EDGE GUIDED SINGLE DEPTH IMAGE SUPER RESOLUTION

Jun Xie*, Rogerio Schmidt Feris**, Ming-Ting Sun*

*Electrical Engineering Department, University of Washington, **IBM T. J. Watson Research Center

ABSTRACT

Recently, consumer depth cameras have gained significant popularity due to their affordable cost. However, the limited resolution and quality of the depth map generated by these cameras are still problems for several applications. In this paper, we propose a novel framework for single depth image super resolution guided by a high resolution edge map constructed from the edges in the low resolution depth image via a Markov Random Field (MRF) optimization. With the guidance of the high resolution edge map, the high resolution depth image is up-sampled via a joint bilateral filter. The edge guidance not only helps avoid artifacts introduced by direct texture prediction, but also reduces the jagged artifacts and preserves the sharp edges. Experimental results demonstrate the effectiveness of our proposed algorithm compared to previously reported methods.

Index Terms— Single Depth Image, Super Resolution, Edge Guided, Joint Bilateral Up-sampling

1. INTRODUCTION

During recent years, we have witnessed a rapid progress in the field of 3D imaging. The birth of low-cost 3D scanning devices such as Microsoft Kinect and Time-of-Flight (TOF) cameras has opened the door for new applications in different research disciplines, including computer vision, graphics, human computer interaction and virtual reality. However, the limited resolution and low quality of the depth map generated by these cameras still pose serious issues for various 3D applications. For example, the resolution of SwissRange SR4000 depth camera and PMD Camcube camera are only about 200*200. Even for Kinect, the resolution of the depth image is 640*480, which is much lower compared to that of its corresponding color image (1280*1024). In this work, we aim to enhance the resolution of depth images with a single depth image as the input, which offers unique challenges. In its essence, single image super resolution (SR) requires the prediction of a large amount of unknown pixels based on the limited input pixels. Moreover, although depth maps contain less texture compared to color images, the depth captured by existing consumer cameras is usually degraded by noises due to the inaccurate scanning hardware or difficulties in calculating the disparity. In this paper, we propose a novel framework for single depth image SR which can produce better results than previous reported methods as shown in Fig. 1.

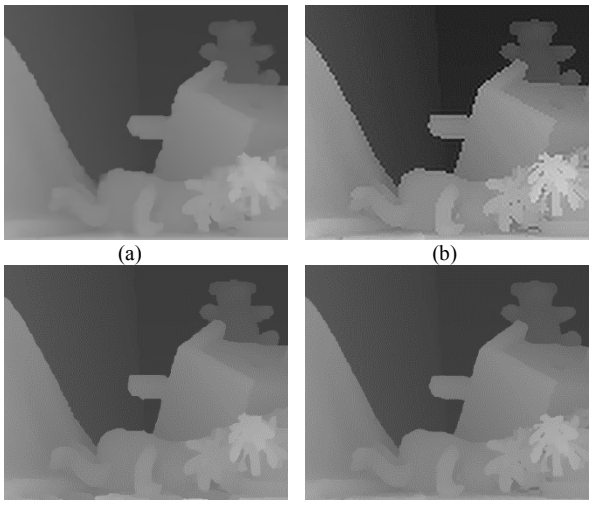

Fig. 1. Visual Result on the Middlebury dataset up-scaled by a factor of 4. (a) Aodha et. al [14], (b) Yang et. al [13], (c) Hornacek et. al [15], (d) Ours.

Traditional depth SR methods are focused on fusing multiple complimentary low resolution (LR) depth maps to get a high resolution (HR) depth image [1, 2]. However, they rely on the assumption that multiple range images are available with small camera movement, which may not be true for many practical applications. It was also proposed to use a pre-aligned HR color image to help upscale the depth map [5-11], since the high frequency components in color images such as edges can be utilized. For example, in [5, 7, 11], joint color and depth up-sampling is proposed to get the discontinuity information from the HR color image. In [10], a nonlocal means filter is utilized to regularize depth image in order to maintain the detailed structure. However, notwithstanding the appealing results that such approaches could generate, in many cases, the HR color image fully registered with the depth image may not be available.

For single image SR, in [12, 14], it formulates the problem as a multiclass MRF model with each hidden node representing the label of a HR patch. In [15], it searches for HR patches by identifying patch correspondences within the depth map itself. However, since the reconstruction is highly biased to examples externally or internally, it will not provide reliable results when no correspondence could be established. In [13], HR/LR patches are reconstructed as sparse linear combinations of learned coupled dictionary atoms based on the assumption that HR and LR patches should share the same reconstruction coefficients. However, sometimes it's difficult to learn the mapping between HR and

LR patches, which is many to one, yielding reconstruction problems such as blurry or ringing artifacts. Among the edge based methods, a gradient profile prior is proposed to improve color image SR in [16]. In [17], a multi-scale edge representation is produced to guide the process of color image SR, subject to the back-projection constraint. But these works still generate blurry artifacts at the edges and are not suitable for the depth image SR case.

Another existing problem in depth SR is to deal with noises. Since depth images usually are noisy, directly up-sampling them will also magnify the noises, and produce artifacts along edges. For the SR purpose, in order to preserve depth edges, a bilateral filter is utilized in the pre-processing step for noise reduction in [14]. However, from our observation, not only the noises, but the jagged artifacts around depth discontinuities caused by inaccurate sampling and heavy quantization of the disparity in the original low resolution image are also magnified.

In this paper, we propose a novel framework for single depth image SR guided by a reconstructed HR edge map. We convert the SR problem from HR texture prediction to HR edge prediction, which is motivated by the essence that edges are of particular importance in the textureless depth image. Guided by the predicted smooth HR edge map, a modified joint bilateral filter is applied to reconstruct the HR depth textures. The edge guidance not only helps avoid artifacts introduced by direct texture prediction, but also reduces the jagged artifacts and preserves the sharp edges. With simulations, we demonstrate the effectiveness of the proposed approach in terms of image quality both visually and quantitatively.

The rest of the paper is organized as follows: In Section 2, we present our proposed approach. In Section 3, we perform simulations to show the effectiveness of the proposed approach. In Section 4, we conclude the paper.

2. PROPOSED APPROACH

Our algorithm is motivated by the color assisted joint up-sampling approaches, in which the HR color image provides the discontinuity guidance so that pixels in a local region with different depth could be weighted differently in the upscaling process. It also infers that if only a HR depth edge image is provided, we could still follow the same idea to reconstruct the HR depth image. Therefore, instead of directly reconstructing the HR depth image, we propose to construct a HR edge map in the first step and then use the edge map as a guidance to reconstruct a HR depth image via a modified joint bilateral filter. Compared to the example or learning based SR methods, the guidance of HR edges could alleviate artifacts generated by direct depth value prediction such as blurring or ringing around edges. Moreover, the constructed HR edge map is smoothed without jagged artifacts, which could be magnified by methods such as nearest neighbor interpolation or some learning based methods. Thus, guided by the edge map, the HR depth image will contain smooth and sharp edges.

In the following discussion, we denote upscaling by a factor of g as up-sampling the image by $g*g$. We denote d_l and d_h as the input LR and the output HR depth image. We also have an external dataset containing a collection of HR and corresponding LR images. We first apply bicubic interpolation to upscale d_l to the same resolution of d_h and then extract edges using Canny edge detector to obtain an edge map e_r. Usually, e_r is not smooth and contains jagged edges. To have a higher quality HR edge map, we apply a Shock filter [19] to the interpolated depth map before edge detection and get a smoother edge map e_s. Together with e_r and the prior knowledge from the external dataset, we will refine e_s into a smooth HR edge map e_h described in the following. Guided by e_h, d_h will be produced by up-sampling d_l via a modified joint bilateral filter.

2.1. High Resolution Edge Map Construction

Given the jagged edge map e_r and the smoothed edge map e_s, we construct e_h by minimizing a discrete MRF energy function. Our method is patch based: for each edge pixel p_i in e_r, we extract w by w patches in e_r and e_s, centered at the position of p_i, denoted as x_r^i and x_s^i, respectively. To reduce the computation complexity, we discard patches which have a larger overlap (i.e. with an overlap area $>(w-2k)^2$, where k is a constant) with previously extracted patches. We denote $\boldsymbol{X}=\{\boldsymbol{x_r}, \boldsymbol{x_s}\}$ as a collection of stacked patches $\boldsymbol{x_r}$ and $\boldsymbol{x_s}$. In the external dataset we go through the same process to get the w by w jagged edge patches y_r^i from the given LR images and smooth edge patches y_s^i from the given HR images. We denote $\boldsymbol{Y}=\{\boldsymbol{y_r}, \boldsymbol{y_s}\}$, containing a collection of stacked patches $\boldsymbol{y_r}$ and $\boldsymbol{y_s}$.

The basic idea of MRF is to obtain HR edge patches from the external dataset under some likelihood and coherence constraints. Instead of directly obtaining depth intensity patches like [14], our intuition is based on the fact that binary edge patterns are much simpler than intensity patterns especially in depth images. Thus, it could give better matches for the edge patch, making the reconstruction less biased to the dataset. In our Markov grid model, each $\boldsymbol{X^i}$ forms the node and the hidden label corresponds to an edge patch $\boldsymbol{Y^i}$ from the dataset. The total energy of this MRF is:

$$E(y) = E_1 + w_1 E_2 + w_2 E_3 \quad (1)$$

where E_1 and E_2 are data terms and E_3 is the smoothness term. The first data likelihood term measures the similarity between the edge patch in $\boldsymbol{x_r}$ and the candidate edge patches in $\boldsymbol{y_r}$ in terms of the Euclidean difference of their corresponding distance transforms:

$$E_1 = \sum_{i=1}^{N} \| d(x_r^i) - d(y_r^i) \|^2 \quad (2)$$

where $d(*)$ stands for the distance transform [3] of the edge patch. The introduction of distance transform is to give a better similarity measurement of the binary patterns.

For E_2, it measures the similarity of the smoothed HR edge patches. Besides the first term E_1, the purpose of this term is to ensure that the HR edge patch candidates should

have consistent similarity measurement both in terms of the corresponded original and the smoothed edge pattern.

$$E_2 = \sum_{i=1}^{N} \|x_s^i - y_s^i\|^2 \quad (3)$$

The smoothness term E_3 enforces coherence in the overlapping regions between the neighboring edge patch candidates, where O_{ij} is an overlap operator that extracts the region of overlap between patch y_s^i and y_s^j:

$$E_3 = \sum_{x_r^i \cap x_r^j \neq \emptyset}^{N} \|O_{ij}(d(y_s^i)) - O_{ji}(d(y_s^j))\|^2 \quad (4)$$

We use belief propagation [18] to minimize Eqn. (1). As a result, for each $\boldsymbol{X_i}$, its discrete label which corresponds to an HR edge patch in $\boldsymbol{Y}$ can be inferred. Finally, y_s^i are put together by averaging pixel values in the overlapped region and then thresholding to the binary edge map e_h as shown in Fig. 2. From it we can see that in our result, the edges are smoothed (straight in the zoomed-in region) without jaggy or wavy noises. Note that for some parts of the edges, the edge width is more than one pixel (thicker edge) due to the averaging of overlapping patches. However, it could be handled properly as discussed in the next section.

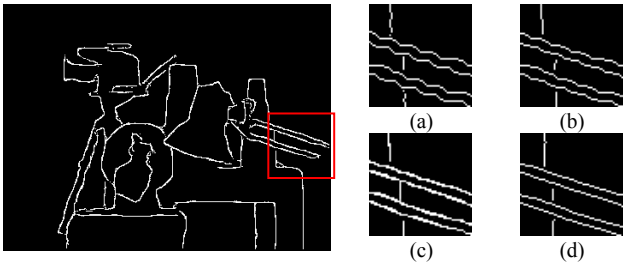

Fig. 2. Constructed edge map with an upscale factor of 4. Zoomed in results of (a) Edges of the bicubic upsampled depth. (b) Edges of the bicubic upsampled depth after using Shock filter. (c) Edges of our result. (d) The ground truth edges.

2.2. Edge Guidance for Single Depth Image Super Resolution via Joint Bilateral Upsampling

Once e_h is constructed, d_h can be recovered via a modified joint bilateral filer, in which the range kernel is replaced by an indicator function guided by e_h:

$$d_h(p) = \frac{1}{k_p} \sum_{q \subseteq N(p)} d_l(q_\downarrow) \cdot f_d(\|p_\downarrow - q_\downarrow\|) \cdot g_{e_h}(p, q) \quad (5)$$

where *N(p)* is an s by s supporting window centered at pixel p. $p\downarrow$ and $q\downarrow$ denote the corresponding pixel location in the LR image. Note that $p\downarrow$ and $q\downarrow$ take only integer coordinates in the LR image. Therefore, the guidance edge map is only sparsely sampled. k_p is a normalizing factor. $f_d(*)$ is a zero mean spatial Gaussian kernel and *g(*)* is a binary indicator function defined as:

$$g_{e_h}(u,v) = \begin{cases} 1 & \text{if } u \text{ and } v \text{ are at the same side of the edge } e_h \\ 0 & \text{if } u \text{ and } v \text{ are at the different sides of the edge } e_h \end{cases} \quad (6)$$

With the guidance provided by the HR edge map, only pixels at the same side of the edge will be considered during averaging so that edges could be well preserved. Within a patch (centered at p), to determine whether two pixels (p and q) are at the same side of the edge, we first dilate the patch to its 4-connected neighbors, the dilation result is denoted as D. If pixel i is on or next to the edge, *D(i)* is *1*, otherwise, *D(i)* is *0*. Then we construct a graph G so that each node N_G is corresponding to a pixel in D and the edge E_G is formed by connecting the 8-connected neighbors of each node. The edge weighting w_G between pixel i and j is determined as:

$$w_G(i,j) = dist(i,j) \cdot (\max(D(i), D(j)) + 1), \forall (i,j) \in E_G \quad (7)$$

dist(*) stands for the Euclidean distance of the coordinate of pixel i and j. Based on the assigned edge weights, we then compute a shortest path S between p and q in G. Basically we add more weights on pixels near the edge towards finding the shortest path so that the path will not touch the edge as much as possible. We draw S in the patch by including the 4-connected neighbors around each pixel along the path (except p and q) denoted as S', as shown in Fig. 3(a). We will divide our situation into two cases for discussion: *1) p is not an edge pixel, 2) p is an edge pixel.*

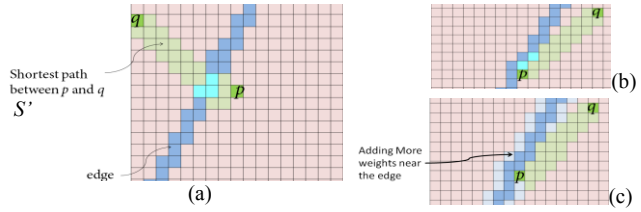

Fig. 3. (a) Illustration of two pixels at different sides of the edge. (b) A special case of two pixels mistakenly classifies as at the different sides of the edge. (c) Adding weights near the edge avoids situation in (b). (Best viewed in color)

CASE 1: p is NOT an edge pixel. As shown in Fig. 3(a), if S' covers the edge pixels (cyan pixels), p and q can be classified as at two sides of the edge, otherwise, p and q are at the same side. It should be noted that in some special cases, since we also add 4-connected neighbors along S, p and q could be mistakenly classified as at opposite sides because S' covers the edge pixel as shown in Fig. 3(b). However, because we are adding more weights on pixels near the edge, the calculated shortest path will avoid going through pixels near the edge. As a result, S' will not cover the edge as shown in Fig. 3(c).

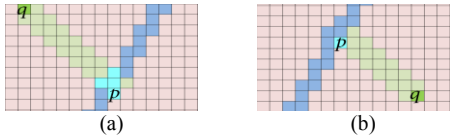

Fig. 4. Illustration of the case that p is on the edge. (a) p and q are at different sides, (b) p and q are at the same side. (Best viewed in color)

CASE 2: p is an edge pixel. If p is on the edge, it is ambiguous to decide whether p and q are on the same side or not. Thus, we simply exam the number of edge pixels along S' in order to reduce the error caused by the thicker edge case. If the number of edge pixels along S' is larger than *1*, p and q are classified as at different sides. Otherwise, p and q are at the same side as shown in Fig. 4(a) and (b).

Once we determine whether two pixels in the support are at the same side of the edge, the HR image can be

interpolated using Eqn. (5). In case that there is no pixel on the same side with p (i.e. $g_{eh}(p, q)=0, \forall q \in N(p)$), which rarely happens, we simply use the corresponded pixel value from bicubic interpolation as the up-sampled value for p.

3. EXPERIMENTAL RESULTS

In this section, we perform experiments on depth images obtained from multiple sources such as TOF camera and Middlebury Stereo data [4]. For training dataset, we use the synthesized HR depth data mentioned in [14]. We get the LR counterparts by down-sampling the HR images. The size of edge patches are set as 21 by 21. We set the parameters in all of our experiments as $w_1=8$ and $w_2=1$ in Eqn. (1), and $k=7$. The supporting window size for the bilateral filter is $s=scale*4+1$, and $sigma\_d=0.5$. We will show some of the experimental results using our algorithm both visually and quantitatively.

3.1. Quantitative Results

We compare our results with the learning based method [13] and the example based methods [14, 15] quantitatively on the Middlebury dataset. We first down-sample and smooth the ground truth using nearest neighbor interpolation to create the LR depth images. RMSE and SSIM results of different methods with different upscaling factors are listed in Table 1. From the result, we can see that our proposed algorithm outperforms other approaches. Note that we scale the depth images in the Middlebury dataset to the range of [0, 1], which is different from that reported in [14].

3.2. Visual Results

We compare our results with others visually on the Middlebury dataset as shown in Fig. 1 and Fig. 5. From the result, we can see that for the basis learning method [13] (Fig. 1(b) and Fig. 5(b)), since it involves direct HR texture prediction, artifacts around edges introduce the undesirable visual result. For the example based methods [13, 15] (Fig. 1 (a, c) and Fig. 5(a, c)), each HR patch is directly replaced by patches externally or internally, leading to artifacts in some region which has unique patterns (e.g. right bottom corner in Fig. 5(a) and upper right region in Fig. 5(c)). Our results, however, do not have jagged artifacts since our constructed HR edge map is smooth. Besides, our results have the most similar visual structures to the ground truths because our method is based on the bilateral filtering in which the HR pixels are interpolated from the original LR counterpart.

We also compare the visual results of the depth map captured by TOF camera (Camcube Camera) and the results are shown in Fig. 6. In this case, our method also generates more desirable visual results compared to other methods.

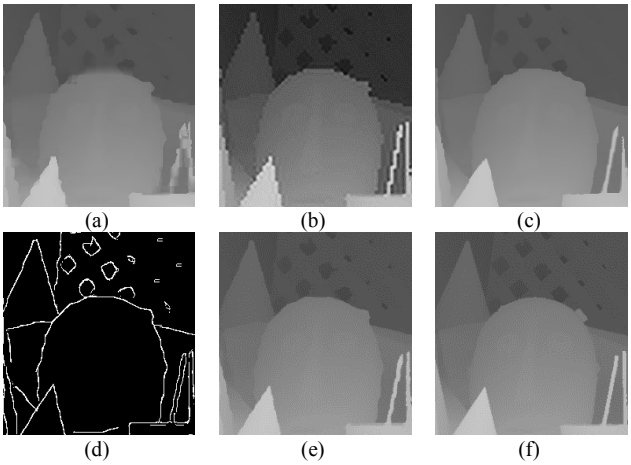

Fig. 5. Visual comparison on the Middlebury stereo data up-scaled by a factor of 4. (a) Aodha et. al [14], (b) Yang et. al [13], (c) Hornacek et. al [15], (d) Reconstructed edge map. (e) Our result, (f) Ground truth.

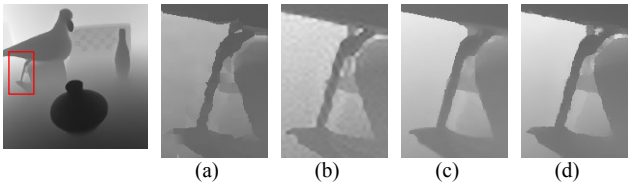

Fig. 6. Visual comparison on the depth captured by TOF camera up-scaled by a factor of 4. (a) Aodha et. al [14], (b) Yang et. al [13], (c) Hornacek et. al [15], (d) Our result.

4. CONCLUSION

In this paper, we present a novel framework for single depth image SR guided by a constructed HR edge map. We convert the SR problem from HR texture prediction to HR edge prediction, which is motivated by the essence that edges are of particular importance in the textureless depth image. We construct the HR edge map by posing it as a MRF labeling problem. Then guided by the edge map, the HR depth image is up-sampled using a joint bilateral filter. Experimental results demonstrate that our method not only has better objective performance, but also helps avoid artifacts introduced by direct texture prediction, reduces the jagged artifacts, and preserves sharp edges.

Table 1 RMSE and SSIM Comparison on Middlebury Data with Different Methods

	RMSE Comparison Scaled *4				SSIM Comparison Scaled *4				RMSE Comparison Scaled *3			
	Cones	Venus	Teddy	Tsukuba	Cones	Venus	Teddy	Tsukuba	Cones	Venus	Teddy	Tsukuba
Nearest Neighbor	1.498	0.367	1.348	0.832	0.886	0.954	0.895	0.833	1.172	0.309	0.925	0.672
Yang et. al [13]	2.169	1.017	1.582	0.840	0.869	0.924	0.871	0.777	1.292	0.421	1.133	1.505
Aodha et. al [14]	1.481	0.337	1.280	0.833	0.982	0.961	0.902	0.839	1.319	0.312	0.987	0.844
Hornacek et. al [15]	1.399	0.450	1.196	0.727	0.911	0.954	0.906	0.850	*NaN*[1]			
Ours	**1.148**	**0.272**	**0.871**	**0.662**	**0.927**	**0.974**	**0.930**	**0.870**	**0.839**	**0.226**	**0.703**	**0.544**

[1] The super resolution result upscaled by 3 is missing from the results provided by the author.

5. REFERENCES

[1] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," CVPR, 2009.

[2] A. N. Rajagopalan, A. Bhavsar, F.Wallho, and G. Rigoll, "Resolution enhancement of pmd range maps," DAGM, 2008.

[3] P. Felzenszwalb and D. Huttenlocher, "Distance Transforms of Sampled Functions," Technical Report 2004-1963, Cornell Univ. CIS, 2004.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," IJCV, vol. 47, pp. 7–42, 2002.

[5] K. Lo,Y. F. Wang, K. H, "Joint Trilateral Filtering For Depth Map Super-Resolution," VCIP 2013.

[6] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," CVPR, 2007.

[7] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," ACM SIGGRAGH, 2007.

[8] J. Lu, D. Min, R. S. Pahwa, and M. N. Do, "A revisit to mrfbased depth map super-resolution and enhancement," ICASSP, 2011.

[9] Y. Li, T. Xue, L. Sun, and J. Liu, "Joint example-based depth map super-resolution," ICME, 2012.

[10] J. Park, H. Kim, Y.W. Tai, M. Brown, I. Kweon, "High Quality Depth Map Upsampling for 3D-TOF Cameras," ICCV 2011.

[11] M. Liu, O. Tuzel, Y. Taguchi, "Joint Geodesic Upsampling of Depth Images," CVPR 2013.

[12] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-based super-resolution," Computer Graphics and Applications, 2002.

[13] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image superresolution as sparse representation of raw image patches," CVPR, 2008.

[14] O. M. Aodha, N. D. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image superresolution," ECCV, 2012.

[15] M. Hornacek, C. Rhemann, M. Gelautz, C. Rother, "Depth Super Resolution by Rigid Body Self-Similarity in 3D," CVPR 2013.

[16] J. Sun, J. Sun, Z. Xu, H. Y. Shum, "Image Super-Resolution using Gradient Profile Prior," CVPR 2008.

[17] Y.W. Tai, W. S. Tong, C. K. Tang, "Perceptually-Inspired and Edge-Directed Color Image Super-Resolution," CVPR 2006.

[18] R. S. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A Comparative Study of Energy Minimization Methods for Markov Random Fields", PAMI, vol. 30 (6), pp. 1068-1080, 2008.

[19] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Image Enhancement and Denoising by Complex Diffusion Processes," PAMI, vol. 26(8), pp. 1020–1036, 2004.