Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Depth Map Inpainting Using a Fully Convolutional Network*

Rong Xiong[1,3,4], Guodong Liu[1,2,3,4], Yangyang Qu[1,2,3,4], and Yongsheng Ou[1,3,4,†]

***Abstract*— Emerging consumer depth cameras, such as Microsoft Kinect and ASUS Xtion Pro, are low-cost depth sensors which can capture the depth maps of the scene, and obtaining color images of the corresponding scenes at the same time. Depth information is very important for computer vision applications, and to acquire the accurate depth information is challengeable. However, due to the limitations of imaging principle, depth maps acquired by such devices are often noisy and will lose depth values for certain areas. In this paper, we present a novel method for depth inpainting problem using a Fully Convolutional Network (FCN). The network has two input branches. One branch is a color image that has been processed by a convoluted hidden layer, and the other one is the missing depth map. Besides, we add a shortcut in order to improve the feature extraction capability of the network. We have validated our method in the Middlebury Stereo Dataset and real-world scenes captured by Kinect 2.0. Experiments show that our method has better applicability and reliability than traditional methods in depth inpainting.**

***Index Terms*— Depth inpainting, Fully Convolutional Network (FCN), Kinect**

I. Introduction

Depth inpainting has always been an important part of computer vision. In many research fields and practical applications, depth information is essential and must be accurate, as which obtained by processing low quality input images often lead to serious problems. Specific to the depth map, the pixel values depict the physical distance between the object and the camera in the 3D scene, which plays a vital role in many applications such as 3D reconstruction [1], 3D video [2], human-computer interaction [3] and autonomous navigation.

*This work was jointly supported by National Natural Science Foundation of China (Grant No. U1613210), Guangdong Special Support Program (2017TX04X265), Science and Technology Planning Project of Guangdong Province (2019B090915002), and Shenzhen Fundamental Research Program (JCYJ20170413165528221).

[1]R. Xiong, G. Liu, Y. Qu and Y. Ou are with Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), 1068 Xueyuan Blvd, Shenzhen, China.

[2]G. Liu and Y. Qu are also with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, 1068 Xueyuan Blvd, Shenzhen, China.

[3]R. Xiong, G. Liu, Y. Qu and Y. Ou are also with Guangdong Provincial Key Lab of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences1068 Xueyuan Blvd, Shenzhen, China.

[4]R. Xiong, G. Liu, Y. Qu and Y. Ou are also with CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, 1068 Xueyuan Blvd, Shenzhen, China.

†Yongsheng Ou is the corresponding author ys.ou@siat.ac.cn

(a) (b)

Fig. 1. Data captured by Microsoft Kinect2 (a) RGB image and, (b) corresponding depth map.

The wide use of low-cost and small depth cameras, such as Microsoft Kinect [4], ASUS Xtion Pro and Obbec Astra, has brought a lot of benefit for related research work. However, as shown in Fig. 1, depth maps obtained by such devices often contain noise, small holes and large missing regions. The main reasons for the depth missing are as follows: (1) Due to occlusion, objects close to depth camera may obstruct objects behind it, which will cause depth missing, since infrared light cannot be projected onto the latter; (2) the second situation occurs when the object is transparent, or whose surface is light absorbing material or very smooth; (3) the actual measured distance exceeds the effective distance limit of the depth camera, which causes depth missing when the object is too close or too far. For example, the limit of effective measurement distance of Kinect2 is $0.5 \sim 4.5$m. Practice has shown that the depth value can be obtained normally in the range of $0.5 \sim 8.0$m, which means the depth value outside this range cannot be obtained. To address the above problems, one way is to enhance depth sensing technology by upgrading the hardware. Although depth sensing technology has been developed over these years, the quality of depth maps is not comparable to the color images. On the other hand, the captured depth data can be processed by using depth inpainting and image restoration algorithms, which can obtain high quality depth images without increasing depth sensor accuracy.

Over the past years, researchers have used a variety of traditional methods for image restoration and depth map inpainting. In recent years, the application of neural networks for depth inpainting has been a big step forward. In this paper, we proposed a novel inpainting method based on Fully Convolutional Network (FCN), which can restore large empty areas in depth maps, as well as small noise areas. The remainder of this paper is organized as follows. Section II

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

introduces plenty of previous work in the image restoration and depth inpainting. Section III presents our method of the overall network architecture and network parameters. And then the manufacturing and augmentation of the dataset are proposed in Section IV. Besides, the comparison experiment between our networks and the traditional methods is carried out in this section. In addition, we verified our method on the real-world depth maps. Section V draws the conclusion.

II. Related Work

Depth inpainting is an ill-posed problem that requires additional information to achieve satisfactory performance. The technique of multi-depth image sequence fusion has attracted more and more attention in recent years, and many methods attempt to combine multi-frame depth maps into a higher quality depth map. Schuon et al. [5] proposed a method that combines several low resolution noisy depth images of a static scene from slightly displaced viewpoints, and merges them into a high-resolution depth image. Hahne et al. [6] combined several exposures with varying integration times of a time-of-flight camera in order to enhance the quality of the depth maps. Choi et al. [7] proposed to improve the quality of multi-view depth maps by improving spatial resolution and enhancing coherence. This method can simultaneously suppress noise and is also sensitive to occlusion. However, it ignores the effect of the depth map on the composite view. Due to the limited information used, this category of methods tends to produce blurred artifacts or jagged artifacts. Telea [8] proposed a fast matching method (FMM), which applied the idea of level set to image patching and achieved good results. This method is very simple and fast to implement, but the blurring will be produced when inpainting regions thicker than $10 \sim 15$ pixels.

Another research direction is to use color images to guide the depth inpainting. The associated high quality color images are used to improve the low precision of depth maps. This research direction can be divided into two categories: filter-based methods and color-guided optimization methods. Common filtering methods, such as Bilateral Filtering (BF) [9], Joint Bilateral Filtering (JBF) [10], Joint Trilateral Filtering (JTF) [11], Guided Image Filtering [12], etc. In filtering methods, the pioneering work is joint bilateral filtering upsampling (JBU) [13], where the bilateral weight is based on guided color images. Filtering methods are convenient and simple for depth inpainting, but the results are not satisfactory for large missing regions. The color-guided optimization methods are also gradually used for depth inpainting. Liu et al. [14] proposed depth inpainting by color-guided fast marching method (GFFM) which extends the original FMM. This method improved the efficiency and quality of depth inpainting to a certain extent, but the repair ability for large areas is still limited. Yang et al. [15], [16] proposed a RGB-D recovery scheme based on autoregressive (AR) model. The AR model has achieved great success in natural image interpolation. In the AR model calculation, color information is introduced as a guide to improve the robustness against noise. However, this method only considered the local depth map smoothness and cannot well preserve the edges of large-scale missing regions.

III. Methodology

In recent years, deep learning has achieved great success in the field of image restoration. Our method will be based on a fully convolutional network for depth inpainting, which will consist of the following steps.

A. Mask map generation

One of the difficulties in depth inpainting is that there is no ground truth of the missing regions, which bring some obstacles for the method based on deep learning. In this paper, we artificially create some missing regions in the non-missing position of the depth map, and the original values at these regions are used as the ground truth for training the network. In this paper, the process of deep inpainting is divided into two stages: fitting on dataset and validating on scenic images.

The main difference between these two stages is the generation way of the depth mask map. In the dataset fitting stage, the depth values of the artificially created regions are set to the maximum value in order to distinguish between the artificial missing regions and the missing areas of the images themselves. When generating the mask map, the pixel value of the mask map is set to 1 at the position corresponding to the maximum depth value of the depth map, and the rest of the mask map is 0. However, in the real scene validation stage, the pixel value of the mask map is set to 1 at the position corresponding to the minimum depth value (i.e., the missing position), and the rest of the mask map is 0. The mask maps will be used to calculate the loss in this paper.

B. Network structure

Depth inpainting can be regarded as a regression problem, which can perform pixel-by-pixel regression for the missing regions based on the alignment of the RGB color image and the missing depth map. Since the fully convolutional network can maintain the spatial position information of the pixel [17], the pixel-to-pixel classification or regression problem can be effectively completed. This paper proposes a fully convolutional network with dual-branch input, which consists of five layers of convolution hidden layer and one layer of deconvolution output layer. The network structure is shown in Fig. 2. The network has two input branches. One is RGB color image and the other is the missing depth map. Note that the RGB color image is combined with the missing depth map after passing through a convolutional hidden layer. This has two advantages: first, the resolution of the color image can be reduced to the same as that of the depth map

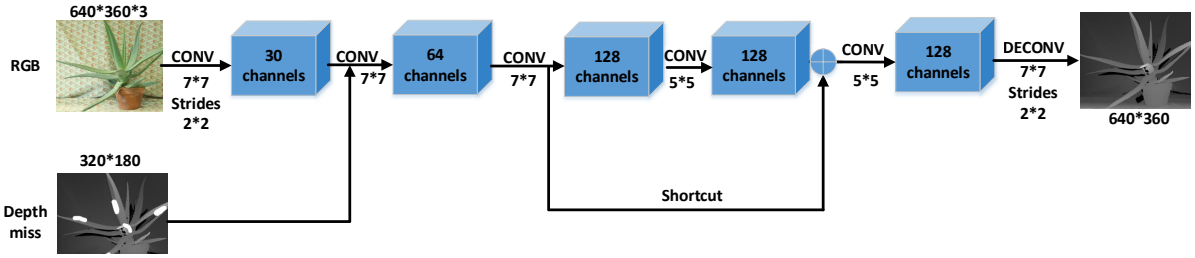

Fig. 2. Overview of network structure.

through the first hidden layer, while retaining its own useful information. Second, the network can automatically learn the weights of color images in the training process, which is similar to the frequency domain attention [18]. The following convolutional layers are used to extract features. In order to make the features extraction more effective, we added a shortcut from the third to the fifth layer. The shortcut is inspired by ResNet [19], which makes the hidden layer in the shortcut only need to learn the residuals. By adopting this idea, feature extraction ability of the network can be enhanced. Each hidden layer contains Batch Normalization (BN) [20], which can accelerate the convergence speed of the network, increase the robustness of the network, and improve the final results. The real depth information is very important. In order to make use of such information, the receptive fields of hidden layers in the convolutional network needs to be large enough. Therefore, in our network, the convolution kernel size of each hidden layer is slightly larger than that of general networks. Last but not least, a deconvolution output layer restores the resolution of the output depth map to that of the input RGB image.

From the network structure, we can see that the output of the network is still a whole image, so a step needs to be added after the network processing is completed: copy the pixel value of the network output image at the mask regions to the corresponding position of the missing depth map.

All the operations and parameters in the network are shown in TABLE I.

TABLE I
THE OPERATIONS IN THE NETWORK

Layers	Operations	Parameters
Hidden Layer 1	Convolution	Kernel=7*7*30, Strides=2*2
	Batch normalization	/
	Activate function	Leaky ReLU
Hidden Layer 2	Concatenate	Depth image
	Convolution	Kernel=7*7*64
	Batch normalization	
	Activate function	Leaky ReLU
Hidden Layer 3	Convolution	Kernel=7*7* 128
	Batch normalization	
	Activate function	Leaky ReLU
Hidden Layer 4	Convolution	Kernel=5*5*128
	Batch normalization	
	Activate function	Leaky ReLU
Hidden Layer 5	Convolution	Kernel=5*5*128
	Batch normalization	
	Shortcut	Layer3+BN5
	Activate function	Leaky ReLU
Output Layer	Deconvolution	Kernel=7*7*1
		Strides=2*2

C. Loss function

We calculate the loss for the missing regions by inputting the mask map of the loss function, and do not calculate the position where the ground truth is 0 (the value of 0 means the ground truth at this position is wrong). The loss of each pixel is shown in Eqn (1):

$$loss_{eachpix} = \begin{cases} 0 & \text{if } y \leq 0 \\ p_{mask}(y^* - y)^2 & \text{if } y > 0 \end{cases} \quad (1)$$

where y is the ground truth, and $y*$ represents the predict results of the network, and p_{mask} refers to the value of corresponding position on the mask map. N means the total number of calculation points in each image where $y > 0$, and the average loss is calculated by Eqn (2):

$$loss = \frac{\sum loss_{eachpix}}{N} \quad (2)$$

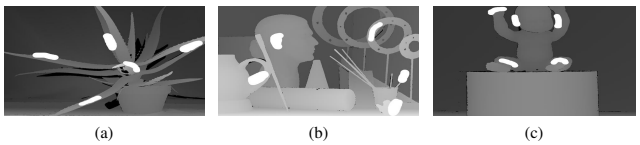

Fig. 3. Examples of man-made missing regions for the Middlebury Stereo Datasets.

IV. EXPERIMENTS

A. Make Dataset

As mentioned above, we artificially add several holes at the Non-missing area of the depth graph based on Middlebury Stereo Datasets (MSD) [21], [22], [23], which generates the dataset of our task. Especially, the positions and shapes of man-made missing regions are very important, which must

be similar to those of real missing regions. The closer the artificial simulation is, the better the network trained from the dataset can be generalized to the real scene restoration. Since the depth map in the dataset itself contains few missing areas, which are the black part in the images, we choose white as the color of the artificially added missing regions to distinguish the missing regions of the image itself. Fig. 3 shows several examples of man-made missing regions.

B. Data Augmentation

Due to the limitation of the number of images in the dataset, the following 7 methods are adopted in this paper for data augmentation.

(1) Enlarge the image at random scale s between 1.1 and 1.5, and cut out the original size at random position in the enlarged image. And the depths are divided by s.

(2) Perform a random rotation transformation r of $-10^\circ \sim +10^\circ$ on the image.

(3) Randomly change the overall brightness of the image by $0.8 \sim 1.2$ times.

(4) Add Gaussian white noise with a standard deviation of 0.002 to the image.

(5) Add random Gaussian Blur to the image with $sigma = 1.1 \sim 1.8$.

(6) Sharpen the image by Laplacian with a 75% probability.

(7) Horizontal flips with a 75% probability.

Note that the operations (1), (2), (7) are performed on both color images and depth maps, whereas the operations $(3) \sim (6)$ are just for color images.

C. Training

We validate our approach and compare it with other methods using both the Middlebury Stereo datasets and real-world Kinect 2 data. In this paper, the dataset is divided into 80% training set and 20% test set. We set a small learning rate, and set a large batch size due to the existence of BN. Our inpainting algorithm takes the training on a dual Intel Xeon E5 server. And then test the results on the test set.

D. Results

Fig. 4 shows the results of our depth inpainting method on the Middlebury Stereo datasets 2005 & 2006. The first column is color images, the second column is the missing maps, the third column is output of our entire network, and the last column is the ground truth. It can be seen from these examples that the inpainting of the artificial missing regions basically achieves satisfactory results, which can not only recover the large missing regions, but also retain the edge information of the missing regions. Since the size of the image is reduced after typesetting, the difference between the recovered depth map and ground truth seems not obvious. In Fig. 5, we have selected a partially enlarged detail map to see the results of the experiment more clearly. We also have a quantitative comparison between our methods with JBF and FMM on the MSD. We evaluate each method using four error functions from prior work [24], including absolute relative difference, squared relative difference, root mean square error (RMSE(linear)) and root mean square error log (RMSE (log)). We only calculate these errors for the missing regions by using a recovered depth map and ground truth. Table II shows the results of comparison of the various methods on the MSD. The calculated data show that our method performs better than other methods.

We further verified our method on real-word depth maps acquired by Kinect2. Fig. 6 shows the results of the depth maps inpainting by using aligned color images.

TABLE II
COMPARISONS ON MSD WITH PREVIOUS WORK

	JBF	FMM	Our Network
Abs relative difference	0.714	0.790	0.188
Squared relative difference	5.284	5.897	1.603
RMSE(linear)	35.019	36.556	7.359
RMSE(log)	0.820	0.899	0.276

V. CONCLUSION

In this paper, we have presented a novel approach to depth inpainting problem. Different from traditional filtering methods, a fully convolution network was proposed for depth inpainting. The network has two input branches. One is the RGB map and the other one is the missing depth map. The RGB map is combined with the missing depth map after a convolutional hidden layer. The missing depth map was generated by artificially manufacturing several missing regions, and a shortcut was added to increase the convergence speed and robustness of the network. We validated our algorithm on the MSD and the real-world scenes from Kinect 2. Experimental results showed that our method gives good performance on noises, small holes and large missing areas. Furthermore, the edge details are better retained.

However, our method also has some shortcomings. Firstly, the design of missing location is not ideal which needs to be improved in the future. Secondly, transfer learning can be applied to the network. Specifically, the dataset applied in this paper can be improved by another appropriate one acquired from a specific camera.

REFERENCES

[1] A. Anwer, S. S. Azhar Ali, A. Khan, and F. Mriaudeau, "Underwater 3-d scene reconstruction using kinect v2 based on physical models for refraction and time of flight correction," *IEEE Access*, vol. 5, pp. 15 960–15 970, 2017.

[2] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Stereoscopic Displays and Virtual Reality Systems XI*, M. T. Bolas, A. J. Woods, J. O. Merritt, and S. A. Benton, Eds., vol. 5291, International Society for Optics and Photonics. SPIE, 2004, pp. 93 – 104. [Online]. Available: https://doi.org/10.1117/12.524762

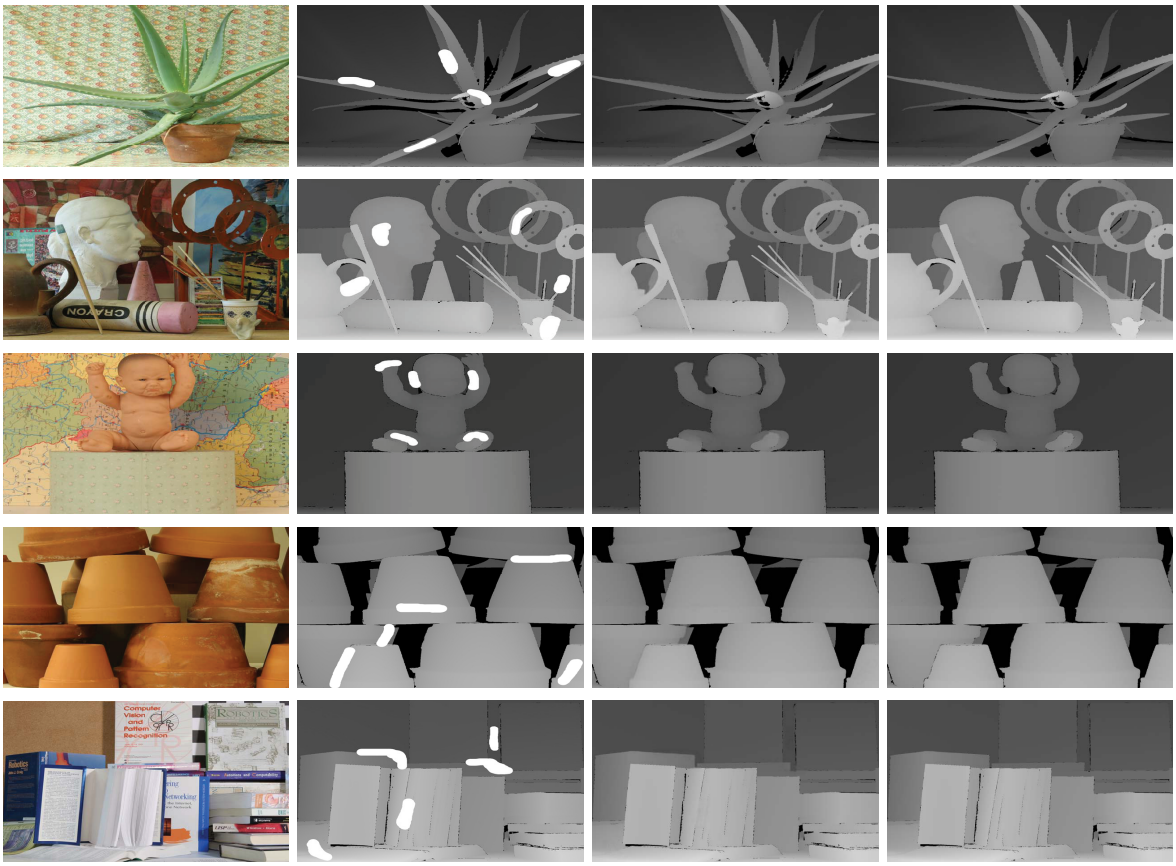

Fig. 4. Result in Middlebury Stereo Datasets 2005 & 2006. The first columns are color maps. The second columns are the missing maps. The third columns are output of our entire network. The last columns are the ground truth. We mainly compare the repaired results of the white area between the third column and the fourth column. It can be seen from the comparison that the repaired part is very close to ground truth.

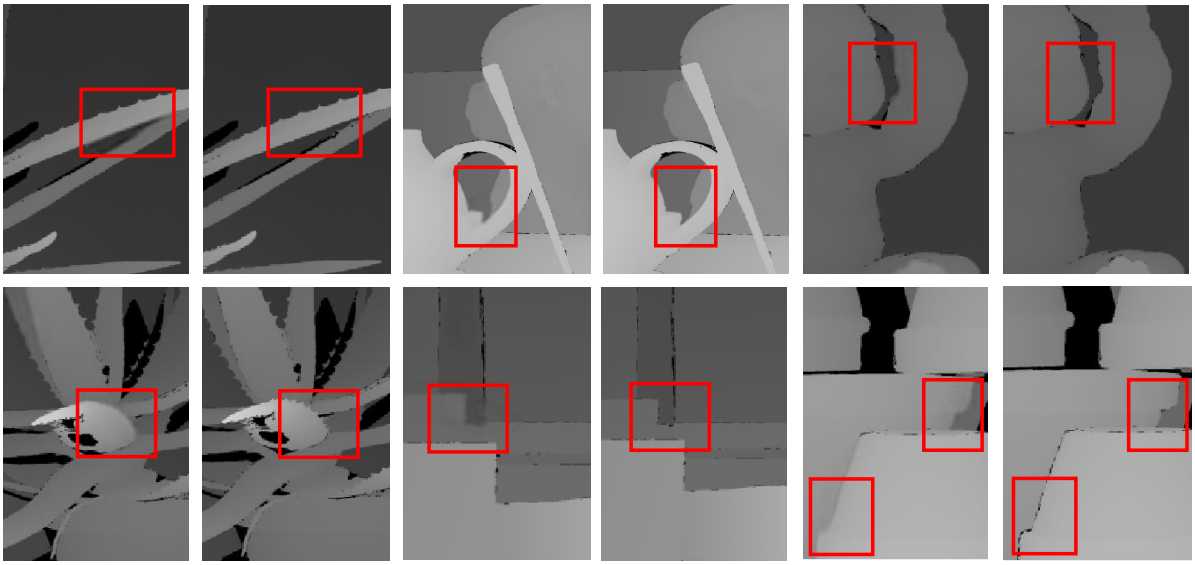

Fig. 5. Details of the depth inpainting result. For each group, the left is the result of our network and the right is ground truth.

(a) (b) (c)

Fig. 6. Inpainting result of real-word depth maps acquired by Kinect 2. (a) The RGB image. (b) The corresponding depth map. (c) Our Inpainting result.

[3] S. Saha, R. Lahiri, A. Konar, B. Banerjee, and A. K. Nagar, "Hmm-based gesture recognition system using kinect sensor for improvised human-computer interaction," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 2776–2783.

[4] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb 2012.

[5] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 343–350.

[6] U. Hahne and M. Alexa, "Exposure fusion for time-of-flight imaging," *Comput. Graph. Forum*, vol. 30, pp. 1887–1894, 2011.

[7] J. Choi, D. Min, and K. Sohn, "Reliability-based multiview depth enhancement considering interview coherence," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 603–616, April 2014.

[8] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004. [Online]. Available: https://doi.org/10.1080/10867651.2004.10487596

[9] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Jan 1998, pp. 839–846.

[10] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 664–672, Aug. 2004. [Online]. Available: http://doi.acm.org/10.1145/1015706.1015777

[11] K. Lo, Y. F. Wang, and K. Hua, "Joint trilateral filtering for depth map super-resolution," in *2013 Visual Communications and Image Processing (VCIP)*, Nov 2013, pp. 1–6.

[12] 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, June 2013.

[13] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007)*, vol. 26. Association for Computing Machinery, Inc., August 2007. [Online]. Available: https://www.microsoft.com/en-us/research/publication/joint-bilateral-upsampling/

[14] J. Liu, X. Gong, and J. Liu, "Guided inpainting and filtering for kinect depth maps," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 2055–2058.

[15] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *ECCV*, 2012.

[16] 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, Aug 2014.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 7132–7141.

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

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

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

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

[23] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition (GCPR 2014*, 2014.

[24] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *2014 Neural Information Processing Systems NIPS*, 2014.