DEPTH UPSAMPLING BY DEPTH PREDICTION

Atsuhiko Tsuchiya, Daisuke Sugimura and Takayuki Hamamoto

Graduate School of Engineering, Tokyo University of Science, Katsushika-Ku, Tokyo, 125-8585, Japan

ABSTRACT

We propose a method for depth upsampling with the aid of high-resolution color image. The key novelty of our method is to exploit a spatio-temporal coherency between the color and depth image *sequences*. It allows us to perform a depth prediction using the responses from motion estimation in the color image sequence. The predicted depth image is able to estimate the scene boundary regions in the color image; it enables to suppress the influences of color image textures in depth upsampling. We synthesize high-resolution depth images with the help of the estimated scene boundary. Our experiments demonstrate the effectiveness of our method.

Index Terms— Depth upsampling, Depth prediction, Spatio-temporal coherency

1. INTRODUCTION

Accurate depth sensing has been considered as one of the fundamental problems in computer vision research fields. In recent years, depth cameras that can simultaneously capture the color and depth information of a scene (e.g., KINECT) have gained in popularity in industrial and research fields. Using the captured color and depth information, many researchers have developed interesting applications such as: virtual reality [1], 3D object detection [2], etc.

In fact, however, there remain well-known problems in using such depth cameras. Typically, the depth images can only be captured with lower spatial resolution than color images because of the restriction of size of depth sensor. Furthermore, the captured depth images are likely to be deteriorated with noises because of limited active illumination energy. These problems make it difficult to develop the above applications with high quality. Therefore, it is important to establish techniques for obtaining high quality (less-noise) depth images with higher spatial resolution.

To achieve this, methods for upsampling low-resolution depth images have been proposed by many researchers. In particular, they exploited a high-resolution color image, which is simultaneously captured with a RGB-D camera, to synthesize high-resolution depth images [3–10].

Ferstl *et al.* [3] have formulated depth upsampling as a convex optimization problem under the assumption that the structure information in the observed color image are likely to correspond to those of depth image to be synthesized. They then effectively solved their optimization problem with a total generalized variation (TGV) prior. On the other hand, researchers of [5–7] have proposed methods for upsampling low-resolution depth images based on a framework of guided-filter [4] with the help of high-resolution color images.

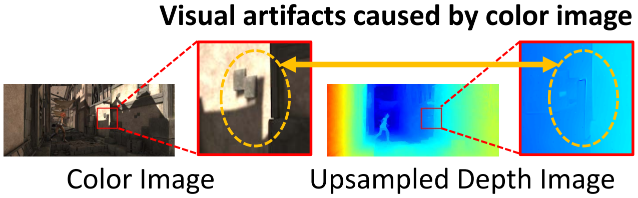


Fig. 1. Problems in typical depth upsampling methods. The left is the observed color image, and the right is the high-resolution depth image synthesized by using a state-of-the-art method [3]. We can see image artifacts on the depth image because of influences of the textures of the color image.

However, these previous methods cause producing image artifacts on the resulting upsampled depth images. This is primarily because they are hard to differentiate the scene boundaries and image textures in the observed color images when performing depth upsampling as shown in Fig.1.

To overcome this problem, we propose a novel approach to depth upsampling by exploiting a spatio-temporal coherency between color and depth image sequences. We can suppose that depth cameras are used for recording the color and depth videos. In that case, temporal consistency in the consecutive frames can be utilized, as well as the spatial coherency between the depth and color images. The temporal consistency of video is useful to predict the images at the next frame. Using this advantage of temporal consistency, we estimate a high-resolution depth image with high quality.

Motion estimation is a promising technique to predict images in the next frame. In the depth image sequence, however, the depth values are unlikely to change in the consecutive frames. It suggests that motion estimation using the depth image sequence would be inadequate to predict reliable images in the subsequent frames. In contrast, accurate motion estimation can be achieved by using the color image sequence, as was done in [11, 12]. Based on the spatial co-

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

herency between the color and depth image, we consider that the motion vectors estimated using the color image sequence would correspond to those in the depth image sequence. We utilize the estimated motion vectors to predict depth images at the subsequent frames.

Using the predicted depth image as a prior, our method enables to estimate only the scene boundaries observed in the color image. It thus achieves suppressing the influences of color image textures in performing depth upsampling.

The main contribution of this study is summarized as follows. We exploit the spatio-temporal coherency between the color and depth image sequences for depth upsampling. Unlike previous methods that were difficult to suppress visual artifacts because of color image textures, the use of the spatio-temporal coherency allows us to exploit only the scene boundary information in the color image for depth upsampling.

2. SCENE BOUNDARY ESTIMATION BY DEPTH PREDICTION

An overview of the proposed method is illustrated in Fig.2. We predict a depth image by using the flow field computed using the consecutive color images. With the predicted depth image, we explore the spatial boundaries of a scene. By exploiting the estimated scene boundaries, we synthesize a high-resolution depth image at the t th frame $\mathbf{D}_t^{\mathrm{H}} = \{d_t^{\mathrm{H}}(\mathbf{p})\}_{\mathbf{p}\in\mathcal{P}^{\mathrm{H}}}$, where $d_t^{\mathrm{H}}(\mathbf{p})$ represents the depth value at the pixel position $\mathbf{p}$, and $\mathcal{P}^{\mathrm{H}}$ denotes a set of pixel positions in the high-resolution depth image.

In this section, we describe the details of our scene boundary estimation.

2.1. Depth Prediction

We predict the depth image at the t th frame. Let $\hat{\mathbf{D}}_t^{\mathrm{H}} = \{\hat{d}_t^{\mathrm{H}}(\mathbf{p})\}_{\mathbf{p}\in\mathcal{P}^{\mathrm{H}}}$ be the predicted depth image at the t th frame. Each component $\hat{d}_t^{\mathrm{H}}(\mathbf{p})$ is the predicted depth value at the pixel position $\mathbf{p}$. Based on the spatial coherency between the color and depth image, we assume that the motions observed in the color image sequence would correspond to those in the depth image sequence. Thus, we predict $\hat{\mathbf{D}}_t^{\mathrm{H}}$ by exploiting the responses from motion estimation using the color image sequence.

Let $\mathbf{I}_t$ be the observed color image at the t th frame. We use a method [11] to compute the flow field $\mathbf{F}_{t-1} = \{\mathbf{f}_{t-1}(\mathbf{p})\}_{\mathbf{p}\in\mathcal{P}^{\mathrm{H}}}$ between $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$. Using the estimated motion vector, $\mathbf{f}_{t-1}(\mathbf{p})$, we predict the corresponding pixel position at the t th frame as: $\hat{\mathbf{p}} = \mathbf{p}+\mathbf{f}_{t-1}(\mathbf{p})$. With the results of this motion prediction, we obtain $\hat{d}_t^{\mathrm{H}}$ as $\hat{d}_t^{\mathrm{H}}(\hat{\mathbf{p}}) = d_{t-1}^{\mathrm{H}}(\mathbf{p})$. For depth values that are not predicted with the motion estimation, we set depth values obtained by interpolating the low-resolution depth image $\mathbf{D}_t^{\mathrm{L}}$. We use a guided filter [4] to interpolate $\mathbf{D}_t^{\mathrm{L}}$. This procedure is illustrated in Fig.3-(i).

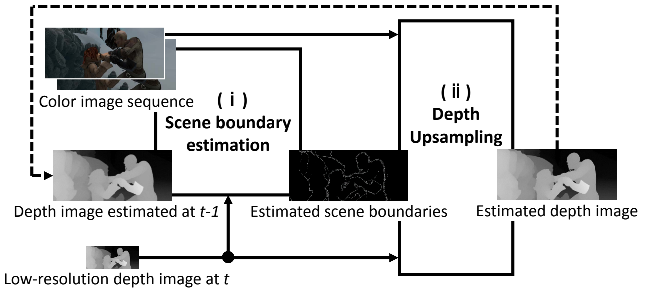

Color image sequence
(i)
Scene boundary estimation
(ii)
Depth Upsampling
Depth image estimated at t-1
Estimated scene boundaries
Estimated depth image
Low-resolution depth image at t

Fig. 2. Overview of the proposed method. (i) Scene boundary estimation. We first compute flow fields using the consecutive color images. We predict a depth image by using the estimated flow field. With the predicted depth image, we explore the scene boundaries in the observed color image. (ii) Depth upsampling. We synthesize a high-resolution depth image by exploiting the predicted scene boundaries.

2.2. Scene Boundary Prediction

Using the predicted depth image $\hat{\mathbf{D}}_t^{\mathrm{H}}$, we construct a binary mask image $\mathbf{M}_t = \{m_t(\mathbf{p})\}_{\mathbf{p}\in\mathcal{P}^{\mathrm{H}}}$ to distinguish the scene boundaries in the depth image at the t th frame. We consider that the scene boundaries will correspond to local regions where the depth values are largely changed. In order to search for such local regions, we consider $r \times r$ pixel-sized small patch centered at $\mathbf{p}$ (we denote this as $\mathbf{R}(\mathbf{p})$). In each local region, we divide the depth values $\{\hat{d}_t^{\mathrm{H}}(\mathbf{p}')\}_{\mathbf{p}'\in\mathbf{R}(\mathbf{p})}$ into two clusters, by using a $K(=2)$-means clustering. We denote the center value in each cluster as $c_1(\mathbf{p})$ and $c_2(\mathbf{p})$. We consider that $\mathbf{p}$ belongs to be a part of scene boundaries where the distance between $c_1(\mathbf{p})$ and $c_2(\mathbf{p})$ is larger than a threshold th. Thus we have $m_t(\mathbf{p})$ representing the positions of scene boundary as

$$m_t(\mathbf{p}) = \begin{cases} 1 & \text{if} \quad |c_1(\mathbf{p}) - c_2(\mathbf{p})| > th \\ 0 & \text{otherwise} \end{cases} . \qquad (1)$$

This procedure is illustrated in Fig.3-(ii).

2.3. Scene Boundary Estimation in the Color Image

Based on the spatial coherency between the depth and color images, $\mathbf{M}_t$ will play an important role as a prior in searching for the scene boundary in the color image $\mathbf{I}_t$. We suppose that the strongest edges in regions masked by $\mathbf{M}_t$ would correspond to the positions of scene boundary in $\mathbf{I}_t$. With this assumption, we obtain a binary mask image representing the scene boundary in $\mathbf{I}_t$, $\mathbf{O}_t = \{o_t(\mathbf{p})\}_{\mathbf{p}\in\mathcal{P}^{\mathrm{H}}}$. Specifically, each

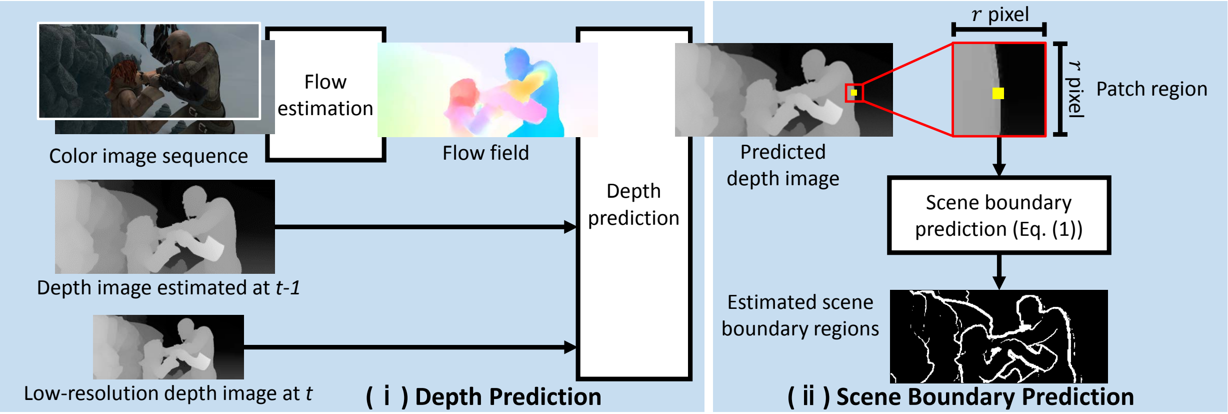

Color image sequence
Flow estimation
Flow field
Depth image estimated at t-1
Low-resolution depth image at t
Depth prediction
(i) Depth Prediction
Predicted depth image
r pixel
r pixel
Patch region
Scene boundary prediction (Eq. (1))
Estimated scene boundary regions
(ii) Scene Boundary Prediction

Fig. 3. Scene boundary prediction. (i) Depth prediction. Exploiting the computed flow field using the consecutive color images, we predict a depth image at the t th frame. (ii) Scene boundary prediction in the depth image. We estimate the pixel positions of scene boundaries by performing a $K(=2)$-means clustering of the predicted depth image.

component $o_t(\mathbf{p})$ is represented as

$$o_t(\mathbf{p}) = \begin{cases} 1 & \text{if} \quad \mathbf{p} = \mathbf{p}^* \\ 0 & \text{otherwise} \end{cases} , \tag{2}$$

where $\mathbf{p}^*$ is the position having the strongest spatial gradients in $\mathbf{I}_t$ that is obtained as

$$\mathbf{p}^* = \arg \max_{\mathbf{p}' \in \mathcal{N}_\mathbf{p}} \left(m(\mathbf{p}') |\nabla \mathbf{I}(\mathbf{p}')| \right) , \tag{3}$$

where $\mathcal{N}_\mathbf{p}$ represents neighboring pixels of pixel position $\mathbf{p}$.

3. DEPTH UPSAMPLING

We synthesize the high-resolution depth image $\mathbf{D}_t^\mathrm{H}$ with the estimated scene boundary regions $\mathbf{O}_t$. To obtain $\mathbf{D}_t^\mathrm{H}$, we solve the following optimization problem,

$$\min_{\mathbf{D}_t^\mathrm{H}} \left(w_1 E_1 \left(\mathbf{D}_t^\mathrm{H} \right) + w_2 E_2 \left(\mathbf{D}_t^\mathrm{H} \right) + w_3 E_\mathrm{TV} \left(\mathbf{D}_t^\mathrm{H} \right) \right) , \tag{4}$$

where w_i $(i = 1, 2, 3)$ is a control parameter.

The first term E_1 is a constraint that models the relation of $\mathbf{D}_t^\mathrm{L}$ and $\mathbf{D}_t^\mathrm{H}$,

$$E_1 \left(\mathbf{D}_t^\mathrm{H} \right) = \| \mathbf{D} \mathbf{B} \mathbf{D}_t^\mathrm{H} - \mathbf{D}_t^\mathrm{L} \|_2^2 , \tag{5}$$

where $\mathbf{B}$ denotes an operator describing a point spread function (PSF), and $\mathbf{D}$ represents a down sampling operator.

The second term E_2 is a constraint that characterizes the correlations of spatial gradients between $\mathbf{D}_t^\mathrm{H}$ and $\mathbf{I}_t$ in the scene boundary regions. Further, E_TV is a regularization term for $\mathbf{D}_t^\mathrm{H}$ based on a total variation prior [13].

We iteratively solve our minimization problem by using an iterative reweighted least squares (IRLS) method [14], similar to [15].

Because one of the key novelties of our method is how the second term E_2 in Eq.(4) is designed, we describe the details of E_2.

3.1. Spatial Correlations in Scene Boundary Regions

We impose a constraint based on spatial correlations between $\mathbf{D}_t^\mathrm{H}$ and $\mathbf{I}_t$. In general, there is a large discrepancy in the magnitude of gradients between the depth and color images. Thus, it is inadequate to use the magnitude of gradients of color image for E_2. In order to address this problem, we exploit the depth image estimated at the $k-1$ th iteration $\mathbf{D}_t^{\mathrm{H},k-1}$ as a guide in estimating $\mathbf{D}_t^\mathrm{H}$. Specifically, we introduce an operator $\mathbf{A}^{k-1} = \{\mathbf{a}^{k-1}(\mathbf{p})\}_{\mathbf{p} \in \mathcal{P}^\mathrm{H}}$ where each component $\mathbf{a}^{k-1}(\mathbf{p})$ is defined as

$$\mathbf{a}^{k-1}(\mathbf{p}) = \sum_{\mathbf{p}' \in \mathcal{N}_\mathbf{p}} \nabla \mathbf{D}_t^{\mathrm{H},k-1}(\mathbf{p}') . \tag{6}$$

We finally model E_2 at the k-th iteration as

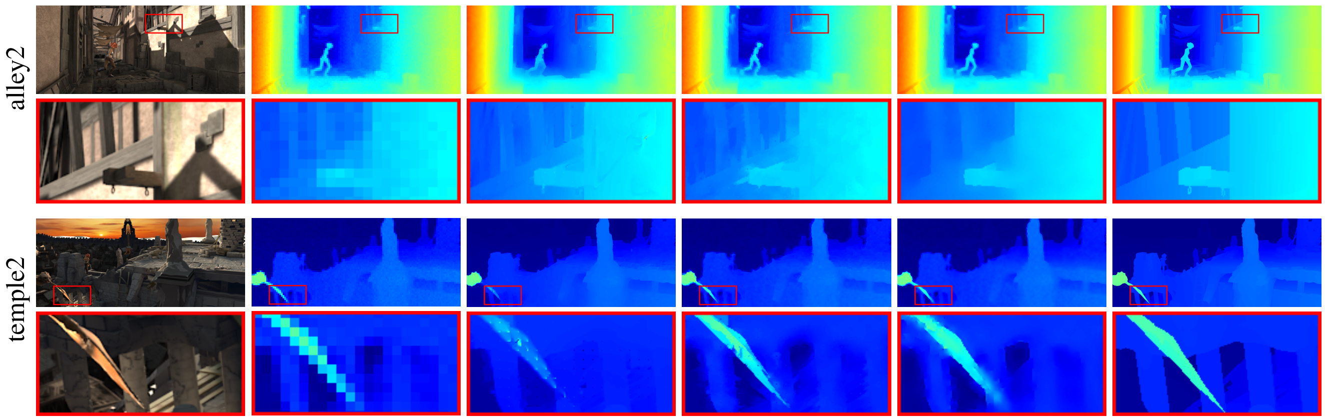
$$E_2 \left(\mathbf{D}_t^\mathrm{H} \right) = \| \mathbf{M}_t \nabla \mathbf{D}_t^\mathrm{H} - \mathbf{A}^{k-1} \mathbf{O}_t \|_2^2 . \tag{7}$$

4. EXPERIMENTS

We show experimental results in terms of quantitative and qualitative evaluations. We tested our method using MPI Sintel datasets [16]. We used 6 sequences in this dataset for the performance evaluation. Using image sequences in this dataset, we made low-resolution depth images with different downsampling factors (×4, ×8). We added a zero-mean Gaussian noise with the standard deviation σ to the low-resolution depth image, in order to simulate image observation process. To analyze noise tolerance of our method, we varied σ from 0 to 2. We used such simulated low-resolution depth image and the original color image as the input data for our computation. Our experiments were run on a Windows PC with an Intel Core i7-6700 3.4 GHz and 16 GB RAM. We used Matlab R2015a for our computation. The processing

Table 1. Quantitative comparisons using a RMSE measure on the MPI Sintel datasets [16]. Best results are represented in **bold**, and the second best are with underline.

Alg.	alley 2						bamboo 1						temple 2						ambush 2					
	×4			×8			×4			×8			×4			×8			×4			×8		
	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2
Ours	0.66	**0.78**	**0.94**	**1.00**	**1.17**	**1.45**	3.27	3.39	**3.52**	**5.75**	**5.90**	**6.07**	0.56	0.76	**0.93**	**1.02**	1.30	**1.56**	**1.33**	**1.59**	**1.92**	**2.90**	3.17	3.51
TGV [3]	**0.61**	0.84	1.08	1.71	1.75	1.90	**3.09**	**3.33**	3.60	7.00	7.13	7.34	**0.48**	**0.75**	1.06	1.06	**1.24**	1.56	2.40	1.67	1.93	3.37	**2.39**	**2.60**
JGU [5]	0.86	1.05	1.39	1.75	1.88	2.16	3.81	4.07	4.41	7.48	7.70	8.01	0.71	1.02	1.47	1.46	1.76	2.21	4.13	2.63	3.01	7.12	5.31	5.65
FGI [7]	1.25	1.25	1.25	2.09	2.08	2.08	4.02	4.01	4.01	7.58	7.56	7.56	0.90	0.90	0.95	1.65	1.65	1.72	2.55	2.68	3.07	5.01	5.21	5.49



(a) Color (b) Low-res. depth (c) TGV [3] (d) JGU [5] (e) Ours (f) Ground truth

Fig. 4. Visual comparisons of depth upsampling results (magnification ratio: ×8) on "alley 2" and "temple 2", respectively.

Table 2. Comparisons of averaged RMSE. Note that these values in this table were obtained by averaging all the results on 6 sequences we used.

Alg.	×4			×8		
	σ=0	σ=1	σ=2	σ=0	σ=1	σ=2
Ours	**1.23**	1.43	1.66	**2.20**	**2.44**	**2.73**
TGV [3]	1.32	**1.39**	**1.64**	2.63	2.57	2.79
JGU [5]	1.89	1.86	2.24	3.34	3.43	3.79
FGI [7]	1.81	1.84	2.06	3.54	3.40	3.58

time for our computation without optimizing implementation is approximately several minutes (2-3 m) per frame.

We utilized the following comparison methods in this evaluation: (i) Total Generalized Variation (TGV) [3], (ii) Joint Geodestic Upsampling (JGU) [5] and (iii) Fast Guided Interpolation (FGI) [7]. In all the experiments, we set the parameters of our method such that w_1 = 1.0,w_2 = 0.5, w_3 = 0.5 and $th = 4$. For the comparison methods, we set their parameters such that the authors reported in the respective papers.

We quantitatively evaluated the results for high-resolution depth image synthesis using a root mean squared error (RMSE) measure. Table 1 lists the comparison results obtained using these methods. Note that these RMSE values in Table 1 were obtained by averaging those from 1 to 10 th frame. Table 2 also shows RMSE comparisons obtained by averaging all the results on 6 sequences we used. We would like to state that our method outperformed state-of-the-art methods. Figure 4 shows visual comparisons on "alley 2" and "temple 2". We can observe that our method was able to suppress visual artifacts owing to the color image textures compared to those obtained using the other methods.

5. CONCLUSION

We proposed a method for depth upsampling by exploiting the spatio-temporal coherency between the color and depth image sequences. This spatio-temporal coherency allowed us to perform a depth image prediction using the responses from motion estimation in the color image sequence. By utilizing the predicted depth image as a prior information, our method enabled to leverage only the scene boundaries in the color image in performing depth upsampling. Through the experiments, we observed the effectiveness of our method.

6. REFERENCES

[1] Jayant Thatte, Jean-Baptiste Boin, Haricharan Lakshman, Gordon Wetzstein, and Bernd Girod, "Depth augmented stereo panorama for cinematic virtual reality with focus cues," in *Proeeding of the IEEE International Conference on Image Processing*, 2016, pp. 1569–1573.

[2] Shuran Song and Jianxiong Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 808–816.

[3] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 993–1000.

[4] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[5] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi, "Joint geodesic upsampling of depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 169–176.

[6] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 96, 2007.

[7] Yu Li, Dongbo Min, Minh N Do, and Jiangbo Lu, "Fast guided global interpolation for depth and motion," in *Proeeding of European Conference on Computer Vision*, 2016, pp. 717–733.

[8] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao 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.

[9] James Diebel and Sebastian Thrun, "An application of markov random fields to range sensing," in *Proceedings of Neural Information Processing Systems*, 2005, vol. 5, pp. 291–298.

[10] Xiaowei Deng and Xiaolin Wu, "Sparsity-based depth image restoration using surface priors and rgb-d correlations," in *Proceeding of the IEEE International Conference on Image Processing*, 2015, pp. 3881–3885.

[11] Yinlin Hu, Rui Song, and Yunsong Li, "Efficient coarse-to-fine patchmatch for large displacement optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5704–5712.

[12] Christoph Vogel, Konrad Schindler, and Stefan Roth, "3d scene flow estimation with a piecewise rigid scene model," *International Journal of Computer Vision*, vol. 115, no. 1, pp. 1–28, 2015.

[13] Leonid I Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[14] Rick Chartrand and Wotao Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.

[15] Xibin Song, Yuchao Dai, and Xueying Qin, "Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network," in *Proceeding of Asian Conference on Computer Vision*, 2016, pp. 360–376.

[16] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black, "A naturalistic open source movie for optical flow evaluation," in *Proceeding of European Conference on Computer Vision*, 2012, pp. 611–625.