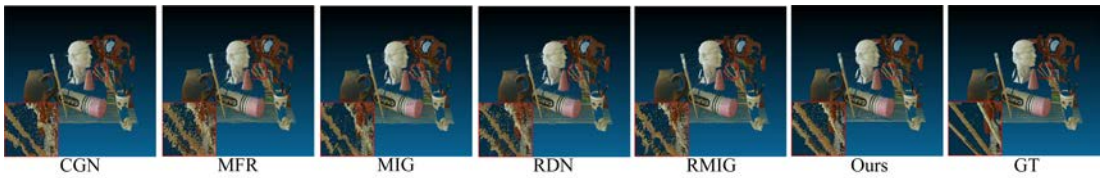

Fig. 5. The visual comparison of the 8x up-sampling 3D results on the Art depth map.

4.3 Performance Comparison on Middlebury RGB-D Dataset

Under diverse super-resolution factors, we compare our method with several state-of-the-art methods, including traditional DMSR methods such as TGV [5], CDLLC [40], CLMF [37], JGU [38], PB [39], and EG [41], as well as many deep learing-based depth map SR methods. These CNN-based methods include CGN [42], MFR [7], DMSG [11], DSR [36], Bridge [34], MIG [29], RDN [14], and RMIG [8]. **Table 2** provides the comparison of the MAD values of these methods under different scale factors (4x, 8x, 16x).

As shown in **Table 2**, EG [41] has the best performance among the traditional methods. Under different scale factors (4x, 8x, 16x), the MAD value of EG [41] in the traditional methods is slightly lower than that of DMSG [11] at 8x and 16x, and DMSG [11] has the worst performance in CNN methods. Therefore, it can be clearly found that the MAD value of traditional methods is usually higher than that of deep learing-based methods. In addition, the objective performance index of the this method is superior to other deep learning-based methods. Specifically, at the 4x DMSR, compared with DSR [36], the MAD value of our method is reduced by 0.021 and the performance is improved by 13%. At 8x depth map SR, compared with MIG [29], the MAD value of the proposed method is decreased by 0.023 and the performance is improved by 7%. When the up-scaling factor is 16 for depth map SR, the proposed method reduces the MAD value by 0.053 compared with RDN [14], and the performance improves by 7%. **Table 2** shows that the proposed method's performance is better than the state-of-the-art methods when these methods are tested on the Middlebury RGB-D dataset.

Next, to further show the advantages of the proposed method, we provide 2D depth map visual comparisons of the proposed method and several state-of-the-art methods such as CGN [42], MFR [7], MIG [29], RDN [14] and RMIG [8]. From the red box in **Fig. 4**, it can be clearly found that the depth maps predicted by different methods have some differences in details. For instance, in depth maps predicted by CGN [42], MFR [7], MIG [29], RDN [14] and RMIG [8], the boundary of the depth map at the intersection of the rod and ring column is blurred. In contrast, this method can well recover depth boundaries of tiny objects. In addition,

to demonstrate the superiority of the proposed method, we use the above-mentioned methods and the **Art** depth map obtained by this method with the corresponding high-resolution color image to reconstruct three-dimensional visual images. As shown in **Fig. 5**, the enlarged image in the red box clearly shows that there are serious twists in the bars for the results of CGN [42], MFR [7], MIG [29], RDN [14] and RMIG [8]. Since the above method reconstructs three-dimensional visual images using the same high-resolution color images, it can be demonstrated that the depth SR images predicted by this method are more similar to the GT images.

Table 3. Ablation studies of the proposed method

Components	CCE	DHLF	CHLF	PSNR	SSIM	RMSE	MAD
Ours-1	×	√	√	40.65	0.9814	2.44	0.99
Ours-2	√	×	√	37.35	0.9765	3.60	1.31
Ours-3	√	√	×	**41.30**	**0.9828**	**2.26**	**0.93**
Ours-4	√	×	×	37.13	0.9762	3.68	1.34
Ours(32)	√	√	√	**41.40**	**0.9830**	**2.24**	**0.92**

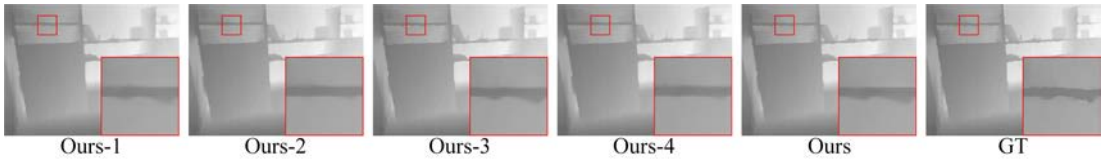

Fig. 6. The visual comparison of ablation studies for the proposed method testing on the 1005-th depth map from NYU-v2 RGB-D Dataset.

4.4 Ablation Studies

To verify the superiority of sampling-based color HF prediction block, depth high-frequency updating block, and criss-cross enhancement (CCE) module, we conduct a series of ablation studies on the NYU-v2 dataset to verify the importance of each module in the proposed method, as shown in **Table 3**. We can get the Ours-1 model when the CCE module is removed and other key modules remain unchanged. PSNR and SSIM of Ours-1 are reduced by 0.75 and 0.0016 respectively. The RMSE and MAD values of Ours-1 are increased by 0.2 and 0.07 respectively. To observe the influence of DHLF feature decomposition and Color High-Low Frequency (CHLF) feature decomposition on network performance. Here, DHLF feature decomposition refers to depth high-frequency updating block, while CHLF feature decomposition denotes sampling-based color HF prediction block. We can get Ours-2 and Ours-3 when DHLF feature decomposition and CHLF feature decomposition are removed respectively. If both of them are removed, we can get Ours-4 model. When DHLF feature decomposition or CHLF feature decomposition is removed, PSNR and SSIM of Ours-2/Ours-3 decrease by 4.05/0.10 and 0.0065/0.0002 respectively. Meanwhile, the RMSE and MAD of Ours-2 increase by 1.36 and 0.39 respectively, while RMSE and MAD of Ours-3 increase by 0.02 and 0.01 respectively. When DHLF feature decomposition and CHLF feature decomposition are removed, PSNR and SSIM decrease by 4.27 and 0.0068. And the RMSE and MAD increase by 1.44 and 0.42 respectively. From these results, it can be found that these three blocks are essential for high-quality depth map SR.

As shown in **Fig. 6**, we compare the visual effects of 8x depth SR under different configurations of our method, when this method is tested on the NYU-v2 dataset. From the enlarged detail of the red box in **Fig. 6**, it is obviously seen that the results of Ours-2 and Ours-4 have excessive smoothing effects, and they lack fine structure and boundary details as compared with the super-resolved depth image predicted by our entire model. Although Ours-1 and Ours-3 are relatively clear, there are still subtle differences of the super-resolved depth images between Ours-1/Ours-3 and the proposed entire model. These visual quality comparisons of the ablation study further demonstrate the effectiveness of each component of the proposed method.

5. Conclusion

In this paper, we propose a color-image guided DMSR method based on iterative depth feature enhancement. Considering the feature difference between high-quality color features and low-quality depth features, we decompose and reorganize the high-resolution color map and low-resolution depth map many times to achieve the purpose of color map guided depth map enhancement. Specifically, the high-frequency detail features obtained from the sampled color high-frequency prediction block are fed into the depth high-frequency updating block stage by stage, so as to realize the update of low-quality depth high-frequency details. Finally, a multi-stage depth reconstruction block is proposed to estimate the final high-quality depth map. A large number of experimental results show that our method has better performance than many advanced depth mapping SR methods. Considering that high-resolution color images may be affected by the dark environment in daily shooting, our future work will study the brightness recovery of DMSR and low-illumination color maps at the same time.

Acknowledgement

This work was supported by National Natural Science Foundation of China Youth Science Foundation Project (No.62202323), Fundamental Research Program of Shanxi Province (No.202103021223284), Taiyuan University of Science and Technology Scientific Research Initial Funding (No.20192023, No.20192055) and National Natural Science Foundation of China (No.62072325).

References

[1] Poole, Ben, and J. T. Barron, “The fast bilateral solver,” in *Proc. of the European conference on computer vision (ECCV)*, 2016. Article (CrossRef Link)

[2] He, Kaiming, Jian Sun, and Xiaoou Tang, “Guided image filtering,” in *Proc. of European conference on computer vision*, pp. 1-14, 2010. Article (CrossRef Link)

[3] Kopf, Johannes, et al, “Joint bilateral upsampling,” *ACM Transactions on Graphics (ToG)*, 26.3, 96-es, 2007. Article (CrossRef Link)

[4] Ham, Bumsub, Minsu Cho, and Jean Ponce, “Robust guided image filtering using nonconvex potentials,” *IEEE transactions on pattern analysis and machine intelligence*, 40.1, 192-207, 2017. Article (CrossRef Link)

[5] Ferstl, David, et al, “Image guided depth upsampling using anisotropic total generalized variation,” in *Proc. of the IEEE international conference on computer vision*, 2013. Article (CrossRef Link)

[6] Diebel, James, and Sebastian Thrun, “An application of markov random fields to range sensing,” *Advances in neural information processing systems*, 18, 2005.

[7] Zuo, Yifan, et al, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Transactions on Circuits and Systems for Video Technology*, 30.2, 297-306, 2019. Article (CrossRef Link)
[8] Zuo, Yifan, et al, "Depth map enhancement by revisiting multi-scale intensity guidance within coarse-to-fine stages," *IEEE Transactions on Circuits and Systems for Video Technology*, 30.12, 4676-4687, 2019. Article (CrossRef Link)
[9] Li, Yijun, et al, "Joint image filtering with deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, 41.8, 1909-1923, 2019. Article (CrossRef Link)
[10] Kim, Beomjun, Jean Ponce, and Bumsub Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, 129.2, 579-600, 2021. Article (CrossRef Link)
[11] Hui, Tak-Wai, Chen Change Loy, and Xiaoou Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. of European conference on computer vision*, pp. 353-369, 2016. Article (CrossRef Link)
[12] He, Lingzhi, et al, "Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. Article (CrossRef Link)
[13] Chen, Yunpeng, et al, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2019. Article (CrossRef Link)
[14] Zuo, Yifan, et al, "Residual dense network for intensity-guided depth map enhancement," *Information Sciences*, 495, 52-64, 2019. Article (CrossRef Link)
[15] Li, Yijun, et al, "Deep joint image filtering," in *Proc. of European conference on computer vision*, pp. 154-169, 2016. Article (CrossRef Link)
[16] Ye, Xinchen, et al, "Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution," *IEEE Transactions on Image Processing*, 29, 7427-7442, 2020. Article (CrossRef Link)
[17] Sandler, Mark, et al, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2018. Article (CrossRef Link)
[18] Ma, Ningning, et al, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. of the European conference on computer vision (ECCV)*, 2018. Article (CrossRef Link)
[19] Wang, Xiaochuan, Kai Wang, and Xiaohui Liang, "Single Depth Map Super-resolution with Local Self-similarity," in *Proc. of the 2018 the 2nd International Conference on Video and Image Processing*, pp. 198-202, 2018.Article (CrossRef Link).
[20] Chen, Baoliang, and Cheolkon Jung, "Single depth image super-resolution using convolutional neural networks," in *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018. Article (CrossRef Link)
[21] Chen, Yuantao, et al, "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, 51.7, 4367-4380, 2021. Article (CrossRef Link)
[22] Chen, Wenlong, et al, "Multi-scale feature aggregation network for image super-resolution," *Applied Intelligence*, 52.4, 3577-3586, 2022. Article (CrossRef Link)
[23] Du, Xiaobiao, "Single image super-resolution using global enhanced upscale network," *Applied Intelligence*, 52.3, 2813-2819, 2022. Article (CrossRef Link)
[24] Huang, Liqin, et al, "Pyramid-structured depth map super-resolution based on deep dense-residual network," *IEEE Signal Processing Letters*, 26.12, 1723-1727, 2019. Article (CrossRef Link)
[25] Song, Xibin, et al, "Channel attention based iterative residual learning for depth map super-resolution," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. Article (CrossRef Link)
[26] Wu, Guoliang, Yanjie Wang, and Shi Li, "Single depth map super-resolution via a deep feedback network," *International Journal of Wavelets, Multiresolution and Information Processing*, 19.02, 2050072, 2021. Article (CrossRef Link)
[27] Ye, Xinchen, et al, "Depth super-resolution via deep controllable slicing network," in *Proc. of the 28th ACM International Conference on Multimedia*, pp. 1809-1818, 2020. Article (CrossRef Link)

[28] Yang, Yuxiang, et al, "Depth map super-resolution using stereo-vision-assisted model," *Neurocomputing*, 149, 1396-1406, 2015. Article (CrossRef Link)
[29] Zuo, Yifan, et al, "MIG-net: Multi-scale Network Alternatively Guided by Intensity and Gradient Features for Depth Map Super-resolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3506-3519, 2021. Article (CrossRef Link)
[30] Wang, Zhihui, et al, "Depth upsampling based on deep edge-aware learning," *Pattern Recognition*, 103, 107274, 2020. Article (CrossRef Link)
[31] Zhang, H., X. Tan, and X. Li, "Towards Lighter and Faster: Learning Wavelets Progressively for Image Super-Resolution," in *Proc. of MM '20: The 28th ACM International Conference on Multimedia ACM*, pp. 2113-2121, 2020. Article (CrossRef Link)
[32] Guo, Chunle, et al, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Transactions on Image Processing*, 28.5, 2545-2557, 2018. Article (CrossRef Link)
[33] Cao, Xiang, et al, "DAEANet: Dual auto-encoder attention network for depth map super-resolution," *Neurocomputing*, 454, 350-360, 2021. Article (CrossRef Link)
[34] Tang, Qi, et al, "BridgeNet: A Joint Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation," in *Proc. of the 29th ACM International Conference on Multimedia*, pp. 2148-2157, 2021. Article (CrossRef Link)
[35] Park, Jaesik, et al, "High quality depth map upsampling for 3d-tof cameras," in *Proc. of 2011 International Conference on Computer Vision*, IEEE, 2011. Article (CrossRef Link)
[36] Sun, Baoli, et al, "Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. Article (CrossRef Link)
[37] Lu, Jiangbo, et al, "Cross-based local multipoint filtering," in *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012. Article (CrossRef Link)
[38] Liu, Ming-Yu, Oncel Tuzel, and Yuichi Taguchi, "Joint geodesic upsampling of depth images," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2013. Article (CrossRef Link)
[39] Mac Aodha, Oisin, et al, "Patch based synthesis for single depth image super-resolution," in *Proc. of European conference on computer vision*, pp. 71-84, 2012. Article (CrossRef Link)
[40] Xie, Jun, et al, "Single depth image super resolution and denoising via coupled dictionary learning with local constraints and shock filtering," in *Proc. of 2014 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2014. Article (CrossRef Link)
[41] Xie, Jun, Rogerio Schmidt Feris, and Ming-Ting Sun, "Edge-guided single depth image super resolution," *IEEE Transactions on Image Processing*, 25.1, 428-438, 2015. Article (CrossRef Link)
[42] Zuo, Yifan, et al, "Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensity-guided refinement," *IEEE Transactions on Multimedia*, 23, 772-783, 2020. Article (CrossRef Link)

Lijun Zhao received his M.S. degree from Taiyuan University of Science and Technology in 2015 and PhD degree from Beijing Jiaotong University in 2019. He is currently an associate professor in the Institute of Digital Media and Communication, Taiyuan University of Science and Technology. His research interests include image coding, multiple description coding, 3D video processing, pattern recognition, computer vision, etc.

Ke Wang received his M.S. degree from Taiyuan University of Science and Technology in 2023. His research interests include depth map super-resolution and depth estimation.

Jinjing Zhang received her B.S. and PhD degree from North University of China in Taiyuan, China, in 2014 and 2022. She is now a lecturer in North University of China. Her research interests include image segmentation and image enhancement.

Jialong Zhang is now studying in Taiyuan University of Science and Technology. His research interests include depth map super-resolution and multi-modality image enhancement.

Anhong Wang received B.S. and M.S. degrees from Taiyuan University of Science and Technology, China respectively in 1994 and 2002, and PhD degree in Institute of Information Science, Beijing Jiaotong University (BJTU) in 2009. She became an associate professor with TYUST in 2005 and became a professor in 2009. She is now the director of Institute of Digital Media and Communication, Taiyuan University of Science and Technology. Her research interests include image and video coding and secret image sharing, etc.