Learning A Cascaded Non-Local Residual Network for Super-resolving Blurry Images

Haoran Bai, Songsheng Cheng, Jinhui Tang, Jinshan Pan*
Nanjing University of Science and Technology

Abstract

Deblurring low-resolution images is quite challenging as blur exists in the images and the resolution of the images is low. Existing deblurring methods usually require high-resolution input while the super-resolution methods usually assume that the blur is known or small. Simply applying the deblurring and super-resolution does not solve this problem well. In this paper, we develop an effective cascaded non-local residual network which cascades the deblurring module and super-resolution module to estimate latent high-resolution images from blurry low-resolution ones. The network first uses the deblurring module to generate intermediate clear features and then develops a non-local residual network (NLRN) as the super-resolution module to generate clear high-resolution images from the intermediate clear features. To better constrain the network and reduce the training difficulty, we develop an effective constraint based on image gradients for edge preservation and adopt the progressive upsampling mechanism. We train the proposed network in an end-to-end manner. Both quantitative and qualitative results on the benchmarks demonstrate the effectiveness of the proposed method. Moreover, the proposed method achieves top-3 performance on the low-resolution track of the NTIRE 2021 Image Deblurring Challenge.

1. Introduction

Recently, blurry image super-resolution (SR) is attracting widespread attention, and it aims to super-resolve the blurry low-resolution (LR) images to sharp high-resolution (HR) ones with rich details and clear textures. The degradation process of the blurry image SR problem can be modeled as:

$$\mathrm{y} = B(\mathrm{x}) \downarrow_s + \mathrm{n}, \tag{1}$$

where y, x, and n denote the blurry LR image, sharp HR image, and noise, respectively; $B(\cdot)$ denotes the blur process; $\downarrow_s$ denotes the downsampling operation with scale factor s.

According to the image formation (1), the degradation process contains two parts: blurring and downsampling. Thus, the direct way to tackle this challenge is to divide it into two sub-problems: deblurring and super-resolution. The recent years have witnessed significant advances in both image deblurring and image super-resolution. The success of these methods is mainly due to the use of kinds of deep neural networks [11, 23, 12, 4, 3, 8, 14, 28, 35]. Although both the image deblurring methods and the image SR methods can generate decent results, simply combining existing deblurring and super-resolution methods does not super-resolve blurry images well.

To address this issue, conventional blind image SR methods [16, 26, 19] simultaneously estimate the latent HR image and blur kernels. Although decent results have been achieved, these methods usually need to design hand-crafted priors which lead to complex optimization problems.

Instead of using hand-crafted priors, several methods develop deep convolutional neural networks (CNNs) to estimate blur kernels for single image SR [1, 6] and achieve better results than conventional methods. In addition, some methods [30, 34] exploit the relation of the image super-resolution and image deblurring problems and jointly solve them in an unified framework. However, these methods either focus on the face and text images [30], or the uniform blur [34, 1, 6]. They are not generalized well to natural images. Recent method aims to super-resolve blurry images [33] which achieves better performance than existing methods. However, the network modules for the image super-resolution and image deblurring are independent, which does not solve the images with significant blur.

Instead of estimating the super-resolution and image deblurring separately, we develop a cascaded deep neural network which cascades the deblurring module and super-resolution module in a unified framework and simultaneously solves these two modules for better image restoration. Specifically, the deblurring module is first used to generate intermediate clear features. As the intermediate clear features are estimated from low-resolution images, we then use the super-resolution module to generate clear high-resolution images from the intermediate clear features. To generate high-quality images, we develop a non-local resid-

*Corresponding author

ual network (NLRN) as the super-resolution module so that more useful features can be better explored. During the network training, we develop an effective constraint based on image gradients for edge preservation and adopt the progressive upsampling mechanism to better constrain the network and reduce the training difficulty. We solve the proposed network in an end-to-end manner and quantitatively and qualitatively evaluate it on the benchmarks to demonstrate its effectiveness.

The main contributions are summarized as follows:

- We develop a cascaded neural network which cascades the image deblurring module and super-resolution module in a unified framework and develops a non-local residual network (NLRN) as the SR module to boost the performance of blurry image SR.
- We develop an effective constraint based on image gradients for edge preservation and adopt the progressive upsampling mechanism to better constrain the network and reduce the training difficulty.
- Both quantitative and qualitative results on the benchmarks demonstrate the effectiveness of the proposed method, and it achieves top-3 performance on the low-resolution track of the NTIRE 2021 Image Deblurring Challenge [18].

2. Related Work

In this section, we briefly review the most related methods and put this work in proper context.

Image deblurring. Conventional image deblurring methods [2, 20, 21, 29] always assume that the blur is uniform. However, the motion blur in real-world images are often caused by camera shakes or fast-moving objects, which is much more complex and non-uniform. In [19], Pan et al. focus on low-resolution image deblurring and injects a super-resolution component for spatially-variant kernel estimation. Although non-uniform blur can be handled, the hand-crafted priors used in these methods often lead to a complex optimization problem which limits the deblurring performance. With the rapid development of CNNs, several deep learning-based methods [11, 23, 12, 4] propose effective networks to bypass the kernel estimation and directly recover sharp images from the blurry ones for better non-uniform blur removal.

Image super-resolution. As the image super-resolution is a highly ill-posed problem, conventional methods [31, 5, 24] develop image priors to solve this problem, whose performance is limited due to their complex optimization problems. Recently, deep CNNs-based methods [3, 8, 9, 13, 36, 14, 28, 15, 7, 35] have achieved significant improvement over these conventional methods. The SRCNN method [3] introduces CNNs into the SR problem and generates notable results. The method VDSR [8] increases the network depth for better performace by using residual architecture, and SRGAN [14] simultaneously trains a generator and a discriminator by adversarial learning for better visual perception. Especially, the method RCAN [35] develops the residual-in-residual architecture and introduces an attention mechanism on feature channels for better network representation ability. Although decent results have been achieved, these methods are designed for clean images or images with known or small blur, and cannot work well on images with motion blur.

Jointly image deblurring and super-resolution. Separately solving the deblurring and super-resolution subproblems always leads to a sub-optimal solution where the errors in these two steps may be accumulated. Conventional blind image SR methods [16, 26, 19] usually involve the blur kernel estimation and the latent HR image restoration based on hand-crafted image priors. In particular, [16] explores the internal patch recurrence for these two steps; [26] proposes an effective probabilistic combination model based on a patch-based image synthesis constraint; and [19] estimates spatially variant kernels based on the exemplars. However, using hand-crafted image priors for constraint needs to solve complex optimization problems. With the development of CNNs, several blind SR methods use CNNs for blur kernels estimation for single image SR [1, 6]. And some methods [30, 34] jointly solve the deblurring and SR in an unified manner. The method [30] trains a generative adversarial network for the blurry face and text images super-resolution, but it focuses on the face and text images and does not have sufficient capacity to handle natural images. The method [34] develops an end-to-end network for this joint problem, but it is designed for images with uniform blur, while the motion blur in captured images is always non-uniform. The method GFN [33] tries to solve the image deblurring and SR in a parallel manner and is able to handle non-uniform motion blur. It develops a gated fusion mechanism to aggregate the deblurred features and super-resolved ones to reconstruct the HR image, but the super-resolved features are still generated from the blurry LR input whose quality is limited. In contrast, we develop an end-to-end trainable deep CNN model which cascades the image deblurring and super-resolution in a unified framework, where the SR module takes the deblurred features as input so that the effects of the motion blur and noise in LR images can be reduced.

3. Proposed Algorithm

Given a blurry LR image y, the goal of the proposed method is to estimate the clear HR images x from y. To this end, we develop an effective cascaded non-local residual network which cascades the deblurring module and super-resolution module. The deblurring module first takes the blurry LR image y as input and generates intermediate clear

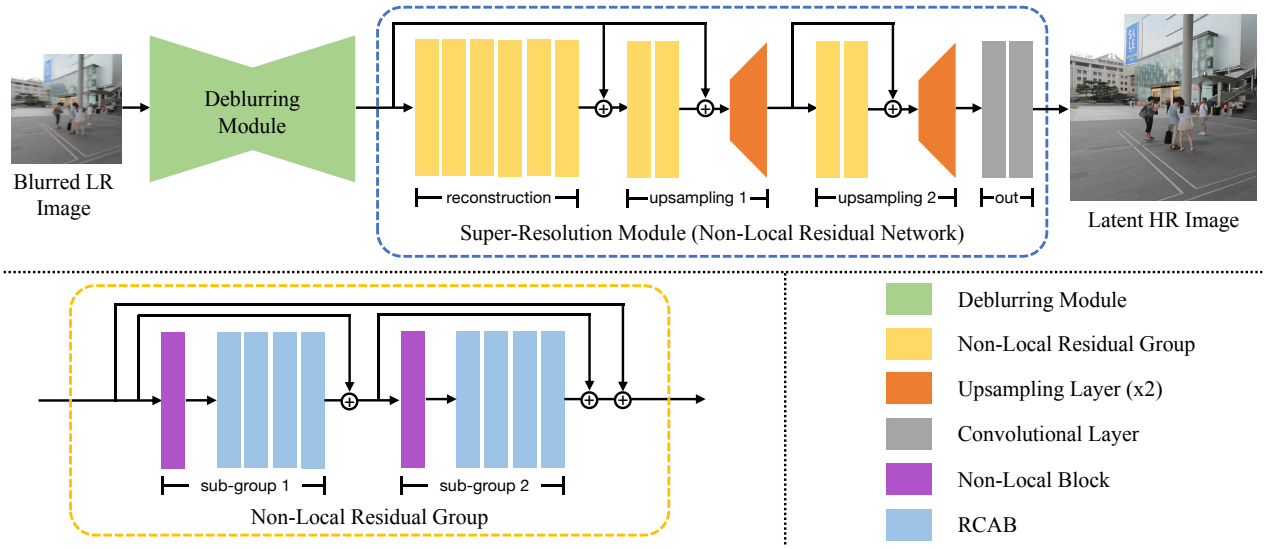


Figure 1. An overview of the proposed cascaded non-local residual network. It cascades the deblurring module and super-resolution module. The deblurring module first takes the blurry LR image as input and generates intermediate clear features. As the intermediate clear features are estimated from the low-resolution image, the super-resolution module then restores the clear high-resolution image from the intermediate clear features. Furthermore, we develop a non-local residual network (NLRN) as the super-resolution module to better generate high-quality images. The NLRN mainly consists of four parts: a reconstruction module, two upsampling modules, and an output module. In the NLRN, the non-local residual group is adopted as the basic unit. We present the detailed description of the deblurring module and super-resolution module in Section 3.1 and Section 3.2.

features. As the intermediate clear features are estimated from the low-resolution image, the super-resolution module then restores clear the high-resolution image x from the intermediate clear features. Furthermore, we develop a non-local residual network (NLRN) (which contains a reconstruction module, two upsampling modules, and an output module) as the super-resolution module to better generate high-quality images. During the training process, we further develop an effective constraint based on image gradients to preserve the edges of the recovered latent HR image. An overview of the proposed algorithm is shown in Figure 1. In the following, we first present the detailed description of the deblurring module and super-resolution module, and then explain the training strategy of the proposed method.

3.1. Deblurring Module

As the LR input contains motion blur, it is necessary to develop a deblurring module for the motion blur removal, so that the following super-resolution process can avoid the effects of motion blur. Given the blurry image y, the deblurring module generates the intermediate clear features by:

$$\mathrm{F}_{deblur} = \mathcal{N}_{deblur}(\mathrm{y}), \tag{2}$$

where $\mathcal{N}_{deblur}$ denotes the deblurring network and F_{deblur} denotes the deblurred intermediate clear features.

For the deblurring network $\mathcal{N}_{deblur}$, we adopt the similar encoder-decoder architecture as [23] as it is effective for image restoration. However different from [23], we do not use the recurrent mechanism and remove the ConvLSTM module from the deblurring network.

3.2. Super-Resolution Module

With the deblurred intermediate clear features F_{deblur}, the goal of the super-resolution module is to restore the clear HR image x. To better generate high-quality images, we develop a non-local residual network (NLRN) as the super-resolution module so that more useful features can be better explored. The proposed NLRN takes the intermediate clear features F_{deblur} as input and recovers the latent HR image x by:

$$\mathrm{x} = \mathcal{N}_{sr}(\mathrm{F}_{deblur}), \tag{3}$$

where $\mathcal{N}_{sr}$ denotes the NLRN which will be described in the following.

Non-Local Residual Network (NLRN). For the NLRN $\mathcal{N}_{sr}$, it has four parts: a reconstruction module, two upsampling modules, and an output module. The detailed architecture is shown in Figure 1.

The reconstruction module is used to refine the intermediate feature F_{deblur} for better image restoration, which is achieved by:

$$\mathrm{F}_{low} = H_{recons}(\mathrm{F}_{deblur}) + \mathrm{F}_{deblur}, \tag{4}$$

where F_{low} denotes the reconstructed low-resolution features; H_{recons} denotes the reconstruction module which consists of six non-local residual groups.

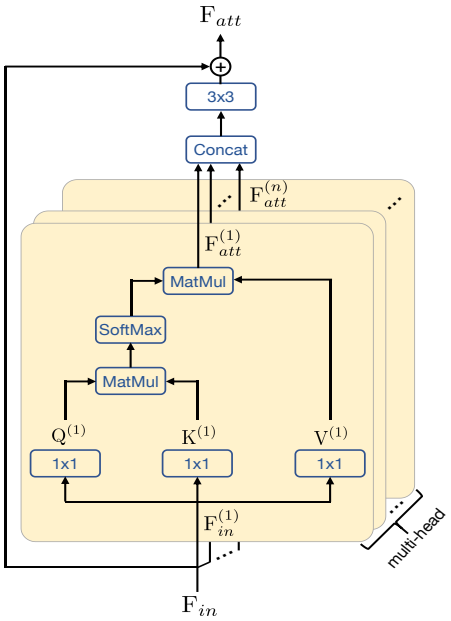


Figure 2. The architecture of the multi-head self-attention in non-local block.

Then, the two upsampling modules are used to progressively increase the resolution of the reconstructed features. In each upsampling module, it contains a refinement operation and a upsampling operation. The low-resolution features are first refined by the refinement operation and then upsampled by the upsampling operation. Thus, the upsampled features can be obtained by:

$$\begin{cases} \mathrm{F}_{up1} = H_{ps1}(H_{rf1}(\mathrm{F}_{low}) + \mathrm{F}_{low}), \\ \mathrm{F}_{up2} = H_{ps2}(H_{rf2}(\mathrm{F}_{up1}) + \mathrm{F}_{up1}), \end{cases} \tag{5}$$

where F_{up1} and F_{up2} denote the upsampled features; H_{rf1}, H_{rf2}, H_{ps1}, and H_{ps2} denote the refine operations and upsampling operations in these two upsampling modules, respectively; The refinement operation consists of two non-local residual groups, and the upsampling operation uses the pixel-shuffle followed by a convolutional layer.

Finally, the image are reconstructed by the upsampled features:

$$\mathrm{x} = H_{out}(\mathrm{F}_{up2}), \tag{6}$$

where H_{out} is the output module. It contains two convolutional layers.

Non-Local Residual Group. In the non-local residual network $\mathcal{N}_{sr}$, we use the non-local residual group as the basic unit. As shown in Figure 1, each of the non-local residual group is a residual-in-residual structure and contains two sub-groups. Each sub-group contains a non-local block and four residual channel attention blocks (RCAB from [35]).

The non-local block is effective for modeling the global information which is able to help the residual blur removal and further improve the performance of super-resolution. In the non-local block, we adopt the self-attention to explore the relationships between each image patch. Given the input features F_{in}, it first calculates the queries vector Q, keys vector K, values vector V by using 1×1 convolutions and the self-attention is represented as:

$$\mathcal{A} = Att(\mathrm{Q}, \mathrm{K}, \mathrm{V}), \tag{7}$$

where $Att(\mathrm{Q}, \mathrm{K}, \mathrm{V}) \triangleq \mathcal{S}(\mathrm{Q}\mathrm{K}^T)\mathrm{V}$, and $\mathcal{S}(\cdot)$ denotes the *softmax* operation.

In addition, we further introduce the multi-head mechanism [25] to make the non-local block focus on more diverse global correlation. Thus, the attention features F_{att} are obtained by:

$$\mathrm{F}_{att} = H_{att}(C(\mathcal{A}^1, \ldots, \mathcal{A}^j, \ldots, \mathcal{A}^n)) + \mathrm{F}_{in}, \tag{8}$$

where H_{att} denotes a convolution operation with filter kernel size of 3×3 pixels; $C(\cdot)$ denotes a concatenation operation; $\mathcal{A}^j = Att(\mathrm{Q}^{(j)}, \mathrm{K}^{(j)}, \mathrm{V}^{(j)})$; $\mathrm{Q}^{(j)}$, $\mathrm{K}^{(j)}$, $\mathrm{V}^{(j)}$ denote the query vector, key vector, value vector in j-th head, respectively; n heads are used. The architecture of the multi-head self-attention is shown in Figure 2. To reduce the computational complexity of the self-attention, we first divide the input features into 4×4 patches, and then perform the self-attention operation on these patches instead of pixels.

As the attention features are obtained from the divided patches, there may exist some errors in F_{att}, such as block artifacts. To avoid this problem, we further design a feed-forward module after the multi-head self-attention to compensate these errors:

$$\mathrm{F}_{out} = H_{feed\_forward}(\mathrm{F}_{att}) + \mathrm{F}_{att}, \tag{9}$$

where $H_{feed\_forward}$ denotes the feed-forward module which contains two convolutional layers; F_{out} denotes the output features of the non-local block.

3.3. Training Strategy

To ensure that the deblurring module and the super-resolution module could play their expected roles, the training strategy also matters.

To train the proposed network for better latent HR image restoration, we develop a simply yet effective training method which mainly contains pre-training of the deblurring module and joint training.

Pre-training of the Deblurring Module. The goal of the deblurring module is to remove the motion blur from the blurry LR images. As we know the ground truth HR image, we use the bicubic downsampled images of sharp HR images as the supervision of the deblurring module. However, the deblurring module is used to estimate features instead of images. We cannot explicitly constrain the features F_{deblur}

from the deblurring module. To overcome this problem, we introduce an additional network $\mathcal{N}_{additional}$ to generate an intermediate image from F_{deblur} so that we can use $\mathcal{N}_{additional}(\mathrm{F}_{deblur})$ and the bicubic downsampled images of sharp HR images to constrain the network training of deblurring module:

$$\mathcal{L}_{pretrain} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{N}_{additional}(\mathrm{F}^i_{deblur}) - \mathcal{B}(\mathrm{x}^i_{gt})\right\|_1, \tag{10}$$

where $\mathcal{L}_{pretrain}$ denotes the pre-training loss function; $\mathcal{N}_{additional}$ denotes the additional network which contains a convolutional layer; F^i_{deblur} denotes the intermediate clear features of the i-th blurry LR image; $\mathcal{B}(\mathrm{x}^i_{gt})$ denotes the bicubic downsampled image of i-th sharp HR image x^i_{gt}; N denotes the number of the training images.

The pre-training loss function $\mathcal{L}_{pretrain}$ and the additional network $\mathcal{N}_{additional}$ are only used in pre-training step.

Joint Training. After pre-training of deblurring module, we jointly train the deblurring module and the super-resolution module in an end-to-end manner. We adopt the widely-used pixel-wise loss function to constrain the recovered latent HR images:

$$\mathcal{L}_{pixel} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathrm{x}^i - \mathrm{x}^i_{gt}\right\|_1, \tag{11}$$

where $\mathcal{L}_{pixel}$ denotes the pixel-wise loss function; x^i denotes the i-th recovered latent HR image. However, only using (11) does not preserve the structural details well. We further develop an effective constraint based on image gradients,

$$\mathcal{L}_{grad} = \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla\mathrm{x}^i - \nabla\mathrm{x}^i_{gt}\right\|_1, \tag{12}$$

where $\mathcal{L}_{grad}$ denotes the gradient loss function; ∇ is the image gradient operator. Thus, the loss function for joint training step is:

$$\mathcal{L}_{joint} = \mathcal{L}_{pixel} + \lambda\mathcal{L}_{grad}, \tag{13}$$

where $\mathcal{L}_{joint}$ denotes the joint training loss function, and λ is the weight parameter.

4. Experimental Results

4.1. Parameter settings and datasets

Parameter settings. To ensure the deblurring module and the super-resolution module could play their expected roles, we train the proposed method in two steps, where the deblurring network $\mathcal{N}_{deblur}$ is trained firstly and then the two networks $\mathcal{N}_{deblur}$ and $\mathcal{N}_{sr}$ are jointly trained. In the training process, we adpot the ADAM optimizer [10] where the parameters β_1, β_2, and ϵ are set to be the default values of 0.9, 0.999, and 10^{-8}, respectively. For the training data, we use a similar data augmentation method to [37]. The size of the input LR patch is 64×64, and the size of the mini-batch is set to 32. In the first training step, the learning rate for the network $\mathcal{N}_{deblur}$ is set to $1e^{-4}$, and in the second step, the learning rates for the network $\mathcal{N}_{deblur}$ and $\mathcal{N}_{sr}$ are set to $1e^{-5}$ and $2e^{-4}$. The Cosine Annealing learning rate scheme [27] is adopted. The loss function weight λ is set to 0.1. We implement the proposed algorithm based on the PyTorch. The source code and trained models are available at https://github.com/csbhr/CNLRN.

Datasets. In this work, we apply the REDS dataset [17] to train and evaluate the proposed method. REDS dataset has 240 training clips, 30 validation clips, and 30 testing clips (each of them contains 100 frames). We use the 240 training clips for training, but each frame is considered as a single input without using the neighbor information. Then, we extract 300 frames (denoted by Val300) from the 30 validation clips for quantitative evaluations, and 300 frames (denoted by Test300) from the 30 testing clips for qualitative evaluations as their ground-truth images are not available. In addition, we further adopt the DVD test dataset [22] to verify the generalization ability of the proposed algorithm. As the DVD test dataset contains 10 clips with high-resolution, we use the bicubic downsampling operation to generate LR images and extract 100 frames (denoted by DVD100) from these 10 clips for quantitative evaluations.

4.2. Quantitative evaluations

The proposed method aims to super-resolve the LR images which contain motion blur, but a few methods are designed for this problem. To evaluate the proposed method, we first compare it against the image SR algorithms (RCAN [35], RRDBNet (the generator of [28])) and blind image SR method (IKC [6]). Then we compare with the method GFN [33] which jointly solves image deblurring and SR problems. As the image SR methods (RCAN [35], RRDBNet [28]) are designed for clean LR images, directly comparing with these methods may be unfair. To avoid this problem, we use the existing image deblurring method (i.e., SRN [23]) and restoration method (i.e., MPRNet [32]) to operate deblurring on blurry LR images, and then use these image SR methods to generate clear high-resolution images. And, these two steps are jointly trained. We retrain or fine-tune these methods on the REDS train dataset to choose the best results for fair comparisons, except IKC [6] which is evaluated with the provided pre-trained model. To evaluate the quality of the recovered images, we use PSNR and SSIM as the evaluation metrics.

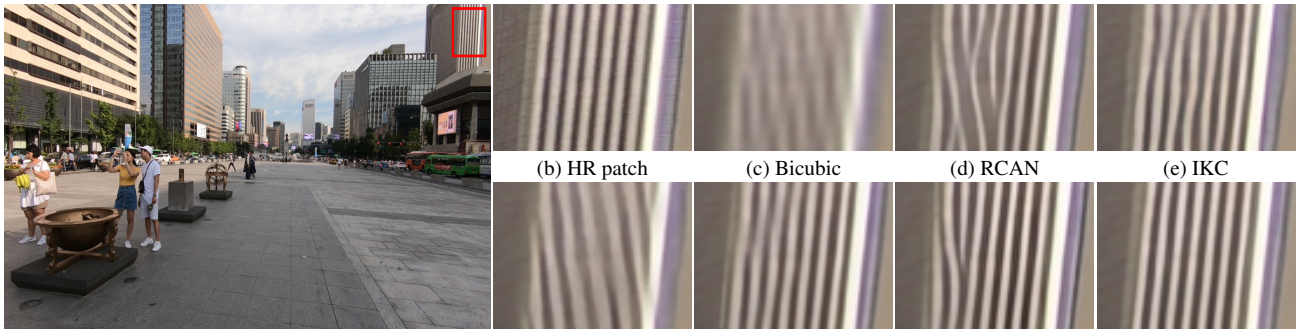
Table 1 shows the quantitative results on the Val300 dataset in terms of PSNR and SSIM. We note that directly using the image SR methods (RCAN [35], RRDBNet [28]) do not perform well as the existed motion blur

Table 1. Quantitative evaluations on the Val300 dataset in terms of PSNR and SSIM. * denotes the results generated by provided pre-trained model; # denotes the results generated by the method with self-emsemble.

Methods	Bicubic	RRDBNet	RCAN	IKC*	GFN	MPRNet+RCAN	SRN+RRDBNet	SRN+RCAN	Ours	Ours#
PSNR	23.848	27.224	27.338	24.368	26.635	27.550	27.395	27.610	**27.770**	**27.922**
SSIM	0.6481	0.7647	0.7661	0.6913	0.7447	0.7740	0.7667	0.7745	**0.7784**	**0.7813**

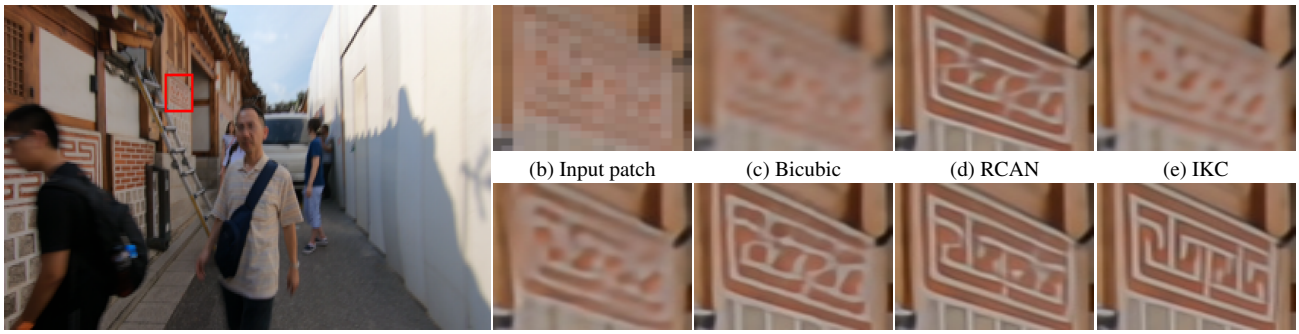
Table 2. Quantitative evaluations on the DVD100 dataset in terms of PSNR and SSIM. * denotes the results generated by provided pre-trained model.

Methods	Bicubic	RRDBNet	RCAN	IKC*	GFN	MPRNet+RCAN	SRN+RRDBNet	SRN+RCAN	Ours
PSNR	24.481	25.584	25.671	25.776	25.628	26.032	25.771	25.907	**26.089**
SSIM	0.7153	0.7769	0.7870	0.7746	0.7852	0.7876	0.7847	0.7888	**0.7974**



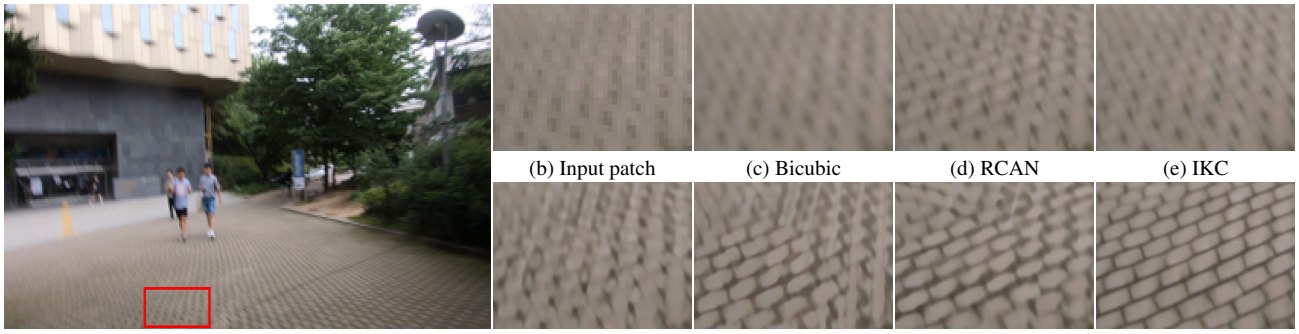
(a) Ground truth HR image (b) HR patch (c) Bicubic (d) RCAN (e) IKC (f) GFN (g) SRN+RRDBNet (h) SRN+RCAN (i) Ours

Figure 3. Comparison of the SR results on the Val300 dataset. Our method recovers high-quality images with clearer structures.



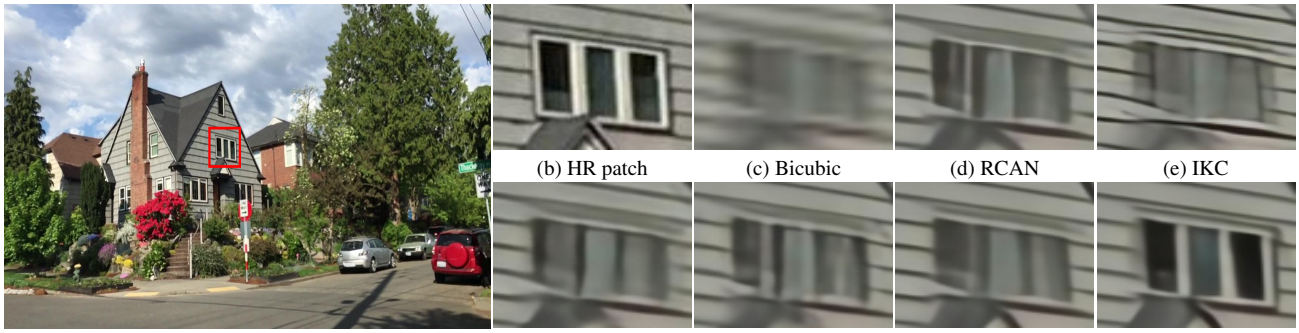
(a) Input blurry LR image (b) Input patch (c) Bicubic (d) RCAN (e) IKC (f) GFN (g) SRN+RRDBNet (h) SRN+RCAN (i) Ours

Figure 4. Comparison of the SR results on the Test300 dataset. Our method recovers high-quality images with fewer artifacts and clearer structures.



(a) Input blurry LR image (b) Input patch (c) Bicubic (d) RCAN (e) IKC (f) GFN (g) SRN+RRDBNet (h) SRN+RCAN (i) Ours

Figure 5. Comparison of the SR results on the Test300 dataset. Our method recovers clearer structures.



(a) Ground truth HR image (b) HR patch (c) Bicubic (d) RCAN (e) IKC (f) GFN (g) SRN+RRDBNet (h) SRN+RCAN (i) Ours

Figure 6. Comparison of the SR results on the DVD100 dataset. Compared with other evaluated methods, the proposed algorithm recovers better structural details.

Table 3. Results of top methods in the NTIRE 2021 Challenge on Image Deblurring.

Team	Method	PSNR	SSIM
our team	ours	**28.51**	**0.8172**
other teams	method1	28.44	0.8158
	method2	28.44	0.8135
	method3	28.42	0.8132

amplifies the difficulty of super-resolution. And the PSNR value of our method is at least 0.43dB higher than these SR methods. The blind image SR method IKC [6] does not solve the blurry image SR problem well due to that it is designed for handling the uniform blur. Although the GFN [33] method solves the deblurring and SR problems in parallel, its SR process is still operated on blurry LR images. Thus, its performance is limited. In addition, the PSNR and SSIM values of the cascaded methods (MPRNet [32]+RCAN [35], SRN [23]+RRDBNet [28], SRN [23]+RCAN [35]) are lower than our method, due to that their SR networks [28, 35] cannot effectively remove the residual blur in intermediate deblurred images. In contrast, the proposed SR network (NLRN) embeds the non-local blocks and progressive upsampling mechanism, which is able to capture global information for better residual blur removal, and further facilitate the super-resolution. Thus, it can generate favorable results against these evaluated methods.

We further evaluate our method on the DVD100 dataset. To verify the generalization ability of the proposed method, we directly apply the models which are trained on the REDS dataset. Table 2 shows that the proposed method performs favorably against the other evaluated algorithms.

In Table 3, we include the top-6 methods from the released contest results, and it shows that the proposed method is among the top-performing methods in the low-resolution track of the NTIRE 2021 Image Deblurring Challenge [18].

Table 4. Effectiveness of the deblurring on the Val300 dataset.

Methods	w/o deblurring	w/ deblurring(Ours)
PNSR	27.532	**27.770**
SSIM	0.7712	**0.7784**

Table 5. Effectiveness of the joint training on the Val300 dataset.

Methods	SRN+RCAN	SRN+NLRN	SRN+NLRN(Ours)
Joint training	✗	✗	✓
PNSR	27.012	27.052	**27.770**
SSIM	0.7543	0.7549	**0.7784**

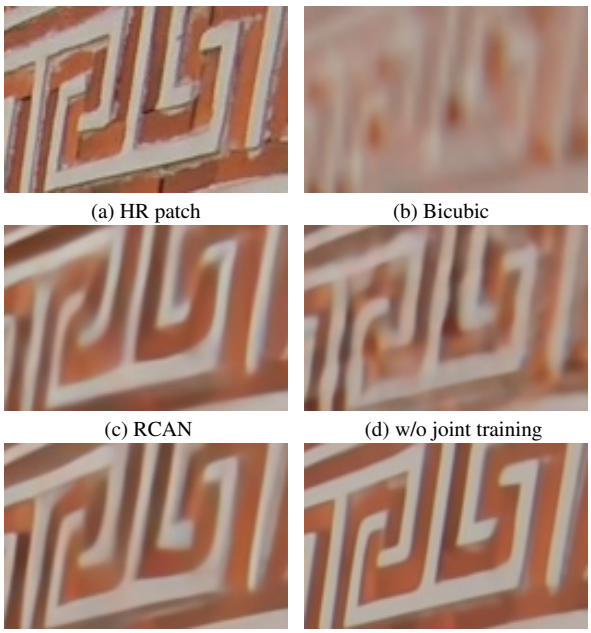
4.3. Qualitative evaluations

Figure 3 shows some visual comparisons of recovered results generated by the evaluated methods on the Val300 dataset. We note that the SR method RCAN [35] does not generate clearer results as it is mainly designed for super-resolving clean LR images which cannot handle motion blur well (Figure 3(d)). The blind image SR method IKC [6] also does not perform well as it is designed for uniform blur which is less effective for motion blur (Figure 3(e)). Although the method GFN [33] involves the motion blur removal and solves the deblurring and SR problems in parallel, its SR branch and deblurring branch are independent which means the SR process is still affected by the motion blur. Thus, its generated results still contain significant blur as shown in Figure 3(f). Furthermore, we compare with the cascaded methods (SRN [23]+RRDBNet [28], SRN [23]+RCAN [35]) where the deblurring and SR processes are jointly trained. Although these cascaded methods remove most of the blur, their generated results still contain residual blur which cannot be removed by the existed SR networks [28, 35] (Figure 3(g)-(h)). In contrast, our method generates results with clearer structures and more details.

Figure 4-6 show the visual results on the Test300 dataset and DVD100 dataset. The proposed method generates much clearer results with better structural details.

5. Ablation Studies

Deblurring. To reduce the effects of the motion blur in LR images, we develop a deblurring module before the super-



(a) HR patch (b) Bicubic (c) RCAN (d) w/o joint training (e) w/o deblurring (f) Ours

Figure 7. Effectiveness of the deblurring and the joint training.

resolution process. One may wonder whether the deblurring module helps the latent HR image recovery. To answer this question, we remove the deblurring module and retrain this baseline using the same training settings. Table 4 shows that the deblurring module is able to improve the performance of the super-resolution. And Figure 7(c) and Figure 7(e) show that the deblurring module is able to reduce the effects of motion blur and generate much clear results.

To demonstrate the effectiveness of the joint training, we separately train the deblurring module and super-resolution module. Table 5 shows that the proposed method with joint training outperforms the methods which are separately trained by a large margin in terms of PSNR and SSIM. Figure 7(d) further demonstrates that without joint training will accumulate and amplify the errors which lead to significant artifacts.

Non-local block. Although the most blur has been removed by the deblurring module, the deblurred intermediate features may still contain residual blur, which requires the following SR module to further remove the residual blur and explore more useful features. To this end, we develop the NLRN embedded the non-local blocks which can capture global information for better residual blur removal. To demonstrate the effectiveness of the non-local block, we remove the non-local blocks from the NLRN and retrain this baseline. Table 6 shows that the method without the non-local block ("baseline1" in Table 6) is less effective than the proposed method.

Progressive upsampling mechanism. Most of existed SR algorithms apply the post-upsampling mechanism, but it is less effective for the large upsample scale ($\times 4$, $\times 8$) as it makes the deep model hard to learn. Thus, we adopt the progressive upsampling mechanism to reduce the training difficulty of the SR model. To demonstrate the effectiveness of the progressive upsampling mechanism, we replace the progressive upsampling with the post-upsampling and retrain this baseline. Table 6 shows that the progressive upsampling mechanism is able to help the blurry image super-resolution.

Table 6. Ablation study of the key components of the proposed method on the Val300 dataset.

Methods	baseline1	baseline2	baseline3	Ours
non-local block		✔	✔	✔
progressive upsampling	✔		✔	✔
gradient loss	✔	✔		✔
PNSR	27.541	27.588	27.660	**27.770**
SSIM	0.7726	0.7736	0.7762	**0.7784**

Gradient loss. The widely-used loss functions (L_1-norm, L_2-norm) treat all pixels in the recovered image equally. This may lose some details like edges, especially if the input LR image contains significant blur. To avoid this problem, we develop the gradient loss function to preserve the edges of the recovered latent HR image. One may wonder whether the gradient loss helps the latent HR image recovery. To answer this question, we remove the gradient loss and retrain this baseline using the same training settings. Table 6 shows that the method without the gradient loss ("baseline3" in Table 6) is less effective than the proposed method, which demonstrates the effectiveness of the gradient loss.

6. Conclusions

We have proposed an effective cascaded non-local residual network which cascades the deblurring module and super-resolution module to estimate latent high-resolution images from blurry low-resolution ones. The proposed network first uses the deblurring module to generate intermediate clear features and then develops a non-local residual network as the super-resolution module to generate clear high-resolution images from the intermediate clear features. In addition, we have developed an effective constraint based on image gradients for edge preservation and adopted the progressive upsampling mechanism to reduce the training difficulty. The proposed network is trained in an end-to-end manner, and both quantitative and qualitative results on the benchmarks demonstrate its effectiveness. Moreover, the proposed method achieves top-3 performance on the low-resolution track of the NTIRE 2021 Image Deblurring Challenge.

Acknowledgement. We thank the organizers of the NTIRE 2021 Image Deblurring Challenge for the invitation. This work has been supported in part by the Fundamental Research Funds for the Central Universities (No. 30920041109).

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Neural Information Processing Systems*, pages 284—-293, 2019. 1, 2

[2] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. *ACM Transactions on Graphics*, 28(5):145, 2009. 2

[3] 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*, 38(2):295–307, 2015. 1, 2

[4] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019. 1, 2

[5] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *IEEE International Conference on Computer Vision*, pages 349–356, 2009. 2

[6] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 1, 2, 5, 7

[7] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018. 2

[8] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 1, 2

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016. 2

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5

[11] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018. 1, 2

[12] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *IEEE International Conference on Computer Vision*, pages 8878–8887, 2019. 1, 2

[13] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017. 2

[14] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 1, 2

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 2

[16] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *IEEE International Conference on Computer Vision*, pages 945–952, 2013. 1, 2

[17] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenges on video deblurring and super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, page 1996–2005, 2019. 5

[18] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, and Kyoung Mu Lee. Ntire 2021 challenge on image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2021. 2, 7

[19] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring low-resolution images. In *Asian Conference on Computer Vision Workshops*, pages 111–127, 2017. 1, 2

[20] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016. 2

[21] Uwe Schmidt, Kevin Schelten, and Stefan Roth. Bayesian deblurring with integrated noise estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2632, 2011. 2

[22] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017. 5

[23] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 1, 2, 3, 5, 7

[24] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126, 2014. 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 4

[26] Qiang Wang, Xiaoou Tang, and Harry Shum. Patch based blind image super resolution. In *IEEE International Conference on Computer Vision*, pages 709–716, 2005. 1, 2

[27] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1954–1963, 2019. 5

[28] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In

European Conference on Computer Vision Workshops, pages 63–79, 2018. 1, 2, 5, 7

[29] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1107–1114, 2013. 2

[30] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *IEEE International Conference on Computer Vision*, pages 251–260, 2017. 1, 2

[31] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[32] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. *CoRR*, abs/2102.02808, 2021. 5, 7

[33] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. In *British Machine Vision Conference*, page 153, 2018. 1, 2, 5, 7

[34] Xinyi Zhang, Fei Wang, Hang Dong, and Yu Guo. A deep encoder-decoder networks for joint deblurring and super-resolution. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1448–1452, 2018. 1, 2

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 1, 2, 4, 5, 7

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

[37] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *IEEE International Conference on Computer Vision*, pages 2482–2491, 2019. 5