Super-Gaussian Fields: A Novel Approach to Image Deblurring

Anonymous authors
Paper under double-blind review

Abstract

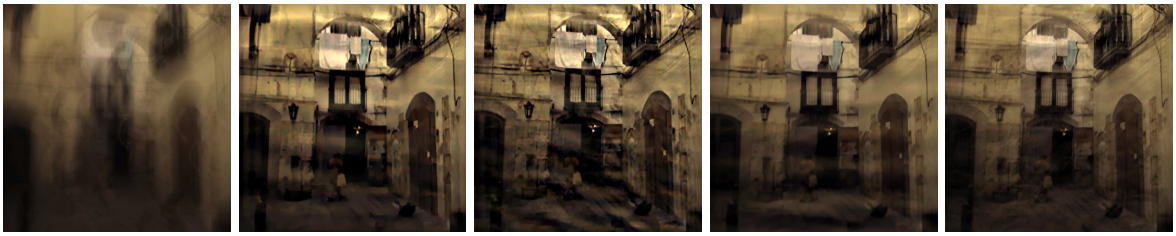
Blind image deblurring is a challenging problem due to its ill-posed nature, of which the success is closely related to a proper image prior. Although most of sparsity-based priors on the gradient filters have been successfully applied, they are inherently limited by the fact that they only explore local coherence in natural image statistics and thus cannot model more complicated structures. We aim to leverage Markov random fields (MRFs) to break the limitation. Due to the intractable partition function, however, traditional MRFs often learn universal filters for various images, resulting in unsatisfactory performance. Motivated by this, we propose a novel MRF-based image prior, referred to as Super-Gaussian Fields. Specifically, we depict potentials by using super-Gaussian distributions, leading to image-specific filters. Relying on the prior and Bayesian MMSE, we proposed an effective image deblurring method. Theory analyses show that the proposed method can effectively avoid local minimum, and can learn image-adaptive sparsity-promoting filters that highlight image structures for kernel estimation. Most importantly, with the theory support, the proposed method can be extended to various scenarios, e.g., face, text, and low-illumination image deblurring. Extensive experiments demonstrate the theoretical advantages and practical effectiveness of the proposed method.

1 Introduction

Blind image deblurring (BID) aims to estimate a sharp image when given a blurred observation. Generally, the imaging model of the blurred image can be formalized as follows:

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

where the blurred image $\mathbf{y}$ is generated by convolving the latent image $\mathbf{x}$ with a blur kernel $\mathbf{k}$, $\otimes$ denotes the convolution operator, and $\mathbf{n}$ denotes the noise. The task of BID includes estimation of $\mathbf{x}$ and the corresponding $\mathbf{k}$. Since this problem is highly ill-posed, to obtain a meaningful solution, appropriate priors on the latent image $\mathbf{x}$ or/and the blur kernel $\mathbf{k}$ are necessary to regularize the solution space.



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

Figure 1: Deblurring results of a challenging example from Köhler et al. (2012). From left to right: Blurred image (PSNR: 22.386), Cho and Lee Cho & Lee (2009) (PSNR: 22.611), Xu and Jia Xu & Jia (2010) (PSNR: 22.782), Pan et al. Pan et al. (2016) (PSNR: 26.259), Ours (PSNR: 27.5427).

One of the most representative statistics characteristics of natural images is the local spatial coherence. With such coherence, natural images have sparse statistics in arbitrary zero-mean gradient domain Schmidt et al. (2010), e.g., $[-1, 1]$. Inspired by this, extensive prevailing methods Fergus et al. (2006); Levin et al. (2011a); Babacan et al. (2009); Perrone & Favaro (2014); Tzikas et al. (2007); Babacan et al. (2012); Krishnan et al. (2011); Xu et al. (2013); Ge et al. (2009) have developed various sparse priors or regularization models to emphasize the sparsity on the latent images in gradient domain for blind image deblurring. A brief review will be included in Section 2. Although these methods have made such remarkable progress, they are restricted to modeling local characteristics. And thus their performances are unsatisfactory and still can be improved. Specifically, 1) the gradient domain only records the response to several basic filters, which are insufficient to capture structures more complex than local coherence, such as the long-range correlation among pixels. In general, those complex structures often benefit from recovering more details in the deblurred results. 2) Most of the existing sparse priors (e.g., Laplace prior) or regularizers (e.g., ℓ_p-norm, $0 \leq p \leq 1$) model each gradient element independently and thus cannot sufficiently capture the complex correlations. Such models often result in some unnatural artifacts, as shown in Fig. 1.

To simultaneously address these two problems, we resort to establishing an appropriate image prior with high order Markov random fields (MRFs) model, which is motivated by the two advantages of MRFs. Firstly, high-order MRFs can be formulated and learned with an ensemble of high-order filters to model the image distribution by considering the complex image structures. Secondly, MRFs integrate the potentials defined on each clique (i.e., centering at each pixel) into a probabilistic joint distribution (without simple independent assumption), which can further help to capture the long-range correlation.

Traditional MRFs models, e.g., Fields of experts (FoE) Roth & Black (2009) and Gaussian scale mixture FoE (GSM-FoE) Schmidt et al. (2010), learn universal filters and the corresponding distributions from clean images, as a general image prior. The learned priors are then used in separate image restoration processes. The learned filters and distributions are assumed to be general for distinguishing clean and degenerated versions of all images. However, the response of filters on images may vary due to the changes of the contents and scales. For example, the learned filters lead to heavy-tailed sparse distributions on the finest scale while Gaussian-like distributions on the coarsest scale. This results in a *distribution mismatching* issue, i.e., applying the same MRFs to the images with different and mismatching distributions. Such mismatching issue makes the performances of the traditional MRFs unsatisfied. In a simple case, the traditional MRFs may not perform well in the commonly used multi-scale (coarse-to-fine) framework in BID. The learned filters only can capture the common statistics among all training images and fail to depict the image-specific characteristics (from scale and content variation) in various images effectively.

To overcome the difficulty above, we propose a novel MRF-based prior, referred to as Super-Gaussian Fields (SGF). With the proposed SGF, we can estimate image-specific adaptive filters and model the image distribution in the filtered space using super-Gaussian distributions. The SGF-based image prior can get rid of the problems in the traditional MRF-based prior with theoretical support. Instead of estimating parameters from only a separate set of clean images as a static image prior, we we predict both image-specific filters and the partition functions for the SGF from each observed image during the kernel estimation and image updating process. The proposed image prior can thus adaptively match different image contents and different image scales in the multi-scale framework. Relying on the proposed SGF-based image prior, we propose a novel image deblurring method with Bayesian Minimum Mean Squared Error (MMSE) estimator. With the estimated image-specific filters and the SG-based sparsity promoting distribution, the proposed SGF-based prior can highlight the representative image characteristics and discriminate the clear and blurry patterns in image deblurring. We conduct the theory analyses to show that the proposed method inherits the advantage of the traditional variational Bayesian method on avoiding troublesome local minima and can learn image-adaptive sparsity-promoting filters that highlight the image structures for kernel estimation. With the theory support, the proposed method can be naturally extended to various scenarios, e.g., face, text, and low-illumination image deblurring.

A preliminary version of this work appeared in Liu et al. (2018). In this work, 1) We provide rigorous theoretical justification for the success of the proposed method on blind image deblurring (Section 5). 2) Built on these theoretical results, we explore the ability of the proposed method to handle various scenarios such as natural, face, text, and low-illumination images (Section 6). 3) Further, we improve the performance

of the proposed method to handle non-blind deblurring and extend the proposed method for non-blind deblurring to handle text images (Section 7). 4) We also extend the theoretical justification for blind image deblurring to non-uniform blind image deblurring (Section 8).

2 Related Work

2.1 Blind Image Deblurring

Due to the pioneering work of Fergus et al. Fergus et al. (2006) that imposes sparsity on image in the gradient spaces, sparse priors have attracted attention Fergus et al. (2006); Levin et al. (2011a); Babacan et al. (2009); Perrone & Favaro (2014); Tzikas et al. (2007); Babacan et al. (2012); Krishnan et al. (2011); Xu et al. (2013); Ge et al. (2009); Gong et al. (2018). For example, a mixture of Gaussian models is early used to chase the sparsity due to its excellent approximate capability Fergus et al. (2006); Levin et al. (2011a). A total variation model is employed since it can encourage gradient sparsity Babacan et al. (2009); Perrone & Favaro (2014). A student-t prior is utilized to impose the sparsity Tzikas et al. (2007). A super-Gaussian model is introduced to represent a general sparse prior Babacan et al. (2012). Those priors are limited by the fact that they are related to the l_1-norm. To relax the limitation, many of the l_p-norm (where $p < 1$) based priors are introduced to impose sparsity on image Krishnan et al. (2011); Xu et al. (2013); Ge et al. (2009). For example, Krishnan et al. Krishnan et al. (2011) propose a normalized sparsity prior (l_1/l_2). Xu et al. Xu et al. (2013) propose a new sparse l_0 approximation. Ge et al. Ge et al. (2009) introduce a spike-and-slab prior that corresponds to the l_0-norm. However, all those priors are limited by the fact that they assume the coefficients in the gradient spaces are mutually independent.

Besides the above-mentioned sparse priors, a family of blind deblurring approaches explicitly exploits the structure of edges to estimate the blur kernel Cho & Lee (2009); Xu & Jia (2010); Cho et al. (2011); Joshi et al. (2008); Sun et al. (2013); Lai et al. (2015); Zhou & Komodakis (2014). Joshi et al. Joshi et al. (2008) and Cho et al. Cho et al. (2011) rely on restoring edges from the blurry image. However, they fail to estimate the blur kernel with large size. To remedy it, Cho and Lee Cho & Lee (2009) alternately recover sharp edges and the blur kernel in a coarse-to-fine fashion. Xu and Jia Xu & Jia (2010) further develop this work. However, these approaches heavily rely on empirical image filters. To avoid it, Sun et al. Sun et al. (2013) explore the edges of natural images using learned patch prior. Lai et al. Lai et al. (2015) predict the edges by learning prior. Zhou and Komodakis Zhou & Komodakis (2014) detect edges using a high-level scene-specific prior. All those priors only explore the local patch in the latent image but neglect the global characters.

Rather than exploiting edges, there are many other priors. Komodakis and Paragios Komodakis & Paragios (2013) explore the quantized version of the sharp image by a discrete MRF prior. Their MRF prior is different from the proposed SG-FoE prior which is a continuous MRF prior. Michaeli and Irani Michaeli & Irani (2014) seek sharp images by the recurrence of small image patches. Gong et al. Gong et al. (2016; 2017a) hire a subset of the image gradients for kernel estimation. Pan et al. Pan et al. (2016), and Yan et al. Yan et al. (2017) explore dark and bright pixels for BID, respectively. More recently, the work in Bai et al. (2019) considers a latent structure prior of the unknown sharp image for image deblurring. Chen et al. Chen et al. (2019) propose a local Maximum Gradient (LMG) prior, which is inspired by the observation that the maximum value of a local patch gradient will diminish after blurring process. The work in Pan et al. (2019) studies the BID in the frequency domain instead of the filtered space, which exploits the phase-only image of the input blurry image. Zhang et al. Zhang et al. (2022b) utilize Bayes posterior estimation to screen through the intermediate image and exclude those unfavorable pixels to reduce their influence on kernel estimation. The work in Chen et al. (2021) introduces a new blur model to fit both saturated and unsaturated pixels for considering the informative pixels during the deblurring process. Compared with the above conventional methods relying on domain knowledge-based regularization and optimization, some recent works explore learning-based regularization methods for image deblurring, which leverages the ability of deep neural networks to learn knowledge from image data Nimisha et al. (2017); Gong et al. (2017b); Kupyn et al. (2019); Su et al. (2017); Wang et al. (2019); Zhou et al. (2019); Lin et al. (2020); Ren et al. (2020); Song et al. (2019); Shen et al. (2018); Tran et al. (2021;?); Ma et al. (2022); Li et al. (2022); Zhang

et al. (2022b). A more detailed survey for learning-based methods for image deblurring can be found in reference Zhang et al. (2022a).

2.2 High-Order MRFs

Since gradient filters only model the statistics of first derivatives in the image structure, high-order MRF generalizes traditional based-gradient pairwise MRF models, e.g., cluster sparsity field Zhang et al. (2018), by defining linear filters on large maximal cliques. Based on the Hammersley-Clifford theorem Besag (1974), high-order MRF can give the general form to model image as follows:

$$p(\mathbf{x};\Theta) = \frac{1}{Z(\Theta)} \prod_{c\in C} \prod_{j=1}^{J} \phi(\mathbf{J}_j \mathbf{x}_c), \tag{2}$$

where C is the set of the maximal cliques, $\mathbf{x}_c$ are the pixels of clique c, $\mathbf{J}_j$ are the linear filters and $j = 1, ..., J$, $Z(\Theta)$ is the partition function with parameters Θ that depend on ϕ and $\mathbf{J}_j$, ϕ are the potentials. In contrast to previous high-order MRF in which the model parameters are hand-defined, FoE Roth & Black (2009), a class of high-order MRF, can learn the model parameters from an external database, and hence has attracted high attention in image denoising Schmidt et al. (2010); Weiss & Freeman (2007), NBID Schmidt et al. (2011) and image super-resolution Zhang et al. (2012).

3 The Proposed Image-Specific SGF Prior

Traditional MRF-based image priors Roth & Black (2009) learn a set of filters and the corresponding partition functions from an external image database, which are used in the processing of various images as a general prior. In the following, we will show that such models may suffer from the *distribution mismatching* issue easily. Specifically, we comprehensively investigate a typical high-order MRF model, Gaussian scale mixture-FoE model (GSM-FoE) Schmidt et al. (2010), and reveal how the generally learned filters hinder the process of image deblurring. Motivated by this, we will propose a novel image-specific MRF-based prior model, named super-Gaussian Fields (SGFs). Different from previous MRFs, the filters of the proposed SGF can be adaptively estimated and updated for each specific image.

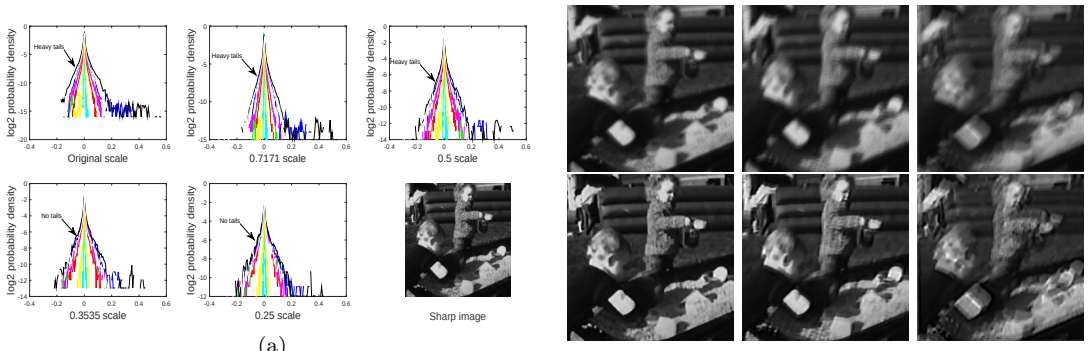
3.1 Mismatching Problem of the Traditional GSM-FoE

According to Schmidt et al. (2010), GSM-FoE follows the general MRFs form in Eq. equation 2 and defines each potential with GSM as follows:

$$\phi(\mathbf{J}_j \mathbf{x}_c; \alpha_{j,k}) = \sum\nolimits_{k=1}^{K} \alpha_{j,k} \mathcal{N}(\mathbf{J}_j \mathbf{x}_c; 0, \eta_j / s_k), \tag{3}$$

where $\mathcal{N}(\mathbf{J}_j \mathbf{x}_c; 0, \eta_j / s_k)$ denotes the Gaussian probability density function with zero mean and variance η_j / s_k. s_k and $\alpha_{j,k}$ denote the scale and weight parameters, respectively. It has been shown that GSM-FoE can well depict wide heavy-tailed distributions Schmidt et al. (2010). Similar to most previous MRFs models, the partition function $Z(\Theta)$ for GSM-FoE is generally intractable since it requires integrating over all possible images. However, evaluating $Z(\Theta)$ is necessitated to learn all the model parameters, e.g., $\{\mathbf{J}_j\}$ and $\{\alpha_{j,k}\}$ (η_j and s_k are generally constant). To sidestep this difficulty, most MRFs models, including GSM-FoE, learn model parameters by maximizing the likelihood in equation 2 on an external image database Roth & Black (2009); Schmidt et al. (2010); Weiss & Freeman (2007), and then apply the learned model in the following applications for different images.

However, directly applying the pre-learned GSM-FoE to BID leads to unsatisfied results. BID commonly adopts a coarse-to-fine framework. The latent images' responses at different scales to these pre-learned filters are different and may mismatch the universal pre-learned GSM distribution. To show this point clearly, we apply the learned filters in GSM-FoE to an example image and show the responses of an image across various scales in Fig. 2a. We can find that the response obtained in the finest scale (e.g., the original resolution) exhibits obvious sparsity and heavy tails, while the response obtained in more coarse scales (e.g., 0.3536



(a) (b)

Figure 2: (a) The 8 distributions with different colors of outputs by applying the 8 learned filters from Schmidt et al. (2010) to the sharp image (the bottom right in (a)) at different scales. The 0.7171, 0.5, 0.3536, and 0.25 denote different downsampling rates. (b) The top: Blurred images with different kernel sizes (Successively, 13×13, 19×19, 27×27). The bottom: Corresponding deblurred images using GSM-FoE.

and 0.25, the down-sampling rates) exhibits a Gaussian-like distribution. Thus the filters (leaned at the finest scale) do not show strong representative ability on other scales. And the Gaussian-like response in coarse-scale cannot be well represented by the GSM-FoE, which prefers to fitting sparse and heavy-tailed distribution. A similar observation is also reported in Schmidt et al. (2010). To further demonstrate the negative effect of such kind of distribution mismatch on BID, we embed the pre-learned GSM-FoE prior into the Bayesian MMSE framework introduced in the following Section 4.1 to deal with an example image blurred with different kernel sizes. The deblurred results are shown in Fig. 2b. Generally, a blurred image with a larger kernel size requires deblurring at a coarser scale. For example, deblurring image with 13×13 kernel requires deblurring at 0.5 scale and obtains a good result, since the filter response exhibits sparsity and heavy tails at 0.5 scale shown as in 2a. However, deblurring image with 19×19 kernel obtains an unsatisfactory result, since it requires deblurring at 0.3536 scale where the filter response mismatches the sparse and heavy-tailed distribution depicted by GSM-FoE shown as 2a. More artifacts are generated in the deblurred results when the kernel size is 27×27 since it requires deblurring at a 0.25 scale, which produces more serious distribution mismatch issues.

We have shown that a universal prior will cause the distribution mismatching problem across different scales above. Maybe a straightforward idea for handling it is to train GSM-FoE models in different scales specifically (GSM-MultiScale). However, as shown in Fig. 8, the performance is still unsatisfactory. Although the GSM-FoE models can be trained with the dataset at different scales separately, they still can only capture the universal statistics among all training images. They cannot sufficiently depict the image-specific character for the candidate blurred image. The multi-scale GSM-FoE prior is difficult to recover the images properly and then impairs the kernel estimation. In contrast, the proposed SGF (See Section 3.2) can be directly learned (or estimated) from the blurred observation in each scale in a Bayesian estimation scheme. Thus the proposed SGF-based prior can capture the image-specific characteristics, leading to better image recovery and kernel estimation.

3.2 Image-Specific SGF Prior

Instead of training GSM-FoE on multiple scales, we turn to a data-adaptive model in which all the model parameters (including filters) can be updated adaptively across different scales. In this work, we propose a novel MRF-based prior model, termed super-Gaussian fields (SGF), which defines each potential in Eq. equation 2 as a super-Gaussian distribution Babacan et al. (2012); Palmer et al. (2005) as follows:

$$\phi(\mathbf{J}_j\mathbf{x}_c) = \max_{\gamma_{j,i} \geq 0} \mathcal{N}(\mathbf{J}_j\mathbf{x}_c; 0, \gamma_{j,i}), \tag{4}$$

where i is the index over image pixels and corresponds to the center of clique c, $\gamma_{j,i}$ denotes the variance. Similar to GSM, SG also can depict sparse and heavy-tailed distributions Palmer et al. (2005). Unlike GSM-

FoE and most MRFs models, the partition function in super-Gaussian fields can be ignored during parameter estimation. More importantly, with such an advantage, it is possible to learn its model parameters directly from the blurred observation on each scale. Thus the proposed super-Gaussian fields can be seamlessly embedded into the coarse-to-fine deburring framework. In the following, we give the theoretical results to show that the partition function can be ignored.

Property 1. *The potential ϕ of SGF is related to $\mathbf{J}_j$ and $\mathbf{x}_c$, but not $\gamma_{j,i}$. Hence, the partition function $Z(\Theta)$ of SGF just depends on the linear filters $\mathbf{J}_j$.*

Proof. As shown in Eq. equation 4, $\gamma_{j,i}$ can be determined by $\mathbf{J}_j$ and $\mathbf{x}_c$. Hence, the potential ϕ in Eq. equation 4 is related to only $\mathbf{J}_j$ and $\mathbf{x}_c$. Furthermore, because $Z(\Theta) = \int \prod_{c\in C} \prod_{j=1}^{J} \phi(\mathbf{J}_j \mathbf{x}_c) \mathbf{dx}$, the partition function $Z(\Theta)$ just depends on the linear filters $\mathbf{J}_j$ once the integral is done. Namely, $\Theta = \{\mathbf{J}_j | j = 1, ..., J\}$. □

Property 1 shows that the parameter γ can be naturally ignored while evaluating the partition function. In the following, we show the condition for ignoring the filers $\mathbf{J}_j$'s. For any filter $\mathbf{J}_j$, we define a $\mathbf{V}_{\mathbf{J}_j}$ to represent its vector version obtained via starching the filter $\mathbf{J}_j$. We have the following property.

Property 2. *Given any set of J orthonormal vector, such as $\{\mathbf{V}_{\mathbf{J}_j}\}$ and $\{\mathbf{V}_{\mathbf{J}'_j}\}$, the value of the partition function of SGF keeps constant, i.e., $Z(\{\mathbf{V}_{\mathbf{J}_j}\}) = Z(\{\mathbf{V}_{\mathbf{J}'_j}\})$.*

Proof. Before proving *Property* 2, we first introduce the following theory:

Theorem 1 (Weiss & Freeman (2007))**.** *For any image $\mathbf{x}$, let $\mathbf{T}_\mathbf{x}$ denote the corresponding Toeplitz convolution matrix, i.e., $\mathbf{J}_j \otimes \mathbf{x} = \mathbf{V}_{\mathbf{J}_j}^T \mathbf{T}_\mathbf{x}$. Let $E(\mathbf{V}_{\mathbf{J}_j}^T \mathbf{T}_\mathbf{x})$ be an arbitrary function of $\mathbf{V}_{\mathbf{J}_j}^T \mathbf{T}_\mathbf{x}$ and define $Z(\mathbf{V}) = \int e^{-\sum_j E(\mathbf{V}_{\mathbf{J}_j}^T \mathbf{T}_\mathbf{x})} \mathbf{dx}$ with $\mathbf{x}$. Then $Z(\mathbf{V}) = Z(\mathbf{V}')$ for any sets of J orthonormal vectors $\{\mathbf{V}_{\mathbf{J}_j}\}$ and $\{\mathbf{V}'_{\mathbf{J}_j}\}$.*

Since the partition function $Z(\Theta)$ of SGF just depends on the linear filters $\{\mathbf{J}_{\mathbf{J}_j}\}$ as mentioned in **Property 1** and the potential ϕ in Eq. equation 4 also perfectly meets the form of $E(\mathbf{V}_{\mathbf{J}_j}^T \mathbf{T}_\mathbf{x})$ in **Theorem** 1, it is easy to proof **Property 2**. □

Based on **Property 1**, we do not need to evaluate the partition function $Z(\Theta)$ of SGF to straightforward update $\gamma_{j,i}$, since $Z(\Theta)$ do not depend on $\gamma_{j,i}$. Further, based on **Property 2**, we do not need to evaluate $Z(\Theta)$ of SGF to update $\mathbf{J}_j$, if we limit updating $\mathbf{J}_j$ in the orthonormal space.

4 Image Deblurring with the Proposed SGF

Based on the proposed SGF model in Eq. equation 2 and Eq. equation 4, we propose an iterative approach for BID in this section. Given a blurred image, the proposed deblurring method attractively updates the intermediate latent image and the blur kernel while the other one is fixed. We will introduce the process for updating image and blur in the following subsections, respectively. Different to many previous methods only focusing on kernel estimation, the proposed image prior and the method can be used for both estimating blur kernel and recovering the final sharp image (i.e., non-blind deblurring), which will be further discussed in Section 7.

4.1 Recovering Latent Image Given the Blur Kernel

Given the blur kernel, a conventional approach to recover latent image is Maximum a Posteriori (MAP) estimation. However, MAP favors the no-blur solution due to the influence of image size Levin et al. (2011b). To overcome it, we introduce Bayesian MMSE to recover the latent image. MMSE can eliminate the influence by integration on image. In the following, to simplify the representation, we slightly abuse the notation and let $\mathbf{J}$ denote the set of the filters $\{\mathbf{J}_j\}$ and γ denote the set $\{\gamma_{j,i}\}$. The MMSE formulation can

be written as follows Murphy (2012):

$$\begin{aligned} &\arg\min_{\tilde{\mathbf{x}}} \int \|\tilde{\mathbf{x}} - \mathbf{x}\|^2 p(\mathbf{x}|\mathbf{y}, \mathbf{k}, \mathbf{J}, \gamma) \mathbf{dx} \\ =& \arg\min_{\tilde{\mathbf{x}}} \ E(\mathbf{x}|\mathbf{y}, \mathbf{k}, \mathbf{J}, \gamma). \end{aligned} \tag{5}$$

This equation is equivalent to the mean of the posterior distribution $p(\mathbf{x}|\mathbf{y}, \mathbf{k}, \mathbf{J}, \gamma)$. However, computing the posterior distribution is generally intractable. Conventional approaches that resort to sum-product belief propagation or sampling algorithms often face high computational costs. To reduce the computational burden, we use a variational posterior distribution $q(\mathbf{x})$ to approximate the true posterior distribution $p(\mathbf{x}|\mathbf{y}, \mathbf{k}, \mathbf{J}, \gamma)$. The variational posterior $q(\mathbf{x})$ can be found by minimizing the Kullback-Leibler divergence $KL(q(\mathbf{x})||p(\mathbf{x}|\mathbf{y}, \mathbf{k}, \mathbf{J}, \gamma))$. This optimization is equivalent to the maximization of the lower bound of the free energy:

$$\begin{aligned} \max_{q(\mathbf{x}), \mathbf{J}, \gamma, \delta^2} &\int q(\mathbf{x}) \log p(\mathbf{x}, \mathbf{y}|\mathbf{k}, \mathbf{J}, \gamma) \mathbf{dx} \\ &- \int q(\mathbf{x}) \log q(\mathbf{x}) \mathbf{dx}, \end{aligned} \tag{6}$$

where δ^2 is the variance of the noise term in Eq. equation 1.

In Eq. equation 6, $p(\mathbf{x}, \mathbf{y}|\mathbf{k}, \mathbf{J}, \gamma)$ should be equivalent to $p(\mathbf{y}|\mathbf{x}, \mathbf{k}, \mathbf{J}, \gamma)p(\mathbf{x})$. We empirically introduce a weight parameter λ to regularize the influences of prior and likelihood similar to Roth & Black (2009); Schmidt et al. (2010). In this case, $p(\mathbf{x}, \mathbf{y}|\mathbf{k}, \mathbf{J}, \gamma) = p(\mathbf{y}|\mathbf{x}, \mathbf{k}, \mathbf{J}, \gamma)p(\mathbf{x})^\lambda$. Without loss of generality, we assume that the noise in Eq. equation 1 obeys i.i.d. Gaussian distribution with zero mean and δ^2 variance. To solve the problem in Eq. equation 6, we iteratively estimate $q(\mathbf{x})$, $\{\mathbf{J}_j\}$ and $\{\gamma_{j,i}\}$ and δ^2. The details are in the following.

Estimating $q(\mathbf{x})$: By setting the partial differential of Eq. equation 6 with respect to $q(\mathbf{x})$ to zero and omitting the details of derivation, we obtain:

$$-\log q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{7}$$

with $\mathbf{A} = \delta^{-2}\mathbf{T}_\mathbf{k}^T\mathbf{T}_\mathbf{k} + \sum_j \lambda \mathbf{T}_{\mathbf{J}_j}^T \mathbf{W}_j \mathbf{T}_{\mathbf{J}_j}$, $\mathbf{b} = \delta^{-2}\mathbf{T}_\mathbf{k}^T\mathbf{y}$, where the image $\mathbf{x}$ is vectored, $\mathbf{W}_j$ denotes the diagonal matrices with $\mathbf{W}_j(i,i) = \gamma_{j,i}^{-1}$, $\mathbf{T}_\mathbf{k}$ and $\mathbf{T}_{\mathbf{J}_j}$ denote the Toeplitz (convolution) matrix with the filter $\mathbf{k}$ and $\mathbf{J}_j$, respectively. Similar to Babacan et al. (2012); Levin et al. (2011a), to reduce computational burden, the mean of $q(\mathbf{x})$, i.e., $\langle\mathbf{x}\rangle$, can be found by solving the linear system $\mathbf{A}\langle\mathbf{x}\rangle = \mathbf{b}$. The covariance of $q(\mathbf{x})$, $\mathbf{A}^{-1}$, can be approximated by inverting only the diagonals of $\mathbf{A}$. $\mathbf{A}^{-1}$ will be used in Eq. equation 9 for estimating the variance γ.

Estimating $\mathbf{J}_j$: $\mathbf{J}_j$'s are related to the intractable partition function $Z(\Theta)$. According to Property 2, we can restrict the estimation of $\mathbf{J}_j$ in the orthonormal space, for which $Z(\Theta)$ is constant. That is, we can define a set $\{\mathbf{B}_j\}$, and then consider all possible rotations of a single basis set of filters $\mathbf{B}_j$ as the solution space of $\mathbf{J}_j$. To this end, we denote by $\mathbf{B}$ a matrix whose j-th column is $\mathbf{B}_j$, and denote by $\mathbf{R}$ an orthogonal matrix. In this case, we can ensure $Z(\mathbf{B}) = Z(\mathbf{RB})$, and give the solution of updating $\mathbf{J}_j$ by maximizing Eq. equation 6 under the condition that $\mathbf{R}$ is any orthogonal matrix as follows (More details can refer to Weiss & Freeman (2007)):

$$\begin{aligned} &\mathbf{V}_{\mathbf{J}_j} = \mathbf{B}\mathbf{R}_j \\ &\mathbf{R}_j = \text{eig}\min(\mathbf{B}^T\langle\mathbf{T}_\mathbf{x}\mathbf{W}_j\mathbf{T}_\mathbf{x}^T\rangle\mathbf{B}), \end{aligned} \tag{8}$$

where the operator $\text{eig}\min(\cdot)$ denotes the eigenvector of $\cdot$ with minimal eigenvalue, $\mathbf{T}_\mathbf{x}$ denotes the Toeplitz (convolution) matrix with $\mathbf{x}$. We require that $\mathbf{R}_j$ be orthogonal to the previous columns $\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_{j-1}$.

Estimating $\gamma_{j,i}$: Updating $\gamma_{j,i}$ is more straightforward (then updating $\mathbf{J}_j$) since $Z(\Theta)$ is naturally not related to $\gamma_{j,i}$ as mentioned in Property 1. We can get the solution of updating $\gamma_{j,i}$ by setting the partial differential of Eq. equation 6 with respect to $\gamma_{j,i}$ to zero, as follows:

$$\gamma_{j,i} = \langle (\mathbf{J}_j \mathbf{x}_c)^2 \rangle. \tag{9}$$

In summary, given the blur kernel, the proposed method recovers the intermediate latent image $\mathbf{x}$ by iteratively updating Eq. equation 7-equation 9.

Estimating δ^2: Unlike most traditional methods in which noise level is given manually and fixed, the proposed method can adaptively estimate the noise level δ^2 by solving the Bayesian variational inference problem in Eq. equation 6. By setting the partial differential of Eq. equation 6 with respect to δ^2 to zero, δ^2 can be updated via $\delta^2 = \frac{\langle \|(\mathbf{y}-\mathbf{k}\otimes\mathbf{x})\|^2 \rangle}{n}$, where $\langle\rangle$ denotes the expectation calculator. However, directly updating it this way leads to problematic optimization diverges with the estimated noise level decreasing too much, which was also discussed in Levin et al. (2009). This problem is due to that the size of the sharp image $\mathbf{x}$ is larger than that of the blurred image $\mathbf{y}$, as mentioned in Wipf & Zhang (2014), and can be remedied by updating δ^2 with as the follows with an additional hyper-parameter d:

$$\delta^2 = \frac{\langle \|(\mathbf{y} - \mathbf{k} \otimes \mathbf{x})\|^2 \rangle}{n} + d, \tag{10}$$

where d acts as an interpretable barrier preventing δ^2 from ever going below d, n is the size of image, and $\langle \|(\mathbf{y} - \mathbf{k} \otimes \mathbf{x})\|^2 \rangle$ denote the expectation of $\|(\mathbf{y} - \mathbf{k} \otimes \mathbf{x})\|^2$.

4.2 Recovering the Blur Kernel with the Latent Image

Similar to existing approaches Levin et al. (2011a); Xu et al. (2013); Sun et al. (2013), given $\langle \mathbf{x} \rangle$, we obtain the blur kernel estimation by solving:

$$\min_{\mathbf{k}} \|\nabla \mathbf{x} \otimes \mathbf{k} - \nabla \mathbf{y}\|_2^2 + \beta \|\mathbf{k}\|_2^2, \tag{11}$$

where $\nabla \mathbf{x}$ and $\nabla \mathbf{y}$ denote the latent image $\langle \mathbf{x} \rangle$ and the blurred image $\mathbf{y}$ in the gradient spaces, respectively. To speed up computation, FFT is used as derived in Cho & Lee (2009). After obtaining $\mathbf{k}$, we set the negative elements of $\mathbf{k}$ to 0, and normalize $\mathbf{k}$. The proposed approach is implemented in a coarse-to-fine manner similar to state-of-the-art methods. Algorithm 1 shows the pseudo-code of the proposed approach.

Algorithm 1 Image deblurring with the proposed SGF

Input: Blurred image $\mathbf{y}$
Output: The blur kernel $\mathbf{k}$
  Initialize: $\mathbf{k}, \mathbf{x}, \mathbf{J}, \mathbf{B}, \delta^2, \gamma$, λ, β and d
  while stopping criterion is not satisfied **do**
    Estimate $\mathbf{x}$, $\mathbf{J}$, γ and δ^2 by Eq. equation 7-Eq. equation 10
    Update for blur kernel $\mathbf{k}$ by Eq. equation 11
  end while

4.3 Initialization of the Filters

As introduced in Sec. 4.1, the proposed method estimates image-adaptive filters to capture the image characteristics. In practice, we need to initialize the filters $\mathbf{J}_j$'s and $\mathbf{B}$ (i.e., the set of $\mathbf{B}_j$) in Eq. equation 8 at the start of the optimization. Although the proposed method works well with random initialization, we observed that it can make the optimization process more effective and efficient by learning the initialization of the filters with additional training on *only sharp images* of the corresponding scenario. By default, we use the clean natural images from Martin et al. (2002) to learn the initialization. Given a set of clean and sharp images, we first downsample the images with 0.5 scale (for reducing noise) to obtain training images

and then train the model in Eq. equation 2-equation 3 to obtain the filters, by using the auxiliary-variable Gibbs sampler and contrastive divergence Hinton (2002). The more detailed training process can be found in Schmidt et al. (2010). The obtained 8 3×3 filters $\mathbf{J}_j$'s are used as the initialization for the proposed method. To initialize basis set $\mathbf{B}$, we use the shifted versions of the whitening filter whose power spectrum equals the mean power spectrum of $\mathbf{J}_j$. More details of the process can be found in Weiss & Freeman (2007). The initialization can be adjusted for the extension to some specific scenarios, as shown in Sec. 6.

5 Theoretical Analysis

In this section, we demonstrate the theoretical advantages of the proposed method for blind image deblurring. To this end, we first formulate the updating process of the latent image, e.g., equations Eq. equation 7-Eq. equation 10, as the sum of a fitting term and a penalty function term. Then, we give detailed analyses about this penalty function term, to show the theoretical advantages of the proposed method compared with traditional MAP with sparse penalty term, e.g., ℓ_1-norm, as well as traditional variational Bayesian (VB) methods with sparse penalty term, e.g., Levin et al. (2011a); Babacan et al. (2012); Wipf & Zhang (2014).

5.1 Penalty Function in the Proposed Method

As shown above, the proposed method recovers the intermediate latent image $\mathbf{x}$ by iteratively updating Eq. equation 7-Eq. equation 10. This updating process can be reformulated as an objective function, including a fitting term and a penalty function term:

Theorem 2. *Consider the objective function*

$$\mathcal{L}(\mathbf{J}, \mathbf{x}, \delta^2) = \frac{1}{\delta^2}||\mathbf{y} - \mathbf{k} \otimes \mathbf{x}||_2^2 + \sum_c g(\mathbf{J}, \mathbf{x_c}, \delta^2), \tag{12}$$

where

$$\begin{aligned} g(\mathbf{J}, \mathbf{x}_c, \delta^2) = \min_{\gamma_{j,i} \geq 0} \lambda \sum_j (\frac{(\mathbf{J}_j \mathbf{x}_c)^2}{\gamma_{j,i}} + \log \gamma_{j,i}) \\ + \log(\lambda \delta^2 \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + ||\mathbf{k}||_2^2) \end{aligned} \tag{13}$$

subject to any two vectors in $\{\mathbf{V}_{\mathbf{J}_j}\}$ are orthonormal. The updating process Eq. equation 7-Eq. equation 10 for recovering the latent image with given the blur kernel is equivalent to coordinate descent minimization of Eq. equation 12 over $\mathbf{x}, \mathbf{J}, \gamma, \delta^2$.

Proof is left in Section 2 in supplemental materials. Here $\gamma_{j,c}$ denote the elements corresponding to $\mathbf{x}_c$, $\mathbf{V}_{\gamma_{j,c}^{-1}}$ denote the vectored version of $\gamma_{j,c}^{-1}$. Given **Theorem** 2, we can analyse the image penalty $g(\mathbf{J}, \mathbf{x}_c, \delta^2)$, which is quite different from traditional image regularizers, e.g., ℓ_1 norm, and discuss some of its relevant properties as below.

Theorem 3. *With any given filters $\mathbf{J}$, $g(\mathbf{J}, \mathbf{x}_c, \delta^2)$ is a concave non-decreasing function of $|\mathbf{J}_j \mathbf{x}_c|$.*

Proof can be found in Section 3 in supplemental materials. According to Chen et al. (2017), introducing such a concave non-decreasing penalty function $g(\mathbf{J}, \mathbf{x}_c, \delta^2)$ into the objective is beneficial to promote the sparsity of the solution. Thus, **Theorem** 3 explicitly stipulates that a strong, sparsity promoting $\mathbf{J}_j \mathbf{x}_c$ penalty is produced by the proposed method.

5.2 Comparison with MAP Methods

Although **Theorem** 3 can give theoretical justification to obtain sparse estimation for the latent image in the filter domains, it cannot reveal the theoretical advantages compared with traditional MAP method with sparse regularization, e.g., ℓ_1 norm. We will give deeper discussions on the penalth in the following.

For the sake of brevity, note that because $g(\mathbf{J}, \mathbf{x}_c, \delta^2)$ is a symmetric function with respect to $\mathbf{J}_j\mathbf{x}_c$, we can only examine its concavity/curvature properties in the positive domain, i.e., $\mathbf{J}_j\mathbf{x}_c \geq 0$.

Theorem 4. *Given any set of filters* $\mathbf{J}$*, there are two results as in the following.*

1) For all δ_1^2 *and* δ_2^2*,* $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c, \delta_1^2) \to 0$ *as* $\mathbf{J}\mathbf{x}_c \to \infty$*. Therefore,* $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2)$ *and* $g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$ *penalize large magnitudes of* $\mathbf{J}\mathbf{x}_c$ *equally.*

2) Let $\delta_2^2 \geq \delta_1^2$*, then if* $\mathbf{J}\mathbf{x}_c' \succ \mathbf{J}\mathbf{x}_c$*, we have* $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c, \delta_1^2) \geq g(\mathbf{J}, \mathbf{x}_c', \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c', \delta_1^2)$*. Therefore, as* $\mathbf{J}\mathbf{x_c} \to 0$*,* $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$ *is maximized, implying that* $g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$ *favors zero-valued coefficients more heavily than* $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2)$*.*

The proof follows from several extensions of the proof for Theorem 2 in Zhang et al. (2014). **Theorem** 4 implies that δ^2 represents a form of shape parameter that modulates the sparsity favor ability of the penalty $g(\mathbf{J}, \mathbf{x}, \delta^2)$, which is the essential advantage compared with traditional MAP methods.

At the beginning of the proposed method, δ^2 is initialized as a large value, regardless of the true noise level. Such a large value means that we have a penalty $g(\mathbf{J}, \mathbf{x}, \delta^2)$ that highly behaves like a convex (less sparse) function with respect to $\mathbf{J}_j\mathbf{x}_c$. As a result, one can effectively avoid local minima to a certain extent, and get a desirable optimization solution. As the iterations proceed, δ^2 is reduced, and the penalty function is made less convex (more sparse). In this case, the risk of local minima tends to be more serious. However, the risk can be tremendously ameliorated since we are likely to be already in the neighborhood of a good solution. In contrast, traditional MAP methods' penalty function has no such shape-modulated ability to reduce the risk.

Such similar shape-modulated ability has been reported in traditional VB methods Babacan et al. (2012); Wipf & Zhang (2014). Namely, **Theorem** 4 fails to explain why the proposed method is superior to traditional VB methods. In the next section, we will further discuss the theoretical advantage of the proposed method.

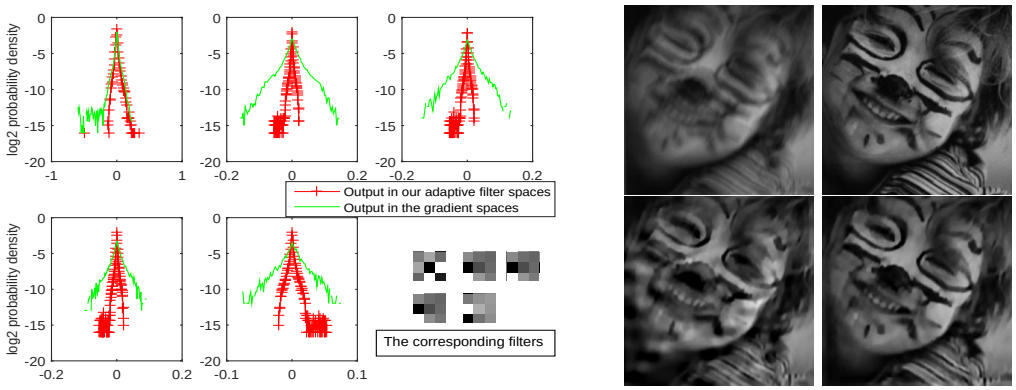
5.3 Comparison with Other VB Methods

In traditional variational Bayesian methods, they often employ basic filters (e.g., gradient filters [1,-1]) to promote kernel estimation. Unlike them, the proposed method can adaptively learn filters, highlighting the theoretical advantage of the proposed method.

Theorem 5. *The learned filters in the proposed SGF are sparse-promoting.*

Proof. As mentioned in Eq. equation 8, the filters in SGF are estimated as the eigenvector of $\langle \mathbf{T}_\mathbf{x}\mathbf{W}_j\mathbf{T}_\mathbf{x}^T \rangle$ with minimal eigenvalue, viz., the filters are the singular vector $\langle \mathbf{T}_\mathbf{x}(\mathbf{W}_j)^{\frac{1}{2}} \rangle$ with minimal singular value. This implies that the proposed method seeks filters $\mathbf{J}_j$'s which lead to the corresponding $\mathbf{V}_{\mathbf{J}_j}^T\mathbf{T}_\mathbf{x}(\mathbf{W}_j)^{\frac{1}{2}}$ being as sparse as possible where $\mathbf{V}_{\mathbf{J}_j}$ denotes the vectorized $\mathbf{J}_j$. Since $(\mathbf{W}_j)^{\frac{1}{2}}$ is a diagonal matrix which only scales each column of $\mathbf{T}_\mathbf{x}$, the sparsity of $\mathbf{V}_{\mathbf{J}_j}^T\mathbf{T}_\mathbf{x}(\mathbf{W}_j)^{\frac{1}{2}}$ is mainly determined by $\mathbf{V}_{\mathbf{J}_j}^T\mathbf{T}_\mathbf{x}$. Consequently, the proposed approach seeks filters $\mathbf{J}_j$ which lead to the corresponding response $\mathbf{V}_{\mathbf{J}_j}\mathbf{T}_\mathbf{x}$ of the latent image being as sparse as possible. □

The results in **Theorem** 5 can be further illustrated by the visual results in Fig. 3(a) where the distribution of image response to these learned filters and gradient filters (e.g., [1,-1]) are plotted. It can be seen that those learned filters lead to a sparser response than that on gradients. To further show the advantage of sparse-promoting filters, we recover the latent image with the proposed method and a typical VB based method in Babacan et al. (2012). The corresponding deblurred results are shown in Fig. 3(b). We can find that these learned filters lead to more clear and sharp results. These results demonstrate that the proposed method with sparse-promoting filters is more powerful than the traditional VB method with the basic filters for image deblurring.



(a) Output (b) Results

Figure 3: Comparison of the outputs in the gradient spaces and our adaptive filter spaces at different scales. (a) The distributions of the filter outputs of sharp image (The top right in (b)) in gradient spaces and our adaptive spaces (by using filters corresponding to the bottom right in (a)) at different scales. From top to bottom and from left to right: the original, 0.7171 (sampling rate), 0.5, 0.3536, 0.25 scales. The bottom right is our final obtained filters corresponding to the different scales. (b) From left to right: blurred image, sharp image, result by the method in Babacan et al. (2012) and result by the proposed method.

6 Extension to Various Specific Scenarios

In this section, we explore the ability of the proposed method to handle various scenarios, such as face, text, and low-illumination images. The images in the special scenarios can have specific characteristics. Some deblurring methods are specifically designed to handle different particular scenarios. Benefiting from the proposed image-adaptive SGF prior, the proposed method can be naturally and conveniently generalized to different scenarios with a universal framework and minor adjustments, as discussed in the following. For example, to effectively handle the images with very special characteristics, we can obtain the initialization of the filters (with details in Sec. 4.3) using the clean images from the corresponding scenarios, e.g., text images in Yao et al. (2012) and low-light images from Loh & Chan (2019), as introduced in the following.

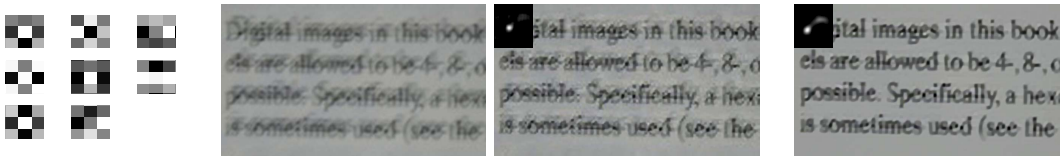
Face images: Deblurring face images is a challenging task because few strong edges in blurry face images can be easily extracted for kernel estimation. Existing methods use explicit edge detection processes for kernel estimation, which are restricted to pre-defined low-order filter space and not general for all images.The proposed method can be directly applied to face images without adjustment. Although the default initialization of the filters $\mathbf{J}_j$ and set $\mathbf{B}$ are designed for natural images, the proposed method can perform well as shown in Fig. 4, since it uses the long-range filters (3×3) to capture the data structure more prominently than the traditional gradient spaces (i.e., [1, -1]). And the proposed method can further highlight the useful structures by updating the image-specific prior distribution attentively, as shown in **Theorem** 5 as suggested in Weiss & Freeman (2007).

Text images: Previous priors for natural images deblurring, which unitize sparse gradient statistics of natural images, are less effective for cases with text contents. This is because text images often are composed of high-order structures in a larger spatial range, which makes sparse priors in gradient space to be inaccurate Xiao et al. (2016) as shown in Fig. 5c. A feasible method to handle this problem is exploring sparsity in high-order filter spaces Xiao et al. (2016) capturing the statistical characteristics in a larger range. The proposed SGF naturally integrates the high-order filters as defined in Eq. equation 2 and Eq. equation 4, and explores sparsity on images in these high-order filter spaces as mentioned in **Theorem** 3. Differing from the method in Xiao et al. (2016), the proposed method can learn image-specific filters, while the method in Xiao et al. (2016) uses pre-trained filters for all blurred images. As a result, the proposed method can be easily extended to other domain-specific images, e.g., natural images and face images. In addition, compared with the method in Xiao et al. (2016), the proposed method is supported by the advantages in theory. To enable the proposed method to deblur text images, we initialize the filters by learning from external MSRA

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

Figure 4: Deblurring a challenging real face example. (a) The initialization of the 8 filters $\mathbf{J}_j$, (b) Blurred face image, (c) Xu et al. Xu et al. (2013) and (d) Ours. (e) The estimated sets of filters at different scales, from left to right and from top to bottom. The bottom right is our final obtained filters corresponding to the original scale.

text images Detection 500 dataset Yao et al. (2012) which enables the initialized filters to learn the high-order structures in text images. Although those initialized filters are then updated by Eq. equation 8 as the proposed method proceeds, the initialized basis set $\mathbf{B}$ in Eq. equation 8 have recorded the high-order statistics characteristic of those initialized filters by power spectrum as mentioned above, and regularize the update of the filters as shown in Eq. equation 8. As the proposed method proceeds, the updated filters are still effective in capturing the high-order structure in text images. As a result, the proposed method performs well for deblurring text images, as shown in Fig. 5d.



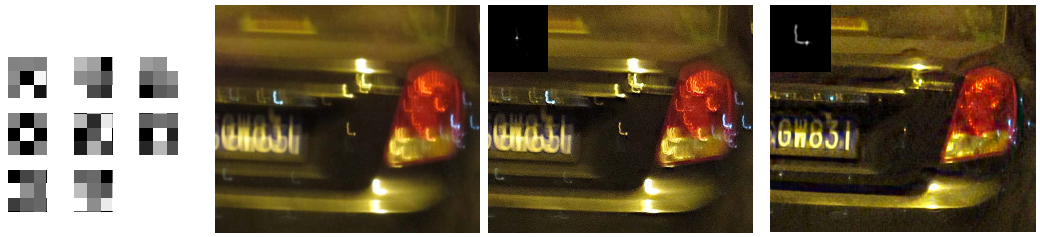
(a) The initialization of $\mathbf{J}_j$ (b) Blurred image (c) Xu et al. Xu et al. (2013) (d) Ours

Figure 5: Deblurring a challenging real text example. (a) shows the filter initialization obtained for text images.

Low-illumination images: Deblurring images with low-illumination scene images is also a challenging task, for which saturated pixels often appear and interfere with kernel estimation. These saturated pixels often enable most traditional deblurring methods, which explore the sparsity of images in the gradient filter spaces, tending to obtain a delta kernel estimation, as shown in 6c. Pan et al. have shown that enforcing sparsity on light streaks in low-illumination images can be effective to deblur these images Jinshan Pan & Yang (2014). The proposed method naturally integrates detecting light streaks (by sparsity-promoting filters mentioned in **Theorem** 5) and enforcing sparsity (**Theorem** 3) into a unified model. As a result, the proposed method can be able to promote kernel estimation. To verify this point, the proposed method learns the initialized filters from external low-illumination images dataset Loh & Chan (2019). This is a very large low-illumination image dataset. In our implementation, we randomly extracted 360 images for training. Consequently, the proposed method performs well on low-illumination images, as shown in Fig. 6d.

7 Non-blind Image Deblurring with SGF

Most of the deblurring methods apply different image priors or regularizers in kernel estimation and non-blind deblurring, for capturing different image characteristics. For example, regularizations or priors for promoting more significant image structures are used for kernel estimation Xu & Jia (2010); Gong et al. (2016), and non-blind deblurring requires the priors to capture more image details in natural images, given a fixed blur kernel. The proposed SGF-based image prior can be feasible for both kernel estimation and



(a) The initialized $\mathbf{J}_j$ (b) Blurred image (c) Xu et al. Xu et al. (2013) (d) Ours

Figure 6: Deblurring a real low-illumination example. (a) shows the filter initialization obtained for low-illumination images.

non-blind deblurring with a unified form. In this section, we discuss the details of non-blind deblurring with the proposed SGF-based image prior and also show the extension on text image non-blind deblurring.

In the preliminary work Liu et al. (2018), we directly used the proposed SGF-based prior for non-blind image deblurring by iteratively computing Eq. equation 7-equation 9 with the given kernel $\mathbf{k}$, which is similar as the operation for updating $\mathbf{x}$ during kernel estimation. The only difference is that we need more iteration times to update Eq. equation 7-equation 9 for non-blind image deblurring than for blind deblurring. The proposed SGF-based prior can work very well and produce high-quality images, as shown in Fig. 7. We observe that such straightforward implementation may lead to the result images lacking some fine details (See Fig. 7c). In the following, we further discuss the reason and the new solutions to improve the non-blind deconvolution quality in the proposed method.

The filters in SGF are initialized as a set of filters pre-leaned from natural images in Schmidt et al. (2010) and are then updated via Eq. equation 8. As statements in **Theorem** 5, the updated filters can promote the sparsity strongly so that the proposed method adaptively learns a set of filters $\{\mathbf{J}_j\}$, which ensures the filtered responses $\mathbf{J_j x_c}$ sparse as much as possible. Further, the proposed method promotes the sparsity of the filtered responses $\mathbf{J_j x_c}$ by the concave non-decreasing regularization term $g(\mathbf{J}, \mathbf{x}_c, \delta^2)$ as shown in **Theorem** 3, regardless of which $\mathbf{J}_j$. Under these supports in theory, the proposed method would obtain highly sparse solution $\mathbf{J}_j\mathbf{x_c}$ by default, which strongly differentiates the shape images and the blurred images. It can be very helpful for kernel estimation. Although the images converge towards the natural images during the optimization, in practice, some fine details in the images may be overly suppressed when the proposed method discriminates the blurry and non-blur patterns too strongly, as depicted by Fig. 7b.

We propose two ways to handle the above problem in non-blind deblurring. 1) Recall that promoting the sparsity of $\mathbf{J}_j\mathbf{x}$ in the proposed method for enhancing the discrimination of the sharp image characteristics. To keep more representative details while promoting sparsity, we can apply more filters with a larger size in non-blind deblurring to capture the more intricate details in the natural images, despite more computations. To verify this point, we train 24 5×5 filters as the initialization and keep updating the filters during optimization, which obtains satisfactory results (See Fig. 7d). 2) Another simple way to relieve too strong sparsity promotion is that we may update the filters less in optimization. For example, in the optimization process of non-blind deblurring, we can directly use the filters learned from natural images and keep them fixed while optimizing for $\mathbf{x}$. We found that these initialization filters can capture natural image statistics when a well-estimated blur kernel is given (only in non-blind deconvolution). And the proposed method can obtain better results with the fixed filters (See Fig. 7b).

Similar to Section 6, the proposed method can also be naturally extended to handle text images in non-blind deblurring. The proposed method can obtain the desired results with the strong sparsity-promoting filter updating process for text images with fewer details. This can be verified by experimental results in Fig. 16.

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

Figure 7: Deblurring results of a challenging example from dataset Levin et al. (2009). From left to right: Blurred image (SSD:574.38), Ours with fixed filters (SSD:35.64), Ours with updating filters (SSD:44.61, aka Liu et al. (2018)), and Ours with 24 5×5 filters (SSD:33.44). The previous method in Liu et al. (2018) neglects many fine details, while the proposed method with 24 5×5 filters obtains better results.

8 Non-uniform Blind Deblurring

In this section, we will show how to use the proposed method to handle non-uniform deblurring. We also demonstrate that the theoretical advantages of the proposed method in uniform blind deblurring can be naturally extended to non-uniform blind deblurring, including avoiding troublesome local minima and image-adaptive sparsity-promoting filters.

The proposed method can also be directly extendted to handle the non-uniform blind deblurring where the blur kernel varies across spatial domain Whyte et al. (2012); Hirsch et al. (2011). According to Whyte et al. (2012), we formulate non-uniform blind deblurring problem as follows:

$$\mathbf{V_y} = \mathbf{D}\mathbf{V_x} + \mathbf{V_n}, \quad \textbf{or} \quad \mathbf{V_y} = \mathbf{E}\mathbf{V_k} + \mathbf{V_n}, \tag{14}$$

where $\mathbf{V_y}$, $\mathbf{V_x}$ and $\mathbf{V_n}$ denote the vectored forms of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ in Eq. equation 1. $\mathbf{D}$ is a large sparse matrix, where each row contains a local blur filter acting on $\mathbf{V_x}$ to generate a blurry pixel and each column of $\mathbf{E}$ contains a projectively transformed copy of the sharp image when $\mathbf{V_x}$ is known. $\mathbf{V_k}$ is the weight vector which satisfies $\mathbf{V_k}_t \geq 0$ and $\sum_t \mathbf{V_k}_t = 1$. Based on Eq. equation 14, the proposed approach can handle the non-uniform blind deblurring problem by alternatively solving the following problems:

$$\begin{aligned} \max_{q(\mathbf{V_x}),\mathbf{J},\gamma,\delta^2} & \int q(\mathbf{V_x}) \log q(\mathbf{V_x}) \mathbf{d}\mathbf{V_x} \\ & - \int q(\mathbf{V_x}) \log p(\mathbf{V_x}, \mathbf{V_y} | \mathbf{D}, \mathbf{J}, \gamma) \mathbf{d}\mathbf{V_x}, \end{aligned} \tag{15}$$

$$\min_{\mathbf{V_k}} \|\nabla \mathbf{E}\mathbf{V_k} - \mathbf{V}_{\nabla \mathbf{y}}\|_2^2 + \beta \|\mathbf{V_k}\|_1. \tag{16}$$

Here, Eq. equation 16 employs l_1-norm to encourage a sparse kernel as Whyte et al. (2012). The optimal $q(\mathbf{V_x})$ in Eq. equation 15 can be computed by using formulas similar to Eq. equation 7-Eq. equation 9 in which $\mathbf{k}$ is replaced by $\mathbf{D}$. In addition, the efficient filter flow Hirsch et al. (2010) is adopted to accelerate the implementation of the proposed approach.

In the preliminary work Liu et al. (2018), we implement the SGF-based non-uniform blind deblurring method with the fixed noise parameter δ^2. In this work, we update δ^2 during the optimization process. The following theoretical analysis for the proposed method on non-uniform blind deblurring shows the significance of updating the noise δ^2.

The objective function Eq. equation 15 is similar to the objective function Eq. equation 6. Their difference is in that the Eq. equation 15 uses the non-uniform kernel formulated with $\mathbf{D}$. In this case, optimizing for Eq. equation 15 can be achieved by replacing $\mathbf{k}$ in Eq. equation 7-Eq. equation 10 by $\mathbf{D}$. As a result, analogying with the inferring processing for **Theorems** 2-5, it can be easily show that: 1) Eq. equation 15 explicitly shows a strong, sparsity promoting penalty on $\mathbf{J}_j \mathbf{V}_{\mathbf{x}_c}$ for non-uniform blind deblurring. 2) Updating noise in Eq. equation 15 has a shape-modulated ability to avoid optimization risk, which often exists in

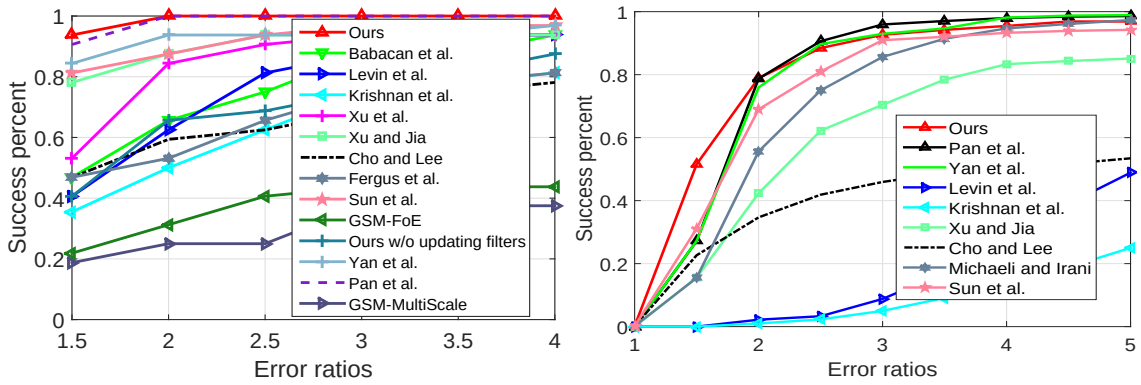


Figure 8: Evaluations on Levin et al. (left) datasets and Sun et al. (right) datasets.

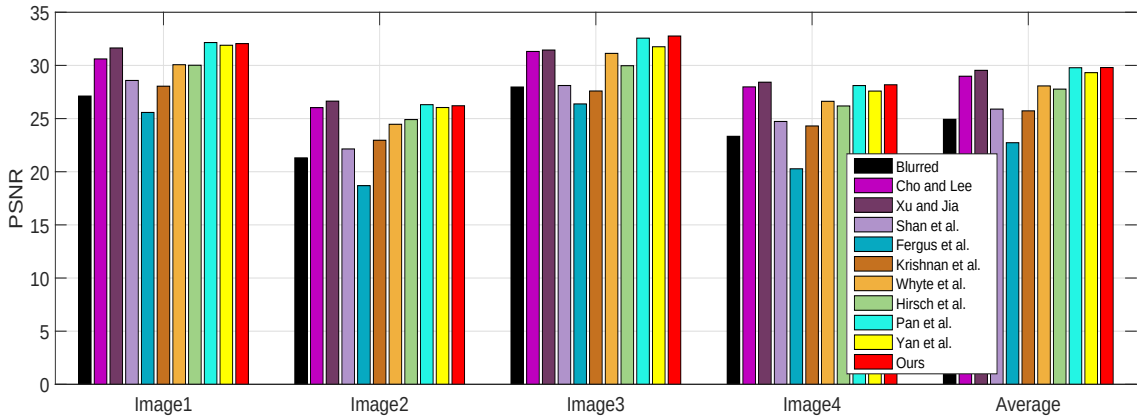


Figure 9: Evaluations on Dataset Köhler et al. (2012).

most traditional methods with sparse regularization, e.g., Whyte et al. (2012). 3) The learned filters in the proposed are sparse-promoting, which are more powerful than the basic filters for non-uniform blind deblurring.

9 Experiments

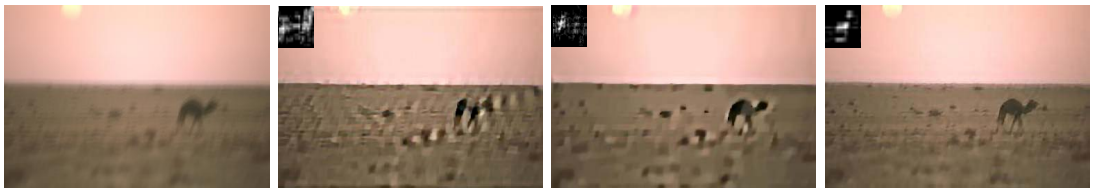
In this section, we illustrate the capabilities of the proposed method for blind, non-blind, and non-uniform image deblurring. We first evaluate its performance for blind image deblurring onnatural, text, low-illumination images datasets and some real images. We then evaluate its performance for non-blind image deblurring. Finally, we report results on blurred images undergoing a non-uniform blur kernel.

Experimental Setting: In all experiments unless especially mentioned, we set $\delta^2 = 0.002$, $\beta = 20$, $\gamma_{j,i} = 1e^{-3}$ and $d = 1e^{-4}$. To initialize the filters $\mathbf{J}_j$, we first downsample all images (grayscale) from the dataset Martin et al. (2002) to reduce noise, then train 8 3×3 filters $\mathbf{J}_j$ on the downsampling images using the method proposed in Schmidt et al. (2010) as the initialization. λ is set as $1/8$. To initialize basis set $\mathbf{B}$, we use the shifted versions of the whitening filter whose power spectrum equals the mean power spectrum of $\mathbf{J}_j$ as suggested in Weiss & Freeman (2007). We use the proposed non-blind approach in Section 7 to give the final sharp image unless otherwise mentioned. We implement the proposed method in MATLAB and evaluate the performance on an Intel Core i7 CPU with 8GB of RAM. Our naive implementation processes images of 255×255 pixels in about 27 seconds.

9.1 Experiments on Blind Image Deblurring

Dataset from Levin et al. (2009): The proposed method is first applied to a widely used dataset Levin et al. (2009), which consists of 32 blurred images, corresponding to 4 ground truth images and 8 motion blur kernels. We compare it with state-of-the-art approaches Fergus et al. (2006); Levin et al. (2011a); Babacan et al. (2012); Cho & Lee (2009); Xu & Jia (2010); Krishnan et al. (2011); Xu et al. (2013); Sun et al. (2013); Yan et al. (2017); Pan et al. (2016). To verify the mismatching problem of GSM-FoE, we implement GSM-FoE for BID by integrating the pre-learned GSM-FoE prior into the Bayesian MMSE framework introduced in Section 4.1. Furthermore, we also try to adapt GSO-FoE to handle characteristics at different scales, termed GSM-MultiScale, in which a specific GSM-FoE is trained to handle the images at a specific scale. We also verify the performance of the proposed method without updating filters to illustrate the necessity of updating filters. For a fair comparison, after estimating blur kernels using different approaches, we use the non-blind deconvolution algorithm Levin et al. (2007) with the same parameters in Levin et al. (2011a) to reconstruct the final latent image. The deconvolution error ratio, which measures the ratio between the Sum of Squared Distance (SSD) deconvolution error with the estimated and correct kernels, is used to evaluate the performance of the different methods above. Fig. 8 shows the cumulative curve of the error ratio. Fig. 8 shows that the performance of GSM-MultiScale and GSM-FoE is unsatisfactory. This may be because 1) Both are general image prior models learned only on clean images, which are not optimized to handle image-specific characteristics and are designed without considering the blind kernel estimation task. 2) GSM-FoE faces with the mismatching problem across different scales, as mentioned in Section 3.1. Furthermore, the results show that the proposed method obtains the best performance in terms of success percent 100% under error ratio 2.

Dataset from Sun et al. (2013): In a second set of experiments we use dataset from Sun et al. (2013), which contains 640 images synthesized by blurring 80 natural images with 8 motion blur kernels borrowed from Levin et al. (2009). For a fair comparison, we use the non-blind deconvolution algorithm of Zoran and Weiss Zoran & Weiss (2011) to obtain the final latent image as suggested in Sun et al. (2013). We compare the proposed approach with Levin et al. (2011a); Cho & Lee (2009); Xu & Jia (2010); Krishnan et al. (2011); Sun et al. (2013); Michaeli & Irani (2014). Fig. 8 shows the cumulative curves of the error ratio. Our results are visually competitive with others.

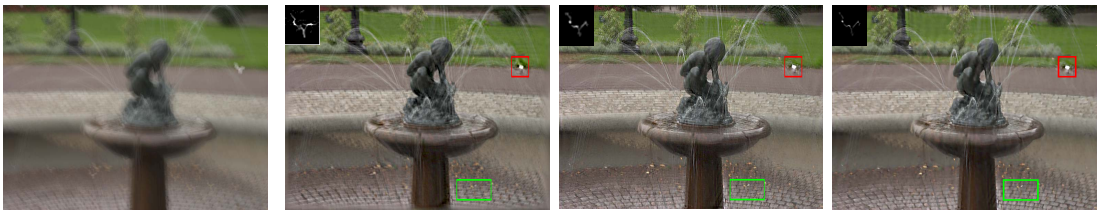


(a) Blurred (b) Result by Pan et al. (2016) (c) Result by Yan et al. (2017) (d) Ours

Figure 10: A challenging example with sky patches.

Dataset from Köhler et al. (2012): We further implement the proposed method on Köhler et al. dataset, which is blurred by space-varying blur, borrowed from Köhler et al. (2012). Although real images often exhibit spatially varying blur kernel, many approaches that assume the shift-invariant blur kernel can perform well. We compare the proposed approach with Fergus et al. (2006); Cho & Lee (2009); Xu & Jia (2010); Krishnan et al. (2011); Shan et al. (2008); Whyte et al. (2014); Hirsch et al. (2011); Pan et al. (2016). The peak-signal-to-noise ratio (PSNR) is used to evaluate their performance. Fig. 9 shows the PSNRs of the different approaches above. We can see that our results are superior to state-of-the-art approaches.

Deblurring image with sky patches: As shown in Figs. 8 and 9, the proposed method performs on par with the method with dark channel prior in Pan et al. (2016) on datasets Levin et al. Levin et al. (2009) and Köhler et al. Köhler et al. (2012). However, the method in Pan et al. (2016) obtains unsatisfactory results for blurred images that do not satisfy the condition of the dark channel prior, e.g., images with sky patches He et al. (2011). To alleviate this limitation, Yan et al. Yan et al. (2017) replace dark channel prior with



(a) Blurred (b) Fergus et al. (2006) (c) Xu et al. (2013) (d) Ours

(e) Blurred (f) Xu & Jia (2010) (g) Sun et al. (2013) (h) Ours

Figure 11: Two example images with unknown camera shake from Fergus et al. (2006) and Xu & Jia (2010).

extreme channels prior. However, its performance is still severely affected by complex brightness, as shown in Fig. 10. By contrast, the proposed method shows impressive results.

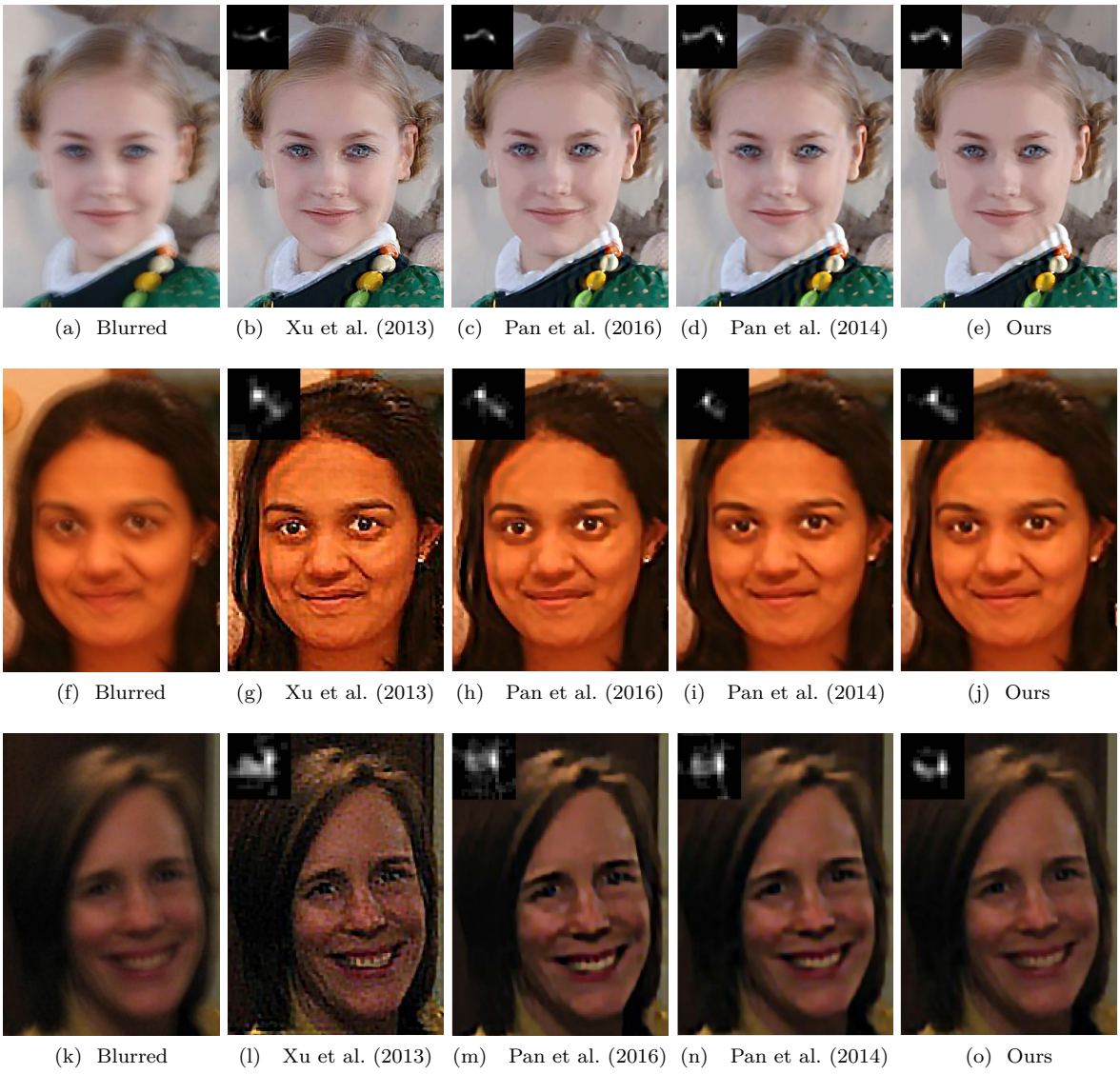
Real natural images: We test the proposed method using two real natural images. In Fig. 11 we show two comparisons on real photos with unknown camera shakes. For the blurred image 11a, Xu et al. (2013) and the proposed method produce high-quality images. For the blurred image 11e, the proposed method produces sharper edges around the texts than Xu & Jia (2010), and Sun et al. (2013).

Real face images: Deblurring face images is a challenging task because few edges in blurry face images can be used for kernel estimation. Existing methods, which implicitly or explicitly detect edges in fixed filter spaces for kernel estimation, are often unsatisfactory. In contrast, the proposed method can highlight useful edges in face images by the sparse-promoting filters, obtaining significant results. We compare the proposed method with existing methods Xu et al. (2013); Pan et al. (2016) and Pan et al. (2014) on two real face images. As shown in Fig. 12, the proposed method is superior to traditional MAP-based method Xu et al. (2013), and is competitive with the method Pan et al. (2014) specializing for deblurring face images.

Methods	PSNR	Methods	PSNR
Blurred	17.35	Xu et al. (2013)	26.21
Cho & Lee (2009)	23.80	Pan et al. (2016)	27.94
Krishnan et al. (2011)	20.86	Xiao et al. (2016)	27.56
Levin et al. (2011a)	24.90	Jinshan Pan & Yang (2014)	28.79
Zhong et al. (2013)	19.05	Ours	27.55

Table 1: Average PSNRs on text dataset Jinshan Pan & Yang (2014).

Text images: For deblurring text images, we initialize the filters $\mathbf{J}_j$ from external MSRA text images Detection 500 dataset. We verify the performance of the proposed method on text images. We compare the proposed method with existing methods, including Cho & Lee (2009); Krishnan et al. (2011); Levin et al. (2011a); Xu et al. (2013); Zhong et al. (2013); Pan et al. (2016); Xiao et al. (2016); Jinshan Pan & Yang (2014) on the text images dataset Jinshan Pan & Yang (2014). As shown in Table 1, The PSNR of the proposed algorithm is higher than that of traditional sparsity-based methods Cho & Lee (2009); Krishnan et al. (2011); Levin et al. (2011a); Xu et al. (2013); Zhong et al. (2013). The proposed method performs on par with the method of Xiao et al. Xiao et al. (2016) specifically designed for text image deblurring. On real blurry text images, as shown in Fig.13, the proposed method outperforms the traditional sparsity-based

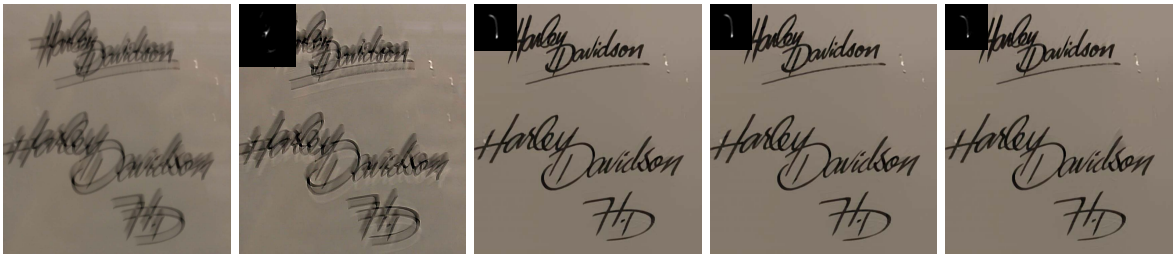


(a) Blurred (b) Xu et al. (2013) (c) Pan et al. (2016) (d) Pan et al. (2014) (e) Ours

(f) Blurred (g) Xu et al. (2013) (h) Pan et al. (2016) (i) Pan et al. (2014) (j) Ours

(k) Blurred (l) Xu et al. (2013) (m) Pan et al. (2016) (n) Pan et al. (2014) (o) Ours

Figure 12: Deblurring challenging real face examples.



(a) Blurred (b) Xu et al. (2013) (c) Pan et al. (2016) (d) Pan et al. (2014) (e) Ours

Figure 13: Deblurring a challenging real text example. The proposed method generates competitive result visionally.

method Xu et al. (2013), and generates competitive results comparable to methods designed specifically for text images.

Methods	PSNR	Methods	PSNR
Blurred	22.07	Xu et al. (2013)	22.31
Cho & Lee (2009)	22.65	Zhe Hu & Yang (2014)	22.77
Krishnan et al. (2011)	22.43	Jinshan Pan & Yang (2014)	24.16
Xu & Jia (2010)	22.73	Ours	24.22

Table 2: Average PSNR on low-illumination dataset Jinshan Pan & Yang (2014).

Low-illumination images: For deblurring low-illumination images, we initialize the filters $\mathbf{J}_j$ from external low-illumination images dataset Loh & Chan (2019). Due to saturated pixels in low-illumination images, most existing methods often obtain a delta kernel estimation. To verify the performance of the proposed method on low-illumination images, we compare the proposed method with existing methods, including Cho & Lee (2009); Xu & Jia (2010); Krishnan et al. (2011); Xu et al. (2013); Zhe Hu & Yang (2014); Jinshan Pan & Yang (2014) on the text images dataset Jinshan Pan & Yang (2014). As shown in Table 2, the PSNR of the proposed algorithm is the highest. Further, Fig.14 shows that residual blur and ringing artifacts exist in the methods Zhe Hu & Yang (2014); Jinshan Pan & Yang (2014); Pan et al. (2016), and the proposed method generates the best result, even compared with the method Zhe Hu & Yang (2014) designed specifically for low-illumination images.

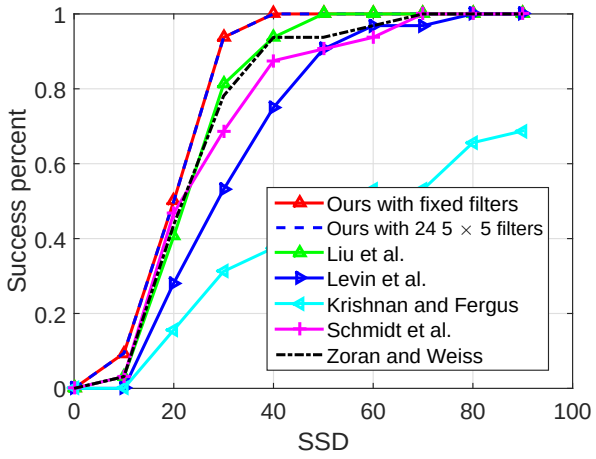
(a) Blurred (b) Zhe Hu & Yang (2014) (c) Pan et al. (2016) (d) Pan et al. (2014) (e) Ours

Figure 14: Deblurring a challenging real low-illumination example. The proposed method generates the clearest result.

9.2 Experiments on Non-blind Image Deblurring

We use the dataset from Levin et al. (2009) to verify the performance of the proposed method on non-blind image deblurring where the blur kernel is given by reference to Levin et al. (2009). In previous work Liu et al. (2018), the proposed method is directly applied for non-blind image deblurring by Eq. equation 7-Eq. equation 9 with the kernel $\mathbf{k}$ given beforehand. However, as mentioned in Section 7, the sparse-promoting filters neglect many fine details, e.g., texture. In this work, the proposed method employs two ways to handle this problem: 1) keeping the initialized filters unchanged in the iterations of the proposed method, 2) applying larger filters size. The Sum of Squared Distance (SSD) deconvolution error between the deblurred image and groundtruth is used to evaluate the performance of the different methods above.

Natural images: We compare the proposed method against Levin et al. Levin et al. (2007) with the same parameters in Levin et al. (2011a), Krishnan and Fergus Krishnan & Fergus (2009), Zoran and Weiss Zoran & Weiss (2011), Schmidt et al. Schmidt et al. (2011) and the previous work Liu et al. (2018). Fig. 15 shows the cumulative curve of SSD on dataset Levin et al. (2011a). We can see that the proposed method produces competitive results. Additionally, as shown in Table 3, compared with Zoran and Weiss Zoran & Weiss (2011), the proposed method obtains the lowest average SSD and less run time. Compared with GSM-FoE based Schmidt et al. Schmidt et al. (2011), the proposed SGF-based method acquires better results and requires much less run time. Compared with the previous work Liu et al. (2018), the proposed method gets more fine details and better performance.



(a)

Figure 15: Non-blind deblurring evaluation on dataset Levin et al. (2009).

Methods	Average SSD	Time (s)
Levin et al. (2007)	30.20	109
Zoran & Weiss (2011)	24.60	3093
Schmidt et al. (2011)	25.43	>10000
Krishnan & Fergus (2009)	82.35	**6**
Liu et al. (2018)	21.77	485
Ours with fixed filters	18.90	485
Ours with 24 5×5 filters	**18.00**	490

Table 3: Average SSD and run time on dataset Levin et al. (2009). For the case where 24 5×5 filters are used, we empirically employ fewer number of iterations to accelerate processing time.

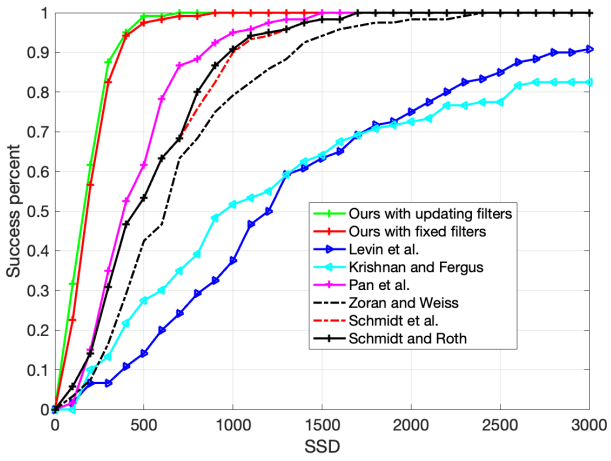
Text images: We compare the proposed method against Levin et al. Levin et al. (2007) with the same parameters in Levin et al. (2011a), Krishnan and Fergus Krishnan & Fergus (2009), Zoran and Weiss Zoran & Weiss (2011), Pan et al. Jinshan Pan & Yang (2014), Schmidt et al. Schmidt et al. (2010), and Schmidt and Rolf Schmidt & Roth (2014). Fig. 16 shows the cumulative curve of SSD on text dataset Jinshan Pan & Yang (2014) and deblurred results by different approaches on a challenging example. We can see that the proposed method obtains significant results for non-blind text deblurring. In addition, unlike natural images, since text images contain fewer details, the proposed method with adaptively sparsity-promoting filters can obtain better results than fixed filters.

9.3 Experiments on Non-uniform Image Deblurring

In the last experiment, we evaluate the performance of the proposed approach on blurred images with non-uniform blur. In the previous work Liu et al. (2018), β in Eq. equation 16 is set as 0.01 for non-uniform deblurring. However, this setting loses sight of updating the noise mentioned in Section 8. In this work, we adaptively update noise. We compare the proposed method with Whyte et al. Whyte et al. (2012), Xu et al. Xu et al. (2013) and the previous work Liu et al. (2018). Fig. 17-18 shows real natural images with non-uniform blur kernel and deblurred results. The proposed method generates images with fewer artifacts and more details.

10 Conclusions

To explore effective image priors for blind image deblurring, we deeply investigated the inherent limitation of the traditional high-order MRFs model, i.e., a set of universal filters. To break the limitation, we propose



(a)

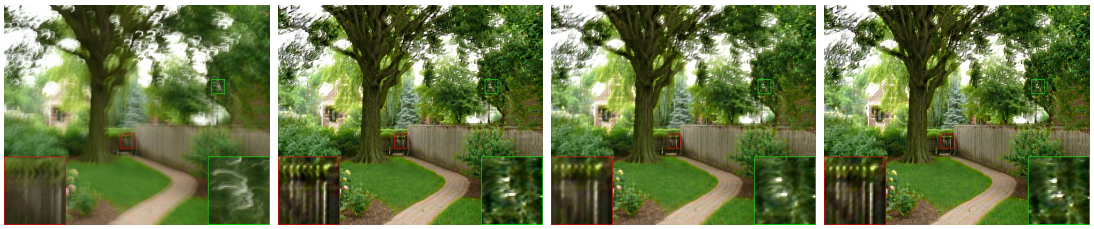
(b) (c) (d) (e) (f) (g)

Figure 16: Quantitative and qualitative evaluation on dataset Jinshan Pan & Yang (2014). (a) Cumulative histograms of SSD. (b)-(g) show a challenging example. From left to right: Blurred SSD:7675.32, Levin et al. (2007) SSD:871.37, Krishnan & Fergus (2009) SSD:3010.54, Zoran & Weiss (2011) SSD:306.06, Pan et al. (2014) SSD:257.07 and the proposed method SSD:108.84.

(a) Blurred (b) Xu et al. (2013) (c) Liu et al. (2018) (d) Ours

(e) Blurred (f) Xu et al. (2013) (g) Liu et al. (2018) (h) Ours

Figure 17: Deblurring two challenging examples with non-uniform blur kernel.



(a) Blurred (b) Whyte et al. (2012) (c) Liu et al. (2018) (d) Ours

Figure 18: Deblurring a challenging example with a non-uniform blur kernel.

a novel supper-Gaussian fields model, referred to as Super-Gaussian Fields (SGF), by defining supper-Gaussian as potential. This model contains two exciting properties, **Property** 1 and **Property** 2 introduced in Section 3.2, which provide theoretical support for image-specific filters. Relying on the proposed SGF prior and Bayesian MMSE, this work proposes a novel method to image deblurring. We theoretically show that the proposed method can avoid troublesome local minima to some extent and learn image-adaptive sparsity-promoting filters. Most importantly, with the theory support, the proposed method can be naturally extended to various scenarios, e.g., face, text, and low-illumination image deblurring. Extensive experiments demonstrate the theoretical advantages and practical effectiveness of the proposed method. It is interesting to exploit adaptive sparse-promoting filter spaces by other methods for BID in the future.

References

S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. Variational bayesian blind deconvolution using a total variation prior. *IEEE Transactions on Image processing*, 18(1):12–26, 2009.

S. Derin Babacan, Rafael Molina, Minh N. Do, and Aggelos K. Katsaggelos. Bayesian blind deconvolution with general sparse image priors. In *European Conference on Computer Vision*, pp. 341–355, 2012.

Yuanchao Bai, Huizhu Jia, Ming Jiang, Xianming Liu, Xiaodong Xie, and Wen Gao. Single-image blind deblurring using multi-scale latent structure prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2033–2045, 2019.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

Liang Chen, Faming Fang, Tingting Wang, and Guixu Zhang. Blind image deblurring with local maximum gradient prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1742–1750, 2019.

Liang Chen, Jiawei Zhang, Songnan Lin, Faming Fang, and Jimmy S Ren. Blind deblurring for saturated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6308–6316, 2021.

Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, and Hao Yin. Strong NP-hardness for sparse optimization with concave penalty functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 740–747, 2017.

Sunghyun Cho and Seungyong Lee. Fast motion deblurring. In *ACM SIGGRAPH Asia 2009 Papers*, pp. 145:1–145:8, 2009.

Taeg Sang Cho, S Paris, B. K. P Horn, and W. T Freeman. Blur kernel estimation using the radon transform. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 241–248, 2011.

Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman. Removing camera shake from a single photograph. *Acm Transactions on Graphics*, 25(25):787–794, 2006.

D. Ge, J. Idier, and E. Le Carpentier. Enhanced sampling schemes for mcmc based blind bernoulli gaussian deconvolution. *Signal Processing*, 91(4):759–772, 2009.

Dong Gong, Mingkui Tan, Yanning Zhang, Anton Van Den Hengel, and Qinfeng Shi. Blind image deconvolution by automatic gradient activation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1827–1836, 2016.

Dong Gong, Mingkui Tan, Yanning Zhang, Anton van den Hengel, and Qinfeng Shi. Self-paced kernel estimation for robust blind image deblurring. In *International Conference on Computer Vision*, pp. 1661–1670, 2017a.

Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2319–2328, 2017b.

Dong Gong, Mingkui Tan, Qinfeng Shi, Anton Van Den Hengel, and Yanning Zhang. MPTV: Matching pursuit-based total variation minimization for image deconvolution. *IEEE Transactions on Image Processing*, 28(4):1851–1865, 2018.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2011.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

M Hirsch, S Sra, B Scholkopf, and S Harmeling. Efficient filter flow for space-variant multiframe blind deconvolution. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 607–614, 2010.

Michael Hirsch, Christian J. Schuler, Stefan Harmeling, and Bernhard Scholkopf. Fast removal of non-uniform camera shake. In *International Conference on Computer Vision*, pp. 463–470, 2011.

Zhixun Su Jinshan Pan, Zhe Hu and Ming-Hsuan Yang. Deblurring text images via l0regularized intensity and gradient prior. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Neel Joshi, Richard Szeliski, and David J. Kriegman. Psf estimation using sharp edge prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *European Conference on Computer Vision*, pp. 27–40. Springer, 2012.

Nikos Komodakis and Nikos Paragios. Mrf-based blind image deconvolution. In *Asian Conference on Computer Vision*, pp. 361–374, 2013.

Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Advances in Neural Information Processing Systems*, pp. 1033–1041, 2009.

Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 233–240. IEEE, 2011.

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8878–8887, 2019.

Wei Sheng Lai, Jian Jiun Ding, Yen Yu Lin, and Yung Yu Chuang. Blur kernel estimation using normalized color-line priors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 64–72, 2015.

Anat Levin, Rob Fergus, Fredo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. *Acm Transactions on Graphics*, 26(3):70, 2007.

Anat Levin, Yael Weiss, Frederic Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1964–1971, 2009.

Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Efficient marginal likelihood optimization in blind deconvolution. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2657–2664, 2011a.

Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding blind deconvolution algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2354, 2011b.

Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *European Conference on Computer Vision*, pp. 736–753. Springer, 2022.

Songnan Lin, Jiawei Zhang, Jinshan Pan, Yicun Liu, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning to deblur face images via sketch synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11523–11530, 2020.

Yuhang Liu, Wenyong Dong, Dong Gong, Lei Zhang, and Qinfeng Shi. Deblurring natural image using super-gaussian fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 452–468, 2018.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: https://doi.org/10.1016/j.cviu.2018.10.010.

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12861–12870, 2022.

David R. Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to. *Proc.int¡ ¯l Conf.computer Vision*, 2(11):416–423 vol.2, 2002.

Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *European Conference on Computer Vision*, pp. 783–798, 2014.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, MA, 2012.

TM Nimisha, Akash Kumar Singh, and AN Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *The IEEE International Conference on Computer Vision*, volume 2, 2017.

J. A. Palmer, D. P. Wipf, K. Kreutz-Delgado, and B. D. Rao. Variational em algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*, pp. 1059–1066, 2005.

Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *European conference on computer vision*, pp. 47–62. Springer, 2014.

Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming Hsuan Yang. Blind image deblurring using dark channel prior. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1628–1636, 2016.

Liyuan Pan, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Phase-only image based kernel estimation for single image blind deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6034–6043, 2019.

Daniele Perrone and Paolo Favaro. Total variation blind deconvolution: The devil is in the details. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2909–2916, 2014.

Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3341–3350, 2020.

Stefan Roth and Michael J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205, 2009.

U. Schmidt, K. Schelten, and S. Roth. Bayesian deblurring with integrated noise estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2632, 2011.

Uwe Schmidt and Stefan Roth. Shrinkage fields for effective image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2774–2781, 2014.

Uwe Schmidt, Qi Gao, and Stefan Roth. A generative perspective on mrfs in low-level vision. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1751–1758, 2010.

Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *Acm Transactions on Graphics*, 27(3):15–19, 2008.

Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8260–8269, 2018.

Yibing Song, Jiawei Zhang, Lijun Gong, Shengfeng He, Linchao Bao, Jinshan Pan, Qingxiong Yang, and Ming-Hsuan Yang. Joint face hallucination and deblurring via structure generation and detail enhancement. *International journal of computer vision*, 127(6):785–800, 2019.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1279–1288, 2017.

Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *IEEE International Conference on Computational Photography*, pp. 1–8, 2013.

Phong Tran, Anh Tuan Tran, Quynh Phung, and Minh Hoai. Explore image deblurring via encoded blur kernel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11956–11965, 2021.

D. Tzikas, A. Likas, and N. Galatsanos. Variational bayesian blind image deconvolution with student-t priors. In *IEEE International Conference on Image Processing*, pp. 109–112, 2007.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Yair Weiss and William T Freeman. What makes a good model of natural images? In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *International journal of computer vision*, 98(2):168–186, 2012.

Oliver Whyte, Josef Sivic, and Andrew Zisserman. Deblurring shaken and partially saturated images. *International Journal of Computer Vision*, 110(2):185–201, 2014.

David Wipf and Haichao Zhang. Revisiting bayesian blind deconvolution. *The Journal of Machine Learning Research*, 15(1):3595–3634, 2014.

Lei Xiao, Jue Wang, Wolfgang Heidrich, and Michael Hirsch. Learning high-order filters for efficient blind deconvolution of document photographs. In *European Conference on Computer Vision*, 2016.

Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *European Conference on Computer Vision*, pp. 157–170, 2010.

Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1107–1114, 2013.

Yanyang Yan, Wenqi Ren, Yuanfang Guo, Rui Wang, and Xiaochun Cao. Image deblurring via extreme channels prior. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6978–6986, 2017.

Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1083–1090. IEEE, 2012.

H. Zhang, Y. Zhang, H. Li, and T. S. Huang. Generative bayesian image super resolution with natural image prior. *IEEE Transactions on Image processing*, 21(9):4054–4067, 2012.

Haichao Zhang, David Wipf, and Yanning Zhang. Multi-observation blind deconvolution with an adaptive sparse prior. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1628–1643, 2014.

Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022a.

Lei Zhang, Wei Wei, Yanning Zhang, Chunhua Shen, Anton van den Hengel, and Qinfeng Shi. Cluster sparsity field: An internal hyperspectral imagery prior for reconstruction. *International Journal of Computer Vision*, pp. 1–25, 2018.

Meina Zhang, Yingying Fang, Guoxi Ni, and Tieyong Zeng. Pixel screening based intermediate correction for blind deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5892–5900, 2022b.

Jue Wang Zhe Hu, Sunghyun Cho and Ming-Hsuan Yang. Deblurring low-light images with light streaks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014.

Lin Zhong, Sunghyun Cho, Dimitris Metaxas, Sylvain Paris, and Jue Wang. Handling noise in single image deblurring using directional filters. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 612–619, 2013.

Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2482–2491, 2019.

Yipin Zhou and Nikos Komodakis. A map-estimation framework for blind deblurring using high-level edge priors. In *European Conference on Computer Vision*, pp. 142–157, 2014.

Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *IEEE International Conference on Computer Vision*, pp. 479–486, 2011.

Supplemental Materials

This supplementary material provides details of the derivation, some discussion of kernel size, and theory proofs in the main paper.

1 Derivation of the optimization method

According to Section 4.1, the proposed method recovers latent image given the blur kernel by optimizing the following objective function:

$$
\begin{aligned}
&\max \int q(\mathbf{x}) \log p(\mathbf{x}, \mathbf{y}|\mathbf{k}) \mathbf{dx} - \int q(\mathbf{x}) \log q(\mathbf{x}) \mathbf{dx} \\
=&\max \int q(\mathbf{x}) \log (p(\mathbf{x})^{\lambda} p(\mathbf{y}|\mathbf{x}, \mathbf{k})) \mathbf{dx} - \int q(\mathbf{x}) \log q(\mathbf{x}) \mathbf{dx} \\
=&\max \int q(\mathbf{x}) \log ((\frac{1}{Z(\Theta)} \prod_{c \in C} \prod_{j=1}^{J} \max_{\gamma_{j,c} \geq 0} \mathcal{N}(\mathbf{J}_j \mathbf{x}_c | 0, \gamma_{j,c}))^{\lambda} \mathcal{N}(\mathbf{y}|\mathbf{k} \otimes \mathbf{x}, \delta^2 \mathbf{I}_N)) \mathbf{dx} - \int q(\mathbf{x}) \log q(\mathbf{x}) \mathbf{dx} \qquad (1) \\
\geq&\max \int q(\mathbf{x}) \log ((\frac{1}{Z(\Theta)} \prod_{c \in C} \prod_{j=1}^{J} \mathcal{N}(\mathbf{J}_j \mathbf{x}_c | 0, \gamma_{j,c}))^{\lambda} \mathcal{N}(\mathbf{y}|\mathbf{k} \otimes \mathbf{x}, \delta^2 \mathbf{I}_N)) \mathbf{dx} - \int q(\mathbf{x}) \log q(\mathbf{x}) \mathbf{dx},
\end{aligned}
$$

Estimating $q(\mathbf{x})$: Setting the partial differential of Eq. equation 1 with respect to $q(\mathbf{x})$ to zero, we have:

$$
\begin{aligned}
\log q(\mathbf{x}) &= \log ((\frac{1}{Z(\Theta)} \prod_{c \in C} \prod_{j=1}^{J} \mathcal{N}(\mathbf{J}_j \mathbf{x}_c | 0, \gamma_{j,c}))^{\lambda} \mathcal{N}(\mathbf{y}|\mathbf{k} \otimes \mathbf{x}, \delta^2 \mathbf{I}_N)), \qquad (2) \\
&= \lambda \log (\frac{1}{Z(\Theta)}) + \lambda \log (\prod_{c \in C} \prod_{j=1}^{J} \mathcal{N}(\mathbf{J}_j \mathbf{x}_c | 0, \gamma_{j,c})) + \log (\mathcal{N}(\mathbf{y}|\mathbf{k} \otimes \mathbf{x}, \delta^2 \mathbf{I}_N)), \qquad (2a) \\
&= \mathbf{x}^T (\sum\nolimits_j \lambda \mathbf{T}_{\mathbf{J}_j}^T \mathbf{W}_j \mathbf{T}_{\mathbf{J}_j}) \mathbf{x} + \mathbf{x}^T (\delta^{-2} \mathbf{T}_{\mathbf{k}}^T \mathbf{T}_{\mathbf{k}}) \mathbf{x} - (\delta^{-2} \mathbf{T}_{\mathbf{k}}^T \mathbf{y})^T \mathbf{x} + const, \qquad (2b)
\end{aligned}
$$

Collecting together terms in $\mathbf{x}$, we have:

$$
-\log q(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} + const. \qquad (3)
$$

Estimating J: Collecting together terms with respect to $\mathbf{J}_j$ in Eq. equation 1, we obtain:

$$
\max (\frac{1}{Z(\{\mathbf{J}_j\})} + \lambda \sum\nolimits_j \sum\nolimits_c \frac{\langle (\mathbf{J}_j \mathbf{x}_c)^2 \rangle}{-2\gamma_{j,c}}). \qquad (4)
$$

Based on **Property 2**,limiting the updating of $\mathbf{J}_j$ in the orthonormal space can lead to:

$$
\max_{\mathbf{J}_j \in ortho} \lambda \sum\nolimits_j \sum\nolimits_c \frac{\langle (\mathbf{J}_j \mathbf{x}_c)^2 \rangle}{-2\gamma_{j,c}}. \qquad (5)
$$

Deleting λ, we obtain:

$$
\mathbf{R}_j = eig \min (\mathbf{B}^T \langle \mathbf{T}_{\mathbf{x}} \mathbf{W}_j \mathbf{T}_{\mathbf{x}}^T \rangle \mathbf{B}) \quad , \quad \mathbf{V}_{\mathbf{J}_j} = \mathbf{B} \mathbf{R}_j. \qquad (6)
$$

Estimating γ: Collecting together terms with respect to $\gamma_{j,c}$ in Eq. equation 1, we obtain:

$$
\max (\log (\gamma_{j,c})^{-\frac{1}{2}} + \sum\nolimits_j \sum\nolimits_c \frac{\langle (\mathbf{J}_j \mathbf{x}_c)^2 \rangle}{-2\gamma_{j,c}}). \qquad (7)
$$

Setting the partial differential with respect to $\gamma_{j,c}$ to zero, we have:

$$
\gamma_{j,c} = \langle (\mathbf{J}_j \mathbf{x}_c)^2 \rangle. \qquad (8)
$$

Estimating δ^2 is similar to $\gamma_{j,c}$.

2 Proof of Theorem 2

Given the objective function equation 12, we can give an equivalent objective as follows:

$$\mathcal{L}(\mathbf{x},\mathbf{J},\gamma,\delta^2)=\frac{1}{\delta^2}||\mathbf{y}-\mathbf{k}\otimes\mathbf{x}||_2^2+\sum_c[\lambda\sum_j(\frac{(\mathbf{J}_j\mathbf{x}_c)^2}{\gamma_{j,i}}+\log\gamma_{j,i})+\log\delta^2+\log(\lambda\sum_j(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{V}_{\gamma_{j,c}^{-1}}+\frac{||\mathbf{k}||_2^2}{\delta^2})], \quad (9)$$

We aim to directly minimize $\mathcal{L}(\mathbf{x},\mathbf{J},\gamma,\delta^2)$ over $\mathbf{x},\mathbf{J},\gamma,\delta^2$. To do that, we independently update each variable while holding the other three variables fixed. For updating $\mathbf{x}$, after collecting relevant terms in equation 9, we have:

$$\min\frac{1}{\delta^2}||\mathbf{y}-\mathbf{k}\otimes\mathbf{x}||_2^2+\lambda\sum_c\sum_j\frac{(\mathbf{J}_j\mathbf{x}_c)^2}{\gamma_{j,i}}, \quad (10)$$

for which a convenient closed-form solution $\mathbf{x}^{opt}$, denotes the optimal solution, can be given by:

$$\mathbf{x}^{opt}=\mathbf{A}^{-1}\mathbf{b}, \quad (11)$$

where $\mathbf{A},\mathbf{b}$ are the same as that defined in equation 7.

Then, we consider updating $\mathbf{J}_j$, after collecting relevant terms in equation 9, we have:

$$\min_{\mathbf{J}_j}\sum_c[\lambda\sum_j\frac{(\mathbf{J}_j\mathbf{x}_c)^2}{\gamma_{j,i}}+\log(\lambda\sum_j(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{V}_{\gamma_{j,c}^{-1}}+\frac{||\mathbf{k}||_2^2}{\delta^2})] \quad (12)$$

Because no closed-form solution for equation 12 is available, similar to Wipf & Zhang (2014), we instead use basic principles from convex analysis to form a strict upper bound that will facilitate subsequent optimization. In particular, we use:

$$(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{z}_j-\Phi^*(\mathbf{z}_j)\geq\log(\lambda\sum_j(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{V}_{\gamma_{j,c}^{-1}}+\frac{||\mathbf{k}||_2^2}{\delta^2}), \quad (13)$$

where $\Phi^*(\boldsymbol{\alpha})$ is the concave conjugate of the concave function $\Phi(\boldsymbol{\alpha})=\log(\lambda\boldsymbol{\alpha}^T\mathbf{V}_{\gamma_{j,c}^{-1}}+\lambda\sum_{j'\neq j}(\mathbf{V}_{\mathbf{J}_{j'}}^2)^T\mathbf{V}_{\gamma_{j',c}^{-1}}+\frac{||\mathbf{k}||_2^2}{\delta^2})$. It can be shown that equality in equation 13 is achieved when:

$$\mathbf{z}_j^{opt}=\frac{\partial\Phi}{\partial\boldsymbol{\alpha}}|_{\boldsymbol{\alpha}=\mathbf{V}_{\mathbf{J}_j}^2}=\frac{\lambda\mathbf{V}_{\gamma_{j,c}^{-1}}}{\lambda\sum_j(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{V}_{\gamma_{j,c}^{-1}}+\frac{||\mathbf{k}||_2^2}{\delta^2}}=\lambda\mathbf{V}_{\gamma_{j,c}^{-1}}A_{i,i}^{-1}, \quad (14)$$

where $A_{i,i}^{-1}$ denote the inverse of the diagonal element with index (i,i) of $\mathbf{A}$. Plugging equation 13 into equation 12, we obtain the revised problem

$$\min_{\mathbf{J}_j}\sum_c[\lambda\frac{(\mathbf{J}_j\mathbf{x}_c)^2}{\gamma_{j,i}}+(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{z}_j], \quad (15)$$

which gives the following updated equation:

$$\min_{\mathbf{J}_j}\mathbf{V}_{\mathbf{J}_j}^T\langle\mathbf{T}_\mathbf{x}\mathbf{W}_j{\mathbf{T}_\mathbf{x}}^T\rangle\mathbf{V}_{\mathbf{J}_j}. \quad (16)$$

We now examine optimization over $\gamma_{j,i}$. Isolating terms in equation 9, this requires that we solve:

$$\min_{\gamma_{j,i}}\sum_c[\lambda\sum_j(\frac{(\mathbf{J}_j\mathbf{x}_c)^2}{\gamma_{j,i}}+\log\gamma_{j,i})+\log(\lambda\sum_j(\mathbf{V}_{\mathbf{J}_j}^2)^T\mathbf{V}_{\gamma_{j,c}^{-1}}+\frac{||\mathbf{k}||_2^2}{\delta^2})]. \quad (17)$$

Note that $\gamma_{j,i}$ not only appears in the center of the clique c, but also in the corresponding location of other cliques, due to convolution operation. Optimizing Eq. equation 17 means optimizing the following objective:

$$\min_{\gamma_{j,i}} \lambda \sum_j (\frac{(\mathbf{J}_j \mathbf{x}_c)^2}{\gamma_{j,i}} + \log \gamma_{j,i}) + \sum_{c_i} [\log(\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c_i}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2})]. \tag{18}$$

where c_i denotes the cliques that include the location i. For each clique c_i, we resort to similar bounding techniques as before, adopting:

$$\boldsymbol{\upsilon}_{j,i} \gamma_{j,i}^{-1} - \psi^*(\boldsymbol{\upsilon}_{j,i}) \geq \log(\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c_i}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2}), \tag{19}$$

where holds for all $\boldsymbol{\upsilon}_{j,i} \geq 0$ (That is, the left of Eq. (19) provides a upper bound for the right.), and $\psi^*(\boldsymbol{\alpha})$ is the concave conjugate of the concave function $\psi(\boldsymbol{\alpha}) = \log(\lambda(\mathbf{J}_{j,i}^2)\boldsymbol{\alpha} + \lambda \sum_{i' \neq i} (\mathbf{J}_{j,i'}^2)^T \gamma_{j,i'}^{-1} + \lambda \sum_{j' \neq j} (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c_i}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2})$, here $\mathbf{J}_{j,i}$ denote the element of the filter J_j related to $\gamma_{j,i}^{-1}$ in the clique c_i. It can be shown that equality in equation 19 is achieved when:

$$\upsilon_{j,i}^{opt} = \frac{\partial \psi}{\partial \boldsymbol{\alpha}}|_{\boldsymbol{\alpha} = \gamma_{j,i}^{-1}} = \frac{\lambda \mathbf{J}_{j,i}^2}{\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c_i}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2}} = \lambda (\mathbf{J}_{j,i}^2) A_{c_i,c_i}^{-1}. \tag{20}$$

Plugging equation 19 into equation 18, we obtain the revised problem

$$\min_{\gamma_{j,i}} \lambda (\frac{(\mathbf{J}_j \mathbf{x}_c)^2}{\gamma_{j,i}} + \log \gamma_{j,i}) + \sum_{c_i} \boldsymbol{\upsilon}_{j,i} \gamma_{j,i}^{-1} \tag{21}$$

which gives the following updated equation

$$\gamma_{j,i}^{opt} = (\mathbf{J}_j \mathbf{x}_c)^2 + \mathbf{J}_j^2 A_{c,c}^{-1} = \langle (\mathbf{J}_j \mathbf{x}_c)^2 \rangle. \tag{22}$$

Next, we consider optimization over δ^2. Again, we first isolate terms in equation 9:

$$\mathcal{L}(\mathbf{x}, \mathbf{J}_j, \gamma_{j,i}, \delta^2) = \frac{1}{\delta^2} ||\mathbf{y} - \mathbf{k} \otimes \mathbf{x}||_2^2 + \sum_c [\log \delta^2 + \log(\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2})], \tag{23}$$

then resort to similar bounding techniques as before, adopting:

$$\delta^{-2} \vartheta - \psi^*(\vartheta) \geq \log(\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2}), \tag{24}$$

where $\psi^*(\vartheta)$ is the concave conjugate of the concave function $\psi(\vartheta) = \log(\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + ||\mathbf{k}||_2^2 \vartheta)$. It can be shown that equality in equation 24 is achieved when:

$$\upsilon_{j,i}^{opt} = \frac{\partial \psi}{\partial \boldsymbol{\alpha}}|_{\boldsymbol{\alpha} = \delta^{-2}} = \frac{||\mathbf{k}||_2^2}{\lambda \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c_i}^{-1}} + \frac{||\mathbf{k}||_2^2}{\delta^2}} = ||\mathbf{k}||_2^2 A_{i,i}^{-1}. \tag{25}$$

Plugging equation 24 into equation 23, we obtain:

$$\min_{\gamma_{j,i}} \lambda (\frac{(\mathbf{J}_j \mathbf{x}_c)^2}{\gamma_{j,i}} + \log \gamma_{j,i}) + \sum_c [\log \delta^2 + \boldsymbol{\upsilon}_{j,i} \delta^{-2}] \tag{26}$$

Finally, we can obtain:

$$\delta^2 = \frac{||(\mathbf{y} - \mathbf{k} \otimes \mathbf{x})||^2 + \sum_c ||\mathbf{k}||_2^2 A_{i,i}^{-1}}{n} = \frac{\langle ||(\mathbf{y} - \mathbf{k} \otimes \mathbf{x})||^2 \rangle}{n}. \tag{27}$$

As mentioned before, by introducing the hyper-parameter d into equation 27, we can achieve equation 10 in the main paper.

3 Proof of Theorem 3

Eq. equation 12 can be expressed as:

$$g(\mathbf{J}, \mathbf{x}_c, \delta^2) = \min_{\gamma_{j,i} \geq 0} \lambda \sum_j (\frac{(\mathbf{J}_j \mathbf{x}_c)^2}{\gamma_{j,i}}) + \lambda \sum_j (\log \gamma_{j,i}) + \log(\lambda \delta^2 \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + ||\mathbf{k}||_2^2) \tag{28}$$

Since *log* is a concave, non-decreasing function of $\gamma_{j,i}$, it can be see that $\lambda \sum_j (\log \gamma_{j,i}) + \log(\lambda \delta^2 \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + ||\mathbf{k}||_2^2)$ is also a concave, non-decreasing function of $\gamma_{j,i}$, as a result, we can always express $\varphi(\gamma_{j,i}) = \lambda \sum_j (\log \gamma_{j,i}) + \log(\lambda \delta^2 \sum_j (\mathbf{V}_{\mathbf{J}_j}^2)^T \mathbf{V}_{\gamma_{j,c}^{-1}} + ||\mathbf{k}||_2^2)$ as:

$$\varphi(\gamma_{j,i}) = \min_{z \geq 0} z \gamma_{j,i} - \varphi^*(z), \tag{29}$$

where $\varphi^*(z)$ is the concave conjugate of $\varphi(\gamma_{j,i})$. In this case, optimizing over $\gamma_{j,i}$ for fixed $(\mathbf{J}_j \mathbf{x}_c)$ and z, the optimal solution is:

$$\gamma_{j,i}^{opt} = z^{-1/2} |(\mathbf{J}_j \mathbf{x}_c)|, \tag{30}$$

which implies that:

$$g(\mathbf{J}, \mathbf{x}_c, \delta^2) = \min_{z \geq 0} 2 z^{1/2} |(\mathbf{J}_j \mathbf{x}_c)| - \varphi^*(z). \tag{31}$$

Any function expressible in this form is necessarily concave, and also non-decreasing for $(\mathbf{J}_j \mathbf{x}_c)$ since $z \geq 0$, as mentioned in Wipf & Zhang (2014).

4 Proof of Theorem 4

Part (1) is very straightforward, similar to Theorem 2 in Zhang et al. (2014). As $\mathbf{J}_j \mathbf{x}_c \to \infty$, the optimizing $\gamma_{i,j}$ will become arbitrarily large regardless of the value of δ^2. In the regime where $\gamma_{i,j}$ is sufficiently large, the difference $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$ must converge to zero. It then follows that the difference between the corresponding minimizing $\gamma_{i,j}$ values, and therefore the cost function difference, converges to zero. Part (2) is also very similar to the proof of Theorem 2 in Zhang et al. (2014), assume that $\delta_2^2 \geq \delta_1^2$, it can be easily shown that $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2) \geq g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$. Then consider a second point $\mathbf{J}\mathbf{x}_c' \succ \mathbf{J}\mathbf{x}_c$. Because the gradient at every intermediate point moving from $g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$ to $g(\mathbf{J}, \mathbf{x}_c', \delta_1^2)$ is greater than the associated gradients moving from $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2)$ to $g(\mathbf{J}, \mathbf{x}_c', \delta_2^2)$, it must be the case that $g(\mathbf{J}, \mathbf{x}_c, \delta_1^2)$ increased at a faster rate than $g(\mathbf{J}, \mathbf{x}_c', \delta_2^2)$, and so it follows that $g(\mathbf{J}, \mathbf{x}_c, \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c, \delta_1^2) \geq g(\mathbf{J}, \mathbf{x}_c', \delta_2^2) - g(\mathbf{J}, \mathbf{x}_c', \delta_1^2)$.