Learning Spatially-Variant MAP Models for Non-blind Image Deblurring

Jiangxin Dong
MPI Informatics
Saarland Informatics Campus

Stefan Roth
TU Darmstadt
& hessian.AI

Bernt Schiele
MPI Informatics
Saarland Informatics Campus

Abstract

The classical maximum a-posteriori (MAP) framework for non-blind image deblurring requires defining suitable data and regularization terms, whose interplay yields the desired clear image through optimization. The vast majority of prior work focuses on advancing one of these two crucial ingredients, while keeping the other one standard. Considering the indispensable roles and interplay of both data and regularization terms, we propose a simple and effective approach to jointly *learn these two terms, embedding deep neural networks within the constraints of the MAP framework, trained in an end-to-end manner. The neural networks not only yield suitable image-adaptive features for both terms, but actually predict per-pixel* spatially-variant *features instead of the commonly used spatially-uniform ones. The resulting spatially-variant data and regularization terms particularly improve the restoration of fine-scale structures and detail. Quantitative and qualitative results underline the effectiveness of our approach, substantially outperforming the current state of the art.*

1. Introduction

The goal of single image deblurring is to estimate a desirable clear image from a blurry input. Mathematically, the process leading to the image blur is frequently formulated as

$$y = x * k + n, \tag{1}$$

where y, x, k, and n denote blurry observation, latent clear image, blur kernel, and image noise, respectively; $*$ is the convolution operator. Significant progress [*e.g.*, 20, 40, 42, 48] has been made in blind image deblurring, which aims to estimate the latent clear image when the blur kernel is unknown. When the blur kernel can be obtained or estimated, this problem reduces to *non-blind image deblurring*, which has been an active area of research since the pioneering work of Richardson and Lucy [29]. Other classical approaches include the Wiener filter [46].

Non-blind image deblurring is a well-known ill-posed problem. Most existing methods formulate it as a *maximum a-posteriori* (MAP) estimation problem [17, 21]:

$$x^* = \arg\max_x p(y \mid x, k)\, p(x), \tag{2}$$

where $p(y \mid x, k)$ denotes the likelihood that measures how consistent the estimated x is with the observation of y and the known k under the model in Eq. (1); $p(x)$ denotes the prior on the latent clear image x, which is used to regularize the problem. Equation (2) can be equivalently reformulated as

$$x^* = \arg\min_x \mathcal{D}(y, x, k) + \mathcal{R}(x), \tag{3}$$

where $\mathcal{D}(\cdot)$ denotes the data term and $\mathcal{R}(\cdot)$ denotes the regularization term [32, 33, 44]. Effectively solving non-blind image deblurring within the MAP framework thus requires carefully designing both $\mathcal{D}(\cdot)$ and $\mathcal{R}(\cdot)$.

To restore high-quality clear images using Eq. (3), numerous approaches have been proposed. One family of methods focuses on advancing the data term to better measure the image reconstruction error. Starting from the most commonly used ℓ_2 norm [17], data terms have been carefully designed for specific types of outliers [1, 6] or even discriminatively learned [11, 28]. A second family of approaches focuses on developing effective regularization terms/image priors to ensure desirable properties of the estimated clear image. This includes modeling statistical properties, *e.g.* by employing Laplacian/hyper-Laplacian priors [17, 21]. Learning effective image priors based on data-driven methods has been a dominant research theme for non-blind image deblurring, *e.g.*, using Gaussian mixture models [52], fields of experts [31], or deep learning [49, 50]. Therefore, most existing non-blind image deblurring methods focus on improving *either* the data term *or* the regularization term. However, as the data and regularization terms play different but indispensable roles in non-blind image deblurring, only improving one of these two terms will limit the power of the MAP framework in Eq. (3) and result in deblurred images with artifacts, see Fig. 1(b)–(d). In contrast, we *jointly* learn *both* the data term and the regularization term to build a more expressive deblurring model in which these two terms can benefit from their interplay, resulting in higher fidelity results as shown in Fig. 1(e).

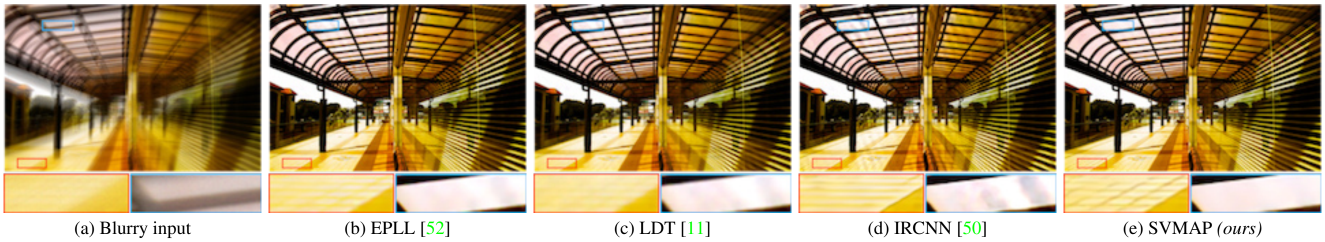

(a) Blurry input (b) EPLL [52] (c) LDT [11] (d) IRCNN [50] (e) SVMAP *(ours)*

Figure 1. *Visual comparison with state-of-the-art non-blind image deblurring methods.* The results in *(b)–(d)* exhibit severe artifacts or do not effectively restore fine-scale structures. In contrast, our approach can recover a clearer image with finer detail as shown in *(e)*.

We further note that existing data or regularization terms are mostly designed or learned to be spatially invariant. While this yields a compact model, the local structures differ notably across the image, *e.g.*, in flat *vs*. textured areas. Moreover, saturated regions or outliers can occur locally. Using uniform data and regularization terms for the whole image thus cannot effectively characterize the spatially-variant properties of the image, hindering the restoration of finer-scale structures and detail (Fig. 1(b)–(d)). To enhance the model expressivity, we propose a *spatially-variant MAP model* (SVMAP) by predicting a set of pixel-dependent filters to adjust the regularization behavior and the treatment of the data-term residuals to the local requirements.

We make the following contributions: *(i)* We propose an expressive filter-based MAP deblurring framework, in which the data and regularization terms are jointly learned based on deep neural networks. A detailed analysis shows that our model is more effective at restoring high-quality images compared to existing ones that focus on improving either the data or the regularization term. *(ii)* To improve the goodness-of-fit and capture the properties of clear images, we construct spatially-variant data and regularization terms by predicting a set of pixel-dependent filters. In contrast to spatially-uniform formulations, our learned pixel-dependent ones are able to model the spatially-variant property of the image structures, facilitating finer-scale structure and detail restoration. *(iii)* We develop an end-to-end learning approach to better capture the spatially-variant properties, integrating the MAP-based optimization framework as a constraint for the deep neural network. *(iv)* Finally, we both quantitatively and qualitatively demonstrate the effectiveness of our method and show that it is able to generate better deblurred results for blurry images with Gaussian noise as well as outliers (*e.g.*, saturated pixels).

2. Related Work

Data term modeling. The data term measures the image reconstruction error and models the image noise distribution. Early methods [1, 17, 32] assume that the image noise conforms to a Gaussian or Laplacian distribution and consequently use an ℓ_2 or ℓ_1 norm-based data term. This has been extended to the case when the noise level is unknown [15, 37]. However, the assumptions of Gaussian or Laplacian noise do not hold in certain challenging environments (*e.g.*, in low light conditions), as the captured image may contain outliers or non-Gaussian noise [6], whose distribution is mixed and not easily modeled by a fixed parametric distribution. To address this, Cho *et al*. [6] analyze the properties of various types of outliers and use them to classify inliers *vs*. outliers. The estimated inliers are used to define the data term. To obtain more flexible and robust data terms, Dong *et al*. [11] learn discriminative shrinkage functions to indirectly model complex noise distributions, inspired by work on learning regularizers [36].

Regularization term modeling. The regularization term determines the property of the restored images, making MAP-based non-blind image deblurring well-posed. Early methods are mostly derived based on statistical prior knowledge, *e.g*., total variation [32, 33, 44], hyper-Laplacian [17, 21] or image-adaptive Laplacian priors [30], internal patch recurrence [25], nonlocal patch-wise image modeling [7], *etc*. To better capture the characteristics of clear images, various methods learn the regularization term rather than adopting a hand-defined one. Early work focuses mostly on Markov random fields (MRFs) to learn generic image priors [31, 34]. Zoran and Weiss [52] learn a prior model of natural image patches based on Gaussian mixture models (GMMs). Over the years, conditional random fields (CRFs) have grown increasingly popular, including Gaussian CRFs [14, 43] or cascades of Gaussian CRFs [35]. Schmidt and Roth [36] derive such a cascade based on more general shrinkage functions. Note that mean-field inference for CRFs can even be integrated in deep networks [51]. Such (discriminative) learning-based methods yield more powerful regularization terms than traditional methods, but the underlying features are usually not image-adaptive. Alternatively, deep learning provides a powerful tool for non-blind image deblurring. Several methods [30, 38, 49, 50] decompose the problem into image deconvolution and denoising, adopting deep networks to solve the sub-problem w.r.t. the regularization term. Kruse *et al*. [18] generalize an FFT-based deconvolution by using a powerful regularization term based on CNNs. Gong *et al*. [12] indirectly model the regularization term by integrating deep neural networks into a parameterized gradient descent scheme.

Joint modeling. The above methods focus on improving either the data or regularization term alone, thus do not fully leverage the power of the MAP framework as both terms play indispensable roles for non-blind image deblurring. Ren *et al.* [28] recently propose to simultaneously learn the data and regularization terms. However, their fixed feature extractors are not image-adaptive, limiting the ability to capture complex noise distributions and image properties. In addition, previous work mostly models and exploits spatially invariant characteristics, thus ignoring locally unique structures. In contrast, our approach *(i) jointly learns image-adaptive data and regularization terms* through deep neural networks, and *(ii)* leverages *pixel-dependent features* owing to the spatially variant properties of image structure and detail. This allows us to build a more expressive deblurring model, benefitting from the interplay of both key ingredients in an image- and spatially-adaptive way.

Other related work. As it is hard to find one solution that can equally address various non-blind image deblurring scenarios, several methods utilize domain knowledge for specific types of noise or outliers. Whyte *et al.* [45] modify the Richardson-Lucy algorithm to prevent error propagation caused by saturated pixels. Xu *et al.* [47] exploit kernel information for deconvolution and embed it into a deep neural network, which performs well for images with saturated pixels. Hu *et al.* [13] explicitly model the property of light streaks in low-light images and detect useful light streaks to help deblur low-light images. Since these methods depend on specific domain knowledge, they cannot be easily extended to handle other types of noise and outliers. Our method, in contrast, is based on a filter-based MAP framework that can adaptively learn a suitable model from the training data, addressing a range of scenarios.

3. Spatially-Variant MAP Model

Our goal is to learn a flexible non-blind image deblurring model for high-quality image restoration. Different from existing methods that focus on improving either the data or the regularization term, we jointly learn these two terms in a unified MAP framework. We formulate the underlying learning task as a bilevel optimization problem [8, 9, 19]:

$$\min_{\mathcal{D}^s,\mathcal{R}^s} \mathcal{L}(\hat{x}, x_{\text{gt}}) \tag{4a}$$

$$\text{s.t.} \quad \hat{x} = \arg\min_{x} \mathcal{D}^s(y,x,k) + \mathcal{R}^s(x), \tag{4b}$$

where $\mathcal{D}^s(\cdot)$ and $\mathcal{R}^s(\cdot)$ denote spatially-variant data and regularization terms, which we will model using deep neural networks; $\mathcal{L}(\cdot)$ denotes the loss function and x_{gt} is the ground truth clear image. Once $\mathcal{D}^s(\cdot)$ and $\mathcal{R}^s(\cdot)$ are learned, we can use the proposed spatially-variant model from Eq. (4b) to restore latent clear images. In the following, we will present how to model and learn the spatially-variant $\mathcal{D}^s(\cdot)$ and $\mathcal{R}^s(\cdot)$ for non-blind image deblurring.

3.1. Data term

In the MAP-based image deblurring model of Eq. (3), previous methods [1, 17] usually define the data term as

$$\mathcal{D}(y,x,k) = \rho(y - x * k), \tag{5}$$

where $\rho(\cdot)$ denotes a penalty function [2, 3]. As pointed out by [11, 28], the data term defined with only the intensity information and a fixed form of $\rho(\cdot)$ is not expressive enough to model the image noise well, and usually leads to deblurred results with artifacts. Thus, more powerful feature information has been introduced to the data term [11] as

$$\mathcal{D}(y,x,k) = \sum_{i=1}^{M} \mathcal{D}_i\big(f_i * (y - x * k)\big), \tag{6}$$

where f_i is the i-th feature extractor, $\mathcal{D}_i(\cdot)$ is the corresponding penalty function, and k is the spatially uniform blur kernel.

However, the noise distribution can be very complex in real applications. Not all areas of the image may be equally affected by noise or outliers (*e.g.*, localized saturation) and the characteristics of the image are also not uniform across the image plane (*e.g.*, smooth *vs.* textured areas). Hence a combination of fixed feature extractors f_i, *i.e.* filters that are the same for all input images, limits the expressivity of the model. Similarly, assuming spatial invariance of the feature extractor for the whole residual image $(y - x * k)$ as in Eq. (6) limits the local adaptivity of the data term. The resulting formulation is thus not effective in fine-scale structure restoration as shown in Fig. 1(c).

To enhance the modeling capacity, we propose a more expressive data term with *image-adaptive and spatially-variant feature extractors*, predicted by a learned non-linear deep neural network. Our learnable data term, assuming a known, spatially uniform blur kernel k, is defined as

$$\mathcal{D}^s(y,x,k) = \sum_{i=1}^{M} \sum_{p\in\mathbb{P}} \mathcal{D}_{i,p}\big(f_{i,p} * (y - x * k)_{(p)}\big), \tag{7}$$

where $f_{i,p}$ and $\mathcal{D}_{i,p}(\cdot)$ denote the i-th pixel-dependent feature extractor for pixel p and the corresponding penalty function to be predicted. $(y - x * k)_{(p)}$ denotes the image patch centered at the p-th pixel of the residual image $(y - x * k)$, and $\mathbb{P}$ is the set of all pixels. Compared to the data term in Eq. (6) with spatially-invariant feature extractors and penalty functions, our learnable data term (Eq. 7) is more flexible, and can generate better results (Sec. 5).

3.2. Regularization term

Existing methods [*e.g.*, 21, 36] usually formulate the regularization term as

$$\mathcal{R}(x) = \sum_{j=1}^{N} \mathcal{R}_j(g_j * x), \tag{8}$$

where g_j is the j-th filter to extract useful features and $\mathcal{R}_j(\cdot)$ is the corresponding penalty function to model desirable properties of the feature $(g_j * x)$. Classical methods usually take the filters $\{g_j\}$ to be horizontal and vertical image derivative operators and model $\mathcal{R}_j(\cdot)$ as the ℓ_1 or ℓ_p $(p < 1)$ norm to impose Laplacian [17] or hyper-Laplacian [21] priors. To overcome the limitations of first-order image gradients and hand-crafted priors, several methods [4, 28, 35, 36] learn linear filters g_j and penalty functions $\mathcal{R}_j(\cdot)$ from training data based on discriminative learning. Although decent image quality has been achieved, various limitations remain: *(i)* the feature extractors g_j are not image-adaptive, and *(ii)* they are spatially-invariant. This limits the expressive power of the regularizer, *e.g.* it cannot capture the spatially-variant characteristics of images, hindering fine-scale detail restoration (Fig. 1(b) and (d)).

To better describe the properties of clear images, we formulate our *spatially-variant regularization term* as

$$\mathcal{R}^s(x) = \sum_{j=1}^{N} \sum_{p \in \mathbb{P}} \mathcal{R}_{j,p}\big(g_{j,p} * x_{(p)}\big), \tag{9}$$

where $\mathcal{R}_{j,p}$ and $g_{j,p}$ are specified per pixel. As demonstrated in Sec. 5, the proposed spatially-variant regularization term predicted by learned deep neural networks is more expressive and effective for non-blind image deblurring.

3.3. Learning and inference

Let $\{y^l, x^l_{\text{gt}}, k^l\}_{l=1}^{L}$ denote a set of L training samples. Putting together Eqs. (4), (7) and (9), the crucial components $\{\mathcal{D}_{i,p}, f_{i,p}\}, \{\mathcal{R}_{j,p}, g_{j,p}\}$ to define our model can thus be learned by solving

$$\min_{\{\mathcal{D}_i, f_i\}, \{\mathcal{R}_j, g_j\}} \sum_{l=1}^{L} \mathcal{L}(\hat{x}^l, x^l_{\text{gt}}), \tag{10a}$$

$$\text{s.t.} \quad \hat{x}^l = \arg\min_{x^l} \sum_{p \in \mathbb{P}} \Bigg[\sum_{i=1}^{M} \mathcal{D}_{i,p}\big(f_{i,p} * (y^l - x^l * k^l)_{(p)}\big) + \sum_{j=1}^{N} \mathcal{R}_{j,p}\big(g_{j,p} * (x^l)_{(p)}\big) \Bigg]. \tag{10b}$$

We here use the ℓ_1 norm to define the robust image loss $\mathcal{L}(\cdot)$. Next, we will discuss how to solve the inner optimization problem in the constraint of Eq. (10b) to obtain the latent image. Moreover, we develop a method to train the required learnable components in an end-to-end manner.

The constraint in Eq. (10b) is a highly non-convex optimization problem. To solve it, we adopt the Iteratively Reweighted Least Squares (IRLS) method and iteratively solve the weighted quadratic problem

$$\min_{x} \sum_{p \in \mathbb{P}} \Bigg[\sum_{i=1}^{M} \omega^d_{i,p} \big| f_{i,p} * (y - x * k)_{(p)} \big|^2 + \sum_{j=1}^{N} \omega^r_{j,p} \big| g_{j,p} * x_{(p)} \big|^2 \Bigg], \tag{11}$$

where $\{\omega^d_{i,p}\}$ and $\{\omega^r_{j,p}\}$ are the pixel-wise weights for the data and regularization terms. For ease of notation, we omit the training sample index l here. Note that we assume a spatially uniform blur kernel k.

As the filters $\{f_{i,p}\}$ and $\{g_{j,p}\}$ are pixel-dependent in our formulation, the weights $\{\omega^d_{i,p}\}$ and $\{\omega^r_{j,p}\}$ can be absorbed into $\{f_{i,p}\}$ and $\{g_{j,p}\}$, respectively. Let $\mathbf{F}_i$, $\mathbf{G}_j$, $\mathbf{y}$, $\mathbf{K}$, and $\mathbf{x}$ denote the vector/matrix forms of f_i, g_j, y, k, and x. We can then reformulate Eq. (11) as

$$\min_{\mathbf{x}} \sum_{p \in \mathbb{P}} \Bigg[\sum_{i=1}^{M} \big\| \mathbf{F}_{i,p}(\mathbf{y} - \mathbf{K}\mathbf{x})_{(p)} \big\|^2 + \sum_{j=1}^{N} \big\| \mathbf{G}_{j,p} \mathbf{x}_{(p)} \big\|^2 \Bigg]. \tag{12}$$

In other words, learning $\{\mathcal{D}_{i,p}, f_{i,p}\}, \{\mathcal{R}_{j,p}, g_{j,p}\}$ is equivalent to learning $\{\mathbf{F}_{i,p}\}$ and $\{\mathbf{G}_{j,p}\}$. Similar to the classical IRLS method used in [21], we alternatingly update $\{\mathbf{F}_{i,p}\}$, $\{\mathbf{G}_{j,p}\}$ and solve Eq. (12), noting that Eq. (12) is a least squares problem whose solution can be easily obtained. What remains to be addressed is to specify deep neural networks to effectively predict $\{\mathbf{F}_{i,p}\}$ and $\{\mathbf{G}_{j,p}\}$.

To this end, we develop the networks $\mathcal{N}_f$ and $\mathcal{N}_g$ to predict the filters $\{\mathbf{F}_{i,p}\}$ and $\{\mathbf{G}_{j,p}\}$, respectively, in an image-adaptive fashion. Consequently, the estimate of the latent image from the current IRLS iteration is fed as input to the networks $\mathcal{N}_f$ and $\mathcal{N}_g$; at the first iteration, the deconvolved result with an ℓ_2 norm and a Gaussian prior is used. Both networks share the same architecture, consisting of 6 convolutional layers followed by a ReLU, except for the last layer. The filter size is 3×3 pixels; their stride is 1. We use 64 features in the first 5 layers. Let us assume that filters $\{\mathbf{F}_{i,p}\}$ and $\{\mathbf{G}_{j,p}\}$ have $s_f \times s_f$ and $s_g \times s_g$ pixels. Then similar to [26], we let the networks $\mathcal{N}_f$ and $\mathcal{N}_g$ predict $s_f^2 M$ and $s_g^2 N$ channels. We reshape the outputs of the networks $\mathcal{N}_f$ and $\mathcal{N}_g$ to M filters of size $s_f \times s_f$ and N filters of $s_g \times s_g$ pixels. With the predicted pixel-dependent filters, we finally use the conjugate gradient method to solve Eq. (12). Note that we use the same network architecture in different IRLS iterations, but the network parameters are not shared across iterations. During training, we update the network parameters of $\mathcal{N}_f$ and $\mathcal{N}_g$ by minimizing the loss function in Eq. (10a) based on the final solution of Eq. (10b).

At test time, the latent clear image is estimated by solving the spatially-variant MAP model in Eq. (10b) with the same IRLS method used in the training phase, *i.e.* iteratively updating $\mathbf{x}$ via Eq. (12). In each iteration, the filters $\{\mathbf{F}_{i,p}\}$ and $\{\mathbf{G}_{j,p}\}$ are predicted by the learned models $\mathcal{N}_f$ and $\mathcal{N}_g$, whose parameters are iteration-specific as mentioned above.

4. Experimental Results

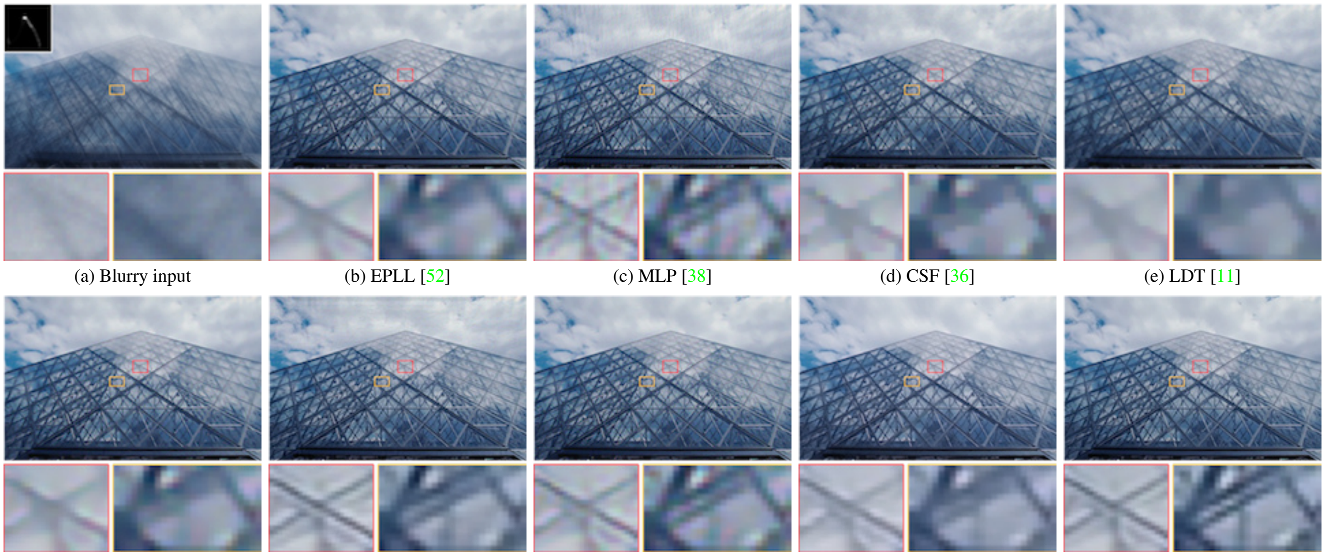
Next, we discuss the datasets and implementation details for our proposed SVMAP approach. Then we evaluate our method on images with simulated Gaussian noise and saturated pixels, as well as on real-world images. More results are included in the supplemental material.

Table 1. *Quantitative comparison to state-of-the-art methods* on the datasets of [24] and [41] with Gaussian noise.

Dataset	Noise level		EPLL [52]	MLP [38]	CSF [36]	LDT [11]	FCN [49]	IRCNN [50]	FDN [18]	FNBD [39]	RGDN [12]	SVMAP *(ours)*
Martin et al. [24]	1%	PSNR (dB)	29.81	28.47	29.00	28.20	29.51	30.63	29.93	30.92	29.51	**31.89**
		SSIM	0.8385	0.7977	0.8230	0.7922	0.8339	0.8645	0.8555	0.8799	0.8616	**0.8973**
	5%	PSNR (dB)	24.66	24.01	24.93	24.90	25.45	25.65	25.93	25.49	25.33	**27.25**
		SSIM	0.6276	0.5619	0.6428	0.6358	0.6771	0.6640	0.6943	0.6589	0.6688	**0.7550**
Sun et al. [41]	1%	PSNR (dB)	32.48	31.47	31.52	30.52	32.36	33.57	32.63	31.22	31.25	**34.51**
		SSIM	0.8815	0.8535	0.8622	0.8399	0.8853	0.8977	0.8887	0.8860	0.8869	**0.9273**
	5%	PSNR (dB)	26.78	24.65	26.62	26.71	27.67	27.64	27.75	27.63	26.93	**29.20**
		SSIM	0.6975	0.5198	0.6735	0.6694	0.7340	0.6884	0.7319	0.7010	0.7161	**0.7940**

Table 2. *Quantitative comparison to state-of-the-art methods* on a dataset with saturated pixels (see text for details).

	EPLL [52]	MLP [38]	CSF [36]	LDT [11]	FCN [49]	IRCNN [50]	FDN [18]	FNBD [39]	RGDN [12]	Whyte [45]	Cho [6]	SVMAP *(ours)*
PSNR (dB)	29.78	28.60	29.28	30.52	29.14	29.92	28.20	27.48	28.61	28.14	33.02	**33.91**
SSIM	0.8950	0.8652	0.8931	0.9167	0.8789	0.9089	0.8560	0.8739	0.8919	0.8824	0.9388	**0.9529**



(a) Blurry input (b) EPLL [52] (c) MLP [38] (d) CSF [36] (e) LDT [11]

(f) FCN [49] (g) IRCNN [50] (h) FDN [18] (i) RGDN [12] (j) SVMAP *(ours)*

Figure 2. *Example with simulated blur (1% noise level) from the dataset of [24].* The deblurred images by [18, 38] exhibit severe artifacts, *cf*. *(c)* and *(h)*. For other methods, fine-scale detail is not effectively recovered, see *(b)*, *(d)–(g)*, and *(i)*. Compared to existing methods, our approach can effectively preserve finer detail as shown in *(j)*. (Best viewed on high-resolution displays.)

4.1. Datasets and implementation details

Training dataset. We collect 400 images from the Berkeley segmentation dataset [24] and 4744 images from the Waterloo Exploration dataset [23] for training. Following previous work [49, 50], we randomly crop image patches of 240×240 pixels from the clear images and convolve each crop with a simulated realistic kernel [35], where the kernel size ranges from 13×13 to 35×35 pixels. Then we add 1% or 5% Gaussian noise, respectively, to each blurry image.

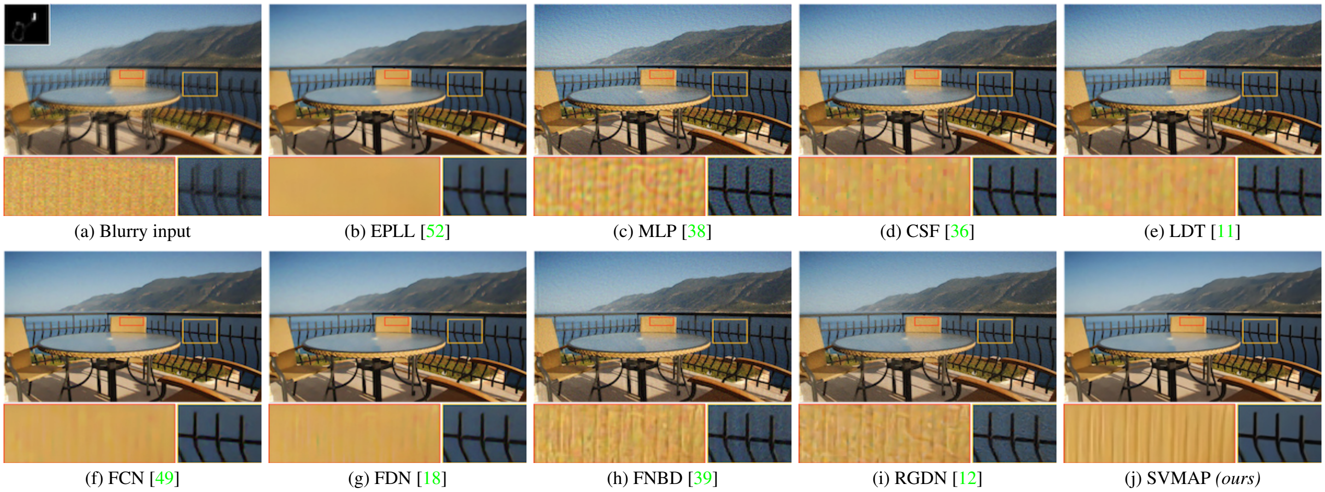
For training the model to handle images with saturated pixels, we collect 500 low-light images from Flickr. Similar to [27], we first enlarge the intensity range of the clear images with a factor of 1.2 and convolve with a simulated blur kernel [35] to generate the blurry image. Afterwards, 0–1% Gaussian noise is added. Finally, both clear and blurry images are clipped to the intensity range of $[0, 1]$.

Test dataset. To evaluate our approach on images with Gaussian noise, we first test on the dataset of [24], which contains 100 clear images; the blurry input is generated as described above. We then validate our method on the dataset of [41] (not trained on), containing 80 clear images with 8 blur kernels from [22]. We test our model on these test datasets with 1% and 5% Gaussian noise, respectively.

To evaluate the effect of the proposed model on images with saturated pixels, we collect 44 low-light images from the literature [6, 10, 27, 47] and use the same method as above to generate the saturated blurry images.

Implementation details.[1] The network is trained using the Adam optimizer [16] with default parameters. The batch size is set to 2. The learning rate is initialized as 5×10^{-5} and halved every 200 epochs. We empirically use $M = 3$ and $N = 5$ pixel-dependent filters for the data term and

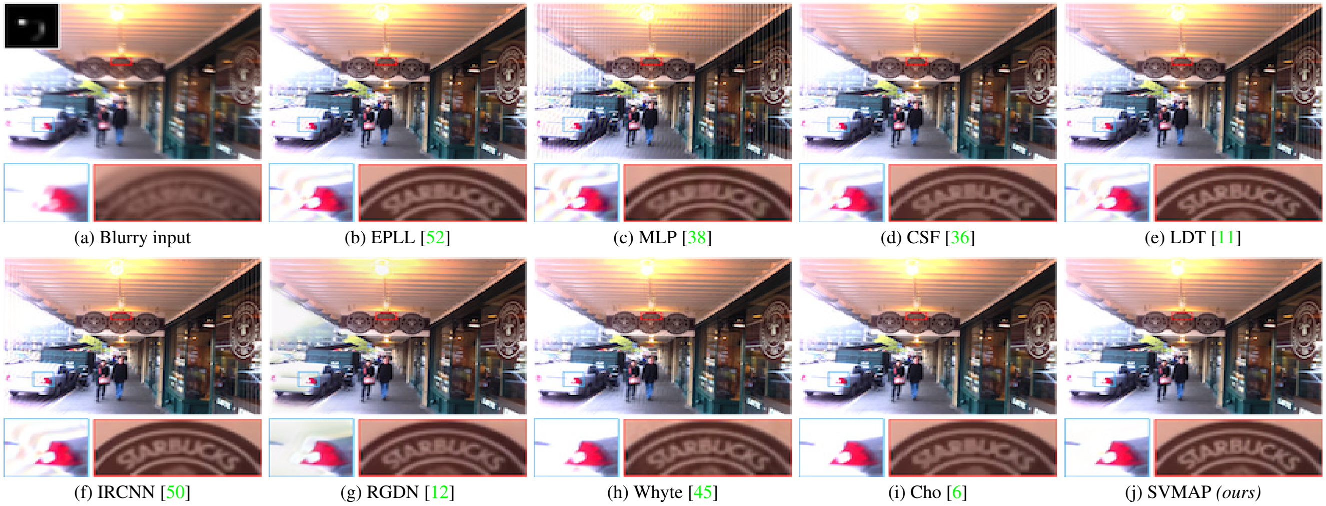
[1] PyTorch code and trained models are available at gitlab.mpi-klsb.mpg.de/jdong/svmap. See also supplemental material.



(a) Blurry input (b) EPLL [52] (c) MLP [38] (d) CSF [36] (e) LDT [11]

(f) FCN [49] (g) FDN [18] (h) FNBD [39] (i) RGDN [12] (j) SVMAP *(ours)*

Figure 3. *Example with simulated blur* (5% *noise level) from the dataset of [41].* The results by [11, 12, 36, 38, 39] contain significant artifacts, see *(c)–(e), (h),* and *(i)*. The methods from [18, 49, 52] oversmooth fine-scale structures, *cf*. *(b), (f)*, and *(g)*. In contrast, our approach effectively restores a clear image with finer detail as shown in *(j)*.



(a) Blurry input (b) EPLL [52] (c) MLP [38] (d) CSF [36] (e) LDT [11]

(f) IRCNN [50] (g) RGDN [12] (h) Whyte [45] (i) Cho [6] (j) SVMAP *(ours)*

Figure 4. *Example with simulated blur and saturated pixels from [47].* The results in *(b)–(g)* exhibit severe artifacts or color distortions and do not effectively restore small-scale structures. In contrast, our approach can preserve finer detail as shown in *(j)*.

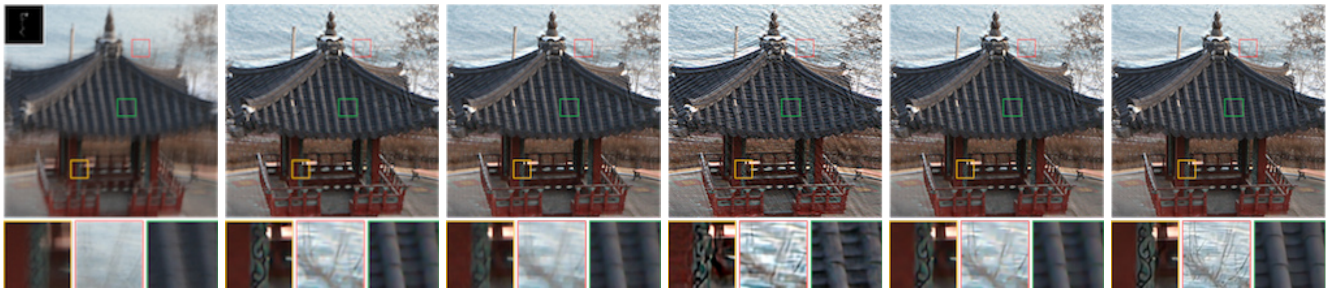
the regularization term, respectively, and set $s_f = s_g = 5$. Balancing effectiveness and efficiency, the number of IRLS iterations, *i.e*. iteratively updating $\{\mathbf{F}_{i,p}\}$ as well as $\{\mathbf{G}_{j,p}\}$ and estimating the latent clear image, is set to 2 unless specified otherwise. We apply 5 conjugate gradient iterations to solve Eq. (12) at each IRLS step.

4.2. Results with simulated blur

We compare our SVMAP approach with state-of-the-art non-blind image deblurring methods. For fair comparison, we finetune all learning-based methods using the same training dataset as ours.

Blurry images with Gaussian noise. We first evaluate our approach on the datasets of Martin *et al*. [24] and Sun *et al*. [41] (not trained on) in Tab. 1. Our approach significantly outperforms the competing methods on images with various noise levels, improving the PSNR by at least 0.94dB (1% Gaussian noise) and 1.32dB (5% noise), respectively. It generalizes well to the unseen dataset of [41]. Among the competing methods, the approaches [36, 38, 49, 50, 52] mainly focus on learning effective priors and [11] proposes to learn robust data terms. All these methods only utilize spatially-invariant feature extractors. In contrast, our method jointly optimizes both the data and regularization terms with spatially-variant image-adaptive filter learning, which facilitates these two terms adapting to each other and leads to higher quality results.

Figure 2 shows an example from [24] with 1% Gaussian noise. The results generated by [18, 38] contain significant artifacts as shown in Fig. 2(c) and (h). The methods of [11, 12, 36, 49, 50, 52] do not effectively restore fine-scale image detail in the deblurred images as seen in Fig. 2(b),



(a) Blurry input (b) EPLL [52] (c) LDT [11] (d) IRCNN [50] (f) SFARL [28] (g) SVMAP *(ours)*

Figure 5. *Example with real camera shake from [5].* The result obtained by [50] in *(d)* has severe artifacts. The methods [11, 28, 52] do not recover small-scale structures, *cf*. *(b), (c)*, and *(f)*. Compared to competing methods, our approach can preserve finer detail as seen in *(g)*.

(d)–(g), and (i). Compared to these competing methods, our approach recovers a much clearer image with finer detail, *cf*. Fig. 2(j). We further show visual comparisons on an image from [41] with 5% Gaussian noise in Fig. 3. The competing methods are not able to recover fine-scale structures, *e.g*., the chair back in Fig. 3(b)–(i). In contrast, our result in Fig. 3(j) is much clearer with finer detail.

Blurry images with saturated pixels. Next, we evaluate our approach on the dataset with saturated pixels. Table 2 shows that our method continues to outperform previous work by a significant margin. The average PSNR from our approach is at least 0.89dB higher than for the competing methods. As the approaches of [18, 36, 38, 49, 50, 52] adopt a fixed ℓ_2 norm-based data term, they are not robust to images with saturated pixels. The method of [6] iteratively estimates outliers and latent images based on the EM algorithm and performs better than other previous methods. Compared to existing methods, even specialized ones, our approach can adaptively learn a more effective deblurring model. Figure 4 shows one example from [47], where our result shows much clearer characters.

4.3. Results with real blur

We further evaluate our method on images with real camera shake and unknown noise or outliers. Figure 5 shows a real captured image from [5]. The method of [50] results in severe artifacts in the deblurred result, see Fig. 5(d). The results obtained by the methods of [11, 28, 52] are oversmoothed as shown in Fig. 5(b), (c), and (f). In contrast, our approach restores much clearer images with finer detail, *e.g*., the branches and patterns in Fig. 5(g).

5. Analysis and Discussion

Table 3. *Effectiveness of jointly learning the data and regularization terms.* All methods are evaluated on the dataset of [24] with 1% Gaussian noise and the dataset with saturated pixels (see text), where the kernel size ranges from 13×13 to 35×35 pixels.

	Dataset with 1% Gaussian noise				
	Data term $\mathcal{D}$		Regularization term $\mathcal{R}$		PSNR (dB) /SSIM
	ℓ_2 norm	learned	hyper-Laplacian	learned	
Fix $\mathcal{D}$&$\mathcal{R}$	✓	✗	✓	✗	29.21/0.8260
Learn $\mathcal{D}$ & Fix $\mathcal{R}$	✗	✓	✓	✗	29.88/0.8525
Fix $\mathcal{D}$ & Learn $\mathcal{R}$	✓	✗	✗	✓	31.59/0.8917
SVMAP *(ours)*	✗	✓	✗	✓	**31.89/0.8973**

	Dataset with saturated pixels				
	Data term $\mathcal{D}$		Regularization term $\mathcal{R}$		PSNR (dB) /SSIM
	ℓ_2 norm	learned	hyper-Laplacian	learned	
Fix $\mathcal{D}$&$\mathcal{R}$	✓	✗	✓	✗	29.15/0.8917
Learn $\mathcal{D}$ & Fix $\mathcal{R}$	✗	✓	✓	✗	30.58/0.9183
Fix $\mathcal{D}$ & Learn $\mathcal{R}$	✓	✗	✗	✓	29.47/0.9015
SVMAP *(ours)*	✗	✓	✗	✓	**33.91/0.9529**

Effect of learning the data and regularization terms. To demonstrate the effectiveness of the proposed method that jointly learns the data and regularization terms, we compare with the baseline methods that omit learning either the data term or the regularization term. Specifically, we respectively replace the learnable data term in Eq. (10b) with the commonly used ℓ_2 norm-based data term (*Fix $\mathcal{D}$ & Learn $\mathcal{R}$* for short) or replace the learnable regularization term in Eq. (10b) with a hyper-Laplacian prior (*Learn $\mathcal{D}$ & Fix $\mathcal{R}$* for short). In addition, we also compare with a classical non-learned method that adopts an ℓ_2 data term and a hyper-Laplacian prior (*Fix $\mathcal{D}$&$\mathcal{R}$* for short). We evaluate all baseline methods on the dataset of Martin *et al*. [24] with 1% Gaussian noise. Table 3 shows the quantitative results, where our approach outperforms all baseline methods, increasing the PSNR value by at least 0.30dB. This demonstrates the importance of jointly learning the data and regularization terms and taking advantage of their interplay.

Note that the ℓ_2 data term is theoretically the most suitable one to model the Gaussian noise underlying the dataset. However, Tab. 3 shows that our method still performs better than the baseline method with the ℓ_2 data term and a learned regularization term. This demonstrates that jointly optimizing the data and regularization terms can help them in compensating each other's limitations.

We further evaluate our approach and the baseline methods on the dataset with saturated pixels, see Sec. 4.1. As the ℓ_2 data term cannot model the distribution of saturated pixels well as demonstrated by [6, 10, 27], the methods based

Table 4. *Effectiveness of learning spatially-variant filters.* All methods are evaluated on the dataset of [24] with 1% Gaussian noise (kernel size from 13 × 13 to 35 × 35 pixels).

	SIMAP	SVMAP *(ours)*
[24] with 1% Gaussian noise	31.25/0.8861	**31.89/0.8973**
[24] with 5% Gaussian noise	26.72/0.7302	**27.25/0.7550**

on the ℓ_2 data term cannot effectively handle blurry images with saturated pixels, *cf*. Tab. 3. Hence, learning a proper data term can yield a significant improvement. Table 3, moreover, shows that in this challenging scenario learning *both* the data and regularization terms is essential for high-quality deblurred results ($>$ 3dB difference).

Effect of predicting spatially-variant filters. To demonstrate the effectiveness of our spatially-variant formulation over spatially-invariant data and regularization terms, we compare with a baseline method that learns *spatially-invariant* $f_i, \mathcal{D}_i, g_j$, and $\mathcal{R}_j$ for each i and j (SIMAP for short). Note that when the filters are learned to be spatially invariant, the weights in Eq. (11) cannot be merged into the filters. Thus, for this baseline method, we actually need to learn both the filters $\{f_i, g_j\}$ and the weights $\{\omega_i^d, \omega_j^r\}$. For fair comparison, we use the same optimization method and experimental settings to train and evaluate this baseline. Table 4 shows that learning spatially-variant filters performs notably better than learning spatially-invariant ones, especially when the blurry images contain significant noise.

Visualization of predicted spatially-variant filters. We visualize some predicted filters in Fig. 6. Since the data term measures the goodness-of-fit, the filters f_i predicted for the data term vary depending on the image reconstruction error (in the sense of Eq. (5)). We note that the initial latent image in Fig. 6(b) exhibits significant errors in saturated areas. Comparing this to the predicted filters in Fig. 6(e), we observe that the trained network predicts quite different filters for saturated areas than in non-saturated areas (with smaller reconstruction errors). This improves the quality of the latent image (Fig. 6(c)), but saturated pixels continue to violate the underlying convolutional assumption of the data term (as stated in Sec. 3.1). Thus, even further iterations show different filters being predicted for areas with and without saturation (Fig. 6(f)). Similarly, the predicted filters g_j for the regularization term are based on the latent image and can adapt to the image content. Hence, differing g_j are predicted for different image structures, *e.g*., flat and textured areas in Fig. 6(g) and (h). Thus, both the filters predicted for the data and regularization terms can effectively capture the spatially-variant image characteristics. More analyses are included in the supplemental material.

Closely-related methods. The recent work of [28] simultaneously learns the data and regularization terms using a linear combination of fixed Gaussian RBFs and spatially-invariant filters. In contrast, we learn a spatially-variant MAP model, where the data and regularization terms are modeled with deep neural networks. In addition and following [11, 36], using a linear combination of fixed Gaussian RBFs may not suffice to model the noise distribution or the spatially-variant properties of the image structure and detail. As [28] does not provide training code, for fair comparison, we compare our result with the reported result of [28] on one real example. Figure 5 shows that our approach recovers a clearer result with finer detail than [28], highlighting the effectiveness of our SVMAP approach.

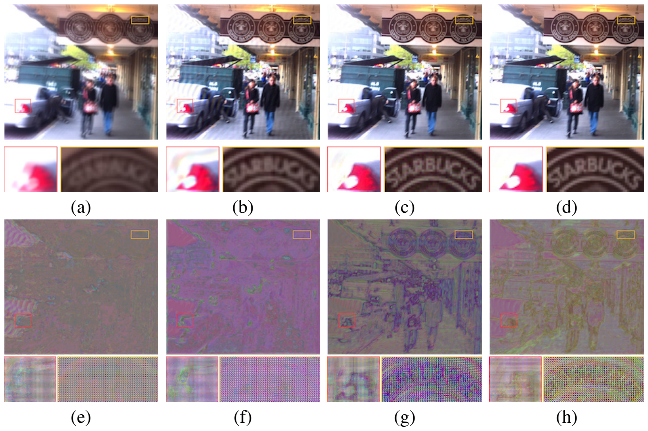


Figure 6. *Visualization of predicted spatially-variant filters. (a)* Blurry input. *(b)–(d)* Results of iterations 0, 1, 2 (*i.e*. our final result), respectively. *(e), (f)* and *(g), (h)* f_i and g_j predicted from *(b), (c)* for iterations 1 and 2. By learning spatially-variant filters for both the data term *(e), (f)* and the regularization term *(g), (h)*, our approach can effectively leverage pixel-dependent properties of image structure and generate a much clearer image with finer-detail in *(d)*. (Best viewed on high-resolution displays.)

6. Conclusion

In this paper, we present an approach for jointly learning spatially-variant data and regularization terms within the MAP framework for non-blind image deblurring. We show that jointly learning both terms is more effective than learning only one term alone; the difference becomes even more striking in challenging scenarios. We further demonstrate that predicting spatially-variant filters instead of the usual spatially-invariant ones better captures the properties of clear images and facilitates finer detail restoration. Taking the MAP-based optimization framework as a constraint for deep neural networks, our proposed model can be trained in an end-to-end manner. Quantitative and qualitative evaluations on benchmark datasets and real-world images demonstrate that our approach achieves substantially better image quality than the current state of the art.

Acknowledgements. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008).

References

[1] Leah Bar, Nahum Kiryati, and Nir Sochen. Image deblurring in the presence of impulsive noise. *IJCV*, 70(3):279–298, 2006. 1, 2, 3

[2] Jonathan T Barron. A general and adaptive robust loss function. In *CVPR*, pages 4331–4339, 2019. 3

[3] Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *IJCV*, 19(1):57–91, 1996. 3

[4] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE TPAMI*, 39(6):1256–1272, 2017. 4

[5] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. *ACM TOG*, 28(5):145, 2009. 7

[6] Sunghyun Cho, Jue Wang, and Seungyong Lee. Handling outliers in non-blind image deconvolution. In *ICCV*, pages 495–502, 2011. 1, 2, 5, 6, 7

[7] Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. BM3D frames and variational image deblurring. *IEEE TIP*, 21(4):1715–1728, 2011. 2

[8] Stephan Dempe. *Bilevel optimization: theory, algorithms and applications*. TU Bergakademie Freiberg, Fakultät für Mathematik und Informatik, 2018. 3

[9] Justin Domke. Generic methods for optimization-based modeling. In *AISTATS*, pages 318–326, 2012. 3

[10] Jiangxin Dong, Jinshan Pan, Zhixun Su, and Ming-Hsuan Yang. Blind image deblurring with outlier handling. In *ICCV*, pages 2478–2486, 2017. 5, 7

[11] Jiangxin Dong, Jinshan Pan, Deqing Sun, Zhixun Su, and Ming-Hsuan Yang. Learning data terms for non-blind deblurring. In *ECCV*, volume 11, pages 748–763, 2018. 1, 2, 3, 5, 6, 7, 8

[12] Dong Gong, Zhen Zhang, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, and Yanning Zhang. Learning deep gradient descent optimization for image deconvolution. *IEEE TNNLS*, 31(12):5468–5482, 2020. 2, 5, 6

[13] Zhe Hu, Sunghyun Cho, Jue Wang, and Ming-Hsuan Yang. Deblurring low-light images with light streaks. In *CVPR*, pages 3382–3389, 2014. 3

[14] Jeremy Jancsary, Sebastian Nowozin, and Carsten Rother. Loss-specific training of non-parametric image restoration models: A new state of the art. In *ECCV*, volume 7, pages 112–125, 2012. 2

[15] Meiguang Jin, Stefan Roth, and Paolo Favaro. Noise-blind image deblurring. In *CVPR*, pages 3510–3518, 2017. 2

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[17] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-Laplacian priors. In *NIPS*, pages 1033–1041, 2009. 1, 2, 3, 4

[18] Jakob Kruse, Carsten Rother, and Uwe Schmidt. Learning to push the limits of efficient FFT-based image deconvolution. In *ICCV*, pages 4586–4594, 2017. 2, 5, 6, 7

[19] Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM J. Imaging Sciences*, 6(2):938–983, 2013. 3

[20] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, pages 1701–1709, 2016. 1

[21] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM TOG*, 26(3):70, 2007. 1, 2, 3, 4

[22] Anat Levin, Yair Weiss, Fredo Durand, and William T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, pages 1964–1971, 2009. 5

[23] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE TIP*, 26(2):1004–1016, 2016. 5

[24] David Martin, Charless 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 416–423, 2001. 5, 6, 7, 8

[25] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *ECCV*, volume 3, pages 783–798, 2014. 2

[26] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, pages 2502–2510, 2018. 4

[27] Jinshan Pan, Zhouchen Lin, Zhixun Su, and Ming-Hsuan Yang. Robust kernel estimation with outliers handling for image deblurring. In *CVPR*, pages 2800–2808, 2016. 5, 7

[28] Dongwei Ren, Wangmeng Zuo, David Zhang, Lei Zhang, and Ming-Hsuan Yang. Simultaneous fidelity and regularization learning for image restoration. *IEEE TPAMI*, 43(1):284–299, 2021. 1, 3, 4, 7, 8

[29] William Hadley Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972. 1

[30] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM J. Imaging Sciences*, 10(4):1804–1844, 2017. 2

[31] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *CVPR*, pages 860–867, 2005. 1, 2

[32] Leonid I. Rudin and Stanley Osher. Total variation based image restoration with free local constraints. In *ICIP*, pages 31–35, 1994. 1, 2

[33] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992. 1, 2

[34] Kegan G.G. Samuel and Marshall F. Tappen. Learning optimized MAP estimates in continuously-valued MRF models. In *CVPR*, pages 477–484, 2009. 2

[35] Uwe Schmidt, Jeremy Jancsary, Sebastian Nowozin, Stefan Roth, and Carsten Rother. Cascades of regression tree fields for image restoration. *IEEE TPAMI*, 38(4):677–689, 2016. 2, 4, 5

[36] Uwe Schmidt and Stefan Roth. Shrinkage fields for effective image restoration. In *CVPR*, pages 2774–2781, 2014. 2, 3, 4, 5, 6, 7, 8

[37] Uwe Schmidt, Kevin Schelten, and Stefan Roth. Bayesian deblurring with integrated noise estimation. In *CVPR*, pages 2625–2632, 2011. 2

[38] Christian J. Schuler, Harold Christopher Burger, Stefan Harmeling, and Bernhard Schölkopf. A machine learning ap-

proach for non-blind image deconvolution. In *CVPR*, pages 1067–1074, 2013. 2, 5, 6, 7

[39] Hyeongseok Son and Seungyong Lee. Fast non-blind deconvolution via regularized residual networks with long/short skip-connections. In *ICCP*, pages 23–32, 2017. 5, 6

[40] Maitreya Suin, Kuldeep Purohit, and A.N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, pages 3606–3615, 2020. 1

[41] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *ICCP*, pages 1–8, 2013. 5, 6, 7

[42] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 1

[43] Marshall F. Tappen, Ce Liu, Edward H. Adelson, and William T. Freeman. Learning Gaussian conditional random fields for low-level vision. In *CVPR*, pages 1–8, 2007. 2

[44] Yilun Wang, Junfeng Yang, Wotao Yin, and Yin Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sciences*, 1(3):248–272, 2008. 1, 2

[45] Oliver Whyte, Josef Sivic, and Andrew Zisserman. Deblurring shaken and partially saturated images. *IJCV*, 110(2):185–201, 2014. 3, 5, 6

[46] Norbert Wiener. Extrapolation, interpolation, and smoothing of stationary time series: With engineering applications. *MIT Press*, 113(21):1043–54, 1949. 1

[47] Li Xu, Jimmy S.J. Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *NIPS*, pages 1790–1798, 2014. 3, 5, 6, 7

[48] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, pages 5978–5986, 2019. 1

[49] Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson W.H. Lau, and Ming-Hsuan Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *CVPR*, pages 3817–3825, 2017. 1, 2, 5, 6, 7

[50] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 1, 2, 5, 6, 7

[51] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H.S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015. 2

[52] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, pages 479–486, 2011. 1, 2, 5, 6, 7