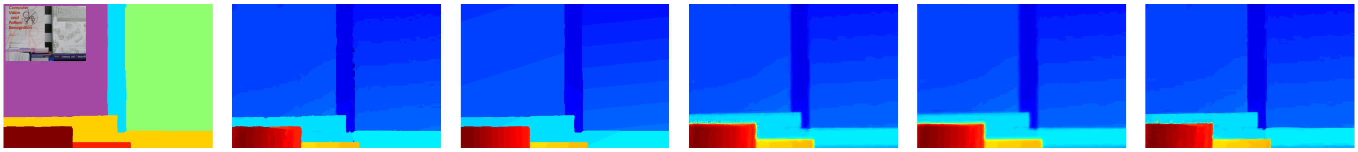
Joint Depth Map Interpolation and Segmentation with Planar Surface Model

Shibiao Xu[1], Longquan Dai[*,2], Jiguang Zhang[1,3], Jinhui Tang[2], G.Hemanth Kumar[3], Yanning Zhang[4], and Xiaopeng Zhang[†,1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[3]Department of Computer Science, University of Mysore, Mysore, India
[4]School of Computer Science, Northwestern Polytechnical University, Xi'an, China



(a) Our partition result (b) Ground truth (c) Our upsampling result (d) [Liu et al. 2013] (e) [Yang et al. 2012] (f) [Park et al. 2011]

Figure 1: *The $8X$ depth upsampling results comparison. We compare our algorithm with two global methods AR[Yang et al. 2012], MRF[Park et al. 2011] and a local method GF[Liu et al. 2013], and only ours can eliminate edge blurring and texture copying artifacts.*

Abstract

Depth map interpolation and segmentation has been a long-standing problem in computer vision. However, many people treat them as two independent problems. Indeed, the two problems are complementary. The results of one problem can aid in improving the results of the other in powerful ways. Assuming that the depth map consists of planar surfaces, we propose a unified variational formula for joint depth map interpolation and segmentation. Specifically, our model uses a multi-label representation of the depth map, where each label corresponds to a parametric representation of the planar surface on a segment. Using alternating direction method, we are able to find the minimal solution. Experiments show our algorithm outperforms other methods.

Keywords: depth map interpolation, object segmentation

Concepts: •**Computing methodologies** → *Image segmentation; Reconstruction;*

1 Introduction

Recently, a variety of depth sensors such as laser scanners, TOF cameras and passive stereo systems have been developed. These are all significant tools for 3D scene understanding. However, these sensors suffer from various problems. Unlike conventional optical cameras, the resolution of active depth sensors is extremely low. For instance, the state-of-the-art 3D-TOF camera Swiss Ranger can only capture a 100×100 depth map. In contrast, the resolution of mainstream optical cameras can be even higher than 1000×1000. Kinect as a representative passive stereo system can capture a 640×480 depth image, but it often loses depth information and forms black holes in the depth map because some infrared lights emitted from Kinect are absorbed and occluded by the objects in the scene.

Although many methods are proposed to upscale the low-resolution depth map, and to recover the missing depths in the literature [Chan et al. 2008; Liu et al. 2013; Park et al. 2011; Yang et al. 2012], there are still some problems. Specifically, these existing algorithms interpolate missing depths under the guidance of a registered RGB image and are based on the assumption that neighboring pixels with similar colors are likely to have similar depth values. To utilize the guidance information of RGB images, various edge-preserving weights have been designed, which are computed from the guidance image, and these weights have been integrated into the interpolation models. However, the structure of depth maps and color images are often mismatched, the defective weights thus inevitably introduce undesirable artifacts into final results, such as edge blurring and texture copying, which create a problem between the smoothing and edge-preserving. While the color images totally agree with the depth maps, we can detect color edges and assign zero to the weights between crossing color edge pixels to indicate that their depths are irrelevant. However, in the real world, depth edges usually are a subset of color edges. If adjacent pixels are in the same depth region but cross color edges, removing the connections between them will block depth diffusion from the seeds to the interpolated pixels. As a compromise, existing methods set color similarities of coupled pixels as their weights. Strong edge-preserving weights methods will copy the guidance image's textures into slant depth surfaces while keeping sharp depth edges. In contrast, weak edge-preserving weights methods will blur depth edges in order to remove texture copying artifacts.

To avoid edge blurring and texture copying, we adopt a novel strategy. Guided by a color photo, the depth map is partitioned into different segments. For each interpolated pixel, we only use the seeds in the same region to interpolate its depth. In this way, we can eliminate edge blurring completely. To recover the texture copying degradation, we assume that the depth surface forms a slant plane for each segment and use that assumption to estimate the missing depths on each segment instead of using edge-preserving weights.

[*]Shibiao Xu and Longquan Dai share the first authorship.
[†]Corresponding author: Xiaopeng Zhang (xiaopeng.zhang@ia.ac.cn).

SA'16 Technical Briefs , December 05 - 08, 2016, , Macao
©2016 ACM. ISBN 978-1-4503-4541-5/16/12 $15.00
DOI: http://dx.doi.org/10.1145/3005358.3005365

In our algorithm, the segmentation and interpolation are no longer two independent processes; we couple them with each other and use a unified variational formula to describe the coupling.

2 Related work

The most related algorithms are depth upsampling methods which focus on upscaling a depth map without considering segmentation. These methods are divided into the global method and the local method. Global methods minimize MRF [Kalra et al. 1999] and AR [Zhang and Wu 2008] models which produce a large cost for the coupled pixels that have similar colors but different depths; Local methods interpolate missing depths by averaging the seeds' depth contribution which is proportional to the color similarities between pixels and seeds. Specifically, Diebel [Diebel and Thrun 2005] first introduced the MRF model to upscale depths; Park [Park et al. 2011] designed the NLM edge-preserving coefficients to improve the performance of the MRF model; Instead of using the MRF model, Yang [Yang et al. 2012] presented an AR interpolation model. In contrast to these methods [Diebel and Thrun 2005; Park et al. 2011; Yang et al. 2012] which fall into the global method, Kopf [Kopf et al. 2007] invented the first filtering-based depth upsampling algorithm. Chan [Chan et al. 2008] indicated the texture copying artifacts in the upsampled depth map and proposed a noise-aware filter to eliminate the artifacts. To reduce running time, Liu [Liu et al. 2013] introduced a novel joint geodesic upsampling filter for real-time processing tasks.

Segmentation methods contains two types of methods: supervised segmentation and unsupervised segmentation. Potts model [Nieuwenhuis et al. 2013] is a popular supervised segmentation model. To find the optimal solution, many convex relaxations of Potts model were proposed. In contrast, K-means, Mean shift and Normalized cuts are three major unsupervised segmentation methods. Whether a method needs the description parameters for each segment determines which category it belongs to. Here, we apply Potts model to unsupervised segmentation. The partition and the depth plane parameters for each segment are calculated simultaneously.

In our algorithm, the two complementary processes of interpolation and segmentation are solved jointly, a 3D scene is represented as a collection of planar surfaces. According to the the parameters a_i, b_i, c_i of each planar surface $a_i x + b_i y + c_i$, we can partition image domain Ω into several parts Ω_i. Based on these parts, we can interpolate the depths d and estimate the parameters of each planar surfaces.

3 Object Function

Our goal is to perform depth map interpolation and segmentation under the guidance of a registered color photo I. Let $1 \leq i \leq L$ and L is the number of segments. We partition the image domain Ω into several segments Ω_i and estimate the parameters a_i, b_i, c_i of each planar surface $a_i x + b_i y + c$ and interpolate depths d by optimization. The object function consists of five terms which will be discussed in the following paragraphs.

Perimeter Per(Ω_i): Per(Ω_i) denotes the boundary length of Ω_i; we consider that for a satisfactory partition, its perimeter should be as small as possible. Let $u_i = \mathbb{1}_{\Omega_i}$ which is the indicator function of Ω_i, we have Per(Ω_i)=$\int_\Omega |\nabla u_i|$, but the form does not employ the edge information of the guidance image I. To pull the segmentation boundaries towards strong image edges of I, which should be possible depth edges, we add a g-weighting term to $\int_\Omega g|\nabla u_i|$ and consider that the perimeter Per(Ω_i) should be as small as possible in the sense of $\int_\Omega g|\nabla u_i|$, where $g(x) = \exp(-a|\nabla I(x)|)$.

Label cost prior $\sum_{i=1}^{L} \|u_i\|_\infty$: The minimum description length principle [Rissanen 1978] demands that we should use fewer symbols than needed to describe the data. Therefore the number of segments should be as small as possible, otherwise too many labels will cause over-segmentation. To overcome the shortcoming, we exploit the label cost prior $\sum_{i=1}^{L} \|u_i\|_\infty$ to constrain the maximium partition number.

Residual $r_i(S)$: The residual $r_i(S) = \int_{\Omega_i} T(|a_i x + b_i y + c_i - d^0|, S)dp$ quantifies how a plane $a_i x + b_i y + c_i$ fits the observer data d^0 on the region Ω_i, where S denotes the position of the observer data and $T(\cdot, S)$ is an interpolation operator which uses the interpolating values $|a_i x + b_i y + c_i - d^0|$ on S to estimate the missing values on $\widetilde{S} = \Omega \setminus S$. Here, we define $T(\cdot, S)$ as the joint bilateral filter [Kopf et al. 2007] because it is aware of the guidance image I's structure while interpolating. For proper plane parameters, $r_i(S)$ approaches zero. There are two reasons: 1) our depth plane assumption is rational and flexible since most objects in a scene reside in several depth planes; even for the surface which is not a plane, it can also be approximated by several planes; 2) $r_i(S)$ is a weighted L_2 norm of the values $|a_i x + b_i y + c_i - d^0|$ on the region $\Omega_i \bigcap S$, thus smaller values $|a_i x + b_i y + c_i - d^0|$ imply smaller $r_i(S)$.

Smooth term s_i and Data term $t_i(S)$: The smooth term $s_i = s_{i_1}(S) + s_{i_2}(\widetilde{S}) = \int_{\Omega_i \bigcap S} |\partial_x d - a_i| + |\partial_y d - b_i| dp + \int_{\Omega_i \bigcap \widetilde{S}} |\partial_x d - a_i| + |\partial_y d - b_i| dp$ and data term $t_i(S) = \int_{\Omega_i \bigcap S} (d - d^0)^2 dp$ measure how the depths d fit a plane $a_i x + b_i y + c_i$ and the observed data d^0, respectively. When we minimize $s_{i_1}(S)$, the depths on S are regularized based on our planar surface assumption; when we minimize $s_{i_2}(\widetilde{S})$, the values of d is extended from S to $\widetilde{S}$; when we minimize $t_i(S)$, d would not deviate the observed data d^0 too much.

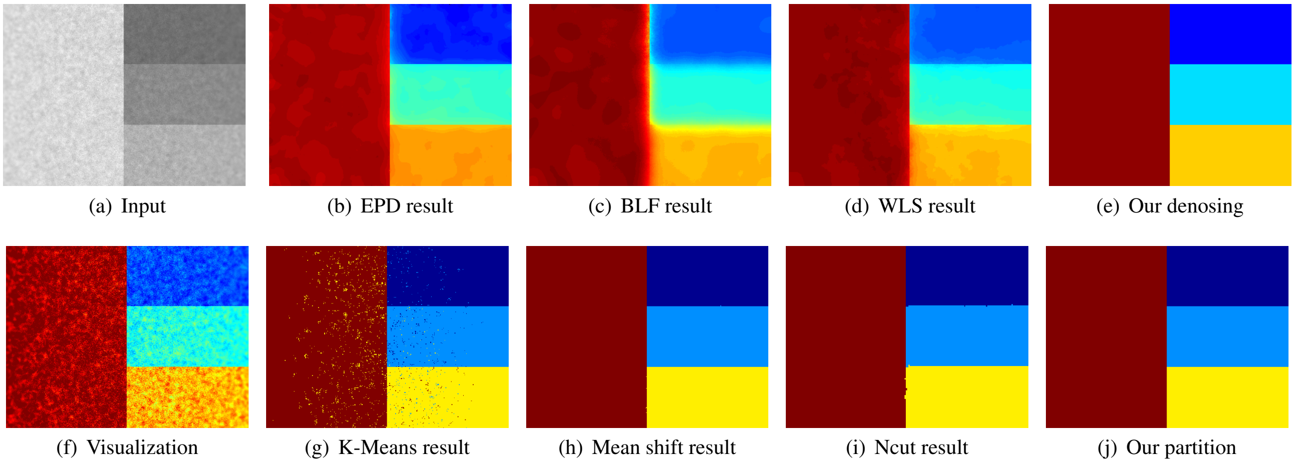
By putting all terms together, computing u_i, a_i, b_i, c_i and d boils down to the minimization of Equ. (1), where $f_i = f_i^s(S) + \varsigma s_{i_2}(\widetilde{S}) = (\beta r_i + \varsigma s_{i_1} + \tau t_i)(S) + \varsigma s_{i_2}(\widetilde{S})$

$$\begin{aligned} &\min_{u_i, a_i, b_i, c_i, d} \sum_{i=1}^{L} \left\{ \int_\Omega g|\nabla u_i| + \gamma \|u_i\|_\infty + f_i \right\} \\ &s.t. \quad \sum_{i=1}^{L} u_i = 1, \quad u_i \in \{0, 1\} \end{aligned} \tag{1}$$

4 Optimization

The optimization in Equ. (1) poses a difficult non-convex optimization problem. If we relax u_i to $[0, 1]$ and take a closer look, the model is convex both in u_i and in a_i, b_i, c_i, d. Hence, we can split up the optimization into two subproblems Equ. (2) and Equ. (3), and use an alternating direction method [Boyd et al. 2011] to find optimal results. Specifically, Equ. (2) is used to calculate segment u_i^k for fixed a_i^{k-1}, b_i^{k-1}, c_i^{k-1} and d^{k-1}:

$$\begin{aligned} &u_i^k = \arg\min_{u_i} \sum_{i=1}^{L} \left\{ \int_\Omega g|\nabla u_i| + \gamma \|u_i\|_\infty + f_i \right\} \\ &s.t. \quad \sum_{i=1}^{L} u_i = 1, \quad u_i \geq 0 \end{aligned} \tag{2}$$


(a) Input (b) EPD result (c) BLF result (d) WLS result (e) Our denosing

(f) Visualization (g) K-Means result (h) Mean shift result (i) Ncut result (j) Our partition

Figure 2: *Joint restoration and segmentation. (a) and (f) are the input image and its color visualization. From left to right, the first row shows the denosing results of EPD [Subr et al. 2009], BLF [Tomasi et al. 1998], WLS [Farbman et al. 2008] and ours; the second row demonstrates the partition of K-means, Mean shift, Normalized cuts and ours, where pixels with the same color implies that they belong to the same region.*

Equ. (3) targets at computing a_i^k, b_i^k, c_i^k and d^k for fixed u_i^k, where k is the number of iterations:

$$a_i^k, b_i^k, c_i^k, d^k = \arg\min_{a_i, b_i, c_i, d} \sum_{i=1}^{L} \left\{ f_i^s(S) + \varsigma s_{i_2}(\widetilde{S}) \right\} \quad (3)$$

The cost of iterative computing Equ. (2) (3) can be reduced. Indeed, only the results of Equ. (3) at the last step are accepted as final results. If we can reduce the cost at the intermediate steps, the overall cost will be reduced. We discover that the region indicator function f_i consists of two terms $f_i^s(S)$ and $s_{i_2}(\widetilde{S})$ which depend on the values of two disjoint regions S and $\widetilde{S}$, respectively. On the region S, d is constrained by the observed data d^0 and thus cannot be freely chosen. Hence the value of $f_i^s(S)$ is usually very large. On the contrary, $s_{i_2}(\widetilde{S})$ usually is very small because the values of d on the region $\widetilde{S}$ are in freedom. So $f_i^s(S)$ is the dominating factor and we can get rid of $s_{i_2}(\widetilde{S})$ without affecting final results. In addition, the guidance image I's structure information which is embedded into Per(u_i) and r_i, is more reliable than d^{k-1} as an indication of the possible depth edges on the region $\widetilde{S}$. This is because the edges of d^{k-1} determined by partition u_i^{k-1} are not as reliable as the color edges of I. Generally, removing s_{i_2} from Equ. (2) (3), we could save the cost of computing the depths on the region $\widetilde{S}$ without introducing negative effect. Furthermore, in Equ. (3), $\sum$ can be removed because f_i^s and f_j^s are independent for $i \neq j$, and each f_i^s only determines $d_{i^s}^k$ denoting the part of d^k on the region $\Omega_i^k \bigcap S$, where $u_i^k = \mathbb{1}_{\Omega_i^k}$.

Jointly minimizing Equ. (2) (3) by the alternating direction method [Boyd et al. 2011] iteratively, we can find the minimal solution of Equ. (1). Firstly, Equ. (2) uses partial depths d^0 on S and a registered guidance image I to iteratively partition the image domain into different parts. After that, Equ. (3) employs the partition and the partial depths to produce our interpolation results.

5 Experiment

We use matlab to implement our algorithm. The upsampling performance of our method is rather stable when the explicit parameters γ, ς, β, τ of Equ. (1) are in the range $[10, 100]$, $[5, 20]$, $[5, 15]$, $[20, 50]$. For convenience, we consistently keep the parameter setting (i.e. $\gamma = 30$, $\varsigma = 10$, $\beta = 10$, $\tau = 30$) of all the data sets used in the experiments unchanged.

Fig. 2 illustrates the ability of joint restoration and segmentation of our algorithm for a noisy image. In this case, all pixels in the image domain Ω are seeds (i.e. $\Omega = S$) and interpolation reduces to restoration. We use the noisy image itself to guide the joint restoration and segmentation. Original image Fig 2(a) is gray. For clarity, we visualize all images using a color map. Our method produces a denoised image Fig 2(e) and corresponding segmentation Fig 2(j). The results are compared with three state-of-the-art denoising methods [Subr et al. 2009; Tomasi et al. 1998; Farbman et al. 2008] and three recently proposed segmentation methods. In the first row of Fig. 2, EPD [Subr et al. 2009], BLF [Tomasi et al. 1998], WLS [Farbman et al. 2008] copy the undulating surfaces into final results and blur the edges between different parts because these methods could not suppress the texture copying effect and prohibit the information diffusion between edges; the segmentation results of K-means, Mean shift, Normalized cuts in the second row of Fig. 2 also are influenced by the noise. Unlike previous methods, our results could exactly keep the sharp edges without any blurring and the noise in the image does not affect final segmentation.

Our algorithm can be applied to various interpolation tasks such as random missing, structural missing and upsampling. In these cases, seeds S are a subset of the image domain Ω. Using the famous Middlebury stereo datasets [Scharstein and Szeliski 2002] and RGB-D object dataset [Lai et al. 2011], we synthesize three depth maps (i.e. structure missing with possion noise, 5% random missing and $8X$ upsmapling) to evaluate the performance of our method for these interpolation tasks. The experimental results are listed in Fig 3 and Fig.1. We also colorize these figures for clarity. We can observe that Equ. (1) could suppress the noise in the seeds and use the partial depth information of seeds to partition the entire image domain. The interpolation results are comparable with the ground truth, proving that Equ. 1 can interpolate satisfactory results. The quality of depth ground truth in the second row is not as good as the quality of depth ground truth in the first row, because the depth of the second row is captured by Kinect. However, the interpolation results are superior to the captured depth map due to our model's regularization ability which employs several planes to reconstruct the entire depth map.

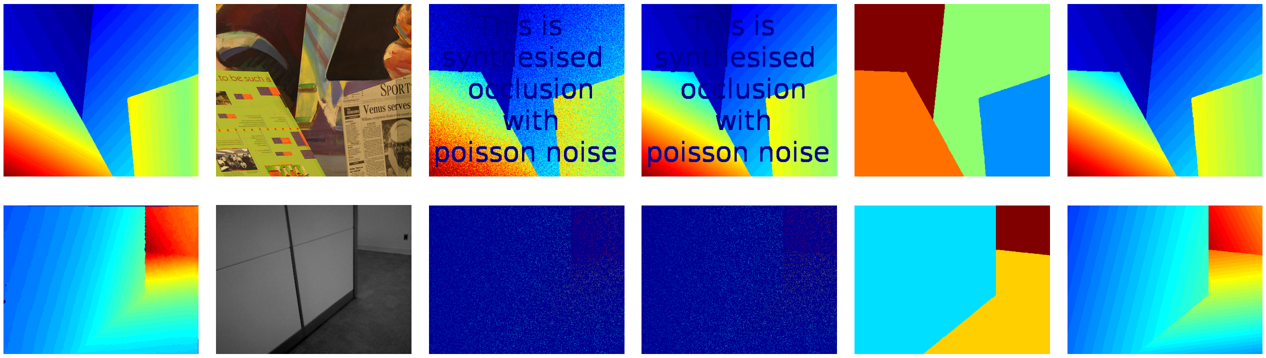


Figure 3: *Joint interpolation and segmentation. From left to right: ground truth images, guidance images, synthesized noisy images, our restored depths by Equ.* (1)*, our partition results by Equ.* (1)*, our depth interpolation results by Equ.* (1)*.*

Depth map upsampling is an important application which attracts much attention. We compare our algorithm with two global methods (AR [Yang et al. 2012], MRF [Park et al. 2011]) and a local method GF [Liu et al. 2013] with the 8X depth upsampling results in Fig. 1. We can observe that only our algorithm can eliminate edge blurring and texture copying artifacts. In contrast, we can easily find the degradation in the results of the optimization based MRF and AR models [Park et al. 2011; Yang et al. 2012] and the geodesic filtering based method [Liu et al. 2013].

6 Conclusion

We presented a variational approach for joint depth map interpolation and segmentation. In our model, the interpolation and segmentation complement each other. The boundary information from segmentation is used to avoid blurring edges while interpolating depths; the recovered depths are employed to remove the negative effect of noise while partition. Furthermore, the partition number is determined automatically by our algorithm. Our method also explicitly models the surfaces, and so can suppress the texture copying and edge blurring artifacts as well as removing noise in the image. Experiments show that our approach is robust to different types of interpolation and delivers good results for a wide range of images.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (Nos. 61620106003, 61332017, 61671451, 61571400), and in part by the China National 863 Program with No. 2015$AA$016402.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn. 3*, 1, 1–122.

Chan, D., Buisman, H., Theobalt, C., and Thrun, S. 2008. A Noise-Aware Filter for Real-Time Depth Upsampling. In *M2SFA2*.

Diebel, J., and Thrun, S. 2005. An application of markov random fields to range sensing. In *NIPS*.

Farbman, Z., Fattal, R., Lischinski, D., and Szeliski, R. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. In *ACM SIGGRAPH*, ACM, 67:1–67:10.

Kalra, S., Krishnan, D., and Chong, M. 1999. A mrf model based scheme for accurate detection and adaptive interpolation of missing data in nightly corrupted image sequences. In *ICIP*, vol. 2, 890–893 vol.2.

Kopf, J., Cohen, M. F., Lischinski, D., and Uyttendaele, M. 2007. Joint bilateral upsampling. In *ACM SIGGRAPH 2007 Papers*, SIGGRAPH '07.

Lai, K., Bo, L., Ren, X., and Fox, D. 2011. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 1817–1824.

Liu, M.-Y., Tuzel, O., and Taguchi, Y. 2013. Joint geodesic upsampling of depth images. In *CVPR*, 169–176.

Nieuwenhuis, C., Tppe, E., and Cremers, D. 2013. A survey and comparison of discrete and continuous multi-label optimization approaches for the potts model. *IJCV 104*, 3, 223–240.

Park, J., Kim, H., Tai, Y.-W., Brown, M., and Kweon, I. 2011. High quality depth map upsampling for 3d-tof cameras. In *ICCV*, 1623–1630.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica 14*, 5, 465 – 471.

Scharstein, D., and Szeliski, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV 47*, 1-3 (Apr.), 7–42.

Subr, K., Soler, C., and Durand, F. 2009. Edge-preserving multiscale image decomposition based on local extrema. In *ACM SIGGRAPH Asia*, 147:1–147:9.

Tomasi, C., Manduchi, R., and Manduchi, R. 1998. Bilateral filtering for gray and color images. In *ICCV*, IEEE Computer Society, Washington, DC, USA, ICCV '98, 839–846.

Yang, J., Ye, X., Li, K., and Hou, C. 2012. Depth recovery using an adaptive color-guided auto-regressive model. In *ECCV*, Springer-Verlag, 158–171.

Zhang, X., and Wu, X. 2008. Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation. *TIP 17*, 6, 887–896.