Guidance-based improved depth upsampling with better initial estimate

Chandra Shaker Balure* and Ramesh Kini M

Department of Electronics and Communication,
National Institute of Technology Karnataka,
Surathkal, India
Email: balure1986a@gmail.com
Email: rameshkinim@gmail.com
*Corresponding author

Abstract: Like optical images, depth images are also gaining popularity because of its use in many applications like robot navigation, augmented reality, 3DTV and more. The commercially available depth cameras generate depth images which suffer from low spatial resolution, corrupted with noise, and missing regions. Such images need to be super-resolved, denoised and inpainted before using them to have better accuracy. Super-resolution (SR) techniques can be used to produce a high-resolution output. Since SR is an ill-posed inverse problem, a good initial estimate is always a good regulariser to find the optimal solution. We propose an initial estimate as part of our SR pipeline, esp. ×8, which will helps in quick convergence and accurate output. We propose a cascade approach by combining residual interpolation (RI) method with anisotropic total generalised variation (ATGV) method, both uses HR guidance image. The improvements are shown qualitative and quantitative with different levels of noise.

Keywords: super-resolution; depth image; initial estimate; interpolation; cascade.

Reference to this paper should be made as follows: Balure, C.S. and M, R.K. (2021) 'Guidance-based improved depth upsampling with better initial estimate', *Int. J. Computational Vision and Robotics*, Vol. 11, No. 1, pp.109–125.

Biographical notes: Chandra Shaker Balure received his BE in Electronics Engineering from the S.R.T.M.U. Nanded University, MTech in VLSI Design and Embedded Systems from the Visveswaraya Technological University Belgaum, and PhD from the National Institute of Technology Karnataka (NITK) Surathkal in 2007, 2009 and 2019, respectively. He has 1.5 years of industry and 3.5 years of teaching experience in engineering colleges. His current research interests include image processing, computer vision, and machine learning.

Ramesh Kini M has received his BE from the Mysore University in 1984, and MTech from the Mangalore University in 1997. He has received his PhD from the National Institute of Technology Karnataka (NITK), Surathkal. He has 3.5 years of industry experience. In 1990, he joined NITK. His research interests are digital VLSI design, dynamically reconfigurable processor for multimedia applications, multicore OpenRISC processor and image processing.

1 Introduction

Optical images are everywhere and they are easy to capture because the modern optical cameras are handy, portable and real-time and they are able to capture the images and videos at a high-resolution (HR). On contrary to optical cameras, the modern depth cameras are bulky because of its internal image capturing mechanism and it operates at low frame rates and capture images and videos at a very low-resolution.

Depth images are becoming popular because of its demand for applications like robot navigation, autonomous vehicle driving, 3D TV, augmented reality and many more. To get better accuracy, these applications need HR depth images. However, the modern depth cameras are not able to keep in pace with the development of the modern optical cameras, and hence it suffers from low-resolution and noise. The modern high-end depth cameras are still trying to reach the competition in terms of image resolution, but they still fail at the computational speed and compactness, and hence not suitable for real-time applications mentioned earlier. Hence, the commercially available modern depth cameras [e.g., time-of-flight (ToF) cameras, or Kinect camera], which are capable of rendering depth images at a faster frame rate (e.g., ~50 frames per second), are suitable for real-time applications. The problem with these cameras is that it capture images with low spatial resolution, corrupted with noise, and suffer from irregular or regular missing region (called depth holes or depth shadows) because of occlusions. Such images are not suitable for direct use, so the possible solution to increase the resolution without overburdening the hardware is to use some image processing techniques, which gives improved resolution image by still utilise the fast rendering modern depth camera.

Super-resolution (SR) methods comes as a rescue for this problem. SR is a set of technique to increase the spatial resolution of an image by maintaining the image details (e.g., edges, corners, textures, etc.) intact. SR methods either require sequence of LR images (called multiple image SR problem) or single LR image (called single image SR problem). The traditional SR methods generally involve sequence of LR images of same scene which are sub-pixel shifted from the reference image. However, such a process of obtaining these image sequence is tedious and is not practical in real scenario. Recent SR methods use single image as input. A single input without any external cue is difficult to address, however it can be handled using interpolation methods, which is explained later in the section, makes the image blur which results in removing the image details. Hence, such outputs compel to use some external cues like set of training examples to learn HR-LR relationship, or a colour guidance image to learn the image structure. As SR problem is an ill-posed inverse problem, they are regularised to obtain a optimal solution from the infinite solution space.

The classical interpolation methods (e.g., nearest neighbour, bilinear, bicubic, etc.) does a job of super-resolving an image, but the output suffer from blurring artefacts because of its implicit low-pass filtering characteristics. For depth images, the edge discontinuities are more significant, and since interpolation methods does not respect the high-frequency details (e.g., edges, corners, etc.), it is not suitable for the problem of depth image super-resolution. The SR methods for depth images are to be such that, it should able to retain the edge discontinuity while super-resolving them.

Since depth images are mostly smooth or linearly smooth at object surfaces, and the sharp discontinuity at object boundaries, the SR methods can target for larger upsampling factor, thereby definitely retaining the edge discontinuities and depth precision to a large extent. Towards this, in literature, there are variety of methods trying to address the

single depth image SR problem. Most of the SR methods first try to estimate an initial output, where is treated as an initial estimate, and thereby improving upon it to get more accurate results. The initial estimation process varies across different SR methods. Some starts with the sparse LR input (Ferstl et al., 2013), where the input LR image is mapped onto the HR grid of the desired resolution and then estimate the unknown pixels from the know pixels based on some prior information, or some starts with bicubic interpolation of the LR input (Yang et al., 2013). Such initial estimates become inappropriate at higher upsampling factor, which is challenging, as because, at higher upsampling factors there will be more unknown pixels to estimate from a very less know pixels. There are some other methods, which use bicubic interpolation as an initial estimate, but the problem with such estimate is that it does not consider the image details, which results into smoothing image details.

In this paper, we are trying to find some suitable initial estimate which is as fast as classical interpolation, and as good as the expected SR output, especially for noisy cases, which is considered as more challenging task. We have combined two distinct methods of residual interpolation (RI) by Konno et al. (2015) and anisotropic total generalised variation (ATGV) by Ferstl et al. (2013), to improve the SR results for single depth image SR problem. We utilise the RI method to generate an initial estimate (which is fast), and then apply the ATGV method for final depth restoration (which is efficient). As we go higher on upsampling factors, the proposed combination of RI and ATGV does a good job in producing the satisfying results.

We have tested our proposed method on standard Middlebury dataset (Scharstein and Szeliski, 2003) for four different upsampling factors (i.e., ×2, ×4, ×8 and ×16). We have shown results for noisy images, as this is more challenging, and the results are compared for qualitative and quantitative analysis against classical interpolation methods (bilinear and bicubic), and also against RI method (Konno et al., 2015), and ATGV (Ferstl et al., 2013).

The rest of the paper is organised as follows: in Section 1.1 we review some of the existing work on guidance image-based depth image SR, followed by the detailed description of the proposed method in Section 2. The results and its discussion are provided in Section 3, followed by the conclusion in Section 4.

1.1 Related work

We can classify depth image SR methods into three categories, i.e., multiple image depth SR, mingle image depth SR, and guidance image-based depth SR. For the sake of brevity and relatedness, we will discuss only the guidance image-based depth image SR.

The problem definition for guidance image-based depth image SR is that, for a given HR colour guidance image and an LR depth image, the SR method has to estimate an HR depth image whose target resolution is equivalent to the resolution of the HR guidance image. Such methods are becoming popular because of the easy availability of a rig with two cameras placed side-by-side, with one as LR depth camera and other as HR optical camera [e.g., rig of three cameras in Li et al. (2008)]. Since, the viewpoint of both the cameras are different which results in misalignment of image frames, but it can be taken care by well recognised image registration methods using calibration techniques. Hence, for our work and for most of the work in literature, we assume that the HR intensity

image and LR depth image aligned such that these two image are co-aligned and the prominent edges coincide on each other.

The thought of guidance image came into existence when Tomasi and Manduchi (1998) proposed bilateral filter (BF) for filtering the noisy image along with edge preservation. This edge preserving smoothing BF filter estimate the *domain* and *range* kernel based on geometric closeness and photometric similarity respectively. The domain kernel refers to closeness of pixel values, and range kernel refers to similarity of pixel values. However, it estimate these kernels from the same input. Later, He et al. (2010) proposed a guided image filter (GIF) which uses guidance image under the assumption that there is a local linear model between the guidance input and the filtered output. Here, the guidance image can be the input image itself or another different image.

Based on the concept of BF filter, Kopf et al. (2007) proposed a joint bilateral upsampling (JBU) method, which uses another image (second image) as guidance image for range kernel estimation, which helps in combining the high frequencies from one image and low frequencies from another. But the JBU method, which essentially splits the kernel into intensity part and depth part, does suffer from the problem of copying the texture from colour image into a smoother region where the depth reading contains huge random noise. To deal with texture copying problem in the heavy noisy region, Chan et al. (2008) proposed noise aware filter for depth upsampling (NAFDU), which formulate the objective function such that it consider behaves as JBU at less noisy region and behaves as BF at heavy noisy region dampening the effect of guidance image. Yang et al. (2013) proposed a method which combines median filtering and BF filtering, named joint bilateral weighted median filter (JBM), for the problem of depth upsampling in an hierarchical fashion, which claims it to be improving the upsampling accuracy and reduces the computational complexity.

Later, Garcia et al. (2010) proposed an extension to the JBU method to get away with the texture copying problem. They found that limiting the prior information only from the guidance image itself is not sufficient, hence they used depth image values also in estimating the range kernel. They proposed an addition factor for the filter kernel, called credibility map (CM) which is based on the gradient information of the LR depth input by assigning lower weights to the pixels along the strips of depth edges. By using their pixel weighted average strategy (PWAS), they fuse the depth data together for depth upsampling. Further, Garcia et al. (2011) proposed a filter which uses credibility weight of a pixel to decide whether to use the PWAS filter which uses only guidance image or to use the same PWAS filter but with considering only depth information, and Garcia et al. (2015) proposed a unified multi-lateral (UML) filter where the reliability weight decides whether to consider kernel with intensity image (PWASI) or kernel with depth image (PWASD), and thereby improving the accuracy within smooth regions. Kim et al. (2010) proposed an additional kernel term to the JBU filter which weigh the similarity in the input depth image.

The work of Diebel and Thrun (2006) uses Markov random field (MRF) for the problem of generating HR images by combining the LR depth image along with the acquired registered HR intensity image. The mode of the probability distribution defined by the MRF provide us the HR depth image. Yang et al. (2007) built a cost volume of depth probability which is based on the probability distribution of depth, to which BF filter is applied to generate the HR output after sub-pixel refinement. Hua et al. (2016) exploit local gradient information of input depth image to deal with texture copying problem of JBU. Park et al. (2014) address depth map upsampling and completion

problem by combining the non-local structure regularisation with edge weighting scheme. Liu et al. (2013) proposed joint geodesic depth upsampling method which compute geodesic distances from each unknown pixels on the HR grid to all the known pixels from LR input. Yang and Wang (2012) combines GIF approach and reconstruction constraints to generate the final HR depth image. Lu and Forsyth (2015) uses HR colour guidance image to extract segments boundaries and corresponding depth boundaries from the co-aligned depth image, and each segment in depth image is reconstructed independently using their smoothing method. Xiao et al. (2015) proposed defocus deblurring and SR of ToF depth image by regularising the solution in amplitude and depth space directly.

Recently, there has been some work on SR of depth and colour images using convolutional neural network. The work of Tai et al. (2017) proposed deep recursive residual network (DRRN) which is 52 layer deep network which require enormous parameter learning. Lim et al. (2017) proposed enhanced deep super-resolution (EDSR) network which uses conventional residual networks. They also propose a new multi-scale deep super-resolution (MDSR) system to reconstruct HR image of different upsampling factors. There has also been sufficient work on depth image SR also using CNN. The work of Song et al. (2016) proposed deep depth SR method by exploiting the depth field statistics and local correlation between depth image and colour image. There has been some work on super-resolving colour image and depth image simultaneously, for which Zhao et al. (2019) proposed a method which use generative adversarial networks (GAN) to enhance a pair of low-quality colour-depth images by merging the features of both the images.

Our proposed method falls under the category of guidance-based depth SR methods. In our proposed method, we have combined RI method (Konno et al., 2015) and ATGV method (Ferstl et al., 2013), and both these methods require HR intensity image as guidance image.

2 Proposed method

The problem statement is, given an LR depth image d and a guidance image as HR colour image I, the SR method is required to estimate the HR depth image $\hat{D}$, whose spatial resolution is equal to the spatial resolution of the guidance image, which is close to the ground truth (GT) image D.

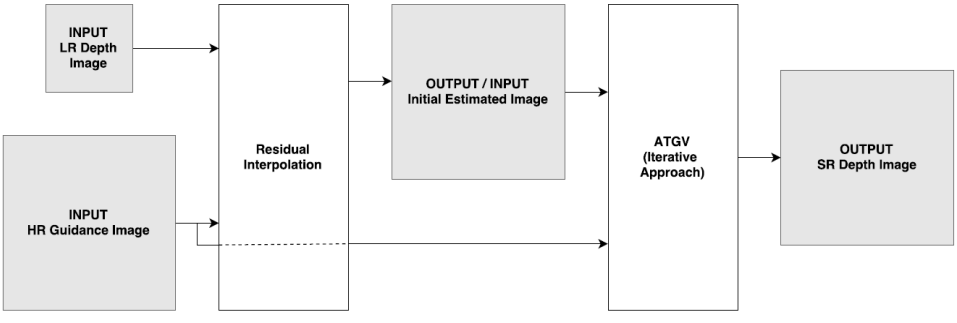
Like other guidance image-based SR methods, we too assume that the input LR depth and HR colour images are co-aligned at each pixel. This assumption is valid, as we have seen in literature that, capturing the intensity image is a low-cost operation and easy, and it can be captured along with the depth images mounted on a same rig. Under this constraint, we assume that the prominent edges in depth image coincide with the edges in the intensity image.

The RI method (Konno et al., 2015) is inspired by GIF approach, where it assumes the local linear mapping between the guidance image and the output image. RI method operates in the residual domain, where the residual is the different between the tentative estimated HR depth map and the LR depth map. This residual is then interpolated added with the tentative estimate to recover the final HR depth map. As this process is easy and fast, we consider this as our initial estimate for the modules in our SR pipeline.

The output from RI method is then fed as an input to the ATGV module (Ferstl et al., 2013). ATGV use anisotropic diffusion tensor, calculated from HR guidance image, is used to guide the upsampling process.

The overall proposed methods block diagram is shown in Figure 1. This figure shows the complete work flow of our proposed method which is a cascade of two approaches combined in a single framework to get a sharp and accurate HR depth output. To this whole SR pipeline, we feed in two inputs, one is the LR depth image and other is the HR guidance image. As we are proposing the guidance-based method for depth image SR, we have considered a RI method as the first stage in our SR pipeline which takes LR input and the HR guidance image to produce an initial estimate of the SR depth image. This output is then fed as an initial input, as opposed to other sparse LR depth-based inputs (Ferstl et al., 2013) or the bicubic interpolated LR depth input (Yang et al., 2013), to the second stage in our SR pipeline which is the ATGV module, which also makes use of the same HR guidance image as used in the first stage. So, the problem statement for our SR problem is stated as follows: Given an LR depth image d of size $m \times n$ and an HR colour guidance image I of size $\alpha m \times \alpha n$, our proposed SR method tries to obtain a super-resolved depth image $\hat{D}$ of size of size $\alpha m \times \alpha n$, equivalent to the resolution of the guidance image, which needs to be as close as the GT image D.

Figure 1 Block diagram of proposed ATGV mod method combining RI and ATGV in cascade form



In the following subsections, we will briefly discuss about the two modules used in our work, i.e., the RI module (Konno et al., 2015) and ATGV module (Ferstl et al., 2013).

2.1 RI method

RI method (Konno et al., 2015) is the initial module in our SR pipeline. It takes the LR depth input d and the HR guidance image I. The complete flow of RI method is shown in the block diagram in Figure 2, where the LR depth image is represented by d and the HR colour guidance image represented by I.

The complete process of RI is mainly processed in residual domain, where residual means the difference between the tentative depth output and the LR input. The tentative depth output is generated using the popularly known GIF method by He et al. (2010) approach, which considers that the dominant edges in the input depth image coincides with the edges in the colour guidance image by considering the local linear relationship

between d and I. The local linear combination between the input guidance image and the tentative output is given in equation (1):

$$t_i = a_k I_i + b_k, \quad \forall i \in w_k \tag{1}$$

where I and t represent the HR colour guidance image and the HR tentative depth output, and i represent all the pixel location in those images, and w_k denotes local window centred at pixel location k, and a_k and b_k are the local linear coefficients. These linear coefficients for each pixel location at calculation is calculated by minimising the cost function $E(\cdot)$ which is given in equation (2):

$$E(a_k, b_k) = \sum \left(a_k I_i^M + b_k - d_i^2 \right) + \eta a_k^2 \tag{2}$$

where I_i^M is the pixel value of the masked HR intensity image, and d_i is the corresponding LR depth value, and η is the regularisation parameter. The linear coefficients for a pixel location are obtained by weighted averaging given in equation (3), instead of just averaging:

$$\hat{a}_k = \frac{\sum_{i \in w_k} W_i a_i}{\sum_{i \in w_k} W_i}, \; \hat{b}_k = \frac{\sum_{i \in w_k} W_i b_i}{\sum_{i \in w_k} W_i} \tag{3}$$

where the weight W is determined by the cost of GIF as in equation (4):

$$W_i = \frac{1}{\max\left(\frac{1}{|\omega_i|} \sum \left(a_i I_i^M + b_i - d_j \right)^2, \delta \right)} \tag{4}$$

where δ is the threshold parameter to avoid the *divide-by-zero* situation. The tentative estimate is finally calculated as given in equation (5):

$$t_i = \hat{a}_k I_i + \hat{b}_k, \quad \forall i \in w_k \tag{5}$$

The tentative output is then masked in accordance to the sparse LR depth input to obtain the residuals. The residual is the result of the image different between the masked tentative output and the sparse LR input. The tentative output will be sharper than the LR input, and hence their different will give us the high-frequency information. The residual image is then bicubic interpolated to estimate the missing pixels from the residual image grid. The interpolated image is then added back to the tentative estimated depth image of earlier step to get the sharper output image.

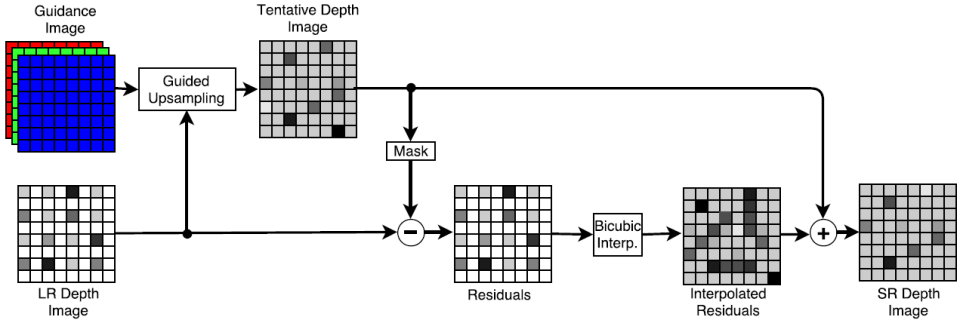
The RI output is more accurate in terms of the edge sharpness and depth precision, and hence, it can be the best initial estimate in our SR pipeline. There are few benefits of considering RI output as the input to the next cascaded ATGV module.

1 Firstly, with good initial solution, the convergence will be faster.

2 Secondly, the output will be more accurate as opposed to the approaches which used no such initial estimate, instead they start from the sparse LR depth input itself as in the case of ATGV (Ferstl et al., 2013) method.

The RI output, which is fast in operation and visually sharp is given to the second module of our SR pipeline (i.e., ATGV module). The RI depth output along with the same HR colour guidance image is passed as input to finally obtain the super-resolved depth image. Even for higher magnification factor, the initial solution from RI method does a better job of preserving the edge information in the final output and converge to the solution faster.

Figure 2 Block diagram of RI method (see online version for colours)



Source: Konno et al. (2015)

2.2 ATGV method

ATGV method (Ferstl et al., 2013) was proposed by keeping in mind the problems of modern depth cameras. As are variety of depth cameras, as discussed earlier, which can measure the depth/distance of the object from the camera position based on the principle of time-off light, ATGV tries to solve their problems of capturing the low-resolution depth image by adding information from the HR guidance image in a variational optimisation framework. The complete work flow is shown in the block diagram in Figure 3.

They proposed a convex optimisation problem which has two terms involved in it, one is the data term and second one is the regularisation term. The data term enforces the output to look similar to the input measurements, and the regularisation term enforces piecewise solution by preserving the edges and reducing the noise. The regularisation term, they use higher order total generalised variation (TGV) regularisation which is weighted according to the texture in the intensity image by an anisotropic diffusion tensor.

With the formation of core convex energy functional, the whole upsampling process is divided into three steps, which are:

1 first task is to register the LR depth image and HR guidance image into one common coordinate system

2 then formulating the convex energy function with higher order regularisation function

3 then solving the optimisation function with first-order primal-dual optimisation scheme.

For coordinate mapping, one image plane has to be considered as a reference plane on which the other image is projected back. Here, the HR guidance image plane is

considered as a reference plane with known intrinsic and extrinsic camera parameters. The LR depth image d at each pixel location $x_{i,j} = [i, j, 1]^T$ is projected onto the HR image plane to a new 3D pixel location $\hat{x}_{i,j}$, which is represented as in equation (6):

$$\hat{d}_{i,j} = C_L + d_{i,j} \frac{P_L^{\dagger} x_{i,j}}{\left\| P_L^{\dagger} x_{i,j} \right\|} \tag{6}$$

where C_L is the depth camera centre and $P_L^{\dagger}$ is the pseudoinverse of depth camera projection matrix. This projected image gives the sparse HR depth image as the mapping from depth image space to guidance image space is on-to-one to avoid the problem of averaging, whereas the unknown pixels are interpolated.

From the sparse HR depth image and with additional cue from HR guidance image, the dense HR depth image is give by equation (7):

$$\hat{D} = \arg\min_{u} \left\{ G\left(u, \hat{d}\right) + \alpha F(u) \right\} \tag{7}$$

where $G(u, \hat{d})$ is the data term that measures quality of u to the input $\hat{d}$ and $F(u)$ is the regularisation term with prior knowledge of smoothness of the final solution, and G and F are the convex lower semi-continuous functions, and α variable is to balance between the data term and the regulariser. The data term is represented by equation (8) as:

$$G(u, D_S) = \int_{\Omega_H} w \left| u - \hat{d} \right|^2 dx \tag{8}$$

where w is a weighter operator between [0, 1] $\in$ ¡ $_{\Omega_H}$ which is zero at unmapped image points.

The whole burden is on the regularisation term to produce a sharp depth output. Earlier, regularisation terms were of first-order smoothness, for example total variation semi norm with L1 norm gives $\|\nabla u\|_1$, but this regulariser could not be used for depth images resulting in piecewise fronto parallel depth reconstruction. Hence, a more generalised regularisation model called TGV is used, which is composed of polynomials of arbitrary order which results in piecewise polynomial depth reconstruction. An order of k favours solutions composed of polynomials of order $k - 1$, so for depth images second-order TGV suffice, which is given by equation (9):

$$TGV_{\alpha}^{2} = \min_{v} \left\{ \alpha_1 \int_{\Omega} |\nabla u - v| \, dx + \alpha_0 \int_{\Omega} |\nabla v| \, dx \right\} \tag{9}$$

where α_0 and α_1 are scalar weights. To produce accurate HR depth output at the edge discontinuities, an anisotropic diffusion tensor $T^{\frac{1}{2}}$ is computed using the HR guidance image, which is calculated as shown by equation (10):

$$T^{\frac{1}{2}} = \exp\left(-\beta \left| \nabla I_H \right|^{\gamma}\right) nn^T + n^{\perp} n^{\perp T} \tag{10}$$

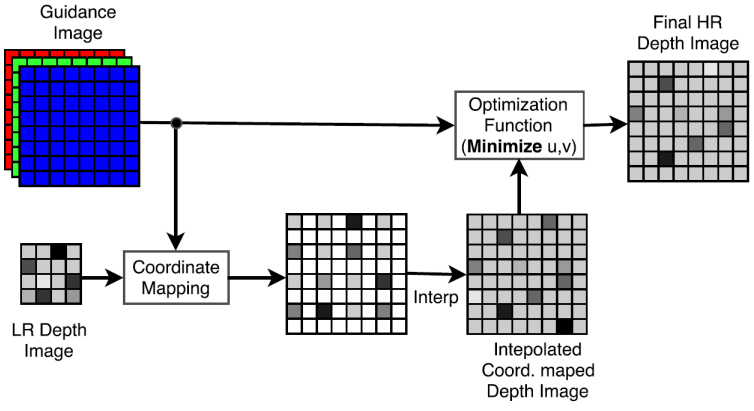
where n is the direction of the gradient, and $n^{\perp}$ is the normal vector to the gradient, and β and γ adjust the direction and sharpness of the tensor.

The final energy is defined as a combination of data term [equation (8)] and TGV term [equation (9)] with anisotropic diffusion [equation (10)] is represented in equation (11) as:

$$\min_{u,v}\left\{\alpha_1\int_{\Omega_H}\left|T^{\frac{1}{2}}(\nabla u - v)\right|dx + \alpha_0\int_{\Omega_H}|\nabla u|\,dx + \int_{\Omega_H} w\left|u-\hat{d}\right|^2 dx\right\} \tag{11}$$

To find the solution to this convex optimisation problem, they use primal-dual energy minimisation scheme which runs iteratively for all pixels individually.

Figure 3 Block diagram of ATGV method (see online version for colours)



Source: Ferstl et al. (2013)

3 Results and discussion

We have evaluated our proposed HR guidance image-based depth image SR method on depth image from standard Middlebury dataset (Scharstein and Szeliski, 2003). We chose this dataset as it has registered depth and colour images. We have experimented our work on noisy LR depth images with various levels of noise with its standard deviation ranging from σ = 1, 2, 3, 4 and 5). We have tested over three different upsampling factors of 2, 4 and 8.

Before we start discussing the results, we would like to mention about the LR image were generated using the LR modelling which is represented in equation (12). To the GT depth image D, it first blurs it with Gaussian blurring filter of kernel size 7 × 7 with mean 0 and standard deviation 1.6. We then downsample it to simulate the low-resolution depth image, which is then added with additive Gaussian noise with different noise levels.

$$d = \mathbb{S}\mathbb{B}D + \eta \tag{12}$$

where $\mathbb{S}$ is the sub-sampling matrix, $\mathbb{B}$ is the blurring operator, and η is the additive Gaussian noise. We use this model only to generate our observed LR images, and it is not used anywhere in our depth image reconstruction stage.

Table 1 MSE results of depth SR by factor ×2, ×4 and ×8 for lowest and highest noise levels, i.e., $\sigma = 1$ and $\sigma = 5$

Images	*Bic*	*RI*	*ATGV*	*ATGV mod*	*Bic*	*RI*	*ATGV*	*ATGV mod*	*Bic*	*RI*	*ATGV*	*ATGV mod*
			×2_σ1				*×4_σ1*				*×8_σ1*	
Aloe	3.92	3.82	*3.61*	3.62	4.25	3.81	*3.75*	3.76	5.19	*3.95*	4.40	4.45
Art	2.00	1.86	*1.60*	1.61	2.43	1.84	1.71	*1.70*	3.68	*2.01*	2.78	2.76
Baby	1.33	1.29	*1.01*	1.02	1.46	1.27	*0.99*	1.00	1.84	1.31	*1.22*	1.25
Books	2.23	2.20	1.91	*1.91*	2.37	2.22	1.98	*1.95*	2.80	2.31	2.35	*2.26*
Bowling	2.58	2.52	2.23	*2.22*	2.78	2.50	2.36	*2.34*	3.40	*2.59*	2.81	2.72
Cones	4.76	4.72	4.45	*4.45*	4.92	4.72	4.54	*4.52*	5.34	4.78	4.94	*4.71*
Moebius	2.16	2.13	1.84	*1.84*	2.31	2.14	1.89	*1.89*	2.74	*2.23*	2.31	2.29
Plastic	1.29	1.26	0.90	*0.89*	1.39	1.25	0.92	*0.90*	1.68	1.31	1.10	*1.01*
Reindeer	2.05	1.99	1.71	*1.71*	2.28	1.98	1.75	*1.75*	2.98	*2.08*	2.27	2.21
Teddy	4.26	4.23	3.96	*3.96*	4.39	4.24	*3.99*	4.00	4.73	4.31	4.23	*4.23*
Average	2.65	2.60	2.32	*2.32*	2.85	2.59	2.38	*2.38*	3.43	*2.68*	2.84	2.78
			×2_σ5				*×4_σ5*				*×8_σ5*	
Aloe	6.29	6.17	*5.42*	5.43	6.55	6.10	*4.80*	4.81	7.31	6.17	5.24	*5.20*
Art	4.40	4.25	*3.34*	3.35	4.75	4.16	2.59	*2.59*	5.78	4.27	3.61	*3.57*
Baby	3.84	3.76	*2.83*	2.84	3.94	3.69	1.84	*1.84*	4.23	3.70	1.87	*1.85*
Books	4.71	4.65	3.70	*3.70*	4.80	4.61	2.87	*2.83*	5.10	4.66	3.01	*2.90*
Bowling	5.04	4.95	3.95	*3.95*	5.18	4.88	3.18	*3.16*	5.69	4.93	3.57	*3.48*
Cones	7.16	7.08	*6.20*	6.21	7.29	7.04	5.41	*5.39*	7.59	7.05	5.50	*5.37*
Moebius	4.62	4.56	3.64	*3.64*	4.72	4.52	2.75	*2.75*	5.03	4.56	3.12	*3.08*
Plastic	3.81	3.75	2.65	*2.65*	3.88	3.69	1.65	*1.63*	4.10	3.73	1.65	*1.57*
Reindeer	4.52	4.42	3.50	*3.50*	4.70	4.37	2.59	*2.59*	5.26	4.44	2.97	*2.91*
Teddy	6.70	6.64	5.75	*5.75*	6.79	6.60	4.82	*4.82*	7.01	6.63	4.80	*4.76*
Average	5.10	5.02	*4.09*	4.10	5.26	4.96	3.25	*3.24*	5.71	5.01	3.53	*3.46*

Table 2 MSE results of depth SR by factor ×2, ×4 and ×8 for different noise levels, i.e., $\sigma = 2$, 3 and 4

Images	*Bic*	*RI*	*ATGV*	*ATGV mod*	*Bic*	*RI*	*ATGV*	*ATGV mod*	*Bic*	*RI*	*ATGV*	*ATGV mod*
			×2_σ2				*×4_σ2*				*×8_σ2*	
Aloe	4.50	4.40	*4.07*	4.08	4.80	4.38	*4.04*	4.05	5.68	4.49	*4.62*	4.63
Art	2.59	2.45	*2.05*	2.06	2.98	2.41	1.96	*1.95*	4.16	2.56	3.00	*2.98*
Baby	1.95	1.91	1.49	*1.49*	2.07	1.87	1.23	*1.23*	2.42	1.91	*1.40*	1.41
Books	2.84	2.81	2.37	*2.37*	2.96	2.81	2.23	*2.20*	3.33	2.89	2.51	*2.42*
Bowling	3.19	3.13	2.66	*2.66*	3.36	3.10	2.61	*2.59*	3.94	3.17	3.00	*2.92*
Cones	5.35	5.30	4.89	*4.89*	5.50	5.30	4.79	*4.77*	5.88	5.34	5.06	*4.89*
Moebius	2.77	2.73	*2.30*	2.31	2.89	2.73	2.14	*2.14*	3.28	2.80	2.53	*2.51*
Plastic	1.91	1.89	1.35	*1.35*	2.00	1.87	1.14	*1.11*	2.26	1.92	1.26	*1.16*
Reindeer	2.66	2.59	*2.16*	2.17	2.87	2.57	1.99	*1.98*	3.51	2.66	2.45	*2.40*
Teddy	4.86	4.83	4.42	*4.42*	4.97	4.83	4.23	*4.23*	5.27	4.89	4.38	*4.38*
Average	3.26	3.20	*2.77*	2.78	3.44	3.18	2.63	*2.62*	3.97	3.26	3.02	*2.97*
			×2_σ3				*×4_σ3*				*×8_σ3*	
Aloe	5.10	4.99	*4.52*	4.53	5.38	4.95	4.30	*4.30*	6.21	5.05	4.83	*4.83*
Art	3.19	3.05	2.49	*2.49*	3.56	2.99	2.17	*2.17*	4.68	3.12	3.21	*3.19*
Baby	2.58	2.52	1.94	*1.94*	2.69	2.48	1.44	*1.44*	3.01	2.50	1.56	*1.55*
Books	3.46	3.42	2.81	*2.81*	3.57	3.41	2.45	*2.41*	3.91	3.48	2.68	*2.58*
Bowling	3.80	3.74	3.09	*3.09*	3.97	3.69	2.81	*2.79*	4.51	3.76	3.21	*3.12*
Cones	5.95	5.90	5.33	*5.33*	6.09	5.88	5.00	*4.99*	6.44	5.91	5.22	*5.06*
Moebius	3.38	3.34	2.75	*2.75*	3.50	3.32	2.35	*2.35*	3.85	3.39	2.73	*2.71*
Plastic	2.55	2.51	1.79	*1.79*	2.63	2.48	1.32	*1.30*	2.87	2.53	1.39	*1.31*
Reindeer	3.28	3.20	2.61	*2.61*	3.47	3.16	2.19	*2.19*	4.08	3.26	2.63	*2.57*
Teddy	5.48	5.44	*4.86*	4.87	5.57	5.42	4.43	*4.43*	5.84	5.47	4.51	*4.51*
Average	3.87	3.81	*3.21*	3.22	4.04	3.77	2.84	*2.83*	4.54	3.84	3.19	*3.14*
			×2_σ4				*×4_σ4*				*×8_σ4*	
Aloe	5.70	5.58	*4.97*	4.98	5.97	5.52	*4.55*	4.56	6.75	5.61	5.04	*5.03*
Art	3.80	3.64	2.92	*2.92*	4.16	3.57	2.38	*2.38*	5.23	3.70	3.41	*3.38*
Baby	3.21	3.14	2.39	*2.39*	3.32	3.08	1.64	*1.64*	3.62	3.10	1.71	*1.70*
Books	4.09	4.04	3.26	*3.26*	4.18	4.01	2.66	*2.62*	4.50	4.07	2.84	*2.74*
Bowling	4.42	4.34	3.52	*3.52*	4.58	4.28	3.00	*2.98*	5.10	4.35	3.40	*3.31*
Cones	6.55	6.49	*5.76*	5.77	6.69	6.46	5.21	*5.19*	7.02	6.48	5.36	*5.22*
Moebius	4.00	3.95	*3.19*	3.20	4.11	3.92	2.56	*2.56*	4.44	3.97	2.93	*2.90*
Plastic	3.18	3.13	2.22	*2.22*	3.26	3.08	1.48	*1.46*	3.49	3.13	1.52	*1.44*
Reindeer	3.90	3.81	*3.05*	3.06	4.08	3.76	2.40	*2.39*	4.67	3.85	2.79	*2.74*
Teddy	6.09	6.04	5.31	*5.31*	6.18	6.01	4.63	*4.63*	6.42	6.05	4.66	*4.63*
Average	4.49	4.41	*3.65*	3.66	4.65	4.36	3.05	*3.04*	5.12	4.43	3.36	*3.30*

For comparison purpose, we have chosen RI method and ATGV method as state-of-the-art guidance-based depth SR methods, and we also compare with classical bicubic interpolation method. To show the quantitative performance of our method, we have computed mean squared error (MSE) between the SR output image and the GT image, which is as shown in equation (13):

$$MSE = \frac{1}{p}\sum_{i=1}^{p}\left(\hat{D}_i - D_i\right)^2 \tag{13}$$

where D is the GT depth image and $\hat{D}$ is the estimated SR depth output, and $p = \alpha m \times \alpha n$ is the dimension of the desired output image.

Figure 4 shows the SR results for various upsampling factors 2, 4 and 8 respectively on a *noisy* depth inputs *Cones* with additive Gaussian noise of standard deviation $\sigma = 5$, the higher noise we considered in our experiment. As one can see that our output shown in the last column of Figure 4 are more sharper than all other comparative methods. The first row shows the SR output for upsampling factor 2, and the subsequent rows for upsampling factors 4 and 8 respectively. The ATGV Mod output is not much distinguishable in ×2 case, but we can notice carefully the head and the stick region in our output is sharper and noise free as compared to other methods. In yet another Figure 5 of SR outputs on depth image *art*, we can see that the output produced by our method shown in last column are more sharper for all the upsampling factors.

For better visual representation, we show the outputs generated by our proposed method on depth image *teddy* with its small region cropped and zoomed. The results shown are produced under different noise levels of $\sigma = 1, 2, 3, 4$ and 5, and for different upsampling factors of ×2, ×4 and ×8 in Figure 6, Figure 7 and Figure 8 respectively. We can notice from the cropped region (*teddy head*) clearly that our method is able to suppress the noise of various noise levels and is able to maintain the depth precision and edge discontinuities in the generated output.

Figure 4 SR results comparison of *noisy* ($\sigma = 5$) depth image *cones*, (a) row 1: SR by factor ×2 (b) row 2: SR by factor ×4 (c) row 3: SR by factor ×8

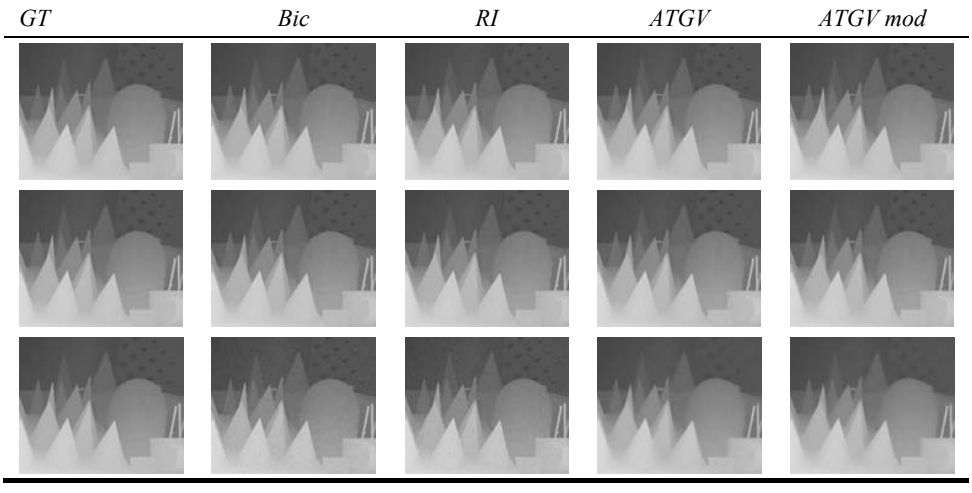


Figure 5 SR results comparison of *noisy* ($\sigma = 5$) depth image *art*, (a) row 1: SR by factor ×2 (b) row 2: SR by factor ×4 (c) row 3: SR by factor ×8

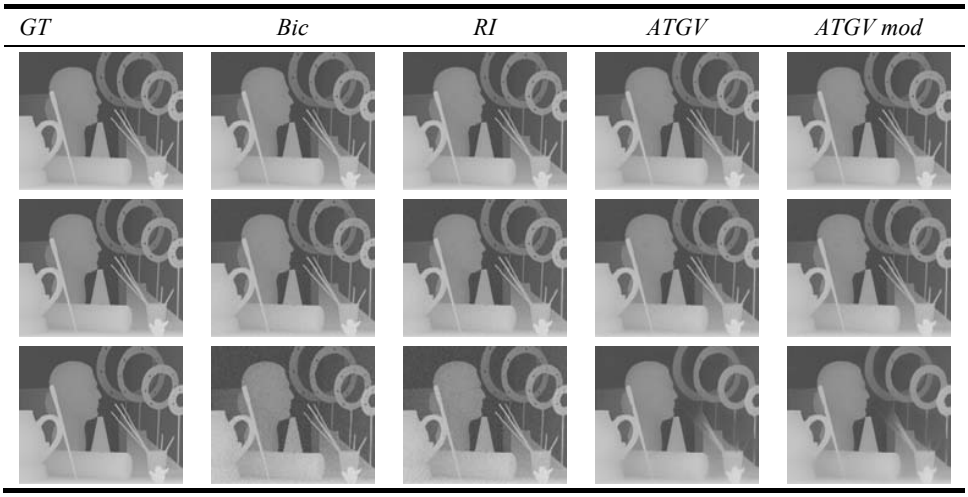


Table 1 and Table 2 shows the MSE performance metric of SR methods on few selected test depth images from Middlebury dataset. We even show the average MSE on the selected test images with different noise levels and for all the upsampling factors. Overall we can see that our method performs better or comparable with other competitive methods for various upsampling factor under different levels of noise. Table 1 shows MSE performance metric on image with lowest ($\sigma = 1$) and highest ($\sigma = 5$) noise level. It shows that our proposed cascade method overall performs well, especially for higher upsampling factor which is more important. For lower level of noise, our method is not able to perform better, however our results are still comparable to other SR methods. Whereas, with higher level of noise, our method is able to show good results for higher upsampling factors. Table 2 shows MSE performance metric on same set of test images but with other levels of added noise, i.e., $\sigma = 2$, 3 and 4. This experiment was performed to see how our method performs under different level of noise. In this scenario also we perform better than other SR methods for SR factors 4 and 8.

Figure 6 SR results of our method under different noise levels for upsampling factors ×2, (a) row 1: outputs from proposed method (b) row 2: cropped and zoomed region of their corresponding top images (see online version for colours)

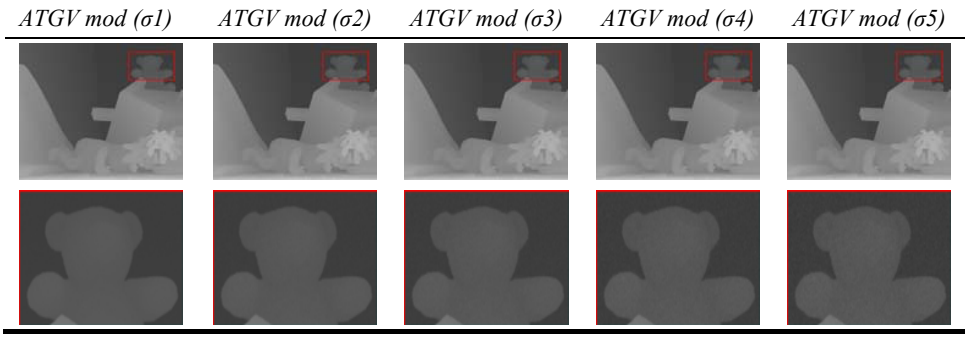


Figure 7 SR results of our method under different noise levels for upsampling factors ×4, (a) row 1: outputs from proposed method (b) row 2: cropped and zoomed region of their corresponding top images (see online version for colours)

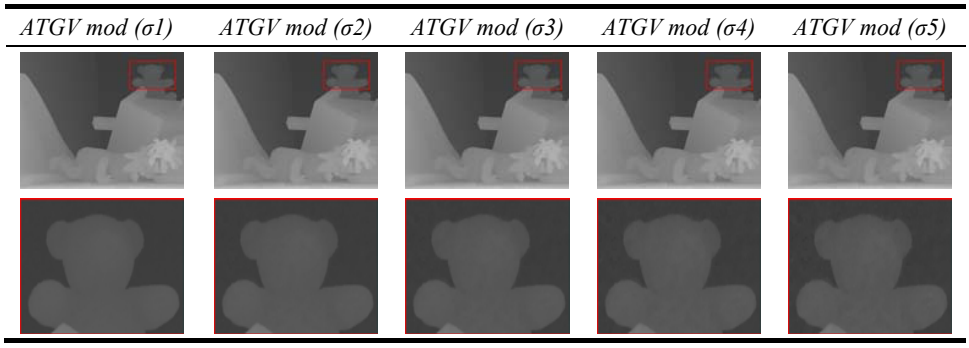
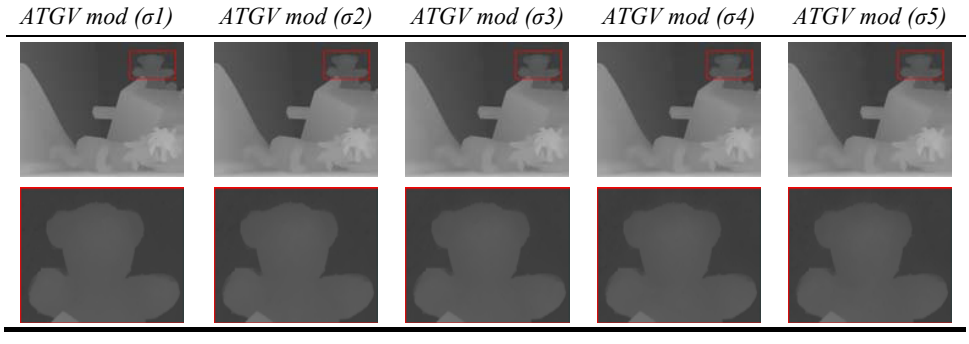


Figure 8 SR results of our method under different noise levels for upsampling factors ×8, (a) row 1: outputs from proposed method (b) row 2: cropped and zoomed region of their corresponding top images (see online version for colours)



4 Conclusions

We have proposed a guidance image-based depth image SR method, which is a combination of RI method and ATGV method in the SR pipeline. We had strong intuition that the initial estimate is a strong cue to improve the SR accuracy, and hence we chose to utilise the RI output as an initial estimate for the second module (ATGV module) in our SR pipeline, which not only helps in faster convergence, but also leads to better accuracy. This initial estimate is better because it is as fast as any other interpolation method. We have experimented our method on various depth images, and we have shown qualitatively and quantitatively that our proposed method performs equally well for upsampling factor 2, but does a better job of maintaining the depth precision and the edge discontinuities as compared to other SR methods, especially for higher upsampling factors 4 and 8 which is a good sign.

References

Chan, D., Buisman, H., Theobalt, C. and Thrun, S. (2008) 'A noise-aware filter for realtime depth upsampling', in *Workshop on Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications-M2SFA2*.

Diebel, J. and Thrun, S. (2006) 'An application of Markov random fields to range sensing', in *Advances in Neural Information Processing Systems*, pp.291–298.

Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M. and Bischof, H. (2013) 'Image guided depth upsampling using anisotropic total generalized variation', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.993–1000.

Garcia, F., Aouada, D., Mirbach, B., Solignac, T. and Ottersten, B. (2011) 'A new multi-lateral filter for real-time depth enhancement', in *2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, IEEE, pp.42–47.

Garcia, F., Aouada, D., Mirbach, B., Solignac, T. and Ottersten, B. (2015) 'Unified multilateral filter for real-time depth map enhancement', *Image and Vision Computing*, Vol. 41, pp.26–41.

Garcia, F., Mirbach, B., Ottersten, B., Grandidier, F. and Cuesta, A. (2010) 'Pixel weighted average strategy for depth sensor data fusion', in *2010 17th IEEE International Conference on Image Processing (ICIP)*, IEEE, pp.2805–2808.

He, K., Sun, J. and Tang, X. (2010) 'Guided image filtering', in *European Conference on Computer Vision*, Springer, pp.1–14.

Hua, K-L., Lo, K-H. and Wang, Y-C.F.F. (2016) 'Extended guided filtering for depth map upsampling', *IEEE Multimedia*, Vol. 23, No. 2, pp.72–83.

Kim, S-Y., Cho, J-H., Koschan, A. and Abidi, M.A. (2010) 'Spatial and temporal enhancement of depth images captured by a time-of-flight depth sensor', in *2010 20th International Conference on Pattern Recognition (ICPR)*, IEEE, pp.2358–2361.

Konno, Y., Monno, Y., Kiku, D., Tanaka, M. and Okutomi, M. (2015) 'Intensity guided depth upsampling by residual interpolation', in *The International Conference on Advanced Mechatronics: Toward Evolutionary Fusion of IT and Mechatronics: ICAM: Abstracts*, Vol. 2015, pp.1–2.

Kopf, J., Cohen, M.F., Lischinski, D. and Uyttendaele, M. (2007) 'Joint bilateral upsampling', in *ACM Transactions on Graphics (ToG)*, ACM, Vol. 26, p.96.

Li, F., Yu, J. and Chai, J. (2008) 'A hybrid camera for motion deblurring and depth map super-resolution', in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, IEEE, pp.1–8.

Lim, B., Son, S., Kim, H., Nah, S. and Mu Lee, K. (2017) 'Enhanced deep residual networks for single image super-resolution', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Liu, M-Y., Tuzel, O. and Taguchi, Y. (2013) 'Joint geodesic upsampling of depth images', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.169–176.

Lu, J. and Forsyth, D. (2015) 'Sparse depth super resolution', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2245–2253.

Park, J., Kim, H., Tai, Y-W., Brown, M.S. and Kweon, I.S. (2014) 'High-quality depth map upsampling and completion for RGB-D cameras', *IEEE Transactions on Image Processing*, Vol. 23, No. 12, pp.5559–5572.

Scharstein, D. and Szeliski, R. (2003) 'High-accuracy stereo depth maps using structured light', in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003 Proceedings*, IEEE, Vol. 1, p.1.

Song, X., Dai, Y. and Qin, X. (2016) 'Deep depth super-resolution: learning depth superresolution using deep convolutional neural network', in *Asian Conference on Computer Vision*, Springer, pp.360–376.

Tai, Y., Yang, J. and Liu, X. (2017) 'Image super-resolution via deep recursive residual network', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tomasi, C. and Manduchi, R. (1998) 'Bilateral filtering for gray and color images', in *Sixth International Conference on Computer Vision*, IEEE, pp.839–846.

Xiao, L., Heide, F., O'Toole, M., Kolb, A., Hullin, M.B., Kutulakos, K. and Heidrich, W. (2015) 'Defocus deblurring and superresolution for time-of-flight depth cameras', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2376–2384.

Yang, Q., Ahuja, N., Yang, R., Tan, K-H., Davis, J., Culbertson, B., Apostolopoulos, J. and Wang, G. (2013) 'Fusion of median and bilateral filtering for range image upsampling', *IEEE Transactions on Image Processing*, Vol. 22, No. 12, pp.4841–4852.

Yang, Q., Yang, R., Davis, J. and Nistér, D. (2007) 'Spatial-depth super resolution for range images', in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07*, IEEE, pp.1–8.

Yang, Y. and Wang, Z. (2012) 'Range image super-resolution via guided image filter', in *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, ACM, pp.200–203.

Zhao, L., Bai, H., Liang, J., Zeng, B., Wang, A. and Zhao, Y. (2019) 'Simultaneous color-depth super-resolution with conditional generative adversarial networks', *Pattern Recognition*, Vol. 88, pp.356–369.