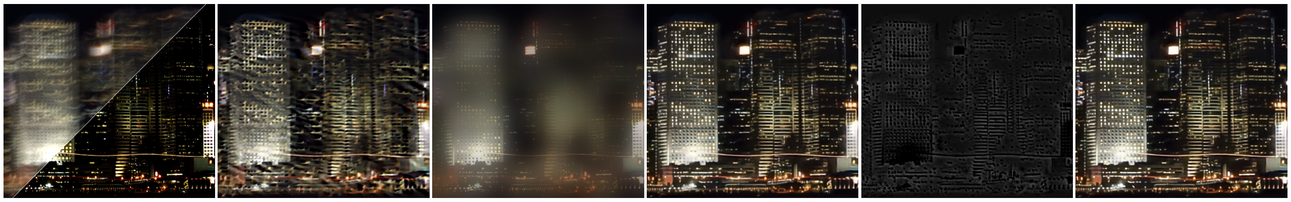
Spatially Variant Linear Representation Models for Joint Filtering

Jinshan Pan[1] Jiangxin Dong[2] Jimmy Ren[3] Liang Lin[4] Jinhui Tang[1] Ming-Hsuan Yang[5,6]
[1]Nanjing University of Science and Technology [2]Dalian University of Technology
[3]SenseTime Research [4]Sun Yat-Sen University [5]UC Merced [6]Google Cloud



(a) Input & guidance (b) Pan et al. [37] (c) He et al. [18] (d) Ours (e) Estimated coefficients

Figure 1. One application of the proposed joint filtering on image deblurring. Our algorithm is based on a spatially variant linear representation model (SVLRM), where the target image (i.e., the deblurred image (d)) can be linearly represented by the guidance image (i.e., the short exposure image in (a)). We develop an efficient algorithm to estimate the linear representation coefficients (i.e., (e)) by a deep convolutional neural network which is constrained by the SVLRM. Our analysis shows that the SVLRM is able to capture the structural details of the input and guidance image well (see (e)). Thus, our method generates better results than those based on the locally linear representation model (e.g., [18]) and favorable results against the state-of-the-art methods on each task (e.g., image deblurring [37]).

Abstract

Joint filtering mainly uses an additional guidance image as a prior and transfers its structures to the target image in the filtering process. Different from existing algorithms that rely on locally linear models or hand-designed objective functions to extract the structural information from the guidance image, we propose a new joint filter based on a spatially variant linear representation model (SVLRM), where the target image is linearly represented by the guidance image. However, the SVLRM leads to a highly ill-posed problem. To estimate the linear representation coefficients, we develop an effective algorithm based on a deep convolutional neural network (CNN). The proposed deep CNN (constrained by the SVLRM) is able to estimate the spatially variant linear representation coefficients which are able to model the structural information of both the guidance and input images. We show that the proposed algorithm can be effectively applied to a variety of applications, including depth/RGB image upsampling and restoration, flash/no-flash image deblurring, natural image denoising, scale-aware filtering, etc. Extensive experimental results demonstrate that the proposed algorithm performs favorably against state-of-the-art methods that have been specially designed for each task.

1. Introduction

Image filters, as fundamental tools in many vision and graphics problems, are mainly used to suppress fine-scale details while preserving primary structures. The linear translation-invariant (LTI) filters usually use spatially invariant kernels such as mean, Gaussian, and Laplacian kernels. As the spatially invariant kernels are independent of image content, these LTI filters usually smooth image structures, details, and noise evenly without discrimination and thus are less effective for preserving main structures [53].

To overcome this problem, joint filtering using additional guidance images has been proposed. Joint filtering aims to transfer the important structural details of the guidance image to the output image so that the important structures of the output image can be preserved in the filtering process. As the guidance image can be the input image itself or the image from different domains [18, 25, 47], joint filtering has been widely used in image editing [27], optical flow [49, 40, 45], stereo matching [40, 43, 31]. Although achieving impressive performance, the joint filtering usually introduces erroneous or extraneous artifacts in the target image when the guidance image and input image are from different domains, such as RGB/depth [56, 39, 12, 25], optical flow/RGB [49, 40, 45], flash/no-flash [54, 18]. Thus, it is of great interest to explore the proprieties of guidance image and input image so that the correct structural information can be transferred in the filtering process.

To explore common structures between the input and guidance images, existing joint filtering methods [43, 17, 16, 22] usually develop kinds of hand-crafted priors to model the structural co-occurrence property. However, using hand-crafted priors usually leads to complex objective functions, which are difficult to solve.

Motivated by the success of deep learning, the joint filtering algorithms based on deep convolutional neural networks (CNNs) [41, 28, 15, 19, 48] have been proposed. These algorithms are efficient and usually outperform conventional methods by large margins. However, they are less effective as using deep CNNs to directly predict the target images may not explore the useful structural details from the guidance image well.

Different from existing methods, we propose a novel joint filtering algorithm based on a spatially variant linear representation model (SVLRM). Instead of directly predicting the target image using a deep CNN, we learn a deep CNN to estimate the spatially variant linear representations coefficients which model the structural information of the guidance and input images and are then used to generate the target image. We show that the proposed algorithm is able to transfer the meaningful structural details of the guidance and input images to the target image and can be applied to a variety of applications, including depth/RGB image upsampling and restoration, flash/no-flash image deblurring, natural image denoising, scale-aware filtering, etc. Figure 1 shows a flash/no-flash image deblurring example where the proposed method generates a clearer image.

The contributions of this work are as follows: (1) we propose the SVLRM for joint filtering, where the target image can be represented by the guidance image with the SVLRM; (2) we develop an efficient optimization method based on a deep CNN (which is constrained by the SVLRM) to estimate the spatially variant linear representation coefficients. Moreover, we analyze that the estimated coefficients model the structural details of input image and guidance well and can determine whether the structures should be transferred to the target image or not; (3) our algorithm achieves state-of-the-art performance on a variety of applications including depth/RGB image upsampling and restoration, flash/no-flash image deblurring, natural image denoising, and scale-aware filtering.

2. Related Work

In this section, we discuss the joint filtering methods most related to this work within proper contexts.

Local joint filtering methods. The representative local joint filtering methods include the bilateral filter (BF) [47, 3, 10], guided filter (GF) [18], weighted median filter (WMF) [31, 59], geodesic distance-based filter [5, 13], weighted mode filter [33], the rolling guidance filter [58], and the mutual structure-based joint filter [43], etc. In these methods, the target image is usually computed by a weighted average of neighboring pixels in the input image. The main success of these methods is due to the use of the locally linear model or different types of affinities among neighboring pixels. For example, the bilateral filter defines the affinity by color difference and spatial distance of the neighboring pixels; the guided filter assumes that the target image can be linearly represented by the guidance image in a local image patch. However, these algorithms usually introduce erroneous or extraneous structures into the target image as only the local structures of the guidance image are explored. Although the common structure has been explored by [43] to solve this problem, this method usually introduces halo artifacts due to the use of the locally linear model [18].

Global joint filtering methods. The global joint filtering algorithms are usually achieved by solving global objective functions. These methods design different handcrafted priors to enforce the target images to have similar structures with the guidance images, e.g., the weighted least squares (WLS) filter [11], total generalized variation (TGV) [12], L_0-regularized prior [51], relative total variation (RTV) [53], scale map scheme [54], and improved RTV [16], etc. Different from the local filters, the algorithms based on such priors are able to exploit global structures in the guidance image. However, using hand-crafted priors may not reflect inherent structural details in the target image. In addition, the objective functions with such priors usually lead to highly non-convex problems which cannot be efficiently solved.

Deep learning-based methods. Image filtering has been significantly advanced due to the use of deep CNNs. Several algorithms use deep CNNs to approximate a number of filters [52, 29, 35, 55, 4, 21, 14]. To deal with depth image upsampling, Hui et al. [19] develop a CNN based on a multi-scale guidance strategy. In [15], Gu et al. use a CNN based on a weighted representation model to dynamically learn structural details from guidance images for the depth image restoration. However, these algorithms are limited to specific application domains. Motivated by the success of GF [18], Wu et al. [48] develop a CNN to approximate the guided image filter. Li et al. [28] propose a joint filter based on an end-to-end trainable network, where the structural information from the guidance image and input image are extracted based on independent CNNs. However, using deep CNNs to estimate target images in a regression way does not always help the structural details transferring as some structural details are smoothed in the convolution process.

Different from existing methods, we propose a joint filtering algorithm based on a SVLRM. The structural information of the guidance image and input image can be effectively transferred to the target image constrained by the spatially variant linear representation coefficients, which are estimated by a deep CNN.

3. Revisiting Guided Image Filtering

To motivate our work, we first revisit the guided image filter [18] and then discuss its role in the filtering process.

Let G, I, and F denote the guidance image, input image, and target image, respectively. The guided filter assumes that the value of the filtered image F at pixel x is represented by the locally linear model,

$$F(\mathrm{x}) = a_k G(\mathrm{x}) + b_k, \quad \mathrm{x} \in \omega_k. \tag{1}$$

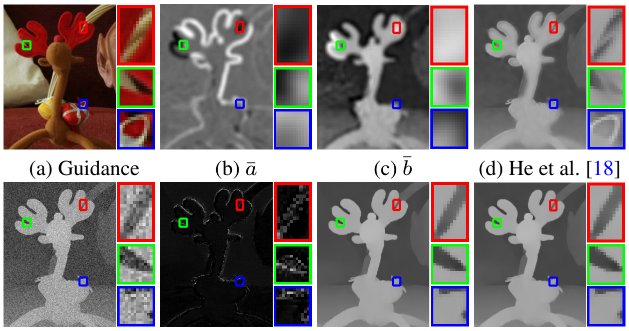


Figure 2. Problems of the locally linear representation model (1) in joint filtering. As shown in (b) and (c), the linear representation coefficients by local image patches do not model the structural information of the guidance image well. By applying the linear representation model (3), the target image (d) contains extraneous textures, and the edges are not preserved well.

The coefficients a_k and b_k are assumed to be constant in each image patch ω_k. By introducing a constraint of a_k (i.e., a_k^2 in [18]), a_k and b_k can be obtained by solving

$$\min_{a_k, b_k} \sum_{\mathrm{x} \in \omega_k} \left((a_k G(\mathrm{x}) + b_k - I(\mathrm{x}))^2 + \gamma a_k^2 \right), \qquad (2)$$

where γ is a positive weight parameter. We note that a_k and b_k can be easily obtained as (2) is a least squares problem. With a_k and b_k in each local image patch, the mean filter is then used to estimate pixel-wise linear coefficients $\bar{a}$ and $\bar{b}$. Finally, the target image is obtained by

$$F(\mathrm{x}) = \bar{a}(\mathrm{x}) G(\mathrm{x}) + \bar{b}(\mathrm{x}). \qquad (3)$$

Although the filtering algorithm with the locally linear model (1) has been demonstrated effective in lots of applications, the assumption of constant a_k and b_k in each image patch usually introduces extraneous textures in the target images (see the parts enclosed in the red and blue boxes of Figure 2(d)). We note that the gradient of the target image and guidance image in each image patch should satisfy

$$\nabla F(\mathrm{x}) = a_k \nabla G(\mathrm{x}), \quad \mathrm{x} \in \omega_k, \qquad (4)$$

according to (1). This constraint ensures that the target image has the similar structures to the guidance image as a_k is a constant. Therefore, the structural details of G are directly transferred to the target image F, which accordingly leads to a target image with extraneous structures from G.

In addition, as the mean filter is further applied to obtain the pixel-wise representation coefficients $\bar{a}$ and $\bar{b}$, this will suppress the high-frequency information which usually corresponds to the important structural details in the guidance image. The parts enclosed in the green boxes in Figure 2(b) and (c) show that the representation coefficients $\bar{a}$ and $\bar{b}$ are over-smoothed. Thus, using such representation coefficients accordingly interferes the structures of the target image (e.g., the edges enclosed in the green box in Figure 2(d) are not sharp).

As the target image F is mainly determined by the representation coefficients $\bar{a}$ and $\bar{b}$, ideal representation coefficients should model the structural details of both guidance and input images well so that they can determine whether the structures of the guidance image G should be transferred to the target image or not.

To solve this problem, we propose a SVLRM and develop a deep CNN to estimate the linear representation coefficients. Figure 2(f) and (g) shows that the estimated spatially variant linear representation coefficients model the structural information of guidance and input image well, which accordingly leads to a better target image.

4. Proposed Algorithm

In this section, we first present the SVLRM and then propose an efficient algorithm based on a deep CNN to solve it for joint filtering.

4.1. Spatially variant linear representation model

Different from the locally linear model (1), we assume that the target image F can be represented by

$$F = \alpha(G, I) G + \beta(G, I), \qquad (5)$$

where $\alpha(G, I)$ and $\beta(G, I)$ are the spatially variant linear representation coefficients which are determined by G and I. The coefficients $\alpha(G, I)$ and $\beta(G, I)$ could determine whether the structural details in G and I should be transferred to F or not (Figure 2(f) and (g)).

4.2. Optimization

Without the assumption of the local constant representation coefficients, estimating $\alpha(G, I)$ and $\beta(G, I)$ from (5) is quite challenging as (5) is highly ill-posed. A common approach is to use the regularization w.r.t. $\alpha(G, I)$ and $\beta(G, I)$ to minimize the following objective function

$$\mathcal{E}(\alpha, \beta) = \|\alpha G + \beta - I\|^2 + \varphi(\alpha) + \phi(\beta), \qquad (6)$$

where $\varphi(\alpha)$ and $\phi(\beta)$ are the constraints of $\alpha(G, I)$ and $\beta(G, I)$. If $\varphi(\alpha)$ and $\phi(\beta)$ are differentiable, the problem (6) can be solved by gradient descent:

$$\alpha^t = \alpha^{t-1} - \lambda \left(\frac{\partial \mathcal{E}(\alpha, \beta^{t-1})}{\partial \alpha} \right)_{\alpha = \alpha^{t-1}}, \qquad (7a)$$

$$\beta^t = \beta^{t-1} - \lambda \left(\frac{\partial \mathcal{E}(\alpha^{t-1}, \beta)}{\partial \beta} \right)_{\beta = \beta^{t-1}}, \qquad (7b)$$

where λ and t denote the step size and the iteration number.

However, it is not trivial to determine $\varphi(\alpha)$ and $\phi(\beta)$ for joint filtering as the properties of $\alpha(G, I)$ and $\beta(G, I)$ are quite different from the statistical properties of natural images [18, 43]. Instead of using hand-crafted constraints for $\alpha(G, I)$ and $\beta(G, I)$, we propose a deep CNN to estimate $\alpha(G, I)$ and $\beta(G, I)$ based on the SVLRM (5).

Table 1. Quantitative evaluations for the depth image upsampling problem on the synthetic benchmark dataset [44] in terms of RMSE.

Methods	Bicubic	MRF [7]	GF [18]	JBU [25]	TGV [12]	3D-TOF [39]	SDF [17]	FBS [1]	DMSG [19]	DJF [28]	Ours
×4	8.16	7.84	7.32	4.07	6.98	5.21	5.27	4.29	3.78	3.54	**1.74**
×8	14.22	13.98	13.62	8.29	11.23	9.56	12.31	8.94	6.37	6.20	**5.59**
×16	22.32	22.20	22.03	13.35	28.13	18.10	19.24	14.59	11.16	10.21	**7.23**

Learning. We note that (7) is in spirit similar to the stochastic gradient descent which is widely used to solve deep CNNs. This motivates us to develop a deep CNN to estimate $\alpha(G, I)$ and $\beta(G, I)$.

Let $\{G^n, I^n, F_{gt}^n\}_{n=1}^N$ denote a set of N training samples and $\mathcal{F}$ denote the deep CNN. Our goal is to learn the network parameters $\Theta = \{\Theta_\alpha, \Theta_\beta\}$ so that $\mathcal{F}_{\Theta_\alpha}$ and $\mathcal{F}_{\Theta_\beta}$ are able to approximate the spatially variant linear coefficients $\alpha(G, I)$ and $\beta(G, I)$.

To this end, we constrain the network $\mathcal{F}$ by the SVLRM (5), which is defined as

$$\mathcal{F}_\Theta(G^n; I^n) = \mathcal{F}_{\Theta_\alpha}(G^n; I^n)G^n + \mathcal{F}_{\Theta_\beta}(G^n; I^n), \quad (8)$$

where $\mathcal{F}_{\Theta_\alpha}(G^n; I^n)$ and $\mathcal{F}_{\Theta_\beta}(G^n; I^n)$ are the results of the network $\mathcal{F}$ w.r.t. parameters Θ_α and Θ_β.

In the training process, we use the L_1-norm as the loss function to constrain the network $\mathcal{F}$, which is defined as

$$\mathcal{L}(\mathcal{F}_\Theta(G^n; I^n); F_{gt}) = \sum_{n=1}^{N} \|\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\|_1. \quad (9)$$

As the L_1-norm is non-differentiable, we use the Charbonnier penalty function $\rho(x) = \sqrt{x^2 + \varepsilon^2}$ to approximate it.

At each training iteration, the gradients of the loss function w.r.t. $\mathcal{F}_{\Theta_\alpha}$ and $\mathcal{F}_{\Theta_\beta}$ are

$$\frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\alpha}} = \sum_{n=1}^{N} \frac{G^n\left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)}{\sqrt{\left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)^2 + \varepsilon^2}}, \quad (10a)$$

$$\frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\beta}} = \sum_{n=1}^{N} \frac{\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n}{\sqrt{\left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)^2 + \varepsilon^2}}. \quad (10b)$$

Based on (10), we update the network parameters by

$$\Theta_\alpha^t = \Theta_\alpha^{t-1} - \lambda \frac{\partial \mathcal{F}_{\Theta_\alpha}}{\partial \Theta_\alpha} \frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\alpha}}, \quad (11a)$$

$$\Theta_\beta^t = \Theta_\beta^{t-1} - \lambda \frac{\partial \mathcal{F}_{\Theta_\beta}}{\partial \Theta_\beta} \frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\beta}}. \quad (11b)$$

After obtaining $\{\Theta_\alpha, \Theta_\beta\}$, we set the spatially variant linear coefficients $\alpha(G, I)$ and $\beta(G, I)$ to be $\mathcal{F}_{\Theta_\alpha}(G; I)$ and $\mathcal{F}_{\Theta_\beta}(G; I)$. Finally, the target image can be obtained by (5). We empirically find that using deep CNNs to estimate $\alpha(G, I)$ and $\beta(G, I)$ is effective (Section 5). More detailed analysis about this is included in Section 6.

Network architecture. Based on above considerations, we can use existing network architectures to define the network $\mathcal{F}$. In this work, we use a CNN with 12 convolution layers. The filter size is set to be 3×3 pixels, and the stride value is set to be 1. The feature number at the first 11 convolution layers are set to be 64. Each convolution layer is followed by ReLU except the final convolution layer.

5. Experimental Results

We evaluate the proposed algorithm on several applications including depth image upsampling, depth image restoration, scale-aware filtering, natural image denoising, and flash image deblurring. The main results are presented in this section, and more results can be found in the supplemental material. The code is publicly available on the authors' websites.

5.1. Parameter settings

In the learning process, we introduce the momentum when updating (11) and use the ADAM optimizer [24] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-4}$. The batch size is set to be 20. The step size λ (i.e., learning rate) is initialized as 10^{-4} which is halved at every minibatch update. The parameter ε is set to be 10^{-3}.

5.2. Depth image upsampling

Training data. For depth image upsampling, we randomly choose 1000 RGB/D image pairs from the NYU depth dataset [44] as the training dataset and follow the protocols of [28] to generate the training data. To evaluate the proposed method, we use the remaining 449 RGB/D image pairs [28] as the test dataset.

We quantitatively and qualitatively evaluate the proposed algorithm against state-of-the-art methods including MRF [7], GF [18], JBU [25], TGV [12], 3D-TOF [39], SDF [17], FBS [1], DMSG [19], and DJF [28]. The quantitative evaluations in Table 1 show that the proposed algorithm performs favorably against state-of-the-art methods.

We show one example from the test dataset in Figure 3. As the GF algorithm [18] is likely to transfer the textures of the guidance image to the depth image according to our analysis in Section 3, the generated result in Figure 3(e) contains extraneous details (e.g., the textures of the flowers). We note that the DJF algorithm [28] uses deep CNNs to learn the dynamic guidance features for joint image upsampling. This algorithm first concatenates the features of the guidance image and input image and then uses a CNN [8] to estimate the target image in a regression way. However, we note the method [8] is less effective for structural details restoration as evidenced by [23]. Thus, the edges of the results by the DJF algorithm [28] are not well estimated as shown in Figure 3(g). Different from the end-to-end trainable CNN-based algorithms, the proposed algorithm explores the SVLRM for joint image filtering and develops a deep CNN to estimate the representation coefficients. Under the guidance of the estimated coefficients, the SVLRM is able to transfer the correct structural details of the

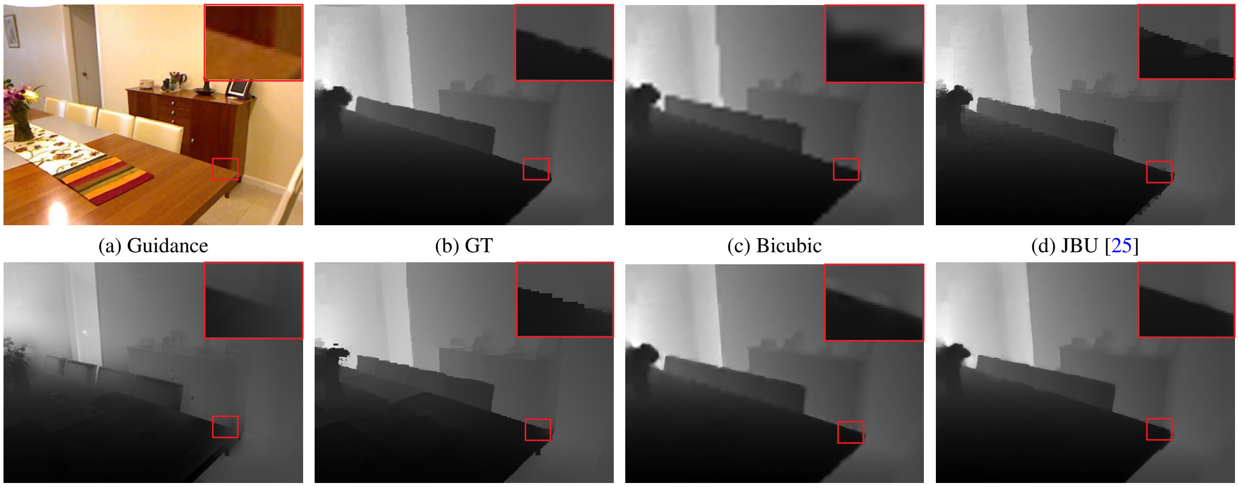


Figure 3. On the depth image upsampling application (×8). The proposed method generates the depth images with sharper boundaries.

guidance image and input image to the target image. Thus, the sharp edges of the super-resolved depth image are preserved well (Figure 3(h)), and the generated results have lower RMSE values (Table 1). All of these indicate the effectiveness of the proposed algorithm.

5.3. Depth image restoration

The proposed algorithm can be applied to depth image restoration.

Training data. To generate the training data for depth image restoration, we use the same training dataset as used in Section 5.2. For each ground truth depth image, we add the Gaussian noise where the noise level ranges from 0 to 10%. We use the test dataset by [30] to evaluate the proposed method, where the training dataset and test dataset do not overlapped. For each test image, we add Gaussian noise with a noise level of 8%.

Table 2 shows the quantitative evaluations against state-of-the-art algorithms. Overall, the proposed method performs favorably against state-of-the-art methods.

Figure 4 shows the depth image denoising results from the evaluated methods. The GF algorithm [18] does not effectively preserve structures as shown in Figure 4(d). We note that the MUJF algorithm by Shen et al. [43] uses mutual structures of the input image and guidance image to avoid the extraneous details in the depth images. However, this method is still based on the locally linear assumption [18] and uses the mean filter to compute the pixel-wise linear representation coefficients. Sharp edges in the restored results are not preserved well (Figure 4(e)) because of less accurate linear representation coefficients. The MUGIF algorithm [16] develops relative structures for joint filtering and generates a better depth image compared to [43]. However, this method does not preserve the sharp edges well as shown in the red boxes of Figure 4(f). The DJF algorithm [28] is able to preserve sharp edges. However, the restored result contains significant artifacts (Figure 4(g)). Instead of using a deep CNN to directly estimates the target image, the proposed algorithm predicts the target image by the SVLRM, where the representation coefficients are estimated by a deep CNN. The generated image contains sharp edges as shown in (Figure 4(h)).

5.4. Scale-aware filtering

With the trained models of the depth image denoising, we show that the proposed algorithm can be straightforwardly applied to scale-aware filtering. Similar to the DJF algorithm [28], we use the input image itself as the guidance image and adopt the rolling guidance strategy [58] to remove small-scale structures and details.

Figure 5 shows an example from [53]. The goal of the scale-aware filtering is to extract meaningful structures from textured surfaces. However, the DJF algorithm [28] and RGF algorithm [58] do not remove the small textures from the input images. The backgrounds of the target images by these two algorithms still contain small scale structures. In contrast, the proposed algorithm removes the small-scale structures from the input images and generates competitive results compared to [53].

5.5. Natural image denoising

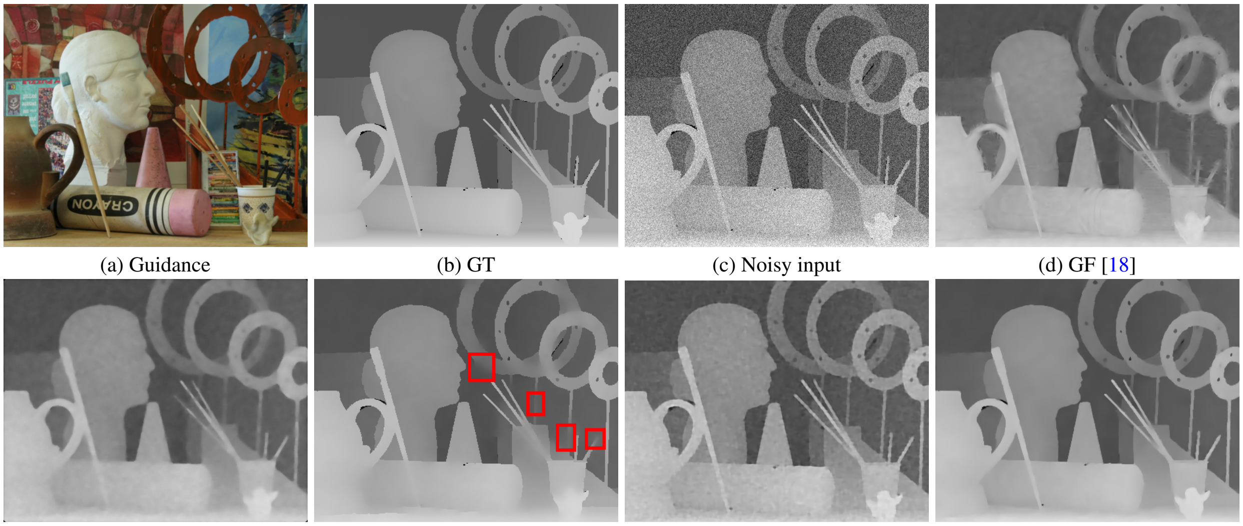
As the guidance image can be the input image itself, we evaluate the proposed algorithm on single natural image denoising.

Training data. To generate training data, we use the training dataset from the BSDS500 dataset [32]. For each clear image, we randomly add the Gaussian noise where the noise level ranges from 0 to 10%. The obtained noisy images are as the inputs of the network. We use the test dataset with 200 clean images by [32] to evaluate the proposed method. The Gaussian noise with a random noise level of 0 to 10% is added to each test image.

We quantitatively and qualitatively evaluate the proposed algorithm against state-of-the-art methods including

Table 2. Quantitative evaluations for the depth image restoration problem on the benchmark dataset [30] in terms of PSNR, SSIM, and RMSE.

Methods	Input	GF	JBU [25]	MUJF [43]	MUGIF [16]	DJF [28]	Ours
Avg. PSNRs	22.03	30.79	26.08	30.67	34.07	32.58	**36.44**
Avg. SSIMs	0.1872	0.9214	0.7820	0.9282	0.9657	0.9016	**0.9762**
Avg. RMSEs	20.18	7.75	12.96	7.76	5.26	6.14	**4.02**


(a) Guidance (b) GT (c) Noisy input (d) GF [18]

(e) MUJF [43] (f) MUGIF [16] (g) DJF [28] (h) Ours

Figure 4. On the depth image restoration application. The parts enclosed in the red boxes in (f) are over-smoothed. The proposed method generates the depth images with sharper boundaries.

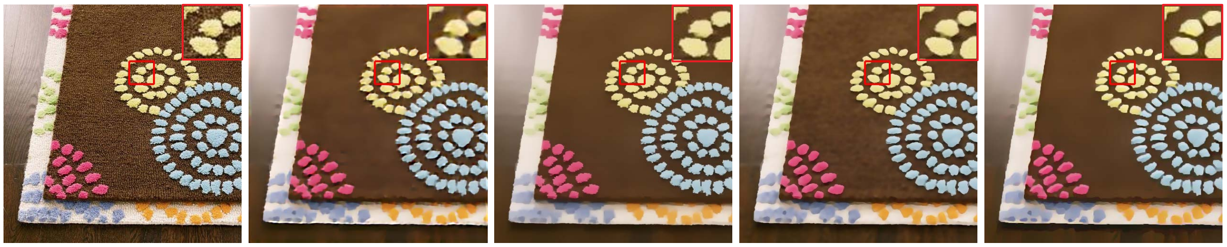

(a) Input image (b) DJF [28] (c) RTV [53] (d) RGF [58] (e) Ours

Figure 5. On the scale-aware filtering application. The comparisons in (b-d) are obtained from the reported results. The proposed algorithm is able to remove small-scale structures while preserving the main sharp edges .

BM3D [6], EPLL [61], CSF [42], MLP [2], and IRCN-N [57]. The quantitative evaluations shown in Table 3 demonstrate that the proposed algorithm is able to generate high-quality images.

Figure 6 shows an example from the test dataset. The structures of the restored results by state-of-the-art methods are over-smoothed. In contrast, the proposed algorithm develops a deep CNN to estimate the spatially variant linear representation coefficients which can determine whether the structural details from the input image are transferred to the target image. Thus, some main structures in the denoised image are preserved well as shown in Figure 6(f).

5.6. Flash image deblurring

In [60], Zhuo et al., propose to deblur a no-flash image under the guidance of its flash image. We show that the proposed method can be applied to this problem. To generate the training data for deblurring, we use the image enhancement dataset by [20] as the flash and no-flash image pairs. We use the algorithm by [9, 36] to generate blur kernels and apply them to the no-flash images to generate blurred images. Finally, we use 100,000 images to train the proposed model.

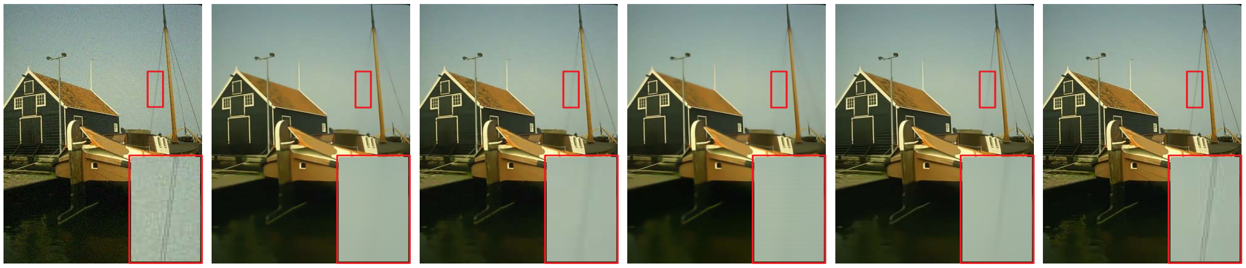
We evaluate the proposed method using real examples by [60] in Figure 7. As the blurred image contains significant blur, single image deblurring methods [50, 26, 34, 38] do not recover clear images. The generated results still contain significant blur and artifacts. We note that using the guidance image is able to help the deblurring problem as shown in Figure 7(g). However, some structural details are not estimated well because only the sparsity of gradient prior is used in image restoration. In contrast, the proposed algorithm is able to remove the blur and generates a clearer image with fine details (Figure 7(h)).

6. Analysis and Discussion

We have shown that the SVLRM with the coefficients learned by a deep CNN for joint filtering outperforms state-

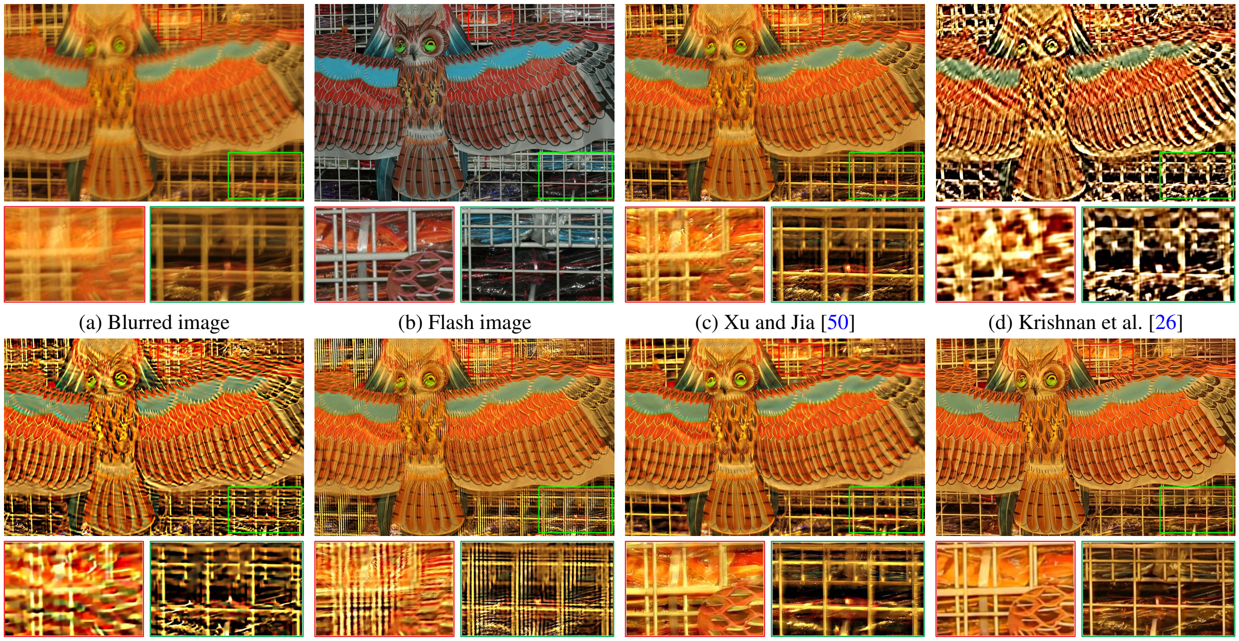
Table 3. Quantitative evaluations for the image denoising problem on the BSDS dataset [32] in terms of PSNR, SSIM, and RMSE.

Methods	Input	BM3D [6]	GF [18]	EPLL [61]	CSF [42]	MLP [2]	IRCNN [57]	Ours
Avg. PSNRs	27.23	31.60	20.35	29.34	30.10	28.91	31.86	**33.04**
Avg. SSIMs	0.6350	0.8765	0.6173	0.8000	0.8164	0.7854	0.8811	**0.8957**
Avg. RMSEs	13.19	6.90	24.85	8.96	8.43	9.39	6.64	**6.32**



(a) Noisy input (b) EPLL [61] (c) CSF [42] (d) MLP [2] (e) IRCNN [57] (f) Ours

Figure 6. On the image denoising application. The proposed method generates the images with clearer structures.



(a) Blurred image (b) Flash image (c) Xu and Jia [50] (d) Krishnan et al. [26]

(e) Pan et al. [34] (f) Pan et al. [37] (g) Zhuo et al. [60] (h) Ours

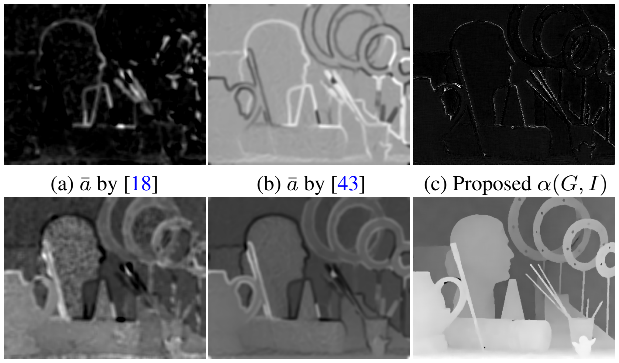
Figure 7. On the image deblurring application. The proposed algorithm is able to generate effective representation coefficients. Thus the deblurred image contains clearer structures and textures.

of-the-art methods on a variety of applications. In this section, we further analyze the effect of the proposed algorithm and compare it with the most related methods.

Relation with locally linear model-based methods. Several notable methods (e.g., [43]) improve the original GF algorithm [18] based on the locally linear model (4). In [43], Shen et al. use the mutual structures of the guidance image and input image to estimate the linear representation coefficients a_k and b_k. The estimated linear representation coefficients contain sharper structures than those of GF [18] and do not introduce additional textures. Thus, the text-copy effect is avoided from the comparisons in Figure 4(d) and (e). However, as the mean filter used in the estimate of linear representation coefficients may smooth the important edge information (Figure 8(b) and (e)), the smoothed structures may affect the sharp edge restoration (Figure 4(e)).

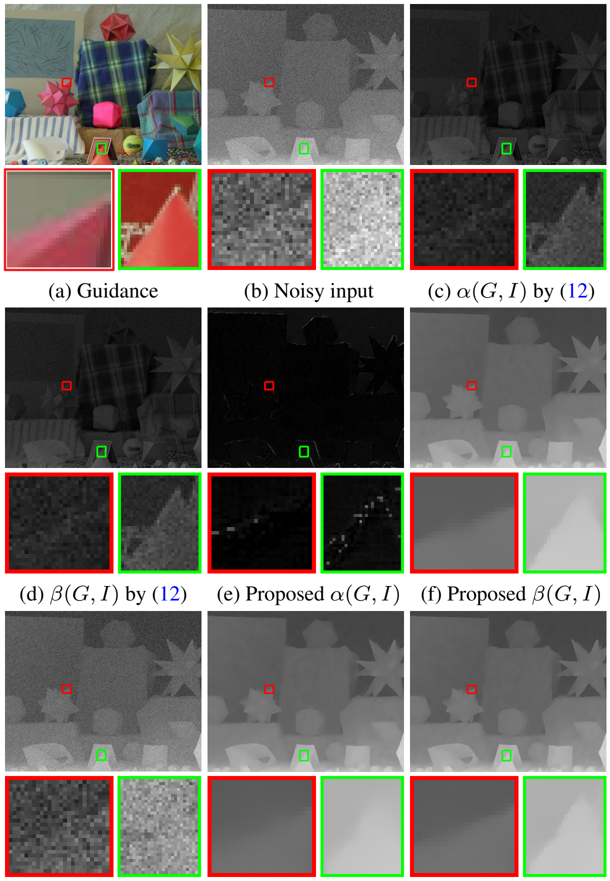
In contrast, the proposed algorithm is based on the SVLRM. The representation coefficients are estimated by a deep CNN. Our estimated linear representation coefficients in Figure 8(c) and (f) are able to better model the structural details of the guidance image and input image, thus facilitating depth image restoration (Figure 4(h)).

Effect of the spatially variant linear representation model. Instead of directly using an end-to-end-trainable network for joint filtering, we propose a new algorithm that learns the SVLRM. The SVLRM is able to capture the structural information of guidance image to help joint filtering. To demonstrate the effect of the proposed linear representation model, we compare it with the method by an end-to-end trainable network (E2ETN). We disable the coefficient learning step and directly estimate the desired output in our implementation to ensure fair comparisons. As


(a) $\bar{a}$ by [18] (b) $\bar{a}$ by [43] (c) Proposed $\alpha(G, I)$

(d) $\bar{b}$ by [18] (e) $\bar{b}$ by [43] (f) Proposed $\beta(G, I)$

Figure 8. The effect of the proposed SVLRM. The guidance image and noisy image are shown in Figure 4(a) and (c). The proposed deep CNN is able to learn the linear representation coefficients which contain the important structural information for joint filtering (Best viewed on high-resolution display with zoom-in).


(a) Guidance (b) Noisy input (c) $\alpha(G, I)$ by (12)

(d) $\beta(G, I)$ by (12) (e) Proposed $\alpha(G, I)$ (f) Proposed $\beta(G, I)$

(g) HC (h) E2ETN (i) Ours

Figure 9. The effect of the proposed SVLRM on depth image denoising (Best viewed on high-resolution display with zoom-in).

shown in Figure 9(h) and (i), the proposed linear representation model learning algorithm generates the results with more sharp edges. In addition, the quantitative evaluations in Table 4 show that the proposed linear representation model consistently improves the performance[1]. All these results concretely demonstrate the effectiveness of the proposed linear representation model learning algorithm.

Table 4. Effect of the proposed SVLRM on image denoising.

	Depth image denoising			Natural image denoising		
	HC	E2ETN	Ours	HC	E2ETN	Ours
Avg. PSNRs	21.84	35.31	**35.98**	24.87	32.44	**33.04**
Avg. SSIMs	0.2044	0.9633	**0.9652**	0.6452	0.8920	**0.8957**

Table 5. Run-time (seconds) performance. All the algorithms are tested on the same machine using the depth image upsampling test dataset.

Methods	JBU [25]	DMSG [19]	DJF [28]	Ours
Avg. run-time	4.8	0.78	1.04	**0.08**

We further note that one alternative approach to estimate $\alpha(G, I)$ and $\beta(G, I)$ is to use hand-crafted priors in (6). Similar to [18, 46], we take $\varphi(\alpha)$ and $\phi(\beta)$ as $\mu\alpha^2$ and $\eta\beta^2$, where μ and η are positive weight parameters. Thus, the solutions of (6) are

$$\alpha = \frac{\eta G I}{\eta G^2 + \mu + \mu\eta}, \qquad \beta = \frac{I - \alpha G}{1 + \eta}. \tag{12}$$

The derivations of (12) and algorithm details of are included in the supplemental material. We empirically set μ and η to be 0.1 on the depth image denoising and natural image denoising problems for fair comparisons. The proposed algorithm with hand-crafted prior (HC for short in Table 4) does not generate better results compared to the method with the deep CNN, indicating the effectiveness of the deep CNN. Moreover, the estimated coefficients in Figure 9(c) and (d) contain significant noise, which accordingly leads to noisy results (Figure 9(g)). In contrast, the proposed algorithm denoises the image well.

Run-time performance. We benchmark the run-time of all methods on a machine with an Intel Core i7-7700 CPU and an NVIDIA GTX 1080Ti GPU. Table 5 shows that the proposed algorithm performs more efficiently than other deep learning-based approaches.

7. Concluding Remarks

In this paper, we have proposed a new joint filter based on the SVLRM and developed an efficient algorithm based on a deep CNN to estimate the linear representation coefficients. The proposed CNN which is constrained by the SVLRM is able to estimate the spatially variant linear representation coefficients. We show that the spatially variant linear representation coefficients model the structural information of both guidance image and input image well. Thus, the linear representation model with the spatially variant representation coefficients is able to transfer meaningful structures to the target image. We show that the proposed algorithm can be effectively applied to a variety of applications and performs favorably against state-of-the-art methods that have been specially designed for each task.

Acknowledgements. This work has been supported in part by the NSFC (No. 61872421, 61732007), the NSF of Jiangsu Province (No. BK20180471), and NSF CAREER (No. 1149783).

[1] The depth image denoising in Table 4 are tested on the dataset by [19], where we add Gaussian noise with a noise level of 8% in each depth image.

References

[1] Jonathan T. Barron and Ben Poole. The fast bilateral solver. In *ECCV*, pages 617–632, 2016. 4

[2] Harold Christopher Burger, Christian J. Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *CVPR*, pages 2392–2399, 2012. 6, 7

[3] Jiawen Chen, Sylvain Paris, and Frédo Durand. Real-time edge-aware image processing with the bilateral grid. *ACM TOG*, 26(3):103, 2007. 2

[4] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *ICCV*, pages 2516–2525, 2017. 2

[5] Antonio Criminisi, Toby Sharp, Carsten Rother, and Patrick Pérez. Geodesic image and video editing. *ACM TOG*, 29(5):134:1–134:15, 2010. 2

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE TIP*, 16(8):2080–2095, 2007. 6, 7

[7] James Diebel and Sebastian Thrun. An application of markov random fields to range sensing. In *NIPS*, pages 291–298, 2005. 4

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014. 4

[9] Jiangxin Dong, Jinshan Pan, Deqing Sun, Zhixun Su, and Ming-Hsuan Yang. Learning data terms for non-blind deblurring. In *ECCV*, pages 777–792, 2018. 6

[10] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *SIGGRAPH*, pages 257–266, 2002. 2

[11] Zeev Farbman, Raanan Fattal, Dani Lischinski, and Richard Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM TOG*, 27(3):67:1–67:10, 2008. 2

[12] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *ICCV*, pages 993–1000, 2013. 1, 2, 4

[13] Eduardo Simoes Lopes Gastal and Manuel M. Oliveira. Domain transform for edge-aware image and video processing. *ACM TOG*, 30(4):69:1–69:12, 2011. 2

[14] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 36(4):118:1–118:12, 2017. 2

[15] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *CVPR*, pages 712–721, 2017. 2

[16] Xiaojie Guo, Yu Li, and Jiayi Ma. Mutually guided image filtering. In *ACM MM*, pages 1283–1290, 2017. 1, 2, 5, 6

[17] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust guided image filtering using nonconvex potentials. *IEEE TPAMI*, 40(1):192–207, 2018. 1, 4, 5

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE TPAMI*, 35(6):1397–1409, 2013. 1, 2, 3, 4, 5, 6, 7, 8

[19] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *ECCV*, pages 353–369, 2016. 2, 4, 8

[20] Andrey Ignatov, Radu Timofte, Thang Van Vu, Tung Minh Luu, and et al. Pirm challenge on perceptual image enhancement on smartphones: Report. In *ECCV Workshops*, 2018. 6

[21] Varun Jampani, Martin Kiefel, and Peter V. Gehler. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. In *CVPR*, pages 4452–4461, 2016. 2

[22] Roy Josef Jevnisek and Shai Avidan. Co-occurrence filter. In *CVPR*, pages 3816–3824, 2017. 1

[23] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 4

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[25] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matthew Uyttendaele. Joint bilateral upsampling. *ACM TOG*, 26(3):96, 2007. 1, 4, 5, 6, 8

[26] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, pages 2657–2664, 2011. 6, 7

[27] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE TPAMI*, 30(2):228–242, 2008. 1

[28] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, pages 154–169, 2016. 2, 4, 5, 6, 8

[29] Sifei Liu, Jinshan Pan, and Ming-Hsuan Yang. Learning recursive filters for low-level vision via a hybrid neural network. In *ECCV*, pages 560–576, 2016. 2

[30] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, pages 3390–3397, 2014. 5, 6

[31] Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *ICCV*, pages 49–56, 2013. 1, 2

[32] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–425, 2001. 5, 7

[33] Dongbo Min, Jiangbo Lu, and Minh N. Do. Depth video enhancement based on weighted mode filtering. *IEEE TIP*, 21(3):1176–1190, 2012. 2

[34] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. L_0-regularized intensity and gradient prior for deblurring text images and beyond. *IEEE TPAMI*, 39(2):342–355, 2017. 6, 7

[35] Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, and Ming-Hsuan Yang. Learning dual convolutional neural networks for low-level vision. In *CVPR*, pages 3070–3079, 2018. 2

[36] Jinshan Pan, Wenqi Ren, Zhe Hu, and Ming-Hsuan Yang. Learning to deblur images with exemplars. *IEEE TPAMI*, 2018. 6

[37] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, pages 1628–1636, 2016. 1, 7

[38] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Deblurring images via dark channel prior. *IEEE TPAMI*, 40(10):2315–2328, 2018. 6

[39] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S. Brown, and In-So Kweon. High quality depth map upsampling for 3d-tof cameras. In *ICCV*, pages 1623–1630, 2011. 1, 4

[40] Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, pages 3017–3024, 2011. 1

[41] Gernot Riegler, David Ferstl, Matthias Rüther, and Horst Bischof. A deep primal-dual network for guided depth super-resolution. In *BMVC*, 2016. 2

[42] Uwe Schmidt and Stefan Roth. Shrinkage fields for effective image restoration. In *CVPR*, pages 2774–2781, 2014. 6, 7

[43] Xiaoyong Shen, Chao Zhou, Li Xu, and Jiaya Jia. Mutual-structure for joint filtering. *IJCV*, 125(1-3):19–33, 2017. 1, 2, 3, 5, 6, 7, 8

[44] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760, 2012. 4

[45] Deqing Sun, Stefan Roth, and Michael J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439, 2010. 1

[46] Jinhui Tang, Xiangbo Shu, Guo-Jun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE TPAMI*, 39(8):1662–1674, 2017. 8

[47] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998. 1, 2

[48] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *CVPR*, pages 1838–1847, 2018. 2

[49] Jiangjian Xiao, Hui Cheng, Harpreet S. Sawhney, Cen Rao, and Michael A. Isnardi. Bilateral filtering-based optical flow estimation with occlusion detection. In *ECCV*, pages 211–224, 2006. 1

[50] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, pages 157–170, 2010. 6, 7

[51] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via L_0 gradient minimization. *ACM TOG*, 30(6):174:1–174:12, 2011. 2

[52] Li Xu, Jimmy S. J. Ren, Qiong Yan, Renjie Liao, and Jiaya Jia. Deep edge-aware filters. In *ICML*, pages 1669–1678, 2015. 2

[53] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM TOG*, 31(6):139:1–139:10, 2012. 1, 2, 5, 6

[54] Qiong Yan, Xiaoyong Shen, Li Xu, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Jiaya Jia. Cross-field joint image restoration via scale map. In *ICCV*, pages 1537–1544, 2013. 1, 2

[55] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM TOG*, 35(2):11:1–11:15, 2016. 2

[56] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér. Spatial-depth super resolution for range images. In *CVPR*, 2007. 1

[57] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 2808–2817, 2017. 6, 7

[58] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. Rolling guidance filter. In *ECCV*, pages 815–830, 2014. 2, 5, 6

[59] Qi Zhang, Li Xu, and Jiaya Jia. 100+ times faster weighted median filter (WMF). In *CVPR*, pages 2830–2837, 2014. 2

[60] Shaojie Zhuo, Dong Guo, and Terence Sim. Robust flash deblurring. In *CVPR*, pages 2440–2447, 2010. 6, 7

[61] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, pages 479–486, 2011. 6, 7