DEEP PROGRESSIVE CONVOLUTIONAL NEURAL NETWORK FOR BLIND SUPER-RESOLUTION WITH MULTIPLE DEGRADATIONS

Jun Xiao *Rui Zhao* *Shun-Cheung Lai* *Wenqi Jia* *Kin-Man Lam**

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

Blind super-resolution (SR) of blurry and noisy low-resolution (LR) images is still a challenging problem in single image super-resolution (SISR). The performance of most existing convolutional neural network (CNN)-based models is inevitably degraded when LR images are corrupted by both blur and noise. For those blind SR methods based on kernel estimation, accurate estimation is barely attained under complex degradations and this gives rise to poor-quality results. To address these problems, we propose a deep progressive network under a probabilistic framework and a novel upsampling method for blind super-resolution with multiple degradations, which effectively utilizes image priors across scales. Experimental results show that the proposed method achieves promising performance on images with multiple degradations.

Index Terms— blind super-resolution, deep progressive network

1. INTRODUCTION

SISR aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart, which is an ill-posed problem in low-level vision tasks. In a typical SISR framework, a LR image is usually generated by downsampling a HR image using bicubic interpolation with anti-aliasing filter $\boldsymbol{H}$, and ignoring extra degradations, such as blur, noise, etc. The downsampling process is formulated as follows:

$$\boldsymbol{I}_{LR} = (\boldsymbol{H} * \boldsymbol{I}_{HR}) \downarrow_s \tag{1}$$

where $\boldsymbol{I}_{LR}$ and $\boldsymbol{I}_{HR}$ represent the synthesized LR image and the HR image. $\downarrow_s$ is the bicubic downsampling kernal with scaling factor s in a typical SISR framework.

Recently, CNN-based models for SISR have achieved significantly improved performance in terms of Peak signal-to-noise ratio (PSNR) [1, 2, 3] and perceptual quality [4, 5, 6]. However, most of these methods follow the degradation-free assumption, and belong to non-blind super-resolution approaches, i.e. downsampling kernel and degradation settings are given. Therefore, these methods achieve poor performance when they are applied to real images or the assumption is violated. To address these problems, some blind SR methods have been proposed recently, in which degradations appearing in LR images are unknown. Wang *et al.* [7] proposed patch-based method to estimate the unknown point spread function (PSF) parameters under a probabilistic framework. Michaeli and Irani [8] proposed a non-parametric approach to estimate the optimal kernel, according to recurrence of patches within a natural image. Furthermore, Shocher *et al.*[9] proposed an image-specified model to blindly super-resolve corrupted images, which uses a deep CNN model to approximate the degradation process. However, accurate kernel estimation is barely attained when the degradation process is complex. In addition to blind methods, some noisy super-resolution approaches [10, 11] have been proposed, but blurring issues are not considered in degraded process.

Besides the structural design of CNN models, learning an effective upsampling method for CNN-based models is another important issue. In general, upsampling methods used in CNN-based models can be classified into three categories, including the interpolation-based method, transposed convolution, and sub-pixel convolution. For example, bicubic interpolation was used in SRCNN [12]. Due to the high computational complexity of the preprocessing stage, FSRCNN [13] adopted transposed convolution to improve performance and speed. Sub-pixel convolution was proposed for ESPCN [14]. Instead of explicitly enlarging the resolution in the height and width channels as transposed convolution does, extra pixels in other channels are utilized and rearranged to construct the HR image. These three upsampling methods tend to generate blurry and over-smoothed images [15], so extra error is easily introduced when they are applied to super-resolving degraded images.

In this paper, we consider multiple degradations, i.e. blur and noise simultaneously, and blindly super-resolving corrupted LR images. Motivated by the cascaded structure of LapSRN [16], we combine a probabilistic graphical model with a deep convolutional network to form a multistage CNN model, which takes blurry and noisy LR images as input and progressively estimate high-frequency information based on output images of previous stages. The structure of our proposed network is different from LapSRN in two aspects. First, instead of using residual learning in LapSR, we adopt the

dense net structure to effectively reuse multi-layer features. Second, we design a novel upsampling module, namely residual upsampling, instead of using transposed convolution as LapSRN does, to further improve the performance.

2. PROPOSED METHOD

2.1. Problem Formulation

To blindly super-resolve corrupted images, a degradation model should first be defined. We follow the degradation model defined in [17], as follows:

$$\boldsymbol{I}_{LR} = ((\boldsymbol{I}_{HR} * \boldsymbol{K}) * \boldsymbol{H}) \downarrow_s + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{I}_{LR}$ and $\boldsymbol{I}_{HR}$ denote the LR image and its corresponding HR image, $*$ is the convolution operation, $\boldsymbol{K}$ represents a blurring kernel, and $\boldsymbol{\epsilon}$ is an additive white Gaussian noise.

2.2. Analysis under a Probabilistic Framework

In typical SISR framework, given a LR image $\boldsymbol{I}_{LR}$, the corresponding HR image $\boldsymbol{I}_{HR}$ is estimated by maximizing the posterior distribution, as follows:

$$\max_{\theta} p_{\theta}(\boldsymbol{I}_{HR}|\boldsymbol{I}_{LR}) \tag{3}$$

where θ denotes the parameters of the model. In our study, blurry and noisy LR images $\tilde{\boldsymbol{I}}_{LR}$ are considered. Instead of maximizing the individual posterior distribution for each scaling factor, we maximize the joint distribution of the clean LR image $\hat{\boldsymbol{I}}_{LR}$, SR$\times$2 image $\hat{\boldsymbol{I}}_{\times 2}$, and SR$\times$4 image $\hat{\boldsymbol{I}}_{\times 4}$, i.e.

$$\max\ p_{\theta}(\hat{\boldsymbol{I}}_{LR}, \hat{\boldsymbol{I}}_{\times 2}, \hat{\boldsymbol{I}}_{\times 4}|\tilde{\boldsymbol{I}}_{LR}) \tag{4}$$

where $\theta = (\theta_1, \theta_2, \theta_3)$ are the model parameters for estimating $\hat{\boldsymbol{I}}_{LR}, \hat{\boldsymbol{I}}_{\times 2}$, and $\hat{\boldsymbol{I}}_{\times 4}$ respectively. The above formulation can be decomposed as follows:

$$\begin{aligned}\max_{\theta_1,\theta_2,\theta_3}\ & p_{\theta_1}(\hat{\boldsymbol{I}}_{LR}|\tilde{\boldsymbol{I}}_{LR}) \times p_{\theta_2}(\hat{\boldsymbol{I}}_{\times 2}|\hat{\boldsymbol{I}}_{LR}, \tilde{\boldsymbol{I}}_{LR}) \\ & \times p_{\theta_3}(\hat{\boldsymbol{I}}_{\times 4}|\hat{\boldsymbol{I}}_{\times 2}, \hat{\boldsymbol{I}}_{LR}, \tilde{\boldsymbol{I}}_{LR})\end{aligned} \tag{5}$$

and can be represented by a direct probabilistic graphical model (DPGM), as illustrated in Fig.1

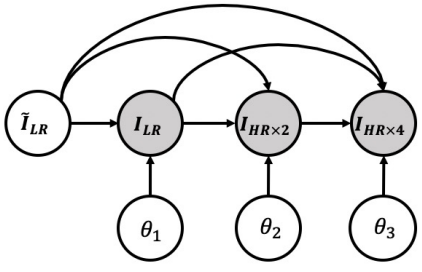


Fig. 1. DPGM model of Eqn 5. Shaded nodes represent distributions of latent images.

In Fig. 1, Eqn.(5) is represented as a progressive estimation method, where p_{θ_1} and p_{θ_2} can be viewed as a prior of p_{θ_3}, and p_{θ_1} is a prior of p_{θ_2}. In other words, $\hat{\boldsymbol{I}}_{\times 2}$ can be inferred based on the prior knowledge of $\hat{\boldsymbol{I}}_{LR}$ and $\tilde{\boldsymbol{I}}_{LR}$ learned from previous stages. Similarly, estimating $\hat{\boldsymbol{I}}_{\times 4}$ only leverages the information from $\hat{\boldsymbol{I}}_{\times 2}, \hat{\boldsymbol{I}}_{LR}$, and $\tilde{\boldsymbol{I}}_{LR}$.

2.3. Proposed Network

From the above analysis, we propose a progressive CNN model for blind super-resolution with multiple degradations. Our proposed model is denoted as PCSR, and the overall structure of PCSR is illustrated in Fig.2

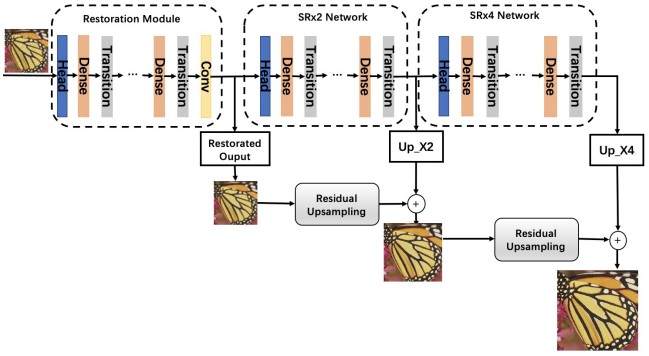


Fig. 2. The overall structure of PCSR net.

PCSR cascades three base modules: the restoration module, the SR$\times$2 module, and the SR$\times$4 module. Two Residual upsampling (Res-up) modules are used to generate super-resolved images. A blurry and noisy LR image is input to the restoration module, which converts a degraded image into a clean image. Then, the two SR modules, each with a Res-up module, are responsible for super-resolution. High-frequency information in HR image is progressively predicted based on the feature maps produced by SR networks, which combined with the estimated image generated in the previous stage to form a clean SR image.

Each base module adopts the dense connected structure, so that feature maps from previous layers can be reused effectively. The output feature map of one module contains information from that stage, and thus it can be propagated to subsequent modules, due to the use of the dense structure. The structure of the restoration module and the SR module are similar, except the output unit. The restoration module uses 1 *conv* layer, with 1×1 kernel size, as the output unit for channel information compression, while the output unit of a SR module is a sub-pixel upsampling layer for super-resolution. The head unit as the input part of each module contains 3 layers of *conv-PReLU-conv* to pre-process the input image. The dense unit stacks 3 *conv* layers alternatively, with 3×3 kernel size, for feature extraction, and they all are fully connected with each other. We apply a transition unit after a dense unit,

which includes 1 *conv* layer with kernel size 1×1 for feature map compression. In PCSR, a batch normalization (BN) layer is adopted after the *conv* layer to stabilize the network and prevent it from divergence. In all the modules, the activation function used is the *PReLU* function, followed by a BN layer.

2.4. Residual Upsampling Module

The structure of the proposed Residual upsampling (Res-up) module is shown in Fig.3. The module consists of one sub-pixel upsampling layer and 4 *conv* layers, with kernel size 3×3, followed by the *PReLU* function. The output is a *conv* layer with 1×1 kernel size. A Short connection is inserted between the output of the sub-pixel upsampling layer and the output of 4 *conv* layers.

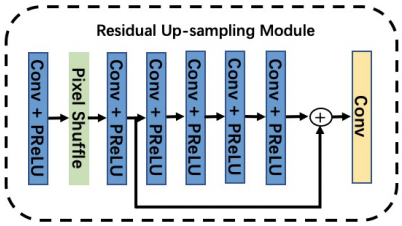


Fig. 3. The structure of residual upsampling layer

The residual upsampling module learns the residues r between the ground-truth HR image $\boldsymbol{I}_{HR}$ and the corresponding estimated SR image $\hat{\boldsymbol{I}}_{SR}$ for the SR networks, i.e.

$$r = \boldsymbol{I}_{HR} - \hat{I}_{SR} \tag{6}$$

2.5. Loss function

To maximize Eqn. (5) is equivalent to maximizing all of its three terms simultaneously. Therefore, the optimization problem can be converted into a sequential optimization problem under the MAP framework, as follows:

$$\theta_1^* = \arg\max_{\theta_1} d(F_1(\tilde{\boldsymbol{I}}_{LR}; \theta_1), \boldsymbol{I}_{LR}) \tag{7}$$

$$\theta_2^* = \arg\max_{\theta_2} d(F_2(\hat{\boldsymbol{I}}_{LR}, f_{\hat{\boldsymbol{I}}_{LR}}; \theta_2), \boldsymbol{I}_{HR\times 2}) \tag{8}$$

$$\theta_3^* = \arg\max_{\theta_3} d(F_3(\hat{\boldsymbol{I}}_{\times 2}, f_{\hat{\boldsymbol{I}}_{\times 2}}; \theta_3), \boldsymbol{I}_{HR\times 4}) \tag{9}$$

where $d(\cdot)$ is a distance metric function. $F_1, F_2,$ and F_3 with parameters θ_1, θ_2 and θ_3 represent the restoration module, SR$\times$2 network with its corresponding Res-up modules, and SR$\times$4 network with its corresponding Res-up modules, respectively. $f_{\hat{\boldsymbol{I}}_{LR}}$ and $f_{\hat{\boldsymbol{I}}_{\times 2}}$ denote the output feature map of the restoration module and the SR$\times$2 module, respectively. $\boldsymbol{I}_{LR}, \boldsymbol{I}_{HR\times 2}$, and $\boldsymbol{I}_{HR\times 4}$ denote the clean LR image, HR$\times$2 image, and HR$\times$4 image. In our method, the L_2 norm function is adopted, and thus the total loss function of our model is

$$\begin{aligned} L(\theta_1, \theta_2, \theta_3) = {} & \lambda \|F_1(\tilde{\boldsymbol{I}}_{LR}; \theta_1) - \boldsymbol{I}_{LR}\|^2 \\ & + (1-\lambda)\|F_2(\hat{\boldsymbol{I}}_{LR}, f_{\hat{\boldsymbol{I}}_{LR}}; \theta_2) - \boldsymbol{I}_{HR\times 2}\|^2 \\ & + (1-\lambda)\|F_3(\hat{\boldsymbol{I}}_{\times 2}, f_{\hat{\boldsymbol{I}}_{\times 2}}; \theta_3) - \boldsymbol{I}_{HR\times 4}\|^2 \end{aligned} \tag{10}$$

where λ is the weight to balance the restoration loss and the super-resolution loss.

3. EXPERIMENTS

3.1. Training Data and Experiment Settings

In this paper, we synthesize a LR image from its HR image according to Eqn. (1). For the blur kernel, we adopt isotropic Gaussian kernel with fixed-size kernel width. Specifically, the range of the kernel width is set to $[0.2, 2]$, with a step size of 0.1. In both the training and testing phases, kernel size is set at 15×15. Bicubic interpolation is used for downsampling, and we adopt the default setting of Matlab function *imresize*, in which an anti-aliasing filer is added before downsamping. The range of noise level, is within $[0, 50]$. We train our model from scratch, without performing any fine-tuning. The data sets include 800 DIV2K dataset [18] and 5,411 images selected from VOC 2012 [19].

We train our network using LR-HR image patch pairs. In the training phase, to synthesize LR images , the kernel width is selected randomly to blur the given HR images, which are then downsampled by bicubic interpolation, with a down scaling factor of s. Finally, Gaussian noise, with noise level σ, is added. The size of the LR image patch is set to 126×126 for both the scale factors 2 and 4. In each epoch, 12,422 LR/HR image pairs were randomly selected for training. The mini-batch size is set to 10. We optimize the loss function using Adam, and run 500 epochs in total. The learning rate is set to 10^{-5} with decay factor 0.1, shrunk in the 300th epoch and 400th epoch. The early-stopping strategy is used if the performance is not improved in 20 epochs. 5 dense units and transition units are stacked alternatively in the restoration module, and 20 (15) dense units and transition units are used in SR$\times$2 (SR$\times$4) modules. Training the model with Pytorch and 2 Nvidia Titan 1080 ti GPUs took about two days.

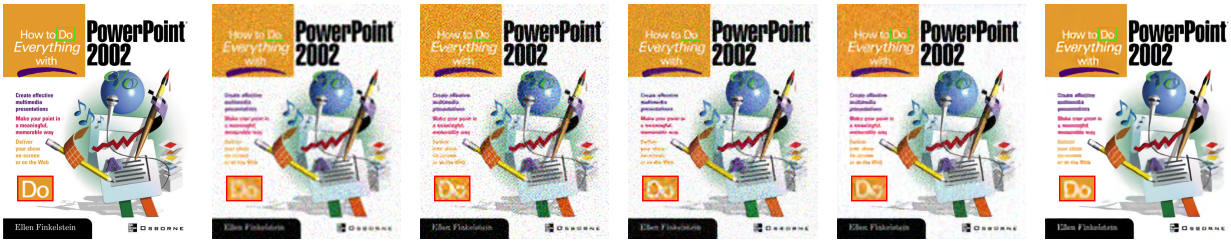
3.2. Experiment on Multiple Degradations

To evaluate the effectiveness of our proposed method for blind super-resolution with blurry and noisy image, isotropic Gaussian kernel with different widths and differnent noise levels are adopted to generate degraded images. The degradation settings are given in Table 1. We compare our proposed PCSR with bicubic interpolation, EDSR[1], ZSSR[9], SRMD[17].

The quantitative results, in terms of PSNR and the structure similarity (SSIM) index with different degradations on Set 14 and with scale factor 4, are tabulated in Table 1. We

Table 1. The average PSNR and SSIM of different methods with different degradation settings on set14, with an upscaling factor 4. The best results are highlighted in bold. **B** denotes blind super-resolution, and **NB** means non-blind super-resolution.

Degradation Setting		Bicubic(**B**)	EDSR(**B**)	ZSSR(**B**)	Ours(**B**)	SRDM(**NB**)	SRMD(**B**)
Kernel Width	Noise Level	PSNR/SSIM					
0.5	0	25.93/0.699	**28.95/0.779**	25.00/0.700	27.25/0.736	28.15/0.770	15.54/0.444
	5	25.75/0.684	**27.47/0.697**	24.63/0.640	26.89/0.704	27.64/0.749	15.01/0.344
	15	24.58/0.595	23.58/0.459	22.78/0.443	**25.14/0.575**	26.43/0.701	13.09/0.248
	50	19.38/0.303	16.01/0.162	16.84/0.160	**19.73/0.294**	23.90/0.609	12.26/0.130
1.5	0	25.23/0.668	26.75/0.714	24.61/0.671	**26.99/0.717**	28.15/0.770	17.59/0.540
	5	25.08/0.653	26.23/0.646	24.27/0.610	**26.47/0.673**	27.30/0.733	16.33/0.381
	15	24.06/0.565	23.32/0.424	22.52/0.413	**24.69/0.674**	25.93/0.676	14.13/0.235
	50	19.21/0.281	15.97/0.143	16.73/0.144	**19.57/0.271**	23.54/0.595	12.19/0.113



(a) GT (b) Bicubic(20.83dB) (c) EDSR(22.13dB) (d) ZSSR(20.45dB) (e) Ours(22.94dB) (f) SRMD(24.58dB)

Fig. 4. Visual results of ground truth (GT), bicubic interpolation, EDSR, ZSSR, PCSR (ours) and SRMD.

can observe the following: 1). When the degradation process becomes more complicated and does not follow the bicubic downsampling assumption, the performance of EDSR deteriorates seriously, and is even worse than bicubic interpolation. 2). PCSR can produce much better results than bicubic, EDSR, and ZSSR under complicated degradations, and is also comparable to SRDM, which is a non-blind CNN-based model. 3).When the kernel information and noisy information are blind, i.e. unknown, the performance of SRDM degrades significantly.

Fig. 4 compares the visual quality of the different methods. We also enlarge the region inside the green rectangle, so we can see the super-resolved results more clearly. The result based on bicubic interpolation contains lots of distortions and fine details are lost, even though it attains a high PSNR compared to ZSSR. Both bicubic interpolation and EDSR amplify the noise effect in the super-resolution process. Our proposed method can effectively reduce noise effects and produce more detailed information, compared to bicubic interpolation, EDSR and ZSSR.

Table 2. The average performance in term of PSNR and SSIM, of different upsampling methods used in super-resolution. The best results are highlighted in bold.

kernle width: 1.0, noise: 5			SP-up	Res-up w/o SC	Res-up
Set 5	×2	PSNR	30.05	30.93	**31.49**
		SSIM	0.898	0.908	**0.909**
	×4	PSNR	27.94	28.90	**29.20**
		SSMI	0.781	0.790	**0.807**
Set 14	×2	PSNR	28.73	29.45	**29.88**
		SSMI	0.843	0.853	**0.856**
	×4	PSNR	25.86	26.58	**26.82**
		SSMI	0.671	0.685	**0.694**

3.3. Ablation Study on Different Upsampling Methods

In this section, we evaluate our proposed method based on different upsampling approaches, i.e. sub-pixel convolution (SP-up), residual upsampling method without short connection (Res-up w/o SC) and residual up-sampling method (Res-up). PSNR and SSIM are measured on Set14 and Set5. The kernel width and noise level are set to 1.0 and 0.5, respectively. The quantitative results are shown in Table 2. We can observe that residual learning improves the performance. Compared with the SP-up method, Res-up combines the sub-pixel upsampling method with residual learning. The learning capacity is increased and adaptive for degradations. Therefore, the Res-up method is the most effective for degraded image super-resolution, and can lead to better performance.

4. CONCLUSION

In this paper, we proposed a progressive CNN-based model from the perspective of direct probabilistic graphical model, for blind super-resolution with multiple degradations. The main contributions in this paper are: 1). Due to the use of dense connection and a progressive strategy, our model can effectively utilize image prior across scales. 2). Combined with residual learning, a novel upsampling method is proposed for blind super-resolution. We have shown that the proposed upsampling method can produce better performance compared with existing upsampling methods. With extensive experiments, our proposed model outperforms bicubic interpolation, EDSR and ZSSR, and achieves comparable performance to a state-of-the-art non-blind model, when the images are of low-resolution and with multiple degradations.

5. REFERENCES

[1] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, 2017, vol. 1, p. 4.

[2] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[3] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1664–1673.

[4] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 105–114.

[5] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4501–4510.

[6] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," *arXiv preprint arXiv:1809.00219*, 2018.

[7] Qiang Wang, Xiaoou Tang, and Harry Shum, "Patch based blind image super resolution," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. IEEE, 2005, vol. 1, pp. 709–716.

[8] Tomer Michaeli and Michal Irani, "Nonparametric blind super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 945–952.

[9] Assaf Shocher, Nadav Cohen, and Michal Irani, "zero-shot super-resolution using deep internal learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3118–3126.

[10] Ding Liu, Zhaowen Wang, Bihan Wen, Jianchao Yang, Wei Han, and Thomas S Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, 2016.

[11] Abhishek Singh, Fatih Porikli, and Narendra Ahuja, "Super-resolving noisy images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2846–2853.

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[13] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[14] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[15] Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang, "Joint sub-bands learning with clique structures for wavelet domain super-resolution," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 165–175.

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate superresolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 2, p. 5.

[17] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, vol. 6.

[18] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.