This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

Learning Spatially Variant Linear Representation Models for Joint Filtering

Jiangxin Dong, Jinshan Pan, Jimmy S. Ren, Liang Lin, Jinhui Tang, Ming-Hsuan Yang

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 approaches that rely on local linear models or hand-designed objective functions to extract the structural information from the guidance image, we propose a new joint filtering method based on a spatially variant linear representation model (SVLRM), where the target image is linearly represented by the guidance image. However, learning SVLRMs for vision tasks is a highly ill-posed problem. To estimate the spatially variant linear representation coefficients, we develop an effective approach based on a deep convolutional neural network (CNN). As such, the proposed deep CNN (constrained by the SVLRM) is able to model the structural information of both the guidance and input images. We show that the proposed approach can be effectively applied to a variety of applications, including depth/RGB image upsampling and restoration, flash deblurring, natural image denoising, and scale-aware filtering. In addition, we show that the linear representation model can be extended to high-order representation models (e.g., quadratic and cubic polynomial representations). Extensive experimental results demonstrate that the proposed method performs favorably against the state-of-the-art methods that have been specifically designed for each task.

Index Terms—Spatially variant linear representation model, convolutional neural network, joint filtering.

1 Introduction

Image filters are of great importance in computer vision and related tasks, which are mainly used to suppress extraneous details while preserving primary structures. The widely used linear translation-invariant (LTI) filters usually adopt spatially invariant kernels such as mean, Gaussian, and Laplacian kernels. Since these spatially invariant kernels do not take image content into account, the LTI filters usually smooth structures, details, and noise evenly without discrepancy and thus do not preserve primary structures well [1].

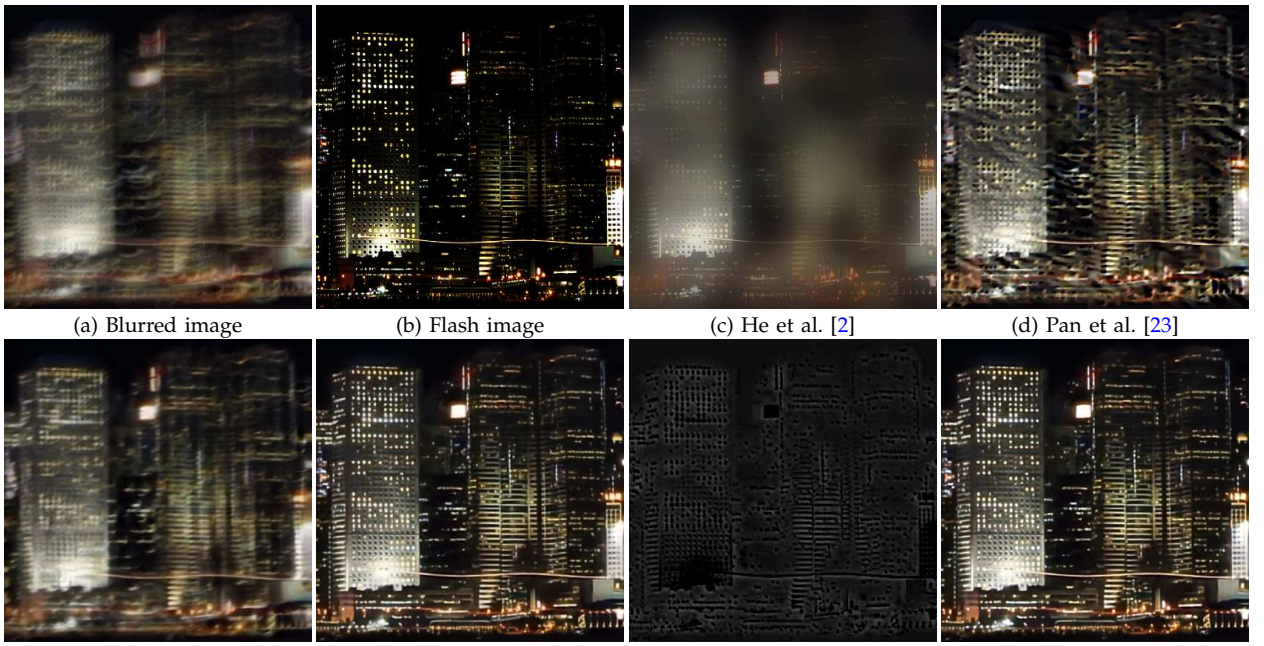
To address this problem, joint filtering methods have been proposed to utilize additional information from given guidance images. The goal of joint filtering is to transfer the important structural details of the guidance image to the output image to preserve the primary structures of the output image in the filtering process. The guidance image can be regarded as the input image itself or the image from different domains [2], [3], [4], and thus joint filtering can be used in a great variety of applications, including image editing [5], optical flow [6], [7], [8], and stereo matching [7], [9], [10]. Although joint filtering methods performs well in numerous tasks, it may introduce erroneous or extraneous artifacts to the output image when the guidance and input images are from different domains, such as RGB/depth [11], [12], [13], [3], optical flow/RGB [6], [7], [8], blurry/flash [14], [2]. Therefore, it is of great importance to explore the proprieties of guidance image and input image such that the correct structural information can be transferred in the filtering process.

One approach to exploit the common structures between the input and guidance images is to explicitly develop hand-crafted priors to model the structural co-occurrence property [9], [15], [16], [17]. However, using hand-crafted priors usually leads to complex objective functions, which are difficult to solve. The advances of deep learning motivate the development of the joint filtering algorithms based on deep convolutional neural networks (CNNs) [18], [19], [20], [21], [22]. While these algorithms perform well against conventional methods, it is less effective to use deep CNNs to explore useful structural details from guidance images for predicting target images.

In this paper, we present a joint filtering method based on a spatially variant linear representation model (SVLRM). Instead of using a deep CNN to directly predict the target image, we first learn a CNN to estimate the spatially variant linear representation that models the structural information of the guidance and input images. The target image is then generated

- *J. Dong, J. Pan and J. Tang are with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China. E-mail: dongjxjx@gmail.com, sdluran@gmail.com, jinhuitang@njust.edu.cn.*
- *J. Ren is with SenseTime Research, Hong Kong and Qing Yuan Research Institute, Shanghai Jiao Tong University. E-mail: jimmy.sj.ren@gmail.com.*
- *L. Lin is with Sun Yat-Sen University, Guangzhou, 510275, China. E-mail: linliang@ieee.org.*
- *M.-H. Yang is with University of California at Merced, Yonsei Yniversity, and Google. E-mail: mhyang@ucmerced.edu.*
- *Jinshan Pan is the corresponding author.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Blurred image (b) Flash image (c) He et al. [2] (d) Pan et al. [23]

(e) Tao et al. [24] (f) SVLRM (g) Coefficients estimated by SVLRM

Fig. 1: One synthetic example of the proposed joint filtering on image deblurring. Our method is based on a spatially variant linear representation model (SVLRM), where the target image (i.e., the deblurred image (f)) can be linearly represented by the guidance image (i.e., the short exposure image in (b)). The proposed approach estimates the linear representation coefficients (i.e., (g)) by a deep convolutional neural network which is constrained by the SVLRM. As the SVLRM is able to capture the structural details of the input and guidance image well (see (f)), our method generates better results than those based on the local linear representation model (e.g., [2]) and the state-of-the-art methods on each task (e.g., image deblurring [23], [24]).

with the estimated SVLRM. We demonstrate that the proposed method is able to transfer the important structural details of the guidance and input images to the target image and performs well in a variety of applications, including depth/RGB image upsampling and restoration, flash deblurring, natural image denoising, and scale-aware filtering. Figure 1 shows a synthetic flash deblurring example where the proposed method generates a clearer image.

The main contributions of this work are summarized as follows: First, we propose the SVLRM for joint filtering where the target image is represented by the guidance image with the SVLRM; Second, an efficient optimization algorithm is developed based on a deep CNN (which is constrained by the SVLRM) to estimate the spatially variant linear representation. We demonstrate that the estimated model coefficients capture the structural details of the guidance and input image well and can determine whether the structures should be transferred to the target image or not; Third, we show that the proposed method performs favorably on a variety of applications including depth/RGB image upsampling and restoration, flash deblurring, natural image denoising, and scale-aware filtering.

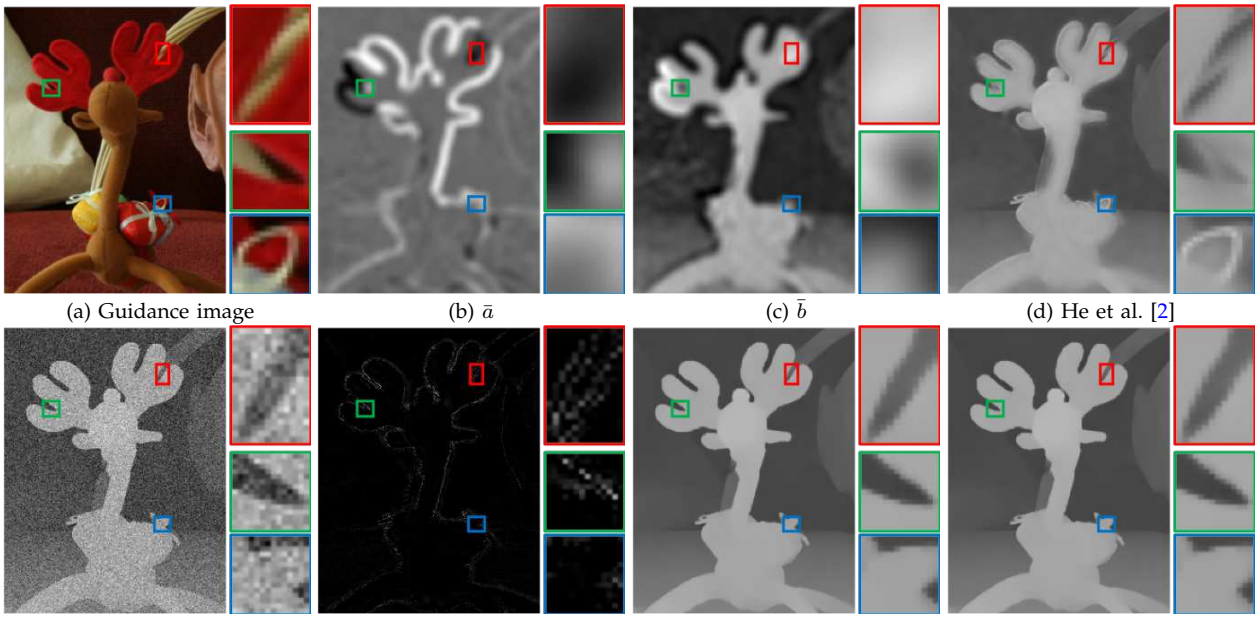
This proposed method is extended from our preliminary work [25] with the following differences. First, we analyze the effect of the network design on the SVLRM. With the proposed SVLRM, even if using a simple network, the proposed method can achieve favorable performance against state-of-the-art methods. Furthermore, we discuss about the network design and show that the proposed SVLRM is robust to the network architectures to some extent. Second, we present more detailed discussions on the SVLRM and show that it can be extended to high-order representation models (e.g., quadratic and cubic polynomial representation). Third, we exploit the property of the learned coefficients of our proposed spatially variant linear representation model and show that our approach can adaptively learn effective coefficients for different tasks but their common goal is to preserve the main structures of the target image while avoid introducing erroneous structures. Finally, we add the comparisons with more recent methods, where our approach still performs favorably against the state-of-the-art methods. We additionally add the comparisons on more real-world images and show that the proposed approach can be extended to other applications, e.g., DSLR-quality image enhancement.

2 Related Work

We review existing joint filtering approaches as local/global and deep learning-based methods.

Local joint filtering methods. The past few decades have witnessed significant advances of local joint filtering methods including the bilateral filter (BF) [4], [26], [27], guided filter (GF) [2], weighted median filter (WMF) [10], [28], geodesic distance-based filter [29], [30], weighted mode filter [31], the rolling guidance filter [32], and the mutual structure-based joint fil-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Guidance image (b) $\bar{a}$ (c) $\bar{b}$ (d) He et al. [2]

(e) Noisy input (f) $\alpha(G, I)$ (g) $\beta(G, I)$ (h) SVLRM

Fig. 2: Problems of the local linear representation model (1) in joint filtering. As shown in (b) and (c), the linear representations computed 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. In contrast, the proposed spatially variant linear representation model (5) can better capture the structural information of the guidance and input images and determine whether the structural details of the guidance image should be transferred to the target image, leading to a better target image in (h).

ter [9], etc. The main success of these methods can be attributed to the use of the local linear model or different types of affinities among neighboring pixels. For instance, the bilateral filter defines the affinity by the 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. Nevertheless, these algorithms may introduce erroneous or extraneous structures into the target image. The main reason is that only the local structures of the guidance image are explored. Although the common structure has been explored [9] to solve this problem, this method may generate halo artifacts due to the limitation of local linear models [2].

Global joint filtering methods. Several approaches optimize global objective functions with hand-crafted priors to enforce the target images to have similar structures with the guidance images, e.g., weighted least squares (WLS) filter [33], total generalized variation (TGV) [13], L_0-regularized prior [34], relative total variation (RTV) [1], scale map scheme [14], and improved RTV [16]. While these methods can better exploit global structures than local approaches, using hand-crafted priors may not well capture the inherent structural details in the target image. In addition, the objective functions with such priors are usually formulated with non-convex optimization problems that cannot be efficiently solved.

Deep learning-based methods. Numerous deep CNNs have been developed to approximate image filters [35], [36], [37], [38], [39], [40], [41]. In [21], Hui et al. learn a CNN based on a multi-scale guidance strategy to deal with depth image upsampling. To dynamically learn structural details from the guidance image for the depth image restoration, Gu et al. [20] develop a CNN based on a weighted representation model. However, these approaches are limited to specific application domains. In [22], Wu et al. propose to use a deep CNN to approximate the guided image filter. On the other hand, Li et al. [19] design 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 directly estimate the target images with regression does not always help transfer the structural details as a result of smoothing in the convolution process.

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

3 Revisiting Guided Image Filtering

We first revisit the guided image filtering [2] and then describe its role in the filtering process.

Let G, I, and F denote the guidance image, input image, and target image, respectively. The guided image filtering defines that the target image F at a pixel x is expressed by a local linear model:

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

where the linear coefficients a_k and b_k are assumed to be constant in each image patch ω_k. To determine the coefficients a_k and b_k, He et al. [2] introduce a constraint of a_k (i.e., a_k^2) and minimize the following objective function:

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

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

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

Although the filtering algorithm based on the local linear model (1) is effective in numerous tasks, it may introduce extraneous textures in the target image (see the parts enclosed in the red and blue boxes of Figure 2(d)) due to the assumption of constant a_k and b_k in each image patch. It is noted that the gradient of the target image and guidance image in each image patch satisfies

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

according to (1). As a_k is a constant, the constraint (4) ensures that the target image has the structures similar to the guidance image. Thus, 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, the mean filter is further applied to compute the pixel-wise representation coefficients $\bar{a}$ and $\bar{b}$, which likely suppress the high-frequency information that corresponds to the important structural details in the guidance image. The regions enclosed in the green boxes in Figure 2(b) and (c) show that the representation coefficients $\bar{a}$ and $\bar{b}$ are over-smoothed. Therefore, using such representations accordingly does not transfer structures to the target image well (e.g., the edges enclosed in the green box in Figure 2(d) are not sharp). Considering that the target image F is mainly determined by the representation coefficients $\bar{a}$ and $\bar{b}$, effective representations should be able to determine whether the structures of the guidance image G should be transferred to the target image or not in order to model the structural details of both the guidance and input images.

To address this problem, we propose an SVLRM where the linear representation coefficients are estimated by a deep CNN. Figure 2(f) and (g) shows that the estimated spatially variant linear representation describes the structural information of the guidance and input images well, which accordingly leads to a better target image.

4 Proposed Method

In this section, we present the SVLRM and estimate the coefficients with a deep CNN for joint filtering.

4.1 Spatially variant linear representation model

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

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

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

4.2 Optimization

A common approach to estimate $\alpha(G, I)$ and $\beta(G, I)$ in (5) is to adopt the regularization w.r.t. $\alpha(G, I)$ and $\beta(G, I)$ to minimize the following objective function:

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

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, minimizing (6) can be solved by the gradient descent:

$$\begin{aligned} \alpha^t &= \alpha^{t-1} - \lambda \left(\frac{\partial \mathcal{E}(\alpha, \beta^{t-1})}{\partial \alpha} \right)_{\alpha=\alpha^{t-1}}, \\ \beta^t &= \beta^{t-1} - \lambda \left(\frac{\partial \mathcal{E}(\alpha^{t-1}, \beta)}{\partial \beta} \right)_{\beta=\beta^{t-1}}, \end{aligned} \quad (7)$$

where λ and t indicate the step size and iteration number, respectively.

Another issue concerning (6) is that it is not straightforward to determine $\varphi(\alpha)$ and $\phi(\beta)$ for joint filtering, as the properties of $\alpha(G, I)$ and $\beta(G, I)$ are different from the statistical properties of natural images [2], [9]. Instead of using hand-crafted priors 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).

Learning. Noting that (7) is in spirit similar to the stochastic gradient descent that is widely used to solve deep CNNs, we propose a deep CNN to estimate $\alpha(G, I)$ and $\beta(G, I)$.

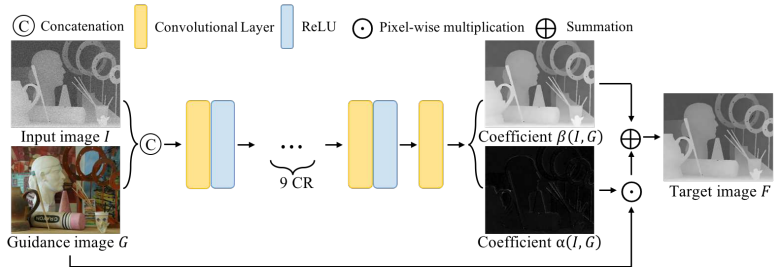
Let $\{G^n, I^n, F_{gt}^n\}_{n=1}^N$ denote a set of N training samples and $\mathcal{F}$ denote the deep CNN. Our goal is to learn the network parameters $\Theta = \{\Theta_\alpha, \Theta_\beta\}$ so that $\mathcal{F}_{\Theta_\alpha}$ and $\mathcal{F}_{\Theta_\beta}$ can approximate the spatially variant linear representation 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 outputs of the network $\mathcal{F}$ w.r.t. the parameters Θ_α and Θ_β.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Network architecture

Layers	Filter size	Stride	Padding
CR1	3×3×2×64	1	1
CR2	3×3×64×64	1	1
CR3	3×3×64×64	1	1
CR4	3×3×64×64	1	1
CR5	3×3×64×64	1	1
CR6	3×3×64×64	1	1
CR7	3×3×64×64	1	1
CR8	3×3×64×64	1	1
CR9	3×3×64×64	1	1
CR10	3×3×64×64	1	1
CR11	3×3×64×64	1	1
C12	3×3×64×2	1	1

(b) Network parameters

Fig. 3: Architecture and parameters of the proposed network based on the SVLRM. "CR" denotes the convolutional layer followed by a non-linear LeakyReLU function and "C" denotes the convolutional layer.

During training, the network $\mathcal{F}$ is constrained by the L_1 loss function defined by

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

Owing that 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

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

Based on (10), the network parameters can be updated by

$$\begin{aligned} \Theta_\alpha^t &= \Theta_\alpha^{t-1} - \lambda \frac{\partial \mathcal{F}_{\Theta_\alpha}}{\partial \Theta_\alpha} \frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\alpha}}, \\ \Theta_\beta^t &= \Theta_\beta^{t-1} - \lambda \frac{\partial \mathcal{F}_{\Theta_\beta}}{\partial \Theta_\beta} \frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\beta}}. \end{aligned} \quad (11)$$

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

Network architecture. Based on above considerations, we can use existing network architectures to define the network $\mathcal{F}$. In this work, the network $\mathcal{F}$ is realized by 12 convolution layers, each of which is followed by the LeakyReLU except for the last convolution layer. We set the feature number of the first 11 convolution layers as 64. The filter size and the stride value are set to be 3×3 pixels and 1, respectively. Figure 3 shows the network architecture and parameters.

The above plain network is able to generate favorable results as demonstrated in our preliminary work [25]. We further analyze the effect of simply stacking more layers in such plain networks and the effect of using more advanced networks on the proposed approach in Section 6.5.

5 Experimental Results

We evaluate the proposed method on a variety of applications including depth image upsampling, depth image restoration, scale-aware filtering, natural image denoising, flash image deblurring, and natural image enhancement.

5.1 Parameter settings

During the training process, we introduce the momentum when updating (11) and use the ADAM optimizer [47] with default parameter values. The batch size is set to be 20. The step size λ (i.e., learning rate) is initialized as 10^{-4}, which is updated by a polynomial decay schedule. The parameter ε is set to be 10^{-6}. The source code will be made available to the public. In the following, we retrain or finetune the deep CNN-based methods using the same training datasets as the proposed method for fair comparisons.

5.2 Depth image upsampling

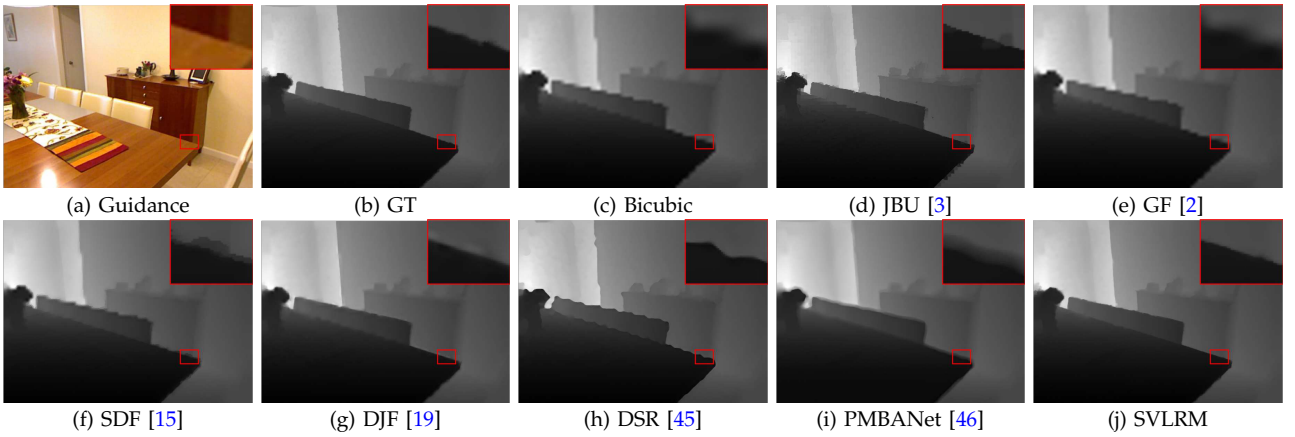
Training data. For depth image upsampling, we use the NYU depth dataset [42] and randomly choose 1000 RGB/D image pairs to generate the training data, following the protocols of [19]. The remaining 449 RGB/D image pairs [42] are used as the test dataset to evaluate the proposed approach.

We quantitatively and qualitatively compare the proposed method against state-of-the-art methods, including MRF [43], GF [2], JBU [3], TGV [13], 3D-TOF [12], SDF [15], FBS [44], DMSG [21], DJF [19], DSR [45], and PMBANet [46]. The quantitative results in Table 1 show that the proposed approach performs favorably against state-of-the-art methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: Quantitative evaluations for the depth image upsampling problem on the synthetic benchmark dataset [42] in terms of RMSE.

Methods	Bicubic	MRF [43]	GF [2]	JBU [3]	TGV [13]	3D-TOF [12]	SDF [15]	FBS [44]	DMSG [21]	DJF [19]	DSR [45]	PMBANet [46]	SVLRM
×4	4.52	3.54	4.28	3.15	2.69	3.91	4.49	2.89	1.94	2.06	3.59	2.75	**1.49**
×8	7.89	5.66	7.75	5.03	4.49	5.56	7.78	5.44	3.42	3.50	5.42	4.24	**2.93**
×16	12.62	9.04	12.55	8.37	7.35	7.10	12.55	8.79	5.82	5.91	7.74	5.77	**5.69**



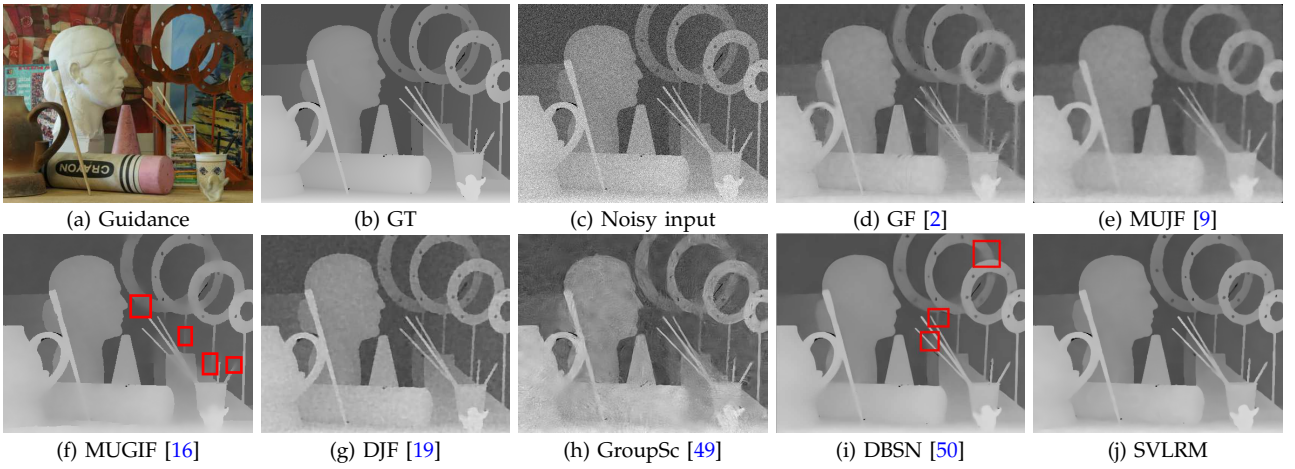
(a) Guidance (b) GT (c) Bicubic (d) JBU [3] (e) GF [2]

(f) SDF [15] (g) DJF [19] (h) DSR [45] (i) PMBANet [46] (j) SVLRM

Fig. 4: On the depth image upsampling application (×8). The proposed method generates the depth images with sharper boundaries.

TABLE 2: Quantitative evaluations for the depth image restoration problem on the benchmark dataset [48] in terms of PSNR, SSIM, and RMSE.

Methods	Input	GF [2]	JBU [3]	MUJF [9]	MUGIF [16]	DJF [19]	GroupSC [49]	DBSN [50]	SVLRM
Avg. PSNRs	27.23	30.79	28.86	30.67	34.03	34.02	36.01	36.84	**37.53**
Avg. SSIMs	0.6350	0.9214	0.9081	0.9282	0.9565	0.9567	0.9294	0.9636	**0.9696**
Avg. RMSEs	13.19	7.75	9.82	7.76	5.29	5.35	4.10	3.80	**3.50**



(a) Guidance (b) GT (c) Noisy input (d) GF [2] (e) MUJF [9]

(f) MUGIF [16] (g) DJF [19] (h) GroupSc [49] (i) DBSN [50] (j) SVLRM

Fig. 5: On the depth image restoration application (8% noise level). The parts enclosed in the red boxes in (f) are over-smoothed. The proposed method generates the depth images with sharper boundaries.

Figure 4 shows one example from the test dataset. As analyzed in Section 3, the GF method [2] is likely to transfer the extraneous structures from the guidance images, which accordingly negatively affects the structures of the super-resolution result as shown in Figure 4(e). To learn the dynamic guidance features for joint image upsampling, the DJF method [19] first concatenates the features of the guidance image and input image, and then uses a CNN [58] to estimate the target image in a regression way. However, as shown in [59], the method [58] is less effective for the structural details restoration. Thus, the edges restored by the DJF method [19] are not well estimated as shown in Figure 4(g). The most recent method [46] proposes a progressive multi-branch aggregation network to progressively recover the degraded depth image. While as shown in Figure 4(i), this method is less effective to preserve sharp edges. In contrast, the proposed approach uses the SVLRM for the joint image filtering and further develops a deep CNN to estimate the representation coefficients. Under the guidance of the estimated coefficients, the SVLRM is able to better transfer the structural details of the guidance image and input image to the target image. Thus, the sharp edges of the super-resolved depth image are preserved well (Figure 4(j)), and the generated results have lower RMSE values (Table 1).

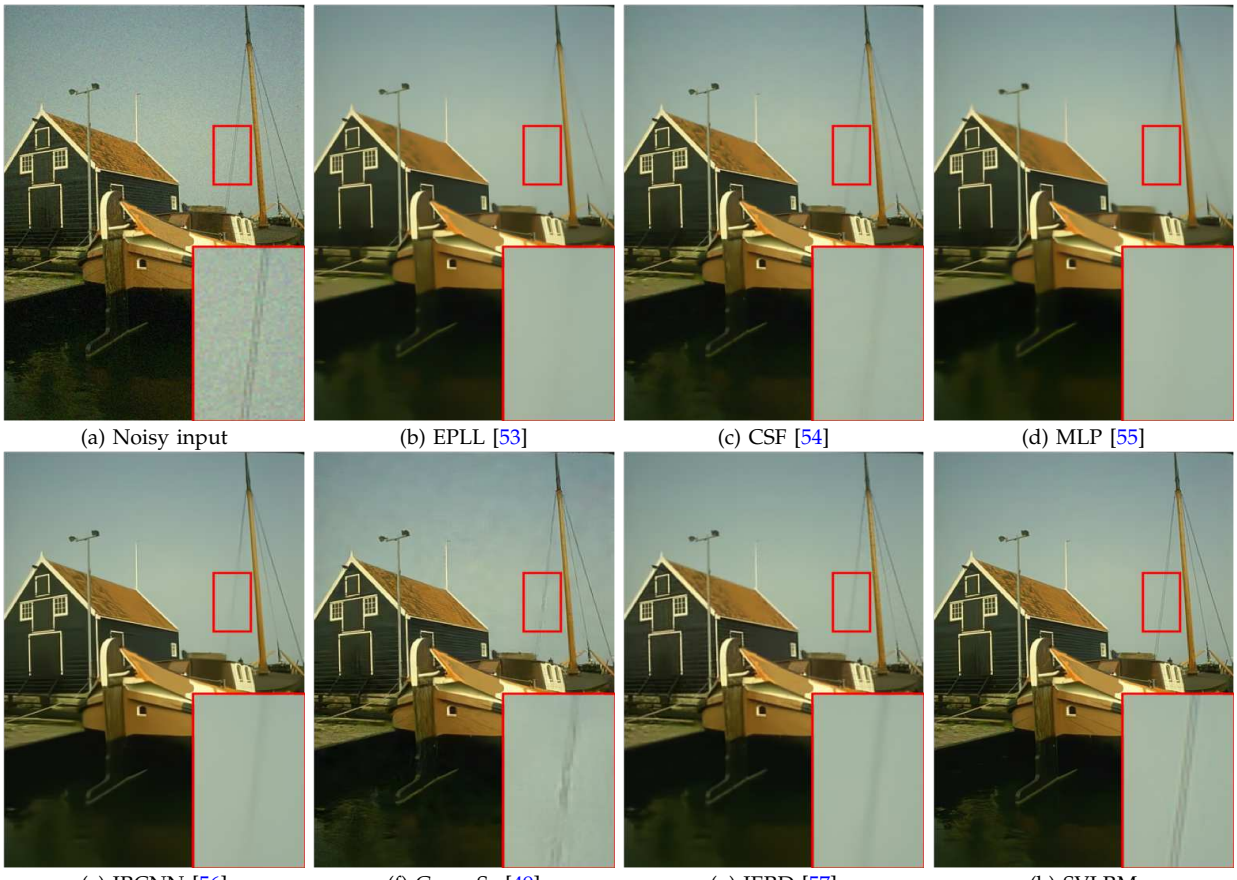
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3: Quantitative evaluations for the image denoising problem on the BSDS dataset [51] in terms of PSNR, SSIM, and RMSE.

Methods	Input	BM3D [52]	GF [2]	EPLL [53]	CSF [54]	MLP [55]	IRCNN [56]	GroupSc [49]	IERD [57]	SVLRM
Avg. PSNRs	27.23	31.60	20.35	29.34	30.10	28.91	31.86	29.73	29.17	**33.08**
Avg. SSIMs	0.6350	0.8765	0.6173	0.8000	0.8164	0.7854	0.8811	0.8163	0.8238	**0.8960**
Avg. RMSEs	13.19	6.90	24.85	8.96	8.43	9.39	6.64	8.58	9.47	**6.29**

(a) Input image (b) DJF [19] (c) RTV [1] (d) RGF [32] (e) SVLRM

Fig. 6: On the scale-aware filtering application. The comparisons in (b-d) are obtained from the reported results. The proposed method is able to remove the small-scale textures while preserving the main sharp structures.



(a) Noisy input (b) EPLL [53] (c) CSF [54] (d) MLP [55]

(e) IRCNN [56] (f) GroupSc [49] (g) IERD [57] (h) SVLRM

Fig. 7: On the image denoising application. The proposed method generates the images with clearer structures.

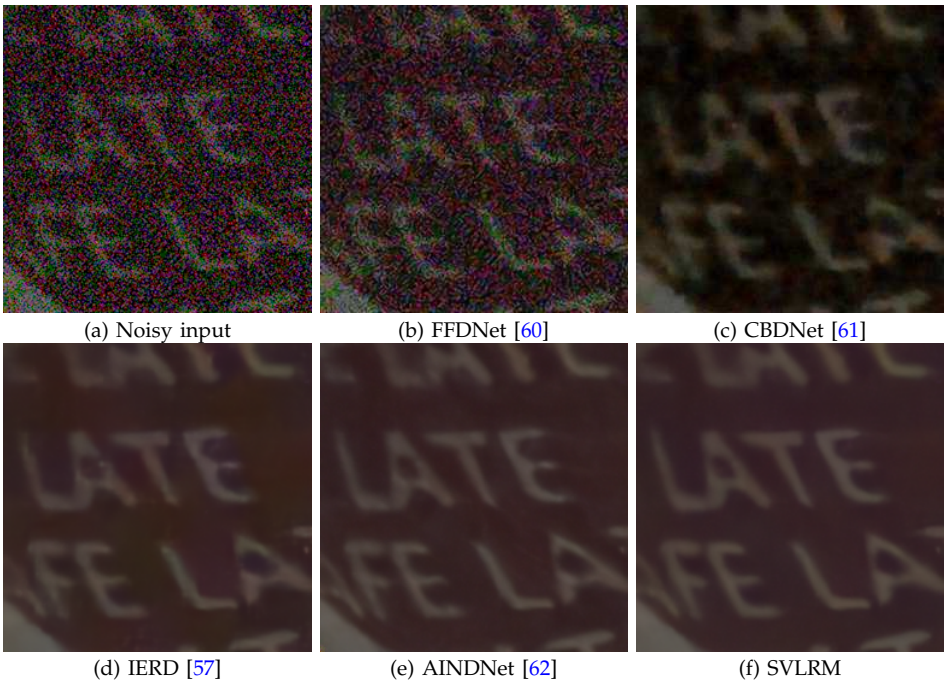
5.3 Depth image restoration

Training data. For depth image restoration, we select the same training dataset as used in Section 5.2. We further add the Gaussian noise to each ground truth depth image, where the noise level ranges from 0 to 10%. The test dataset by [48] is used to evaluate the proposed method to avoid any overlap between the training and test datasets. For each test image, we add the Gaussian noise with a noise level of 8%.

Table 2 shows the quantitative evaluation results where the proposed approach performs favorably against state-of-the-art methods.

Figure 5 shows the depth image denoising results by the evaluated methods. The GF method [2] does not effectively restore the structural details as shown in Figure 5(d). The MUJF method [9] uses the mutual structures of the guidance and input images to avoid introducing extraneous details to the target images. Note that this method still adopts the local linear assumption [2] and uses the mean filter to obtain the final pixel-wise linear representation coefficients. Figure 5(e) shows that the edges restored by [9] are not preserved well due to the less accurate linear

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Noisy input (b) FFDNet [60] (c) CBDNet [61]

(d) IERD [57] (e) AINDNet [62] (f) SVLRM

Fig. 8: One real image denoising example from the SIDD dataset [63]. The proposed approach generates the images with clearer characters.

TABLE 4: Quantitative evaluations for the natural image denoising problem on the SIDD validation dataset [63] in terms of PSNR. The results of [64], [62], [65], [66] are obtained from their reported results and those of [52], [67], [60], [61], [68], [57] are obtained from [57].

Methods	BM3D [52]	DnCNN [67]	FFDNet [60]	CBDNet [61]	RIDNet [68]	IERD [57]	GradNet [64]	AINDNet [62]	DANet [65]	SADNet [66]	SVLRM
Avg. PSNRs	30.88	26.21	29.20	30.78	38.71	38.82	38.34	39.01	39.30	39.46	**39.48**

representation coefficients. By exploring the mutual structures, the MUGIF method [16] generates better results than [9]. However, some edges estimated by [16] are over-smoothed as shown in the red boxes of Figure 5(f). Although the DJF method [19] is able to preserve sharp edges, the obtained result contains significant artifacts (Figure 5(g)). The method [50] presents a two-stage scheme for image restoration by incorporating self-supervised learning and knowledge distillation. However, some structures recovered by [50] are over-smoothed as shown in Figure 5(i). Instead of directly estimating the target image by a deep CNN, the proposed approach predicts the target image by the SVLRM, where the representation coefficients are estimated by a deep CNN. Figure 5(j) shows the proposed method generates a clearer image with sharper edges.

5.4 Scale-aware filtering

With the models trained for the depth image denoising, the proposed approach can be straightforwardly applied to scale-aware filtering. Similar to the DJF approach [19], we use the input image as the guidance and adopt the rolling approach [32] to remove small-scale structures and details.

We show one example from [1] in Figure 6. The scale-aware filtering aims to suppress fine-scale details while extracting meaningful structures from textured surfaces. However, the filtered results by the DJF [19] and RGF [32] methods still contain small-scale structures in the backgrounds. In contrast, the proposed approach is able to remove the small-scale structures from the input images and generates competitive results compared to [1].

5.5 Natural image denoising

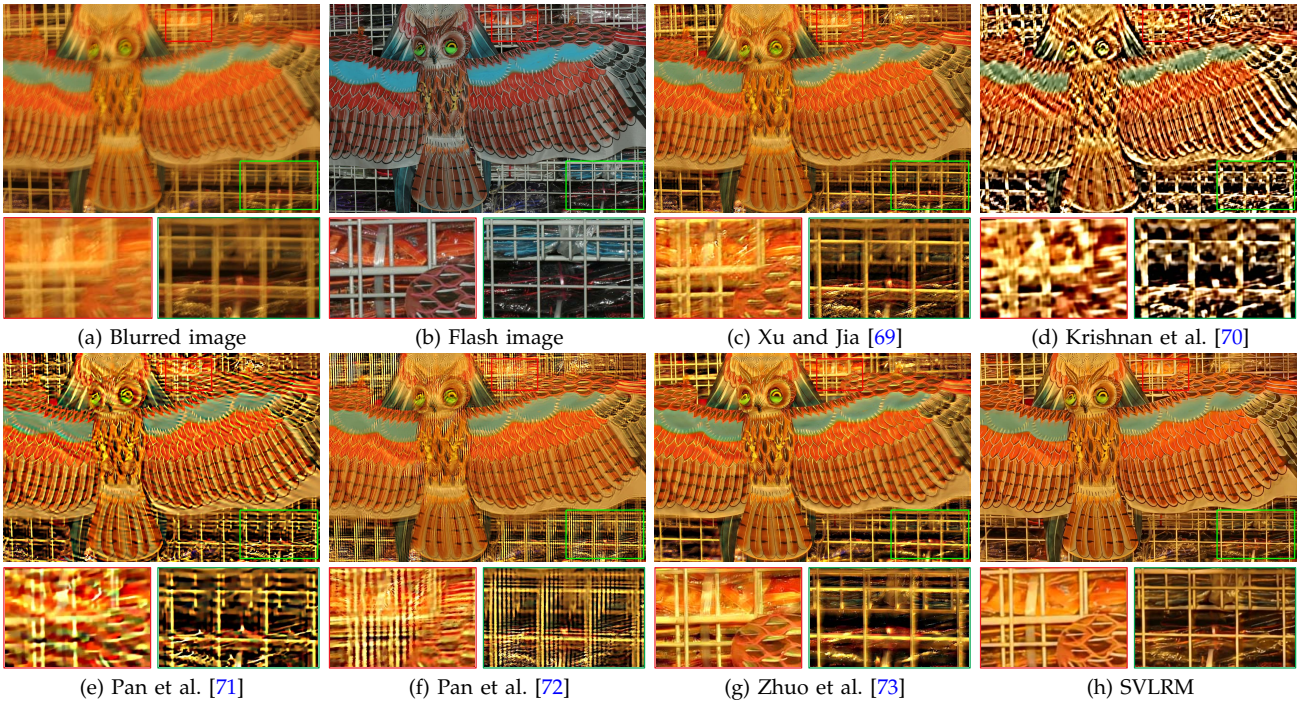
As the guidance image can be the input image itself, the proposed approach can be applied to natural image denoising.

Training data. For natural image denoising, we use the training dataset from the BSDS500 dataset [51] and randomly add the Gaussian noise to each clear image, where the noise level ranges from 0 to 10%. We then evaluate the proposed method on the test dataset by [51]. The Gaussian noise with a random noise level from 0 to 10% is added to each test image.

We quantitatively and qualitatively evaluate the proposed approach against state-of-the-art methods, including BM3D [52], EPLL [53], CSF [54], MLP [55], IRCNN [56], GroupSc [49], and IERD [57]. Table 3 quantitatively demonstrates that the proposed approach is able to generate high-quality images.

Figure 7 shows one example from the test dataset, where the structures restored by state-of-the-art methods are over-smoothed. In contrast, the main structures restored by the proposed method are preserved well as shown in Figure 7(h), as the estimated spatially variant linear representation coefficients can effectively transfer the structural details of the guidance and input images to the target image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Blurred image (b) Flash image (c) Xu and Jia [69] (d) Krishnan et al. [70]

(e) Pan et al. [71] (f) Pan et al. [72] (g) Zhuo et al. [73] (h) SVLRM

Fig. 9: One real example from [73] on image deblurring. The proposed approach is able to estimate effective representation coefficients. Thus the deblurred image contains clearer structures and textures.

(a) Blurred image (b) Flash image (c) Zhuo et al. [73] (d) SVLRM

Fig. 10: One real example from [73] on flash image deblurring. The proposed approach generates a much clearer result.

We evaluate the proposed approach on the real-world image dataset. To do this, we train the proposed model on the training dataset from the NTIRE 2020 contest (i.e., Smartphone Image Denoising Dataset [63]). To validate the performance of our approach, we first quantitatively compare against the state-of-the-art methods on the validation dataset of SIDD. Table 4 shows that our approach performs comparably with the state-of-the-art methods. In addition, we qualitatively compare with the competed methods on one example from the SIDD validation dataset in Figure 8, where the proposed method recovers much clearer characters. All the comparisons demonstrate that our approach is able to handle real-world noise.

5.6 Flash deblurring

In [73], Zhuo et al. propose to deblur a blurry image under the guidance of its flash image. We show that the proposed method can be applied to this task.

Training data. To generate the training data for flash image deblurring, we use the image enhancement dataset by [75] as the flash and no-flash image pairs. We generate the blur kernels by [76], [77] and apply them to the no-flash images to generate blurred images. Finally, a set of 100,000 blurred images is constructed to train the proposed model.

We evaluate the competed methods using one real example from [73] in Figure 9. As the blurred image contains significant blur, single image deblurring methods [69], [70], [71], [72] do not recover clear images. Figure 9(c)-(f) shows that the generated results still contain significant blur and artifacts. Under the guidance of the flash image, the method by Zhuo et al. [73] is able to help the deblurring problem. However, some structural details in Figure 9(g) are not estimated well because only the sparsity of gradient prior is considered in image restoration. In contrast, the proposed approach is able to remove the blur and generates a clearer image with fine details (Figure 9(h)). Figure 10 shows another real example from [73], where our approach recovers much clearer images. The comparisons in Figures 9-10 demonstrate the effectiveness and robustness of our approach on the real application of flash image deblurring.

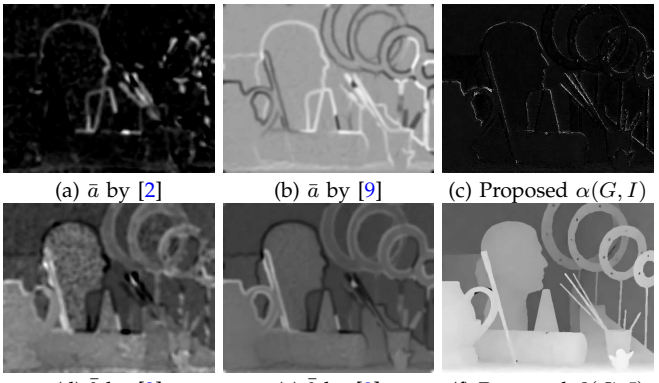
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

(a) Input image (b) Ignatov et al. [74] (c) SVLRM

Fig. 11: Comparisons of the DSLR-quality image enhancement results. The enhanced result by the proposed method is comparable to that by Ignatov et al. [74] (Best viewed on high-resolution display with zoom-in).

5.7 DSLR-quality Image Enhancement

Ignatov et al. [74], [75] propose an effective DSLR-quality image enhancement method for the images captured by mobile phones such that the quality is similar to that of DSLR. We show that the proposed method can also be applied to this problem and achieves comparable performance. We use the same training and test datasets as [74] to train and evaluate the proposed model. Figure 11 shows an real example from [74], where the proposed method generates comparable results.



(a) $\bar{a}$ by [2] (b) $\bar{a}$ by [9] (c) Proposed $\alpha(G, I)$

(d) $\bar{b}$ by [2] (e) $\bar{b}$ by [9] (f) Proposed $\beta(G, I)$

Fig. 12: The effect of the proposed SVLRM. The guidance image and input image are shown in Figure 5(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).

6 Analysis and Discussion

In this section, we analyze the proposed approach based on the SVLRM with comparisons to the most related methods. We explain why the proposed method is effective for joint filtering and discuss the effect of high-order representation models.

Instead of using the local linear model, we propose the SVLRM and directly estimate the pixel-wise representation coefficients by a deep CNN. The estimated linear representation coefficients in Figure 12(c) and (f) are able to better capture the structural details of the guidance and input images, thereby facilitating depth image restoration (Figure 5(h)).

6.1 Comparisons with local linear models

Numerous methods have been developed to improve the linear model (1) of the GF algorithm [2]. Shen et al. [9] estimate the local linear representation coefficients a_k and b_k by exploiting the mutual structures of the guidance and input images. While this method is able to alleviate the text-copy effect, it still uses the mean filter to compute the final pixel-wise coefficients, which may smooth important structural details (Figure 12(b) and (e)) and do not restore the sharp edges well (Figure 5(e)).

On the other hand, Wu et al. [22] propose an effective convolutional guided filtering layer based on [2] to solve image processing problems. Zhu et al. [78] embed the local linear regularization constraint (1) into deep CNNs for depth image restoration. These approaches perform well on depth image restoration against the GF method [2]. We note that the method [22] is inherently based on the local linear model (1). It derives a differentiable convolutional layer based on the solutions of (2) and embeds it into an end-to-end trainable network. The method [78] uses a deep CNN consisting of residual learning, where the network architecture is similar to DJF [19]. These two modules are first applied to guidance images and input images to extract features respectively and then the remaining deep module is used to restore images. Although using different deep modules increases the model capacity, it makes the training process more difficult. In contrast to these methods, our SVLRM has a smaller model size but better performance. Table 5 shows that our method generates better results.

6.2 Property of learned coefficients

To exploit the nature of the learned coefficients, we show the coefficients $\alpha(G, I)$ and $\beta(G, I)$ learned for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5: Comparisons with local linear model-based methods on depth image denoising using the test dataset [48]. For each test image, we add Gaussian noise with a noise level of 3%.

	GF [2]	Zhu et al. [78]	SVLRM
Avg. PSNRs	32.11	33.54	**42.77**
Avg. SSIMs	0.9437	0.9626	**0.9855**
Model size (#M)	–	0.93	**0.37**

TABLE 6: Comparisons with end-to-end methods on image denoising. The depth image denoising is evaluated on the test dataset by [42]. The natural image denoising is evaluated on the test dataset by [51]. For both tasks, we add Gaussian noise with a noise level of 8% to each test image.

	Depth image denoising		Natural image denoising	
	E2ETN	SVLRM	E2ETN	SVLRM
Avg. PSNRs	39.37	**39.69**	28.61	**29.20**
Avg. SSIMs	0.9633	**0.9644**	0.7810	**0.8072**

different tasks (i.e., depth image upsampling, depth image denoising, and natural image denoising) in Figure 13. For the task of depth image upsampling (i.e., the first column of Figure 13), given the bicubic upsampled image (Figure 13(b1)) as the input image I, we aim to estimate the high-quality high-resolution depth image (Figure 13(f1)) under the guidance image G in Figure 13(a1). The learned coefficients $\alpha(G, I)$ and $\beta(G, I)$ are shown in Figure 13(c1) and (d1), where the coefficient $\alpha(G, I)$ contains the common main structures of the guidance and input images. In contrast, the depth image denoising and natural image denoising (i.e., the second to fourth columns of Figure 13) aim to restore the clear images (Figure 13(f2)-(f4)) from the noisy inputs I (Figure 13(b2)-(b4)), using themselves as the guidance images G. Figure 13(c2)-(c4) and (d2)-(d4) shows the learned coefficients $\alpha(G, I)$ and $\beta(G, I)$, which contain the main structure of input image and the image noise, respectively. Therefore, the proposed spatially variant linear representation model can adaptively learn effective coefficients for different tasks. However, their common goal is to effectively determine what structures of the guidance image should be transferred to the output image.

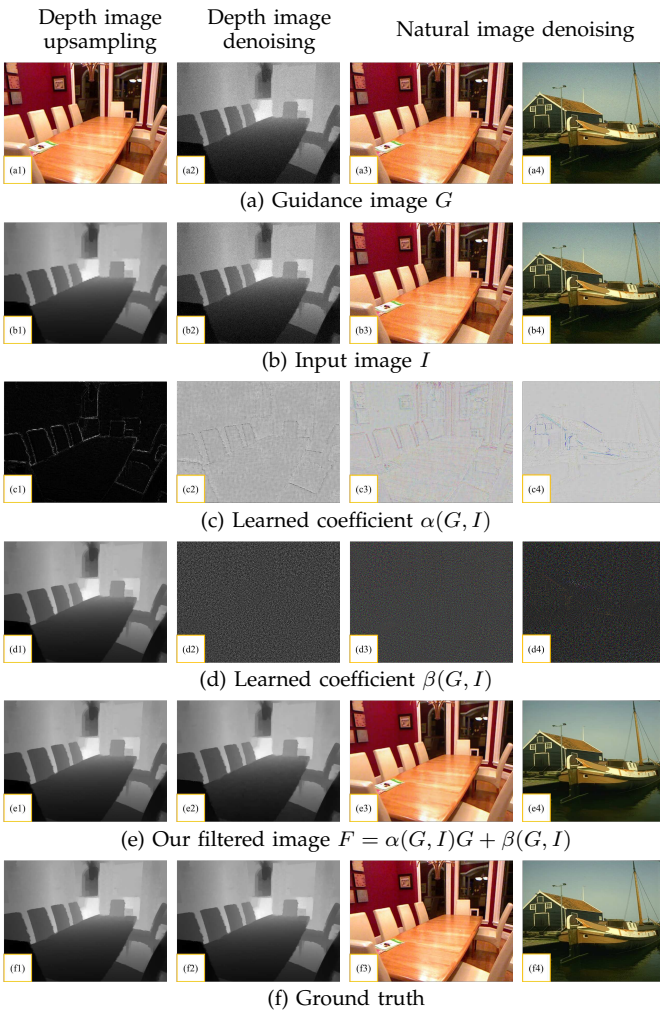


Fig. 13: Learned coefficients for various applications. Each column from left to right corresponds to the task of depth image upsampling, depth image denoising, and natural image denoising (the third column to the fourth column). The values of the learned coefficients in (c) and (d) are rescaled to the ranges of the input image for better visualization.

6.3 Comparisons with end-to-end networks

Instead of using an end-to-end trainable network to directly estimate the target image, we propose the SVLRM, which is able to effectively capture the structural information of both the guidance and input images. To demonstrate the effectiveness of the proposed SVLRM, we compare it with the end-to-end trainable networks. For fair comparisons, we remove the step for learning representations in our implementation and directly estimate the target image, and referred it as E2ETN. As the proposed method estimates the linear representation coefficients instead of the target images, they can effectively transfer the structures of the guidance and input images to the target image as analyzed in Section 3. Therefore, the proposed SVLRM is able to generate sharper edges under the guidance of the estimated representation coefficients (see Figure 14(d)). In addition, the quantitative evaluations in Table 6 show that the proposed method consistently improves the performance. All these results concretely demonstrate the effectiveness of the proposed representation coefficient learning method.

6.4 Comparisons with hand-crafted priors

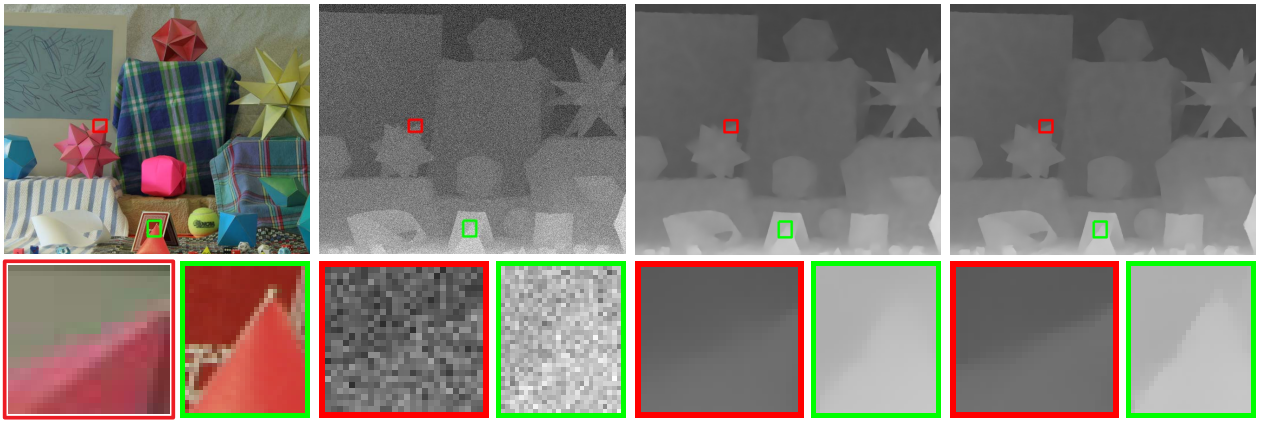
One alternative approach to estimate $\alpha(G, I)$ and $\beta(G, I)$ is to use hand-crafted priors in (6). Similar to [2], we take $\varphi(\alpha)$ and $\phi(\beta)$ as $\mu\alpha^2$ and $\eta\beta^2$ and the objective function (6) becomes

$$\mathcal{E}(\alpha, \beta) = \|\alpha G + \beta - I\|^2 + \mu\alpha^2 + \eta\beta^2, \tag{12}$$

where μ and η are positive weight parameters. The gradients of $\mathcal{E}(\alpha, \beta)$ with respect to α and β are

$$\begin{aligned} \frac{\partial \mathcal{E}(\alpha, \beta)}{\partial \alpha} &= 2G(\alpha G + \beta - I) + 2\mu\alpha, \\ \frac{\partial \mathcal{E}(\alpha, \beta)}{\partial \beta} &= 2(\alpha G + \beta - I) + 2\eta\beta. \end{aligned} \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Guidance (b) Noisy input (c) E2ETN (d) Ours

Fig. 14: Comparisons with end-to-end networks. The proposed approach is able to estimate effective representation coefficients. Thus the restored image contains sharper edges.

TABLE 7: Comparisons with hand-crafted priors on image denoising.

	Depth image denoising			Natural image denoising		
	HC-ℓ_2	HC-grad-ℓ_1	SVLRM	HC-ℓ_2	HC-grad-ℓ_1	SVLRM
Avg. PSNRs	21.84	29.98	**37.53**	24.87	27.48	**33.08**
Avg. SSIMs	0.2044	0.7102	**0.9696**	0.6452	0.7813	**0.8960**

By setting $\frac{\partial \mathcal{E}(\alpha,\beta)}{\partial \alpha} = 0$ and $\frac{\partial \mathcal{E}(\alpha,\beta)}{\partial \beta} = 0$, we can obtain

$$\alpha = \frac{GI - G\beta}{G^2 + \mu}, \qquad \beta = \frac{I - \alpha G}{1 + \eta}. \tag{14}$$

Thus, the solutions of (12) are

$$\alpha = \frac{\eta GI}{\eta G^2 + \mu + \mu\eta}, \qquad \beta = \frac{I - \alpha G}{1 + \eta}. \tag{15}$$

The parameters μ and η are empirically set to be 0.1 on the depth image denoising and natural image denoising problems. Table 6 shows that the method with the hand-crafted priors in (12) (HC-ℓ_2 for short) is less effective compared to the methods with deep CNNs. Furthermore, as shown in Figure 15(b) and (c), the coefficients computed by HC-ℓ_2 contain significant noise, which accordingly leads to noisy results (Figure 15(d)). In contrast, the proposed approach generates a much clearer image (Figure 15(l)).

To determine whether the unsatisfactory artifacts result from the use of $\mu\alpha^2$ and $\eta\beta^2$ as these constraints are less effective to image noise, we use the sparsity of the gradient (i.e., $\|\nabla\alpha\|_1$ and $\|\nabla\beta\|_1$) as the constraint in (6) because this constraint is more robust to image noise and is able to preserve the main structures of the images. Thus, the objective function becomes

$$\mathcal{E}(\alpha,\beta) = \|\alpha G + \beta - I\|^2 + \mu\|\nabla\alpha\|_1 + \eta\|\nabla\beta\|_1. \tag{16}$$

We use the gradient descent method (7) to solve (16). The gradients of $\mathcal{E}(\alpha,\beta)$ with respect to α and β are

$$\begin{aligned}
\frac{\partial \mathcal{E}(\alpha,\beta)}{\partial \alpha} &= 2G(\alpha G + \beta - I) + \mu\left(\frac{\partial_h^\top \partial_h \alpha}{\|\nabla\alpha\|_1} + \frac{\partial_v^\top \partial_v \alpha}{\|\nabla\alpha\|_1}\right), \\
\frac{\partial \mathcal{E}(\alpha,\beta)}{\partial \alpha} &= 2(\alpha G + \beta - I) + \eta\left(\frac{\partial_h^\top \partial_h \beta}{\|\nabla\beta\|_1} + \frac{\partial_v^\top \partial_v \beta}{\|\nabla\beta\|_1}\right).
\end{aligned} \tag{17}$$

We empirically set $t = 200$, and $\mu = \eta = 0.05$ for fair comparisons. We quantitatively compare the proposed approach with the baseline methods based on hand-crafted priors on the tasks of depth image denoising and natural image denoising using the same test dataset as Section 5.3 and Section 5.5. Table 7 demonstrates the effectiveness of leaning a deep neural network to estimate the representation coefficients.

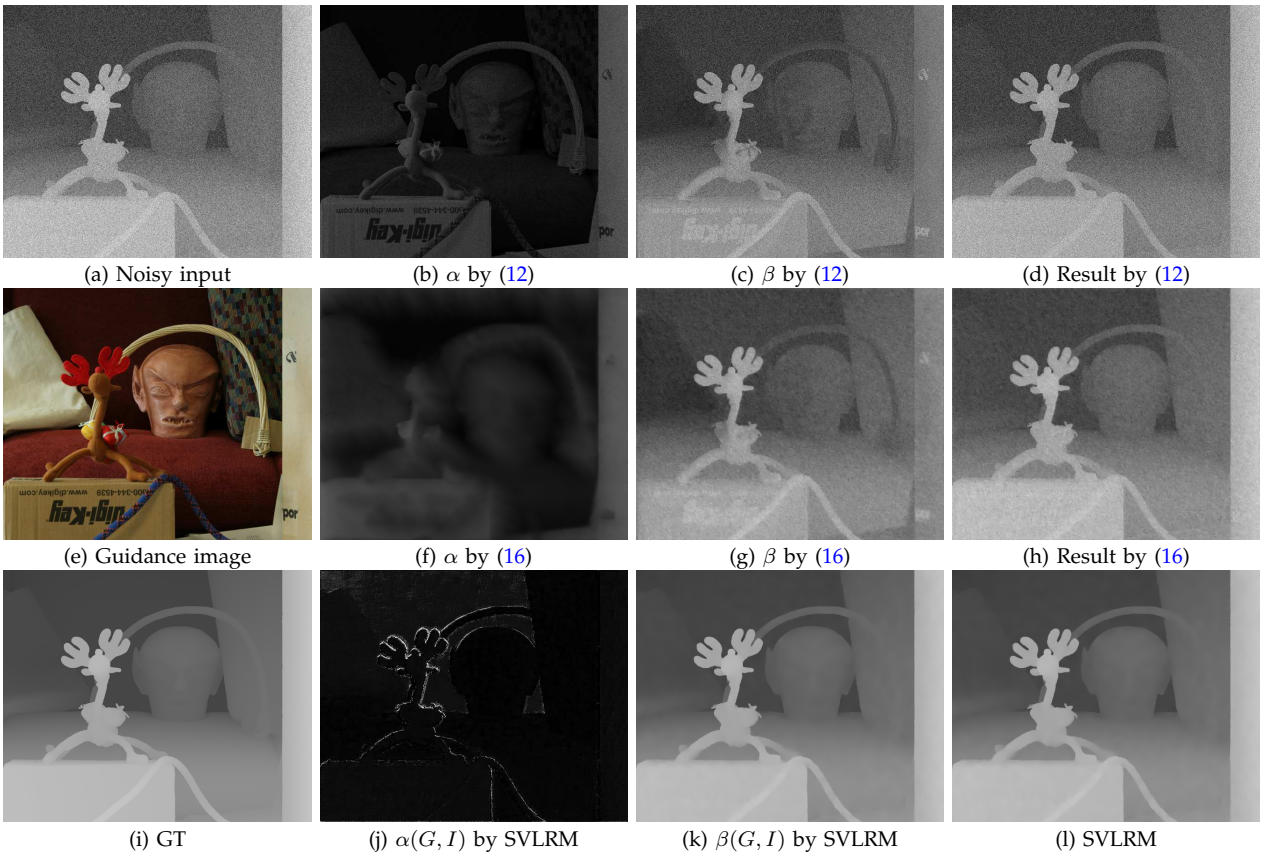
Figure 15 shows the visual comparisons of the proposed method and others with different hand-crafted priors. Although the approach using gradient sparsity as the constraint for the coefficients performs better than those using the priors α^2 and β^2 [2], the generated results still contain significant noise in Figure 15(h). The proposed deep CNN is constrained by the SVLRM and able to effectively estimate the linear representation coefficients (Figure 15(j) and (k)). Thus, the proposed approach is able to remove noise and generate a better restoration result as shown in Figure 15(l).

6.5 Analysis on the network design

The proposed SVLRM uses a fully convolutional neural network with 12 convolutional layers. We further evaluate the effect of model depth by varying the number of the convolutional layer from 6 to 36. Table 8 shows the evaluation results on the depth image denoising task. While stacking more models improves the image quality, the model size and computation cost grow significantly. We empirically use 12 convolutional layers as a trade-off between image quality and efficiency.

In addition to CNNs, it is of great interest to analyze whether using more advanced models can generate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Noisy input (b) α by (12) (c) β by (12) (d) Result by (12)

(e) Guidance image (f) α by (16) (g) β by (16) (h) Result by (16)

(i) GT (j) $\alpha(G, I)$ by SVLRM (k) $\beta(G, I)$ by SVLRM (l) SVLRM

Fig. 15: Comparisons of the depth denoising results with different hand-crafted priors. Modeling the properties of the coefficients by hand-crafted priors is not a trivial task as it is quite difficult to describe the statistical properties of the linear representation coefficients. Thus, the models based on the commonly used hand-crafted priors do not generate clear images. In contrast, we develop a deep CNN which is constrained by the SVLRM to estimate the coefficients. With the estimated linear representation coefficients, the proposed method generates better denoised results (Best viewed on high-resolution display with zoom-in).

TABLE 8: Ablation study on the network design of the proposed SVLRM. "SVLRM-i" ($i = 6, 12, 20, 28, 36$) denotes the SVLRM method with i convolutional layers.

Method	SVLRM-6	SVLRM-12 (Ours)	SVLRM-20	SVLRM-28	SVLRM-36	SVLRM-Enc
Avg. PSNRs	36.13	37.53	37.89	38.04	38.11	37.92
Avg. SSIMs	0.9599	0.9696	0.9711	0.9717	0.9722	0.9714
Avg. RMSEs	4.09	3.50	3.36	3.32	3.29	3.37
Model size (#M)	0.15	0.37	0.67	0.96	1.26	6.81

better results or not. We develop a method that estimates the representation coefficients by an encoder and decoder architecture based on [24] with residual learning [79] (denote it by SVLRM-Enc). However, we do not use the ConvLSTM module and the multi-scale strategy. Table 8 shows that using more advanced models facilitates estimating more effective representation coefficients, while the performance improvement is not significant considering the increased model size. As such, we use CNN with SVLRM in this work.

6.6 Higher-order representation models

We extend the proposed method with a second-order scheme to analyze the performance for joint filtering. Let target image F be

$$F = \gamma(G, I)G^2 + \alpha(G, I)G + \beta(G, I), \tag{18}$$

where $\gamma(G, I)$, $\alpha(G, I)$ and $\beta(G, I)$ are the spatially variant representation coefficients that are determined by G and I. Similar to the optimization method used in the linear representation model, we modify the constraint (8) based on (18) and constrain the network $\mathcal{F}$ by

$$\begin{aligned}\mathcal{F}_\Theta(G^n; I^n) =& \mathcal{F}_{\Theta_\gamma}(G^n; I^n)(G^n)^2 \\ &+ \mathcal{F}_{\Theta_\alpha}(G^n; I^n)G^n + \mathcal{F}_{\Theta_\beta}(G^n; I^n),\end{aligned} \tag{19}$$

where Θ_γ, Θ_α, and Θ_β are the network parameters. We use $\mathcal{F}_{\Theta_\gamma}(G^n; I^n)$, $\mathcal{F}_{\Theta_\alpha}(G^n; I^n)$, and $\mathcal{F}_{\Theta_\beta}(G^n; I^n)$ as the representation coefficients. We use the same loss function (9) and experimental settings as the SVLRM to train the network $\mathcal{F}$. The network parameters are updated by

$$\Theta_v^t = \Theta_v^{t-1} - \lambda \frac{\partial \mathcal{F}_{\Theta_v}}{\partial \Theta_v} \frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_v}}, \quad v \in \{\gamma, \alpha, \beta\}, \tag{20}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3102575, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 9: Effect of the proposed SVLRM and the higher-order representation models on depth image denoising in terms of PSNR, SSIM, and RMSE.

Methods	SVLRM	SVLRM-2nd	SVLRM-3rd	SVLRM-Enc	SVLRM-Enc-2nd	SVLRM-Enc-3rd
Avg. PSNRs	37.53	37.68	37.60	37.92	38.18	38.34
Avg. SSIMs	0.9696	0.9699	0.9697	0.9714	0.9722	0.9726
Avg. RMSEs	3.50	3.44	3.47	3.37	3.26	3.21

TABLE 10: Run-time (seconds) performance. All the methods are evaluated on the same machine using the depth image denoising test dataset.

Methods	JBU [3]	MUJF [21]	MUJIF [21]	DJF [19]	SVLRM	SVLRM-Enc
Avg. run-time	5.18	0.96	5.44	0.76	**0.003**	0.009

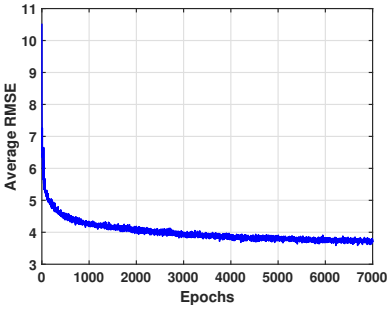


Fig. 16: Quantitative evaluation of the convergence property on the depth image denoising test dataset used in Section 5.3. The deep CNN used for the spatially linear representation coefficient estimation converges well.

where

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\gamma}} &= \sum_{n=1}^{N} \frac{(G^n)^2 \left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)}{\sqrt{\left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)^2 + \varepsilon^2}}, \\
\frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\alpha}} &= \sum_{n=1}^{N} \frac{G^n \left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)}{\sqrt{\left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)^2 + \varepsilon^2}}, \\
\frac{\partial \mathcal{L}}{\partial \mathcal{F}_{\Theta_\beta}} &= \sum_{n=1}^{N} \frac{\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n}{\sqrt{\left(\mathcal{F}_\Theta(G^n; I^n) - F_{gt}^n\right)^2 + \varepsilon^2}}.
\end{aligned} \tag{21}$$

Similar to the SVLRM, we use the same settings to compute the representation coefficients. The model (18) can also be extended to a third-order representation model. We evaluate the proposed SVLRM, SVLRM-Enc, the second-order representation model (SVLRM-2nd, SVLRM-Enc-2nd for short), and the third-order representation model (SVLRM-3rd, SVLRM-Enc-3rd for short) on the depth image denoising task. Table 9 shows the evaluation results. While using the higher-order representation is likely to have better approximation than the linear representation model, the results in Table 9 show that only minor improvements can be achieved with these schemes. These results demonstrate that using the proposed linear representation is effective for numerous joint filtering tasks considered in this work.

6.7 Convergence and Run-time

To quantitatively evaluate the convergence properties of the proposed algorithm, we evaluate our method on the depth image denoising test dataset used in Section 5.3. Figure 16 shows that the proposed network converges well.

We benchmark the run-time of all methods on a machine with an Intel Xeon E5-2650 v4 CPU and an NVIDIA TITAN Xp GPU. Table 10 shows that the proposed approach performs more efficiently than other deep learning-based approaches.

7 Concluding Remarks

In this paper, we propose a joint filter based on the SVLRM and develop an efficient approach 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 model captures the structural information of both guidance image and input image well. Thus, the proposed model is able to transfer meaningful structures to the target image for joint filtering. We show that the proposed approach can be effectively applied to a variety of applications and performs favorably against the state-of-the-art methods that have been specially designed for each task.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (No. 2018AAA0102001), the National Natural Science Foundation of China (Nos. 61922043, 61872421, 61732007), the Natural Science Foundation of Jiangsu Province (No. BK201xispell80471), and the Fundamental Research Funds for the Central Universities (No. 30920041109). M.-H. Yang is supported in part by the National Science Foundation CAREER award (No. 1149783).

References

[1] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM TOG*, vol. 31, no. 6, pp. 1–10, 2012.
[2] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE TPAMI*, vol. 35, no. 6, pp. 1397–1409, 2013.
[3] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM TOG*, vol. 26, no. 3, p. 96, 2007.

[4] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, 1998, pp. 839–846.

[5] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE TPAMI*, vol. 30, no. 2, pp. 228–242, 2008.

[6] J. Xiao, H. Cheng, H. S. Sawhney, C. Rao, and M. A. Isnardi, "Bilateral filtering-based optical flow estimation with occlusion detection," in *ECCV*, 2006, pp. 211–224.

[7] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *CVPR*, 2011, pp. 3017–3024.

[8] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *CVPR*, 2010, pp. 2432–2439.

[9] X. Shen, C. Zhou, L. Xu, and J. Jia, "Mutual-structure for joint filtering," *IJCV*, vol. 125, no. 1-3, pp. 19–33, 2017.

[10] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *ICCV*, 2013, pp. 49–56.

[11] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *CVPR*, 2007.

[12] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I.-S. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *ICCV*, 2011, pp. 1623–1630.

[13] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *ICCV*, 2013, pp. 993–1000.

[14] Q. Yan, X. Shen, L. Xu, S. Zhuo, X. Zhang, L. Shen, and J. Jia, "Cross-field joint image restoration via scale map," in *ICCV*, 2013, pp. 1537–1544.

[15] B. Ham, M. Cho, and J. Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE TPAMI*, vol. 40, no. 1, pp. 192–207, 2018.

[16] X. Guo, Y. Li, and J. Ma, "Mutually guided image filtering," in *ACM MM*, 2017, pp. 1283–1290.

[17] R. J. Jevnisek and S. Avidan, "Co-occurrence filter," in *CVPR*, 2017, pp. 3816–3824.

[18] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof, "A deep primal-dual network for guided depth super-resolution," in *BMVC*, 2016.

[19] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *ECCV*, 2016, pp. 154–169.

[20] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, "Learning dynamic guidance for depth image enhancement," in *CVPR*, 2017, pp. 712–721.

[21] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *ECCV*, 2016, pp. 353–369.

[22] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *CVPR*, 2018, pp. 1838–1847.

[23] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Blind image deblurring using dark channel prior," in *CVPR*, 2016, pp. 1628–1636.

[24] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *CVPR*, 2018, pp. 8174–8182.

[25] J. Pan, J. Dong, J. S. J. Ren, L. Lin, J. Tang, and M.-H. Yang, "Spatially variant linear representation models for joint filtering," in *CVPR*, 2019, pp. 1702–1711.

[26] J. Chen, S. Paris, and F. Durand, "Real-time edge-aware image processing with the bilateral grid," *ACM TOG*, vol. 26, no. 3, p. 103, 2007.

[27] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *SIGGRAPH*, 2002, pp. 257–266.

[28] Q. Zhang, L. Xu, and J. Jia, "100+ times faster weighted median filter (WMF)," in *CVPR*, 2014, pp. 2830–2837.

[29] A. Criminisi, T. Sharp, C. Rother, and P. Pérez, "Geodesic image and video editing," *ACM TOG*, vol. 29, no. 5, pp. 134:1–134:15, 2010.

[30] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM TOG*, vol. 30, no. 4, pp. 69:1–69:12, 2011.

[31] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE TIP*, vol. 21, no. 3, pp. 1176–1190, 2012.

[32] Q. Zhang, X. Shen, L. Xu, and J. Jia, "Rolling guidance filter," in *ECCV*, 2014, pp. 815–830.

[33] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," *ACM TOG*, vol. 27, no. 3, pp. 67:1–67:10, 2008.

[34] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via L_0 gradient minimization," *ACM TOG*, vol. 30, no. 6, pp. 174:1–174:12, 2011.

[35] L. Xu, J. S. J. Ren, Q. Yan, R. Liao, and J. Jia, "Deep edge-aware filters," in *ICML*, 2015, pp. 1669–1678.

[36] S. Liu, J. Pan, and M.-H. Yang, "Learning recursive filters for low-level vision via a hybrid neural network," in *ECCV*, 2016, pp. 560–576.

[37] J. Pan, S. Liu, D. Sun, J. Zhang, Y. Liu, J. Ren, Z. Li, J. Tang, H. Lu, Y.-W. Tai, and M.-H. Yang, "Learning dual convolutional neural networks for low-level vision," in *CVPR*, 2018, pp. 3070–3079.

[38] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM TOG*, vol. 35, no. 2, pp. 11:1–11:15, 2016.

[39] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *ICCV*, 2017, pp. 2516–2525.

[40] V. Jampani, M. Kiefel, and P. V. Gehler, "Learning sparse high dimensional filters: Image filtering, dense CRFs and bilateral neural networks," in *CVPR*, 2016, pp. 4452–4461.

[41] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM TOG*, vol. 36, no. 4, pp. 118:1–118:12, 2017.

[42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *ECCV*, 2012, pp. 746–760.

[43] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *NIPS*, 2005, pp. 291–298.

[44] J. T. Barron and B. Poole, "The fast bilateral solver," in *ECCV*, 2016, pp. 617–632.

[45] Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li, "Depth upsampling based on deep edge-aware learning," *Pattern Recognition*, vol. 103, p. 107274, 2020.

[46] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, "PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution," *IEEE TIP*, vol. 29, pp. 7427–7442, 2020.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[48] S. Lu, X. Ren, and F. Liu, "Depth enhancement via low-rank matrix completion," in *CVPR*, 2014, pp. 3390–3397.

[49] B. Lecouat, J. Ponce, and J. Mairal, "Fully trainable and interpretable non-local sparse models for image restoration," in *ECCV*, 2020.

[50] X. Wu, M. Liu, Y. Cao, D. Ren, and W. Zuo, "Unpaired learning of deep image denoising," in *ECCV*, 2020, pp. 352–368.

[51] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001, pp. 416–425.

[52] K. Dabov, A. Foi, V. Katkovnik, and K. O. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE TIP*, vol. 16, no. 8, pp. 2080–2095, 2007.

[53] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *ICCV*, 2011, pp. 479–486.

[54] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *CVPR*, 2014, pp. 2774–2781.

[55] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *CVPR*, 2012, pp. 2392–2399.

[56] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *CVPR*, 2017, pp. 2808–2817.

[57] S. Anwar, C. P. Huynh, and F. Porikli, "Identity enhanced residual image denoising," in *CVPR Workshops*, 2020, pp. 520–521.

[58] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014, pp. 184–199.

[59] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

[60] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE TIP*, vol. 27, no. 9, pp. 4608–4622, 2018.
[61] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *CVPR*, 2019, pp. 1712–1722.
[62] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *CVPR*, 2020, pp. 3482–3492.
[63] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *CVPR*, 2018, pp. 1692–1700.
[64] Y. Liu, S. Anwar, L. Zheng, and Q. Tian, "Gradnet image denoising," in *CVPR Workshops*, 2020, pp. 508–509.
[65] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *ECCV*, 2020, pp. 41–58.
[66] M. Chang, Q. Li, H. Feng, and Z. Xu, "Spatial-adaptive network for single image denoising," in *ECCV*, 2020, pp. 171–187.
[67] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE TIP*, vol. 26, no. 7, pp. 3142–3155, 2017.
[68] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *CVPR*, 2019, pp. 3155–3164.
[69] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *ECCV*, 2010, pp. 157–170.
[70] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *CVPR*, 2011, pp. 2657–2664.
[71] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "L_0-regularized intensity and gradient prior for deblurring text images and beyond," *IEEE TPAMI*, vol. 39, no. 2, pp. 342–355, 2017.
[72] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Deblurring images via dark channel prior," *IEEE TPAMI*, vol. 40, no. 10, pp. 2315–2328, 2018.
[73] S. Zhuo, D. Guo, and T. Sim, "Robust flash deblurring," in *CVPR*, 2010, pp. 2440–2447.
[74] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. V. Gool, "DSLR-quality photos on mobile devices with deep convolutional networks," in *ICCV*, 2017, pp. 3297–3305.
[75] A. Ignatov, R. Timofte, T. V. Vu, T. M. Luu, and et al., "PIRM challenge on perceptual image enhancement on smartphones: Report," in *ECCV Workshops*, 2018.
[76] J. Dong, J. Pan, D. Sun, Z. Su, and M.-H. Yang, "Learning data terms for non-blind deblurring," in *ECCV*, 2018, pp. 777–792.
[77] J. Pan, W. Ren, Z. Hu, and M.-H. Yang, "Learning to deblur images with exemplars," *IEEE TPAMI*, 2018.
[78] J. Zhu, J. Zhang, Y. Cao, and Z. Wang, "Image guided depth enhancement via deep fusion and local linear regularizaron," in *ICIP*, 2017, pp. 4068–4072.
[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

Jiangxin Dong is a postdoctoral researcher at the Max Planck Institute for Informatics. She did her Ph.D. at Dalian University of Technology. Her research interests include image deblurring and denoising.

Jinshan Pan is a professor of School of Computer Science and Engineering, Nanjing University of Science and Technology. He received the Ph.D. degree in computational mathematics from the Dalian University of Technology, China, in 2017. He was a joint-training Ph.D. student in School of Mathematical Sciences at and Electrical Engineering and Computer Science at University of California, Merced, CA, USA from 2014 to 2016. His research interest includes image deblurring, image/video analysis and enhancement, and related vision problems.

Jimmy S. Ren is a research director at SenseTime Research. His research interests include computational photography, computational imaging and deep learning. He is also an adjunct faculty member of Qing Yuan Research Institute, Shanghai Jiao Tong University. He got his Ph.D. degree from City University of Hong Kong.

Liang Lin is CEO of DMAI Great China and a full professor of Computer Science in Sun Yat-sen University. He served as the Executive Director of the SenseTime Group from 2016 to 2018, leading the R&D teams in developing cutting-edge, deliverable solutions in computer vision, data analysis and mining, and intelligent robotic systems. He is an associate editor of IEEE Trans, Human-Machine Systems and IET Computer Vision, and he served as the area/session chair for numerous conferences, such as CVPR, ICME, ICCV, ICMR. He was the recipient of Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Dimond Award for best paper in IEEE ICME in 2017, ACM NPAR Best Paper Runners-Up Award in 2010, Google Faculty Award in 2012, award for the best student paper in IEEE ICME in 2014, and Hong Kong Scholars Award in 2014. He is a Fellow of IET.

Jinhui Tang is a professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. He received the BE and PhD degrees from the University of Science and Technology of China, in 2003 and 2008, respectively. He received the Best Student Paper Award in MMM 2016, and Best Paper Awards in ACM MM 2007, PCM 2011, and ICIMCS 2011. He is a senior member of the IEEE and a member of the ACM.

Ming-Hsuan Yang is affiliated with University of California at Merced, Yonsei University, and Google. He received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign, USA, in 2000. He served as an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, and is an Associate Editor of the International Journal of Computer Vision, Image and Vision Computing, and Journal of Artificial Intelligence Research. He received the NSF CAREER Award in 2012, and the Google Faculty Award in 2009. He is a fellow of the IEEE.