Deep Color Guided Coarse-to-Fine Convolutional Network Cascade for Depth Image Super-Resolution

Yang Wen, Bin Sheng, Ping Li, Weiyao Lin, and David Dagan Feng, *Fellow, IEEE*

***Abstract*—Depth image super-resolution is a significant yet challenging task. In this paper, we introduce a novel deep color guided coarse-to-fine convolutional neural network (CNN) framework to address this problem. First, we present a data-driven filter method to approximate the *ideal* filter for depth image super-resolution instead of hand-designed filters. Based on large data samples, the filter learned is more accurate and stable for upsampling depth image. Second, we introduce a coarse-to-fine CNN to learn different sizes of filter kernels. In coarse stage, larger filter kernels are learned by CNN to achieve crude high-resolution depth image. As to fine stage, the crude high-resolution depth image is used as the input so that smaller filter kernels are learned to gain more accurate results. Benefit from this network, we can progressively recover the high frequency details. Third, we construct a color guidance strategy that fuses color difference and spatial distance for depth image upsampling. We revise the interpolated high-resolution depth image according to the corresponding pixels in high-resolution color maps. Guided by color information, the depth of high-resolution image obtained can alleviate texture copying artifacts and preserve edge details effectively. Quantitative and qualitative experimental results demonstrate our state-of-the-art performance for depth map super-resolution.**

***Index Terms*—Depth super-resolution, color guidance, coarse-to-fine convolutional neural network, filter kernel learning.**

I. Introduction

IMAGE super-resolution (SR) refers to construct a potential high-resolution (HR) image from a low-resolution (LR) depth image, and has been widely applied in medical images [1], [2], digital image enhancement [3], [4], video surveillance [5], [6], etc. Since deep neural networks have been demonstrated very effective for many computer vision tasks by extracting useful semantic information from abundant data, diverse deep neural networks based SR methods have been developed. Typically, Super-Resolution Convolutional Neural Network (SRCNN) [7], Super-Resolution using Very Deep convolutional networks (VDSR) [8], and Deep Edge Guided REcurrent rEsidual (DEGREE) [9] algorithms use deep CNN to learn the end-to-end mapping between the low/high-resolution images for color image super-resolution. With the rapid development in the 3D imaging fields, reliable depth information generated by the consumer 3D scanning devices has captured the attention of researchers in various applications, e.g., interactive free-viewpoint video [10], 3D reconstruction [11], semantic scene analysis [12], [13], and human pose recognition [14], [15]. It mainly contains two major classes of depth measuring methods to obtain depth information, which are passive and active sensors [16], [17].

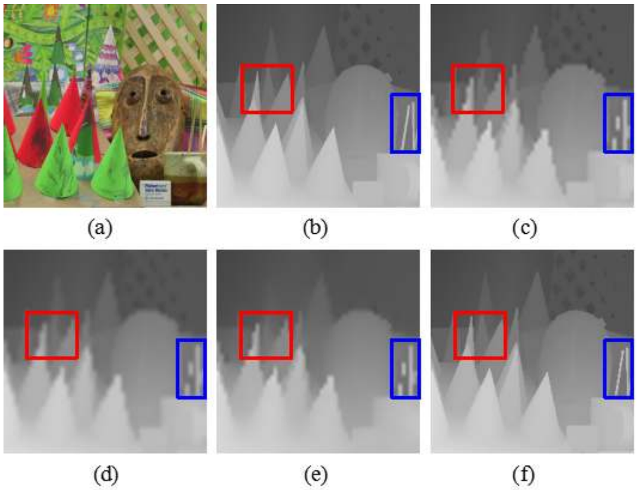


Fig. 1: Ambiguity and discontinuities in upsampling depth map. (a) Color image. (b) Ground-truth. (c) (Enlarged) LR depth map downsampled by a factor of 8. (d) Guided image Filtering (GF) [18]. (e) SRCNN [7]. (f) Our proposed method without edge ambiguity and discontinuities.

For passive sensors, the most famous and widely used stereo matching methods [19], [20] are always time consuming and imprecise in textureless or occluded regions instead of active sensors which can produce more accurate results. Nowadays, the two popular active sensors are Laser scanners and Time-of-Flight (ToF) sensors. Even though Laser scanners can generate the depth map with high-quality, they can only measure a single point at a time and their applications are subject to environmental restrictions. Compared with Laser scanners, ToF sensors are cheaper and have the function of capturing the depth maps of the fast-moving objects. Although these advantages make ToF sensors get more attention, they are still limited in resolution and random noise. For instance, the resolution of MESA SR 4000 and PMD CamCube 3.0 are only 176×144 and 200×200, respectively [16]. To promote the use of depth information and meet the actual needs, Depth image Super-Resolution (DSR) is proposed to manufacture a visually satisfactory high-resolution depth image from a low-resolution depth input based on traditional super-resolution methods.

Most of the recent DSR methods propose to utilize an additional aligned high-resolution color image of the same scene

Manuscript received July 15, 2018; revised September 08, 2018.

Y. Wen, B. Sheng, and W. Lin are with the School of Electronics, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (Email: shengbin@sjtu.edu.cn).

P. Li is with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China (Email: pli@must.edu.mo).

D. D. Feng is with the Biomedical and Multimedia Information Technology Research Group, School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia (Email: dagan.feng@sydney.edu.au).

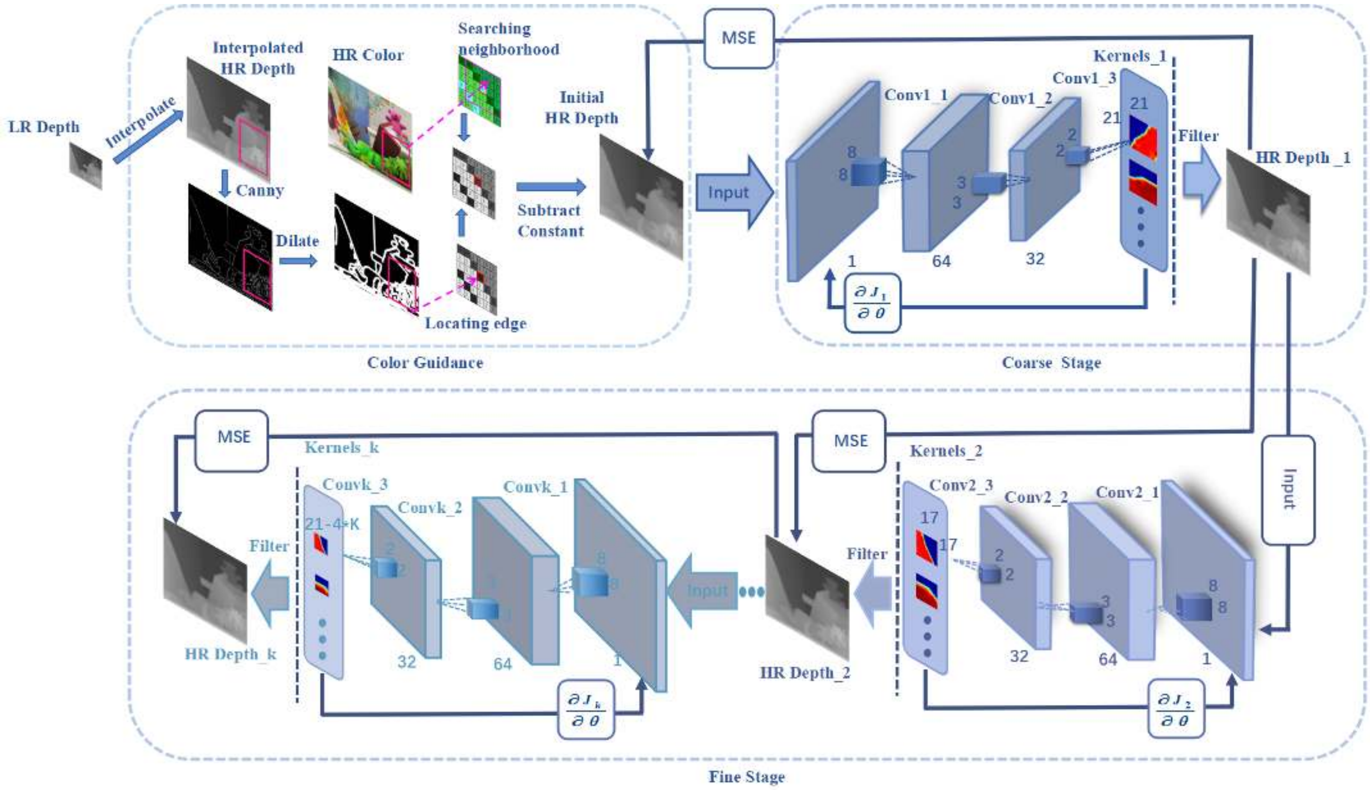


Fig. 2: Conceptual illustration of our framework. It mainly includes three parts: color guidance, coarse stage, and fine stage. In the color guidance part, we first interpolate the LR depth image by bicubic algorithm to obtain the initial HR depth image, then we detect the edges of the interpolated high-resolution depth map by canny operator and dilate the edges by five pixels to determine the marginal areas. To calculate the approximate depth value, both the neighborhood information determined by HR color image and the depth information of LR depth image are utilized. Finally, we subtract the approximate depth from the marginal area of the interpolated high-resolution image. In the coarse stage, the modified interpolated HR depth map is used as the input of CNN to generate the approximate ideal filter, which is named Kernels_1 in the figure. Then the coarse HR depth_1 map can be constructed by filter_1. In order to recover better details, the coarse HR depth_1 is used as input for the fine CNN. Through the fine CNN, smaller kernels are obtained to consider smaller neighborhood and recover better resolution HR depth_k. MSE is used to judge the convergence of network, the network can be considered convergent if MSE is small enough.

captured by the RGB-D cameras as assistance [18], [21]–[27]. For example, in [28], a manually-designed edge-preserving filtering kernel is applied to propagate the local color information to the depth image with the assumption that the color edges aligned well with the depth edges. However, artifacts including edge blurring and texture copying may occur when the assumption is violated. First, if smooth regions having rich color texture, texture information would be transferred to the upsampled object surface. Second, if color edges do not align well with the depth edges, the upsampled object boundary would be ambiguous via inappropriate color engagement. In fact, the optimal guidance is indeed the ground-truth high-resolution depth image. However, due to the inevitable noise of the equipment, the theoretical precision of the optimal guidance cannot be achieved in practical applications.

Motivated by the fact that color information can provide valuable guidance for image super-resolution and the success of deep CNN models in modeling complex image details, we present a novel deep color guided coarse-to-fine CNN to learn *ideal* filter for depth map SR. For a LR depth map, we employ bicubic interpolation to estimate its initial HR version firstly. Since pixels in depth discontinuity regions are regarded as uncertain ones, we modify them according to the pixels of the corresponding positions in HR color image so that both color and depth images are considered. Then, we learn an edge-preserving filter kernel by deep CNN on an external database to upsample the low-resolution depth images instead of traditional handicrafts. The use of deep CNN can better extract the mutual information between the LR depth image and the HR color image. This is the reason why the proposed data-driven filters can better approximate the ground-truth high-resolution depth image. For example, Fig. 1 shows several popular edge-preserving filter based upsampling methods. Fig. 1(a) and Fig. 1(b) present the high-resolution color image and the ground-truth map respectively. Fig. 1(c)–Fig. 1(e) are HR depth maps obtained by other filter methods, Fig. 1(f) is the result of our proposed method. Fig. 1 demonstrates that our performance is out-performed and can effectively reduce the texture copying effect and edge discontinuity effect (especially in color regions). Our work makes the following three main contributions:

- *Significant Color Guidance:* We integrate both HR image's color prior [29] and LR image's depth prior to guide low-resolution depth map upsampling. Since the edge information of the marginal areas in the interpolated HR depth image is always inaccurate, we first use canny operator to detect and expand the edge of the interpolated HR depth image to obtain the marginal areas, and then obtain the approximate depth value by HR's color information and the corresponding LR's depth information. The color guidance can improve the quality of the interpolated depth image and optimize the learned filter by combining HR' color information and LR' depth information. It can also accelerate the network convergence by subtracting the approximate depth value from the interpolated depth image based on the idea of residual learning [8], [9].

- *Efficient Deep Cascade Structure:* We introduce a deep coarse-to-fine network cascade model to solve depth image super-resolution problems. In the coarse stage, a convolutional neural network is designed to obtain larger filter kernels. Benefiting from the network, the data-driven filter learned improves the quality of depth image SR significantly. In order to achieve more details, we learn a series of smaller kernels in the fine stage to reduce filtering range. Combining two stages can give a recovered HR depth image with high-quality and sharp high-frequency details.
- *Especial Data-Driven Upsample Filter:* We propose the concept of *ideal* filtering and design an edge-preserving filter via deep convolutional neural networks for depth map upsampling. The filter we learned can obviously avoid complicated artificial designs and approximate the *ideal* filter effectively. Experiments show that it achieves better performance compared to state-of-the-art methods.

II. Related Work

A. Single Image Super-Resolution

Image super-resolution is one of the most active research topics in computer vision. Generally, there are mainly two different types of approaches for image-super resolution. (1) *Single image super-resolution methods* generally rely on image priors to generate a HR image. Although the calculation is simple, they cannot restore the high-frequency details very well. For instance, Yan *et al.* [30] employed gradient profile sharpness to realize SR. (2) *Image super-resolution with external data* methods usually learn dictionaries, regression functions or end-to-end mapping between the HR images and their down-sampled LR versions. In [31]–[34], sparse representation methods learned a couple dictionary to represent LR image patches. Deep neural networks based methods deal with SR problems in various ways. Moreover, Dong *et al.* [7] directly learned an end-to-end mapping between the low/high-resolution images using deep CNN. Wang *et al.* [35] incorporated the sparse prior into CNN by exploiting a learned iterative shrinkage and thresholding algorithm. Though these methods have a significant effect on single image super resolution, they are not very suitable for depth maps because they will still generate blurry artifacts at the edges and cannot deal with depth image super-resolution problems very well.

B. Depth Image Super-Resolution

Depth map SR methods can be grossly divided into learning based and filtering based methods. For learning-based ones, Diebel and Thrun [36] conjugated a Markov Random Field (MRF) and gradient method to upsample LR depth map. Ferstl *et al.* [37] considered depth map upsampling as a convex optimization issue with higher order regularization. It is demonstrated that an additional high-resolution color image is very useful for depth SR. However, Learning-based methods are often limited in application because of their high computational complexity. For filtering-based ones, Guided Filter (GF) [18] was used as an edge-preserving smoothing operator like the popular Bilateral Filter (BF) [22] which calculated the edge-smoothing output via both the spatial and intensity domain information. Although GF [18] could keep the edges well and compute easily, it also suffered from halo artifacts sometimes. Gradient domain guided filter [38], [39] could keep edge better by adding an explicit first order edge sensing constraint. Joint Bilateral Filter (JBF) [29] employed an additional guidance to improve the quality of the input target image taken from a dark or noisy environment. Hua *et al.* [40] approximately applied the filtering procedure with local gradient information of the depth image with the guidance of a HR color image. Yang *et al.* [24] employed edge-preserving filters like JBF to upsample a depth image with an additional color image. These methods are based on the assumption that local pixels with similar color will have similar depth value. However, sometimes this assumption is unfounded: (i) texture copying artifact may occur in textured color and textureless depth; (ii) blurry edges will occur on textureless color and textured depth or when the color and depth edges are not aligned well. Chan *et al.* [21] proposed a noise-aware filter and use the input depth values as guidance in geometrically smooth region and color image as guidance in depth discontinuities. It can suppress texture copying but is still suffering from blurry edge.

Recently, color guidance pre-processing aims to employ a pre-aligned high-resolution color image to guide the low-resolution depth map upsampling. For instance, image guided depth upsampling using anisotropic Total Generalized Variation (TGV) [37] and high-quality depth map upsampling for 3D-ToF cameras Non-Local Means (NLM) [41] are also very classical color assisted depth image super-resolution approaches. Anisotropic Total Generalized Variation Network (ATGV-Net) [42] modelled the piecewise affine structures apparent by a variational method. Song *et al.* [43] used both the statistics of the depth field and the local correlation between the color map and the depth map. In [44], Hui *et al.* proposed a Multi-Scale Guided convolutional Network (MSG-Net) for depth map super resolution. Since color information is vital to depth image SR for providing the edge guidance, we have applied the high-resolution color image as supporting information so that pixels with different depth can be weighted differently according to the color value during the upsampling process. In view of the excellent performance of CNN on depth map upsampling, it is now gradually combining color image to solve DSR problems. Since traditional filter based DSR method cannot recover high frequency details effectively, here we use the cascade CNN network and additional HR color images to train the *ideal* upsampling filter.

III. Proposed Method

In this paper, we propose a meaningful framework to deal with the low-resolution depth image upsampling issue. Fig.

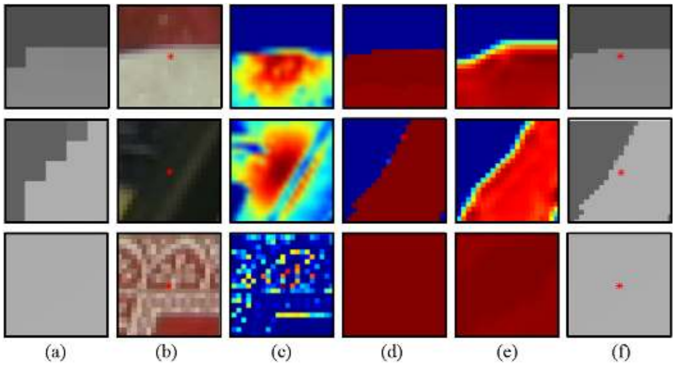


Fig. 3: Filter kernels. The first row presents the database without well-aligned color or depth edges. (a), (b) and (f) are the input depth image, color image, and ground truth depth image, respectively. In (c), the Joint Bilateral Filter (JBF) (with the color as the guidance) can well approximate the *ideal* filter kernel (*Refer to Section III-A for details*); In (d), kernels computed with the guidance of the ground-truth depth in (f). The color represents the filter weights. Red color corresponds to large weights while blue represents small value. The last two rows present two databases when the color edges are different from the depth edges, it can be seen that the joint bilateral filter kernel will be quite different from (d). While our proposed method can learn a much better approximation of the *ideal* filter kernel as shown in (e).

2 shows the proposed architecture. It mainly includes three components: color guidance pre-processing, coarse stage for DSR, fine stage for DSR. In the color guidance part of Fig. 2, an input low-resolution depth map is interpolated to be an initial high-resolution version firstly. Then, since the high frequency components of color images such as edges are propitious to assist the depth pixel prediction, we modify the uncertain interpolated depth values in marginal areas to solve the edge discontinuity and texture copying problems. Firstly, we obtain the edge by Canny operator and then dilated it by 5 pixels. After that, local real depth constants are computed according to the minimum difference between the central pixel and neighboring pixels in color image. At last, the local real depth constants are subtracted from the initial high-resolution depth version based on the residual learning idea. For coarse stage, since learning-based methods are very powerful to generate plausible details from external database meanwhile suppress artifacts, we learn larger filter kernels by a three layers CNN model, and then the kernels are used to reconstruct the high-resolution depth image named HR Depth_1. As shown in Fig. 2, the fine stage is prone to generate more fine details based on the HR Depth_1 for the reason that more shape edges can be preserved by smaller local area. Since HR Depth_1 has higher resolution than the original version, if as fine network input, the higher depth map can be achieved via the fine network, and at the same time, the MSE between different HR depth images can be used to judge whether the network converges or not.

In this section, we first briefly introduce edge-preserving filters methods for depth image super-resolution in Section III-A, and then we present the proposed color guided CNN-based DSR in Section III-B, and illustrated how to build deep coarse-to-fine convolutional network cascade for depth image SR in Section III-C. Finally, the details of the proposed network model are discussed in Section III-D.

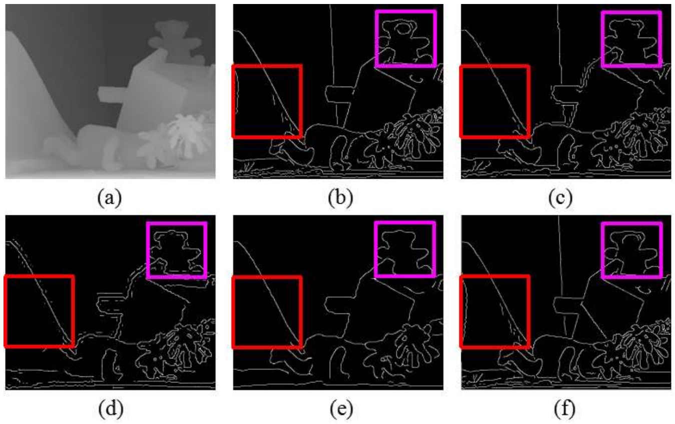


Fig. 4: Constructed edge maps with an upscale factor of 2. (a) and (b) are the ground truth depth image with its edge map. (c) Edge map of bicubic upsampled depth. (d) Edge map of bicubic upsampled depth after using guided filter [18]. (e) Edge map of bicubic upsampled depth after using gradient domain guided filter [38]. (f) Edge map of our method.

A. Filtering-Based Depth Super Resolution

According to the Subsection II-B, it has been demonstrated that filter-based methods with the guidance of an additional high-resolution color image have a remarkable performance for upsampling a low-resolution depth image. In this paper, we introduce a concept of the *ideal* filter for upsampling and design an approximate *ideal* filter to realize the depth map super-resolution. Universally, the filtering-based DSR can be formulated as follows:

$$D_p = \sum_{q_\downarrow \in \Omega_p} (w_{p,q} L_{q_\downarrow}) / \sum_{q_\downarrow \in \Omega} w_{p,q} \qquad (1)$$

where D and L denote the upsampled depth image and the input low-resolution depth image respectively, q denotes the coordinate of pixels in image D, $q_\downarrow$ denotes the corresponding (possibly fractional) coordinates of q in the image L. $w_{p,q}$ represents the edge-preserving filter kernel Ω_p centered at pixel p. If G denotes the guidance image, then the joint bilateral filter kernel can be described as:

$$w_{p,q} = exp(-\frac{\parallel p, q \parallel}{2\sigma_S^2}) exp(-\frac{\parallel G_p, G_q \parallel}{2\sigma_R^2}) \qquad (2)$$

where σ_S and σ_R are two constants to adjust the spatial similarity and range (intensity/color) similarity. $\parallel , \parallel$ denotes the distance between two constants. Fig. 3(c) presents the joint bilateral filter kernel computed with high-resolution color image in Fig. 3(b) when σ_S = 10 and σ_R = 10.

Since the joint bilateral filter has limitations when the color edge is not consistent with the depth edge, as shown in Fig. 3(c). We introduce the *ideal* filter kernel shown in Fig. 3(d), it is defined as that when the ground-truth high-resolution depth image is used as guidance. Fig. 3(d) presents the corresponding filter kernel when $\sigma_S = +\infty$ and $\sigma_R = 5$. Theoretically, σ_S/σ_R should be infinitely large/small to maximize/minimize the contribution from correct/incorrect depth seeds. However, σ_R is set to a relatively small value to suppress inevitable depth noise in practice and σ_S is set to infinitely large to ignore the spatial similarity. Recently, the deep convolutional neural network has been demonstrated to be very effective for extracting useful features with better performance than most manually-designed features. Inspired by this, we aim to use deep CNN to learn the *ideal* filter kernel for DSR. Especially, the end-to-end mapping relationship between $patch_p^H$ and its corresponding filter kernel w_p based on deep CNN can be directly described as:

$$w_p = f_{CNN}(patch_p^H) \tag{3}$$

where H denotes the bicubically-upsampled version of input low-resolution depth image L, while $patch_p^H$ denotes a p-centered block of image H. The above formulation is a direct application of CNN for filtering based DSR. It sounds like a natural solution. However, its performance is far from expectations because that depth images are normally noisy without color information.

B. Color Guidance with High-Resolution Color Image

To solve the problem of lack of color guidance, as discussed in Sec III-A, we use an additional HR color image as guidance for pre-processing and approximating the *ideal* filter. The color guidance process mainly consists of five steps. First, we obtain the initial HR depth image by bicubic interpolation method from LR depth image. Second, we mainly detect the edge of the initial HR depth image by canny operator. Third, we inflate the edges detected by the canny to determine the marginal areas. Fourth, for a neighborhood centered on a pixel p in the marginal area of the initial HR depth image, we find a set of pixels with similar colors in the corresponding region of the HR color image and determine the location of the pixel q that has the most similar color with the pixel p. Finally, we find the pixels corresponding to the position of pixel q in the LR depth image, and d_p is the depth value of the pixel p.

Let $patch_p^G$ denotes a local patch of the high-resolution guidance image G (e.g., the color image) centered at pixel p, one simple solution is feeding both $patch_p^H$ and $patch_p^G$ to the networks, and it can be simply described as:

$$w_p = f_{CNN}(patch_p^H, patch_p^G) \tag{4}$$

In order to reduce the amount of computation with the help of color image information, we draw on the idea of residual network [8], [45]. we modify the uncertain interpolated depth value by subtracting a value represented by d_p from $patch_p^H$. Theoretically, d_p should be the real depth value of the center pixel p in $patch_p^H$, so the guided patch can be described as:

$$Gpatch_p^H = patch_p^H - d_p \tag{5}$$

where $patch_p^H$ denotes the interpolated high-resolution depth image centered at pixel p, d_p denotes the approximately real depth value obtained by HR color image and LR depth image, $Gpatch_p^H$ denotes the interpolated depth image patch guided by HR color image. However, the ground-truth depth image is not available in practice and thus an approximation of d_p is proposed in this paper.

Let $\widetilde{L}_p$ denotes a candidate set and the $p_\downarrow$ denotes the corresponding coordinate of pixel p in the low-resolution image L. $\widetilde{L}_p$ is filled by the pixels around $p_\downarrow$ in L:

$$q_\downarrow \in \widetilde{L}_p\textbf{, if } \parallel q_\downarrow, p_\downarrow \parallel \leq 2 \tag{6}$$

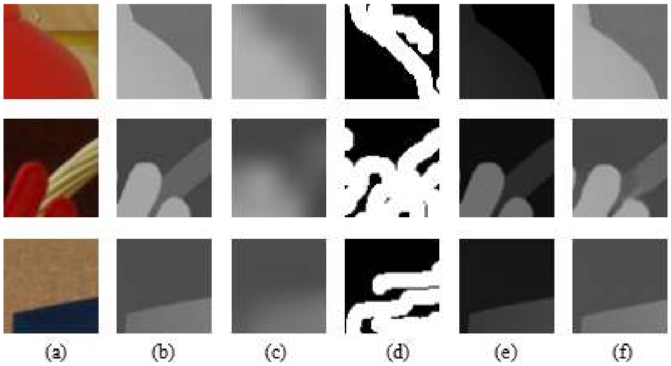


Fig. 5: Illustration of color guidance procedure. (a) and (b) are the ground truth high-resolution color image and low-resolution depth image. (c) The initial high-resolution depth map after bicubic interpolation with 16× upsampling factor. (d) Expanded edge map of the initial high-resolution depth image. (e) The guided initial high-resolution depth image. (f) Super-resolution depth image of our method.

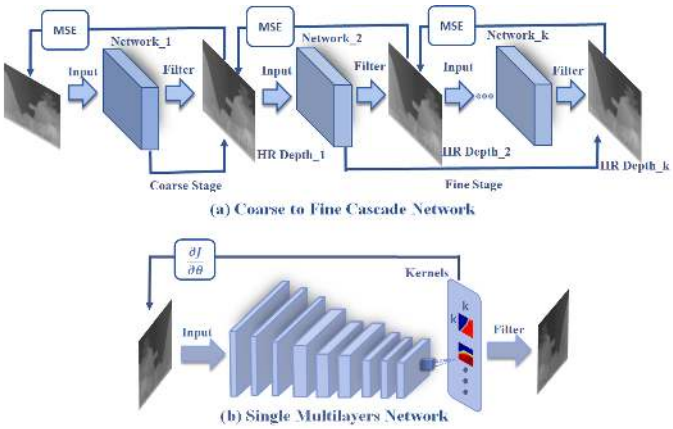
Then the color differences between these candidates are used to find the best approximation of d_p:

$$d_p = L_{\underset{i\downarrow \in \widetilde{L_p}}{arg\,\min}|G_i - G_p|} \tag{7}$$

where $i_\downarrow$ is the corresponding coordinate of pixel i in the low-resolution image L and G is the original high-resolution color image. As shown in Fig. 4, especially in the red box, we can see that edges extracted from the depth map interpolated by bicubic contains obvious jagged edges, while the edge maps of depth images after using guided filter and gradient domain guided filter [38], [39] cannot be recognized. It shows our result is most similar to the ground graph relatively. Nevertheless, Section IV also demonstrates that this simple integration outperforms the current state-of-the-art DSR methods. The guiding depth modification process is shown in detail in the Fig. 5. As shown in the Fig. 5, Fig. 5(a) and Fig. 5(b) are the ground truth high-resolution color image and low-resolution depth image, respectively. Fig. 5(c) shows the result of the initial high-resolution depth map after bicubic interpolation with 16× upsampling factor. Fig. 5(d) is the expanded edge map of the initial high-resolution depth image. Fig. 5(e) is the guided initial high-resolution depth image after subtracting the approximate depth value. Fig. 5(f) shows the super-resolution depth image of our method. From Fig. 5, we can clearly see that the proposed color guidance method effectively protects the edge information of the depth map.

C. Deep Coarse-to-Fine Cascade Architecture

As discussed above, with the color guidance, approximate *ideal* filter kernels can be learned via a convolutional neural network to reconstruct the HR depth image. However, the filter size may have a relationship with the quality of reconstructed HR image. For example, some depth values may only relate to the values in very small neighborhoods, and large kernels may affect the upsampling results sometimes. Besides, it is generally believed that the more convolution layers, the more accurate the reconstruction results are. Based on this assumption, we consider using deep coarse-to-fine network to



(a) Coarse to Fine Cascade Network

(b) Single Multilayers Network

Fig. 6: Compare cascade network and single deep network. (a) is our proposed cascade network, (b) is the single deep network with the same layers as (a).

increase the layer of the network and change the kernel size for depth map upsampling.

The architecture of our proposed network mainly includes coarse stage and fine stage. In both stages, the convolutional neural networks are composed of three convolutional layers, each layer followed by an element-wise activation function layer applies a linear convolution to its input. Every convolutional layer has a filter bank W of size $s_1 \times s_2 \times s_3 \times s_4$ and an s_4-dimensional biases vector B, where s_1 is the number of its input feature maps, s_4 is the number of its output feature maps, $s_2 \times s_3$ is the spatial support of the convolutional kernel. Intuitively, the convolutional layer applies s_4 convolutions on image. Each convolution has a kernel size of $s_1 \times s_2 \times s_3$, and each element of B is associated with a convolutional kernel (which is an output feature map). The Rectified Linear Unit (ReLU) is used as the activation function [46] so that it can converge much faster while maintain high quality [47], [48]. We refer to a convolutional layer and its following activation layer as a block, and our single CNN has three blocks. The i-th block can be expressed as a function f_i:

$$f_i(x_i) = x_{i+1} = \max(W_i * x_i + B_i, 0) \tag{8}$$

where x_i is the output of block $i-1$ and input of block i, W_i and B_i is the learned convolutional kernel and the biases vector of block i. Finally, the networks can be expressed as:

$$\begin{aligned} w_p &= f_{CNN}(Gpatch_p^H) \\ &= f_{CNN}(patch_p^H - d_p) \\ &= f_3(f_2(f_1(patch_p^H - d_p))) \end{aligned} \tag{9}$$

The mapping function f_{CNN} is represented by parameters $\theta = \{W_1, B_1, W_2, B_2, W_3, B_3\}$ which are learned by minimizing the loss between the output w_p of the network F_i and the *ideal* filter kernel w_p^{GT}. The Mean Squared Error (MSE) is used as the cost function:

$$\begin{aligned} J(\theta) &= \frac{1}{2n}\sum_p \left\| w_p - w_p^{GT} \right\|^2 \\ &= \frac{1}{2n}\sum_p \left\| f_{CNN}(patch_p^H - d_p) - w_p^{GT} \right\|^2 \end{aligned} \tag{10}$$

where, n is the number of training patches.

Algorithm 1 Generating the Training Data

```
Input: Low-resolution depth image $L$, high-resolution color image $G$
Output: Filter Kernel $w_p$
Interpolate image $L$ to achieved initial high-resolution depth image $H$;
Obtain the edge of image $H$ by "Canny" descriptor and dilate the edge by five pixels to achieve the marginal areas;
for each pixel in marginal area of $H$ do
    Extract each patch $patch_p^H$ centered at $p$ from the marginal area of image $H$;
    Compute approximate depth $d_p$ via Eq. (6) & Eq. (7);
    Subtract $d_p$ from each pixel $p$ of $patch_p^H$;
    For each point $q$ that has the distance less than 10 from the point $p$ in the selected patch, obtain the weight of pixel $p$ regarded as the ground truth kernel $w_p^{GT}$;
end for
Combine each patch pair{$patch_p^H$, $w_p^{GT}$ } at a selected pixel position from the external dataset;
```

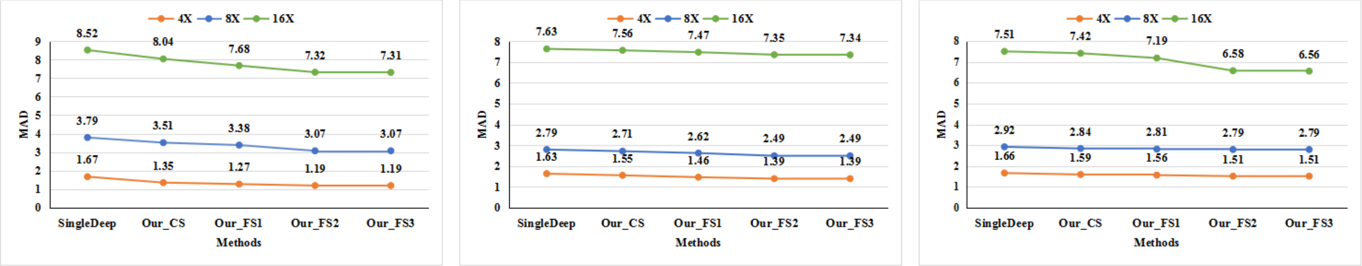
In this paper, we first use a CNN network to learn some larger kernels to approximate the *ideal* filter kernels. With the help of larger kernels, high-resolution depth version is obtained. Since some depth values in the marginal area only have relationship with small local neighborhood pixels, we learn smaller filter kernels through the network. Besides, since the input of the fine CNN is the better resolution depth map, the reconstructed high-resolution depth image will have better quality than the input one. With the help of the different small size filter kernels, the optimal SR results can be obtained when the MSE between the input and the output is small enough.

$$\begin{aligned} J(\theta_k) &= \frac{1}{2n}\sum_p \left\| w_p^k - w_p^{GT} \right\|^2 \\ &= \frac{1}{2n}\sum_p \left\| f_{CNN_k}(patch_p^{H_k} - d_p^k) - w_p^{GT} \right\|^2 \end{aligned} \tag{11}$$

where k is the index of the neural networks, w_p^k is the k^{th} filter obtained, $patch_p^{H_k}$ and d_p^k denote the patch and real depth constant of the k^{th} high-resolution depth image respectively. As shown in Fig. 6, each network can obtain one size kernel to construct corresponding HR depth map. Since the convolution operator, the kernels learned are smaller and smaller while the input image of each network is better than that of the previous network. However, for the single deep networks with almost the same layers, only single size kernels can be obtained from Fig. 6(b). Besides, the model of Fig. 6(b) is more complex to be trained than Fig. 6(a). Experimental results in Fig. 7 also demonstrate that the network architecture is more effective.

D. Implementation Details

In the training stage of color guidance part, a patch pair $\{patch_p^H, w_p^{GT}\}$ will be extracted at a selected pixel position, and all the training patch pairs selected are around depth edges. Depth edges of the ground-truth depth images are obtained from Canny edge detector. They are dilated by 5 pixels to



(a) (b) (c)

Fig. 7: Quantitative comparisons on the Middlebury dataset 2005 in *MAD*. (a), (b) and (c) are the Book scene, Laundry scene and Reindeer scene seperately. Three upsampling factors 4×, 8× and 16× are marked in orange, blue and green, respectively. SingleDeep is the result of the single deep neural network (see Fig. 6) for DSR. Our_CS is the result of our coarse stage, while Our_FS1, Our_FS2 and Our_FS3 are the results of fine stages. All three scenes show that Our results are better than that of SingleDeep network, and the MAD in our fine stages are also smaller than that of coarse stage.

locate the depth discontinuities. The training patch pairs are then extracted from a pixel p only when p is inside these regions with a stride of 6 pixels. Finally, over 40,000 training patch pairs are extracted, and three CNNs corresponding to three different upsampling factors (4×, 8×, 16×) are trained.

For the coarse stage, the size of $patch_p^H$ is set to 31×31. The size of the filter bank of the three convolutional layers are $1 \times 8 \times 8 \times 64$, $64 \times 3 \times 3 \times 32$, and $32 \times 2 \times 2 \times 1$ respectively. Due to the convolutional operation, the output feature map of a block will be smaller than the input feature map. According to the filter spatial support of the three blocks, the size of the output filter kernel w_p will be 21×21. As a result, the size of the ground-truth filter kernel will be also 21×21. The cost function is minimized using Stochastic Gradient Descent (SGD) and the parameters $\theta = \{W_1, B_1, W_2, B_2, W_3, B_3\}$ are updated at step t as follows:

$$\theta^t = \theta^{t-1} - r\frac{\partial J(\theta)}{\partial \theta} \tag{12}$$

where r is the learning rate. We set the learning rate to 0.00001 without decay in our training. The weights in each convolutional layer ($\{W_1, W_2, W_3\}$) are initialized from a zero-mean Gaussian distribution with standard deviation 0.01. The biases ($\{B_1, B_2, B_3\}$) are initialized with constant 0. For the fine stage, the convolutional layers and the size of filter kernels are the same as that of coarse stage. Since the output image patches reconstructed by coarse stage are smaller than the original input ones, smaller kernels will be learned if the input patches are smaller. By this way, the filter kernels learned via the fine network will be smaller and smaller so that smaller range can be considered when filtering. The training data are generated using Algorithm 1.

IV. Experimental Results

Dataset and Parameter Setting: To evaluate the performance of our proposed method, we conducted experiments on Middlebury 2003 datasets (including 4 scenes) [49], Middlebury 2005 datasets (including 6 scenes) [50], [51], and ToFMark databases (including 3 scenes) [37]. Each scene contains two views (left view and right view) with a depth image and its aligned color image in one view. The color images in the datasets are acquired by passive RGB-D cameras and supposed to be available in both training and testing stages. The *ideal* filter kernel will be computed from the high-resolution depth images using Eq. (2) (by setting σ_S to $+\infty$ and σ_R to 5) as discussed in Section III-A. The low-resolution depth images are obtained from the collected high-resolution depth images using nearest-neighbor downsampling. The input of our networks will be computed using Eq. (5) and Eq. (7) as discussed in Section III-A. We collected 60 RGBD images from Middlebury databases (6, 21 and 33 images are from 2001, 2006 and 2014 datasets respectively)) with deviation 1.5 and threshold of 0.35 for canny detector, and the input image patches in first network is 31×31 for the scale = 4, 8, 16. Other parameters can be found in Section III-D.

Baseline Methods: Our DSR method was quantitatively and qualitatively compared with the state-of-the-art methods. These methods can be separated into two categories: *(1) color assisted depth SR methods*: JBF [22], Tree [25], AutoRegressive (AR) [52], Guided [18], TGV [37], Joint Geodesic Filtering (JGF) [53], Edge [41], Cross-based Local Multipoint Filtering (CLMF) [54], Coupled Dictionary Learning with Local Constraints (CDLLC) [55], Joint Super Resolution and Denoising (JSRD) [56], MSG-Net [44], Xie *et al.* [26]; *(2) single depth image upsampling methods*: bicubic interpolation method, Patch Based method (PB) [57], SRCNN [7], Huang *et al.* [58], Super-Resolution via Sparse coding (ScSR) [33], Wang *et al.* [35]. Most of the results from these state-of-the-art methods are generated using the source code provided by the authors. For the training-based methods PB [57], SRCNN [7], MSG-Net [44], ATGV-Net [42], Song [43] and Wang *et al.* [35], we adopt the released model trained by the authors.

A. Quantitative Evaluation

For quantitative evaluation of cascade networks, we first evaluate our results on Middlebury 2005 databases [50], [51] with factors of 4, 8 and 16, respectively. To obtain LR depth images, we firstly smooth and downsample ground truth images. The evaluation metrics are two popular disparity error measurement metrics: percentage of error pixels (PE) and mean absolute difference (MAD). For both metrics, the smaller the better. Fig. 7 shows the comparison of our approach with different numbers of fine network with three factors, Fig. 7(a)–Fig. 7(c) show that different scenes have the similar tendency. From Fig. 7, we can see that: (1) Comparing the SingleDeep and Our_FS1 in Fig. 7, we can conclude that our cascade

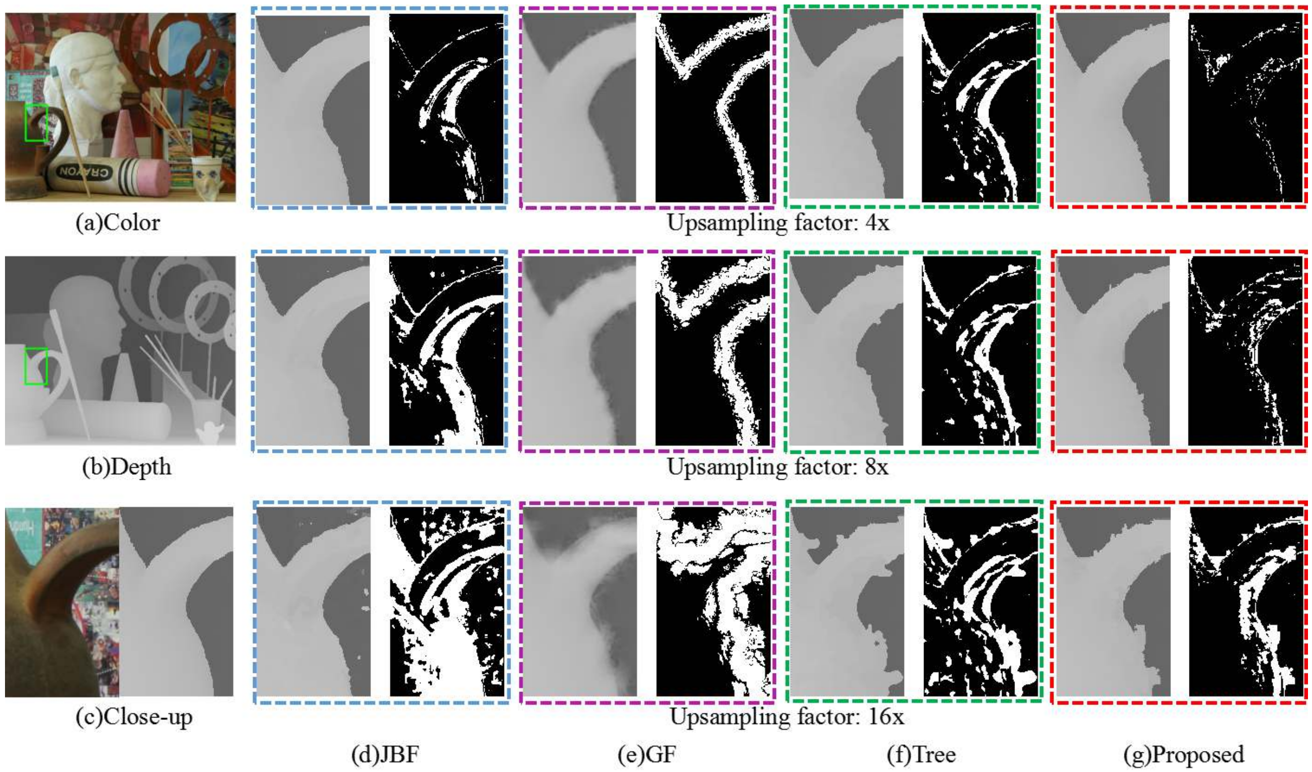


Fig. 8: Edge preserving quality. Visual comparison with popular edge-preserving filtering based upsampling methods [18], [22], [25] were conducted. (a) presents the high-resolution color image and the ground-truth disparity map. (b) and (c) are close-ups from (a). (d)-(g) are the close-ups of the disparity maps upsampled using different methods and the corresponding disparity error maps (obtained with error threshold 1). Note that although edge-preserving filters can all effectively maintain the edges, the accuracy is quite different. As can be seen from the binary error maps, the proposed method achieves the best accuracy around depth discontinuities.

architecture has better performance than the single deep neural network (see Fig. 6) for DSR; (2) Comparing the Our_CS and Our_FS1 to Our_FS3, the values of the fine stage are smaller than that of the coarse stage. This phenomenon proves that the cascade structure is effective; (3) Comparing the Our_FS1, Our_FS2 and Our_FS3, we can see that both indicators get better with the number of networks increases. However, when the number of fine networks is three, the performance gains are small while time-consuming increases too large. Thus, considering the balance between performance and time, we use only two networks at the fine stage in practice.

In Table I and Table II, our proposed method is compared with other 17 kinds of the state-of-the-art depth image super-resolution methods on Middlebury dataset 2005. According to Table III and Table IV, our results are compared with other methods on Middlebury dataset 2003. If the disparity error of a pixel is larger than 1, it is treated as an error pixel. The best performance in all the tables is marked in bold. From the four tables, we can see that the proposed method almost outperforms all the others on Middlebury datasets with all three upsampling factors. MAD in Table I and Table III measures mainly focus on average absolute error between reconstructed HR depth maps and ground truth ones. As the result shows, our approach is almost superior to other methods, including both traditional filtering-based methods and learning-based methods. Firstly, our color guidance manner can help to maintain the edge of depth. Secondly, the filter we learned is closest to the *ideal* filter for upsampling. Only very few values are not optimal for individual smooth area, that is because that the training patches are mostly selected around depth discontinuities areas, and our method is especially effective for edge discontinuous regions.

PE in Table II and Table IV measures the percentage of error pixels and thus all the inliers should be very accurate. As the result shows, the performance of the classical filter-based methods like Bicubic interpolation method will be not very good, especially around depth edges. That is because these filter-based methods cannot preserve the edges very well. The performance of the proposed method is almost better than other methods, including 4× and 8×. Only very few values are slightly lower than the MSG-Net with the factor of 16×. The reason is that our training data contains relatively little smooth area information. Table V shows a quantitative comparison on the ToFMark dataset [37] in MAD under 4 upsampling factors. As shown in the table, our approach has the best performance on all the three scenes of ToFMark dataset [37]. It also proves that our approach always outperforms on all three datasets comparing with other methods.

TABLE I: Quantitative comparison on the Middlebury dataset 2005 in *MAD* with three upsampling factors

	Art			Book			Dolls			Laundry			Moebius			Reindeer		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	10.58	19.59	35.15	3.80	8.15	16.41	4.73	9.60	19.49	7.42	14.82	26.77	4.52	9.38	18.51	5.20	9.88	19.13
CLMF0 [54]	7.57	16.72	33.32	3.17	7.25	16.93	3.97	9.65	18.36	6.11	12.57	25.35	4.03	8.40	17.60	4.60	9.71	18.26
CLMF1 [54]	8.12	17.28	33.25	3.27	7.25	16.09	4.04	8.76	18.32	5.50	12.67	25.40	4.13	8.42	17.27	4.65	9.96	18.34
TGV [37]	5.14	10.51	21.37	2.48	4.65	11.20	4.45	11.12	45.54	6.99	16.32	53.61	3.68	6.84	14.09	4.67	11.22	43.48
Guided [18]	9.97	15.53	28.43	3.68	6.52	13.07	4.46	7.63	15.87	6.33	11.90	20.26	4.78	7.88	14.84	5.16	8.11	15.71
JBF [22]	3.36	8.73	21.69	4.05	10.18	19.94	3.98	12.86	29.72	2.39	5.64	13.72	3.19	7.43	15.78	3.89	13.94	27.15
Edge [41]	6.82	13.49	25.90	3.35	8.50	19.32	2.90	6.84	17.97	2.82	5.46	13.57	3.72	7.36	14.05	2.67	6.22	16.80
JGF [53]	3.25	7.39	14.31	2.14	5.41	12.05	3.23	7.29	15.87	2.60	4.54	8.69	3.36	6.45	12.33	2.27	5.17	11.84
AR [52]	4.13	5.58	21.67	1.88	4.16	9.25	4.07	6.62	11.50	3.51	5.19	11.12	2.14	5.57	10.87	3.64	5.76	9.40
Tree [25]	3.96	5.24	9.74	5.77	7.22	11.48	4.60	6.36	13.02	2.27	3.94	8.87	3.52	4.90	8.67	3.97	5.76	2.77
KSVD [34]	3.46	5.18	8.39	2.13	3.97	8.76	4.53	6.18	12.98	2.19	3.89	8.79	2.08	4.86	8.97	2.19	5.76	12.67
CDLLC [55]	2.86	4.59	7.53	1.34	3.67	8.12	4.61	5.94	12.64	2.08	3.77	8.25	1.98	4.59	7.89	2.09	5.39	11.49
JSRD [56]	2.57	4.35	6.79	1.27	3.16	7.93	2.78	5.67	12.19	1.98	2.98	7.98	1.87	4.32	7.64	2.03	4.39	9.83
Xie [26]	2.48	3.31	**5.88**	1.23	3.09	7.58	2.72	5.59	12.06	1.62	2.86	7.87	1.88	4.29	7.63	1.97	4.31	9.27
PB [57]	3.12	6.18	12.34	1.39	3.34	8.12	3.99	6.22	12.86	2.68	5.62	11.76	1.95	4.12	8.32	6.04	12.17	21.35
SRCNN [7]	7.61	14.54	23.65	2.88	7.98	15.24	3.93	8.34	16.13	6.25	13.63	24.84	3.63	7.28	14.53	3.84	7.98	14.78
ATGV-Net [42]	3.78	3.78	9.68	5.48	7.16	10.32	4.55	6.27	12.64	2.07	3.78	8.69	3.47	4.81	8.56	3.82	5.68	2.63
Song [43]	2.39	3.28	5.82	1.21	2.98	7.48	2.59	5.47	11.78	1.56	2.75	7.64	1.86	4.15	7.52	1.86	3.92	8.67
Wang [35]	7.83	15.21	31.32	3.19	8.52	16.73	4.74	9.53	19.37	6.19	12.86	22.96	3.89	8.23	16.58	3.59	7.23	14.12
MSG-Net [44]	2.31	4.31	8.78	1.21	3.24	7.85	2.39	4.86	9.94	1.68	2.78	7.62	**1.79**	4.05	7.48	1.73	2.93	7.63
Our_CS	2.28	4.27	8.61	1.35	3.51	8.04	2.01	4.53	10.90	1.55	2.71	7.56	2.25	3.98	7.41	1.59	2.84	7.42
Our_FS	**2.23**	**3.59**	7.28	**1.19**	**3.07**	**7.32**	**1.98**	**4.49**	**9.84**	**1.39**	**2.49**	**7.35**	2.18	**3.91**	**7.41**	**1.51**	**2.79**	**6.58**

TABLE II: Quantitative comparison on the Middlebury dataset 2005 in *PE* with three upsampling factors

	Art			Book			Dolls			Laundry			Moebius			Reindeer		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	0.97	1.85	3.59	0.29	0.59	1.15	0.36	0.66	1.18	0.54	1.04	1.95	0.30	0.59	1.13	0.55	0.99	1.88
CLMF0 [54]	0.74	1.37	2.95	0.28	0.51	1.06	0.34	0.66	1.02	0.50	0.82	1.66	0.29	0.52	1.01	0.51	0.84	1.51
CLMF1 [54]	0.76	1.44	2.87	0.28	0.51	1.02	0.34	0.60	1.01	0.50	0.80	1.67	0.29	0.51	0.97	0.51	0.84	1.55
TGV [37]	0.65	1.17	2.30	0.27	0.42	0.82	0.33	0.70	2.20	0.55	1.22	3.37	0.29	0.49	0.90	0.49	1.03	3.05
Guided [18]	0.96	1.57	3.05	0.35	0.58	1.06	0.36	0.56	1.01	0.51	0.89	1.65	0.34	0.55	1.00	0.54	0.83	1.64
JBF [22]	0.55	1.08	2.26	0.38	0.71	1.40	0.41	0.82	1.80	0.33	0.61	1.33	0.33	0.68	1.44	0.45	0.90	1.77
Edge [41]	0.65	1.03	2.11	0.30	0.56	1.03	0.31	0.56	1.05	0.32	0.54	1.14	0.29	0.51	1.10	0.37	0.63	1.28
JGF [53]	0.47	0.78	1.54	0.24	0.43	0.81	0.33	0.59	1.06	0.36	0.64	1.20	0.25	0.46	0.80	0.38	0.64	1.09
AR [52]	0.49	0.64	2.01	0.22	0.37	0.77	0.34	0.50	0.82	0.34	0.53	1.12	0.20	0.40	0.79	0.40	0.58	1.00
Tree [25]	0.67	0.84	1.49	0.46	0.55	0.84	0.48	0.58	0.94	0.41	0.56	0.95	0.40	0.49	0.82	0.48	0.62	1.04
KSVD [34]	0.64	0.81	1.47	0.23	0.52	0.76	0.34	0.56	0.82	0.35	0.52	1.08	0.28	0.48	0.81	0.47	0.57	0.99
CDLLC [55]	0.53	0.76	1.41	0.19	0.46	0.75	0.31	0.53	0.79	0.30	0.48	0.96	0.27	0.46	0.79	0.43	0.55	0.98
JSRD [56]	0.51	**0.70**	1.37	0.17	0.39	0.72	0.29	0.51	0.76	0.29	0.47	0.94	0.24	0.43	0.76	0.39	0.53	0.96
Xie [26]	0.48	0.71	**1.35**	0.15	**0.36**	0.70	0.27	0.49	**0.74**	0.28	0.45	0.92	0.23	0.42	0.75	0.36	0.51	**0.95**
PB [57]	0.93	0.79	1.98	0.16	0.43	0.79	0.83	0.53	0.99	1.13	1.89	2.87	0.17	0.47	0.82	0.56	0.97	1.89
SRCNN [7]	0.63	1.21	2.34	0.25	0.52	0.97	0.29	0.58	1.03	0.40	0.87	1.74	0.25	0.43	0.87	0.35	0.75	1.47
ATGV-Net [42]	0.65	0.81	1.42	0.43	0.51	0.79	0.41	0.56	0.52	0.89	0.37	0.94	0.38	0.45	0.80	0.41	0.58	1.01
Song [43]	0.47	0.70	1.38	0.17	0.38	0.72	0.26	0.48	0.76	0.27	0.44	0.93	0.24	0.45	0.75	0.34	0.50	0.96
Wang [35]	0.73	1.56	3.03	0.28	0.61	1.31	0.32	0.65	1.45	0.45	0.98	2.01	0.31	0.59	1.26	0.42	0.84	1.73
MSG-Net [44]	0.46	0.76	1.53	**0.15**	0.41	0.76	**0.25**	0.51	0.87	0.30	0.46	1.12	**0.21**	0.43	0.76	0.31	0.52	0.99
Our_CS	0.45	0.74	1.55	0.22	0.39	0.74	0.27	**0.46**	0.82	0.26	0.44	0.94	0.25	0.41	0.74	0.31	0.48	0.97
Our_FS	**0.43**	0.72	1.50	0.17	**0.36**	**0.69**	**0.25**	**0.46**	0.75	**0.24**	**0.41**	**0.71**	0.23	**0.39**	**0.73**	**0.29**	**0.46**	**0.95**

TABLE III: Quantitative comparison on the Middlebury dataset 2003 in *MAD* with three upsampling factors

	Tsukuba			Venus			Teddy			Cones		
	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
Edge [41]	2.35	4.44	6.95	0.44	0.90	2.65	3.12	6.27	13.73	3.26	7.18	14.38
TGV [37]	1.79	3.08	5.31	0.41	0.60	1.76	2.31	3.72	7.51	2.54	4.34	8.17
ScSR [33]	3.27	6.15	9.17	0.71	1.43	3.42	3.76	7.79	15.86	4.43	9.33	17.35
KSVD [34]	2.48	4.30	6.78	0.59	1.22	3.15	2.97	5.17	8.93	3.97	6.45	12.51
SRCNN [7]	2.99	5.52	8.64	0.71	1.30	3.23	3.98	6.92	14.12	4.99	8.64	16.18
CDLLC [55]	2.41	4.15	6.59	0.71	1.18	3.08	2.99	4.72	9.13	3.68	5.79	11.23
Huang *et al.* [58]	3.53	6.20	9.32	0.67	1.45	3.61	3.88	7.37	15.24	4.52	8.44	15.38
PB [57]	1.57	2.52	3.69	0.39	0.66	1.83	4.13	8.03	17.90	4.35	9.73	17.69
JSRD [56]	1.4	2.37	3.52	0.38	0.59	1.69	1.71	3.13	6.23	1.96	3.23	6.53
Xie [26]	1.27	2.36	3.50	0.37	0.54	1.62	1.61	3.11	6.18	1.72	3.09	6.27
ATGV-Net [42]	1.52	2.41	3.59	0.40	0.63	1.76	5.35	5.37	7.62	4.63	5.74	7.36
Song [43]	1.25	2.23	3.49	0.39	0.53	1.60	1.63	3.10	4.52	1.71	3.05	4.37
Wang [35]	3.12	3.24	5.68	0.68	1.21	2.87	3.92	4.27	5.67	4.83	8.72	9.35
MSG-Net [44]	1.22	2.21	3.44	0.35	0.51	1.58	1.59	3.07	3.69	1.68	2.98	3.73
Our_CS	1.24	2.23	3.46	0.34	0.53	1.62	1.59	3.07	3.67	1.71	2.92	3.71
Our_FS	**1.16**	**2.18**	**3.42**	**0.33**	**0.51**	**1.56**	**1.58**	**2.98**	**3.58**	**1.64**	**2.89**	**3.70**

TABLE IV: Quantitative comparison on the Middlebury dataset 2003 in *PE* with three upsampling factors

	Tsukuba			Venus			Teddy			Cones		
	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
Edge [41]	0.61	0.77	1.32	0.23	0.29	0.56	0.78	1.08	2.13	1.03	1.52	2.98
TGV [37]	0.53	0.71	1.18	0.17	0.24	0.43	0.75	0.83	1.62	0.83	1.13	2.23
ScSR [33]	0.64	0.82	1.62	0.29	0.38	0.64	0.90	1.18	2.31	1.15	1.45	2.84
KSVD [34]	0.51	0.66	1.09	0.23	0.30	0.59	0.70	0.92	2.07	0.91	1.15	2.28
SRCNN [7]	0.64	0.79	1.43	0.28	0.34	0.61	0.88	1.10	2.35	1.12	1.41	2.91
CDLLC [55]	0.48	0.61	0.98	0.21	0.27	0.53	0.67	0.85	1.59	0.85	1.07	2.12
Huang *et al.* [58]	0.66	0.87	1.73	0.29	0.39	0.69	0.90	1.23	2.68	1.15	1.48	2.88
PB [57]	0.62	0.86	1.71	0.30	0.38	0.62	0.89	1.26	2.73	1.18	1.56	3.11
JSRD [56]	0.47	0.71	1.21	0.18	0.29	0.51	0.64	0.97	1.56	0.81	1.24	2.32
Xie [26]	0.45	0.67	1.09	0.19	0.29	0.49	0.63	0.95	1.51	0.76	1.16	2.14
ATGV-Net [42]	0.46	0.72	0.88	0.23	0.31	0.52	0.69	1.03	1.6	0.83	1.27	2.42
Song [43]	0.43	0.66	0.89	0.17	0.37	0.56	0.68	0.91	1.72	0.75	1.12	2.13
Wang [35]	0.65	0.68	0.83	0.26	0.34	0.69	0.75	1.24	3.01	1.856	1.35	4.86
MSG-Net [44]	0.41	0.62	0.75	0.14	0.34	0.57	0.65	0.82	2.76	0.73	1.06	2.22
Our_CS	0.43	0.65	0.73	0.14	0.26	**0.44**	0.63	0.80	1.67	0.93	1.08	2.31
Our_FS	**0.39**	**0.61**	**0.71**	**0.12**	**0.25**	**0.44**	**0.61**	**0.79**	**1.42**	**0.71**	**1.05**	**2.09**

TABLE V: Quantitative comparison on the ToFMark databases [37] using MAD metric with 4× upsampling factor

	Tree [25]	PB [57]	SRCNN [7]	GF [18]	JBF [22]	TGV [37]	KSVD [34]	CDLLC [55]	Xie [26]	ATGV-Net [42]	Song [43]	Wang [35]	MSG-Net [44]	Our_CS	Our_FS
Books	5.34	3.91	3.59	3.50	3.38	3.19	3.11	2.86	2.74	2.83	2.64	3.62	2.61	2.54	**2.51**
Devil	15.07	3.44	2.59	2.47	2.85	2.44	2.32	2.27	2.22	2.16	2.18	2.63	2.11	2.07	**2.01**
Shark	13.72	5.37	4.48	4.57	3.95	4.07	3.98	3.77	3.46	3.53	3.37	4.67	3.27	3.29	**3.26**
Average	11.38	4.24	3.56	3.51	3.39	3.23	3.13	2.97	2.79	2.86	2.69	3.64	2.66	2.63	**2.60**

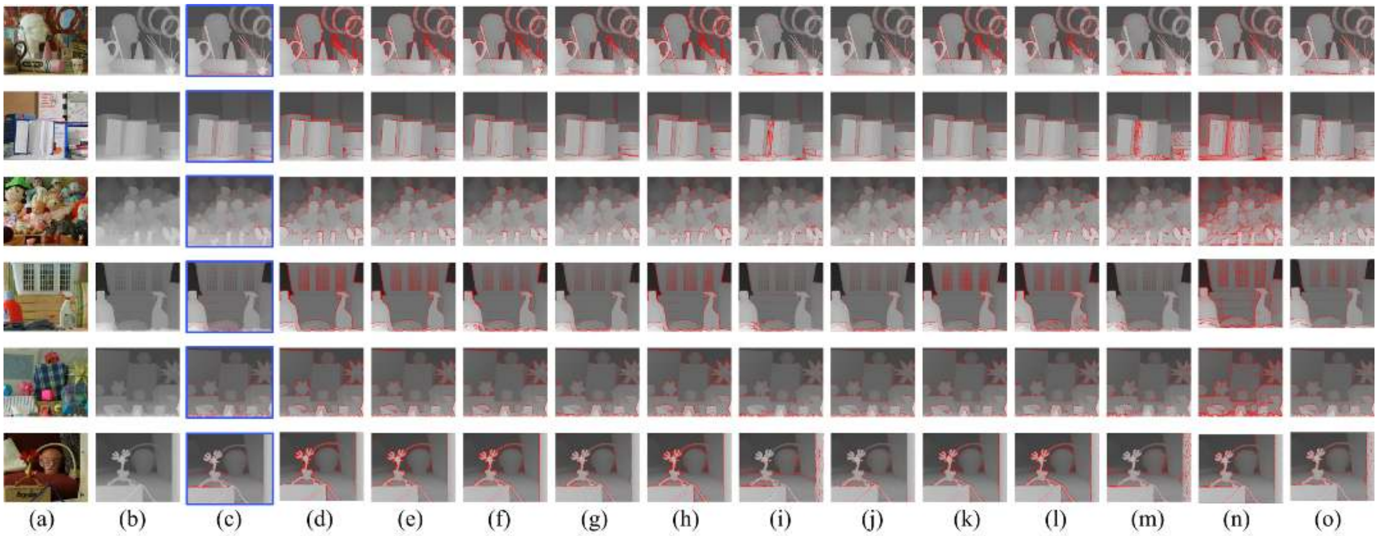


Fig. 9: Visual comparison of upsampling images on Middlebury database (scaling factor = 4), the upsampling pixel errors are marked with red. (a) Color image. (b) Ground truth. (c) Our Proposed. (d) AR [52]. (e) Bicubic. (f) CLMF0 [54]. (g) CLMF1 [54]. (h) Edge [41]. (i) Guided [18]. (j) JBF [22]. (k) JGF [53]. (l) TGV [37]. (m) Tree [25]. (n) ATGV-Net [42]. (o) Song [43].

B. Qualitative Evaluation

Fig. 8 visually compares edge-preserving quality on several popular edge-preserving filtering based DSR methods [18], [22], [25] using the *Art* database. Fig. 8(a) presents registered color image and ground-truth high-resolution disparity image. Fig. 8(b)–Fig. 8(c) present the color and ground-truth disparity values of a close-up region where the color and depth edges are not consistent. Fig. 8(d)–Fig. 8(g) are the disparity values (of the close-up) upsampled using different methods and the corresponding disparity errors (obtained with error threshold 1). The methods [18], [22], [25] mainly relies on the color edges in the registered high-resolution color image to preserve the depth edges. The accuracy drops when the color edges are not always aligned well with the depth edges. The proposed method uses CNN to learn a data-driven combination of the color and depth information and thus is more accurate around depth discontinuities. Fig. 9, Fig. 10 and Fig. 11 show the comparison with a great many popular super-resolution methods. To make the comparison clearer, we use red color to mark the super-resolution error pixel. The less red dots, the better. Among the figures, Fig. 9, Fig. 10 and Fig. 11 all show that the quality of our reconstructed super-resolution depth image is better than that of other methods on the Middlebury dataset. We also can see that our approach generates more visually appealing results than the previous ones, Especially, in edge areas, our reconstruction effects are better. Besides, our method not only has better reconstruction effect than most current methods, but also has comparable running speed. If just single scale CNN is used, the average running speed of dataset Middlebury 2003 is 1.24 second. As cascading more CNN, the reconstruction quality will improve at the expense of running time. But the increased running time is acceptable since we use a three-layer lightweight CNN.

V. Conclusion

In this paper, we propose to solve the depth super-resolution problem via a cascade coarse-to-fine convolutional neural network. First, we propose the concept of the *ideal* filter and use the deep network to approach it. Through the coarse network, large edge-preserving filters are learned to approximate the *ideal* filters to obtain a rough depth map. Then, smaller filtering kernels are learned to optimize results so that better high-resolution depth image can be achieved progressively. Besides, we use an additional registered high-resolution color image as guidance to modify the uncertain interpolated depth value so that it can achieve a better combination of the high-resolution color and the low-resolution depth information. Numerous experiments on different databases have demonstrated the effectiveness of our proposed approach. In the future, we will work on more challenging tasks such as super-resolution problems with noisy depth inputs, we will also study better color guidance for even high-quality effects generation.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (61572316, 61671290, 61872241), the National Key Research and Development Program of China (2016YFC1300302, 2017YFE0104000), the Science and Technology Commission of Shanghai Municipality (16DZ0501100, 17411952600).

References

[1] F. Jiang, H. Li, X. Hou, B. Sheng, R. Shen, X.-Y. Liu, W. Jia, P. Li, and R. Fang, "Abdominal adipose tissues extraction using multi-scale deep neural network," *Neurocomputing*, vol. 229, pp. 23–33, 2017.

[2] A. Masood, B. Sheng, P. Li, X. Hou, X. Wei, J. Qin, and D. Feng, "Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images," *Journal of Biomedical Informatics*, vol. 79, pp. 117–128, 2018.

[3] P. Li, H. Sun, C. Huang, J. Shen, and Y. Nie, "Interactive image/video retexturing using GPU parallelism," *Computers & Graphics*, vol. 36, no. 8, pp. 1048–1059, 2012.

[4] P. Li, H. Sun, B. Sheng, and J. Shen, "Image stylization with enhanced structure on GPU," *Science China Information Sciences*, vol. 55, no. 5, pp. 1093–1105, 2012.

[5] Y. Nie, C. Xiao, H. Sun, and P. Li, "Compact video synopsis via global spatiotemporal optimization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 10, pp. 1664–1676, 2013.

[6] Y. Nie, H. Sun, P. Li, C. Xiao, and K. Ma, "Object movements synopsis via part assembling and stitching," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 9, pp. 1303–1315, 2014.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[8] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[9] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5895–5907, 2017.

[10] C. Kuster, T. Popa, C. Zach, C. Gotsman, and M. Gross, "FreeCam: A hybrid camera system for interactive free-viewpoint video," in *Vision, Modeling, and Visualization*, 2011, pp. 1–8.

[11] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun, "Multi-view image and ToF sensor fusion for dense 3D reconstruction," in *IEEE International Conference on Computer Vision Workshops*, 2009, pp. 1542–1549.

[12] D. Holz, R. Schnabel, D. Droeschel, J. Stückler, and S. Behnke, "Towards semantic scene analysis with time-of-flight cameras," in *RoboCup 2010: Robot Soccer World Cup XIV*, 2011, pp. 121–132.

[13] A. Karambakhsh, A. Kamel, B. Sheng, P. Li, P. Yang, and D. D. Feng, "Deep gesture interaction for augmented anatomy learning," *International Journal of Information Management*, pp. 1–9, 2018.

[14] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–14, 2018.

[15] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[16] Y. Yang, M. Gao, J. Zhang, Z. Zha, and Z. Wang, "Depth map super-resolution using stereo-vision-assisted model," *Neurocomputing*, vol. 149, pp. 1396–1406, 2015.

[17] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," in *IEEE International Conference on Multimedia and Expo*, 2012, pp. 152–157.

[18] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[19] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, 2013.

[20] F. Tombari, S. Mattoccia, and L. Di Stefano, "Segmentation-based adaptive support for accurate stereo correspondence," in *Pacific-Rim Symposium on Image and Video Technology*, 2007, pp. 427–438.

[21] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008, pp. 1–12.

[22] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[23] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 96:1–96:5, 2007.

[24] Q. Yang, N. Ahuja, R. Yang, K.-H. Tan, J. Davis, B. Culbertson, J. Apostolopoulos, and G. Wang, "Fusion of median and bilateral filtering for range image upsampling," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4841–4852, 2013.

[25] Q. Yang, "Stereo matching using tree filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 834–846, 2015.

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

[27] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*, 2014, pp. 31–42.

[28] K.-H. Lo, Y.-C. F. Wang, and K.-L. Hua, "Joint trilateral filtering for depth map super-resolution," in *Visual Communications and Image Processing*, 2013, pp. 1–6.

[29] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *International Conference on Computer Vision*, 1998, pp. 839–846.

[30] Q. Yan, Y. Xu, X. Yang, and T. Q. Nguyen, "Single image superresolution based on gradient profile sharpness," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3187–3202, 2015.

[31] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.

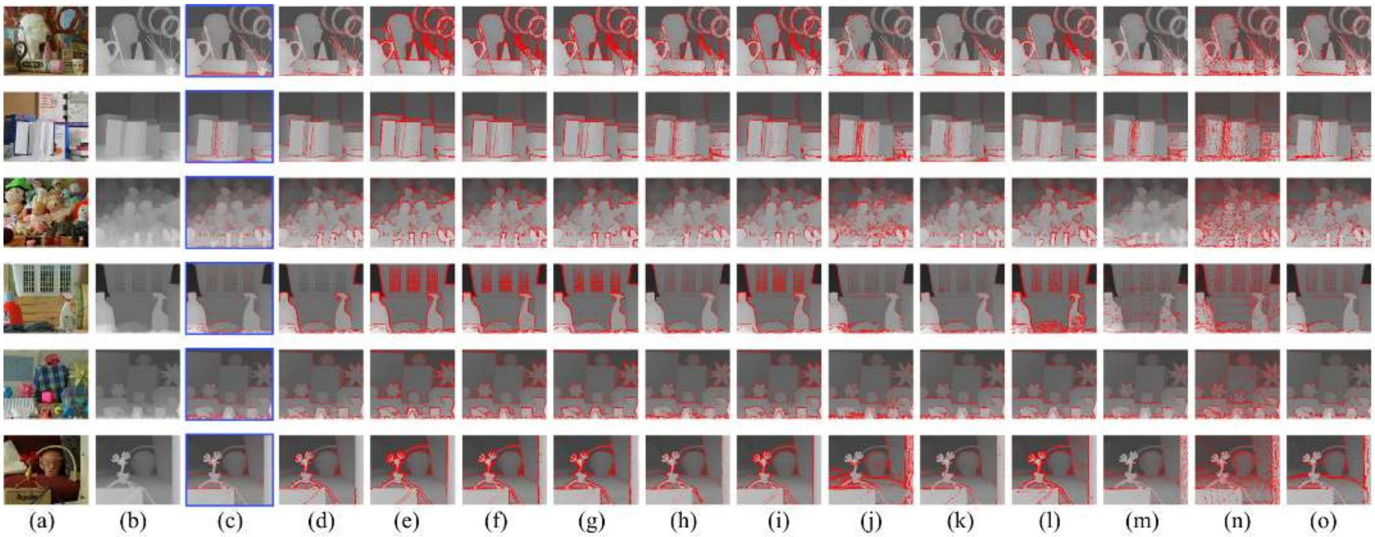


Fig. 10: Visual comparison of upsampling images on Middlebury database (scaling factor = 8), the upsampling pixel errors are marked with red. (a) Color image. (b) Ground truth. (c) Our Proposed. (d) AR [52]. (e) Bicubic. (f) CLMF0 [54]. (g) CLMF1 [54]. (h) Edge [41]. (i) Guided [18]. (j) JBF [22]. (k) JGF [53]. (l) TGV [37]. (m) Tree [25]. (n) ATGV-Net [42]. (o) Song [43].

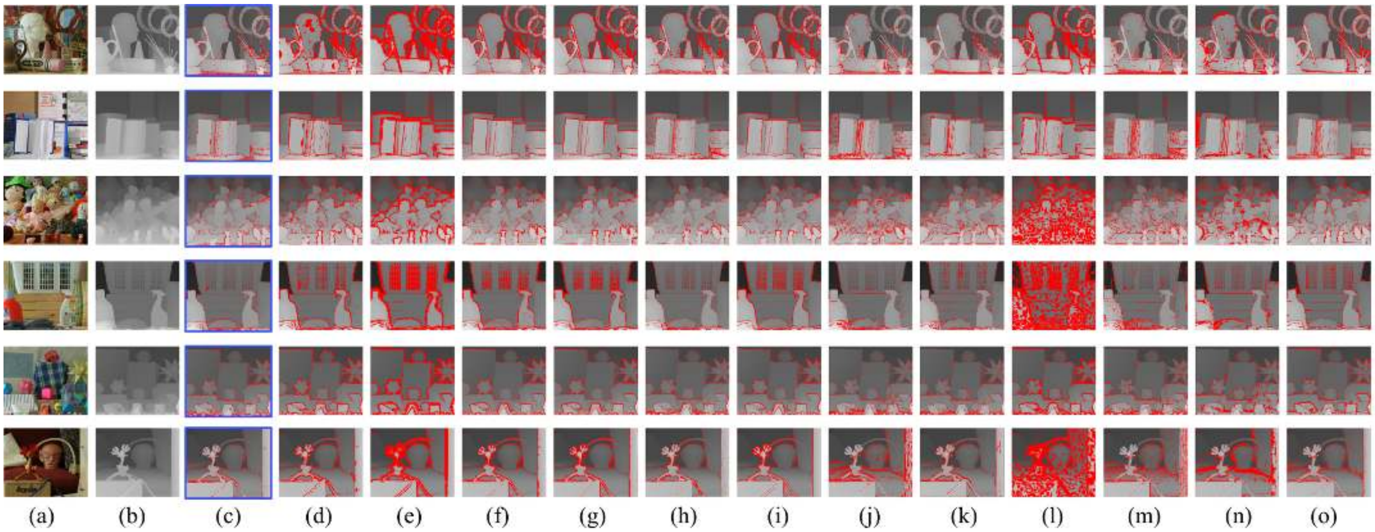


Fig. 11: Visual comparison of upsampling images on Middlebury database (scaling factor = 16), the upsampling pixel errors are marked with red. (a) Color image. (b) Ground truth. (c) Our Proposed. (d) AR [52]. (e) Bicubic. (f) CLMF0 [54]. (g) CLMF1 [54]. (h) Edge [41]. (i) Guided [18]. (j) JBF [22]. (k) JGF [53]. (l) TGV [37]. (m) Tree [25]. (n) ATGV-Net [42]. (o) Song [43].

[32] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, “Coupled dictionary training for image super-resolution,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

[33] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[34] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in *International Conference on Curves and Surfaces*, 2012, pp. 711–730.

[35] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, “Deep networks for image super-resolution with sparse prior,” in *IEEE International Conference on Computer Vision*, 2015, pp. 370–378.

[36] J. Diebel and S. Thrun, “An application of Markov random fields to range sensing,” in *Neural Information Processing Systems*, 2005, pp. 291–298.

[37] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in *IEEE International Conference on Computer Vision*, 2013, pp. 993–1000.

[38] F. Kou, W. Chen, C. Wen, and Z. Li, “Gradient domain guided image filtering,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4528–4539, 2015.

[39] C. A. Ochotorena, C. N. Ochotorena, and E. Dadios, “Gradient-guided filtering of depth maps using deep neural networks,” in *International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management*, 2015, pp. 1–8.

[40] K.-L. Hua, K.-H. Lo, and Y.-C. F. Wang, “Extended guided filtering for depth map upsampling,” *IEEE MultiMedia*, vol. 23, no. 2, pp. 72–83, 2016.

[41] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, “High quality depth map upsampling for 3D-TOF cameras,” in *International Conference on Computer Vision*, 2011, pp. 1623–1630.

[42] G. Riegler, M. Rüther, and H. Bischof, “ATGV-Net: Accurate depth super-resolution,” in *European Conference on Computer Vision*, 2016, pp. 268–284.

[43] X. Song, Y. Dai, and X. Qin, “Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network,” in *Asian Conference on Computer Vision*, 2016, pp. 360–376.

[44] T.-W. Hui, C. C. Loy, and X. Tang, “Depth map super-resolution by deep multi-scale guidance,” in *European Conference on Computer Vision*, 2016, pp. 353–369.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012, pp. 1097–1105.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[49] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 195–202.

[50] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[51] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[52] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3443–3458, 2014.

[53] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 169–176.

[54] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, "Cross-based local multipoint filtering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 430–437.

[55] J. Xie, C.-C. Chou, R. Feris, and M.-T. Sun, "Single depth image super resolution and denoising via coupled dictionary learning with local constraints and shock filtering," in *IEEE International Conference on Multimedia and Expo*, 2014, pp. 1–6.

[56] J. Xie, R. S. Feris, S.-S. Yu, and M.-T. Sun, "Joint super resolution and denoising from a single depth image," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1525–1537, 2015.

[57] O. Mac Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *European Conference on Computer Vision*, 2012, pp. 71–84.

[58] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

Yang Wen received the M.Eng. degree in computer science from the Xidian University, Xi'an, China. She is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include depth image super-resolution, convolutional neural networks, image/video processing, computer graphics, and computer vision.

Bin Sheng received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, China. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include image-based rendering, machine learning, virtual reality, and computer graphics.

Ping Li received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, China. He is currently an Assistant Professor with the Macau University of Science and Technology, Macau, China. His current research interests include image/video stylization, big data visualization, GPU acceleration, and creative media. He has one image/video processing national invention patent, and has excellent research project reported worldwide by *ACM TechNews*.

Weiyao Lin received the B.Eng. and M.Eng. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2005, respectively, and the Ph.D. degree from the University of Washington, Seattle, USA, in 2010. He is currently an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include image/video processing, video surveillance, and computer vision.

David Dagan Feng (F'03) received the M.Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), Los Angeles, CA, in 1985 and 1988, respectively, where he received the Crump Prize for Excellence in Medical Engineering. He is currently the head in the School of Information Technologies, the director in the Biomedical & Multimedia Information Technology Research Group, and the research director in the Institute of Biomedical Engineering and Technology at the University of Sydney, Sydney, Australia. He has published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his field. More importantly, however, is that many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He has served as the chair in the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems, has organized/chaired over 100 major international conferences/symposia/workshops, and has been invited to give over 100 keynote presentations in 23 countries and regions. He is a fellow of the IEEE and Australian Academy of Technological Sciences and Engineering.