Journal of Real-Time Image Processing (2020) 17:1787–1800
https://doi.org/10.1007/s11554-020-00976-x

SPECIAL ISSUE PAPER

Fast simultaneous image super-resolution and motion deblurring with decoupled cooperative learning

Heng Liu[1] · Jiajun Qin[1] · Zilin Fu[1] · Xue Li[1] · Jungong Han[2]

Received: 2 December 2019 / Accepted: 9 April 2020 / Published online: 11 May 2020

Abstract
In recent years, deep convolutional neural networks (CNNs) have been widely applied to handle low-level vision problems. However, most existing CNN-based approaches can either handle single degeneration each time or treat them jointly through feature entangling, thus likely leading to poor performance when the actual degradation is inconsistent with hypothetical degradation condition. Furthermore, feature coupling will bring a large amount of computation, which may make the methods impractical to real-time mobile scenarios. In order to address these problems, we propose a deep decoupled cooperative learning model which can not only develop the corresponding recover network to deal with each degradation, but also flexibly handle multiple degradations at the same time. Thus, our approach can achieve disentangling and synthesizing single image super-resolution and motion deblurring, which has high practicability. We evaluate the proposed approach on various benchmark datasets, covering both natural images and synthetic images. The results demonstrate its superiority, compared to the state-of-the-art, where image SR and motion deblurring can be accomplished effectively concurrently. The source code of the work is available at https://github.com/hengliusky/Cooperative-Learning-Deblur-SR.

Keywords Image super-resolution · Motion deblurring · Decoupled cooperative learning

1 Introduction

As is well known, resolution reduction and motion blurring are two main manifestations of image degradation, in which the former is usually caused by down-sampling, while the latter arises when the image recorded within a single exposure changes due to rapid movement. In contrast to such two degradation processes, image super-resolution (SR) and motion deblurring are the corresponding reverse processes—reconstructing high resolution (HR) images from low resolution (LR) counterparts, and recovering sharp images from blurred ones.

For single image SR (SISR), the goal of super-resolving a low-resolution (LR) image is to recover the missing high-frequency details in the original HR image. In theory, a typical resolution degradation model consists of a serial of degenerating operations, such as smoothing, down-sampling and adding additive noise, which can be denoted as:

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

where x is the HR image and y is the LR image, $*$ represents the convolution (blur) operator, k denotes the blur (smoothing) kernel, $\downarrow_s$ indicates a down-sampling operator with factor s, and n refers to additive white Gaussian noise. Because the denoising task in the low-level vision can be handled individually, most existing image SR works usually ignored the involved noise, that is, $n = 0$ in the equation.

Instead of imposing classical image priors (e.g., sparsity prior [32], non-local prior [4] and pixel consistency prior [34]), recent deep learning-based methods, especially convolutional neural networks (CNNs)-based SISR [2, 3], pay more attention to obtaining end-to-end implicitly embedded mappings from LR image patches to their corresponding HR ones. Thanks to its favorable performance in terms of effectiveness and convenience, CNNs-based SISR approach has attracted considerable attention.

Similar to image SR, image deblurring is also an ill-conditioned problem. Ignoring the nonlinear camera response

✉ Heng Liu
hengliusky@gmail.com

[1] School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China

[2] Warwick Manufacturing Group, The University of Warwick, Conventry CV4 7AL, UK

function (CRF) effect, the degeneration model of image blur can be described as:

$$y = (x * h) + n. \tag{2}$$

Here, y is the blur observation, x is the sharp image, h is the blur kernel (also named point spread function, PSF) and n again represents the noise model. The general objective of image deblurring is to recover the sharp image x directly (such deblurring is called blind deblurring), or to get the blur kernel h first and then to recover the image x (named as non-blind one in this instance) from the blur observation y. Owning to no down-sampling operation, the process of image blurring seems to be simpler than that caused by resolution reduction. Actually, unlike the resolution degradation where the smoothing kernel is usually uniform and linear, the blur one is likely non-uniform and non-linear, which makes the estimate of the accurate kernel for image deblurring difficult. However, as a typical form of blur, motion blur is usually modeled by uniform linear motion. Recently, several CNNs based works [15, 21, 22] have emerged for blind deblurring.

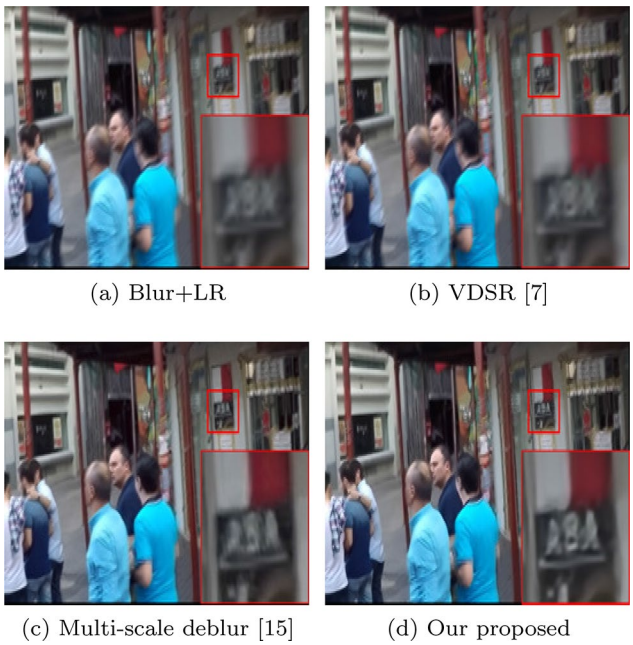
In practice, image degradation is more like a combination of multiple degradation factors. Thus, conducting image SR or motion deblurring alone does not help that much. In addition, it is not clear whether the image recovery way suitable for one degradation can still work for another degeneration. Especially in a mobile scenario, due to the limited computation capacity, there is an urgent demand for a flexible method that is able to achieve image SR and motion deblurring efficiently. A possible idea is to decouple the composite reconstruction task into the direct summation of corresponding simple independent recovery sub-tasks, which can easily meet the real-time or fast requirements in mobile applications.

In virtue of these issues, the following questions need to be investigated: (1) can a single-factor recovery CNN model be easily and effectively extended to handle multiple factors degradation (such as resolution reduction and motion blur) simultaneously? Although the recent work of Zhang et al. [36] has proposed a gated fusion CNN approach for joint image SR and deblurring, their model is elaborately designed with complex feature coupling structure, which means each recovery sub-task is not independent at the same time. Thus, the second question should be further investigated: (2) Can the learned joint image SR and motion deblurring model be easily decomposed into corresponding independent sub-task networks, i.e., is the system able to acquire the independent decoupled features of the sub recovery tasks while achieving them at the same time? This work aims to give a preliminary attempt to answer these two questions.

To do so, we begin with the discussion of two widely used degeneration models of image resolution reduction and motion blur. By analyzing the process of multiple degenerations, we are conscious of that the degradation actions of different order actually give rise to diversified comprehensive degeneration effects. This implies that it might be difficult and computationally expensive, if not impossible, to design a single convolutional recovery model for general purpose in the existence of either completely mismatched or mixed multiple degradation actions.

In view of this, by analyzing the maximum likelihood estimate (MLE) solution for simultaneous SISR and motion deblurring, we find that the estimation of the HR and sharp image can be treated as the process of decoupling cooperative learning. Thus, this acknowledgment becomes our guide for designing the multiple degradations recovery model. The benefits of employing such a decoupling cooperative learning enabled CNN model typically lie in: (1) adapting to the multiple degradations combination in any order and (2) maintaining the independence of sub-tasks. An example of a simultaneous super-resolved and deblurred image recovered by the proposed model from compound 4× down-sampling and motion blur is illustrated and compared to VDSR SR [7] and multi-scale deblur [15] in Fig. 1.

To the best of our knowledge, the attempts dedicated to handle both resolution reduction and motion blurring via CNNs are few, needless to say, considering the independence of each sub-task in one model. The main contributions of this work are summarized in the following:



(a) Blur+LR (b) VDSR [7]

(c) Multi-scale deblur [15] (d) Our proposed

Fig. 1 The details in the super-resolved (4×) and deblurred image produced by the proposed model (right bottom) are much sharper than the ones produced by VDSR [7] (right upper) SR and multi-scale deblur (left bottom) [15]

- We propose a decoupled cooperative learning-based end-to-end CNN model for simultaneous image SR and motion deblurring. The proposed model gets out of the unrealistic assumption that only one single type of degeneration exists, and adapts our model to different types of degenerate images with the aid of the decoupling and cooperative learning. This seems to be a better solution toward developing a CNN-based image clearer for real-time applications.
- We give the rationality of the proposed approach through analyzing the MLE estimation of the multiple degenerates equivalence. In addition, through in-depth analysis on the effect of each sub-task network of the overall model, we demonstrate the advantages of the proposed decoupled cooperative learning approach: It not only can accomplish joint tasks, but also is capable of completing each sub-task independently when necessary.
- We verify and evaluate our proposed model not only on the widely recognized public dataset, but also on the dataset that is completely different from training images, e.g., synthetic images. We show that through decoupled cooperative learning our approach is able to produce the competitive results against the state-of-the-art SISR and motion deblurring methods on natural LR and blurred images. More importantly, it gives rise to visually plausible results on synthetic non-uniform blurred and LR images.

2 Related works

Recent deep learning ways, especially CNNs, have succeeded from middle–high-level tasks such as saliency detection [24, 31] to traditional low-level vision tasks such as denoising [17]. While for SISR, pioneering CNN-based method is SRCNN [2, 3], which used a three-layer convolutional network to learn the mapping from LR images to the HR ones. Following the way of SRCNN but increasing the depth, Kim et al. [7] proposed a very deep SR network (VDSR) with residual connection which provides a significant performance improvement. Noting that no image priors were used in the previous CNN-based SR methods, Yang et al. [33] integrated an explicit Sobel edge prior with HR images to jointly supervise the learning of their so-called recurrent CNN. Unlike Dong's [2, 3] way of making LR input image bi-cubic interpolation to the size of HR image, Shi et al. [19] directly learned up-scaling filters with the proposed sub-pixel convolutional layer to up-sample the LR feature maps into HR images. Aiming to improve the perceptual quality, Ledig et al. [9] proposed a generative adversarial network [5]-based image SR (SRGAN) method, in which two sub-pixel convolution layers were used to upscale the LR input efficiently. In addition to the blur kernel, recently Zhang et al. [35] considered the effect of additional noise on image resolution reduction separately and fully utilized the great number of LR images generated based on various size blur kernels and multiple level noise to learn a CNN model for robust image SR. Moreover, noticing that the hierarchical features from all convolution layers are not well utilized in SR, Zhang et al. [38] proposed residual dense block (RDB) to extract abundant local features and took global fusion to learn holistic hierarchical features.

CNN also performs well for motion deblurring. Xu et al. [27] made use of CNN to recover the blurred images in a non-blind setting. Sun et al. [21] parameterized the non-uniform motion blur kernel and estimated it at the patch level through a deep layered architecture. Schuler et al. [18] firstly used a CNN network to mimic a classical optimization-based deblurring process. Then, recently, Li et al. [10] treated the CNN learning as the latent process of maximum a posterior (MAP) and then utilized a MAP framework to deblur the degraded images blindly. Nah et al. [15] adopted an end-to-end multi-scale CNN to predict the latent blur-free images. Based on this, Tao et al. [22] proposed a scale-recurrent motion deblurring network with a simpler structure and fewer parameters.

Only a few works enable to exploit CNNs for concurrent SISR and motion deblurring. Xu et al. [29] addressed super-resolve blurry faces through GAN. However, their work was limited to face images and training GAN to achieve good performance is not a trivial task in practical applications. Zhang et al. [37] presented to use a deep encoder–decoder network to perform joint image SR and deblurring. Very recently, Zhang et al. [36] again put forward a gated fusion network for joint image deblurring and super-resolution. Despite reported good performance, the part of image SR in their network is coupled with the deblurring part, implying that their method may not be able to perform image SR or motion deblurring individually.

In general, recent works on joint SISR and motion deblurring are limited. Actually, there appears a desire for a CNN style model that not only can handle image SR and deblurring simultaneously, but also can deal with any task of them independently.

3 Methodology

3.1 Degradation model

For SISR and motion deblurring, we usually use the below degenerate models to induce the CNN-based image recovery, which can be given separately as

$$y = (x\downarrow_s)\uparrow_s + n, \tag{3}$$

$$y = \left(\sum_{i=1}^{N} x_i\right)/N + n. \tag{4}$$

Here, Eq. 3 denotes the process of spatial sampling-based resolution reduction, while Eq. 4 represents a typical motion blur accumulation process given the sampled sharp frames. Here, x in Eq. 3 is the HR image, N and x_i in Eq. 4 are the number of sampled frames and the ith sharp frame image captured during the exposure time, y is either the LR or motion-blurred image, n models the additional noise (usually is white Gaussian noise), $\downarrow_s$ and $\uparrow_s$ are down-sampling and up-sampling operators with scale factor s. Note that Eq. 3 is different from Eq. 1 at the point that there is no explicit blur kernel and $\downarrow_s$ and $\uparrow_s$ can be any form of sampling operation, such as the bi-cubic sampling operator. In addition, Eq. 4 is actually a special blur model of Eq. 2 when the frame averaging represents moving blur act on all sampled frame images.

Based on Eqs. 3 and 4, a widely used multifactor degradation model can be described as

$$y = \left(\left(\sum_{i=1}^{N} x_i)/N\right)\downarrow_s\right)\uparrow_s + n, \tag{5}$$

where frame averaging produces the effect of motion blurring and twice bi-cubic sampling (down-sampling first and up-sampling then) leads to resolution reduction. Obviously, changing the order of action between motion blurring and spatial sampling operations will result in other different multiple degradation models. Such another typical one can be denoted as

$$y = \left(\left(\sum_{i=1}^{N} x_i\downarrow_s\right)/N\right)\uparrow_s + n, \tag{6}$$

where a down-sampling is accomplished first and then the motion blurring is followed. Since the multifactor degradation model contained in Eq. 5 is often used in practical real-time camera imaging scenarios, in this work, we will focus on the multiple degradations shown in Eq. 5 and incorporate this model for constructing our CNNs to recover the degraded images. Moreover, in the following, we will show that the proposed decoupling cooperative learning-based model can also handle other different multiple degradations as shown in Eq. 6. The broad applicability of our model demonstrates its high practicability.

3.2 Decoupled cooperative learning

When conducting bi-cubic down-sampling, if spatial coordinate correspondence of bi-cubic interpolation is omitted, bi-cubic sampling will become a linear convolution against the HR image. Thus, twice bi-cubic samplings (down-sampling first and up-sampling then) are fully equivalent to a convolution kernel, denoted as h. As for motion blur, if assuming the blur kernel is k and taking the sharp latent image x corresponding to each blurry one as the mid-frame among the sharp frames $(\ldots, x_i, \ldots)$ that are used to generate the blurry image, then according to Eqs. 2 and 5, we have

$$y = (x * k) * h + n, \tag{7}$$

where $*$ is the convolution operator and y is the multiple degraded observation. Based on the associative law of convolution, Eq. 7 can be written as $x * (k * h) + n$. Then, if defining the kernel convolution of $k * h$ as a new kernel T, finally we can get an equivalent degradation model formalized as

$$y = x * T + n. \tag{8}$$

Equation 8 reveals that the compounded action of multiple degradations is approximately equivalent to making one time convolution with single equivalent blur kernel. Moreover, since the degraded image y is the observation of the sharp and HR image x, we can calculate the residual r between x and y. Assuming all the image samples follow Gaussian distribution, naturally an MLE estimation solution of Eq. 8 is $\tilde{x} = y + r$. Because the blur convolution is always low pass, the observation y in Eq. 8 can be regarded as the estimated low-frequency component of x. Thus, the residual r naturally turns to the high-frequency component of x. Obviously, r is composed of multiple diverse of high-frequency details, such as HR details and motion details. Thus, let HR details be denoted as r_{HR} and motion details be r_{M}, the MLE estimation $\tilde{x} = y + r$ can be expressed through the first-order Taylor expansion as

$$\tilde{x} = y + r_{\mathrm{HR}} + r_{\mathrm{M}}, \tag{9}$$

where $r = r_{\mathrm{HR}} + r_{\mathrm{M}}$. Equation 9 means that if given the composite degradation observation y, the latent sharp and HR image x can be estimated when acquiring r_{HR} and r_{M}.

Based on Eq. 9, if we can learn r_{HR} and r_{M} by CNN model, then the sharp and HR image x can be cooperatively reconstructed as long as the degraded image y is input to the model. Here, r_{HR} and r_{M} can be acquired independently through different CNNs and then they can be summed directly (note that the operator between r_{HR} and r_{M} in Eq. 9 is an addition) to achieve the final reconstruction result.

Suppose $H(y, \Theta)$ and $G(y, \Psi)$ are HR details and motion details recovery CNNs, respectively, the loss function of the proposed decoupled cooperative learning-based model can be formalized as

$$\text{Loss}(\Theta, \Psi) = \sum_{i=1}^{N} \left\| x_i - (y_i + H(y_i, \Theta) + G(y_i, \Psi)) \right\|^2, \quad (10)$$

where N is the number of training samples, x_i and y_i represent one pair of sharp and HR image and the multiple degraded image; Θ and Ψ are the network parameters of the respective image SR and motion deblurring sub-CNNs.

3.3 Network architecture

Our proposed deep network contains two sub-CNNs, where one is for HR recovery and the other one is deployed for motion deblurring. We design a decoupled cooperative learning architecture (Fig. 2) such that the two sub-CNNs are added to reconstruct the final sharp and HR image. Specifically, the proposed model consists of three main modules: (1) the upper sub-CNN module, which recovers image HR details for image SR; (2) the bottom sub-CNN module, which acquires motion details for motion deblurring; (3) cooperative leaning module, which adds the low-frequency degraded input image, the learned HR image details and the acquired motion details to reconstruct the final sharp and HR image. We use different color boxes to indicate diverse operations: the orange-red box indicates convolution layers, the grass green box corresponds to deconvolution layers, the bright yellow box represents PReLU activation operation, the sky blur box refers to sum operation, and the brown box marks batch normalization (BN) layers.

We construct the upper sub-CNN for image SR by utilizing the improved residual structure [16]. Each improved residual unit contains a pair of symmetrical convolution and deconvolution layers with PReLU activation and a scale factor β. In addition, another bypass scale factor α is also introduced in each residual unit. Usually, we make the sum of β and α equal one.

Suppose the input of residual unit l is x_l and the output is t_l, the improved residual learning unit can be formalized as

$$t_l = \alpha h(x_l) + \beta f(x_l, w_l), x_{l+1} = g(t_l), \quad (11)$$

where h is just the identity mapping of the bypass path, f represents the mapping of the principal 'conv–deconv' path, and g is the PReLU activation function. With these residual learning units, the upper sub-CNN module for image SR can be described as

$$H(x, \theta) = f_{\text{rec}}(R^n(R^{n-1}(\cdots R(x) \cdots))), \quad (12)$$

where x is the LR input, θ represents all the parameters of the image SR sub-CNN H, R indicates one improved residual learning unit, n is the number of units and f_{rec} is the network convolution reconstructing operation.

At the same time, we construct the bottom motion deblurring sub-CNN through deep encoder–decoder with a multi-scale recursive architecture. Inspired by the work [20], we propose an encoder–decoder model with novel multi-scale recursive residual structure which can recursively convey the extracted feature information of certain scale to different scale layers of the model. When the blurred image is input, the bottom deblurring sub-CNN module will gradually reconstruct the motion details along the different scale structure of the decoder.

Actually, there are three scales deep recursive residual structure and each scale contains different number residual connections as well as different length 'Conv-BN-PRelu' blocks. The recursive residual architecture of our model is marked with solid lines in Fig. 2, and the dashed lines indicate the bypass hop connections between the encoder and the decoder. An example of the proposed recursive residual structure is shown in Fig. 3. In the figure, the input signal will pass through two 'Conv-BN-PRelu-Conv-BN-PRelu' blocks with recursively residual connections.

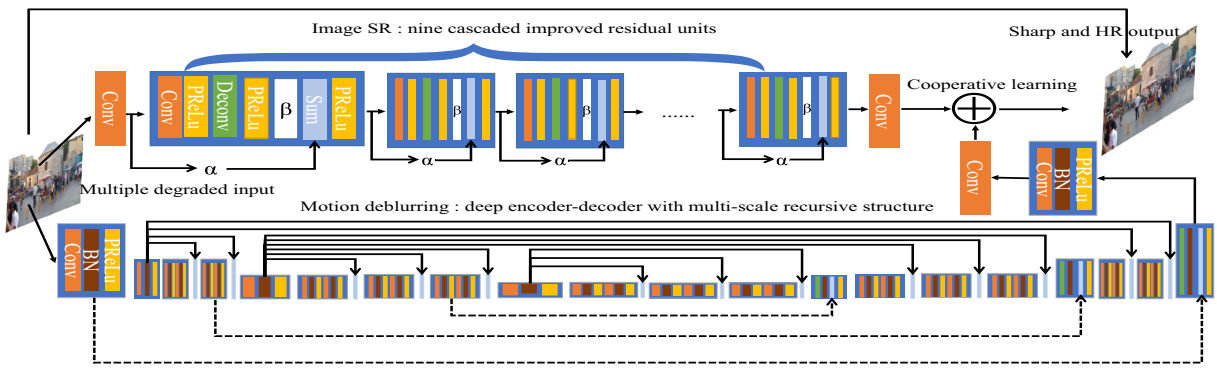


Fig. 2 The proposed decoupled cooperative learning-based joint image SR and deblurring model: the upper sub-CNN is for image SR, and the bottom sub-CNN is for motion deblurring; here, the cooperative learning (not multitask collaborative learning) means the two sub-CNNs can not only accomplish their respective tasks independently, but also achieve SR and deblurring simultaneously

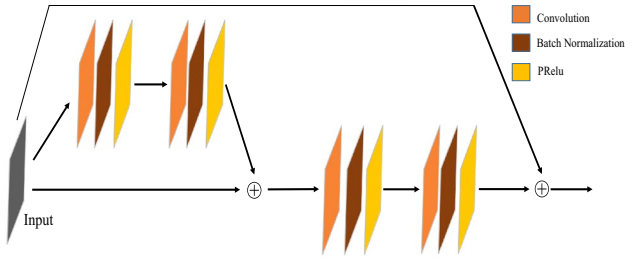


Fig. 3 An example of the recursive residual structure including two 'Conv-BN-PRelu-Conv-BN-PRelu' blocks with recursively bypass connections

Suppose at the scale i, x_u^i represents the input of one recursive residual unit u, t_{u+1}^i is the output of the same unit, and let F reflect the main route transformation mapping in the residual section, then the recursive residual structure in Fig. 3 can be described as

$$t_{u+1}^i = F(F(x_u^i) + x_u^i) + x_u^i. \tag{13}$$

Here, the main route transform F is actually the forward response of the blocks 'Conv-BN-PRelu-Conv-BN-PRelu,' and the input x_u^s can be treated as the output of the last scale recursive residual structure, i.e., x_u^s equals t_o^{s-1}. Finally, the bottom sub-CNN for motion deblurring in Fig. 2 can be described as:

$$G(x, \Psi) = D_3(D_2(D_1(P^3(P^2(P^1(f(x))))) + q^2) + q^1) + q^0, \tag{14}$$

where x is the motion-blurred input, f refers to certain processing (e.g., 'convolution') operation on the degraded input, Ψ describes the parameters of the entire deblurring sub-CNN module G, D_i indicates the whole decoder mapping at scale i, P^i represents the mapping of whole recursive residual structure at scale i , q^i denotes the dashed line connection at scale i from the encoder to the decoder (the shortest dashed line connection is q^2 and q^0 equals to $f(x)$).

4 Experiments and analysis

4.1 Datasets and training details

We perform experiments and compare the algorithm performance on the widely acknowledged motion blurring datasets: GOPRO dataset [15] and the dataset provided by Lai et al. [8]. GOPRO [15] is a natural image sequences dataset which has a total of 2103 training HR image pairs (the blurry and the sharp one) and 1111 test blurry ones, and each image keeps the size of 1280 * 720. On the contrary, the dataset of Lai et al. [8] is a synthetic blurred image dataset. Each degraded image (the size varies from 502×351 to 1280×680) in this dataset comes from the convolution of a corresponding sharp image with a blur kernel. (Its size may range from 21×21 to 75×75.) For a fair comparison, we only use GOPRO dataset [15] to generate the training data, but for performance evaluation both the test blurry images in GOPRO [15] and the dataset of Lai et al. [8] will be employed. Specifically, for deblurring training, we crop HR image pairs in GOPRO dataset [15] into 96×96 patches with a stride of 27 to obtain blurry HR patches and the sharp HR patches. While for the training of joint SR and deblurring, the blurry HR images are firstly imposed the 4× bi-cubic interpolation twice (down-sampling first and up-sampling then) to get the blurry LR images. Then, the blurry LR images and the original sharp HR images are cropped separately in the same way as above to get the blurry LR patches and the label sharp HR patches for final training. Regarding the quality measurement of the reconstructed or recovered images, the well-known PSNR [dB] and SSIM [23] metrics are adopted.

We take two steps to train the proposed model. The first step is to train the deblurring sub-CNN with the blurry HR patches and the sharp HR patches. In the first step, the training procedure is implemented by Adam solver from Caffe package [6] with the learning rate being fixed 0.001 and the batch size being 24. The second step is to fine-tune the entire model with the LR and HR image pairs. Actually, this second step is just the cooperative learning procedure that trains the image SR sub-CNN coordinately with the motion deblurring module so as to acquire the ability of simultaneous image SR and motion deblurring. In such a step, the label images are the sharp HR patches. This step training is also implemented by Adam solver from Caffe package [6] but with different learning rate of 0.0001. With a Nvidia Titan GTX1080ti GPU, training our proposed model by such two steps will cost totally about one day.

4.2 Sub-CNNs versus overall model

Our entire model contains two sub-CNNs, where one is for motion deblurring and the other is for image SR. In the model, the same LR and motion-blurred images are input into the two sub-CNNs, respectively, and the loss function, denoted as Eq. 10, is utilized to supervise the entire model training. In this section, we will investigate whether these two sub-CNNs can perform their own independent tasks and whether they can cooperatively work to improve the overall performance.

The first experiment is to test the SR sub-CNN (4× down-sampling). We input the LR images, the motion-blurred images, the LR and motion-blurred images into the SR sub-network, separately. The SR results are shown in Fig. 4. From the figure, we can easily get that the image

Fig. 4 Performance test of image SR sub-CNN: the first and third columns are the input images, and the second and fourth ones are the SR outputs; from top row to bottom one are LR images, motion-blurred images, and LR and blur images, respectively; PSNR and SSIM values of each image are indicated

SR sub-network works well for LR or LR and blur images. But for motion-blurred images, it has no substantial positive effect. (See the PSNR/SSIM value in the second row.)

The next experiment is to test the deblurring sub-CNN. We also take the above three kinds of degraded images as input to the deblurring sub-network. The corresponding deblurring results are shown in Fig. 5. As can be seen from the figure, the deblurring sub-CNN indeed recovers the blurred images while having zero effect on the resolution degenerated images. Based on these two experiments, it is easy to see that such two sub-networks are efficient for their own target tasks and decoupled independent for their partner task.

The last experiment in this section is to test the overall model. We input the same three types of degraded images into the overall model and observe its output results. The corresponding results of the overall model are shown in Fig. 6.

Comparing Fig. 6 with Fig. 5, we find that the overall model has better deblurring effect on the same blurred inputs. This phenomenon shows that after cooperative learning, the image SR sub-CNN and the deblurring sub-network have indeed become mutually beneficial partners, thus improving the performance of the whole model. In addition, since the overall model is trained only with blurred image inputs and supervised only by the sharp labels, it does not have a much better effect on pure LR degraded images as illustrated in Fig. 4.

4.3 Comparisons and degradation order analysis

We do some comparisons with several recent related methods, including image SR models [9, 11, 38], the deblurring methods [15, 28, 22], the multifactor degradation recovering approaches [35–37] and the combinations of SR algorithms [11, 38], and blind deblurring algorithms [15, 22]. For fair play, we use the public code they provide and for those that are not directly available (for example, SCGAN [29] and ED-DSRN [37]), we get them by retraining the models with our training dataset. The visual comparisons with Xu et al. [28] and multi-scale deblur [15] on GOPRO [15] and Lai et al. [8] datasets are shown in Fig. 7. And the visual comparisons with GFN [36] are illustrated in Fig. 8. At the same time, the comparisons of the average PSNR, the average SSIM, the model parameters, task independence, the training time and the test time, with some methods under such two datasets, are shown in Tables 1 and 2.

From Figs. 7 and 8, it can be observed that our proposed model gets the best degraded image recovery results in most cases and only performs a little worse at certain special scenes than the GFN [36] (Fig. 8b). But note that sometimes the PSNR or SSIM values calculated from some

Fig. 5 Performance test of motion deblurring sub-CNN: the first and third columns are the input images, and the second and fourth ones are the SR outputs; from top row to bottom one are LR images, motion-blurred images, and LR and motion-blurred images, respectively; PSNR and SSIM values of each image are indicated

Fig. 6 Performance test of the overall model: the first and third columns are the input images, and the second and fourth ones are the simultaneous SR and deblur outputs; from top row to bottom one are LR images, motion-blurred images, and LR and blurred images, respectively; PSNR and SSIM values of each image are indicated

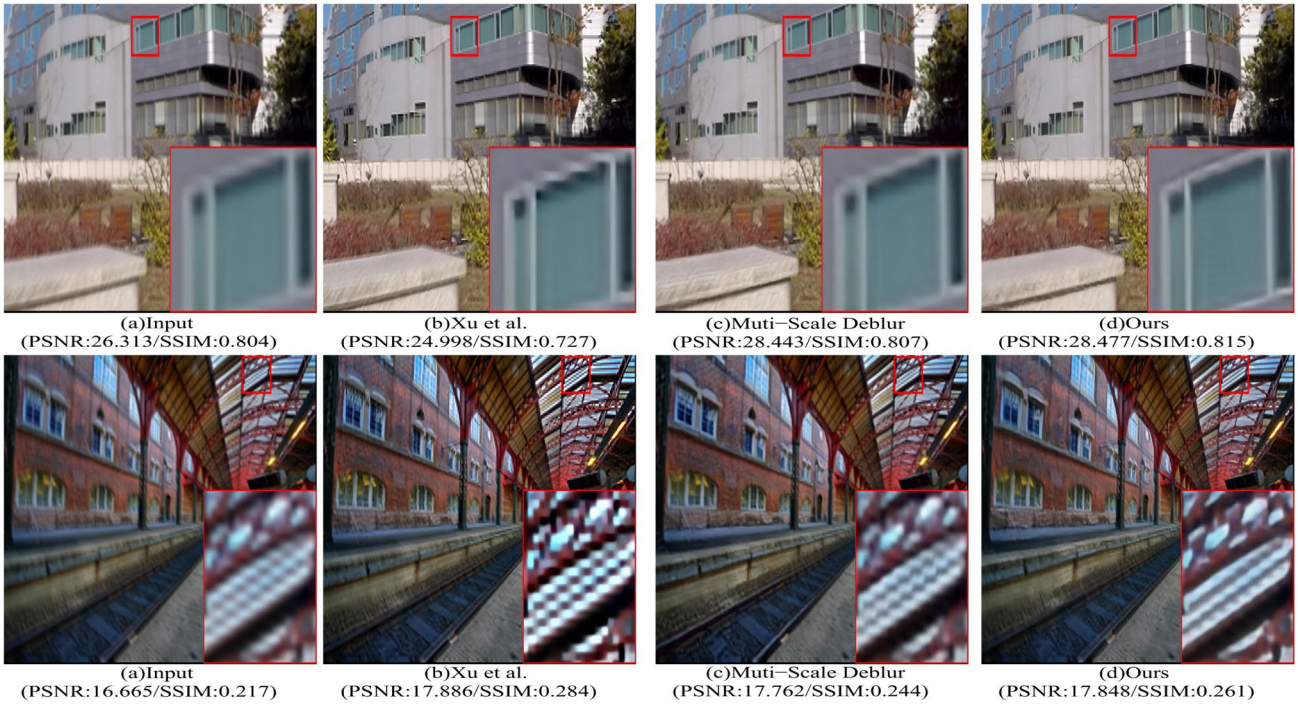


Fig. 7 Visual comparisons to some state-of-the-art methods: **a** the input blur and LR images (the first row input image comes from GOPRO dataset [15] while the second row one is from the dataset of Lai et al. [8]); **b** SR deblur of Xu et al. [28]; **c** multi-scale deblur [15]; **d** our proposed approach

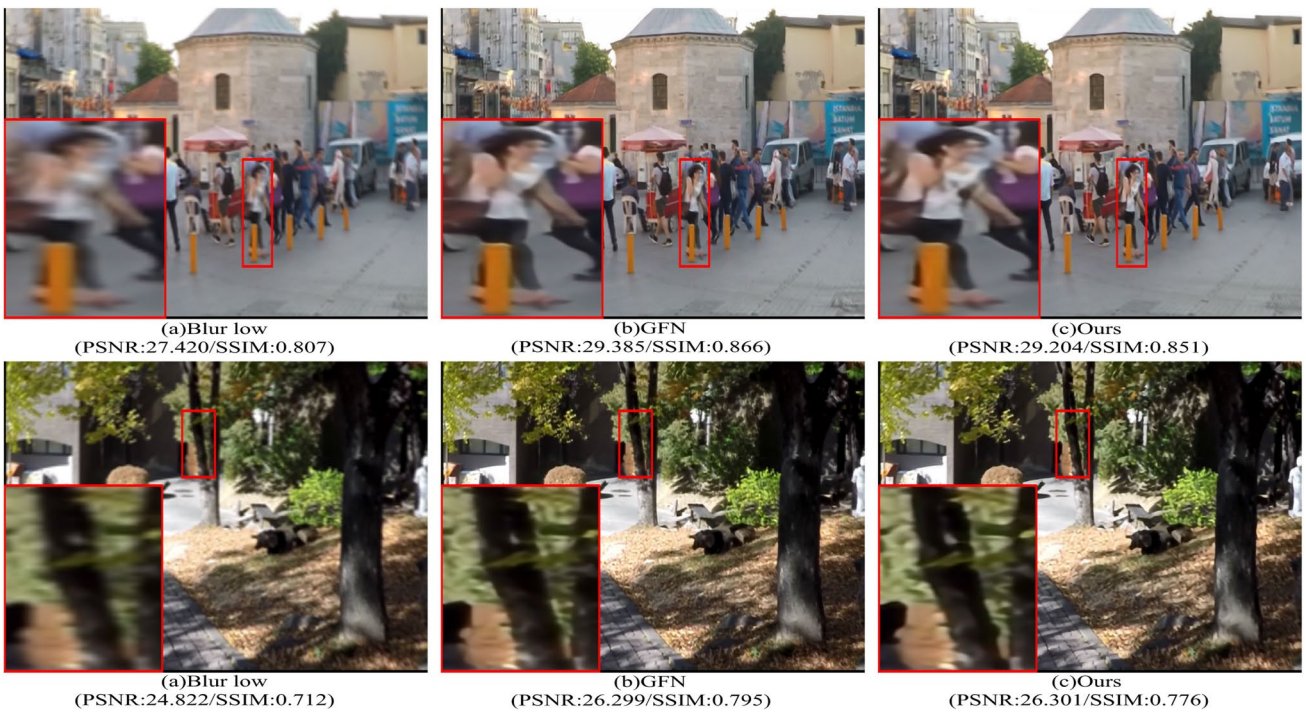


Fig. 8 Visual comparisons to some state-of-the-art methods: **a** the input blur and LR images (from GOPRO dataset [15]); **b** GFN (gated fusion network) of Zhang et al. [36]; **c** our proposed approach

specific images may be deceptive. The image of the second row in Fig. 7b will be an illustration in which our recovered images look visually much better than those generated by Xu et al. [28] though our PSNR is a bit lower. With a view to the quantitative measures in Tables 1 and 2, it is clear that compared to other methods, our proposed model can get the best or the second average performance in terms of PSNR and SSIM even on different blurry and LR datasets. It should be noted that Zhang et al. [35] achieves better performance on Lai's dataset, but poor performance on GOPRO dataset may be due to the fact that the blurring of GOPRO dataset is natural image motion blurring rather than the convolution effect of different convolution kernels as Lai's dataset does (which is exactly the same way as Zhang et al. [35] generates its training data). Moreover, according to Table 1, our proposed approach holds a unique characteristic that our model is decoupled, that is, each sub-task CNN can be taken out directly to serve its independent task.

In addition, for mobile applications, in terms of running speed, model volume and performance integration (see Tables 1 and 2), our approach is also faster, smaller and better as compared to others.

In order to investigate whether the order of multiple degradations affects the performance of the proposed model, we compare the recovery effect of different models trained by different degradation order images. Firstly, we generate the new degraded images with different action order based on Eq. 6 (i.e., first down-sampling, then blurring and finally up-sampling) and then use these new degraded images to train the proposed model. Let us denote the original trained model and the new trained model as CL1 and CL2 separately. The comparisons of the recovery effects with different degradation order inputs are illustrated in Fig. 9. In addition, the recovery comparisons from our original CL1 model and the new trained CL2 model also are shown in Fig. 10.

From Fig. 9, it is clear that the effects of different degeneration order given by Eqs. 5 and 6 on image quality deterioration are almost the same, and they can be dealt with appropriately to obtain equally good recovery images through our approach. From Fig. 10, it is obvious that the model CL1 and the model CL2 can both achieve good and similar reconstruction results, regardless of the degeneration order of input images. This demonstrates that our proposed approach is not sensitive to the degradation order of training data. Thus, our proposed approach holds a great tolerance on multiple degeneration images with different order, which shows excellent generalization performance.

4.4 Ablation study

There are several key components in the proposed decoupled cooperative learning-based model: (1) using the

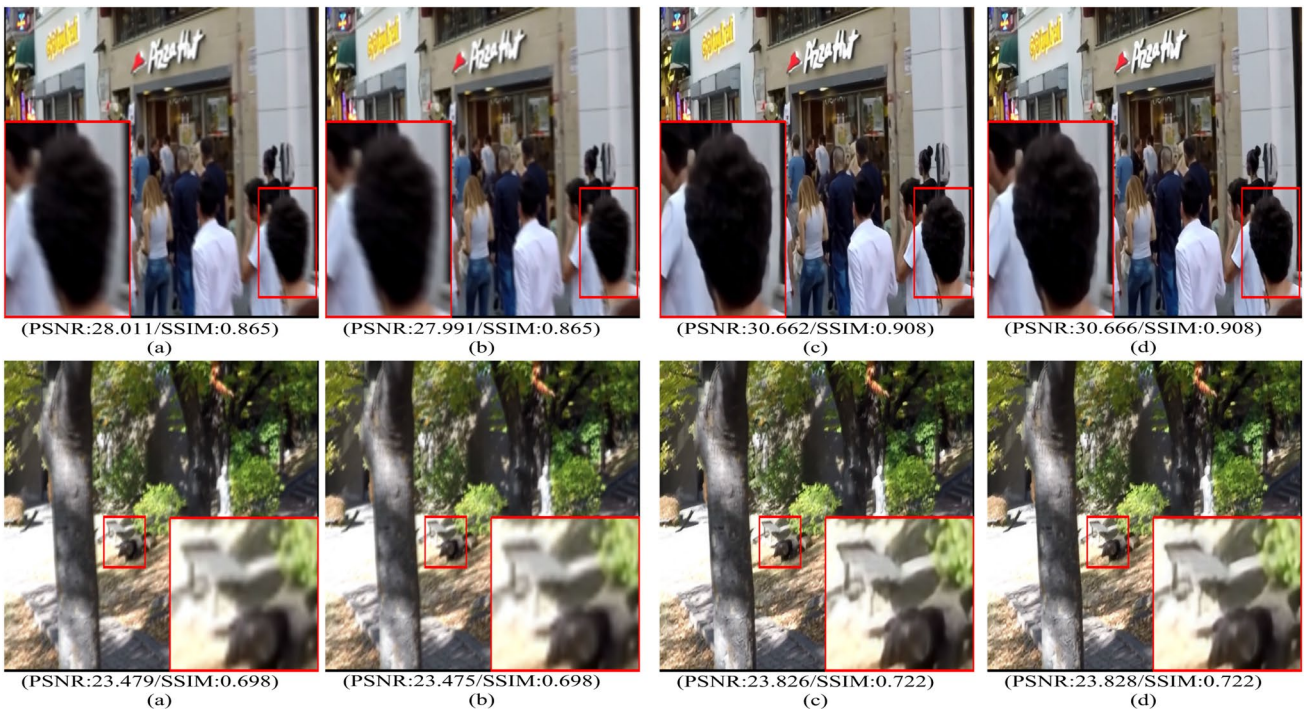


Fig. 9 The recovery comparisons with different degeneration order inputs: **a** are the input images (from GOPRO dataset [15]) with the degeneration order of Eq. 5; **b** are the input images with the degradation order of Eq. 6; **c** and **d** are the respective recovery results of (**a**) and (**b**) by our approach (CL1 model)

Fig. 10 The recovery comparisons of different models trained by different degeneration order images: **a** the input motion blur and LR images; **b** the recovery results of (**a**) through CL1 model; **c** the recovery results of (**a**) through CL2 model (trained by different degeneration order images); **d** the ground truth

sub-SR module to recover the degraded spatial features; (2) using the sub-deblurring module to recover the motion-blurred details; (3) cooperatively fusing features from image SR and deblurring modules with a sum operator instead of concatenation; (4) using a unified loss other than a multitask loss to eliminate the effect of multiple factors co-degradation. Here, we discuss the performance contribution of these components.

We first take the multi-scale deblur (MS deblur) model [15] as a baseline model (with deblurring loss). Then, starting with the SR module, we gradually add these key components to form a new model. We use the same hyper-parameters to train these models produced in the process. The quality assessment results on GOPRO dataset [15] for each model are shown in Table 3. The experimental results in the table demonstrate that the deblurring module plays a more significant role on performance improvement than the image SR module when dealing with multifactor co-degradation, while the implementation speed of the SR module is the fastest. However, according to the table, the multitask loss may lead to the performance compromise and the concatenation of SR and deblurring modules also might be sub-optimal possibly owing to the wrong feature accumulation. Compared with Model 4 (multitask loss), Model 3 (taking the single loss after feature concatenation) acquires 0.16 dB performance improvement and is with the same speed. Finally, we see that our proposed cooperative learning-based model can further boost the performance by 0.18 dB.

5 Conclusion

In this work, we proposed a simple but efficient deep decoupled cooperative learning model to achieve fast and simultaneous image SR and motion deblurring. We explored the principle of decoupled cooperative learning for constructing multiple degradations recovery model and investigated the role of each sub-network in the model when facing different types of degenerated inputs. At the same time, we also investigated the impact of the order of multifactor degradation and the key components on the model performance. Lots of experiments and comparisons are performed to show the good recovery performance as well as the good generalization compared to the other state-of-the-art methods.

Future work will, on the one hand, focus on introducing the edge map guidance [13] to achieve better simultaneous SISR and motion deblurring. On the other hand, we intend to apply the cooperative learning to facilitate applications, such as large-scale cross-retrieval [25, 26], video captioning [30], image classification [1, 14] and gesture biometrics [12].

Table 1 Quantitative performance comparisons under GOPRO dataset [15]

Measures	RDN [38]	SRN [22]	Multi-scale deblur [15]	SCGAN [29]	SRResNet [9]	RDN [38] + SRN [22]	EDSR [11] + MS deblur [15]	ED-DSRN [37]	Zhang et al. [35]	Our proposed
PSNR	24.370	25.829	26.765	22.791	24.430	26.211	26.275	26.331	25.80	**27.048**
SSIM	0.739	0.782	0.786	0.783	0.804	0.792	**0.860**	0.810	0.768	0.811
Parameters	178M	28.8M	12M	**1.1M**	1.5M	305M	13M	25M	7M	9M
Task independent	No	No	No	No	No	No	No	No	No	**Yes**
Training/inference time	1.0 day/2.8 s	3 days/0.4 s	1 day/1.3 s	1.5 days/0.68 s	**0.5 day/0.11 s**	3.8 days/4 s	2 days/2.7 s	1.5 days/0.22 s	2 days/1.3 s	1 day/1.9 s

Best results are indicated in bold

Table 2 Quantitative performance comparisons under Lai et al. dataset [8]

Measures	RDN [38]	SRN [22]	Multi-scale deblur [15]	SCGAN [29]	SRResNet [9]	EDSR [11] + MS deblur [15]	RDN [38] + SRN [22]	ED-DSRN [37]	Zhang et al. [35]	Our proposed
PSNR	17.780	17.444	18.541	18.572	18.785	18.175	18.861	18.791	**19.003**	18.907
SSIM	0.416	0.408	0.456	0.460	0.471	0.478	0.423	0.473	0.466	**0.484**
Inference time (s)	2.3	0.3	1.0	0.50	**0.09**	3.1	2.2	0.20	1.1	1.4

Best results are indicated in bold

Table 3 Analysis on key components in our proposed decoupled cooperative learning model

Components	MS deblur	Model 1	Model 2	Model 3	Model 4	Our model
SR module		√	√	√	√	√
Deblurring module	√		√	√	√	√
Multitask loss			√		√	
Features concatenation				√	√	
Features cooperation						√
PSNR	26.76	25.83	26.51	26.86	26.70	27.04
Inference time (s)	1.3	1.1	1.8	1.9	1.9	1.9

Acknowledgements This work is supported in part by the National Natural Science Foundation of China under Grant No. 61971004, by the Key Project of Natural Science of Anhui Provincial Department of Education under Grant No. KJ2019A0083) and by the Natural Science Foundation of Anhui University of Technology under Grant No. RD18100244.

References

1. Ding, G., Guo, Y., Chen, K., Chu, C., Han, J., Dai, Q.: Decode: deep confidence network for robust image classification. IEEE Trans. Image Process. **28**(8), 3752–3765 (2019)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of European Conference on Computer Vision, pp. 184–199 (2014)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
4. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Trans. Image Process. **22**(4), 1620–1630 (2013)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of ACM International Conference on Multimedia, pp. 675–678 (2014)
7. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
8. Lai, W.S., Huang, J.B., Hu, Z., Ahuja, N., Yang, M.H.: A comparative study for single image blind deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1709 (2016)
9. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
10. Li, L., Pan, J., Lai, W.S., Gao, C., Sang, N., Yang, M.H.: Learning a discriminative prior for blind image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6616–6625 (2018)
11. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
12. Liu, H., Dai, L., Hou, S., Han, J., Liu, H.: Are mid-air dynamic gestures applicable to user identification? Pattern Recognit. Lett. **117**, 179–185 (2019)
13. Liu, H., Fu, Z., Han, J., Shao, L., Hou, S., Chu, Y.: Single image super-resolution using multi-scale deep encoder-decoder with phase congruency edge map guidance. Inf. Sci. **473**, 44–58 (2019)
14. Luan, S., Chen, C., Zhang, B., Han, J., Liu, J.: Gabor convolutional networks. IEEE Trans. Image Process. **27**(9), 4357–4366 (2018)
15. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3883–3891 (2017)
16. Patrick, H.: Super-resolution on satellite imagery using deep learning part 1. The DownLinQ (2016)
17. Qiao, T., Ren, J., Wang, Z., Zabalza, J., Sun, M., Zhao, H., Li, S., Benediktsson, J.A., Dai, Q., Marshall, S.: Effective denoising and classification of hyperspectral images using curvelet transform and singular spectrum analysis. IEEE Trans. Geosci. Remote Sens. **55**(1), 119–133 (2016)
18. Schuler, C., Hirsch, M., Harmeling, S., Scholkopf, B.: Learning to deblur. IEEE Trans. Pattern Anal. Mach. Intell. **38**(7), 1439–1451 (2016)
19. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016)
20. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1279–1288 (2017)
21. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 769–777 (2015)
22. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8174–8182 (2018)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
24. Wang, Z., Ren, J., Zhang, D., Sun, M., Jiang, J.: A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing **287**, 68–83 (2018)
25. Wu, G., Han, J., Guo, Y., Liu, L., Ding, G., Ni, Q., Shao, L.: Unsupervised deep video hashing via balanced code for large-scale video retrieval. IEEE Trans. Image Process. **28**(4), 1993–2007 (2018)
26. Wu, G., Han, J., Lin, Z., Ding, G., Zhang, B., Ni, Q.: Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning. IEEE Trans. Ind. Electron. **66**(12), 9868–9877 (2018)

27. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. In: Advances in Neural Information Processing Systems, pp. 1790–1798 (2014)
28. Xu, L., Zheng, S., Jia, J.: Unnatural l0 sparse representation for natural image deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1107–1114 (2013)
29. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 251–260 (2017)
30. Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., Dai, Q.: Stat: spatial-temporal attention mechanism for video captioning. IEEE Trans. Multimed. (2019)
31. Yan, Y., Ren, J., Sun, G., Zhao, H., Han, J., Li, X., Marshall, S., Zhan, J.: Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. Pattern Recognit. **79**, 65–78 (2018)
32. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
33. Yang, W., Feng, J., Yang, J., Zhao, F., Liu, J., Guo, Z., Yan, S.: Deep edge guided recurrent residual learning for image super-resolution. IEEE Trans. Image Process. **26**(12), 5895–5907 (2017)
34. Zhang, K., Wang, B., Zuo, W., Zhang, H., Zhang, L.: Joint learning of multiple regressors for single image super-resolution. IEEE Signal Process. Lett. **23**(1), 102–106 (2016)
35. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3262–3271 (2018)
36. Zhang, X., Dong, H., Hu, Z., Lai, W.S., Wang, F., Yang, M.H.: Gated fusion network for joint image deblurring and super-resolution. In: BMVC (2018)
37. Zhang, X., Wang, F., Dong, H., Guo, Y.: A deep encoder–decoder networks for joint deblurring and super-resolution. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1448–1452. IEEE (2018)
38. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)

Heng Liu is currently a professor in School of Computer Science and Technology, Anhui University of Technology. He received his B.Sc. degree in automation from Southwest University of Science and Technology in 1993, M.Sc. degree and Ph.D. degrees in Pattern Recognition and Intelligent System from Chongqing University and Shanghai Jiao Tong University in 2004 and 2008, respectively. His research interests include computer vision, deep learning and biometrics. He has contributed more than 70 research papers. He is a member of ACM, IAPR and CCF.

Jiajun Qin received the B.Sc. degree in Anhui University of Technology in 2018. Now, he is pursuing his master degree in computer science and technology. His research interests are deep decoupling learning-based low-level vision.

Zilin Fu received the M.Eng. degree in Anhui University of Technology in 2019. His research interests include deep learning and image super resolution.

Xue Li received her B.S. degree and the Ph.D. degree in computer science and technology from Nanjing University of Science and Technology, Nanjing, China, in 2011 and 2017, respectively. She held research positions with Anhui University of Technology since 2017. Her current research interests include hyperspectral unmixing, deep learning, computer vision and image processing.

Jungong Han is currently an Associate Professor of Data Science, University of Warwick, Coventry, UK. His research interests include video analysis, computer vision and artificial intelligence.