IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 21, 2020, accepted April 4, 2020, date of publication April 7, 2020, date of current version March 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.2986365

Single Image Super-Resolution by Residual Recovery Based on an Independent Deep Convolutional Network

FEI WANG AND MALI GONG

State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instruments, Tsinghua University, Beijing 100084, China
Key Laboratory of Photonic Control Technology, Ministry of Education, Tsinghua University, Beijing 100084, China
State Key Laboratory of Tribology, Department of Precision Instrument, Tsinghua University, Beijing 100084, China

Corresponding author: Mali Gong (gongml@mail.tsinghua.edu.cn)

ABSTRACT In this paper, we propose an independent neural network for single image super-resolution by residual recovery. The network is inspired by the observation that there still exists image residuals between the low-resolution image and the downsampled high-resolution output obtained by a previously proposed super-resolution network. Based on this observation, we design a simple but effective deep convolutional neural network to train the mapping between the image residuals and the corresponding ground-truth residuals. Furthermore, we combine the high-resolution output generated by the previous super-resolution network and the high-resolution residual output by the proposed neural network to yield the final high-resolution image. Extensive experiments on simulated natural images and real time-of-flight (ToF) images demonstrate the effectiveness of the proposed method from the aspects of visual and quantitative performance.

INDEX TERMS Single image super-resolution, independent deep convolutional neural netowork (IDCNN), image residual recovery, ToF images.

I. INTRODUCTION

The main goal of single image super-resolution (SR) is to recover a high-resolution (HR) image from one low-resolution (LR) image while keeping clear image details. In general, the LR image only contains fewer image details than that of the HR image, which promotes us to develop mathematical strategies or approaches to improve the LR image's details. Therefore, how to propose an accurate and fast SR approach to increase image resolution is quite crucial, which is also the main task and challenge in this work.

From the perspective of methodology, existing single image SR approaches can be divided into three categories: 1) interpolation-based method; 2) statistics-based method; 3) learning-based method. In particular, the learning-based method could be roughly divided into two parts. One is the dictionary-based learning method, and the other is the deep learning-based method. The proposed method in the paper belongs to the category of the deep learning-based method.

The interpolation-based method is a kind of classical single image SR approach. It has been studied for several decades. This kind of method is mainly to fill in pixels at unknown locations by some relations in terms of its neighbor points. The most classical interpolation methods for single image SR are nearest-neighbor interpolation and bicubic interpolation. Both methods could yield SR outcomes fastly; however, the nearest-neighbor interpolation generally will lead to a jaggy effect, and bicubic interpolation may result in blur effect. Besides them, recently some state-of-the-art interpolation methods are also proposed, readers are recommended to check the related references, see, e.g., [2]–[5].

The statistics-based method also becomes an active field of the image SR. In general, it mainly contains two important directions, i.e., Maximum a Posterior (MAP) based method and Maximum Likelihood estimator (MLE) based method (see more related references [6]–[8]). In [7], Capel *et al.* proposed two estimators for the resolution enhancement of text images. One was proposing a MAP estimator that was based on a Huber prior, and the other was proposing an estimator using the total variation (TV) regularization. The given method was not only for enhancing image resolution

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

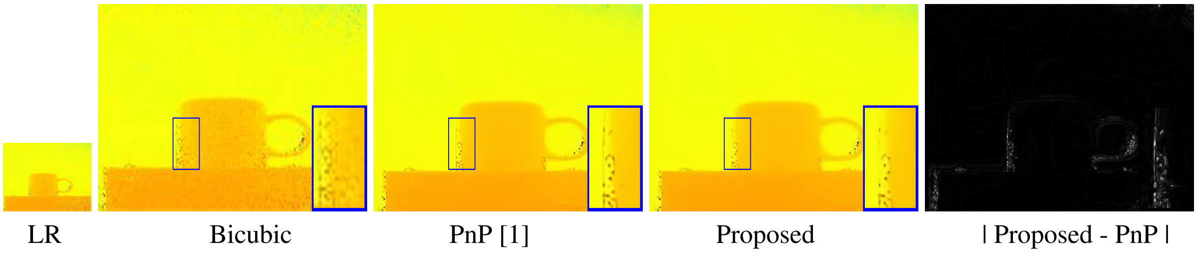

FIGURE 1. The super-resolution results for a real ToF image with a scale factor of 3. The LR image is with low-resolution and additional noise. It is clear that the proposed method holds the better ability of outlier removal than the state-of-the-art PnP method [1] when increasing image resolution (please see the close-up), which indicates the better performance of our method. Moreover, the absolute residual map (shown in the last image) between the PnP and the proposed method demonstrates that our method could pick up image details from the result of the PnP to get a better visual outcome.

but also could work for denoising tasks simultaneously. Based on the MAP, some regularization models are proposed for single image applications, e.g., image super-resolution [9], [10]. In [9], Deng *et al.* proposed a sparse regularization model by reproducing kernel Hilbert space (RKHS) function for single image SR. To pick up more image details, they also designed an iterative scheme for the solution by alternating direction method of multipliers (ADMM). After that, Deng *et al.* [11] presented a ℓ_1 sparse model based on two Heaviside function terms that one is to depict the primary image information and the other is to describe the sparse sharp edges. Experimental results demonstrate that the regularization models could obtain promising performance. Wang and Gong in [10] proposed an RKHS-based regularization model which can realize image SR and denoising simultaneously.

Dictionary-based learning approaches play a crucial role in the field of image SR, as well as show significant improvements than classical methods. Readers are recommended to find more references of this direction, e.g., [12]–[17]. One representative dictionary-based learning method for image SR was proposed by Yang *et al.* [16]. The authors formulated a dictionary-based learning framework for single image SR, which is to utilize a ℓ_0 sparse training model with LR patches and HR patches as input. After getting the relation between the LR patches and HR patches, it could obtain the output HR image by inputting an LR image to the learned relation.

Recently, with the tremendous improvements in hardware devices, deep learning has shown the superpower for image processing, e.g., [18], [19]. For the application of image SR, Dong *et al.* [20], [21] first utilized three layers of convolutional neural network (CNN) to address single image SR, called SRCNN. This network is based on a ℓ_2 loss function and to calculate the parameters on each layer, finally to predict the HR image by the trained nonlinear mapping with any LR image as input. After this work, many literatures based on CNN have been proposed for image SR, e.g., [22]–[27]. Kim *et al.* [24] proposed a deep recursive CNN for single image SR, which mainly has a very deep recursive layer. This recursive CNN will not introduce new parameters; thus, it has a quite fast speed for training and testing. Additionally, Lai *et al.* [26], [27] presented a fast and accurate image SR with a designed deep Laplacian pyramid network. The proposed network could reconstruct the sub-band residuals at multiple pyramid levels. Besides, due to the feature extraction on LR grids, thus the proposed approach has quite low computation. In [1], Zhang *et al.* proposed a deep plug-and-play SR method with arbitrary blur kernels. Especially, the framework of deep plug-and-play is mainly based on a new single image SR degradation model, which could take advantage of existing blind deblurring approaches. Experimental results on several simulated and real examples show that this method obtained the state-of-the-art single image SR performance. Although there are many deep CNN methods for the application of image SR, it still has space for improvements due to the multiscale property of SR. Especially, here we utilize this property of SR to design a deep neural network architecture for single image SR.

In this paper, we observe that there exist image residuals between the LR image and the downsampled HR output yielded by a previously proposed SR network. To utilize the image residuals on LR grids, we independently design a simple deep CNN that is based on ResNet [28] to pick up more image details for the final HR image. In particular, the ground-truth residuals of the independent deep CNN are obtained by the subtraction of the high-resolution output obtained by the previously proposed SR network and the ground-truth HR image. Furthermore, we use a ℓ_2 norm as the loss function. Experimental results on simulated natural images and real ToF images demonstrate the effectiveness of the proposed method. Additionally, Fig. 2 shows the flowchart of proposed deep CNN for single image SR.

In summary, this paper mainly has the following contributions: 1) Unlike the previous deep SR CNN that enforces the error between the network output and the ground-truth as small as possible, the paper is to formulate an independent deep CNN for the residual recovery to pick up more image details of HR images. 2) The proposed deep CNN yields the best performance, especially on the quantitative aspect, compared with modern state-of-the-art SR methods. 3) Our approach could work for real ToF images and get competitive visual performance.

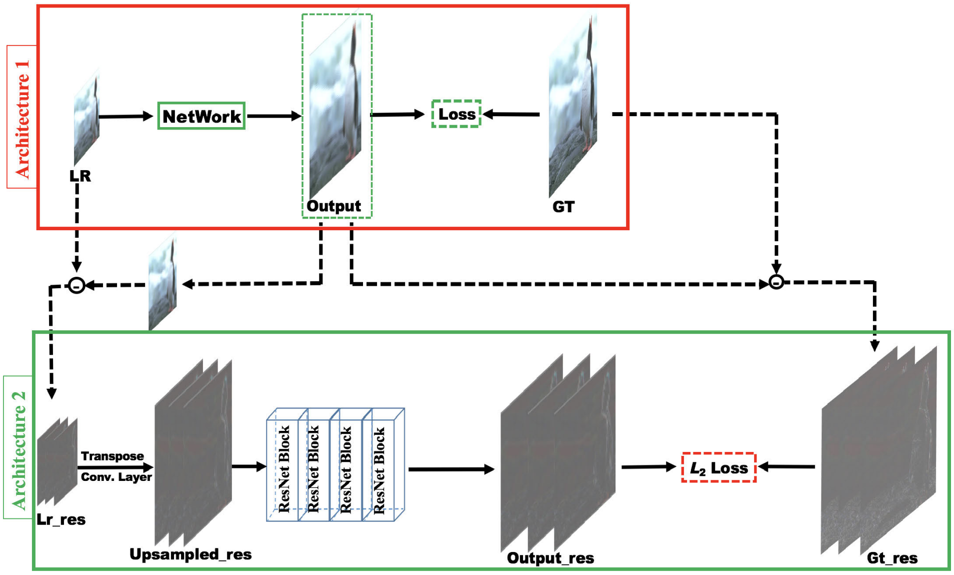

FIGURE 2. The flowchart of the proposed method. "Architecture 1" could be any existed network used for single image SR. After "Architecture 1", we downsample the "output" then calculate the residual between the LR and the downsampled output to get the residual input "Lr_res" for "Architecture 2". Besides, we also compute the ground-truth residual "Gt_res" by the subtraction of the HR output and the ground-truth. The designed network "Architecture 1" involves four ResNet blocks and ℓ_2 loss function. Especially, before entering into ResNet blocks of "Architecture 2", there is an operation of transposed convolution, which could increase the size of "Lr_res" to match the size of "Gt_res". The final SR image is the summation of the "Output" in "Architecture 1" and the "Output_res" in "Architecture 2". More parameter setting for our architecture can be found from the Section III-C.

The paper is outlined as follows. In Section II, we briefly introduce the related works. Section III detailedly presents the proposed deep CNN for single image SR. In Section IV, visual and quantitative results are reported to show the superiority of the proposed method. Also, we apply our method to real ToF images. In particular, we also explain why choosing ToF images as real test data in this Section. Finally, some conclusions are drawn in Section V.

II. PROBLEM FORMULATION AND RELATED WORKS

The proposed method in the work is actually based on the formulation of plug-and-play (PnP) [1], and the outcome of PnP is crucial to the final SR result of our method; thus we mainly review the brief introduction of SR and the formulation of PnP in this section.

Image SR is a critical problem in image processing, which is mainly to increase the spatial resolution of an image such that the processed image can better serve for subsequent applications, *e.g.,* recognition, segmentation, object detection, etc. Especially, the image SR can be mathematically formulated as follows

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

where $\mathbf{y}$ stands for the LR image, $\otimes$ is the convolution between the blur kernel $\mathbf{k}$ and the clean HR image $\mathbf{x}$, $\mathbf{n}$ represents the additive Gaussian white noise. Additionally, $\downarrow_s$ is the downsampling operator with a scaling factor s. This degraded SR model (1) is an ill-posed problem, we may take many strategies to solve it, *e.g.,* regularization-based approaches which have been used in many applications [29]–[34]. If following these regularization methods, some issues will appear. For example, how to estimate the blur kernel accurately. Even though some recent works arise to calculate the blur kernel, it is also difficult to accurately compute it.

Recently, Zhang *et al.* [1] novelly view the formulation (1) as the following SR degraded model,

$$\mathbf{y} = \mathbf{k} \otimes \mathbf{x} \downarrow_s + \mathbf{n}, \tag{2}$$

which means that it first downsamples the HR clean image $\mathbf{x}$ with a scaling factor s, then the downsampled image $\mathbf{x} \downarrow_s$ is blurred by the kernel $\mathbf{k}$. With this novel degraded SR model, two advantages are holden. First of all, the new degradation model follows the bicubic degradation model of $\mathbf{x} \downarrow_s$. Secondly, by this degradation model, many previous excellent blind deblurring approaches can be used for the estimation of blur kernel $\mathbf{k}$.

Based on the new degradation model (2), Zhang *et al.* propose the corresponding regularization model that is applied to the task of single image SR,

$$\min_{\mathbf{x}} \frac{1}{2\sigma^2} \|\mathbf{k} \otimes \mathbf{x} \downarrow_s - \mathbf{y}\|_2^2 + \lambda \Psi(\mathbf{x}), \tag{3}$$

where $\frac{1}{2\sigma^2} \|\mathbf{k} \otimes \mathbf{x} \downarrow_s - \mathbf{y}\|_2^2$ is the fidelity term and $\Psi(\mathbf{x})$ represents the regularization term, σ and λ are the noise level and the regularization parameter, respectively (see more details in [1]).

For the solution of (3), Zhang *et al.* [1] give a strategy that will solve the unknow variables alternatingly to obtain excellent SR outcomes. More details of the solving process can be found in [1] and the corresponding code of this method is also available (see the result section).

Especially, the PnP method in [1] could obtain state-of-the-art single image SR results, also shows the enormous capacity for a variety of images. *However, just like the mentioned before, there still exist visible image residuals between the LR image and the downsampled HR output yielded by a previous SR method, e.g., PnP (see also "Lr_res" in Fig. 2).* Motivated by the image residuals, we intend to design a deep CNN to recovery the lost HR residuals to finally generate better visual results. In what follows, we will present the whole flowchart of our approach detailedly.

III. PROPOSED METHOD

With the considerable development of image SR techniques, especially deep learning techniques, one can obtain very desired SR results even for a different type of images. However, there is no end for the improvement of image SR. We still have room to make SR results better by some new investigations or observations.

In this work, we propose the method based on an observation that there still exist visible image residuals between the LR image and the downsampled HR output generated by a previous SR method that even could be a state-of-the-art approach. In Fig. 2, it is evident that "Lr_res" that is from the subtraction of the LR image and the downsampled HR output still has significant image residuals; thus we attempt to pick up more HR image details from the LR residuals, just like the iterative SR method in [9], [35]. Different from [9], here we do not use a similar strategy to recover HR image details iteratively, but utilize the deep CNN that has been proven as a very efficient and effective technique for image SR in many pieces of literature. Especially if we intend to use the deep CNN for image SR, we have to simulate the training data which mainly includes two kinds of data, *i.e.,* the LR data and the corresponding ground-truth (GT) data. Fortunately, it is not difficult to yield the LR-GT residual-pairs for training in the work. After obtaining LR residual "Lr_res", the corresponding GT residual can be naturally generated by the subtraction between GT and the output of the network.

A. NETWORK ARCHITECTURE FOR THE RESIDUALS

The main goal of image SR is to recover spatial information from the LR image that generally only contains less spatial image details. Also, the spatial image details usually exist in the difference between the LR image and the downsampled estimated HR image. Besides, the deep CNN method, without depending on the pre-defined image priors that are sometimes not so accurate, has shown its significant superiority in image SR. Motivated by the just mentioned, we intend to propose a simple and effective network architecture by considering the spatial details on LR grids and deep CNN.

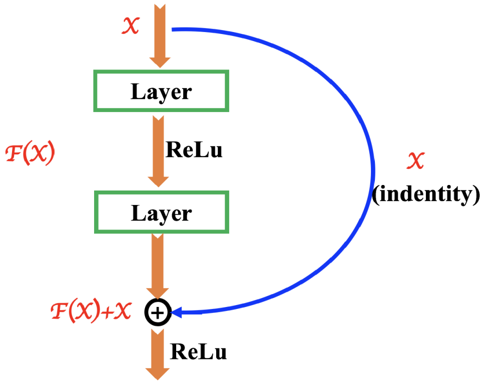

FIGURE 3. An illustration of ResNet block which contains two convolutional layers and two nonlinear functions (*i.e.,* ReLu).

The "Architecture 2" in Fig. 2 is our design for the residual recovery of image SR. From this architecture, it is easy to know that the calculated residual of LR "Lr_res" is taken into the network and will establish a nonlinear mapping f to the GT residual "Gt_res". Therefore, the output of the deep network can be viewed as the following:

$$\textbf{Output\_res} = f_{\Theta}(\textbf{Lr\_res}), \tag{4}$$

where Θ contains the network parameters that mainly include the convolutional filters and bias on each layer. Especially, the input LR residual has high-frequency image details such as edge information, and it is better to select a deep network architecture for the feature extraction. ResNet [28] is a very promising and excellent architecture in a deep convolutional neural network. It can achieve deep layers, which means the network has a more flexible ability to extract and represent image features. Thus we choose ResNet as the main part of our architecture. Specifically, the ResNet can be viewed as the combination of some ResNet blocks. Each ResNet block generally consists of two layers[1] with a nonlinear function ReLU or not, see Fig. 3 for one ResNet block. Especially, we only take four ResNet blocks in this work, since the input of the network is actually similar to the output of the network, the ResNet with few blocks is suitable to learn a transformation like this case. From "Architecture 2" in Fig. 2, it is easy to find that "Output_res", the output of our network, indeed contains some visible image residuals which can be viewed as the lost image details in "Architecture 1".

B. LOSS FUNCTION

After obtaining the output of network, *i.e.,* "Output_res" with the paramter Θ, it is necessary to define the loss function between the "Output_res" and the "Gt_res" so that we may calculate the paramters on each layers by backpropagation. Especially, one conventional loss function for high-frequency image details is ℓ_1 loss function which indicates $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. However, considering the performance in

[1] It is flexible in defining the layer number according to actual requirements.

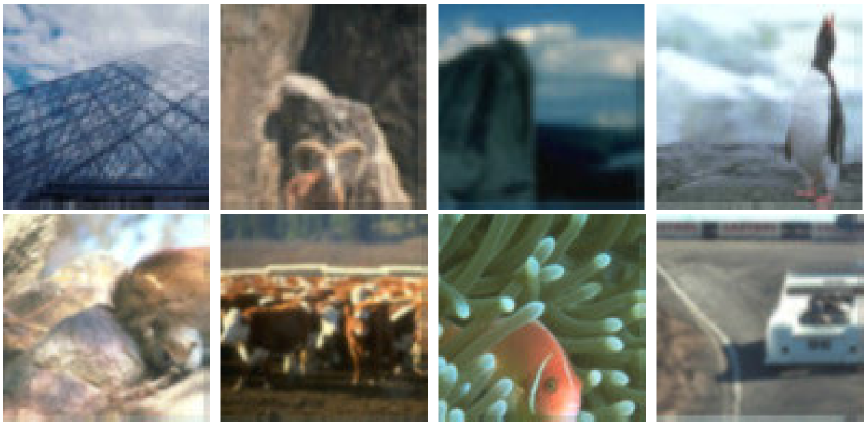

FIGURE 4. Low-resolution images in Fig. 5 (First row) and Fig. 6 (Second row), respectively.

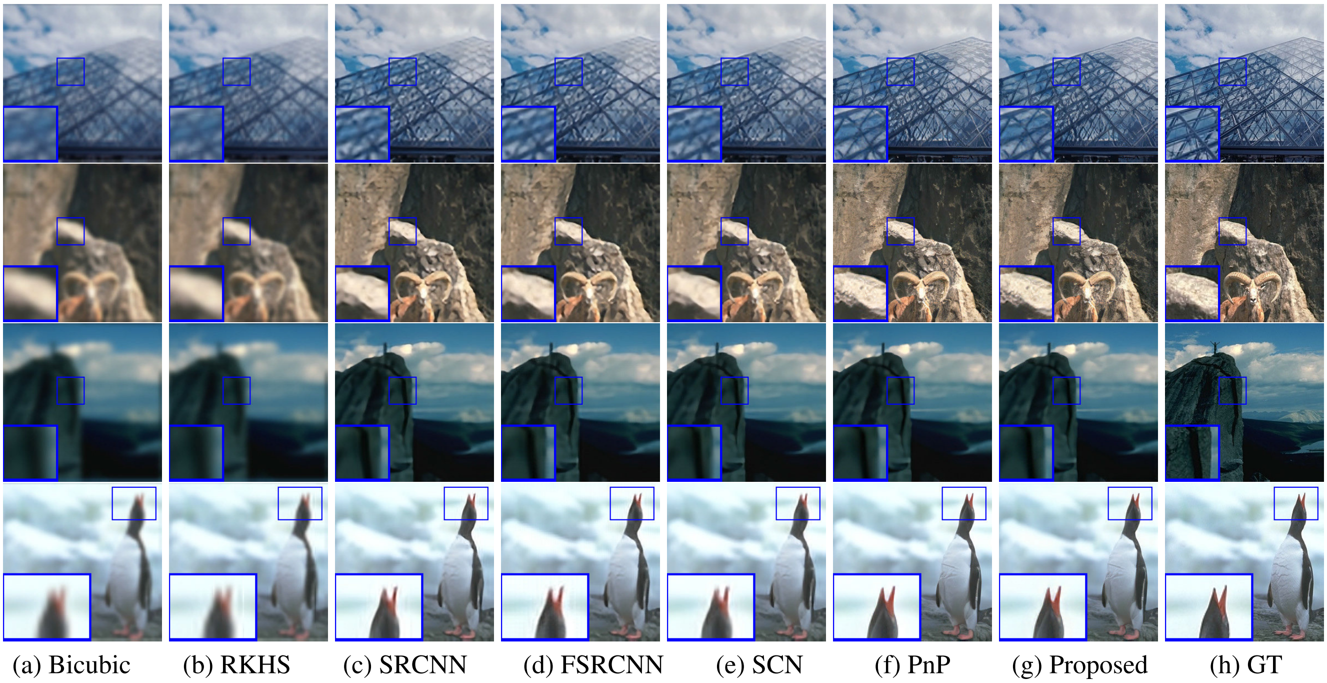

FIGURE 5. Comparsions with some recent state-of-the-art super-resolution approaches on four examples (called "Glass", "Sheep", "Rock", "Penguin"). The upscaling factor is 4. The results of (a) Bicubic, (b) RKHS [9], (c) SRCNN [20], (d) FSRCNN [38], (e) SCN [22], (f) PnP [1], (g) Proposed (IDCNN) and (h) GT.

the experiments, we take another conventional loss function with ℓ_2 norm,

$$\begin{aligned} Loss &= \|\mathbf{Output\_res} - \mathbf{Gt\_res}\|_F^2 \\ &= \|f_\Theta(\mathbf{Lr\_res}) - \mathbf{Gt\_res}\|_F^2, \end{aligned} \tag{5}$$

where $\|\cdot\|_F^2$ norm for matrice (or tensors) is equivalent to ℓ_2 norm for vectors.

The parameters on each layer can be obtained by

$$\widehat{\Theta} = \arg\min_{\Theta} \; Loss = \|f_\Theta(\mathbf{Lr\_res}) - \mathbf{Gt\_res}\|_F^2, \tag{6}$$

where we use the backpropagation to compute them. After defining the loss function, in what follows, we will present how to simulate the training data.

C. TRAINING DETAILS

The proposed network "Architecture 2" is based on a previous network "Architecture 1" in which we employ PnP in this paper. Thus we do not need to re-simulate the training images, *i.e.*, LR-GT image pairs, since the training image pairs have been generated in the previous network "Architecture 1". We only need to simulated "Lr_res" images and "Gt_res" images for our "Architecture 2". In particular, we could generate the "Lr_res" images simply by the subtraction of the original LR images and the downsampled "Output" images that are implemented directly by bicubic downsampling. Also, we could generate the "Gt_res" images by the subtraction of the original GT images and the "Output" images (see Fig. 2 for more details).

TABLE 1. The quantitative results for the four testing examples in Fig. 5, including the average PSNR and SSIM with the corresponding standard deviation (std). (Bold: the best).

	Bicubic	RKHS	SRCNN	FSRCNN	SCN	PnP	Proposed
				"Glass"			
PSNR	21.128	22.345	22.252	23.291	23.282	23.587	**23.656**
SSIM	0.7117	0.7205	0.7458	0.7475	0.7421	0.7463	**0.7578**
				"Sheep"			
PSNR	21.628	21.769	21.960	22.061	22.059	22.179	**22.345**
SSIM	0.5864	0.5932	0.6183	**0.6184**	0.6182	0.6181	0.6141
				"Rock"			
PSNR	27.364	28.094	28.993	29.038	29.060	29.033	**29.202**
SSIM	0.8391	0.8598	0.8644	0.8692	0.8706	0.8718	**0.8767**
				"Penguin"			
PSNR	26.572	26.805	29.203	30.077	30.081	30.041	**30.369**
SSIM	0.8565	0.8586	0.8765	0.8802	0.8813	0.8834	**0.8881**
Average PSNR	24.173±3.250	24.753±3.166	25.602±4.039	26.116±4.027	26.120±4.036	26.210±3.906	**26.393±3.982**
Average SSIM	0.7484±0.1258	0.7580±0.1278	0.7762±0.1206	0.7788±0.1226	0.7780±0.1239	0.7799±0.1244	**0.7841±0.1277**

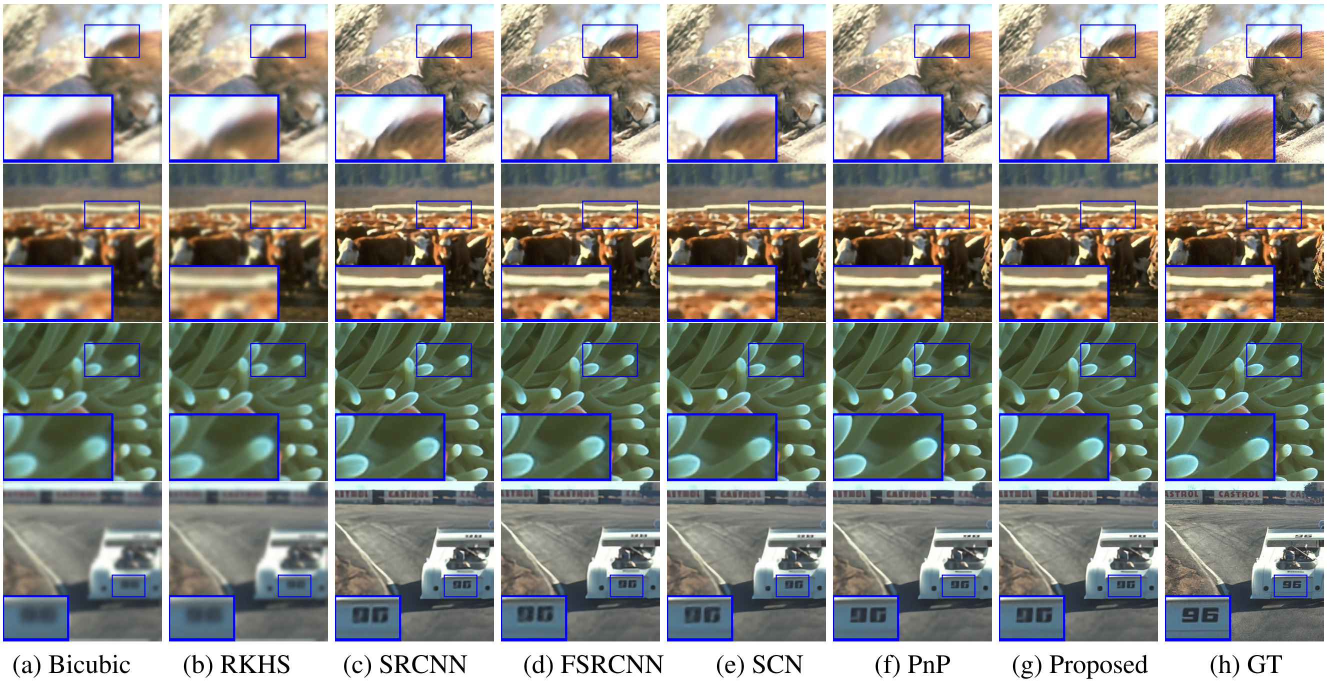
(a) Bicubic (b) RKHS (c) SRCNN (d) FSRCNN (e) SCN (f) PnP (g) Proposed (h) GT

FIGURE 6. Comparsions with some recent state-of-the-art super-resolution approaches on four examples (called "Lion", "Cow", "Coral", "Car"). The upscaling factor is 3. The results of (a) Bicubic, (b) RKHS [9], (c) SRCNN [20], (d) FSRCNN [38], (e) SCN [22], (f) PnP [1], (g) Proposed (IDCNN) and (h) GT.

Especially, the partial training images for "Architecture 2" in the work come from the test dataset, *i.e.,* BSD68 [1], [36], [37] that contains 68 natural images. We simulate the LR images by the following steps: 1) blurring each clean image by Gaussian kernels with eight standard deviations (stds); 2) downsampling the blurred images directly by bicubic interpolation. Thus we may get 544 LR-GT image pairs in this simulation. Particularly, the 544 LR-GT image pairs are divided into 80% (for training) and 20% (for testing), respectively, which indicates we have about 435 LR-GT image pairs for training and 109 LR-GT image pairs for testing. In other words, even though we do not take too many LR-GT image pairs into our network for training, it still obtains competitive results.

Moreover, the more details about the network "Architecture 2" are outlined as follows. Adam optimizer with a learning rate of 1×10^{-4} is employed for computing the network parameters.[2] The kernel size of each ResNet block is 3×3 with 32 filters. The batch size is set as 30, and the total iterations are 10000. Besides, all data are normalized into the range of [0, 1] for use. Moreover, we train the models on Python 3.5.2 with Tensorflow 1.0.1 on an NVIDIA GeForce GTX 1080 GPU with 8GB RAM.

IV. RESULTS

In this section, we compare the proposed method, called IDCNN, with six competitive image SR methods, including:

[2] Other settings for Adam are default.

TABLE 2. The quantitative results for the four testing examples in Fig. 6, including the average PSNR and SSIM with the corresponding standard deviation (std). (Bold: the best).

	Bicubic	RKHS	SRCNN	FSRCNN	SCN	PnP	Proposed
“Lion”							
PSNR	22.045	22.205	23.645	23.989	24.571	24.771	**24.804**
SSIM	0.7031	0.7510	0.8212	0.8162	0.8172	**0.8276**	0.8272
“Cow”							
PSNR	23.539	23.897	30.029	30.403	30.136	31.241	**31.562**
SSIM	0.8879	0.9148	0.9670	0.9678	0.9640	0.9697	**0.9707**
“Coral”							
PSNR	26.845	27.388	35.872	35.439	35.101	35.977	**36.076**
SSIM	0.9265	0.9342	0.9804	0.9798	0.9767	0.9846	**0.9850**
“Car”							
PSNR	21.791	21.884	26.435	27.155	26.635	26.924	**27.192**
SSIM	0.5935	0.5984	0.7574	0.7613	0.7549	0.7691	**0.7700**
Average PSNR	23.555±2.324	23.843±2.522	28.995±5.276	29.246±4.888	29.110±4.607	29.728±4.958	**29.908±4.973**
Average SSIM	0.7777±0.1568	0.7996±0.1573	0.8815±0.1097	0.8812±0.1092	0.8782±0.1095	0.8877±0.1061	**0.8882±0.1062**

1) A classical interpolation method called as “bicubic”; [3] 2) A competitive variational-based method, called “RKHS” [9];[4] 3) A benchmark method for single image SR, called SRCNN which is also the first approach for image SR using CNN [20];[5] 4) The acclerated SRCNN for single image SR, called FSRCNN [38][6]; 5) A novel CNN method with sparse priors, called SCN [22];[7] 6) A recent state-of-the-art image SR method using a plug-and-play strategy, called PnP [1].[8] Especially, we keep all default parameters along with the source codes for fair comparsions.

For the display of visual results and the quantitative evaluations, we implement them on Matlab R2017 on a desktop computer. Furthermore, we employ two popular metrics, *i.e,* peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index [39],[9] to evaluate the quantitative performance of compared approaches.

For fair comparisons, we trained the networks, i.e., SRCNN, FSRCNN, and SCN, on the DIV2K dataset[10] which is also the training dataset of PnP method. Since our IDCNN method is actually based on the PnP method, thus it is also trained on the DIV2K dataset. Note that the DIV2K dataset has 800 HD images for training and 100 HD images for validation, which could provide abundant image features for training.

In what follows, we will exhibit the performance of different compared methods from two aspects: 1) The visual and quantitative results on simulated natural images to evaluate the effectiveness of compared methods; 2) The visual results on real ToF images to validate the practical ability of image SR. Besides, we also make some discussions in this Section to adequately demonstrate the effectiveness and validation of the proposed method.

A. SIMULATED DATA

In this section, we first blur the HR noise-free images (*i.e.,* GT images) by different Gaussian kernels,[11] then downsample the blurred images to generate the simulated LR images that will be tested in the experiments (accordingly Eq. (1)). Fig. 4 exhibits the simulated LR images with different blur kernels. In Fig. 4, the first row is with a scale factor of 4, and the second row is with a factor of 3. Especially, the GT images for the corresponding simulated LR images are displayed in the last column of Fig. 5 and Fig. 6, respectively.

From Fig. 5, it is easy to know that the bicubic interpolation shows significant blur effects since the methodology of interpolation usually overlooks the image spatial details preservation. Similarly, the RKHS method also ignores the spatial details for the obtained SR images, as the given algorithm for solving the RKHS based model does not consider the blur of Gaussian kernel, it only considers the simple bicubic interpolation as a replacement. In particular, SRCNN, FSRCNN, and SCN methods could yield better visual results than the bicubic interpolation and the RKHS method since they are CNN based methods that can capture more image features on each layer, which naturally obtains better visual results. However, the three approaches fail to outperform the PnP method, as the PnP method not only considers the CNN based architecture but also can estimate the blur kernels with some existing kernel estimation approaches due to the novel formulation, *i.e.,* Eq. (2). Especially, the proposed method that is an improvement of PnP could generate better visual performance than the PnP method, as well as enhances the image resolution significantly. Correspondingly, the quantitative metrics in Tab. 1, including PSNR and SSIM, also validate the superiority of the proposed method. From the table, it is clear that our method performs best, which demonstrates

[3] Bicubic is realized by Matlab command “imresize”.

[4] Code link: https://liangjiandeng.github.io/

[5] Code link: http://kaiminghe.com/

[6] Code link: http://mmlab.ie.cuhk.edu.hk/projects/FSRCNN.html

[7] Code link: http://www.ifp.illinois.edu/~dingliu2/iccv15/

[8] Code link: http://www4.comp.polyu.edu.hk/~cslzhang/papers.htm

[9] https://ece.uwaterloo.ca/~z70wang/research/ssim/

[10] Dataset available on https://data.vision.ee.ethz.ch/cvl/DIV2K/

[11] In this work, we only consider Gaussian kernel since it is the most common case.

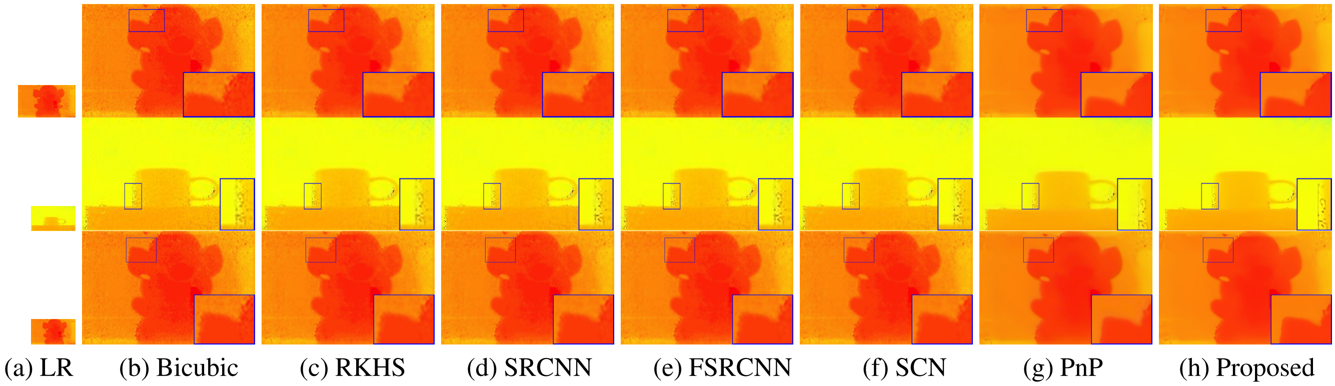

FIGURE 7. Visual results on the real ToF images. The first example is with the scale factor of 3, and the last two examples are all with the scale factor of 4. (a) LR images (with real outliers); (b) bicubic; (c) RKHS [9]; (d) SRCNN [20]; (e) FSRCNN [38]; (f) SCN [22]; (g) PnP [1] and (h) Proposed (IDCNN).

the effectiveness of our improvement to PnP. The results by our method have a larger margin than that by SRCNN, FSRCNN, and SCN, since our method also involves the kernel estimation for image SR, while the three methods are a direct CNN way for the image SR.

Fig. 6 and Tab. 2 respectively present the visual and quantitative results with the scale factor of 3. We have the *similar conclusions* as that for the scale factor of 4, which is just described in the last paragraph. Here, we do not repeat more.

B. REAL ToF DATA

In this section, we choose a particular real data, *i.e.,* ToF images, to validate the effectiveness of the proposed method. ToF image is a kind of image that contains the distance information of the detected object captured by the ToF sensor. Especially, the ToF sensor is a class of scanner-less LIDAR, in which the entire scene is captured with each laser or light pulse, as opposed to point-by-point with a laser beam such as in scanning LIDAR systems. In this work, we choose the ToF images as the real test data since we have already captured ToF images via our designed and made manufactural ToF instruments. However, the captured images often hold low image resolution and additional outliers, which motivates us to increase image resolution by a new SR method. As there are no reference images in the real ToF data, we do not show the quantitative metrics and only present the visual results in this section.

In Fig. 7, we exhibit the visual results of three real ToF images, in which the first example is with the scale factor of 3, and the last two examples are all with the scale factor of 4. Note that the LR images captured by our designed ToF instruments are often with low image resolution and corrupted outliers, thus it is quite essential to propose an efficient SR method to enhance the image resolution and suppress the outliers simultaneously. From Fig. 7, the bicubic interpolation and the RKHS method show blur effects, while the SRCNN, FSRCNN and SCN methods show clearer image structures than the bicubic and the RKHS. However, the other five approaches all fail to suppress the outliers (see Fig. 7) since they were not built for noise removal. They are just for image SR. Constrastly, the PnP method and the proposed method could not only increase the image resolution significantly but also remove the involved real outliers effectively. Notably, our method holds the better ability of outlier removal than PnP (please see the close-up in the second example), which indicates the better performance of our method.

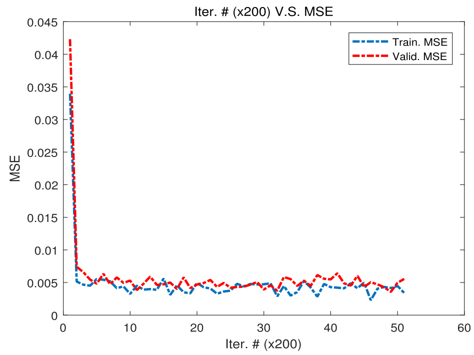

FIGURE 8. The convergence curve (mean square error, MSE) of our neural network architecture for both training dataset and validation dataset.

C. MORE DISCUSSIONS

1) THE CONVERGENCE OF OUR NEURAL NETWORK ARCHITECTURE

In this work, we propose an independent deep CNN to pick up image details to merge into the final HR image. The proposed network, *i.e.,* "Architecture 2", is actually simple but effective, and is trained on the given training dataset (see Section III-C for more details). Therefore, it is necessary to investigate the convergence property of the proposed network. Fig. 8 shows the convergence curve (calculated with the mean square error (MSE)) of our neural network architecture both for the training dataset and the validation dataset. From this figure, it is clear that the given network is converged, as well as there is not overfitting or underfitting happened in the training phase.

2) RESIDUAL RECOVERY BY THE PROPOSED METHOD

Our approach that is actually based on the previous state-of-the-art method, *i.e.,* PnP [1], it could recover more image

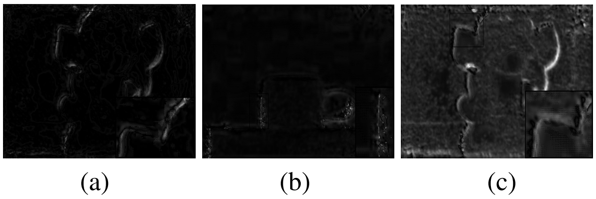

FIGURE 9. **The absolute residual maps between the PnP and the proposed method, *i.e.,* |Proposed - PnP|. The maps in (a), (b) and (c) are the absolute residual maps of the first, the second and the third example in Fig. 7, respectively.**

details based on the PnP. Therefore, it is also necessary to investigate what image details are recovered by the proposed method. Fig. 9 exhibits the absolute residual maps of the three examples in Fig. 7. From Fig. 9, it is easy to know that our method could pick up some image details to improve the quality of the final HR images, which also verifies the motivation of our method.

V. CONCLUSION

In the paper, we proposed an independent deep CNN to recover more image details from the obtained SR image. The work was motivated on an observation that there existed image residuals between the LR image and the downsampled HR output yielded by a previously proposed SR network. Extensive experiments on the simulated and the real ToF data verified the motivation, as well as the proposed method also held competitive outlier removal ability when increasing image resolution significantly. Moreover, the experimental results also validate the two mentioned contributions in the introduction.

In the future, we intend to collect more real ToF images by our designed instruments to construct a benchmark dataset for real ToF image restoration. Based on the dataset, we may design novel and useful deep CNNs for various applications of image restoration.

ACKNOWLEDGMENT

Thanks for the insightful and valuable comments of reviewers.

REFERENCES

[1] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.

[2] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.

[3] N. Mueller, Y. Lu, and M. Do, "Image interpolation using multi-scale geometric representations," *Proc. SPIE*, vol. 6498, Feb. 2007, Art. no. 64980A.

[4] P. Getreuer, "Image interpolation with contour stencils," *Image Process. On Line*, vol. 1, pp. 70–82, Aug. 2011.

[5] L. Wang, H. Wu, and C. Pan, "Fast image upsampling via the displacement field," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5123–5135, Dec. 2014.

[6] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[7] D. Capel and A. Zisserman, "Super-resolution enhancement of text image sequences," in *Proc. 15th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, Sep. 2000, pp. 600–605.

[8] R. Fattal, "Image upsampling via imposed edge statistics," *ACM Trans. Graph.*, vol. 26, no. 3, p. 95, Jul. 2007.

[9] L.-J. Deng, W. Guo, and T.-Z. Huang, "Single-image super-resolution via an iterative reproducing kernel Hilbert space method," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2001–2014, Nov. 2016.

[10] F. Wang and M. Gong, "An iterative robust kernel-based regression method for simultaneous single image super-resolution and denoising," *IEEE Access*, vol. 7, pp. 98161–98173, 2019.

[11] L.-J. Deng, W. Guo, and T.-Z. Huang, "Single image super-resolution by approximated heaviside functions," *Inf. Sci.*, vol. 348, pp. 107–123, Jun. 2016.

[12] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, pp. 25–47, Oct. 2000.

[13] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[14] C. Kim, K. Choi, and J. B. Ra, "Improvement on learning-based super-resolution by adopting residual information and patch reliability," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 1197–1200.

[15] X. Qinlan, C. Hong, and C. Huimin, "Improved example-based single-image super-resolution," in *Proc. 3rd Int. Congr. Image Signal Process.*, vol. 3, Oct. 2010, pp. 1204–1207.

[16] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[17] C. Fernandez-Granda and E. J. Candes, "Super-resolution via transform-invariant group-sparse regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3336–3343.

[18] J. Duan, J. Schlemper, C. Qin, C. Ouyang, W. Bai, C. Biffi, G. Bello, B. Statton, P. Declan O'Regan, and D. Rueckert, "VS-Net: Variable splitting network for accelerated parallel MRI reconstruction," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Shenzhen, China, Oct. 2019, pp. 713–722.

[19] J. Duan, G. Bello, J. Schlemper, W. Bai, T. J. W. Dawes, C. Biffi, A. de Marvao, G. Doumoud, D. P. O'Regan, and D. Rueckert, "Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined Multi-task deep learning approach," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2151–2164, Sep. 2019.

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 184–199.

[21] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[22] 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. Vis. (ICCV)*, Dec. 2015, pp. 370–378.

[23] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016.

[24] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[26] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.

[27] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] L.-J. Deng, T.-Z. Huang, X.-L. Zhao, and T.-X. Jiang, "A directional global sparse model for single image rain removal," *Appl. Math. Model.*, vol. 59, pp. 662–679, Jul. 2018.

[30] J. Liu, T.-Z. Huang, I. W. Selesnick, X.-G. Lv, and P.-Y. Chen, "Image restoration using total variation with overlapping group sparsity," *Inf. Sci.*, vol. 295, pp. 232–246, Feb. 2015.

[31] L.-J. Deng, M. Feng, and X.-C. Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior," *Inf. Fusion*, vol. 52, pp. 76–89, Dec. 2019.

[32] J. Liu, T.-Z. Huang, G. Liu, S. Wang, and X.-G. Lv, "Total variation with overlapping group sparsity for speckle noise reduction," *Neurocomputing*, vol. 216, pp. 502–513, Dec. 2016.

[33] J. Duan, Z. Pan, B. Zhang, W. Liu, and X.-C. Tai, "Fast algorithm for color texture image inpainting using the non-local CTV model," *J. Global Optim.*, vol. 62, no. 4, pp. 853–876, Aug. 2015.

[34] J. Duan, Z. Qiu, W. Lu, G. Wang, Z. Pan, and L. Bai, "An edge-weighted second order variational model for image decomposition," *Digit. Signal Process.*, vol. 49, pp. 162–181, Feb. 2016.

[35] L.-J. Deng, G. Vivone, W. Guo, M. Dalla Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4330–4344, Sep. 2018.

[36] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001, pp. 416–423.

[37] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[38] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 391–407.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

FEI WANG received the B.S. degree from the School of Instrumentation Optoelectronic Engineering, Beihang University, Beijing, China, in 2012. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instruments, Tsinghua University, Beijing. His current research interests include optics, laser, and image processing.

MALI GONG serves as the Director of the Center for Photonics and Electronics, Tsinghua University. As the Chief of the Photonics Expert Team, he is in charge of the national laser development plan. He also serves as the Chief Scientist of the National Keystone Basic Research Program-fundamental researcher on fiber laser. He is in charge of two international collaboration projects, namely violet fiber laser and laser communication technologies applied for intelligent transport systems. He has been engaged in laser technology and application for 20 years. He has presided over 20 research projects. He is also a Pioneer on the study of the eye-safe optical parametric oscillator (OPO) in China. He serves as the Director of Digital Video Systems, Inc.