IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 13 March 2023, accepted 2 May 2023, date of publication 10 May 2023, date of current version 17 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274966

RESEARCH ARTICLE

Improved Upsampling Based Depth Image Super-Resolution Reconstruction

YANMING YE[1,2], MENGXIONG ZHOU[1], ZHANYU WANG[3], AND XINGFA SHEN[2], (Member, IEEE)
[1]School of Information Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[2]School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China
[3]School of Computer Science and Engineering, Changshu Institute of Technology, Changshu 215500, China

Corresponding author: Yanming Ye (yeym@hdu.edu.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62072121, in part by the ''Pioneer'' and ''Leading Goose'' Research and Development Program of Zhejiang under Grant 2023C01143, and in part by the Zhejiang teaching reform project of undergraduate colleges and universities under Grant JG20220744.

ABSTRACT Constrained by current sensing technology, depth camera only acquires a low-resolution depth image that does not meet actual requirements. To solve this problem, this paper take a divide-and-conquer strategy to synthesize a high-resolution depth image from a low-resolution range image under the guidance of a registered high-resolution color image. Initially, the depth image is divided into planar areas and edge regions. For different zones, we exploit different methods to interpolate the missing depths. At planar area, the linear interpolation method is employed to perform upsampling. At edge region, a segmentation-separation upsampling method is used to interpolate the missing values. Then the upsampling result are refined on the Depth CNN that is built in this paper. We conduct extensive experiments on the benchmark database and real world data with various upsampling rates to illustrate the upsampling ability of our method. The comparison with classical super-resolution algorithms demonstrates that our upsampling algorithm achieves the best quality with fewer artifacts and our depth CNN outperforms the most state-of-the-art methods in terms of qualitative and quantitative evaluations.

INDEX TERMS Depth image, upsampling, super-resolution reconstruction, edge guided.

I. INTRODUCTION

In recent years, depth image has been applied more and more widely in the field of computer vision, telemedicine, driverless and security monitoring to improve the performance of products. However, the resolution of the depth image collected by the sensing devices is relatively low due to the equipment's limitation. For example, the resolution of the depth image collected by the Mesa Swiss Ranger 4000 or Microsoft Kinect V2 is only 176×144 or 512×424, which is difficult to meet the actual requirements [1]. Therefore, how to reconstruct high resolution (HR) depth images from low resolution (LR) depth images has become one of the research hotspots.

The upsampling operation of feature image is very important for image restoration in reconstruction. Different upsampling operations may directly affect the quality of reconstructed image. Existing depth image upsampling methods can be roughly categorized into filtering based methods, optimization based methods and deep-learning based methods. Filtering based methods employ the color information of color photo or texture information of texture image with various edge-preserving filters. Chang et al. [2] use potency guided upsampling and adaptive gradient fusion filters to enhance the erroneous depth images to refine the upsampled depth results. Qiao et al. [3] construct a feature-based bilateral filter (FBF) for the interpolation, by using the extracted RGB shallow and multi-layer features to improve the upsampled depth image quality. Liu et al. [4] puts forward a precise three-dimensional (3D) reconstruction method for uncooperative spacecraft based on a low-resolution time-of-flight (ToF) depth camera coupled with a high-resolution optical texture camera. Yang et al. [5] build a novel joint trilateral filter with two different modes: one for the pixels on the edges and the other for the pixels in the smooth regions. Jiang et al. [6] propose a deep edge map guided depth SR

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

method which includes an edge prediction subnetwork and an SR subnetwork. Yang et al. [7] use depth-texture similarity to construct a pixel-level confidence calculation method for 3D view synthesis, and construct a joint guided filter based on confidence, which not only considered the smoothness between depth pixels, but also incorporated depth-texture similarity, improving the performance of sampling on depth images and the quality of synthesized views. Lei et al. [8] propose a view synthesis quality based trilateral depth-map upsampling method, which considers depth smoothness, texture similarity and view synthesis quality in the upsampling filter. Filtering based methods generally use the local or non-local neighborhood relationship of the depth image to estimate the high-resolution depth value. One advantage of filtering-based methods is that they can be easily parallelized on graphics hardware. However, to find enough support for each pixel, large filtering kernels are often used, or the filters have to be performed iteratively, which may lead to over-smoothed depth results.

The second category is the optimization-based methods. Sharma [9] presents a depth image enhancement algorithm based on Riemannian Geometry that performs depth image de-noising and completion simultaneously. Chen et al. [10] propose a new optimization model depending on the relative structures of both depth and color images for both depth image filtering and upsampling tasks. Yan et al. [11] use the non-local mean algorithm to obtain the initial upsampling depth image first, and then optimized it using the edge detection algorithm to improve the reconstruction quality of the depth image. Maxim et al. [12] present a new fuzzy method for creating a depth image based on a combination of Canny detector with a three-level fuzzy system. Jung et al. [13] propose a post-processing algorithm to refine the depth image using super-pixel segmentation and considering the relation between multiple views. Wang et al. [14] propose an RGB-guided depth image recovery method to recover true boundaries in seriously distorted depth images. The optimization-based methods generally use the depth image degradation model and various prior knowledge to transform the reconstruction problem into the cost function optimization problem. One advantage of filtering-based methods is the applicability to multiple degenerate models and only needs to change the data item of the cost function. However, these methods may lead to high computational complexity and the selection of prior knowledge has a great impact on performance.

In recent years, convolution neural network has made remarkable achievements in the field of image super-resolution reconstruction. Dong et al. [15] proposed the end-to-end SRCNN (Super resolution convolution neural network) network, which can learn the mapping relationship from low resolution to high resolution and solve the problem of image super-resolution reconstruction based on depth learning. SRCNN is considered as the first work of the third category that based on deep-learning. Since then, a large number of researchers have tried to realize depth image super-resolution reconstruction using convolutional neural networks. Lim et al. [16] and Shi et al. [17] improve SRCNN and increase the depth of the network. Zhang et al. [18] and Zhou et al. [19] improve the quality of hyper-differential reconstruction and limit the model parameters based on the recursive residual network which effectively reduced the computational complexity. Cao et al. [20] propose a novel dual auto-encoder attention network (DAEANet) which includes two auto-encoder networks, where guidance auto-encoder network (GAENet) and target auto-encoder network (TAENet) aim to extract feature information from intensity image and depth image. Kim et al. [21] propose a novel depth image super-resolution method using guided deformable convolution, which obtains 2D kernel offsets of the depth features from the guidance features to significantly alleviate the texture copying artifacts in the resultant depth image. Guo et al. [22] propose a two-branch network to achieve depth image super-resolution with high-resolution guidance image, which can be viewed as a prior to guide the low-resolution depth image to restore the missing high-frequency details of structures. Deep-learning based methods use manifold learning, sparse coding, depth learning and other strategies to learn the relationship between high-resolution and low-resolution depth images by the training of a large number of data. The advantage of deep-learning based methods is the good reconstruction performance. However, the training process of these methods is time-consuming and the selection of training sets has a great impact on the performance.

In general, the filtering based methods are applicable to the scenarios that need strong real-time performance, and the optimization based methods are applicable to the scenarios with multiple degradation coexisting or serious degradation in the depth image, and the deep-learning based methods are applicable to the scenarios with a large number of training samples. In this paper, we try to use a improved upsampling algorithm that comprehensively utilize both optimization based and filtering based methods to obtain a higher resolution image, and then use the deep neural network to refine it.

The first part of the work focuses on the improved edge guided depth image upsampling. The new upsampling algorithm not only should take advantage of the characteristic of depth image, but also should preserve the sharp edges while upscaling range image with large upsampling rate. First, unlike general nature image, depth image belongs to cartoon image. Generally speaking, cartoon images are composed of planar patches with sharp transitions between the boundaries of planar patches. For high quality upsmapling result, it is necessary to embed the characteristic into upsampling model; Second, the resolutions of depth sensors are extreme low and we often need the same size depth image as optical photo whose solution is at least 1000 × 1000. This situation implies that the upsampling rate is usually very large; Third,

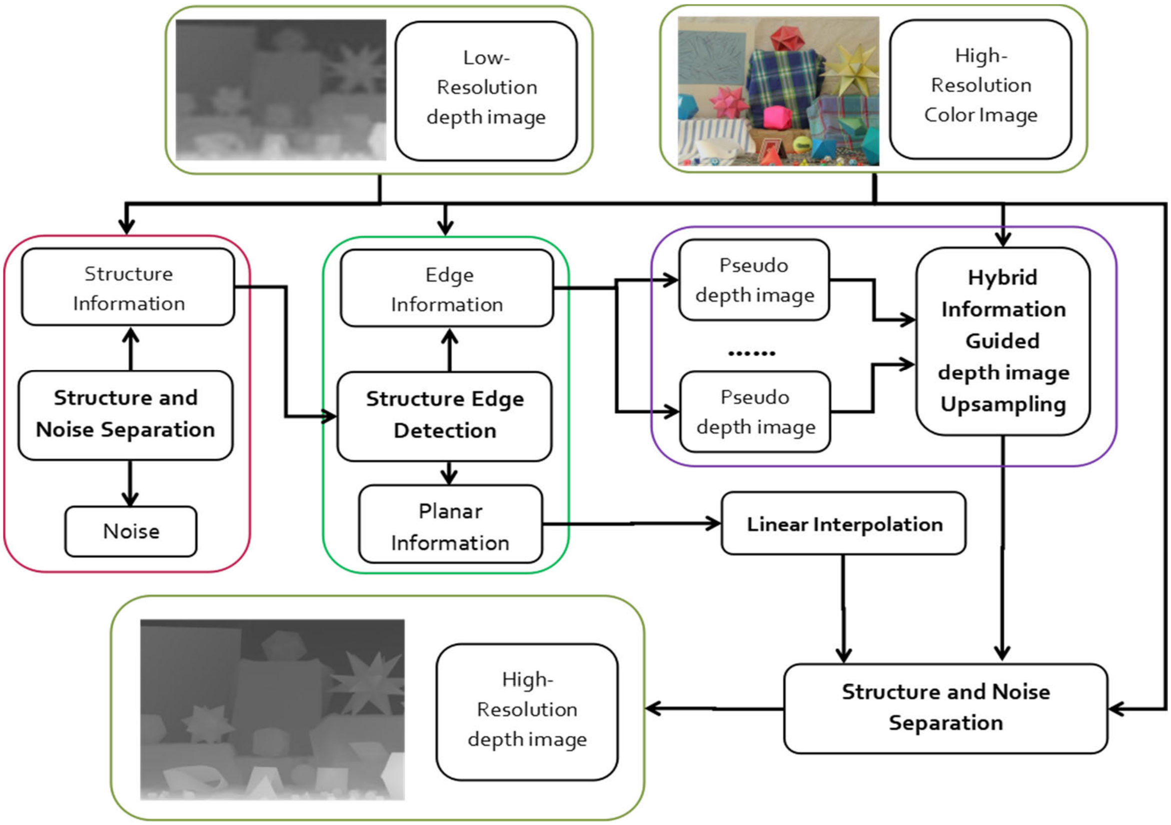

Low-Resolution depth image
High-Resolution Color Image
Structure Information
Structure and Noise Separation
Noise
Edge Information
Structure Edge Detection
Planar Information
Pseudo depth image
......
Pseudo depth image
Hybrid Information Guided depth image Upsampling
Linear Interpolation
Structure and Noise Separation
High-Resolution depth image

FIGURE 1. The pipeline of our upsampling algorithm.

Due to the large upsampling rate, geometric information in the LR depth image is not enough to produce fine details. To supply extra edge information and suppress the blurring effect, people advocates that depth and color boundaries of a scene are closed correlated: abrupt depth transition often leads to abrupt color transition. Thus one way to enhance the resolution of depth image is to use a high resolution optical camera in tandem with the depth sensor. In this way, we can obtain a one-to-one relationship between pixels of HR depth image and color image. Further, the acquired LR depth image is mapped into HR depth image and forms a set of sparse seeds in the HR depth image. We will interpolate the missing depth values under the guidance of the registered color photo.

Previous depth upsampling methods tend to produce artifacts at the planar areas and likely smooth the sharp edges in the edge regions. These methods directly use the registered color photo to indicate the geometric structure of HR depth image. However, color edges do not completely coincide with depth edges. A planar area of HR depth image usually corresponds to many color edges. The mismatching situation will inevitable introduce artifacts into planar areas. Previous methods also do not explicitly model the cartoon-like depth image. On the contrary, they employ weight coefficient to indicate the similarity between coupled pixels. Large weight signifies the coupled pixels are likely in the same region with same depths, otherwise, they belong to different areas and should be assigned different depths. For preserving shape depth edges, we hope the weight coefficients tend to zero, when the coupled pixels belong to different regions. However, the weight coefficient would not be zero, even if the coupled pixels actually across the depth boundary, according to the definition of weight coefficient of previous methods. Therefore these methods will inevitable leak the depth information between different regions and produce intermediate values that could smooth the sharp depth edges.

We model the characteristics of depth image and take a divide-and-conquer strategy to upscale LR depth image. Initially, the LR depth is segmented into planar areas and edge regions, and these zones are mapped into HR depth image. The process not only can form the planar areas and edge regions of depth image, but also could determine the reference areas on the registered color photo for different regions of depth image. We exploit different methods to interpolate the missing depths of different regions. At planar area, we employ the simplest linear interpolation method to perform upsampling, because it could achieve competitive performance with least cost. At edge region, we calculate a set of pseudo depth images and exploit them to compute correlation coefficients of interpolated pixel with respect to given seeds. The most correlative depths of seeds are chosen as interpolation values. This segmentation-like interpolation method unavoidably introduces zigzag artifacts into edge regions. We develop a separation algorithm to separate them. The segmentation-separation edge region upsampling method would not produce any intermediate values, thus it can perfectly preserve the sharp edges.

The second part of the work focuses on the neural network for high resolution reconstruction of depth images. In this paper, we try to build a new CNN network for depth image super- with pre-upsampling structure and dense connections based on DenseNet and ResNet. The main contributions of our work are:

- We compare the upsampling performance of classical algorithms for different areas and find out a way to improve the upsampling quality.
- Our upsampling algorithm takes a divide-and-conquer strategy. For planar area, we exploit the linear interpolation method. For edge region, we take advantage of the segmentation-separation interpolation method.
- The proposed depth CNN will refine the rough image obtained from upsampling result, which significantly reduces the learning difficulty and can take interpolation with arbitrary size and scaling factor.
- No matter the visual quality or the quantity evaluation, the results of our method outperform compared methods.

II. UPSAMPLING ALGORITHM

We comprehensively utilize both optimization-based and filtering-based methods to perform upsampling, and demonstrate the pipeline of our upsampling algorithm in Fig.1. Initially, the geometric structure is separated from the contaminated LR depth image; Then, we detect the edge regions and the planar areas of LR depth image, after that we take customized upsampling methods to upscale each region. At last, the synthesized result is processed again by our separation algorithm to remove nasty artifacts.

The acquired data by current depth capturing devices is somewhat like a degraded version of the underlying ground truth, we thus separate the geometric structure and noise from the LR range map under the guidance of the downsampled photo image, shown in the red box. This step could remarkably stabilize the edge detection in the following process, illustrated in the green box. We use different upsampling methods to interpolate different areas. For planar areas, the simplest linear interpolation could produce comparative results. For edge regions, we take a segmentation separation method to interpolate the missing values. Finally, the synthesized result is processed again by our separation algorithm to remove the nasty artifacts.

A. STRUCTURE AND NOISE SEPARATION

Both acquired LR depth image $D^{\downarrow}$ and upsampled HR depth image D are likely contaminated by various noise and artifacts. We formulate the corrupted depth image as $z = x + y$, where x is the geometrical structure and y is the contaminated noise or artifacts. Both components are of arbitrary magnitude. We use the well-known non-local regularization [23], [24] to separate the geometric structure x and the noise component y from the contaminated z by solving following optimization problem:

$$\begin{aligned}&\sum_{k\in O}\|D_k x\|_1 + \tau\|y\|_1\\ &s.t.\ x + y = z\end{aligned} \tag{1}$$

where D_k represents the weighted differential operator at the direction k, i.e. $D_k x = \sum_{i\in S} w_{ik}^{\downarrow}|x_i - x_{i+k}|$, x_i denotes the value of x at position i, $V^{\downarrow}$ is the pixel set of LR depth image, O denotes the relative offsets of neighborhood pixels with respect to central pixel i and $w_{ik}^{\downarrow} = \exp(-\|k\|^2/\sigma_s^{\downarrow})\exp(-\left\|I_i^{\downarrow} - I_j^{\downarrow}\right\|^2/\sigma_c^{\downarrow})$ is the weight coefficient calculated from downsampled color image $I^{\downarrow}$.

We decompose objective function (1) into three subproblems (2)-(4) that have closed form solutions, using well-known split Bregman method [25], [26].

$$\min_x \left\|x + y^k - b_0^{k+1}\right\|_2^2 + \frac{\lambda_2}{\lambda_1}\sum\left\|D_i x - d_i^k - b_i^{k+1}\right\|_2^2 \tag{2}$$

$$\min_y \tau\|y\|_1 + \frac{\lambda_1}{2}\left\|y + x^{k+1}\text{-}\mathrm{b}_0^{k+1}\right\|_2^2 \tag{3}$$

$$\min_{d_i} \|d_i\|_1 + \frac{\lambda_2}{2}\left\|D_i x^{k+1} - d_i - b_i^{k+1}\right\| \tag{4}$$

Let $b_0^{k+1} = b_0^k + f - (x^k + y^k) b_i^{k+1} = b_i^k - D_i x^k + d_i^k$ and $shrink(y, a) = \mathrm{sgn}(y)\max\{|y| - a, 0\}$, the closed form solutions of subproblem (2)-(4) are given by:

$$x^{k+1} = \frac{(b_0^{k+1} - y^k + \frac{\lambda_2}{\lambda_1}\sum D_i^T(b_i^{k+1} + d_i^k))}{(I + \frac{\lambda_2}{\lambda_1}\sum D_i^T D_i)} \tag{5}$$

$$y^{k+1} = shrink\left(b_0^{k+1} - x^{k+1}, \frac{\tau}{\lambda_1}\right) \tag{6}$$

$$d_i^{k+1} = shrink\left(D_i x^{k+1} - b_i^{k+1}, \frac{1}{\lambda_2}\right) \tag{7}$$

Separating results x and y can be achieved by iteratively computing x^k and y^k according to formula *(5)-(7)*. Finally, we note that the separation algorithm not only is a preprocessing step to suppress the noise in the acquired LR depth image, but also is a post-processing step to remove artifacts produced by our upsampling method.

B. PLANAR AREA UPSAMPLING

In this section, we will show which method is most suitable for the planar area upsampling by evaluating the bad pixel distribution of previous algorithms. The 8X statistics data listed in Table 1 are collected from the results of highly cited classical naive methods [27], [28], [29], [30], [31]. Table 1 not only proves that, for planar area upsampling, linear interpolation method is the most appropriate upsampling algorithm with the highest cost-performance, but also indicates our divide-and-conquer upsampling strategy is a reasonable method.

To determine planar areas and edge regions of HR depth image D, we detect planar areas and edge regions in the denoised LR depth image $D^{\downarrow}$. A pixel $i \in D^{\downarrow}$ is classified as 'planar' if its depth value $D^{\downarrow}(i)$ satisfies $|D^{\downarrow}(i) - D^{\downarrow}(j)| < \lambda, \forall j \in N(i)$, otherwise it falls into the

TABLE 1. Properties of raw materials bad pixels distribution of 8X upsampling.

		PP	PE	PBP	BMP	PBE	BME	RPP
Art	MRF[27]	55.15%	44.85%	9.77%	43209	90.23%	399042	-0.0783
	BF[28]			9.36%	53832	90.64%	521335	0.1482
	GF[29]			9.44%	48786	90.56%	468097	0.0406
	NLM[30]			14.46%	75172	85.54%	444762	0.6034
	AR[31]			13.06%	39712	86.94%	264376	-0.1529
Book	MRF[27]	70.21%	29.79%	19.48%	61084	80.52%	252465	-0.0295
	BF[28]			29.74%	140901	70.26%	332802	1.2386
	GF[29]			19.52%	67463	80.48%	278079	0.0718
	NLM[30]			25.06%	105211	74.94%	314694	0.6716
	AR[31]			21.50%	68041	78.50%	248466	0.0810
Moebius	MRF[27]	63.10%	36.90%	11.82%	39126	88.18%	292022	-0.0459
	BF[28]			28.12%	148012	71.88%	378299	2.6094
	GF[29]			11.79%	40571	88.21%	303503	-0.0106
	NLM[30]			15.76%	57451	84.24%	306988	0.4010
	AR[31]			12.67%	37448	87.33%	258126	-0.0868

edge region, where $N(i)$ is the second order neighborhood of pixel i, and λ is a depth threshold value. For U rate upsampling, a pixel $i \in D^{\downarrow}$ corresponds to a $U \times U$ patch in D. Thus, we can map planar areas and edge regions of $D^{\downarrow}$ into D and determine corresponding planar areas and edge regions of D.

We calculate the percentage of planar areas and edge regions (PP and PE) to reveal which zone dominates depth image. The percentage of bad matching pixels (PBP and PBE) at planar areas and edge regions are given to compare the upsampling quality of different methods, where the threshold is set to 1. Other than PBP and PBE, we also count the bad matching pixel number (BMP and BME) for planar areas and edge regions. To choose the best planar area upsampling method, the relative performances (RPP) of different methods are figured out, using formula (8).

$$RPP_{ref} = \frac{(BPP_{ref} - BPP_{baseline})}{BPP_{baseline}} \tag{8}$$

where BPP_{ref} denotes the BPP performance of comparison methods [27], [28], [29], [30], [31] and $BPP_{baseline}$ is BPP performance of linear interpolation method.

The performance discrepancy of different methods at planar areas is not distinctive. From the data reported at RPP row of Table 1, we can observe that the well-designed methods MRF [27] and AR [31] do not significantly surpass the linear interpolation method, and the performance of BF [28] and NLM [30] are even worse. The upsampling methods [27], [28], [29], [30], [31] assume that similar colors has similar depths, and utilize color information of color photo to indicate the geometric structure of planar areas of depth image. But, a planar area usually corresponds to multiple color edges, the mismatch will introduce artifacts and could degrade upsampled planar regions. Depth surfaces of planar areas are usually planes, thus the linear model used by linear interpolation method is an appropriate upsampling model for planar areas. We conclude that the simplest linear interpolation method is the most appropriate upsampling algorithm for planar area upsampling.

Table 1 also reveals that over half of the bad pixels of each method exist in the edge region of depth image (refer to PBE row), while most of the pixels in the disparity belong to the planar region (refer to PP row). Moreover, there is a connection between BMP and BME. The lower bad pixel number of edge regions suggests lower bad pixels number of the planar areas (refer to BMP and BME rows). We conclude that the restoration quality of the upsampling method is dependent on the restoration ability of upsampling method in the boundary area, thus edge regions deserve more attention and should be carefully handled to get a satisfactory upsampling result.

According to above discussion, we ought to pay more attention on edge region upsampling and use linear interpolation to simplify the computation procedure of planar areas.

C. EDGE REGION UPSAMPLING

Edge region upsampling ability of an algorithm determines the quality of final result. To improve the upsampling quality, we introduce a set of pseudo depth images which can represent the geometric structure of HR depth image and then jointly use these pseudo depth images to compute the correlation coefficients of interpolated pixels with respect to seeds. Finally, the most correlative depth values of seeds are assigned to the missing value pixels of HR depth image.

1) PSEUDO DEPTH IMAGE ESTIMATION

Classical upsampling methods [27], [28], [30], [31] directly encode color edge information into their upsampling models, based on the assumption that color edge with high contrast indicates abrupt depth edge. The upsampling strategy limits further improvement, because edges of color photo do not completely coincide with edges of depth image. Sharp depth edge often corresponds to a blurred color edge with low contrast, thus previous upsampling algorithms inevitable produce intermediate depth value and the depth transition in the boundary will be smoothed by the intermediate depths.

We retrieve pseudo depth images from color photo to indicate the structure of HR depth image, instead of directly using color guidance. The edges of pseudo depth image will coincide with the edges of HR depth image. The values of pseudo depth image are only used to indicate the edges instead of representing the actual depth. Thus we can freely scale the values of pseudo depth image.

We use a segmentation-like interpolation algorithm to compute the pseudo depth image $\tilde{D}$. First, we quantify the depth range of LR depth image and obtain a pseudo LR depth image by mapping disjoint depth interval to its median value. In this way, the contaminated noise can be suppressed and the edges will be enhanced. Second, the pseudo LR depth image is mapped into pseudo depth image and forms a set of seeds. Third, we compute the correlation coefficients between seeds and interpolated pixels, where the correlation coefficient p_{ij} of missing value pixel j with respect to seed i is defined as

$$p_{ij} = \exp\left(-\|i-j\|^2/\sigma_s\right)\exp\left(-\left\|I_i - I_j\right\|^2/\sigma_c\right) \tag{9}$$

Fourth, the quantized depth value of the most correlative seed is chosen as the interpolated depth for each interpolated pixel. It is worth noting that the assignment process of our algorithm can be interpreted as a kind of segmentation process, if we view the quantized depth value as a label for each seed.

Pseudo depth image can preserve sharp edges of LR depth image, even if the corresponding color edges are blurred, for the reason that the depth interpolation, in a sense, is a kind of segmentation that does not yield any intermediate value. Therefore pseudo depth image is better than color photo to represent the geometric structure of depth image.

2) EDGE GUIDED DEPTH IMAGE UPSAMPLING

In this section, we will take advantage of random walk model [32], [33] to compute the most correlative depth under the guidance of both edge information of pseudo depth image and color image. Let N_i^f be the first order neighborhood, w_{ij} be the weight coefficient between i and $j \in N_i^f$ and N_s denote a window neighborhood of seed s, where N_s must be large enough to contain other seeds around s. We partition the pixels of N_s into two sets, v_S (seeds) and v_U (unseeded nodes), and use S_d to represent the set of v_S whose depth values are equal to d in the N_s. Further, we define indicator vector δ^{S_d} of S_d as $\delta_S^{S_d} = 1$ if $s \in S_d$, otherwise $\delta_S^{S_d} = 0$. Then we can use optimization problem (10) to estimate the correlation coefficients p_U of v_U with respect to S_d.

$$\begin{aligned}&\min_{p_U} E(p_U, p_S)\\ &s.t. p_S = \delta^{\mathrm{S_d}}\end{aligned} \tag{10}$$

where

$$E(p_U, p_S) = \left[p_S^T p_U^T\right] L \begin{bmatrix} p_S \\ p_U \end{bmatrix} = \left[p_S^T p_U^T\right] \begin{bmatrix} L_S & B \\ B & L_U \end{bmatrix} \begin{bmatrix} p_S \\ p_U \end{bmatrix} \tag{11}$$

L is the combinatorial Laplacian matrix, which is defined as

$$L_{ij} = \begin{cases} \sum_{j' \in N_i^f} w_{ij'} & \text{if } i = j \\ -w_{ij} & \text{if } j \in N_i^f \\ 0 & otherwise \end{cases} \tag{12}$$

For seeds with depth value d, the minimal point can be computed by $p_u = -L_u^{-1}B^T\delta^{S_d}$ according to the norm equation of $E(p_U, p_S)$. The correlation coefficients with respect to depth value $\{d_1 \ldots d_n\}$ can be figured out by $p_u = -L_u^{-1}B^T\left[\delta^{S_{d_1}} \ldots \delta^{S_{d_n}}\right]$ simultaneously. Finally, we ffnd the maximal correlation coefficient for each row of P_u to determine the most correlative depth.

Our weight coefficient calculating procedure is much simpler than NLM [7] and AR [8]. Instead of computing complicate segmentation, edge saliency, anisotropic structural-aware filter and patched based bilateral filter as NLM [7] and AR [8] do, we only employ a bilateral-filter-like kernel to compute the coefficients w_{ij}^c and w_{ij}^d from color image and pseudo depth image, then add them together to obtain the final weighting coefficients used in our upsampling model, where w_{ij}^c and w_{ij}^d are respectively formulated as

$$\begin{aligned} w_{ij}^c &= \exp\left(-\|i-j\|^2/\sigma_s^c\right)\exp\left(-\left\|I_i - I_j\right\|^2/\sigma_c^c\right) \\ w_{ij}^d &= \exp\left(-\|i-j\|^2/\sigma_s^d\right)\exp\left(-\left\|\tilde{D}_i - \tilde{D}_j\right\|^2/\sigma_d^d\right) \end{aligned} \tag{13}$$

3) WEIGHT COEFFICIENTS FUSION

Single pseudo depth image can not accurately present the position of ground truth boundary, thus weight coefficients estimated from it are not very reliable. As a remedy, we produce a set of pseudo depth images $\{\tilde{D}_l\}$ under different parameter configurations, and synthesize more reliable weight coefficients w_{ij}^d from the weight coefficient set $\{\tilde{w}_{ij}^d\}$ estimated from the pseudo depth images $\{\tilde{D}_l\}$, where $l \in \{1 \ldots L\}$. Here, P is the pixel set of HR depth image and $w_{i,k}^d$ denotes $w_{i(i+k)}^d$. Let $k = j - i$, then $w_{i,k}^d = w_{ij}^d$. According to the index k, w_{ij}^d can be divided into different sets $w_k^d = \{w_{i,k}^d | i \in P\}$ and each w_k^d forms a image that has same size with D under the cyclic boundary condition.

We use TV model (14) to fuse the information of $\{w_k^{d_l}\}$, for each k

$$\min_{w_k^d} \beta((\partial_x w_k^d)^2 + (\partial_y w_k^d)^2) + \sum\nolimits_{l=1}^{L} (w_k^d - \tilde{w}_k^{d_l})^2 \tag{14}$$

where $\partial_x w_k^d$ and $\partial_y w_k^d$ denote weight difference between neighboring pixels along the x and y directions. Let $w_k^{d_a} = \frac{1}{L}\sum_{l=1}^{L} \tilde{w}_k^{d_l}$, formula (14) equals to formula (15).

$$\min_{w_k^d} \beta((\partial_x w_k^d)^2 + (\partial_y w_k^d)^2) + L(w_k^d - w_k^{d_a})^2 \tag{15}$$

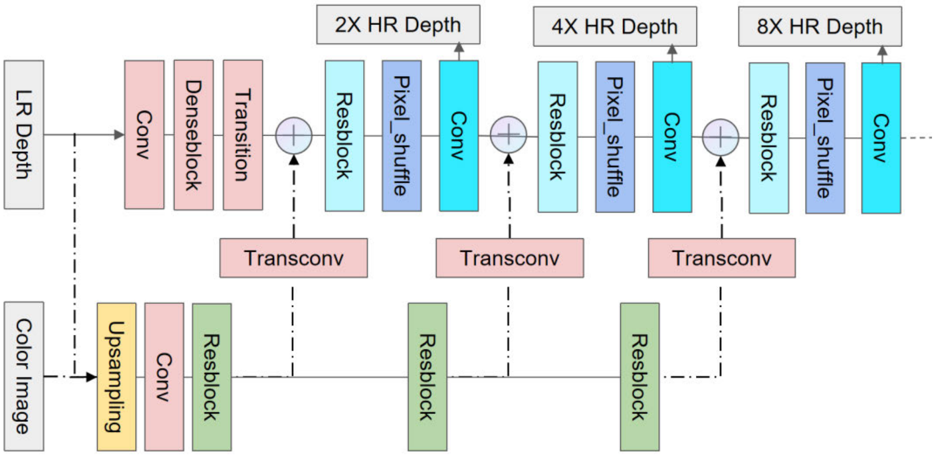


FIGURE 2. **An overview of the proposed depth CNN.**

The energy function of (15) is quadratic and thus has a global minimum. We can use formula (16) to find the optimization solution [34].

$$w_k^d = F^{-1}\left(\frac{LF(w_k^{d_a})}{F(1)+\beta((F(\partial x)^*F(\partial x))+(F(\partial y)^*F(\partial y)))}\right) \tag{16}$$

where F is the *FFT* operator and *F()** represents the complex conjugate and *F(1)* is the Fourier Transform of the delta function.

III. DEPTH CNN

The Depth CNN network proposed in this paper (as shown in Fig.2) is mainly based on "*resblock*" and "*denseblock*", where the "batch-normalization" layer are removed and Its linear conversion function is merged into the convolution layer. To reduce the channel and computation, the denseblock is cascaded a transition module and the combined pair of denseblock and transition can be implemented multiple times according to the actual situation. The depth image is sent to DepthSRNet through trunk network and optimization branch, and then feature fusion and reconstruction processing are carried out to obtain HR depth image.

The low-level feature extraction module is based on residual network structure that includes three 3*3 convolution layers and the residual jump connection is added between the last two convolution layers. The high-level feature extraction module includes a plurality of dense connection layers and an equal number of transition layers which are connected in a cascade at intervals and the connection can be presented as $y_n = h([x_{n-1}, x_{n-2}, \ldots, 1])$, where h represents convolution layer and activation function processing and [...] represents connection operation. The Pixel_shuffle layer can carry out high and wide upsampling processing on feature image that can effectively retain image details. In general, The LR depth image is sent into the trunk of DepthSRNet and the original HR depth image that is upsampled and constructed from LR depth image is sent into optimization branch as the monitoring signal for model training. Then, the loss between the HR depth image output by the neural network model and the original HR depth image is calculated, and the loss function used in training is expressed as:

$$loss = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{17}$$

where, n represents the number of samples, y_i represents the original HR depth image and $\hat{y}_i$ represents the HR depth image that output from the model training. The Adam gradient update algorithm is used in the training, and the exponential decay rate range is (0.9, 0.999).

IV. EXPERIMENTS AND COMPARISONS

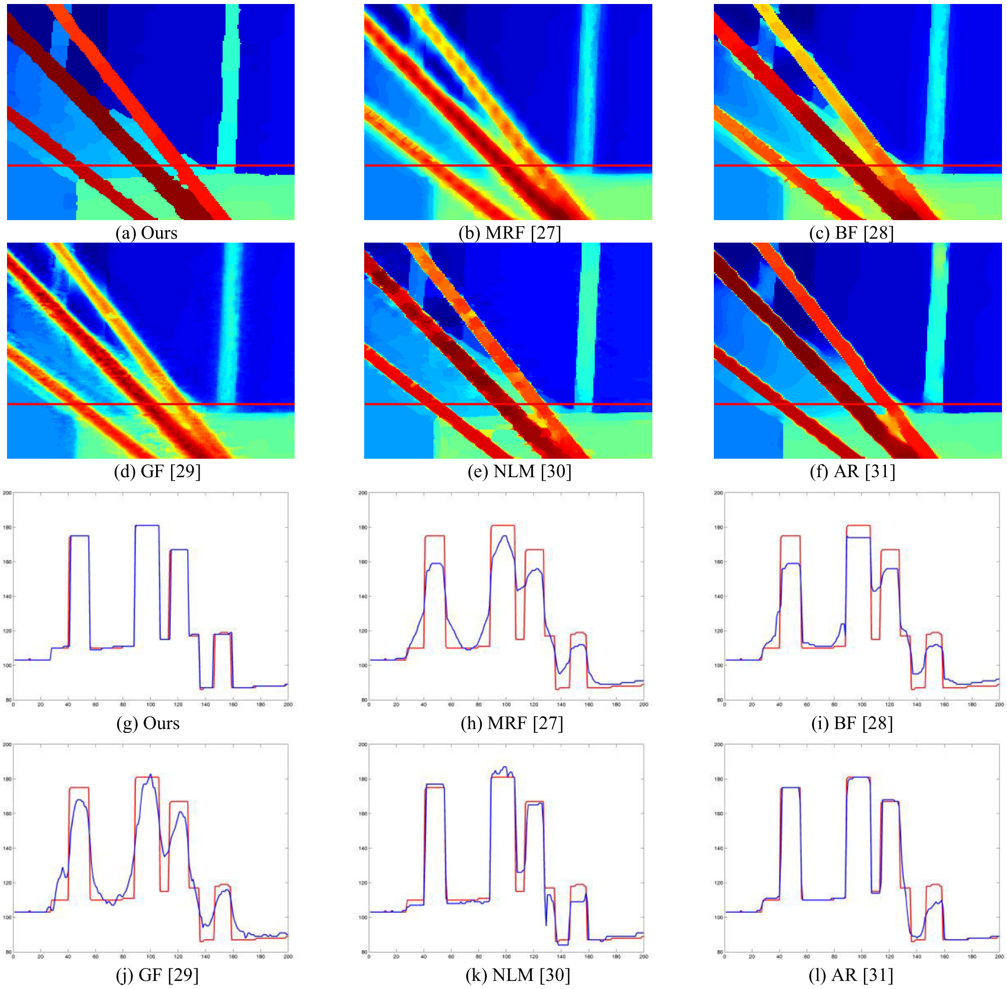
We implement our upsampling program with python 3.6 on a PC. The parameter configuration used in separation solver (1) is $\sigma_C^{\downarrow} = 50, \sigma_S^{\downarrow} = 4, \tau = 1, \lambda_1 = 70, \lambda_2 = 10$. To indicate the edges of depth image, we usually compute two pseudo depth images with parameter settings $\sigma_s = 3$, $\sigma_c = 20$ and $\sigma_s = 3$, $\sigma_c = 80$. Moreover, we find that the performance of computing correlation coefficient (10) is stable when the parameters are within certain ranges: $\beta \in [0.01, 0.2], \sigma_s^c, \sigma_s^d \in [3, 8], \sigma_c^c \in [30, 60], \sigma_c^d \in [70, 125]$. The computational complexity of our upsampling method is same to the *NLM [30]* and *AR [31]*, because the most time consuming calculation is solving the quadratic optimization. Both synthetic examples and real world examples are tested in the experiments. Then the Depth CNN is deployed on the PyTorch framework and the experiment is conducted on a PC with Intel (R) Core (TM) i7-7700HQ CPU@2.80GHz and NVIDIA GeForce GTX 1060 5GB GPU, and we compare the proposed method with the state-of-the-art depth image SR networks (*VDSR [35], TSDR [36], DepthSR-Net [37], MFR-SR [38], RYNet [39]*) in conducting qualitative and quantitative performance by various scaling factors.

A. SYNTHETIC EXAMPLES EVALUATIONS

In the experiments, we employ two depth images, Art and Book from the Middlebury's benchmark, to evaluate the

TABLE 2. MAD comparison of upsampling on middlebury dataset at 2X, 4X, 8X, 16X rates.

		Bicubic	MRF[27]	BF[28]	GF[29]	NLM[30]	AR[31]	Ours
Art	2X	0.47	0.63	0.54	0.65	0.42	0.19	0.21
	4X	0.99	1.13	0.73	1.15	0.73	0.52	0.34
	8X	1.87	1.98	1.51	1.76	1.13	0.69	0.53
	16X	3.52	3.91	3.64	3.61	2.24	2.14	1.23
Book	2X	0.14	0.21	0.32	0.25	0.19	0.12	0.09
	4X	0.28	0.32	0.46	0.34	0.33	0.26	0.15
	8X	0.57	0.61	0.63	0.60	0.58	0.49	0.31
	16X	1.17	1.24	1.43	1.14	1.02	0.83	0.58



(a) Ours (b) MRF [27] (c) BF [28]

(d) GF [29] (e) NLM [30] (f) AR [31]

(g) Ours (h) MRF [27] (i) BF [28]

(j) GF [29] (k) NLM [30] (l) AR [31]

FIGURE 3. Partial enlargements results of six methods for Art 8X upsampling.

upsampling performance. The low resolution version depth images are down-sampled from the ground truth depth images. Original color images are used as the high resolution guided images. We use MAD (mean absolute difference) criterion to evaluate the upsampling quality. The MADs against the ground truth depth images are reported in Table 2. We can observe that our method obtains the lowest MADs for 4X, 8X and 16X upsampling. For low upsampling rate such as 2X, the restoration ability on planar regions is the critical factor. The reason is that our algorithm will degenerate to the simplest linear interpolation method, when the pixel number in the planar regions overwhelms the number in the edge regions. The statistics in Table 2 also show that our method becomes more and more prominent with the upsampling factor increasing or the boundary regions growing. For visual comparison, 8X upsampling depth images of Art are shown in Figure 3. Partial enlargements are drawn in Figure 3(a)-3(f). Figure 3(g)-3(l) are corresponding section curves. We can find out that the compared methods introduce obvious jaggy artifacts along the section curve. Although the result of

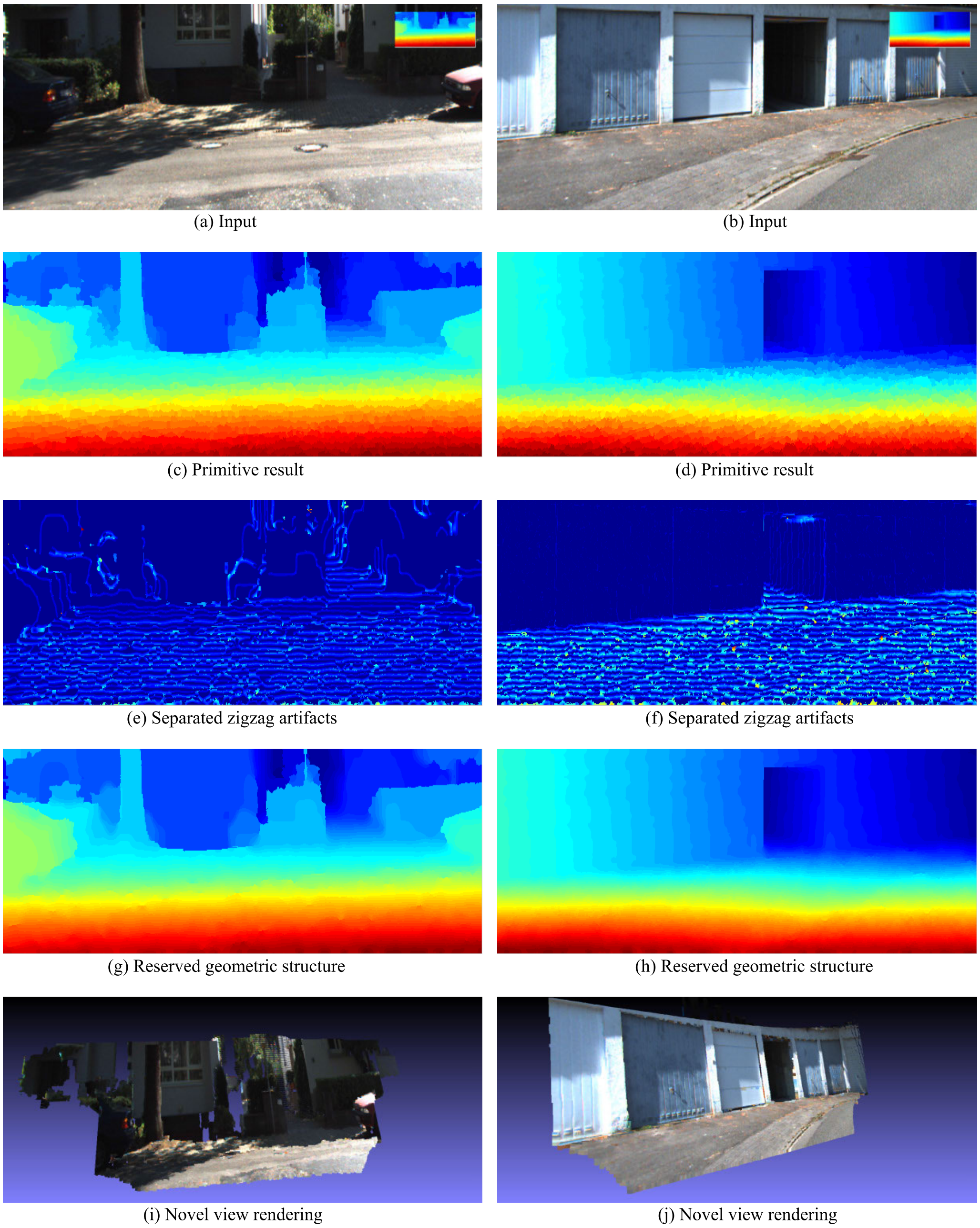

(a) Input
(b) Input
(c) Primitive result
(d) Primitive result
(e) Separated zigzag artifacts
(f) Separated zigzag artifacts
(g) Reserved geometric structure
(h) Reserved geometric structure
(i) Novel view rendering
(j) Novel view rendering

FIGURE 4. **6X upsampling results.**

AR [31] is comparable to ours, its edges are smoothed and blurred.

In the first two rows, we show the close-up of Art. In the next two rows, we illustrate the depth profile along the red section line shown in the first rows, where the red line denotes the groudtruth and the blue line is the produced depths.

B. REALWORLD EXPERIMENTS

We use the standard database of KITTI Vision Benchmark Suite to perform real-world experiments. All of the data in the benchmark is acquired by a standard station wagon with two color and grayscale video cameras. The accurate ground truth is provided by a Velodyne laser scanner. We employ the laser scanner and companied color image to perform our experiments.

It is a tough task to upscale the depth image of KITTI. All images of KITTI are captured from Karlsruhe streets, thus the scenes are very large and the geometric structure of objects is very small, compared with background scene in the image. In contrast to the artificial in-door pictures used in state-of-the-art method [30], [31], the images of KITTI surfer from varied sensing noise. Figure 4 illustrates 6X super-resolution results of our method. The zigzag artifacts are inevitable using maximum correlation principle to fulfill the missing hole,

TABLE 3. RMSE comparison on middlebury dataset at 2X, 4X, 8X, 16X rates.

		VDSR[35]	TSDR[36]	DepthSR-Net[37]	MFR-SR[38]	RYNet[39]	Ours
Art	2X	1.13	-	0.57	0.46	0.28	0.31
	4X	1.87	1.57	1.23	1.26	1.05	1.08
	8X	4.86	2.30	2.24	2.29	2.03	2.05
	16X	6.27	4.30	3.91	3.79	3.32	3.57
Book	2X	0.40	-	0.45	0.24	0.18	0.19
	4X	0.98	1.05	0.63	0.49	0.35	0.37
	8X	1.99	1.06	0.93	0.89	0.77	0.71
	16X	2.47	1.59	1.52	1.69	1.53	1.45

TABLE 4. PSNR comparison on middlebury dataset at 2X, 4X, 8X, 16X rates.

		VDSR[35]	TSDR[36]	DepthSR-Net[37]	MFR-SR[38]	RYNet[39]	Ours
Art	2X	45.32	-	53.77	55.24	60.41	58.16
	4X	43.17	44.19	46.78	46.91	48.35	47.98
	8X	36.19	40.86	41.36	41.21	42.38	41.95
	16X	31.74	35.45	36.37	36.69	37.91	37.76
Book	2X	56.08	-	55.89	61.43	64.02	63.16
	4X	50.92	47.69	52.65	54.58	56.97	56.51
	8X	40.73	47.62	49.35	49.51	51.12	51.39
	16X	37.82	44.10	44.57	43.39	44.54	45.36

since acquired depths in the practical environment vary dramatically and a lot of depth information is lost by downsampling the laser data. To deal with this situation, our structure and noise separation algorithm is used as a post processing step to smooth the final results and departs the stair jump from the structure of depth image. All of results are exhibited in Figure 4. We can observe that our separation algorithm can keep the sharp boundary of depth image Figure 4(g)-4(h) and segregate the jagged artifacts Figure 4(e)-4(f) from primitive results Figure 4(c)-4(d) simultaneously. The rendered images Figure 4(i)-4(j) suggest that our method reliably restores the geometric relationship.The depth-color pairs are shown at their original ratio of size and the image size is given in the captions.

C. DEPTH CNN EVALUATIONS

In the network evaluation phase, 100 RGB-D images are chosen from the Middlebury to form the test dataset, while 82 images for training and 18 images for validation. the root mean square error (RMSE) and peak signal to noise ratio (PSNR) are used to evaluate the super-resolution reconstruction performance of the network. The state-of-the-art depth image super-resolution methods under comparison include VDSR [35], TSDR [36], DepthSR-Net [37], MFR-SR [38], RYNet [39], and the result is list in table 3 and table 4. As we can see, the proposed Depth CNN and RYNet show suboptimal and optimal performance for most test images respectively.

V. CONCLUSION

We introduce a divide-and-conquer upsampling method to upscale the LR depth image with a registered high quality optical image. The depth image is divided into planar areas and edge regions. We exploit different methods to interpolate the missing depths of different areas. The statistic data shows the simplest linear interpolation method can produce completive upsampling results at planar areas, compared with state-of-the-art methods. For edge region upsampling, our segmentation-separation upsampling method outperforms previous methods and yields much better upsampled edge regions. Then we propose a depth CNN to refine upsampled results, and the experiments show that the effect meets the requirements. In future work, we will attempt to add constraints such as prior images to further improve network performance, and also consider other network architectures such as GANs.

REFERENCES

[1] S. Y. Zhang, M. Q. Liu, and C. Yao, "Hierarchical feature feedback network for depth super-resolution reconstruction," *Acta Automatica Sinica*, vol. 48, no. 4, pp. 992–1003, 2022.

[2] T. Chang and J. Yang, "Precise depth map upsampling and enhancement based on edge-preserving fusion filters," *IET Comput. Vis.*, vol. 12, no. 5, pp. 651–658, Aug. 2018.

[3] Y. G. Qiao, L. Jiao, W. Li, C. Richardt, and D. Cosker, "Fast, high-quality hierarchical depth-map super-resolution," in *Proc. 29th ACM Int. Conf. Multimedia*, New York, NY, USA, 2021, pp. 4444–4453.

[4] Z. Liu, H. Liu, Z. Zhu, C. Sun, and J. Song, "Three-dimensional shape reconstruction of uncooperative spacecraft with texture-guided depth upsampling," *Meas. Sci. Technol.*, vol. 32, no. 9, Sep. 2021, Art. no. 095006. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6501/abf9d5

[5] S. Yang, N. Cao, B. Guo, and G. Li, "Depth map super-resolution based on edge-guided joint trilateral upsampling," *Vis. Comput.*, vol. 3, pp. 1–13, Mar. 2021.

[6] Z. Jiang, H. Yue, Y.-K. Lai, J. Yang, Y. Hou, and C. Hou, "Deep edge map guided depth super resolution," *Signal Process., Image Commun.*, vol. 90, Jan. 2021, Art. no. 116040. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0923596520301867

[7] Y. Yang, H. S. Lee, and B. T. Oh, "Depth map upsampling with a confidence-based joint guided filter," *Signal Process., Image Commun.*, vol. 77, pp. 40–48, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S092359651830849X

[8] J. Lei, L. Li, H. Yue, F. Wu, N. Ling, and C. Hou, "Depth map super-resolution considering view synthesis quality," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1732–1745, Apr. 2017.

[9] M. Sharma, "An integrated optimization approach for depth map enhancement on special Riemannian manifold," in *Proc. 11th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2018, pp. 1–9, doi: 10.1145/3293353.3293409.

[10] Y. Chen, Z. Hong, and D. Sun, "Color-guided optimization model with reliable self-structure priors for depth map restoration," *OSA Continuum*, vol. 4, no. 7, pp. 1964–1980, 2021.

[11] X. Yan, A. Ping, Z. Shuai, Z. Yi-Fan, and S. Li-Quan, "Super-resolution reconstruction for depth map based on edge enhancement," *J. Optoelectron. Laser*, vol. 27, no. 4, pp. 437–447,2016.

[12] B. Maxim, A. Arkhipov, S. Emelyanov, and N. Milostnaya, "A method for creating a depth map based on a three-level fuzzy model," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105629.

[13] W. K. Jung and J. K. Han, "Depth map refinement using super-pixel segmentation in multi-view systems," in *Proc. ICCE*, 2021, pp. 1–5.

[14] H. Wang, M. Yang, X. Lan, C. Zhu, and N. Zheng, "Depth map recovery based on a unified depth boundary distortion model," *IEEE Trans. Image Process.*, vol. 31, pp. 7020–7035, 2022.

[15] Y. Gao, H. Li, J. Dong, and G. Feng, "A deep convolutional network for medical image super-resolution," in *Proc. Chin. Autom. Congr. (CAC)*. Zurich, Switzerland: Springer, Oct. 2017, pp. 184–199.

[16] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 136–144.

[17] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 1874–1883.

[18] Y. F. Zhang, Y. Liu, and C. Jiang, "A curriculum learning approach for single image super resolution," *Acta Automatica Sinica*, vol. 46, no. 2, pp. 274–282, 2020.

[19] D. W. Zhou, Z. Li-Juan, D. Ran, and C. Xiao-Liang, "Image super-resolution based on recursive residual networks," *Acta Automatica Sinica*, vol. 45, no. 6, pp. 1157–1165,2019.

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

[21] J.-Y. Kim, S. Ji, S.-J. Baek, S.-W. Jung, and S.-J. Ko, "Depth map super-resolution using guided deformable convolution," *IEEE Access*, vol. 9, pp. 66626–66635, 2021.

[22] J. Guo, R. Xiong, Y. Ou, L. Wang, and C. Liu, "Depth image super-resolution via two-branch network," in *Proc. ICCSIP*, 2021, pp. 200–212.

[23] G. Peyre, S. Bougleux, and L. D. Cohen, "Non-local regularization of inverse problems," in *Proc. Eur. Conf. Comput. Vis.*, vol. 5304. Cham, Switzerland: Springer, 2008, pp. 57–68.

[24] L. Condat, "Semi-local total variation for regularization of inverse problems," in *Proc. EUSIPCO*, 2014, pp. 1806–1810.

[25] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[26] Y. J. Chu and C. M. Mak, "A new QR decomposition-based RLS algorithm using the split Bregman method for L_1-regularized problems," *Signal Process.*, vol. 128, pp. 303–308, Nov. 2016.

[27] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. ANIPS*, 2005, pp. 291–298.

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

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

[30] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1623–1630.

[31] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *Proc. ECCV*. Berlin, Heidelberg: Springer, 2012, pp. 158–171.

[32] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[33] X. Dong, J. Shen, L. Shao, and L. Van Gool, "Sub-Markov random walk for image segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 516–527, Feb. 2016.

[34] X. Li, C. Lu, Y. Xu, and J. Jia, "Image smoothing via L_0 gradient minimization," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–12, 2011.

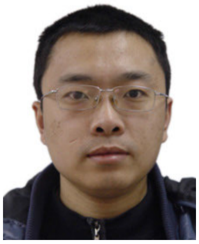
[35] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[36] Z. Jiang, Y. Hou, H. Yue, J. Yang, and C. Hou, "Depth super-resolution from RGB-D pairs with transform and spatial domain regularization," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2587–2602, May 2018.

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

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

[39] T. Li, X. Dong, and H. Lin, "Guided depth map super-resolution using recumbent y network," *IEEE Access*, vol. 8, pp. 122695–122708, 2020.



YANMING YE received the M.S. and Ph.D. degrees in computer science and technology from Zhejiang University, Hangzhou, China, in 2005 and 2014, respectively. He is currently an Associate Professor with the School of Information Engineering, Hangzhou Dianzi University. His current research interests include artificial intelligence, computer vision, and service computing.

MENGXIONG ZHOU was born in Fenghua, Zhejiang, China, in 1981. He received the M.S. degree in engineering from the Harbin University of Science and Technology, Harbin, China, in 2007. After graduation, he was with the Department of Computer Science, College of Information Engineering, Hangzhou Dianzi University, Hangzhou, China. His research interests include software engineering, artificial intelligence, and graphics.

ZHANYU WANG was born in Harbin, China, in 1982. He received the B.S. and M.S. degrees in engineering from the Harbin University of Science and Technology, Harbin, and the Ph.D. degree in engineering from the Harbin Institute of Technology, Harbin, in 2018. Since 2018, he has been with the Changshu Institute of Technology. His research interests include intelligent computing, computer vision, artificial intelligence, and networked control systems.

XINGFA SHEN (Member, IEEE) received the B.S. degree in electrical engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2000 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University. His current research interests include CPS, the Internet of Things, and mobile computing.