Collaborative Deep Learning for Super-Resolving Blurry Text Images

Yuhui Quan, Jieting Yang, Yixin Chen, Yong Xu*, Hui Ji

***Abstract*—Text image, the one with its content dominated by text, is a common type of images seen in many applications. In practice, text images are often degraded by many factors mixed together. This paper aims at recovering degraded text images that suffer from the mixture of multiple degradations, including low resolution, uniform blurring and noise. It is observed that such mixed degradations treat the low-frequency components and high-frequency components of an image in different manners, and the emphasis of recovery is on predicting high-frequency information. Motivated by such an observation, we proposed a neural network that collaboratively works on the prediction of high-frequency information and the recovery of text image with both low and high frequencies. The experiments are conducted on one existing benchmark dataset of document images and one new dataset covering a wide range of text images. The results show that the proposed method noticeably outperformed the existing ones.**

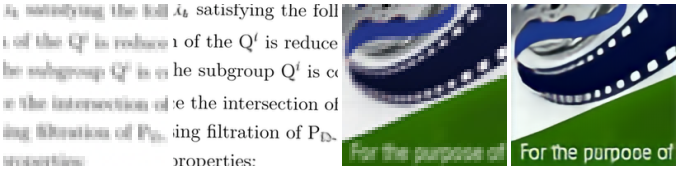
***Index Terms*—Text image processing, Image restoration, Collaborative learning; Deep learning**

I. Introduction

Text is what words, sentences, paragraphs or books are made of. Text images refer to those images whose contents are dominated by texts. Such images are one prevalent type of images in daily life, *e.g.* document images, scanned cards and pictures of class notes. See Fig. 1 for an illustration. Also, the text contents of images provide rich information for a wide range of vision applications, *e.g.* image search, target geolocation, robotic navigation, and human-machine interaction. Thus, many efforts have been made on text extraction [1], localization [2] and recognition [3].

In practice, the visual quality of the text images taken by cameras or extracted from full images is often degraded by many factors, such as low resolution, motion/defocus blurring, and low signal-to-noise ratio, among many others. These degradations significantly decrease the readability of text contents. A text image recovery method that can noticeably improve the visual quality of degraded text images certainly sees its needs in daily life, as well as in the vision applications that involve text processing. In the past, there have been an enduring research effort along this line. Most existing methods for text image recovery focus either on deblurring [4], [5], [6], [7], [8] or on super-resolution [9], [10], [11], [12], [13]. Nevertheless, low resolution and blur often simultaneously occur in real scenarios, especially when the text images are extracted from the pictures of large sizes. Recently, there are a few studies on recovering images with such complex degradations. For instance, Xu *et al.* [14] considered the problem of joint deblurring and super-resolution of text images, and contributed the first practical solution together with a new dataset of document images.

Fig. 1: Samples of text images.



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

Fig. 2: Joint super-resolution and deblurring on text images. (a)&(c): different types of low-resolution blurry text images (zoomed in for easier inspection). (b)&(d): recovered images by our method.

A. Aim

Same as Xu *et al.* [14], this paper aims at recovering text images degraded by two most often-seen factors in the presence of noise: low resolution and uniform blurring, which often make the text contents difficult to recognize. See Fig. 2a and Fig. 2c for an illustration. Let $\boldsymbol{X} \in \mathbb{R}^{N_1 \times N_2}$ denote a high-quality text image and $\boldsymbol{Y} \in \mathbb{R}^{M_1 \times M_2}$ its degraded

Yuhui Quan, Jieting Yang, Yixin Chen and Yong Xu are with School of Computer Science & Engineering at South China University of Technology, Guangzhou 510006, China, and also with the Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, China. Hui Ji is with Department of Mathematics at National University of Singapore, Singapore 119076. (Email: csyhquan@scut.edu.cn; csjietingyang@mail.scut.edu.cn; yx.chen.cs@foxmail.com; yxu@scut.edu.cn; matjh@nus.edu.sg).

This work is supported by National Natural Science Foundation of China (61872151, 61672241, U1611461), Natural Science Foundation of Guangdong Province (2017A030313376, 2020A1515010134), Science & Technology Program of Guangdong Province (2019A050510010, 20140904-160), Science & Technology Program of Guangzhou (201802010055), Fundamental Research Funds for Central Universities of China (x2js-D2181690), and Singapore MOE AcRF (R146000229114, MOE2017-T2-2-156).

Asterisk indicates the corresponding author.

observation. The formation of the degraded image $\boldsymbol{Y}$ can be formulated as

$$\boldsymbol{Y} = (\boldsymbol{K} \otimes \boldsymbol{X}) \downarrow_r + \boldsymbol{N}, \tag{1}$$

where $\otimes$ denotes the discrete convolution operator, $\downarrow_r$ denotes the operation of down-sampling with rate r, $\boldsymbol{K} \in \mathbb{R}^{Z_1 \times Z_2}$ denotes the blur kernel, and $\boldsymbol{N} \in \mathbb{R}^{M_1 \times M_2}$ denotes the noise.

In practice, the blur kernel $\boldsymbol{K}$ can be composed from multiple sources, including de-focus, motion blurring, or the smoothing introduced by the anti-aliasing process during down-sampling. In general, without additional input, it is difficult to estimate $\boldsymbol{K}$ reliably. Therefore, we consider an end-to-end approach that directly estimates $\boldsymbol{X}$ from $\boldsymbol{Y}$.

Inspired by the success of deep learning in many image recovery tasks [15], [16], [17], [18], we propose a neural network (NN) based method to recover a degraded text image of low resolution and blurry appearance so as to greatly improve its readability. See Fig. 2b and Fig. 2d for an illustration. Such a method can be applied to processing document images and text images taken by mobile devices. It also has applications in many vision tasks with text processing modules.

B. Main Idea

Both the downsampling operator and blurring operator will distort the high-frequency components of an image. Taking 1D signal for example. Let $\widehat{(\cdot)}$ denote the discrete Fourier transform. For an image $\boldsymbol{x} \in \mathbb{R}^N$, the downsampling operator with downsampling rate 4, denoted by $\downarrow_4$, will distort its high-frequency parts (known as aliasing):

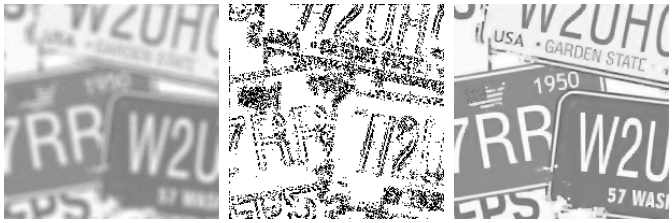
$$\widehat{\boldsymbol{x} \downarrow_4}(\omega) = \sum_{j=0}^{3} \widehat{\boldsymbol{x}}(\omega + \frac{j\pi}{2}). \tag{2}$$

For blurring with a kernel $\boldsymbol{k} \in \mathbb{R}^Z$, we have

$$\widehat{\boldsymbol{k} \otimes \boldsymbol{x}}(\omega) = \widehat{\boldsymbol{k}}(\omega) \cdot \widehat{\boldsymbol{x}}(\omega) \tag{3}$$

according to the convolution theorem [19, Chapter 3.3]. The mixture of these two operations has profound effects on distorting the high-frequency components of $\boldsymbol{x}$. Clearly, how to accurately predict the high-frequency information is the key to recover $\boldsymbol{x}$ in full spectrum from its low-resolution blurry measurement. In short, the prediction of high-frequency information should receive specific treatment when using the NN to recover a low-resolution blurry image.

For a 2D image $\boldsymbol{X}$, its image gradients, denoted by $\nabla \boldsymbol{X}$, are calculated by convolving the image using the high-pass filter $[1, -1]$ horizontally and vertically. Recall that the magnitude spectrum of such a filter in the Fourier domain is $2|\sin \frac{\omega}{2}|$, implying the filter only attenuates low-frequency components and keeps most high-frequency components. Thus, the image gradients $\nabla \boldsymbol{X}$ encode most high-frequency components of an image [20, Chapter 3.4]. However, the difficulty of predicting $\nabla \boldsymbol{X}$ is essentially the same as predicting the image $\boldsymbol{X}$ in full spectrum, up to a constant. The question is then what measurement on image gradients is suitable for our purpose that has following two properties: *(i)* it can be reliably predicted (at least for text images); *(ii)* it contains sufficient high-frequency information for recovering the image in full spectrum.



(a) (b) (c)

Fig. 3: Recovering a text image based on accurate high-frequency information. (a) A low-resolution blurry text image $\boldsymbol{Y}$ (zoomed in for better inspection). (b) The binary map $\boldsymbol{\Sigma}$ calculated on the high-resolution clear version of $\boldsymbol{Y}$, which provides accurate high-frequency information of the clear image. (c) The result $\boldsymbol{X}$ recovered by solving (4). It can be seen that once the high-frequency information is accurately predicted, the recovery result is quite good.

Consider a simple experiment in Fig. 3. A high-resolution clear text image is firstly blurred by a disk kernel with radius 4 pixels, and then downsampled by a factor of 2 to generate the degraded text image $\boldsymbol{Y}$, as shown in Fig. 3a. Then, a binary matrix, denoted by $\boldsymbol{\Sigma}$, is calculated on the high-resolution clear text image. It indicates the locations of the pixels with large magnitude of gradients, as shown in Fig. 3b. That is, $\boldsymbol{\Sigma}(i, j) = 1$ if the magnitude of gradient at the (i, j)-th pixel location is larger than a threshold. Given $\boldsymbol{Y}$ and $\boldsymbol{\Sigma}$, we estimate the truth image $\boldsymbol{X}$ by solving

$$\min_{\boldsymbol{X}} \|\boldsymbol{Y} - (\boldsymbol{K} \otimes \boldsymbol{X}) \downarrow_4 \|_2^2 + \lambda \|(1 - \boldsymbol{\Sigma}) \odot \nabla \boldsymbol{X}\|_2^2, \tag{4}$$

where $\odot$ denotes the element-wise multiplication, and the weight λ is empirically set to 0.4. The problem (4) can be solved by finding its unique stationary point, which requires solving a linear system. The result in Fig. 3c shows that a text image can be fully recovered from its low-resolution blurry version, when only the locations of large image gradients are provided as additional inputs. In short, a weak form of image gradient magnitudes can serve the purpose of predicting high-frequency information for recovering the text image well.

Motivated by the discussion above, we proposed an NN that enables the collaboration between *(i)* the prediction of the information related to high frequencies and *(ii)* the recovery of the image in full frequency spectrum. As the location information of large image gradients is not differentiable, we propose to use the local average of image gradient magnitudes along different orientations as the weak form of high-frequency information in our NN. Concretely, for an image $\boldsymbol{X} \in \mathbb{R}^{N_1 \times N_2 \times C}$ with C channels, the measurement is defined by

$$\boldsymbol{E}_{\boldsymbol{X}}(x_0, y_0) = \frac{1}{C} \sum_{c=1}^{C} \sum_{(x,y) \in \mathbb{N}(x_0, y_0)} |\boldsymbol{X}(x, y, c) - \boldsymbol{X}(x_0, y_0, c)|, \tag{5}$$

where $\mathbb{N}(x_0, y_0)$ denotes the spatial neighborhood of the point (x_0, y_0) (four-connected neighbors in practice).

The collaboration mechanism in the proposed NN is implemented by using a pair of dual convolutional neural networks (CNNs). Briefly, one CNN is trained to predict the measurement $\boldsymbol{E}_{\boldsymbol{X}}$ with the help of the estimate on the image $\boldsymbol{X}$, and

the other one is trained to recover the image $\boldsymbol{X}$ with the help of $\boldsymbol{E_X}$. Such a pair of CNNs is repeated in an iterative manner with interactions, but using different model parameters. See Fig. 4a for the diagram of the proposed NN.

Such an NN architecture has its motivation from the sparsity-based image recovery. In general, the ℓ_0-norm regularization methods (*e.g.* [21], [22]) can be unrolled into three iterative sub-processes: location detection of non-zero entries, least-square fitting, and estimation of degradation parameters (*e.g.* blur kernel estimation in blind deblurring). The locations of non-zero entries of $\nabla \boldsymbol{X}$ are replaced by the differentiable measurement $\boldsymbol{E_X}$. The other two sub-processes are implemented by another CNN that recovers the image given the input of $\boldsymbol{E_X}$.

Although $\boldsymbol{E_X}$ seems to be quite a weak statistical measurement on image gradients, the question is whether it can be accurately predicted via a CNN. Text images have their special characteristics on image gradients, which are much simpler than that of natural images in general [7]. For example, the image gradients of a text image are dominated by the ones with large magnitude and their orientations are sufficiently diverse in local regions. Such special properties are exploited when blindly deblurring text images; see *e.g.* the work of Cho *et al.* [23] as well as Xu and Jia [24].

C. Contributions

This paper proposes a new CNN-based method for recovering the text images degraded by the mixture of multiple factors, including both low resolution and blurring. Different from the existing methods, the proposed one introduces a collaborative mechanism that enables the interaction between the local statistical measurement on high-frequency components and the full frequency spectrum of the image. The motivation comes from the importance of high-frequency information in image recovery, and the intrinsic relationship between the full frequency spectrum of the image and the local statistical measurement on high-frequency components. The proposed approach also can be interpreted from model-based deep learning [25], [26], [27]. Our NN essentially unfolds the ℓ_0-norm regularization methods for image recovery, with the modification on image priors and other non-differentiable terms involved.

In addition, the research on deep learning for text image recovery can greatly benefit from available high-quality text image datasets. Currently, there is only one dataset in the public domain (Hradiš *et al.* [8]), which is limited to document images. This paper contributes a large benchmark dataset with significant variations on the types of text images, which benefits the community who is working on text images or whose work involves text image processing.

The proposed NN is extensively evaluated on both the existing dataset of document images and the new dataset of general text images. The experiments show that the proposed method can greatly improve the visual quality of low-resolution blurry text images and outperforms the existing state-of-the-art methods.

II. Related work

We first give a brief review on the methods that either deblur or super-resolve text images. Then, we review in details the related ones that jointly conduct super-resolution and deblurring on text images. Lastly, we give the discussion on some other related work.

A. Text Image Deblurring

Text image deblurring can be done using general image deblurring methods [28], [29], which usually rely on some statistical priors on image gradients. Such generic priors are not accurate for text images. For example, text images do not obey the heavy-tailed distribution of image gradients, which is often used in general image deblurring. Thus, Chen *et al.* [5] proposed to use the intensity probability density function of sharp document images as the prior for deblurring. Their method only works on the text images with monotone backgrounds. To overcome this weakness, Cho *et al.* [6] proposed to model the background in a text image with natural image statistics. In addition, they assumed high contrasts on text characters and weak gradients inside each character. Cho *et al.*'s method requires the preprocessing of text region detection. Without such a preprocessing, Pan *et al.* [7] proposed a simultaneous sparsity prior on image intensities and image gradients of text images, which leads to an ℓ_0-norm based regularization approach for text image deblurring. Instead of using pre-assumed priors, Hradiš *et al.* [8] leveraged the power of deep learning to directly learn the mapping from degraded text images to the sharp ones. They proposed a CNN for blind deblurring and denoising of document images, and showed that the CNN can well restore text images.

B. Text Image Super-Resolution

Similarly, text image super-resolution can also be done by calling general image super-resolution methods [17], [18], [30], [31], [32]. Owing to the special characteristics of text images, the methods specifically designed for text images can have better performance than the generic ones. Most existing methods for text image super-resolution are exemplar-based approaches, which use the example patches from both high-resolution (HR) and low-resolution (LR) images to guide the recovery process. Park *et al.* [9] proposed a Markov random field framework to learn the priors of text regions and backgrounds using the example pairs of original and degraded patches. Walha *et al.* [10] constructed a dataset of HR/LR patch pairs of character images and used it to learn two dictionaries for HR/LR images. The restoration is done by joint sparse representation of HR/LR patches under the two dictionaries. Abedi and Kabir [11] constructed an HR dictionary for each character and used it to reconstruct HR images with sparse representation. An exemplar set of frequent characters is also constructed by Abedi and Kabir [12] for character super-resolution, and the HR text image is obtained by placing all the super-resolved characters on their corresponding positions. The exemplar-based methods heavily rely on the dataset of example patches, and they are usually

only applicable to document images. Very recently, Xu *et al.* [33] proposed a realistic super-resolution dataset created by simulating the imaging process of digital cameras, which can be helpful to the development of learning-based super-resolution methods.

C. Joint Super-Resolution and Deblurring

It is observed by Xu *et al.* [14] that when recovering images degraded by both low resolution and blurring, sequentially applying deblurring and super-resolution techniques usually does not produce satisfactory results. An early work related to joint super-resolution and deblurring is from Joshi *et al.* [34], which proposed a nonparametric kernel estimation technique for both deblurring and super-resolution, but without a detailed algorithm for recovery. Harmeling *et al.* [35] proposed a method for joint multi-frame blind deconvolution, super-resolution and saturation correction. Liu and Sun [36] proposed a Bayesian method for joint video super-resolution and motion deblurring. These two methods require multiple images as the input and do not work on single image. Michaeli and Irani [37] proposed to estimate the blur kernel using the recurrence of small patches across different scales of the low-resolution image, which further leads to a single-image-based joint upscaling and deblurring method. All the above methods assume small simple motion occurs and thus cannot handle large complex motion blur.

The first practical solution to joint super-resolution and deblurring on text images is given by Xu *et al.* [14]. Taking advantage of deep learning, they proposed a CNN combined with a multi-class generative adversarial network (GAN), as well as proposed a dataset proposed for model training and performance evaluation. Zhang *et al.* [38] proposed a two-steam CNN for joint super-resolution and motion deblurring of general images. To bypass the challenges imposed by learning the mixed degradation all-in-one, they used different streams to learn super-resolution and deblurring separately. Such a separate treatment on super-resolution and deblurring is not optimal in the sense that it omits the fact that both the recovery processes indeed are similar in many aspects and can benefit from the same information, *e.g.* the gradient information.

D. Other Related Work

The collaboration mechanism of the proposed method focuses on the treatment of image gradients, which relates to the edge-guided mechanisms adopted in the NNs designed for other image processing tasks; see *e.g.* Fan *et al.*'s work [39] for reflection removal. Nevertheless, ours are different from Fan *et al.*'s work [39] in several aspects. In Fan *et al.*'s work [39], the predicted gradient map is mainly for guiding the separation of the reflection layer from the latent image layer, as the two layers often have different edge strengths. In our work, the use of the $\boldsymbol{E_X}$ is motivated by *(i)* its essential role of recovering most high-frequency information of degraded images and *(ii)* the possibility of its accurate prediction for text images via an NN. Moreover, the edge-guided mechanism in Fan *et al.* [39] is called only once in their method. In contrast, the prediction of $\boldsymbol{E_X}$ is called in an iterative scheme derived from the ℓ_0-norm regularization, with different model parameters in each iteration.

Collaborative deep learning also see its applications in other domains; see *e.g.* Wang *et al.*' work [40] in recommender systems. The collaborative mechanisms used in those deep-learning-based recommender systems are related to the collaborative filtering. That is, the individuals in a recommendation system help others in filtering information. In contrast, the collaboration mechanism in our method refers to that the two sub-tasks, *i.e.* the prediction of local statistical measurement on high-frequency components, and the recovery of full frequency spectrum of the image, help each other.

III. Proposed method

A. Framework

The proposed NN for super-resolving blurry text images is outlined in Fig. 4a, which is composed of $T+1$ concatenated modules $\mathcal{M}_0, \cdots, \mathcal{M}_T$. Each module contains a pair of dual CNNs: an $\boldsymbol{E_X}$ prediction CNN (denoted by EP-CNN) that estimates local statistical measurement $\boldsymbol{E_X}$ of the desired image, and a guided recovery CNN (denoted by GR-CNN) that takes the degraded image and the estimated $\boldsymbol{E_X}$ as input for recovering the clear image. More specifically, given a low-resolution blurry text image $\boldsymbol{Y} \in \mathbb{R}^{M_1 \times M_2 \times C}$, the NN generates a sequence of high-resolution deblurred images:

$$\{\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}\} \subset \mathbb{R}^{N_1 \times N_2 \times C}, N_1 > M_1, N_2 > M_2$$

by the modules:

$$\mathcal{M}_0 : \boldsymbol{Y} \to \boldsymbol{X}^{(0)}, \quad \mathcal{M}_t : (\boldsymbol{Y}, \boldsymbol{X}^{(t-1)}) \to \boldsymbol{X}^{(t)}, 1 \le t \le T.$$

The design of $\mathcal{M}_0$ is different from other modules $\mathcal{M}_t$ ($0 < t \le T$), as they have different inputs. It is empirically observed that our NN with three modules (*i.e.* $T = 2$) already yields good results in the experiments, and additional iterations still bring minor improvement.

1) The CNN for predicting $\boldsymbol{E_X}$ (EP-CNN): Regarding the EP-CNN, we have two different designs for the initial module $\mathcal{M}_0$ and the remains respectively. See Fig. 4b and Fig. 4c for two designs. There are two types of $\boldsymbol{E_X}$s involved: $\boldsymbol{S}^{(t)}$ and $\boldsymbol{U}^{(t)}$ ($0 \le t \le T$). The $\boldsymbol{S}^{(t)}$ is used as a reference, which is directly calculated on images without learning, and $\boldsymbol{U}^{(t)}$ is the output of EP-CNN, which is obtained by learning.

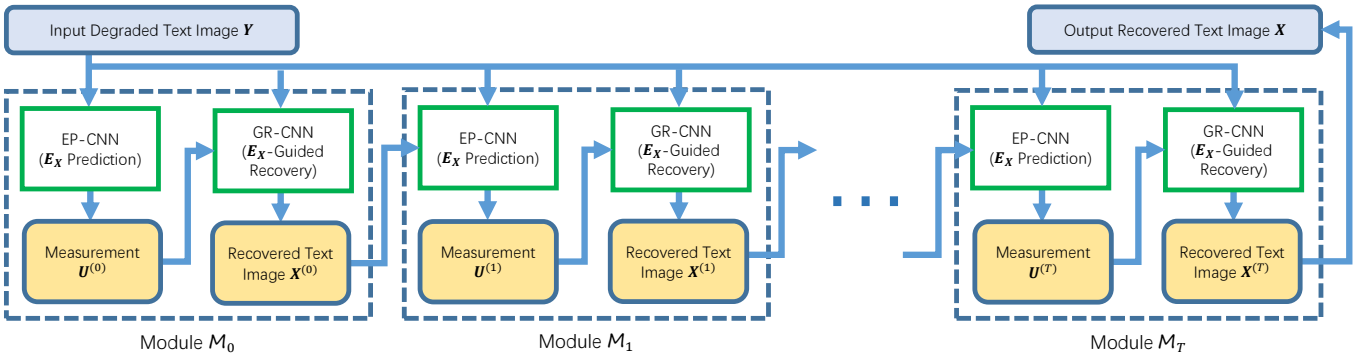
For the module $\mathcal{M}_0$, the EP-CNN is formulated by

$$f_0(\cdot \mid \boldsymbol{\phi}_0) : \boldsymbol{Y} \in \mathbb{R}^{M_1 \times M_2 \times C} \to \boldsymbol{U}^{(0)} \in \mathbb{R}^{N_1 \times N_2}, \qquad (6)$$

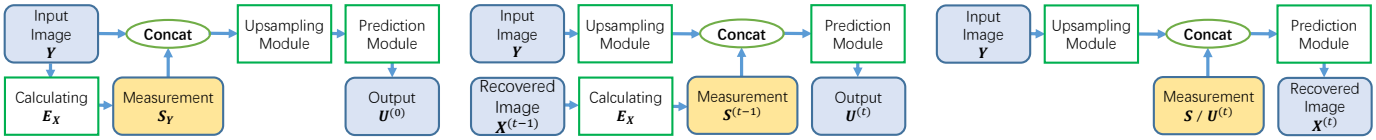
where $\boldsymbol{\phi}_0$ denotes the parameter vector, and $\boldsymbol{U}^{(0)}$ denotes the output $\boldsymbol{E_X}$. The function f_0 first calculates the estimate $\boldsymbol{S_Y} \in \mathbb{R}^{M_1 \times M_2}$ from $\boldsymbol{Y}$ using some predefined high-pass filters. A tensor in $\mathbb{R}^{M_1 \times M_2 \times (C+1)}$ is formed by stacking $\boldsymbol{Y}$ and $\boldsymbol{S_Y}$ together and then fed to an upsampling process:

$$\text{upsampling:} \quad \mathbb{R}^{M_1 \times M_2 \times (C+1)} \to \mathbb{R}^{N_1 \times N_2 \times D},$$

where D denotes the number of convolution kernels in the last layer of the upsampling module, which is set to 64 in



(a) Framework of proposed method.



(b) EP-CNN in Module $\mathcal{M}_0$. (c) EP-CNN in Modules $\mathcal{M}_1, \cdots, \mathcal{M}_T$. (d) GR-CNN in Modules $\mathcal{M}_0, \cdots, \mathcal{M}_T$.

Fig. 4: Framework of proposed method and architectures of EP-CNN and GR-CNN. In (d), the measurement is $\boldsymbol{S}$ calculated from the ground truth image when training GR-CNN separately, and is $\boldsymbol{U}^{(t)}$ predicted by EP-CNN when jointly fine-tuning EP-CNN and GR-CNN.

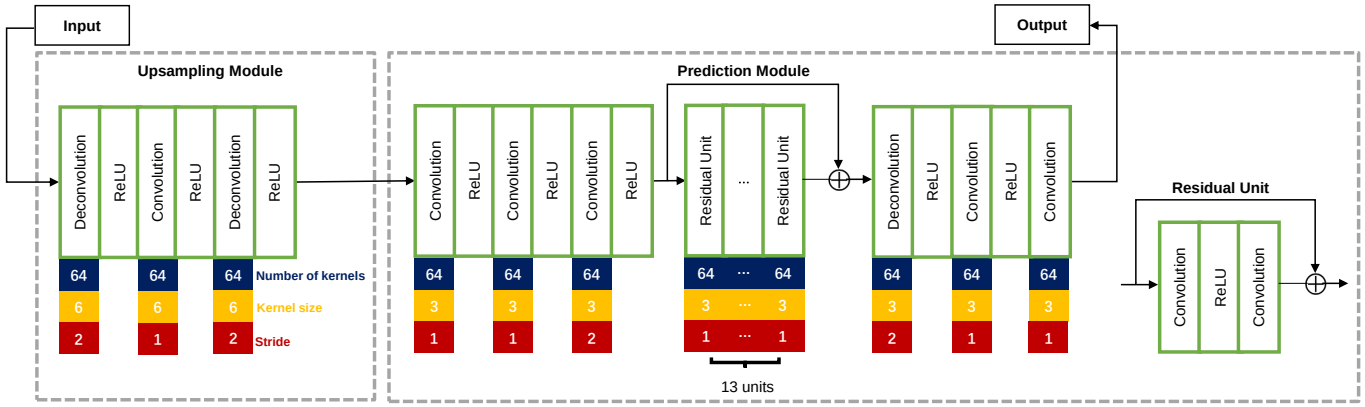


Fig. 5: Illustration of upsampling module (left) and prediction module (right).

our experiments, and $N_1 > M_1, N_2 > M_2$. The output of the upsampling module is then fed to a prediction process:

$$\text{prediction:} \quad \mathbb{R}^{N_1 \times N_2 \times D} \to \mathbb{R}^{N_1 \times N_2},$$

which outputs an estimate $\boldsymbol{U}^{(0)}$.

For the module $\mathcal{M}_t$ with $0 < t \leq T$, there are two inputs: the degraded image $\boldsymbol{Y}$ and an estimate $\boldsymbol{X}^{(t-1)}$. The EP-CNN in $\mathcal{M}_t$ is formulated as the following:

$$f_t(\cdot, \cdot \mid \boldsymbol{\phi}_t) : (\boldsymbol{Y}, \boldsymbol{X}^{(t-1)}) \to \boldsymbol{U}^{(t)}, \tag{7}$$

where $\boldsymbol{\phi}_t$ denotes the parameter vector. The procedure of f_t is as follows. Applying the same pre-defined high-pass filters on $\boldsymbol{X}^{(t-1)}$, we first calculate the estimate $\boldsymbol{S}^{(t-1)}$. Then, we form a tensor in $\mathbb{R}^{N_1 \times N_2 \times (D+1)}$ by stacking $\boldsymbol{S}^{(t-1)}$ and the up-sampled version of $\boldsymbol{Y}$ together. Then the tensor is fed to the prediction module to obtain a new estimate, denoted by $\boldsymbol{U}^{(t)} \in \mathbb{R}^{N_1 \times N_2}$.

The structures of the aforementioned upsampling module and prediction module are illustrated in Fig. 5. The upsampling module sequentially connects a deconvolutional layer, a convolutional layer and a deconvolutional layer. All use 64 convolution kernels with size 6×6, followed by ReLU. The strides of both deconvolutional layers are set to 2 for 4x resolution, *i.e.* $N_1 = 4M_1, N_2 = 4M_2$. The prediction module is much deeper by employing 32 convolutional layers with 64 convolution kernels each. All the layers use kernels of size 3×3, and all the layers except the last one are followed by ReLU. The middle 26 layers are implemented as 13 modified residual units [41] for better performance and convergence [42]. To enlarge the receptive field, a convolutional/deconvolutional layer with stride 2 is used before/after the residual units for downscaling/upscaling [39].

2) The recovery CNN with $\boldsymbol{E_X}$ guidance (GR-CNN): See Fig. 4d for the design of GR-CNN in modules $\mathcal{M}_0, \cdots, \mathcal{M}_T$. With the help from an estimated local statistical measurement on high-frequency components of the ground truth, the image

is recovered by a GR-CNN:

$$g_t(\cdot,\ |\ \boldsymbol{\psi}_t) : (\boldsymbol{Y}, \boldsymbol{U}^{(t)}) \to \boldsymbol{X}^{(t)}, \quad t = 0, 1, \ldots, T. \tag{8}$$

Taking the input degraded image $\boldsymbol{Y} \in \mathbb{R}^{M_1 \times M_2 \times C}$ and the estimate $\boldsymbol{U}^{(t)} \in \mathbb{R}^{N_1 \times N_2}$, GR-CNN first stacks the up-sampled version of $\boldsymbol{Y}$ and $\boldsymbol{U}^{(t)}$ to form a tensor in $\mathbb{R}^{N_1 \times N_2 \times (D+1)}$. The tensor is then input to the prediction module to generate the result $\boldsymbol{X}^{(t)}$. The upsampling and prediction module are the same as those in EP-CNN.

B. Connection to ℓ_0-norm Regularization

The proposed NN can also be interpreted from the view of unrolling the ℓ_0-norm regularization method. Generally, the ℓ_0-norm regularization model for image recovery can be expressed as

$$\min_{\boldsymbol{X}} \|\mathcal{H}_{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{Y}\|_2^2 + \lambda \|\nabla \boldsymbol{X}\|_0, \tag{9}$$

where $\mathcal{H}_{\boldsymbol{\theta}}(\cdot)$ denotes the degradation operator parameterized by $\boldsymbol{\theta}$ and λ is an empirically-selected weight. Based on the pursuit algorithm [43], the above model can be unrolled into the following iterations [21], [22]: $m = 0, 1, \ldots$,

$$\begin{aligned} &\boldsymbol{\Sigma}^{(m+1)}(i,j) = 1 \text{ if } |\nabla \boldsymbol{X}^{(m)}|(i,j) > \tau; \text{ and } 0 \text{ otherwise.} \\ &\boldsymbol{X}^{(m+1)} = \min_{\boldsymbol{X}} \|\mathcal{H}_{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{Y}\|_2^2 + \|(I - \boldsymbol{\Sigma}^{(m+1)}) \odot \nabla \boldsymbol{X}\|_2^2, \end{aligned} \tag{10}$$

where $\tau > 0$ is a small threshold. The first step is to estimate the support $\boldsymbol{\Sigma}^{(m+1)}$ of image gradients $\nabla \boldsymbol{X}$ from $\boldsymbol{X}^{(m)}$. Our method generalizes this process by the EP-CNN and replaces the support by local average of image gradient magnitudes for differentiability. The second step is about solving a ℓ_2-regularized problem with the guidance from the estimated support $\boldsymbol{\Sigma}^{(m)}$, which is analogous to the $\boldsymbol{E}_{\boldsymbol{X}}$-guided recovery mechanism in our GR-CNN. For the recovery with unknown $\boldsymbol{\theta}$, our GR-CNN bypasses the estimation of $\boldsymbol{\theta}$ and directly predicts the truth image.

C. Training

In the proposed NN, each module contains 70 convolutional layers, and the entire network has $70 \cdot T$ convolutional layers. Considering the model size of all modules together, the end-to-end training of the entire NN has very high computational cost. Thus, we choose a sequential and incremental training scheme for balancing the performance gain and the computational efficiency. More specifically, we sequentially train the modules $\mathcal{M}_0, \cdots, \mathcal{M}_T$. Different modules $\mathcal{M}_0, \cdots, \mathcal{M}_T$ do not share parameters with each other. The current module to be trained uses the weights from the previous trained module for initialization.

Let $\{(\boldsymbol{X}_k, \boldsymbol{Y}_k)\}_{k=1}^K$ denote the training image set, where $\boldsymbol{X}_k$ is a truth text image and $\boldsymbol{Y}_k$ is its degraded version with low resolution and blur. Let $\boldsymbol{S}_k$ denote the measurement calculated on $\boldsymbol{X}_k$. In the module $\mathcal{M}_0$, we first train EP-CNN and GR-CNN separately and then jointly fine-tune these two dual CNNs. When training the GR-CNN separately, the ground truth $\boldsymbol{S}_k$ calculated on truth $\boldsymbol{X}_k$ are used, instead of the output of EP-CNN. Let $\boldsymbol{U}_k^{(0)} = f_0(\boldsymbol{Y}_k|\boldsymbol{\phi}_0)$ and $\boldsymbol{X}_k^{(0)} = g_0(\boldsymbol{Y}_k, \boldsymbol{S}_k|\boldsymbol{\psi}_0)$. Then the loss functions $\ell_{f_0}(\cdot)$ and $\ell_{g_0}(\cdot)$ for training the EP-CNN and the GR-CNN respectively are defined by

$$\ell_{f_0}(\boldsymbol{\phi}_0) := \sum_{k=1}^K \|\boldsymbol{U}_k^{(0)} - \boldsymbol{S}_k\|_2^2 + \alpha \|\nabla \boldsymbol{U}_k^{(0)} - \nabla \boldsymbol{S}_k\|_1, \tag{11}$$

$$\ell_{g_0}(\boldsymbol{\psi}_0) := \sum_{k=1}^K \|\boldsymbol{X}_k^{(0)} - \boldsymbol{X}_k\|_2^2 + \beta \|\nabla \boldsymbol{X}_k^{(0)} - \nabla \boldsymbol{X}_k\|_1. \tag{12}$$

In $\ell_{f_0}(\cdot)$ and $\ell_{g_0}(\cdot)$, the first term measures the difference between the output and the ground truth, and the other term measures the discrepancy of their gradients which prevents the CNN favoring blurry outputs [44]. For the joint fine-tuning on EP-CNN and GR-CNN, we denote $\tilde{\boldsymbol{X}}_k^{(0)} = g_0(\boldsymbol{Y}_k, f_0(\boldsymbol{Y}_k|\boldsymbol{\phi}_0)\ |\ \boldsymbol{\psi}_0)$, and the loss function is defined by

$$\begin{aligned} \ell_{(f_0,g_0)}(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0) = \sum_{k=1}^K ((&\|\boldsymbol{U}_k^{(0)} - \boldsymbol{S}_k\|_2^2 + \alpha \|\nabla \boldsymbol{U}_k^{(0)} - \nabla \boldsymbol{S}_k\|_1) \\ &+ \gamma(\|\tilde{\boldsymbol{X}}_k^{(0)} - \boldsymbol{X}_k\|_2^2 + \beta \|\nabla \tilde{\boldsymbol{X}}_k^{(0)} - \nabla \boldsymbol{X}_k\|_1)). \end{aligned} \tag{13}$$

The training of module $\mathcal{M}_t$ $(t = 1, \cdots, T)$ is done as follows. Once the previous modules $\mathcal{M}_0, \cdots, \mathcal{M}_{t-1}$ have been trained, we feed each $\boldsymbol{Y}_k$ to the trained modules and update the estimated of the output image, denoted by $\boldsymbol{X}_k^{(t-1)}$. Let $\boldsymbol{U}_k^{(t)} = f_t(\boldsymbol{Y}_k, \boldsymbol{X}_k^{(t-1)}|\boldsymbol{\phi}_t)$. Then we pre-train the EP-CNN in $\mathcal{M}_t$ with the following loss function:

$$\ell_{f_t}(\boldsymbol{\phi}_t) := \sum_{k=1}^K \|\boldsymbol{U}_k^{(t)} - \boldsymbol{S}_k\|_2^2 + \alpha \|\nabla \boldsymbol{U}_k^{(t)} - \nabla \boldsymbol{S}_k\|_1, \tag{14}$$

For the GR-CNN in $\mathcal{M}_t$, the weights are duplicated from that in the trained $\mathcal{M}_{t-1}$ as pre-training. Afterward, the EP-CNN and GR-CNN in $\mathcal{M}_t$ are jointly trained by minimizing

$$\begin{aligned} \ell_{(f_t,g_t)}(\boldsymbol{\phi}_t, \boldsymbol{\psi}_t) = \sum_{k=1}^K ((&\|\boldsymbol{U}_k^{(t)} - \boldsymbol{S}_k\|_2^2 + \alpha \|\nabla \boldsymbol{U}_k^{(t)} - \nabla \boldsymbol{S}_k\|_1) \\ &+ \gamma(\|\tilde{\boldsymbol{X}}_k^{(t)} - \boldsymbol{X}_k\|_2^2 + \beta \|\nabla \tilde{\boldsymbol{X}}_k^{(t)} - \nabla \boldsymbol{X}_k\|_1)), \end{aligned} \tag{15}$$

where $\tilde{\boldsymbol{X}}_k^{(t)} = g_t(\boldsymbol{Y}_k, f_t(\boldsymbol{Y}_k, \boldsymbol{X}_k^{(t-1)}|\boldsymbol{\phi}_t)\ |\ \boldsymbol{\psi}_t)$.

IV. Experiments

A. Datasets and Training Details

There are few available datasets for joint super-resolution and deblurring on text images [14], and one is proposed by Xu *et al.* [14]. In this dataset, the training set contains over one million 16×16 low-resolution blurry image patches, generated by downsampling the 64×64 blurred patches cropped from the dataset of Hradiš *et al.* [8] with bicubic downsampling by a factor of 4. There are two types of blur in the training data: *(i)* motion blur whose kernel is generated by random walk with kernel size $\in [5, 21]$; and *(ii)* defocus blur implemented by anti-aliased discs with radii uniformly sampled from $[0, 4]$. The Gaussian white noise with *s.t.d.* uniformly sampled from $[0, \frac{7}{255}]$ is then added to the degraded data. The test set contains

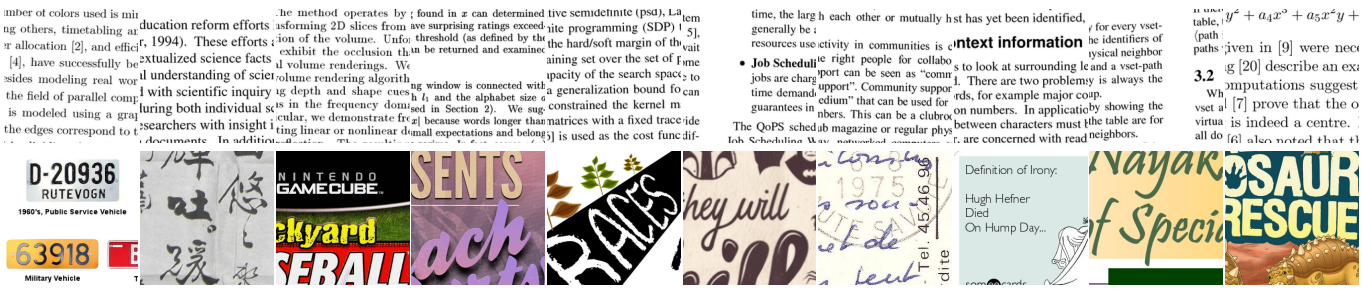


Fig. 6: Sample images in Xu *et al.*'s dataset [14] (top row) and our ComTex dataset (bottom row).

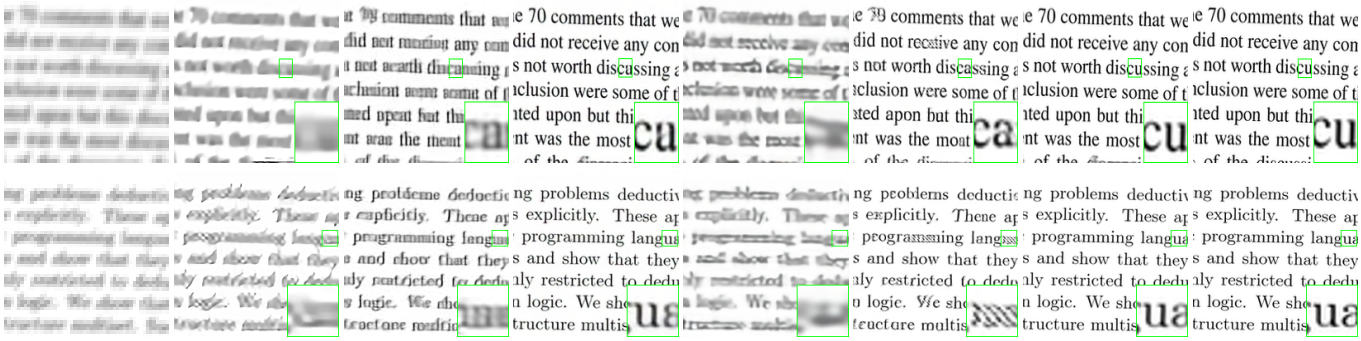


Fig. 7: Visual inspection on the recovery (super-resolution + deblurring) results from Xu *et al.*'s dataset [14].

Fig. 8: Visual inspection on the recovery (super-resolution + deblurring) results from ComTex dataset.

TABLE I: Statistics of ComTex dataset.

# Images	Game Cards	Postcards	Scores	Calligraphies	Posters	License Plates	Covers	Certificates	Slides	Documents	Others	Total
Training	1466	684	625	3965	4871	4796	1964	520	6276	4918	6529	36614
Test	184	108	75	72	46	100	94	90	87	82	200	1138

TABLE II: Performance comparison on Xu *et al.*'s dataset [14].

Method	[48] + [7]	[48] + [46]	[45] + [46]	[45] + [8]	[47] + [16]	GFN [38]	SCGAN [14]	Ours
PSNR (dB)	15.39	15.44	15.54	17.84	24.16	16.57	20.65	**25.16**
SSIM	0.6408	0.6396	0.6651	0.8142	0.9578	0.8078	0.9069	**0.9695**
OCR Acc. (%)	-	-	-	-	-	80.69	90.13	**98.98**

TABLE III: Performance comparison on ComTex dataset.

Method	GFN [38]	SCGAN [14]	Ours
PSNR (dB)	23.52	21.34	**27.64**
SSIM	0.7867	0.7560	**0.8800**

TABLE IV: Comparison on model size and run time.

Method	[47] + [16]	GFN [38]	SCGAN [14]	Ours
Model Size ($\times 10^6$, $\approx$)	45.9	12.2	1.08	2.86
Run Time (s)	0.0562	0.0182	0.0953	0.0252

100 low-resolution blurry text images degraded by the above two types of blur and bicubic downsampling. See Fig. 6 for some sample images in this dataset.

The images in Xu *et al.*'s dataset [14] are all document images with monotone backgrounds. For evaluating our method on a wider range of text images, we constructed a dataset, called *ComTex*, which contains totally 37752 high-resolution clear text images with very diverse types and backgrounds. The construction process of the ComTex dataset is as follows. A large number of images were collected through the Internet using spiders with specific keywords, including cards, scores, calligraphies, posters, license plates, covers, certificates, slides, documents, books, text images and many others. Then, we manually removed those unrelated or low-quality images by visual inspection. The remaining ones are used as the high-resolution clear text images in the dataset.

A fixed training/test split is used in ComTex for the reproducibility of the results. An image subset are randomly picked up from each class to form the test set, and the rest are used as the training set. The resulting training set and test set contain 36614 and 1138 images respectively. The images have varying sizes. For the experiments, the degraded training images are generated by using the blur kernels from Hradiš *et al.* [8] and resized by bicubic downsampling, as well as additive Gaussian white noise with *s.t.d.* uniformly sampled from $[0, \frac{7}{255}]$. In training, over one million 200×200 patches are cropped from the training images and downsampled those patches with bicubic downsampling by a factor of 4. For tuning the hyperparameters, we randomly separated the validation data from the training data of ComTex. Following the same treatment of dataset in Xu *et al.* [14], the test images are generated by using the blur kernels from Hradiš *et al.* [8] and resized by bicubic downsampling. The test images are cropped into over 40000 patches of size 200×200 and those patches are downsampled with bicubic downsampling by a factor of 4 for evaluation. See Table I for the statistics of our ComTex dataset. See also Fig. 6 for some sample images.

All of our models are trained using the Adam optimizer [49]. The initial learning rate and mini-batch size are set to $1e^{-4}$ and 32 respectively when training EP-CNN and GR-CNN separately, and set to $1e^{-5}$ and 16 respectively when fine-tuning the entire network. The learning rate is decayed along with iterations. The weights in the loss functions are $\alpha = 0.5$, $\beta = 2$ and $\gamma = 2.5$.

B. Results

1) Results on Xu et al.'s dataset [14]: The proposed method is compared to three approaches in terms of PSNR, SSIM and OCR accuracy. The first one is the combination of two state-of-the-art methods on text super-resolution [45] and image deblurring [8] with fine-tuning. The remaining two are the most-related ones that jointly conduct super-resolution and deblurring. One is the Xu *et al.*'s SCGAN method [14] which uses a CNN+GAN framework for super-resolving blurry images. The second one is Zhang *et al.*'s GFN method [38] which trains an end-to-end network to recover sharp high-resolution images from the degraded ones. Both methods have their models available, and we fine-tuned the model of GFN on text images as it is originally designed for general images. In addition, we cite the results from Xu *et al.* [14] on some combinations of the deblurring and super-resolution methods for comparison. We also combined the recent super-resolution (Lim *et al.* [47]) and deblurring (Tao *et al.* [16]) methods and fine-tuned it on the dataset for comparison. Same as Hradiš *et al.* [8], the OCR accuracy is computed by first using the recognition function of ABBYY FineReader 12 on the results and then calculating the mean character accuracy.

See Table II for the comparison of all the methods on Xu *et al.*'s dataset [14]. Clearly, our method noticeably outperforms all others, which indicated the effectiveness of our method in super-resolving blurry document images. See Fig. 7 for visual comparison. It can be seen that as expected, sequentially running deblurring (Hradiš *et al.* [8]) and super-resolution (Kim *et al.* [45]), does not produce satisfactory results. Similar phenomenon has also been observed by Xu *et al.* [14]. The performance of GFN is not satisfactory either, as it is originally

designed for natural images. For SCGAN which is specifically designed for text images, the recovered text contents can be easily dominated by the checkerboard effects, resulting in poor visual quality. For comparison, we also list the model size (*i.e.* number of parameters) and average run time per image of different methods in Table IV. The run time is recorded on mapping a 50×50 low-resolution blurry text image into a 200×200 high-resolution clear one. The computational environment is a PC with Intel i7-8700K CPU and NVIDIA TITAN XP GPU. Note that the compared methods are implemented with different deep learning platforms. Thus, the run time depends on not only the model size but also the configurations of both hardware and software platform.

2) Results on ComTex dataset: With its diversity on image types and complex backgrounds, the ComTex dataset is more challenging than Xu *et al.*'s [14]. The proposed method is only compared to two most-related CNN-based ones: Xu *et al.*'s [14] and Zhang *et al.*'s [38]. These two methods for comparison are evaluated in two ways: *(i)* re-training its model with the new data, and *(ii)* fine-tuning the model pre-trained in the original work; and the best result is reported. See Table III for the comparison. It can be seen that our method again outperforms the other two by a large margin in terms of PSNR and SSIM. See Fig. 8 for the visual illustration of some results. Clearly, the images generated by our method are of better visual quality.

3) Results on real degraded images: Note that both datasets above are composed of synthetic degraded images. We also evaluate the performance of recovering real degraded images when using the model of our method trained on the synthetic images. See Fig. 9 for the visualization of some results. It can be seen that compared to other methods, ours has better generalization performance and its recovered images have higher visual quality. Meanwhile, we note that there is still room for improvement on handling significant spatially-varying blur.

C. Ablation Study

1) Influence of number of modules: It is interesting to examine how our method performs when using different numbers of modules on Xu *et al.*'s dataset [14]. See Table V for the results of such a study. It can be seen that the results of the first iteration of our method already show significant improvement over the compared methods, while the other two iterations further improve the results. It can also be seen that the improvement becomes much less when using more than 3 modules. In Fig. 10, we show an example of the measurements $\boldsymbol{E_X}$ predicted by EP-CNN in different iterations. It can be seen that the intermediate $\boldsymbol{E_X}$s, as well as the intermediate recovery results, are of better quality with more details, when passing through more modules.

2) Effectiveness of collaborative deep learning: The introduction of collaborative learning is the key to the success of the proposed method. The local statistical measurement on image gradients of text image is easier to predict, yet it provides sufficient information to help the recovery of full spectrum of a text image. To show the gain from such a collaborative mechanism, another version of the proposed method is constructed which removes the EP-CNNs. To further eliminate the influence of model size, we retrain a GR-CNN-70 whose depth is the same as each module of our CNN (*i.e.* 70 layers) with similar model size. The GR-CNN-70 only takes $\boldsymbol{Y}$ as input. In Table VI, we compare the performance of GR-CNN-70 with the module $\mathcal{M}_0$ of our method. The performance of $\mathcal{M}_0$ is much better than that of GR-CNN-70. Such results have demonstrated the benefits of using $\boldsymbol{E_X}$ to guide the recovery process. When provided with an additional input $\boldsymbol{E_X}$ (*i.e.* $\boldsymbol{S_Y}$) from $\boldsymbol{Y}$ for GR-CNN-70, there is no noticeable improvement observed, which is probably because the measurement is not learned/refined and thus erroneous.

3) Influence of model size: To see how our method performs when using different model sizes, we change the depth (*i.e.* number of layers) and the width (*i.e.* number of kernels in each convolutional layer) of $\mathcal{M}_0$ to generate the models with different sizes, and the resulting models are tested on on Xu *et al.*'s dataset [14]. See Table VII for the results. It can be seen that, when the model size is nearly halved, the performance of our model drops noticeably (around 2.3dB). This is not surprising as the expressibility of our network is reduced too much in such cases.

D. More Analysis on Our Method

1) Performance of text image deblurring: Though our focus is to super-resolve blurry images, our NN can also be applied to the deblurring on text images by simply removing the

TABLE V: Performance comparison of different modules on Xu *et al.*'s dataset [14] and ComTex dataset.

Dataset	# Module	PSNR(dB)	SSIM
Xu *et al.*'s dataset [14]	1	24.85	0.9636
	2	25.06	0.9670
	3	25.16	0.9695
	4	25.18	0.9695
ComTex	1	27.14	0.8725
	2	27.46	0.8768
	3	27.64	0.8800
	4	27.68	0.8801

TABLE VI: Performance comparison of GR-CNN-70 and our method ($\mathcal{M}_0$) on Xu *et al.*'s dataset [14] and ComTex dataset.

Dataset	Method	PSNR(dB)	SSIM
Xu *et al.*'s dataset [14]	GR-CNN-70	23.53	0.9517
	Ours ($\mathcal{M}_0$)	**24.85**	**0.9636**
ComTex	GR-CNN-70	25.48	0.8595
	Ours ($\mathcal{M}_0$)	**27.14**	**0.8725**

TABLE VII: Performance comparison of using our method ($\mathcal{M}_0$) with different model sizes on Xu *et al.*'s dataset [14].

Model Size ($\times 10^6$, $\approx$)	# Filters Each Layer	# Layers	PSNR (dB)	SSIM
2.86	64	70	24.85	0.9636
1.39	64	30	22.55	0.9373
1.62	48	70	23.97	0.9547

Input GFN [38] SCGAN [14] Ours

Fig. 9: Visual results on real text images. The first two images are from Xu *et al.* [14], and the others are collected by the authors.

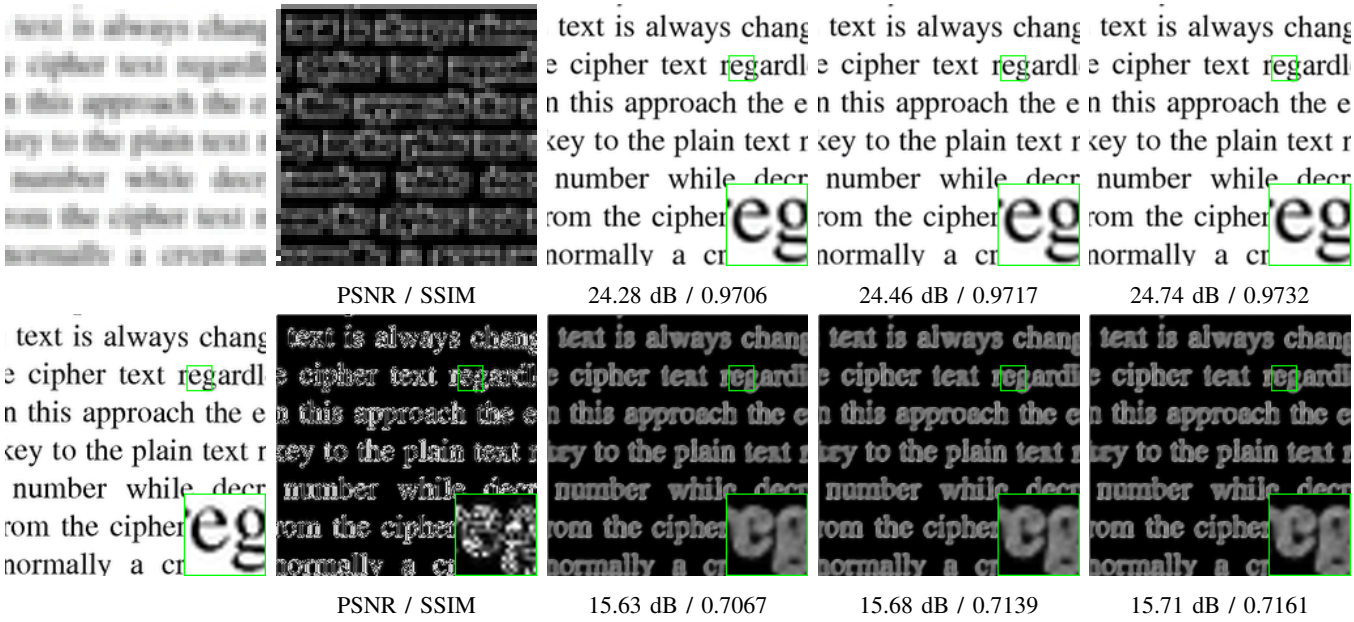
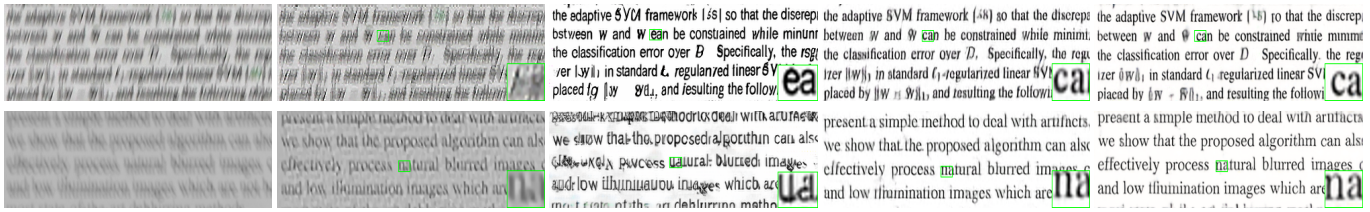


Fig. 10: First row: from left to right are input image, the measurement $\boldsymbol{E_X}$ of input image, and the images output at the 1^{st}, 2^{nd} and 3^{rd} iterations respectively. Second row: truth image, truth $\boldsymbol{E_X}$, the $\boldsymbol{E_X}$ output at 1^{st}, 2^{nd} and 3^{rd} iterations respectively.

TABLE VIII: Average PSNR(dB) and SSIM of restored images on Hradiš *et al.*'s dataset [8].

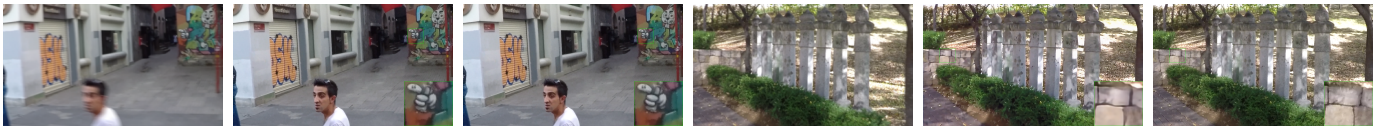
Method	Metric	Standard Deviation of Gaussian Noise							
		0	1	2	3	4	5	6	7
Pan *et al.* [7]	PSNR	15.76	15.73	15.74	15.57	15.33	14.57	14.09	13.38
	SSIM	0.7231	0.7194	0.7173	0.6968	0.6564	0.5514	0.4446	0.3448
Hradiš *et al.* [8]	PSNR	23.99	23.97	23.94	23.87	23.78	23.66	23.53	23.38
	SSIM	0.9476	0.9474	0.9467	0.9454	0.9438	0.9417	0.9391	0.9362
Lu *et al.* [50]	PSNR	17.03	17.03	17.01	16.99	16.97	16.93	16.89	16.85
	SSIM	0.6905	0.6881	0.6811	0.6697	0.6546	0.6367	0.6168	0.5957
Ours	PSNR	**27.13**	**27.14**	**27.11**	**27.05**	**26.96**	**26.83**	**26.70**	**26.56**
	SSIM	**0.9815**	**0.9813**	**0.9808**	**0.9799**	**0.9786**	**0.9769**	**0.9748**	**0.9725**
		8	9	10	11	12	13	14	15
Pan *et al.* [7]	PSNR	12.88	12.28	11.88	11.46	11.14	10.86	10.53	10.27
	SSIM	0.2813	0.2352	0.2106	0.1895	0.1773	0.1695	0.1565	0.1471
Hradiš *et al.* [8]	PSNR	23.21	23.03	22.83	22.63	22.40	22.17	21.93	21.68
	SSIM	0.9328	0.9289	0.9244	0.9192	0.9131	0.9061	0.8980	0.8886
Lu *et al.* [50]	PSNR	16.80	16.74	16.68	16.62	16.56	16.49	16.42	16.34
	SSIM	0.5740	0.5523	0.5310	0.5102	0.4904	0.4714	0.4535	0.4365
Ours	PSNR	**26.41**	**26.26**	**26.09**	**25.92**	**25.75**	**25.56**	**25.38**	**25.20**
	SSIM	**0.9699**	**0.9671**	**0.9641**	**0.9609**	**0.9575**	**0.9541**	**0.9503**	**0.9465**

upsampling module. We give some demos of our method on text image deblurring. We use the dataset from Hradiš *et al.* [8] for training and testing. The training set contains 67742 blurry image (patches) of size 300×300, with two types of blur including realistic motion blur due to camera shake and defocus blur. The motion blur is generated by random walk with kernel size sampled from $[5, 21]$, and the defocus blur is generated by anti-aliased disc with radius uniformly sampled from $[0, 4]$. The Gaussian noise with standard deviation uniformly sampled from $[0, \frac{7}{255}]$ is then added. The test set contains one hundred 200×200 image (patches), which are blurred by different blur kernels that are generated by the above scheme. Following the experimental settings in Hradiš *et al.*'s method [8], the white Gaussian noises with standard deviations sampled in $[0, \frac{15}{255}]$ are added to the test data. Our method is compared with Hradiš *et al.*'s method [8] which is a CNN-based method for blind image deconvolution and denoising, Pan *et al.*'s method [7] which use a simple yet effective L0-regularized prior on intensities and gradients of text images, and Lu *et al.*'s method [50] which is a latest unsupervised powerful approach for domain-specific single-image deblurring. The comparison is in terms of the average PSNR and SSIM computed on the text set of Hradiš *et al.* [8]. The results are listed in Table VIII, where the results of Pan *et al.*'s method are cited from Hradiš *et al.* [8]. See Fig. 11 for the comparison on real blurry text images. It can be seen that our method outperforms other compared methods, in terms of both PSNR and SSIM.



Input Pan *et al.* [7] Lu *et al.* [50] Hradiš *et al.* [8] Ours

Fig. 11: Visual results of deblurring on real blurry text images from Hradiš *et al.* [8].



Input Ours Ground Truth Input Ours Ground Truth

Fig. 12: Visual results of recovery (super-resolving + deblurring) on LR-GOPRO dataset [38].

2) How about recovery of natural images?: The success of the NN largely depends on how accurately the trained module can predict $\boldsymbol{E_X}$. The statistical distribution of image gradients of text images is much simpler, making it possible to learn the function of predicting $\boldsymbol{E_X}$ accurately. However, for natural images with textures, especially texture regions, the complex statistical characteristics of local image gradients make the prediction a much hard task. To see this, we train and test our NN on the LR-GOPRO dataset from Zhang *et al.* [38], a dataset generated from the original GOPRO dataset proposed by Nah *et al.* [41]. The performance of our method is not as good as Zhang *et al.* [38]. See Fig.12 for some results on the LR-GOPRO dataset [38].

V. Summary

In this paper, we proposed a CNN-based approach with a collaborative mechanism on high-frequency information and full spectrum for joint deblurring and super-resolution on degraded text images. The collaborative mechanism is motivated from the special characteristics of text images and the ℓ_0-norm related regularization. The experiments on both document images and general text images showed the noticeable improvement of the proposed method over the state-of-the-art ones, in terms of both quantitative metric and visual quality. Also, the paper contributed a new dataset of general text images which can benefit the development of learning-based methods related to text images. In addition to text image processing, the collaborative mechanism introduced in the proposed method has the potential to be used in other image recovery tasks.

Owing to very different characteristics between text images and natural images, the performance of the proposed method on the images of natural scenes is not as good as that on text images. Also, there is still a lot of room for improvement when processing real degraded text images with complex backgrounds. In future, we would like to investigate how to further improve the performance of the proposed method when processing complex real text images, as well as real images of natural scenes.

References

[1] S. Lee, M. S. Cho, K. Jung, and J. H. Kim, "Scene text extraction with edge constraint and text collinearity," in *Proc. Int. Conf. on Pattern Recognition*, 2010, pp. 3983–3986.

[2] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*. IEEE, 2012, pp. 3538–3545.

[3] C. Yi and Y. Tian, "Assistive text reading from complex background for blind persons," in *Proc. Int. Workshop Camera-Based Document Anal. Recognition*. Springer, 2011, pp. 15–28.

[4] Y. Lou, A. L. Bertozzi, and S. Soatto, "Direct sparse deblurring," *J. Math. Imaging Vision*, vol. 39, no. 1, pp. 1–12, 2011.

[5] X. Chen, X. He, J. Yang, and Q. Wu, "An effective document image deblurring algorithm," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2011, pp. 369–376.

[6] H. Cho, J. Wang, and S. Lee, "Text image deblurring using text-specific properties," in *Proc. European Conf. Comput. Vision*. Springer, 2012, pp. 524–537.

[7] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via l0-regularized intensity and gradient prior," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2014.

[8] M. Hradiš, J. Kotera, P. Zemcık, and F. Šroubek, "Convolutional neural networks for direct text deblurring," in *Proc. British Mach. Vision Conf.*, 2015.

[9] J. Park, Y. Kwon, and J. H. Kim, "An example-based prior model for text image super-resolution," in *Proc. Int. Conf. Document Anal. Recognition*. IEEE, 2005, pp. 374–378.

[10] R. Walha, F. Drira, F. Lebourgeois, and A. M. Alimi, "Super-resolution of single text image by sparse representation," in *Proc. Workshop Document Anal. Recognition*. ACM, 2012, pp. 22–29.

[11] A. Abedi and E. Kabir, "Text-image super-resolution through anchored neighborhood regression with multiple class-specific dictionaries," *Signal, Image and Video Process.*, vol. 11, no. 2, pp. 275–282, 2017.

[12] ——, "Text image super resolution using within-scale repetition of characters and strokes," *Multimedia Tools Appl.*, vol. 76, no. 15, pp. 16 415–16 438, 2017.

[13] A. V. Nasonov and A. S. Krylov, "Text images superresolution and enhancement," in *Proc. IEEE Int. Congr. Image Signal Process.*, 2012, pp. 728–731.

[14] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct 2017.

[15] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.

[16] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.

[17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.
[18] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.
[19] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 2008.
[20] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.
[21] S. Nam, M. Davies, and M. Elad, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30–56, 2013.
[22] H. Ji, Y. Luo, and Z. Shen, "Image recovery via geometrically structured approximation," *Appl. Comput. Harmon. Anal.*, vol. 41, no. 1, pp. 75–93, 2016.
[23] S. Cho and S. Lee, "Fast motion deblurring," in *ACM Siggraph Asia*, 2009, pp. 1–8.
[24] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. European Conf. Comput. Vision*, 2010, pp. 157–170.
[25] B. Xin, Y. Wang, W. Gao, D. Wipf, and B. Wang, "Maximal sparsity with deep networks?" in *Advances in Neural Info. Process. Syst.*, 2016, pp. 4340–4348.
[26] Y. Yang, J. Sun, H. LI, and Z. Xu, "Admm-csnet: A deep learning approach for image compressive sensing," *IEEE Trans. on Pattern Anal. Mach. Intell.*, pp. 1–1, 2018.
[27] J. Sun, H. Li, Z. Xu *et al.*, "Deep admm-net for compressive sensing mri," in *Advances in Neural Info. Process. Syst.*, 2016, pp. 10–18.
[28] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph." *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, 2006.
[29] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–10, 2008.
[30] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.
[31] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.
[32] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2018.
[33] X. Xu, Y. Ma, and W. Sun, "Towards real scene super-resolution with raw images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2019, pp. 1723–1731.
[34] N. Joshi, R. Szeliski, and D. J. Kriegman, "Psf estimation using sharp edge prediction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*. IEEE, 2008, pp. 1–8.
[35] S. Harmeling, S. Sra, M. Hirsch, and B. Schölkopf, "Multiframe blind deconvolution, super-resolution, and saturation correction via incremental em," in *Proc. IEEE Int. Conf. Image Process.* IEEE, 2010, pp. 3313–3316.
[36] C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2011, pp. 209–216.
[37] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *Proc. IEEE Int. Conf. Comput. Vision*, December 2013.
[38] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for joint image deblurring and super-resolution," in *Proc. British Mach. Vision Conf.*, 2018.
[39] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct 2017.
[40] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*. ACM, 2015, pp. 1235–1244.
[41] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, July 2017.
[42] Q. Fan, D. P. Wipf, G. Hua, and B. Chen, "Revisiting deep image smoothing and intrinsic image decomposition," *CoRR abs/1701.02965*, vol. 2, 2017.
[43] G. M. Davis, S. G. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions with matching pursuit," in *Wavelet Applications*, vol. 2242. International Society for Optics and Photonics, 1994, pp. 402–413.
[44] T. Narihira, M. Maire, and S. X. Yu, "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression," in *Proc. IEEE Int. Conf. Comput. Vision*, December 2015.
[45] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2016.
[46] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Blind image deblurring using dark channel prior," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, June 2016.
[47] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. Conf. Comput. Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.
[48] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vision*, December 2015.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[50] B. Lu, J.-C. Chen, and R. Chellappa, "Unsupervised domain-specific deblurring via disentangled representations," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2019, pp. 10 225–10 234.

Yuhui Quan received the Ph.D. degree in Computer Science from South China University of Technology in 2013. He worked as the postdoctoral research fellow in Mathematics at National University of Singapore from 2013 to 2016. He is currently the associate professor at School of Computer Science and Engineering in South China University of Technology. His research interests include image processing, sparse representation and deep learning.

Jieting Yang received the B.Eng degree in Network Engineering from South China University of Technology in 2018. She is currently a M.A candidate in South China University of Technology. Her research interests include computer vision, image processing, and sparse coding.

Yixin Chen received the B.Eng degree in Network Engineering from South China University of Technology in 2017. He is currently a M.A candidate in South China University of Technology. His research interests include computer vision, image processing, and sparse coding.

Yong Xu received the B.S., M.S., and Ph.D. degrees in mathematics from Nanjing University, Nanjing, China, in 1993, 1996, and 1999, respectively. He was a Post-Doctoral Research Fellow of computer science with South China University of Technology, Guangzhou, China, from 1999 to 2001, where he became a Faculty Member and where he is currently a Professor with the School of Computer Science and Engineering. His current research interests include image analysis, video recognition, and image quality assessment. Dr. Xu is a member of the IEEE Computer Society and the ACM.

Hui Ji received the B.Sc. degree in Mathematics from Nanjing University in China, the M.Sc. degree in Mathematics from National University of Singapore and the Ph.D. degree in Computer Science from the University of Maryland, College Park. In 2006, he joined National University of Singapore as an assistant professor in Mathematics. Currently, he is an associate professor in mathematics at National University of Singapore. His research interests include computational harmonic analysis, optimization, computational vision, image processing and biological imaging.