This is the Pre-Published Version.

The following publication S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen and L. Zhang, "Learning Dynamic Guidance for Depth Image Enhancement," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 712-721 is available at https://doi.org/10.1109/CVPR.2017.83.

Learning Dynamic Guidance for Depth Image Enhancement

Shuhang Gu[1], Wangmeng Zuo[2], Shi Guo[2], Yunjin Chen[3], Chongyu Chen[4,1], Lei Zhang[1,*]
[1] The Hong Kong Polytechnic University, [2] Harbin Institute of Technology,
[3]ULSee Inc., [4]Sun Yat-sen University.
{*shuhanggu@gmail.com, cswmzuo@gmail.com, cslzhang@comp.polyu.edu.hk*}

Abstract

The depth images acquired by consumer depth sensors (e.g., Kinect and ToF) usually are of low resolution and insufficient quality. One natural solution is to incorporate with high resolution RGB camera for exploiting their statistical correlation. However, most existing methods are intuitive and limited in characterizing the complex and dynamic dependency between intensity and depth images. To address these limitations, we propose a weighted analysis representation model for guided depth image enhancement, which advances the conventional methods in two aspects: (i) task driven learning and (ii) dynamic guidance. First, we generalize the analysis representation model by including a guided weight function for dependency modeling. The task-driven learning formulation is introduced to obtain the optimized guidance tailored to specific enhancement tasks. Second, the depth image is gradually enhanced along with the iterations, and thus the guidance should also be dynamically adjusted to account for the updating of depth image. To this end, stage-wise parameters are learned for dynamic guidance. Experiments on guided depth image upsampling and noisy depth image restoration validate the effectiveness of our method.

1. Introduction

High quality and dense depth image plays a fundamental role in many real world applications, such as robotics, human-computer interaction, and augmented reality. Traditional depth sensing is mainly based on stereo or laser measurement, which in general is of high computational burden or expensive price. Recently, the wide availability of consumer depth sensing products, e.g., RGB-D cameras and Time of Flight (ToF) range sensors, offers an economic alternative for dense depth measurements. However, the depth image generated by consumer depth sensors usually is of insufficient quality, resulting in depth image with low resolution, noise or missing values.

Depth image enhancement have received considerable recent research interests [38, 15, 4, 28, 6, 20, 24]. One representative solution is to utilize multiple depth images from the same scene to reconstruct a high quality depth image [38, 15]. These methods, however, relies heavily on accurate calibration, and may fail when applied to dynamic environment. Another popular solution is to incorporate a high quality color camera with a depth sensor for depth image enhancement [4, 1, 28, 6, 24]. For most consumer depth sensors, high quality RGB image generally can be simultaneously acquired with the depth image, making it very appealing and natural to exploit color images for guided depth image enhancement.

Modeling dependency between intensity and depth images plays a key role in guided depth image enhancement. Based on the structural co-occurrence between intensity and depth images, filtering methods have been used to transfer the salient structure from intensity image to the enhanced depth map [13, 34]. Certain forms of objective functions have also been adopted for interdependency modeling, resulting several Markov Random Fields (MRF) [4], non-local mean [28]), and variational (e.g., Total Generalized Variation [6] models. The filter-based and model-based approaches, however, usually are ad hoc and limited in characterizing the complex dependency.

Recently, learning-based methods have been studied. Under the sparse representation framework, analysis and synthesis dictionary learning models have been exploited for modeling the statistical relation of intensity and depth images [36, 20]. Motivated by the success of deep learning, deep CNNs [7, 22] are also developed. However, the existing dictionary learning methods simply packed intensity and the associated depth patches to learn dictionaries in a group learning manner. Moreover, along with the enhancement, more details of the depth image will be recovered. Therefore, the guidance should also be dynamically adjusted to cope with the updating of depth image, but few studies have been given to address this issue in the dictionary learn-

*This work is supported by HK RGC General Research Fund (PolyU 152124/15E) and National Natural Science Foundation of China (grant no. 61672446).

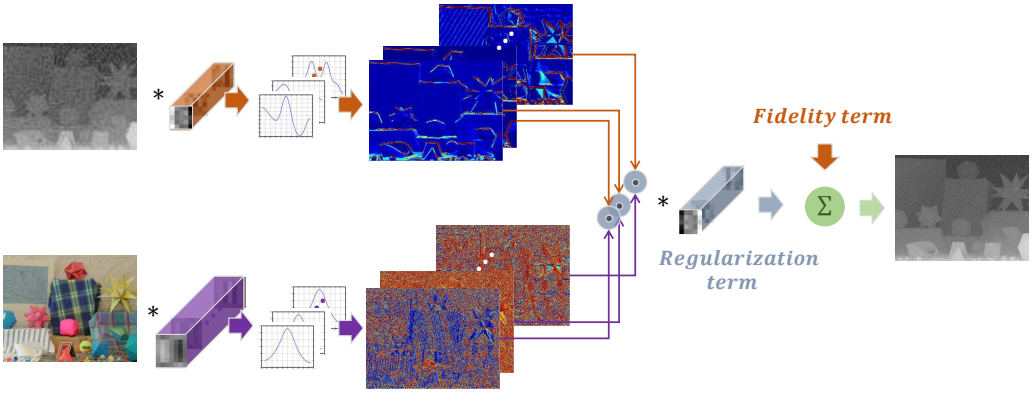


Figure 1. Illustration of our guided depth enhancement method. At stage $t+1$, the current enhancement result $\boldsymbol{x}_t$ and the guided image $\boldsymbol{g}$ are first convolved with the corresponding L analysis filters, respectively. After nonlinear transform, the filtering responses of $\boldsymbol{x}_t$ and $\boldsymbol{g}$ are combined via element-wise product, and further convolved with the L adjoint filters to form the result by regularization term. Finally, the results by regularization and fidelity terms are summarized to obtain the updated result $\boldsymbol{x}_{t+1}$.

ing and CNN-based methods.

In this paper, we investigate the problem of task-driven dynamic guidance learning in the analysis representation framework [30]. Due to its ability in modeling complex local structure, analysis representation model has been adopted in various image restoration tasks [30, 2]. Here we extend it by including a guided weight function for dependency modeling, resulting in our weighted analysis representation model. For task-driven guidance learning, we introduce a bi-level optimization model, which allows us to obtain the optimized guidance tailored to specific enhancement task. For dynamic guidance learning, by referring to [33, 3], we use gradient descent to solve the lower-level problem, and learn stage-wise parameters from training data.

To sum up, our method can not only provide a good way to introduce intensity guidance for depth image enhancement, but also result in good stage-wise parameters specified to certain depth image enhancement task. Fig. 1 illustrates the process of one stage in our method. The contribution of this paper is three-fold:

- By introducing a guided weight function, we extend analysis representation model for guided depth image enhancement. In our model, analysis representation and weight function are combined to characterize the priors and dependency of intensity and depth images.
- A task-driven formulation is suggested to learn weighted analysis representation model by solving a bi-level optimization problem, and dynamic guidance is learned from training data for certain task.
- Experiments are conducted on depth image upsampling and noisy depth image restoration. The results validate the superiority of our method against state-of-the-arts by both quantitative metric and visual quality.

2. Related works

In this section, we first provide a brief survey on analysis representation model, and then review several related methods on explicit guidance modeling.

2.1. Analysis sparse representation

Analysis sparse representation has been widely applied in many image processing and computer vision tasks [32, 35, 30, 33, 40, 3]. It adopts analysis operator [31] on image patches or analysis filters [30] on whole images for modeling the local structure of natural images. Compared with synthesis sparse representation, the analysis model adopts an alternative viewpoint for union-of-subspaces reconstruction by characterizing the complement subspace of signals [5], and usually results in efficient solutions.

Here we only consider convolutional analysis representation, and one representative form can be given by:

$$\boldsymbol{x} = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}) + \sum\nolimits_l \sum\nolimits_i \rho_l((\boldsymbol{k}_l \otimes \boldsymbol{x})_i), \quad (1)$$

where $\otimes$ denotes the convolution operator, and $(\cdot)_i$ denotes the value at position i. The penalty function $\rho_l(\cdot)$ is introduced to characterize the analysis coefficients. $\mathcal{L}(\boldsymbol{x},\boldsymbol{y})$ is the data fidelity term determined by degradation model. For Gaussian image denoising, one can simply let $\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$.

Analysis sparse representation has been studied for several decades. Rudin et al. proposed a total variation (TV) model [32], where the analysis filters are gradient operators and the penalty function is the ℓ_1-norm. Subsequently, a great many of attempts have been made to provide better analysis filters and penalty functions. And an emerging topic is to learn analysis sparse models from training data. Zhu et al. [39] proposed a FRAME model which aims to learn penalty functions for predefined filters. Roth et al. [30] proposed a field-of-expert (FoE) model in which anal-

ysis filters are learned for predefined penalty functions. Although FRAME and FoE are originally introduced from a MRF perspective, they can also be interpreted as the analysis representation models [31]. Recently, Schmidt et al. [33] and Chen et al. [3] suggest to model the related functions with linear combination of Gaussian RBF kernels, and can learn both analysis filters and penalty functions from training data. Moreover, by incorporating with the specific optimization methods, stage-wise parameters can be learned in a task driven manner.

Despite their achievements in image restoration, most existing methods are used for learning analysis representation on images from one single modality, and cannot be applied to guided depth image reconstruction. Kiechle et al. take one step forward by introducing a bimodal analysis model to learn a pair of analysis operators [19]. But the issue of explicit and dynamic guidance from intensity image remains unaddressed in analysis representation learning. In this work, we extend the analysis model by introducing guided weight function for modeling the guidance from intensity image, and adopt a task-driven learning method to learn stage-wise parameters for dynamic guidance.

2.2. Explicit guidance modeling

A number of approaches have been proposed to introduce the guidance information. One representative method is to formulate the input image $\boldsymbol{y}$, output image $\boldsymbol{x}$ and the guidance image $\boldsymbol{g}$ into an optimization model [21, 4, 28, 6, 11]. Here we only focus the explicit guidance models, where the regularization term is represented as the combination of the guidance function on $\boldsymbol{g}$ and the penalty function on $\boldsymbol{x}$. In the MRF-based depth upsampling model [4], the prior potential function is defined as:

$$\sum_i \sum_{j\in\mathcal{N}(i)} \phi_\mu(\boldsymbol{g}_i - \boldsymbol{g}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^2, \tag{2}$$

where i and j are the pixel indexes of image, $\mathcal{N}(i)$ is the set of neighboring index of i, and $\phi_\mu(z) = \exp(-\mu z^2)$. Similar weight function has also been adopted in other models, e.g., non-local mean (NLM) [28], for guided depth enhancement. Besides pixel-wise difference, other cues such as color, segmentation and edge, are also considered to design proper weight function.

Instead of modifying weight function, Ham et al. [11] adopt the Welsch's function to regularize the depth differences:

$$\sum_i \sum_{j\in\mathcal{N}(i)} \phi_\mu(\boldsymbol{g}_i - \boldsymbol{g}_j)(1 - \phi_\nu(\boldsymbol{x}_i - \boldsymbol{x}_j))/\nu. \tag{3}$$

Besides, several hand-crafted high order models are also proposed to model the weight function and depth regularizer [6].

Actually, the models in Eqns. (2) and (3) can be treated as extensions of handcrafted analysis model. In which a group of inter-pixel difference operators are used as the analysis filters, and weight functions on $\boldsymbol{g}$ are introduced for explicit guidance. Motivated by this observation, we propose a generalized analysis representation model to include weight function, and provide a task-driven learning method to learn the weight functions, analysis filters, and penalty functions from training data.

3. Proposed method

In this section, we introduce our guided depth image enhancement method. First, a weighted analysis sparse representation model is suggested to introduce guidance information from intensity image. We then provide a task driven formulation, resulting in a bi-level optimization problem. Finally, stage-wise model parameters are learned from training data.

3.1. Weighted analysis sparse representation

For conventional analysis sparse representation in Eqn. (1), the regularization term is only a function of the output image $\boldsymbol{x}$. To introduce guidance information from intensity image, we refer to the models in Eqns. (2) and (3), and generalize the analysis model by including a weight function. Instead of handcrafted guidance, we adopt a parametric form of the weight function, and the parameters can be learned in a task-driven manner.

We define the weight function for the l-th analysis operator at position i as $w_{l,i}(\boldsymbol{g})$. It is known that the depth and intensity discontinuities often co-occur. Thus, $w_{l,i}(\boldsymbol{g})$ is defined based on the local structure of intensity image, such that $\boldsymbol{w}_i \to 1$ for smooth region, and $\boldsymbol{w}_i \to 0$ when discontinuity occurs. The resulting weighted analysis model will penalize depth discontinuities when the corresponding intensity region is smooth, and allow sharp depth jumps when the intensity region exhibits strong discontinuities.

Although the intensity and the depth images arise from the same scene are strongly dependent, the values in the two images have different physical meaning. For example, a black box in front of a white wall or a gray box in front of a black wall may correspond to the same depth map but totally different edge gradients for intensity images. Therefore, the weight function should be able to avoid the interference of such structure-unrelated intensity information, while extracting useful salient structures to help the depth map locate its discontinuities. To this end, local normalization on intensity map is employed to avoid the effect of different intensity magnitude. Specifically, given the guided intensity image $\boldsymbol{g}$, we introduce the operator $\boldsymbol{R}_i$ to extract the local patch in position i by $\boldsymbol{R}_i\boldsymbol{g}$. The local normalization of $\boldsymbol{R}_i\boldsymbol{g}$ can then be attained by $\boldsymbol{e}_i = \frac{\boldsymbol{R}_i\boldsymbol{g}}{||\boldsymbol{R}_i\boldsymbol{g}||_2}$.

With $\boldsymbol{e}_i$, we define the weight function for the l-th analysis operator $\boldsymbol{\beta}_l$ at position i as,

$$w_{l,i}(\boldsymbol{g}) = \exp\left(-(\boldsymbol{\beta}_l^T \boldsymbol{e}_i)^2\right). \tag{4}$$

The analysis operator $\boldsymbol{\beta}_l$ can serve as a special local structure detector. If the local normalized patch $\boldsymbol{e}_i$ contains local structure such as edges, $w_{l,i}(\boldsymbol{g})$ will be very small to encourage that the depth patch exhibits the corresponding local structure. By introducing the weight function $w_{l,i}(\boldsymbol{g})$, we define the weighted analysis sparse representation as,

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) + \sum\nolimits_i \sum\nolimits_l w_{l,i}(\boldsymbol{g})\rho_l((\boldsymbol{k}_l \otimes \boldsymbol{x})_i)\,. \tag{5}$$

Based on the weighted analysis sparse representation model, we further provide the task-driven formulation for guided depth image enhancement, and suggest to learn the parameters $\{\rho_l, \boldsymbol{\beta}_l, \boldsymbol{k}_l\}_{l=1\cdots L}$ from training data.

3.2. Task driven formulation

In the weighted analysis representation model, the data fidelity term is specified by the depth enhancement task. This work considers two representative tasks, i.e., depth upsampling and hole filling. And their fidelity terms take the following form,

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) = \frac{\tau}{2}\|\boldsymbol{M}^{\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{y})\|_2^2. \tag{6}$$

where $\boldsymbol{M}$ is a diagonal matrix and τ is a tradeoff parameter. For depth upsampling, the diagonal elements in $\boldsymbol{M}$ indicate the corresponding points between high resolution estimation $\boldsymbol{x}$ and aligned low resolution input $\boldsymbol{y}$. For hole filling, the diagonal element in $\boldsymbol{M}$ is binary to indicate whether the pixel is observable or not.

Given $\mathcal{L}(\boldsymbol{x},\boldsymbol{y})$, one natural solution is to perform depth image enhancement by minimizing the model in Eqn. (5). However, the model parameters $\{\rho_l, \boldsymbol{\beta}_l, \boldsymbol{k}_l\}_{l=1\cdots L}$ remain unknown and should be learned from training data. Moreover, the model parameters may vary for different tasks. Thus, we provide a task-driven formulation of weighted analysis sparse representation, which allows us learn model parameters to specific task [25, 2].

Denote by $\mathcal{D} = \{\boldsymbol{y}^{(s)}, \boldsymbol{x}^s_{gt}, \boldsymbol{g}^s\}_{s=1}^S$ a training set of S samples. $\boldsymbol{y}^{(s)}, \boldsymbol{x}^s_{gt}$, and $\boldsymbol{g}^s$ denote the s-th input depth image, ground truth depth image, and ground truth intensity image, respectively. Following [25, 2], the task-driven formulation can be written as a bi-level optimization problem,

$$\begin{aligned}\{\rho_l^*, \boldsymbol{\beta}_l^*, \boldsymbol{k}_l^*\}_{l=1}^L = \arg \min_{\{\rho_l, \boldsymbol{\beta}_l, \boldsymbol{k}_l\}_{l=1}^L} \sum\nolimits_{s=1}^S \|\boldsymbol{x}^s_{gt} - \boldsymbol{x}^s\|_2^2 \\ s.t.\ \boldsymbol{x}^s = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}^s) + \sum\nolimits_l \sum\nolimits_i w_{l,i}(\boldsymbol{g}^s)\rho_l((\boldsymbol{k}_l \otimes \boldsymbol{x})_i)\end{aligned} \tag{7}$$

By solving the model above, we can obtain the task-specific model parameters $\{\rho_l, \boldsymbol{\beta}_l, \boldsymbol{k}_l\}_{l=1\cdots L}$.

3.3. Stage-wise model parameter learning

The lower-level problem in Eqn. (7) defines an implicit function on $\{\rho_l, \boldsymbol{\beta}_l, \boldsymbol{k}_l\}_{l=1\cdots L}$, making the training problem very difficult to optimize. The high non-convexity of the lower-level problem further adds difficulty in obtaining the exact solution. Moreover, along with the enhancement procedure, more details of $\boldsymbol{x}^s$ will be recovered. Thus, instead of employing the same model parameters in all the iterations, dynamic guidance by learning stage-wise parameters may benefit both efficiency and enhancement result. Actually, stage-wise learning has been adopted in different applications. Gregor et al. [8] show that a deep architecture can be obtained based on the truncated version of optimization algorithm with fixed iterations to approximate the sparse coding process. Based on half-quadric splitting (gradient descent), [33] ([3]) learns a few stage-wise operations to deal with natural image restoration problems, and achieves the state-of-the-art performance.

Following [8, 33, 3], we use the gradient descent method to solve the lower-level problem, and adopt a greedy learning strategy to learn stage-wise parameters. Assume that both the model parameters $\{\{\rho_l^1, \boldsymbol{\beta}_l^1, \boldsymbol{k}_l^1\}_{l=1\cdots L}, \cdots, \{\rho_l^t, \boldsymbol{\beta}_l^t, \boldsymbol{k}_l^t\}_{l=1\cdots L}\}$ and the enhancement results $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t\}$ are known. Using gradient descent, the updated result $\boldsymbol{x}_{t+1}$ can be obtained by,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \left(\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{y}) + \sum\nolimits_l \bar{\boldsymbol{k}}_l^{t+1} \otimes (\boldsymbol{W}_l^{t+1}\rho_l^{t+1\prime}(\boldsymbol{k}_l^{t+1} \otimes \boldsymbol{x}_t))\right), \tag{8}$$

where $\boldsymbol{W}_l^{t+1}$ is with the same size of $\boldsymbol{k}_l^{t+1} \otimes \boldsymbol{x}_t$, and its value in position i is $w_{l,i}^{t+1}(\boldsymbol{g})$. $\bar{\boldsymbol{k}}_l^{t+1}$ is obtained by rotating $\boldsymbol{k}_l^{t+1}$ 180 degree.

So far, the penalty function $\rho_l^{t+1}(z)$ is still not parameterized. One possible choice is to use the existing regularizers in literature [32, 35, 30, 4, 28, 6]. Actually, from Eqn. (8), one can see that what we should parameterize is not the penalty function $\rho_{l,t+1}(z)$ but the influence function $\rho_l^{t+1\prime}(z)$. Here we allow the influence function $\rho_l^{t+1\prime}(z)$ to have more flexible shapes by parameterizing it with,

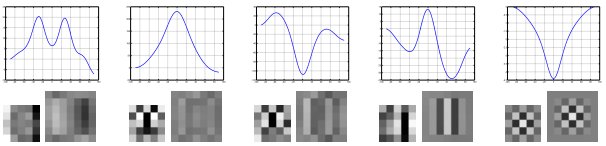
$$\rho_l^{t+1\prime}(z) = \sum\nolimits_j^M \alpha_{l,j}^{t+1} \exp\left(\frac{-(z-\mu_j)^2}{2\gamma_j^2}\right), \tag{9}$$

which is the summation of M Gaussian RBF kernels with center μ_j and scalar factor γ_j. This formulation can provide a group of highly flexible functions for image restoration [33, 3].

Let $\boldsymbol{\alpha}_l^{t+1} = \{\alpha_{l,j}^{t+1}\}_{j=1}^M$. Then $\boldsymbol{x}_{t+1}$ can be explicitly written as a function of $\boldsymbol{\Theta}^{t+1} = \{\tau^{t+1}, \{\boldsymbol{\alpha}_l^{t+1}, \boldsymbol{\beta}_l^{t+1}, \boldsymbol{k}_l^{t+1}\}_{l=1}^L\}$, i.e., $\boldsymbol{x}_{t+1}(\boldsymbol{\Theta}^{t+1})$. Thus, we adopt a greedy training strategy to learn the stage-wise parameters $\boldsymbol{\Theta}^{t+1}$ by solving the following problem,

$$\boldsymbol{\Theta}^{t+1} = \arg\min_{\boldsymbol{\Theta}} \frac{1}{2}\sum\nolimits_{s=1}^S ||\boldsymbol{x}_g^s - \boldsymbol{x}_{t+1}^s(\boldsymbol{\Theta})||_2^2\,. \tag{10}$$

The gradient of the loss function with respect to the parameters $\boldsymbol{\Theta}^{t+1}$ can be calculated by the chain rule. Then the LBFGS method [23, 26] is used to learn the parameters for each stage. We experimentally found that we can get very



(a) (b) (c) (d) (e)

Figure 2. Part of learned parameters. In each sub-figure, the upper part is the regressed penalty function by $\boldsymbol{\alpha}_i$; the lower left part is the analysis filters $\boldsymbol{k}_i$ for depth map; and the lower right part is the analysis filters $\boldsymbol{\beta}_i$ for guided intensity image.

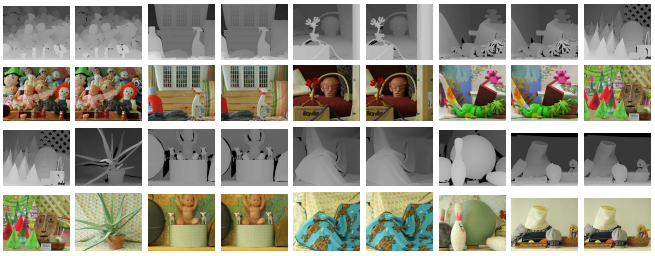


Figure 3. The training samples used in the guided depth upsampling experiments.

good results after only a few stages of process, e.g., T. After greedy learning, joint training is further utilized to learn the parameters of the T stages simultaneously.

3.4. Discussion

Our weighted analysis sparse representation model can provide a flexible model to characterize the complex and dynamic dependency between guided intensity and output depth images. Experiments are conducted to validate this claim by using noisy depth upsampling with factor 4 as an example. The detailed experimental setting will be introduced in Section 4. Fig. 2 shows five of the 24 groups of learned parameters, i.e., penalty function, and analysis filters for intensity and depth images. The analysis filters for intensity and depth images are reshaped for better visualization.

In Eqns. (2) and (3), the same pixel-wise difference operator is used to model the co-occurrence of intensity and depth discontinuities. From Fig. 2, one can see that the correlation between intensity and depth images is more complex, and the corresponding analysis filters for intensity and depth images are quite different. Moreover, previous hand-crafted models usually adopt some monotone shrinkage functions on filter responses to promote smoothness in the estimation. Instead, the penalty functions learned by our models are much more complex. Some learned functions clearly show expansion behaviour, making our model able to generate high quality depth map with sharp edges.

4. Experiments on depth map upsampling

In this section, we compare the proposed method with other depth upsampling methods. Three common used datasets (Middlebury [14], NYU[27] and ToFMark [6]) are utilized to evaluate the depth upsampling performance of the proposed method. Besides the baseline bicubic and bilinear upsampling methods, we compare the proposed methods with a variety of guided upsampling methods. The comparison methods include two filtering based methods [37, 13], some optimization based method: MRF based method [4], non-local mean regularized depth upsampling method [28], total generalized variation (TGV) method [6], the joint static and dynamic filtering(SDF) method [12], and recent proposed CNN-based deep joint filtering method [22]. The root mean square error (RMSE) indexes by recent proposed deep learning based methods are also included. Detailed experimental setting will be introduced in the following subsections.

4.1. Upsampling results on the Middlebury dataset

The *Art*, *Books* and *Moebius* images in the Middlebury dataset [14] have been widely utilized to evaluate depth restoration algorithms. Following the experimental setting of [6], we conduct upsampling experiments with both the noise-free and noisy low resolution depth map on four zooming factors, i.e. 2, 4, 8, 16. For the noise-free experiments, both the training and testing samples are generate by a bicubic resizing of the high quality depth maps. While, for the noisy experiments, the noisy low-resolution depth maps are from [28], and we prepare training noisy low resolution depth map by adding white Gaussian noise with standard variation 6 to the clean low resolution depth images.

To prepare training data, we choose 18 depth and intensity image pairs in the Middlebury data set [14] and extract 300 72×72 small images as the training dataset. The 18 images are shown in Fig. 3, we can see that some images are just a change of viewpoint for the same scene, our training data set actually only contained limited samples. We thus further extend the training data set by flipping and rotating the original image. After extension, we get 1200 small images of resolution 72×72 in the training data set. Although the extension improves the structure variety of the training samples, the training data is still not diverse enough because the original training images only contain limited kinds of colors. In our experiments, instead of using RGB image, we only use gray intensity image to guide the restoration.

Besides the model parameters we aim to learn from training data, there are still some algorithm parameters, e.g. the number of filters and the filter size for the depth map and intensity map. Generally, larger filters are able to model local structure relationship of larger area, with enough training data, they will lead to better performance. However, utilizing large filters often demands large number of filters to model local structural prior, which will greatly increase the computational burden in both the training and testing phase. To take a balance between efficiency and upsampling per-

Table 1. Experimental results (RMSE) on the 3 noise-free test image.

	Art				Books				Moebius			
	×2	×4	×8	×16	×2	×4	×8	×16	×2	×4	×8	×16
Bicubic	2.57	3.85	5.52	8.37	1.01	1.56	2.25	3.35	0.91	1.38	2.04	2.95
Bilinear	2.83	4.15	6.00	8.93	1.12	1.67	2.39	3.53	1.02	1.50	2.20	3.18
GF [13]	2.93	3.79	4.97	7.88	1.16	1.58	2.10	3.19	1.10	1.43	1.88	2.85
MRF [4]	3.12	3.79	5.50	8.66	1.21	1.55	2.21	3.40	1.19	1.44	2.05	3.08
Yang [37]	4.07	4.06	4.71	8.27	1.61	1.70	1.95	3.32	1.07	1.39	1.82	2.49
Park [28]	2.83	3.50	4.17	6.26	1.20	1.50	1.98	2.95	1.06	1.35	1.80	2.38
TGV [6]	3.03	3.79	4.79	7.10	1.29	1.60	1.99	2.94	1.13	1.46	1.91	2.63
SDF [12]	3.31	3.73	4.60	7.33	1.51	1.67	1.98	2.92	1.56	1.54	1.85	2.57
DJF [22]	2.77	3.69	4.92	7.72	1.11	1.71	2.16	2.91	1.04	1.50	1.99	2.95
Ours	**0.89**	**2.00**	**3.84**	**6.16**	**0.47**	**0.91**	**1.68**	**2.67**	**0.45**	**0.84**	**1.54**	**2.34**

Table 2. Experimental results (RMSE) on the 3 noisy test image.

	Art				Books				Moebius			
	×2	×4	×8	×16	×2	×4	×8	×16	×2	×4	×8	×16
Bicubic	5.32	6.07	7.27	9.59	5.00	5.15	5.45	5.97	5.34	5.51	5.68	6.11
Bilinear	4.58	5.62	7.14	9.72	3.95	4.31	4.71	5.38	4.20	4.57	4.87	5.43
GF [13]	3.55	4.41	5.72	8.49	2.37	2.74	3.42	4.53	2.48	2.83	3.57	4.58
MRF [4]	3.49	4.51	6.39	9.39	2.06	3.00	4.05	5.13	2.13	3.11	4.18	5.17
Yang [37]	3.01	4.02	4.99	7.86	1.87	2.38	2.88	4.27	1.92	2.42	2.98	4.40
Park [28]	3.76	4.56	5.93	9.32	1.95	2.61	3.31	4.85	1.96	2.51	3.22	4.48
TGV [6]	3.19	4.06	5.08	7.61	1.52	2.21	2.47	3.54	1.47	2.03	2.58	**3.56**
Chan [1]	3.44	4.46	6.12	8.68	2.09	2.77	3.78	5.45	2.08	2.76	3.87	5.57
SDF [12]	3.36	3.86	4.93	7.85	1.59	1.92	2.60	4.16	1.64	1.85	2.67	4.21
Ours	**1.84**	**2.96**	**4.41**	**7.06**	**1.18**	**1.64**	**2.35**	**3.50**	**1.34**	**1.74**	**2.57**	3.79

Table 3. Experimental results (RMSE) on the 449 NYU test image.

	NYU		
	×4	×8	×16
MRF [4]	4.29	7.54	12.32
GF [13]	4.04	7.34	12.23
JBU [16]	2.31	4.12	6.98
TGV [6]	3.83	6.46	13.49
Park [28]	3.00	5.05	9.73
SDF [11]	3.04	5.67	9.97
DJF [22]	1.97	3.39	5.63
Ours	**1.56**	**2.99**	**5.24**

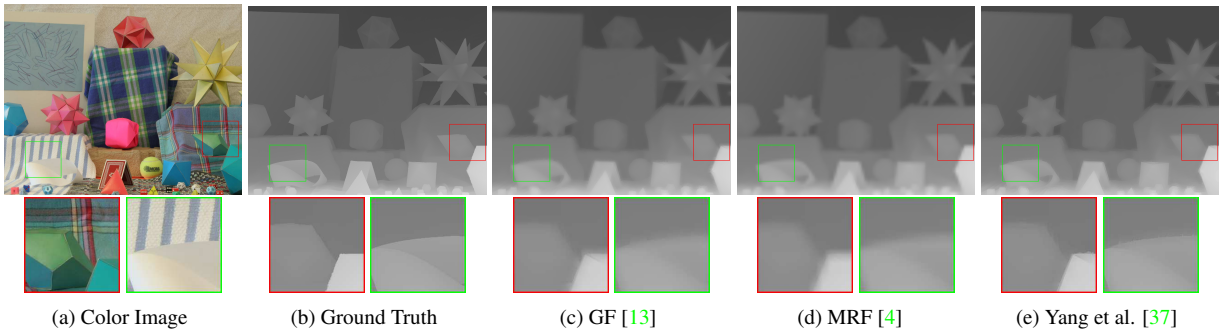
formance, we use 24 5×5 analysis filters $\{\boldsymbol{k}_l\}_{l=1...L}\}$ for depth image. For the filters $\{\boldsymbol{\beta}_l\}_{l=1...L}\}$ used to extract information from intensity image, we set their size as 7×7. For both the noisy and noise-free cases, we use the results by bicubic interpolation as the initialization of $\boldsymbol{x}_0$. Our experimental results show that the proposed model is able to generate very good upsampling results in a few steps. For the noise free upsampling experiemts, we set the stage number for zooming factor 2, 4, 8 and 16 as 4, 5, 6 and 7. While for the noisy upsampling experiments, the stage numbers are set as 6, 8, 10 and 12. Adding extra stages will further improve the training loss, but surfers from more computation burden in both the training and testing phase.

The upsampling experimental results on the 3 noise-free testing images by different methods are shown in table 1. The proposed method consistently shows its advantage over the competing methods, it achieves the best results on all the 3 images with different zooming factors. In Fig. 4, we give visual examples of the upsampling results on the moebius image with zooming factor 16. In the figure we can see that the guided filter method [13] and the MRF method [4] can not generate very sharp edges; while, the results by [37][28] and [6] have some artifacts around the edge area. Our method is able to generate high quality depth map with sharper edges and less artifacts.

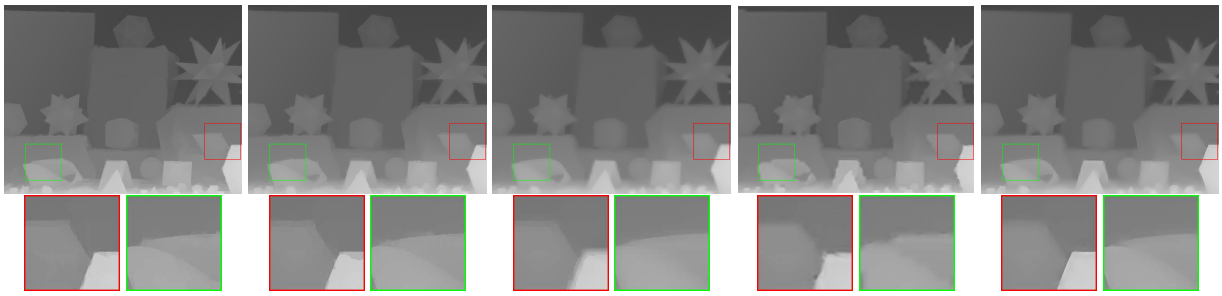
We further evaluate the proposed method by noisy depth maps upsampling experiments. The results by different methods are shown in Table 2, we do not provide the results by DJF [22] because the authors have not provide their network as well as results on such setting. The results by [1] is also included, which is designed to handle noise in depth super-resolution problem. The proposed method again achieves the best results.

4.2. Upsampling results on the NYU dataset

In [22], Li et al. utilize the first 1000 images of NYU dataset [27] as training data, and evaluate their DJF method on the last 449 images of the NYU dataset. In this section, we follow their experimental setting and compare different methods on the 448 images. The results by the other methods are provided by the authors of [22]. To save the training time, we train our model with only the first 100 images from the training data set of [22]. The number and size of the filters for the NYU dataset are the same as our settings on the Middleburry [14] dataset. While, the stage number for all



(a) Color Image (b) Ground Truth (c) GF [13] (d) MRF [4] (e) Yang et al. [37]



(f) Park et al. [28] (g) TGV [6] (h) SDF [12] (i) DJF [22] (j) Ours

Figure 4. Depth restoration results by different methods based on noise-free data (Moebius).

the zooming factors 4, 8 and 16 are set as 4.

The experimental result are shown in Table 3. Compared with other methods, the proposed method achieves the best results in terms of RMSE.

4.3. Upsampling results on real Sensor Data

Besides synthetic data, we also evaluate the proposed method on real sensor dataset [6]. In which a Time of Flight (Tof) and a CMOS camera are used to obtain low resolution depth maps and intensity images, and the ground truth depth images are generated by a structured light scanner.

We utilize the same 18 images in Fig. 3 as the training images. The noise in the low resolution input of ToFMark dataset is different from previous synthetic data. To generate similar low resolution input for training, we first use a t location-scale distribution to fit the residuals between input and groundtruth data, and then generate additive noise by the distribution parameters. Since the missing values in the depth map are represented as zeros, which may be termed as very sharp edge in the depth map. We use a simple masked joint bilateral filtering [29] method to generate initialization values for the unknown points in the depth map. Although such initialization $\boldsymbol{x}_0$ is still very noisy, our method can still generate very good results in just several stages. In addition, we adopt larger size filters (7×7 filters $\boldsymbol{k}_i$ for depth image and 9×9 filters β_i for intensity image) to further improve the performance of the proposed method.

The restoration results are showed in Table 4. We compare our method with other classic or state-of-art methods. Table 4 shows that our method gets better result in terms of the RMSE. From Fig. 5, it is easy to see that our method is capable of generating clean upsampling estimation, while, the results by other methods would copy irrelevant textures from intensity image.

Table 4. Experimental results (RMSE) on the 3 test images in [6]

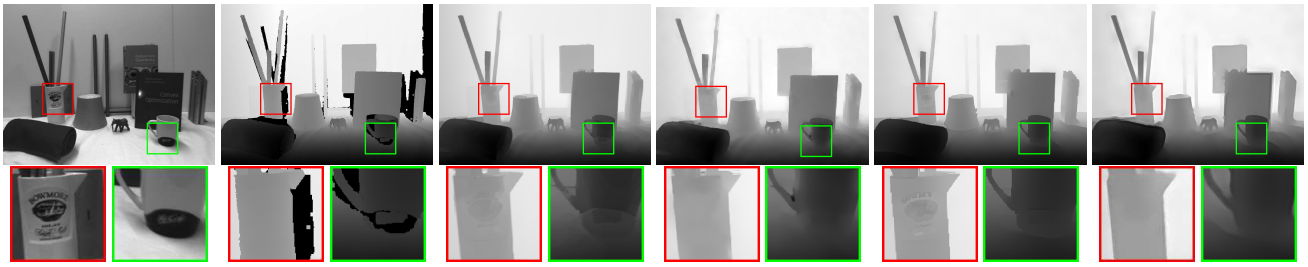
	Books	Shark	Devil
Nearest Neighbor	18.21	21.83	19.36
Bilinear	17.10	20.17	18.66
Kopf [16]	16.03	18.79	27.57
He [18]	15.74	18.21	27.04
TGV [6]	12.36	15.29	14.68
Yang [17]	**12.25**	14.71	13.83
SDF [12]	12.66	14.33	10.68
Ours	12.31	**14.06**	**9.66**

Table 5. Experimental results (RMSE) on the 3 test images in [24].

	Lu et al. [24]	Shen et al. [34]	Ours
Art	6.77	5.65	**4.96**
Books	2.24	2.24	**1.66**
Moebius	2.18	2.27	**1.76**

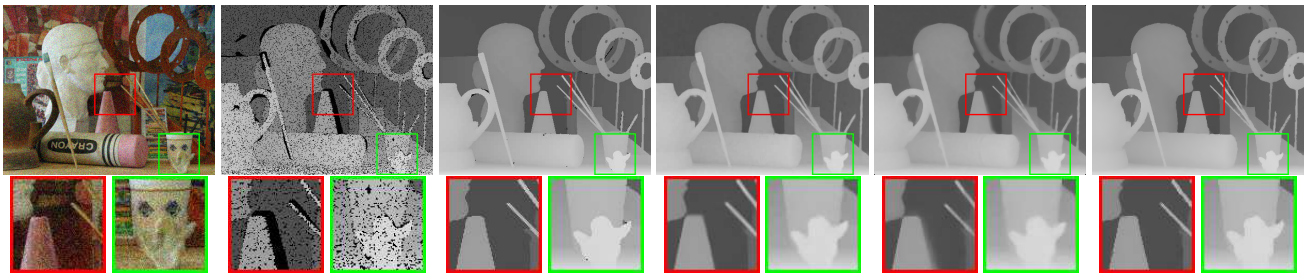
5. Experiments on noisy depth map restoration

In this section, we provide some experimental results on other depth map restoration problems. The dataset in [24] is used to test the proposed method, in which not only additive Gaussian noise but also some missing values are contained in the depth image. In the following subsections, we first introduce our experimental settings which including the



(a) Intensity Image (b) Ground Truth (c) TGV [6] (d) Yang et al. [17] (e) SDF [12] (f) Ours

Figure 5. Depth restoration results by different methods based on real data(books).



(a) Color Image (b) Input (c) Ground Truth (d) Lu et al. [24] (e) Shen et al. [34] (f) Ours

Figure 6. Depth restoration results by different methods.

preparation of training data, the setting of initialization and some algorithm parameters. Then, we compare our method with other methods designed for this task, which include low rank based method [24] and recently proposed mutual-structure joint filtering method [34].

5.1. Experimental setting

Lu et al. [24] provided a synthetic data set to evaluate depth restoration methods. 30 depth and RGB image pairs in the Middlebury database [14] are included in the data set. The size of all the images have been normalized to the same height 370. Zero mean additive Gaussian noise with stand deviation 25 and 5 have been added into the RGB and depth image, respectively. The author [24] have manually set 13% of pixels in depth map as missing values to simulate the depth map acquired from consumer level depth sensor.

To compare the proposed method with other methods, we take the *Art*, *Books* and *Moebius* as testing images, and use the remaining 27 images as the training images. Our proposed method does not consider the noise in the RGB image, for fair comparison, we pre-process the RGB image by a state-of-the-art denoising method [10, 9] and use the denoised image to guide the restoration of depth map. Such a method has been utilized in the original paper [24] to compare with other depth restoration methods.

The filter number and size setting in this noisy depth map restoration experiments is the same as it in the Middleburry upsampling experiment. The same as our setting in the ToF dataset [6], we also adopt JBF [29] to provide initial values for the missing data. Since less training data is provided in this dataset, we only adopt 4 stages to enhance the input depth image.

5.2. Experimental results

The restoration results by different methods are shown in Tabel 5. The results by [24] and [34] are downloaded from the author's websites. The proposed method shows significant advantage over the competing methods in terms of RMSE.

In Fig. 6, we give some visual examples of the restoration results. One can clearly see that our restoration method is able to generate sharp edges as well as remove noise in the input image.

6. Conclusions

To better modeling the dependency between intensity and depth map, we proposed a weighted analysis representation model for guided depth reconstruction. An intensity weighting term and an analysis representation regularization term are combined to model complex relationship between depth image and RGB image. We utilized a task driven training strategy to learn stage-wise parameters for specific task, the proposed model is able to generate high quality depth restoration results in a few stages. Compared with other state-of-the-art methods on both the guided depth upsampling and restoration problems, the proposed model achieved better results with less RMSE value and more pleasant visual quality.

References

[1] D. Chan, H. Buisman, C. Theobalt, and S. Thrun. A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008. 1, 6

[2] Y. Chen, R. Ranftl, and T. Pock. Insights into analysis operator learning: From patch-based sparse models to higher order mrfs. *IEEE Transactions on Image Processing*, 23(3):1060–1072, 2014. 2, 4

[3] Y. Chen, W. Yu, and T. Pock. On learning optimized reaction diffusion processes for effective image restoration. In *CVPR*, 2015. 2, 3, 4

[4] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *NIPS*, 2005. 1, 3, 4, 5, 6, 7

[5] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007. 2

[6] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *ICCV*, 2013. 1, 3, 4, 5, 6, 7, 8

[7] M. R. Gernot Riegler, David Ferstl and H. Bischof. A deep primal-dual network for guided depth super-resolution. In *BMVC*, 2016. 1

[8] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *ICML*, 2010. 4

[9] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang. Weighted nuclear norm minimization and its applications to low level vision. *IJCV*, 2016. 8

[10] S. Gu, L. Zhang, W. Zuo, and X. Feng. Weighted nuclear norm minimization with application to image denoising. In *CVPR*, 2014. 8

[11] B. Ham, M. Cho, and J. Ponce. Robust image filtering using joint static and dynamic guidance. In *CVPR*, 2015. 3, 6

[12] B. Ham, M. Cho, and J. Ponce. Robust image filtering using joint static and dynamic guidance. In *CVPR*, 2015. 5, 6, 7, 8

[13] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013. 1, 5, 6, 7

[14] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007. 5, 6, 8

[15] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, pages 559–568, 2011. 1

[16] D. L. J. Kopf, M. F. Cohen and M.Uyttendaele. Joint bilateral upsampling. In *ACM transactions on graphics*, volume 26, 2007. 6, 7

[17] K. l. C. H. Y. W. J. Yang, X. Ye. Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. In *IEEE*, 2014. 7, 8

[18] J. S. K. He and X.Tang. Guided image filtering. In *ECCV*, 2010. 7

[19] M. Kiechle, S. Hawe, and M. Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *ICCV*, 2013. 3

[20] H. Kwon, Y.-W. Tai, and S. Lin. Data-driven depth map refinement via multi-scale sparse representation. In *CVPR*, 2015. 1

[21] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2008. 3

[22] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep joint image filtering. In *European Conference on Computer Vision*, 2016. 1, 5, 6, 7

[23] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989. 4

[24] S. Lu, X. Ren, and F. Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, 2014. 1, 7, 8

[25] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012. 4

[26] mark schmidt. minfunc, 2013. http://mloss.org/software/view/529/. 4

[27] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 6

[28] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *ICCV*, 2011. 1, 3, 4, 5, 6, 7

[29] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. In *ACM transactions on graphics*, volume 23, pages 664–672, 2004. 7, 8

[30] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, 2009. 2, 4

[31] R. Rubinstein, T. Peleg, and M. Elad. Analysis k-svd: a dictionary-learning algorithm for the analysis sparse model. *IEEE Transactions on Signal Processing*, 61(3):661–677, 2013. 2, 3

[32] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 2, 4

[33] U. Schmidt and S. Roth. Shrinkage fields for effective image restoration. In *CVPR*, 2014. 2, 3, 4

[34] X. Shen, C. Zhou, L. Xu, and J. Jia. Mutual-structure for joint filtering. In *ICCV*, 2015. 1, 7, 8

[35] J.-L. Starck, E. J. Candès, and D. L. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 11(6):670–684, 2002. 2, 4

[36] I. Tosic and S. Drewes. Learning joint intensity-depth sparse representations. *IEEE Transactions on Image Processing*, 23(5):2122–2132, 2014. 1

[37] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *CVPR*, 2007. 5, 6, 7

[38] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1400–1414, 2011. 1

[39] S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998. 2

[40] W. Zuo, D. Ren, S. Gu, L. Lin, and L. Zhang. Discriminative learning of iteration-wise priors for blind deconvolution. In *CVPR*, 2015. 2