SINGLE DEPTH IMAGE SUPER-RESOLUTION USING CONVOLUTIONAL NEURAL NETWORKS

Baoliang Chen and Cheolkon Jung

School of Electronic Engineering, Xidian University, Xian, Shaanxi 710071, China
zhengzk@xidian.edu.cn

ABSTRACT

In this paper, we propose single depth image super-resolution using convolutional neural networks (CNN). We adopt CNN to acquire a high-quality edge map from the input low-resolution (LR) depth image. We use the high-quality edge map as the weight of the regularization term in a total variation (TV) model for super-resolution. First, we interpolate the LR depth image using bicubic interpolation and extract its low-quality edge map. Then, we get the high-quality edge map from the low-quality one using CNN. Since the CNN output often contains broken edges and holes, we refine it using the low-quality edge map. Guided by the high-quality edge map, we upsample the input LR depth image in the TV model. The edge-based guidance in TV effectively removes noise in depth while minimizing jagged artifacts and preserving sharp edges. Various experiments on the Middlebury stereo dataset and Laser Scan dataset demonstrate the superiority of the proposed method over state-of-the-arts in both qualitative and quantitative measurements.

Index Terms— Convolutional neural networks, depth image, edge-guided, super-resolution, total variation.

1. INTRODUCTION

Depth estimation from natural scenes is a challenging task in computer vision. Various depth cameras have been developed for depth acquisition including time-of-flight (TOF) and light-coded cameras. However, depth images captured by these cameras suffer from the limited spatial resolution and much noise. Thus, depth image super-resolution (DISR) has received much attention by researchers. Most DISR methods have used a HR intensity image [1, 2, 3, 4, 5] as an assistant. However, their performance are heavily depending on the assumption that its HR color registered with the depth image should be available, which may not be practical for many applications. Thus, single DISR is more practical, which offers challenges compared to color image-based DISR [6]-[7]. Aodha et al. [8] employed a patch-based Markov random field (MRF) model in DISR. Hornacek et al. [9] proposed to search low and high resolution patch pairs via a rigid body transformation. Xie et al. [10] proposed coupled dictionary learning with a regularized shock filter to reduce jagged noise while sharpening edges. Ferstl et al. [11] generated HR depth edges by learning a dictionary that contains edge priors. Inspired by the edge guidance, Xie et al. [12] constructed HR edge map through an MRF optimization through patch synthesis.

In this paper, we propose single DISR using CNN. We acquire the high-quality edge map from LR depth image based on CNN. Guided by the high-quality edge map, we perform SR reconstruction of the LR depth image using a total variation (TV) model. First, we interpolate the input LR depth image by bicubic interpolation. The interpolation result is blurry and its edge information is not clear. Then, we get its edge map by canny edge detection and extend the edge region. This is because real edges are mostly located at the extended regions. Next, we extract a window around pixels in the edge region as the first input and the corresponding window in the interpolated depth map as the other input. We use CNN on the two inputs to classify whether the center pixel is a real edge or not, thus resulting in the high-quality edge map. Since the CNN output contains broken edges and holes, we refine the edge map with the help of the low-quality edge map. Finally, we use a TV model for depth upsampling that consists of the fidelity and regularization terms. In the TV model, the high-quality edge map is used to adjust the weight of the regularization term. Compared with existing methods, main contributions of this paper are as follows:

- We propose a novel CNN architecture to acquire the high-quality edge map from the low-quality one.
- We utilize the low-quality edge map to connect broken edges and fill holes in the edge map.
- We use the high-quality edge map to adjust the weight of the regularization term in TV.

2. PROPOSED METHOD

Fig. 1 illustrates the entire diagram of the proposed depth image super-resolution method. The proposed method consists

This work was supported by the National Natural Science Foundation of China (No. 61271298) and the International S&T Cooperation Program of China (No. 2014DFG12780).

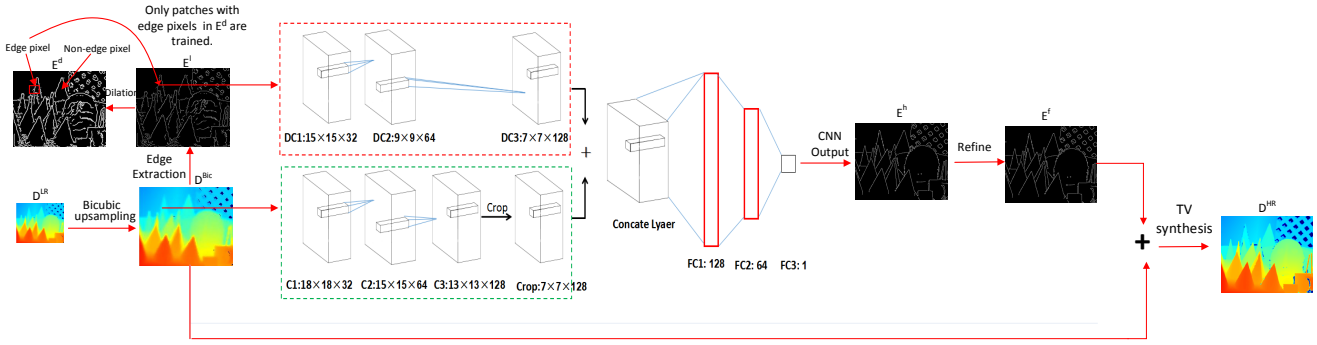

Fig. 1: Entire diagram of the proposed depth image super-resolution method. D^{LR}: Input LR depth image. D^{Bic}: Interpolated depth image by bicubic interpolation. E^l: Low-quality edge map from D^{Bic}. E^d: Dilated edge map from E^l. **DC**: Dilated convolution layer. **C**: Convolution layer. **FC**: Fully connected (FC) layer. TV: Total variation. E^h: High-quality edge map acquired by CNN. E^f: Refined edge map. D^{HR}: Final HR depth image.

of high-quality edge map acquisition, edge map refinement, and edge-guided depth upsampling. Deep conventional neural networks are used to get high-quality edge maps, and the high-quality edge map is used as the weight of the regularization term in the TV model.

2.1. Construction of Training Dataset

We first acquire E^d by dilating the edge region in E^l for each pixel p in E^d. If it belongs to the dilated edge region, then we perform the deep CNN to determine whether the pixel is edge or not. If the pixel does not belong to the dilated edge region, then it is treated as the non-edge pixel in E^h. To construct the training dataset, we extract 65 depth images from Middlebury stereo dataset [13, 14] and Laser Scan dataset [8] provided by Aodha et al. [8]. Among them, 11 depth images are selected as the test images. For each dataset, we first downsample depth images by the scale $S(\times 2,\ \times 4)$ and add an standard Gaussian noise to simulate the initial LR depth image D^{LR}. Then, we interpolate them by bicubic interpolation D^{Bic} and extract its edge map using canny edge detector. Next, we dilate the edge map with a $S \times S$ square kernel. Finally, we extract a patch of size 21×21 around each pixel in the dilated edge region as the first input and its corresponding region in the interpolated depth map as the other input. The label of the input is acquired by the binary value of the center pixel in the edge map. We extract it from the ground truth depth image. After extracting all patches, we randomly select 350000 sub-images for training and 15000 sub-images for testing.

2.2. High-Quality Edge Map Acquisition

Deep convolution neural networks are to do a binary classification. Two inputs of the network is low-quality edge patch and interpolated depth patch, while the output is labels of 1 (edge) and 0 (non-edge). As illustrated in Fig. 1, we use 3 dilated convolution (DC) layers consist of 32 kernels of size $5 \times 5 \times 1$, 64 kernels of size $5 \times 5 \times 32$ and 128 kernels of size $3 \times 3 \times 64$ for the first input, i.e. low-quality edge patch. To accelerate the training process, we add a batch normalization (BN) layer [15] after each DC layer, and the activation function is rectified linear units (ReLU) [16]. We denote the output of the each layer as $DC1$, $DC2$ and $DC3$. This is because we use DC layers to contain most possible edge points in the patch, but not to expand the receptive field of the network. If the receptive field is small, high level features would not be learned. For the second input, i.e. interpolated depth patch, we use 3 convolution layers that consist of 32 kernels of size $5 \times 5 \times 1$, 64 kernels of size $5 \times 5 \times 32$ and 128 kernels of size $3 \times 3 \times 64$. The same as the first input, we add BN layer and ReLU after each convolution layer. The output of the each layer we denoted as $C1$, $C2$ and $C3$. We do not use Pooling layers on the two inputs because the proposed network is pixel-based classification network and the pixel information would be lost when the input maps are downsampled by pooling layers. Since $C3$ and $DC3$ do not have the same size, we crop $C3$ to the same size as $DC1$. Denote the output as Crop1. To synthesize the information of $C3$ and Crop1, we add fully connected (FC) layers denoted as $FC1$, $FC2$, $FC3$ with 1024, 512, and 2 neurons, respectively. Every layer is followed by ReLU except the last layer. We assign a softmax layer to the last layer to get a confidence score which indicates the probability to be edge. During the training process, we minimize binary cross-entropy loss with respect to the weight w that parameterizes the network as follows:

$$\min_{w} \quad -[y_{gt} \log p(y_{gt}, w) + (1 - y_{gt}) \log(1 - p(y_{gt}, w))] \tag{1}$$

where y_{gt} is the binary label value; and $p(y_{gt}, w)$ is the output that indicates the probability to be edge. We train our network using stochastic gradient descent back propagation with AdaGrad [17]. Similar to the moment-based stochastic

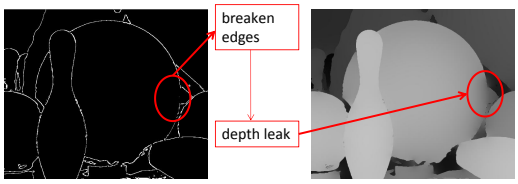

Fig. 2: Left: Edge map with broken edges. Right: Depth SR by the edge map.

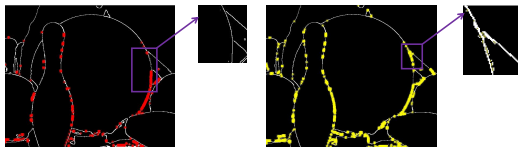

Fig. 3: Left: Connected edges (red: new edge pixels). Right: Filled holes (yellow: new edge pixels).

gradient descent, AdaGrad adapts the gradient based on historical information. Compared with moment-based methods, it emphasizes rare but informative features.

2.3. Edge Map Refinement

Based on CNN, we acquire an high-quality edge map E^h. However, the CNN output includes broken edges and some holes between edges, which may cause depth leak in the SR reconstruction as shown in Fig. 2. Thus, we refine the edge map using LR depth image. Denote the refined edge map as E^f. To refine E^h, we first detect the broken edges and then connect them using E^l. For each pixel p in E^h, if $E^l(p)$ is 1, i.e. an edge pixel, then we extract matrices M_l and M_h of size $S \times S$ in E_l and E_h, where S is the upsampling factor. Finally, we perform AND operation on M_l and M_h. If the sum of AND operation result is zero, then we assign the value of p to 1, otherwise, it remains its value as follows:

$$E^f(p) = \begin{cases} E^l(p) & if\ E^l(p) = 1\ and\ Sum\ (Ml\ \&\ Mh) = 0\ ; \\ E^h(p) & else\ ; \end{cases} \tag{2}$$

By this process, the depth leak caused by broken edges is repaired show as shown in Fig. 3. Moreover, there exist small holes between continuous edge in the CNN output. As shown in Fig. 4, we use four 3×3 edge patterns denoted as $P1$, $P2$, $P3$, and $P4$. To find holes, we extract $P_{h,p}$ of size 3×3 around each non-edge pixel p in E^h, and perform AND operation using four patterns. If all four results denoted as $R1$, $R2$, $R3$, and $R4$ are smaller than 2, then it is truly a non-edge point, otherwise, this pixel is a hole, we assign the value of p to 1 as follows:

$$E^f(p) = \begin{cases} 1 & if\ E^h(p) = 0\ and\ ((R1 > 2)\ or \\ & (R2 > 2)\ or\ (R3 > 2)\ or\ (R4 > 2)); \\ E^h(p) & else\ ; \end{cases} \tag{3}$$

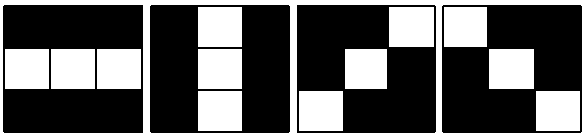

Fig. 4: Edge patterns $P1$, $P2$, $P3$ and $P4$.

Connected edges and filled holes are shown in Fig. 3.

2.4. Edge-Guided Depth Upsampling

Guided by the edge map E^f, we use a variational approach to get the depth SR image D^R. Our variational approach includes two terms of the fidelity and regularization terms as follows:

$$\min_{D^R} E(D^R) = E(D^{Bic}) + \lambda R_{\text{smooth}} \tag{4}$$

where λ is a fixed parameter to adjust two terms. The fidelity term D^{Bic} makes the depth SR result be close to the input LR depth map as follows:

$$E(D^{bic}) = \sum_p \left(D^R(p) - D^{Bic}(p) \right)^2 \tag{5}$$

The regularization term ensures the smoothness of the depth SR result based on total variation (TV) as follows:

$$R_{\text{smooth}} = \sum_p E^l(p) \left[|\partial_x(D^R)|^2 + |\partial_y(D^R)|^2 \right] \tag{6}$$

We use the high-quality edge map as the binary weights of the regularization term. (4) is rewritten into a matrix form as follows:

$$\begin{aligned} \min_{D^R} \ & \left[(D^R - D^{Bic})^T (D^R - D^{Bic}) \right] \\ & + \lambda E^l \begin{bmatrix} \left((\partial_x D^R)^T (\partial_x D^R) \right) \\ + \left((\partial_y D^R)^T (\partial_y D^R) \right) \end{bmatrix} \end{aligned} \tag{7}$$

We get the SR reconstruction result as follows:

$$D^R = (I + \lambda D_x + \lambda D_y)^{-1} * D^{Bic} \tag{8}$$

where D_x and D_y are derivative operators in horizontal and vertical directions, respectively.

3. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed method quantitatively and qualitatively in comparison with state-of-the-art SR methods. We perform experiments on publicly available datasets including Middlebury stereo dataset [14, 21, 22, 23] and the Laser Scan dataset provided by Aodha et al. [8]. We set $\lambda = 10$ in (4) in all experiments. We implement the proposed method on a PC with an Intel I7-6700 3.40 GHz CPU

Table 1: RMSE Evaluation Results

Method	x4				x4		
	Cones	Teddy	Tsukuba	Venus	Scan21	Scan30	Scan42
NN	6.0054	4.5466	12.9083	2.9333	2.6474	2.5196	5.6044
Bicubic	3.8635	2.893	8.7103	1.9403	2.0324	1.9764	4.5813
Park et al.[3]	6.5447	4.3366	12.1231	2.2595	*N/A*	*N/A*	*N/A*
Yang et al.[18]	5.139	4.066	13.1748	2.7559	*N/A*	*N/A*	*N/A*
Ferstl et al.[4]	3.9968	2.808	10.0352	1.6643	*N/A*	*N/A*	*N/A*
NE+NNLS	3.4362	2.4887	7.5344	1.6291	1.7313	1.6849	3.5733
SRCNN [19]	4.219	2.456	8.6643	1.9717	1.6732	1.5141	2.783
Aodha et al.[8]	12.6938	4.1113	12.6938	2.6497	2.5983	2.6267	6.1871
Hornacck ct al.[9]	5.4898	5.0212	11.1101	3.5833	2.8585	2.7243	4.5074
Ferstl et al.[11]	3.568	2.6474	7.5356	1.7771	1.4349	1.4298	3.141
Xie et al.[12]	4.4087	3.2768	9.7765	2.3714	1.3993	1.4101	2.691
Proposed	**3.1742**	**2.1357**	**6.3472**	**0.9955**	**1.2453**	**1.0022**	**1.5431**

Table 2: SSIM Evaluation Results

Method	X4				X4		
	Cones	Teddy	Tsukuba	Venus	Scan21	Scan30	Scan42
NN	0.936	0.945	0.9003	0.98	0.9814	0.9828	0.9679
Bicubic	0.9538	0.9619	0.9205	0.9845	0.9875	0.9879	0.9743
Park et al.[3]	0.942	0.9553	0.8981	0.9862	*N/A*	*N/A*	*N/A*
Yang et al. [18]	0.9624	0.9695	0.9314	0.9879	*N/A*	*N/A*	*N/A*
Ferstl et al.[4]	0.9625	0.9707	0.9245	0.9901	*N/A*	*N/A*	*N/A*
NE+NNLS	0.9424	0.9499	0.8872	0.982	0.9896	0.99	0.9805
SRCNN [19]	0.9379	0.9408	0.8932	0.9766	0.9843	0.9853	0.9822
Aodha et al. [8]	0.9392	0.952	0.908	0.9822	0.9838	0.9838	0.9668
Hornacck ct al. [9]	0.9501	0.9503	0.9137	0.9789	0.9814	0.9825	0.9754
Ferstl et al. [11]	0.9645	0.9716	0.9413	0.9893	0.9918	0.9916	0.9819
Xie et al. [12]	0.9319	0.9331	0.8822	0.973	0.9869	0.9878	0.9899
Proposed	**0.9711**	**0.9883**	**0.9599**	**0.9897**	**0.9959**	**0.997**	**0.9948**

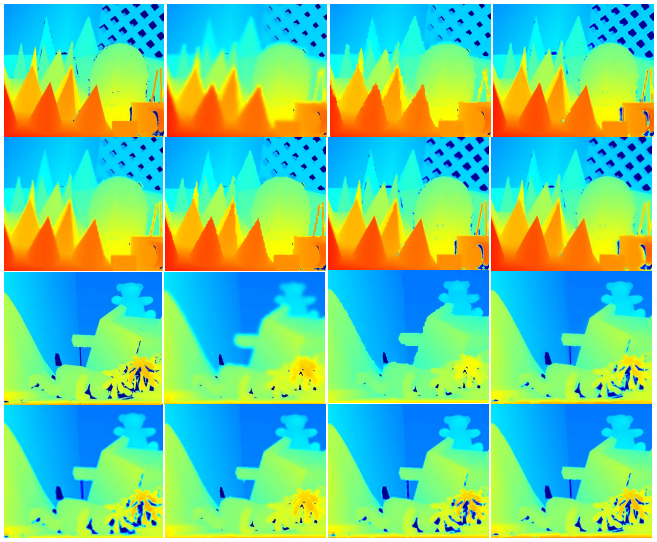

Fig. 5: Depth SR reconstruction results in *Cones* and *Teddy* when the upsampling factor is 4. Up to down, left to right: Ground truth, Bicubic interpolation, Park et al. [3], ANR [20], SRCNN [19], Xie et al. [12], Ferstl et al. [11], and Proposed method.

and 8 GB RAM using Matlab2015b and Tensorflow. We compare our results with three groups of methods: 1) Single depth image SR methods such as SRCNN [19], Ferstl et al. [11] and Xie et al. [12]; 2) Color assisted depth image SR approaches such as Park et al. [3], Yang et al. [18], and Ferstl et al. [4]; 3) Bicubic interpolation.

We provide RMSE and SSIM evaluation results on Middlebury dataset *Cones*, *Teddy*, *Tsukuba* and *Venus*) and Laser Scan dataset (*Scan21*, *Scan30* and *Scan42*) for up-sampling factors $\times 4$ in Tables 1 and 2. As listed in the tables, the proposed method achieves the smallest RMSE compared with other methods. Moreover, the SSIM score of the proposed method is nearly close to 1 for most test images. The results indicate that the proposed method achieves good SR reconstruction in depth while successfully preserving structure information. We also provide visual comparison results in

Fig. 6: Depth SR reconstruction results in Laser Scan data set when the upsampling factor is 4. Left to right: Ground Truth, Bicubic interpolation, ANR [20], SRCNN [19], Xie et al. [12], Ferstl et al. [11], and Proposed method.

depth SR reconstruction on different images by factor $\times 4$ in Figs. 5 and 6. Pseudo-color are employed in the depth maps to show details more clearly. It is obvious that the proposed method produces more visually pleasing results than the others. Above all, object boundaries of the proposed method are sharper than the others along the edge direction, which indicates that the proposed method successfully preserves the structure of the scene.

4. CONCLUSION

In this paper, we have proposed single DISR using CNN. We have addressed the DISR problem by HR edge prediction, instead of HR texture prediction. We have produced a high-quality edge map for depth upsampling based on CNN. Guided by the high-quality edge map, we have reconstructed SR depth images from LR ones using TV regularization. Specifically, the high-quality edge map has been used as the weight of the regularization term in the TV model. Experimental results demonstrate that the proposed method achieves better performance in SR reconstruction than state-of-the-arts including texture prediction-based ones while successfully preventing artifacts such as jagged edges, blurring and ringing.

5. REFERENCES

[1] Qingxiong Yang, Ruigang Yang, James Davis, and David Nister, "Spatial-depth super resolution for range images," in *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[2] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele, "Joint bilateral upsampling," in *Proc. ACM Siggraph*, 2007, p. 96.

[3] Jaesik Park, Hyeongwoo Kim, Yu Wing Tai, Michael S. Brown, and Inso Kweon, "High quality depth map upsampling for 3d-tof cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2012, pp. 1623–1630.

[4] 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.

[5] Wentian Zhou, Xin Li, and Daryl Reynolds, "Guided deep network for depth map super-resolution: How much can color help?," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1457–1461.

[6] William T Freeman, Thouis R Jones, and Egon C Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[7] Kwang In Kim and Younghee Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[8] Oisin Mac Aodha, Neill D. F. Campbell, Arun Nair, and Gabriel J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. European Conference on Computer Vision*, 2012, pp. 71–84.

[9] Michael Hornacek, Christoph Rhemann, Margrit Gelautz, and Carsten Rother, "Depth super resolution by rigid body self-similarity in 3d," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1123–1130.

[10] Jun Xie, Rogerio Schmidt Feris, Shiaw Shian Yu, and Ming Ting Sun, "Joint super resolution and denoising from a single depth image," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1525–1537, 2015.

[11] David Ferstl, Matthias Ruther, and Horst Bischof, "Variational depth superresolution using example-based edge representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 513–521.

[12] Jun Xie, Rogerio Schmidt Feris, and Ming Ting Sun, "Edge-guided single depth image super resolution," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 428–438, 2015.

[13] D. Scharstein and R. Szeliski, "Middlebury stereo vision.," in *http://vision.middlebury.edu/stereo/.[11]*.

[14] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning*, 2015, vol. 37, pp. 448–456.

[16] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2011.

[17] John Duchi, Elad Hazan, and Yoram Singer, *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*, vol. 12, 2011.

[18] Jingyu Yang, Xinchen Ye, Kun Li, and Chunping Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *Proc. European Conference on Computer Vision*, 2012, pp. 158–171.

[19] Chao Dong, Change Loy Chen, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. European Conference on Computer Vision*.

[20] Radu Timofte, Vincent De, and Luc Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.

[21] Daniel Scharstein and Richard Szeliski, "High-accuracy stereo depth maps using structured light," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2003, vol. 1, pp. 195–202.

[22] Daniel Scharstein and Chris Pal, "Learning conditional random fields for stereo," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[23] Heiko Hirschmuller and Daniel Scharstein, "Evaluation of cost functions for stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.