A DEEP ENCODER-DECODER NETWORKS FOR JOINT DEBLURRING AND SUPER-RESOLUTION

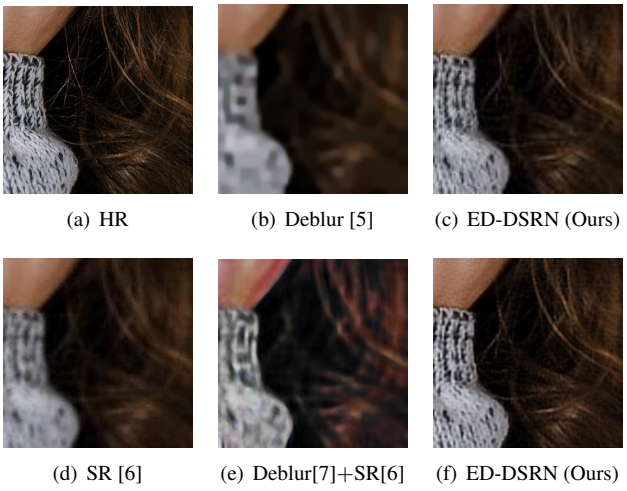
Xinyi Zhang, Fei Wang, Hang Dong, Yu Guo

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China

ABSTRACT

In this paper, we propose an end-to-end convolution neural network (CNN) to restore a clear high-resolution image from a severely blurry image. It's a highly ill-posed problem and brings tremendous challenges to state-of-art deblurring or super-resolution (SR) methods. A straightforward way to solve this problem is to concatenate two types of networks directly. However, experiments show that the concatenation of independent networks increases computation complexity instead of generating satisfying high-resolution images. Consequently, we focus on designing a single deep network to solve the deblurring and SR problems in parallel. Our method, called ED-DSRN, extends the traditional Super-Resolution network by adding a deblurring branch that shares the same feature maps extracted from an encoder-decoder module with the original SR branch. Extensive experiments show that our method produces remarkable deblurred and super-resolved images simultaneously with high efficiency.

***Index Terms*—** Super-Resolution, Blind deblurring, Joint tasks, Encoder-Decoder networks, Parallel branches



(a) HR (b) Deblur [5] (c) ED-DSRN (Ours)

(d) SR [6] (e) Deblur[7]+SR[6] (f) ED-DSRN (Ours)

Fig. 1. Deblurred images and $\times 4$ super-resolution results generated by our method compared with existing algorithms. The 1st row contains the HR images and the deblurred outputs. The 2nd row shows the results of x4 Super-resolution.

1. INTRODUCTION

Super-Resolution, enjoyed much attention and progress in recent years, aims to restore high-resolution (HR) images from low-resolution (LR) images. Super-resolution method not only generates pleasing images but also can be used as a preprocessing step for other image-related tasks, such as detection [1] and face recognition [2].

However, by long-time exposure, camera shake, de-focus and turbulent flow, images are inclined to be degradation and the application of Super-Resolution is also impeded [3]. Gaussian blur is a common model for image degradation [4]. Therefore, real-world images are often low-resolution with significant degradation and blur. It is a valuable research topic to address the problem of super-resolution and blind deblurring from severely blurry low-resolution images without category restriction.

Super-Resolution. The task of single image super-resolution (SISR) has been intensively studied for decades [8]. In recent years, the performance of SR has been promoted to a new stage by incorporating Convolutional Neural Network (CNN). The representative CNN based methods, such as SRResNet [9], LapSRN [10] and EDSR/MDSR [6], can obtain promising results when the downsampling kernel is certain.

Image Deblurring. Blind deconvolution is defined as the task of finding the latent image and a corresponding blur kernel only given an observed blurry image, which is also an ill-posed problem. Recently, deep neural networks have also been used for blind image deblurring [11, 12]. Without involving blur kernel estimation, Hradiš et al. [7] develop a deep convolutional neural network (CNN) model for text image reconstruction. However their network has been designed for deblurring and cannot be easily extended for joint super-resolution and deblurring.

Joint Super-Resolution and Deblurring. Some frameworks are proposed to estimate the downsampling blur kernels [13, 14]. However, these methods do not perform well on low-resolution images with severe blurs. The most related work comes from Xu et al. [3], who train a generative adversarial network (GAN) to learn a category-specific prior to restore a clear high-resolution image from a blurry low-resolution input. Though their method achieves huge improvement on text

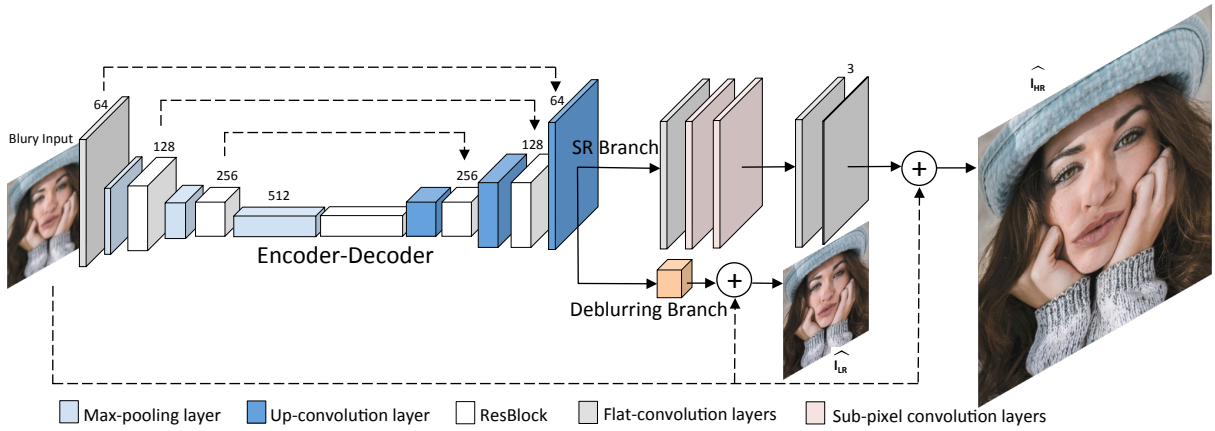


Fig. 2. Architecture of the proposed ED-DSRN model

and face images, their model needs to be retrained when it is applied to image with different category. However, the generative adversarial networks they used often suffer from unstable training process, which make their models difficult to extend to other datasets.

It is a highly ill-posed problem to restore HR images from blurry LR images. From the perspective of deblurring, it is difficult to estimate the blur kernel and restore clear image from severely blurring images, as shown in Fig.1(b). From the perspective of super-resolution, existing super-resolution methods often generate results with little high-frequency detail, as shown in Fig.1(d). Meanwhile, experiments indicate that the output deteriorates when the two tasks are simply concatenated in sequence. This phenomenon can be explained by the following reason: once trained, the super-resolution networks are only effective in one specific degradation process. When the blur kernel deviates from the training one and becomes the complex one between deblurred images and high-resolution images, the SR networks may improperly enhance some structural artifacts generated by the deblurring network. It is obvious Fig.1(e) contains too many artifacts and shadows that do not exist in Fig.1(a).

To tackle this problem, we formulate deblurring and super-resolution as two-stage task and make three enhancements as blew. First, inspired by the architecture of Mask R-CNN [15], we add a deblurring branch in parallel to the original super-resolution branch which can benefit both branches. As Fig.1(c) and Fig.1(f) show, our deblurred and SR outputs exhibit perceptually more convincing solutions compared with other methods. Second, to dispose of the disadvantages of the traditional deblurring method, we establish a deep encoder-decoder network for blind deconvolution. Furthermore, unlike [3], our method extracts the feature maps under the low-resolution scale, thus dramatically reducing the computational complexity. Our method is very suitable for the datasets that contain ultra high-resolution images, such as the newly proposed DIV2K dataset [16].

2. PROPOSED METHOD

2.1. Network Architecture

As Fig.2 shown, our network adopts a two-stage procedure: (i) an encoder-decoder module used for feature extraction over an entire image, and (ii) a two-branch network head for image deblurring and high-resolution image prediction. We refer to this network as Encoder-Decode Deblurring and Super-Resolution Network (ED-DSRN).

2.1.1. Encoder-Decoder Module

We apply one 7×7 flat-convolution layers and three 3×3 ResBlocks to build the encoder part. Three pooling layers are also used in encoder part to expand the receptive field and learn high-dimensional representation. The decoder part, consisting of three 3×3 up-convolution layers and two 3×3 ResBlocks, gradually increase the feature maps size to the original input resolution. We make some improvement in residual blocks and the skip connections as following.

Compare with original ResNet [17], we remove the batch normalization layers in our ResBlock. The batch normalization layer may reduce the flexibility of networks in some tasks [18] by normalizing the features and increase computation. We also add three skip connections between encoder and decoder part. These connections can accelerate the convergence of the network and generate much sharper outputs. In particular, we also deploy a 3×3 flat-convolution layer before element-wise addition to prevent introducing too many blurry features from the encoder side.

2.1.2. Two-Branch Network Head

We extend the super-resolution head by adding a deblurring branch that shares features from encoder-decoder module. The strong representational ability of encoder-decoder module allows us to choose more efficient heads with fewer filters.

In the super-resolution branch, we deploy three flat-convolution layers with 3×3 kernels followed by two trainable 2x sub-pixel convolution layers [9] to increase the resolutions of feature maps. Then, we add a 3×3 convolution layer before the features transform to RGB images to remove artifact as used in [19]. For the deblurring branch, we apply three 3×3 flat-convolutional layers to predict the clear low-resolution images. Due to the two branches sharing the same encoder-decoder module, the deblurring branch can improve the performance of super-resolution branch by encouraging the encoder-decoder module to generate feature maps with more high-frequency details. Furthermore, unlike these directly concatenated methods, our super-resolution branch doesn't use the deblurring image as input, avoiding entirely depending on some unsatisfactory deblurring results.

2.2. Loss function

Given a low-resolution blurry image x, the output of the proposed network can be described as:

$$\widehat{I}_{LR} = P^{db}_{\omega_1}(F_\theta(x)) + x \tag{1}$$

and

$$\widehat{I}_{HR} = P^{sr}_{\omega_2}(F_\theta(x)) + x \tag{2}$$

where F, P^{db} and P^{sr} denote the decoder-encoder module, deblurring branch and super-resolution branch respectively; θ, ω_1 and ω_2 are the network parameters.

The training goal is to make outputs the$\widehat{I}_{HR}$ and $\widehat{I}_{LR}$ close to HR and LR, the bicubic downsampling counterpart of HR, simultaneously. This process can be formulated as:

$$\min_{\theta,\omega_1,\omega_2} \mathcal{L}_{sr}(\widehat{I}_{HR}, HR) + \alpha\mathcal{L}_{db}(\widehat{I}_{LR}, LR), \tag{3}$$

where $\mathcal{L}_{sr}$ and $\mathcal{L}_{db}$ are the loss functions of SR branch and deblurring branch respectively.

Obviously, the definition of loss functions is critical for the performance of our network. The pixel-wise MSE loss is the most widely used optimization target for image super-resolution, and it is calculated as

$$\mathcal{L}_{sr} = \|\widehat{I}_{HR} - I_{HR}\|_2^2. \tag{4}$$

However, solutions of MSE optimization problems often lack high-frequency detail which results in overly smooth textures. This problem becomes worse when it is applied to deblurring branch. In order to acquire high-quality deblurring images, we use the Charbonnier penalty function as the deblurring loss function as LapSRN. The overall loss function is defined as:

$$\mathcal{L}_{db} = \frac{1}{NWH}\sum^{N}\sum^{W}\sum^{H}\sqrt{(\widehat{I}_{LR} - LR)^2 + \varepsilon^2}, \tag{5}$$

where N is the batchsize; W and H denote the dimensions of the low-resolution images; ε is a hyper-parameter where we empirically set it to $1e-3$.

3. EXPERIMENTAL RESULTS

Training. We use DIV2K [16] as training dataset, which contains all types of images. We randomly crop RGB patches of size 256×256 from the HR images and augment the HR patches with randomly rotation and horizontal flip. In order to obtain the clear low-resolution patches, we downsample the HR patches using bicubic interpolation by a factor of 4. Finally, the blurry low-resolution inputs are generated by applying a Gaussian kernel, with the standard deviation sampled from [1,2].

We train our model with 120 epochs. For optimization we use Adam with default setting in Pytorch. The learning rate is initialized to $5e-4$ for all layers and decreases by a factor of 2 for every 30 epochs. Here, the batch size is set to 32 and the trade-off weight α in loss function (3) is set to 0.2. Our Network is implemented with Pytorch framework and trained with NVIDIA 1080Ti GPU. It will takes about 30 hours for training one model.

Inference. We test our proposed networks and the state-of-the-art super-resolution and deblurring methods on the DIV2K validation set and SET14. The PSNR and SSIM of the final output with respect to its original groundtruth are calculated.

We test five state-of-the-art SR algorithms (SRResNet [9], LapSRN [10], VDSR [20], EDSR [6], SCGAN [3]) and the cascading schemes of state-of-the-art deblurring method [7] and SR networks. We also fine-tune the SRResNet and LapSRN to make a full comparison. Finally, we compare the output of our deblurring branch with two state-of-the-art deblurring methods: DarkChannel [5] and DeblurCNN [7].

Results. Table.1 shows that the proposed algorithm performs well in terms of PSNR and structural similarity (SSIM). Our two-branch architectures outperform all the other methods and significantly improve the PSNR 1.757 and 2.042 in DIV2K and Set14 datasets. SCGAN, the most related SR method, doesn't work well in our test because they are specifically designed for text and face images with motion blur. However, our validation sets contain all kinds of images with random Gaussian blur. Notice that our method is much faster than VDSR, SCGAN and EDSR, and is just a little slower than LapSRN and SRResNet.

To better explain the reasons behind the numerical results in Table.1, we present the qualitative results of some representative methods in Fig.3. As one can observe from Fig.3(g), the conventional cascaded method performs worse than directly using state-of-the-art super-resolution networks. This can be explained by that the SR network didn't learn proper parameters for restoring high-resolution images from the deblurred images.

Although visible improvement has been made by fine-tuning the SRResNet and LapSRN, they still cannot compete with our network that is trained without deblurring loss (EDSRN). It's obvious to find that Fig.3(h) contains less texture

Table 1. Quantitative evaluation of state-of-the-art SR algorithms on the DIV2K Validation set and SET14. Red indicates the best performance and blue indicates the second best. The ED-DSRN denotes our two-branch architecture.

Method	DIV2K_VAL 4x			SET14 4x		
	PSNR	SSIM	TIME(s)	PSNR	SSIM	TIME(s)
SRResNet [9]	22.6978	0.5932	0.1704	21.7681	0.5234	0.0301
LapSRN [10]	22.7343	0.6043	0.2274	21.7878	0.5346	0.1497
VDSR [20]	22.7421	0.6048	1.8358	21.7877	0.5341	0.2908
EDSR [6]	22.7508	0.6050	19.037	21.7933	0.5337	5.7320
DeblurCNN [7]+ SRResNet [9]	22.5109	0.5880	-	21.5640	0.5191	-
DeblurCNN [7]+ LapSRN [10]	22.5915	0.5963	-	21.6191	0.5299	-
DeblurCNN [7]+ VDSR [20]	22.5208	0.5970	-	21.6090	0.5292	-
DeblurCNN [7]+ EDSR [6]	22.5363	0.5985	-	21.6200	0.5301	-
SRResNet [9] Fine-tune	23.9431	0.6414	0.1704	22.9713	0.5797	0.0301
LapSRN [10] Fine-tune	23.0812	0.5894	0.2078	22.0166	0.5184	0.0287
SCGAN [3]	21.7681	0.5588	4.4056	21.9028	0.5209	0.1397
ED-DSRN (Ours)	25.7005	0.6994	0.2219	25.0131	0.6503	0.0420
ED-SRN (Ours)	25.4688	0.6927	0.2219	24.8101	0.6426	0.0420

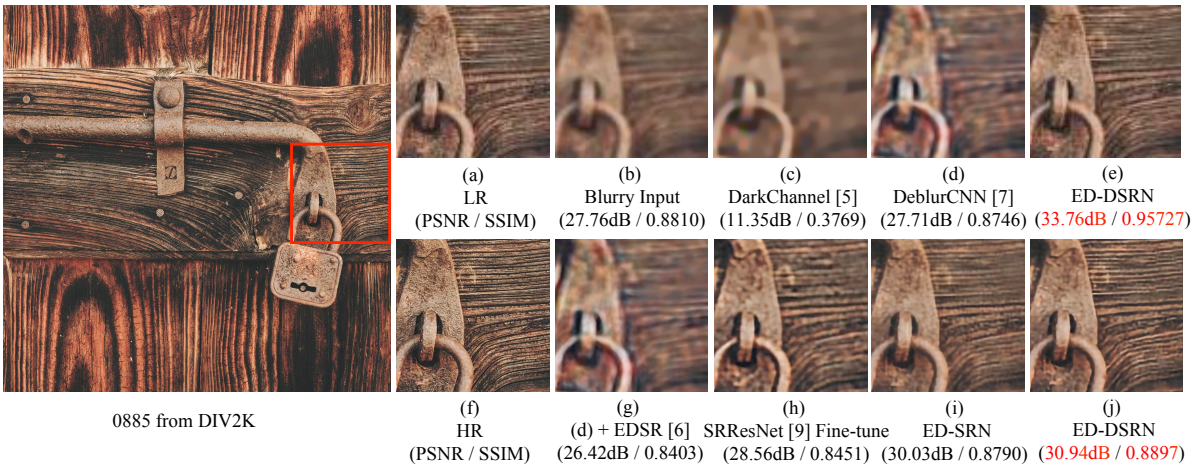


Fig. 3. Qualitative comparison of our models with other works on deblurring (1st row) and $\times 4$ super-resolution (2nd row)

details than Fig.3(i). This performance gaps mainly caused by the lack of encoder-decoder structure, which is a critical architecture when designing a blind deblurring network.

As Fig.3(j) shows, the generated image is clearer than Fig.3(i) and most wood stripes are accurately reconstructed. This improvement can be owed to the deblurring branch, who encourages the shared features contains more detail information.

The last but not the least, we also achieve improved deblurred results through the deblurring branch. As Fig.3(c)-(e) show, our methods outperform DeblurCNN and DarkChannel in both PSNR and perceptual quality. The experiments demonstrate that, by incorporating the deblurring and super-resolution loss functions as (3) shown, the two branches successfully promote each other.

4. CONCLUSION

In this paper, an encoder-decoder network with two paralleled branches is proposed to produce deblurred and high-resolution images simultaneously from low-resolution blurry inputs. We demonstrate how the encoder-decoder module and two-branch architecture can help to obtain perceptually more convincing results. Extensive evaluations show that our model achieves superior performance in terms of PSNR/SSIM and perceptual quality.

Acknowledgement. This research is supported by the National Natural Science Foundation of China under Grant 61603291 and Program of Introducing Talents of Discipline to University B13043.

5. REFERENCES

[1] Haichao Zhang, Jianchao Yang, Yanning Zhang, Nasser M Nasrabadi, and Thomas S Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 770–777.

[2] Clinton Fookes, Frank Lin, Vinod Chandran, and Sridha Sridharan, "Evaluation of image resolution and super-resolution on face recognition performance," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 75–93, 2012.

[3] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang, "Learning to super-resolve blurry face and text images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 251–260.

[4] Robert A Hummel, B Kimia, and Steven W Zucker, "Deblurring gaussian blur," *Computer Vision, Graphics, and Image Processing*, vol. 38, no. 1, pp. 66–80, 1987.

[5] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang, "Blind image deblurring using dark channel prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1628–1636.

[6] 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.

[7] Michal Hradiš, Jan Kotera, Pavel Zemcík, and Filip Šroubek, "Convolutional neural networks for direct text deblurring," in *Proceedings of BMVC*, 2015, vol. 10.

[8] Michal Irani and Shmuel Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[9] 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," *arXiv preprint arXiv:1609.04802*, 2016.

[10] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *arXiv preprint arXiv:1704.03915*, 2017.

[11] Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf, "Learning to deblur," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1439–1451, 2016.

[12] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 769–777.

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

[14] Ce Liu and Deqing Sun, "A bayesian approach to adaptive video super resolution," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 209–216.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," *arXiv preprint arXiv:1703.06870*, 2017.

[16] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al., "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1110–1121.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," *arXiv preprint arXiv:1612.02177*, 2016.

[19] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," *arXiv preprint arXiv:1612.07919*, 2016.

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.