Mathematical Problems of Computer Science 56, 18--34, 2021.

UDC 004.932

Single Image Joint Motion Deblurring and Super-Resolution Using the Multi-Scale Channel Attention Modules

Misak T. Shoyan

National Polytechnic University of Armenia
e-mail: misakshoyan@gmail.com

Abstract

During the last decade, deep convolutional neural networks have significantly advanced the single image super-resolution techniques reconstructing realistic textural and spatial details. In classical image super-resolution problems, it is assumed that the low-resolution image has a certain downsampling degradation. However, complicated image degradations are inevitable in real-world scenarios, and motion blur is a common type of image degradation due to camera or scene motion during the image capturing process. This work proposes a fully convolutional neural network to reconstruct high-resolution sharp images from the given motion blurry low-resolution images. The deblurring subnetwork is based on multi-stage progressive architecture, while the super-resolution subnetwork is designed using the multi-scale channel attention modules. A simple and effective training strategy is employed where a pre-trained frozen deblurring module is used to train the super-resolution module. The deblurring module is unfrozen in the last training phase. Experiments show that, unlike the other methods, the proposed method reconstructs relatively small structures and textural details while successfully removing the complex motion blur. The implementation code and the pre-trained model are publicly available at https://github.com/misakshoyan/joint-motion-deblur-and-sr.

Keywords: Motion deblurring, super-resolution, channel attention.

1. Introduction

Single image super-resolution (SISR) addresses the problem of recovering a sharp, high-resolution (HR) image from a given low-resolution (LR) image. The super-resolution (SR) problem has become very popular during the last decade. Its solution is beneficial for a wide range of applications such as object detection, object recognition, surveillance, etc. Image SR is an ill-posed problem since multiple possible HR images correspond to a single LR image.

In classical SR problems, the LR image is assumed to be the bicubically downsampled version of the HR image with known or small degradations. However, complicated image degradations are inevitable in real-world scenarios, and image blur is a common type of image degradation. During the image capturing process, the LR image may be degraded by various blur effects, such as motion blur, resulting from camera or scene motion during the exposure. Therefore, upscaling LR images with the SR technique will cause distorted HR results. So, it is essential to effectively combine deblurring and SR techniques to address the motion blurry image super-resolution problem. The motion deblurring problem is also highly ill-posed as multiple possible deblurred images correspond to a single motion blurry image. In this work, the problem of restoring the HR sharp image from a given motion blurry LR image is addressed.

The image degradation process consists of two parts for the motion blurry image super-resolution problem: blurring and downsampling. So, to tackle this joint problem, both deblurring and SR problems should be solved. During the last decade, deep convolutional neural networks and transformer-based [1] architectures have significantly advanced the image deblurring [2-7] and SR [8-12] techniques. Although both the image deblurring [2-7] and SR [8-12] methods generate state-of-the-art results, naively cascading them sequentially does not reconstruct blurry images well, as it is shown in [13]. There are several reasons: first, the error estimated from the first module will be magnified by the second module leading to error accumulation. Second, these two tasks are correlated, and it is sub-optimal to employ the feature extraction and image reconstruction phases twice. The features extracted from the deblurring module can be reused in the SR module to reconstruct the spatial details of the image.

Several recent methods jointly solve the motion blurry image super-resolution problem. Zhang et al. [14] proposed a dual-branch architecture to extract deblurring and SR features parallel and fuse them by the recursive gate module. However, the proposed deblurring and SR modules extract independent features, leading to sub-optimal results when blur is significant. Shoyan et al. [13] propose a single branch architecture by reusing the features extracted by the hierarchical layers of MPRNet [3] for image super-resolution. They refine the features extracted by MPRNet [3] and propagate them to the reconstruction module. However, their reconstruction module is not large enough to fully reuse the features extracted from the deblurring module.

Recently, the NTIRE 2021 Image Deblurring Challenge [15] was held where the Image Deblurring track 1 (low resolution) addresses the joint image deblurring and super-resolution problem. Several methods were proposed under the competition track 1 in the challenge. Bai et al. [16] developed a cascaded non-local residual network (CNLRN). The deblurring subnetwork is based on encoder-decoder architecture like SRN [2]. As a super-resolution subnetwork, they develop a non-local residual network to increase the resolution of the reconstructed features progressively. They propose a non-local block based on the self-attention mechanism [1] and use it in the SR subnetwork, thus achieving the 3rd PSNR [17] / SSIM [18] / LPIPS [19] scores in the NTIRE 2021 challenge [15]. However, as shown in Fig. 1, their method fails to remove the complex motion blur and recover enough sharpness for relatively small structures present in the LR blurry image. Also, the proposed network has a large number of parameters (~81M) and, therefore, has high computational complexity. Xi et al. [20] proposed a pixel-guided dual-branch attention network (PDAN). They design a residual spatial and channel attention module (RSCA) for feature extraction. They propose a Hard Pixel Example Mining (HPEM) loss to pay more attention to pixels with complicated degradation. Their method achieves the 2nd PSNR/SSIM scores in the NTIRE 2021 challenge [15]. However, the network performance mainly depends on the deep architecture and synthesized dataset. The network has about 61M parameters and considerable computational complexity as [16]. Also, the source code, pre-trained model, and network implementation details, such as the number of blocks used in feature extraction, reconstruction, and deblurring modules, are not publicly available, making their results non-reproducible. Xu et al. [21] proposed an enhanced deep pyramid network (EDPN). They adjust a

video restoration architecture to the blurry image super-resolution problem. Five replicated images are generated from the input image and fed into the network to extract the degraded LR image's self-scale and cross-scale similarity information. Their proposed pyramid progressive transfer (PPT) module performs feature alignment on the replicated images while the pyramid self-attention (PSA) module fuses the aligned features. The proposed network achieves the 1st PSNR/SSIM/LPIPS scores in the NTIRE 2021 challenge [15]. However, no pre-trained model is available for reproducing the results. Also, the publicly available source code is one of the ablation studies and has inconsistencies with some implementation details described in the paper.

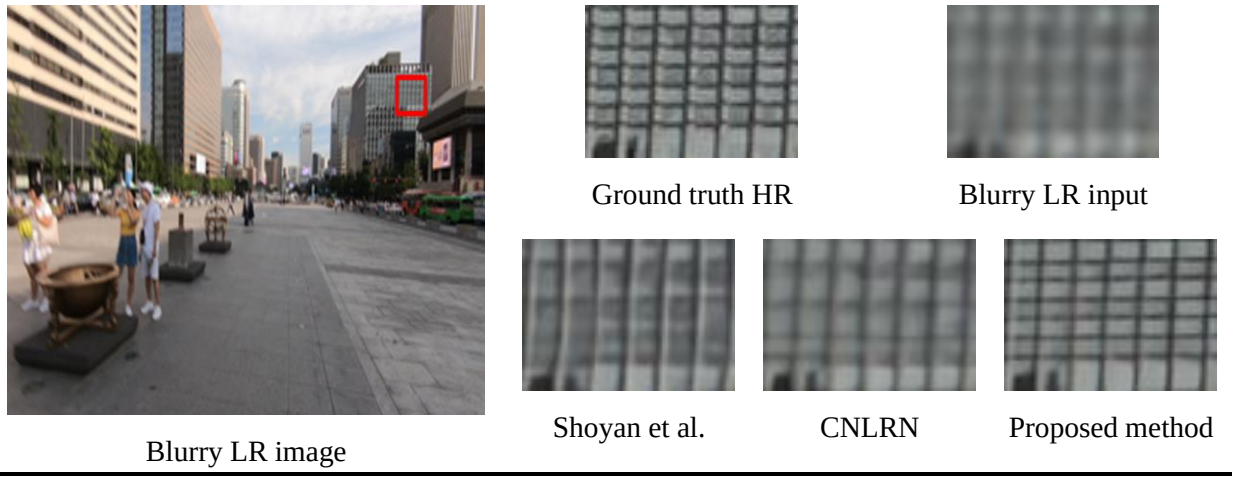

Fig. 1. Qualitative comparison between Shoyan et al. [13], CNLRN [16], and the proposed method. Please zoom in for the best view.

To tackle the problems mentioned above, a single branch architecture is proposed to effectively estimate the HR sharp image from the given motion blurry LR image. As shown in Fig. 1 (as well as in the Results section), in contrast to the other methods, the proposed method removes the complex motion blur from the small structures full of edges and generates an HR sharp image closer to the ground truth. Compared to the architectures suggested in [16] and [20], the proposed method has fewer parameters and, therefore, less computational complexity (see the Results section). Since MPRNet [3] generates contextually and spatially enriched features in its 3rd stage, the proposed architecture reuses these deblurring features in the reconstruction module to recover the HR sharp image. The reconstruction module is designed based on multi-scale channel attention modules (MS-CAM) [22] to process the small structures better. To train the whole network, a simple and effective training strategy is employed where 1) firstly, the deblurring module is trained, then 2) the SR module is trained using the pre-trained frozen deblurring module, and finally, 3) the deblurring module is unfrozen for joint end-to-end training.

The contributions of this paper are as follows:

- An end-to-end single branch architecture is proposed to effectively reconstruct the HR sharp image from the given motion blurry LR image.
- A reconstruction module is designed based on MS-CAM blocks [22] to better process the small structures in the LR blurry image.
- A simple and effective training strategy is exploited where the pre-trained frozen deblurring module is used as a feature extractor to train the SR module. Then the whole network is jointly trained with the unfrozen deblurring module.

The source code and the pre-trained model are publicly available at https://github.com/misakshoyan/joint-motion-deblur-and-sr.

2. Related Work

This section briefly reviews the motion deblurring, super-resolution, motion blurry super-resolution techniques and the related attention mechanism.

Motion deblurring. The motion deblurring problem for dynamic scenes is highly ill-posed since the blur kernel is spatially varying and unknown (non-uniform blind deblurring problem). Conventional blind motion deblurring methods jointly estimate the blur kernel and sharp image by solving computationally expensive optimization problems [23, 24]. Simplified priors are employed to regularize the solution space, such as dark channel prior [24], total-variation prior [23], etc. However, the simplified assumptions on the blur model make these techniques non-generalizable for real-world applications.

Recently, deep convolutional neural networks (CNN) have achieved significant success in image deblurring [2-7]. Since the blur kernel estimation is not practical in real-world applications, recent methods directly map the blurry image to the corresponding sharp image by designing an end-to-end image deblurring architecture. Nah et al. [6] proposed a deep multi-scale CNN called DeepDeblur to mimic the conventional coarse-to-fine optimization approaches. Tao et al. [2] proposed a scale recurrent network (SRN) and improved the DeepDeblur [6] network by sharing the network weights at different scales where each scale employs an encoder-decoder architecture. Zamir et al. [3] proposed a multi-stage progressive image restoration architecture called MPRNet. They employ an encoder-decoder subnetwork in the first and second stages to learn contextually enriched features due to large receptive fields. In contrast to the first and second stages, the last stage does not contain a downsampling operation. It employs a single-scale pipeline to generate spatially enriched features via channel attention blocks (CAB) [3]. A supervised attention module (SAM) is introduced to control the information flow between the stages, which refines the features before propagating them to the next stage. The MPRNet achieves competitive results in motion deblurring, image denoising, and deraining problems. Chu et al. [7] further improved the MPRNet results by exploring the statistics distribution inconsistency between training (with image patches) and testing (with full-size image) phases. They argue that the operations, which aggregate the global spatial statistics along the entire spatial dimension, may lead to a statistics distribution shift between training and testing phases since different size images are used in these phases. This statement mainly refers to the global average pooling operation used in CAB. To address this issue, they propose a test-time local statistics converter (TLSC) mechanism to calculate the mean value in a local window for each pixel rather than calculating a single mean value for the entire spatial dimension. They replace the global average pooling layer of CAB with TLSC-based local average pooling layers for each pixel at the test-time without re-training or fine-tuning the network. Zamir et al. [4] proposed a transformer-based [1] encoder-decoder architecture called Restormer. They calculate the self-attention [1] across channels rather than the spatial dimension, thus implicitly encoding the global context with linear complexity.

Super-Resolution. Like the motion deblurring problem, the SR problem is also highly ill-posed. Early approaches employ interpolation techniques such as bicubic and bilinear interpolations [25, pp. 77-78]. Some methods rely on other techniques such as neighbor embedding [26], sparse coding [27], etc. CNN-based methods [8-10] have achieved unprecedented success in single image SR during the last decade. Zhang et al. [8] designed a very deep residual channel attention network (RCAN). They focus on learning high-frequency information by employing residual in residual structure with long and short skip connections to bypass the abundant low-frequency information. The channel attention (CA) mechanism is proposed to model the interdependencies among feature channels inspired by the Squeeze and Excitation mechanism [28]. They design residual channel attention block (RCAB) using Conv-

RELU-Conv structure followed by CA block and construct the network based on RCABs. Dai et al. [9] designed a second-order attention network (SAN) by introducing second-order channel attention (SOCA) block to learn the channel-wise feature interdependencies better and to focus on more informative features using second-order channel statistics. They exploit share-source skip connections to bypass more abundant low-frequency information present in the LR image. Niu et al. [10] proposed a holistic attention network (HAN) based on RCAN [8]. The proposed layer attention module (LAM) allows the network to learn the interrelationships between features of different layers and focus on more informative layers. A channel-spatial attention module (CSAM) is introduced to learn the inter-channel and intra-channel dependencies for the last layer of the network.

However, the discussed SR methods assume known or small degradations and amplify the blur present in the LR image, as shown in [13]. Therefore, the SR network needs also to incorporate motion deblurring techniques.

Joint motion deblurring and super-resolution. Several recent methods solve the motion blurry image super-resolution problem by incorporating motion deblurring and SR techniques in a unified network. Zhang et al. [14] proposed a gated fusion network (GFN) to extract deblurring and SR features separately by designing a dual-branch architecture. The gate module fuses these features, and the reconstruction module generates an HR sharp image using the fused features. However, in their design, the SR branch operates on the blurry LR image, which limits the reconstruction ability of the network to generate sharp HR results, as shown in [13].

In contrast to GFN, Shoyan et al. [13] proposed a single branch architecture. They suggest reusing the features extracted by the hierarchical layers of MPRNet [3] since it generates contextually and spatially enriched features in its deep layers. These features are then refined and fed to the reconstruction module to generate the four times upscaled sharp HR image. However, it seems that their reconstruction module could be a bit larger to reuse the features extracted from the deblurring module fully.

Several methods were proposed in the scope of NTIRE 2021 Image Deblurring Challenge [15] under the Image Deblurring track 1 (low resolution) to solve the joint motion deblurring and SR problem. Bai et al. [16] cascaded the deblurring and super-resolution modules in a unified network. The deblurring module employs an encoder-decoder architecture like SRN [2], without a recurrent mechanism. As a super-resolution subnetwork, they develop a non-local residual network that contains RCABs [8] and self-attention-based [1] non-local blocks. The non-local block aims to model the global information for residual blur removal. A gradient loss function is developed to preserve the edges of the reconstructed HR image. They also propose a progressive upsampling mechanism to increase the resolution of the reconstructed features progressively and achieve the top-3 PSNR [17] / SSIM [18] / LPIPS [19] scores in the low-resolution track 1 of the NTIRE 2021 challenge [15]. However, as shown in Fig.1 (see also the Results section), this method fails to recover enough sharpness in its HR output when the relatively small structures of the LR blurry image are full of edges. Also, the non-local blocks and progressive upsampling mechanism have considerable computational complexity due to the self-attention mechanism [1]. In addition, the deblurring module has about 70M parameters because of the large convolutional kernel size.

Xi et al. [20] proposed a pixel-guided dual-branch attention network. The proposed residual spatial and channel attention module aims to better extract informative features by fusing cross-channel and spatial information. The deblurring module is based on an encoder-decoder architecture that employs residual blocks. Unlike the other architectures, the deblurring module is used only in the training stage for computational efficiency and helps the network extract and learn more useful deblurring information. The SR module is mainly composed of convolutional layers and pixel shuffling layers [29]. They argue that some pixels of the LR

degraded image may contain a large amount of blur and downsampling degradation. In contrast, the other pixels may be only affected by downsampling, so they propose the Hard Pixel Example Mining loss to pay more attention to pixels with complicated degradation. Their network achieves top-2 PSNR/SSIM scores on track 1 of the NTIRE 2021 challenge [15]. However, the network has considerable computational complexity as [16] since it has about 61M parameters. Unlike other methods, the network's performance depends on the additionally synthesized dataset (~72K images) used to fine-tune the network. In addition, the source code and the implementation details of the network (such as the number of blocks used in feature extraction, reconstruction, and deblurring modules) are not publicly available for researchers to reproduce the results.

Xu et al. [21] adjusted a video restoration architecture to the blurry image super-resolution problem by feeding five replicated images into the network to fully exploit the degraded LR image's self-scale and cross-scale similarity information. Their proposed pyramid progressive transfer (PPT) module employs a pyramid structure and aims to generate attention masks to progressively transfer the self-similarity information. The pyramid self-attention (PSA) module aggregates and reweights the transferred features with a pyramid structure. Their proposed network achieves the best PSNR/SSIM/LPIPS scores in the NTIRE 2021 Image Deblurring Challenge [15]. However, the pre-trained model is not provided for researchers. The publicly available source code is one of the ablation studies of EDPN that has some inconsistencies with the paper. Therefore, the method results are not reproducible, like [20].

Attention mechanism. The attention mechanism aims to mimic the human visual perception to pay attention to the informative part of the input. The self-attention [1] mechanism allows modeling the global dependencies between each word in a sentence or pixel in the image. Hu et al. [28] proposed the Squeeze and Excitation block (SE) to model the interdependencies between the feature channels. Dai et al. [9] designed the second-order channel attention block by replacing the global average pooling operation of the SE block with the global covariance pooling operation [9] to capture higher-order feature statistics for more discriminative representations. To solve the problems of scale variation and small objects, Dai et al. [22] designed the multi-scale channel attention module (MS-CAM) for aggregating both the local and global feature contexts within the channel attention. The global average pooling operation employed in the SE block emphasizes the globally distributed large objects while potentially ignoring the locally distributed small objects. Therefore, in addition to the global branch, a local branch is added into SE to simultaneously aggregate the global and local contexts within a multi-scale channel attention mechanism.

3. Proposed Method

This work proposes a single-branch network architecture to solve the joint motion deblurring and SR problem. Unlike the existing methods, the proposed method successfully reconstructs the relatively small degraded structures (see the Results section). The network reuses the contextually and spatially enriched features generated by MPRNet [3] for image super-resolution. The SR module is designed by employing residual in residual structure based on multi-scale channel attention modules (MS-CAM [22]), which simultaneously aggregate the local and global feature contexts within the channel attention to better process the small objects.

Network Architecture. To solve the motion deblurring and SR problems jointly, the network should extract both contextually and spatially informative features. The encoder-decoder architectures [2, 4] are effective in encoding the contextually informative features, which are helpful for the motion deblurring problem. The single-scale pipelines [8-10] extract

spatially-enriched features that are informative for the SR task. The MPRNet [3] naturally offers a twofold functionality since it employs both encoder-decoder and single-scale pipelines.

The proposed network consists of three main parts: deblurring, feature transformation, and super-resolution (see Fig. 2).

Deblurring Module. The deblurring module is based on MPRNet [3] and employs multi-stage progressive architecture with three stages (see Fig. 2). An encoder-decoder architecture is employed in the first and second stages of MPRNet based on U-Net architecture [30] to generate contextually informative features due to the large receptive field. Each scale of the encoder and decoder networks uses 2 channel attention blocks (CAB) [3]. Also, each skip connection between encoder and decoder uses 2 CABs. The CAB block incorporates Conv-RELU-Conv structure and Squeeze-and-Excitation block (SE) [28] with the residual connection. It reweights the input feature map by emphasizing more informative channels. The 3rd stage employs a single-scale original-resolution subnetwork (ORSNet) [3] and aims to generate spatially informative features without downsampling and upsampling operations. It contains 3 original-resolution blocks (ORB) [3]. Each ORB is a residual group with 8 CABs followed by a convolution layer.

The supervised attention module (SAM) [3] and the cross-stage feature fusion (CSFF) [3] mechanism are employed to control the information flow between two consecutive stages. The SAM module uses the ground-truth supervisory image. It generates attention maps to suppress the less informative features of the current stage and to propagate only the more informative ones to the next stage. The CSFF mechanism propagates the intermediate contextualized features of the previous stage to the next stage. It aims to compensate for the information loss due to repeated upsampling and downsampling operations performed in the first and second stages. Thus, the SAM module and CSFF mechanism allow the ORSNet to combine the generated spatially informative features with the contextually informative features of the previous stage and allow the MPRNet to produce both contextually and spatially enriched features in its 3rd stage.

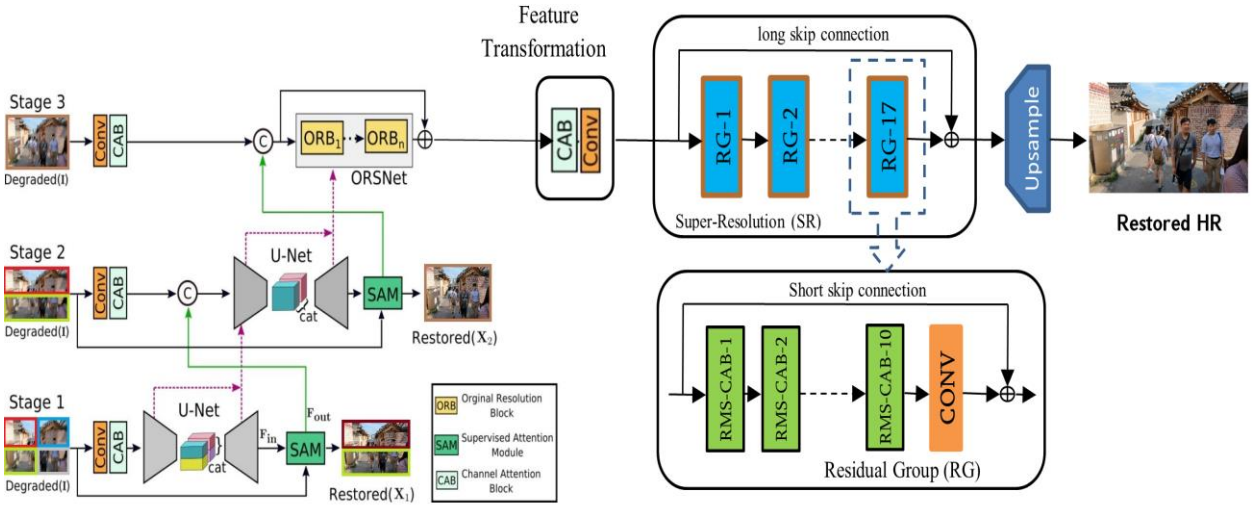

Fig. 2. The architecture of the proposed method. Please, zoom in for the best view.

Feature Transformation Module. The feature transformation module refines the contextually and spatially enriched features generated by MPRNet [3] before propagating them to the super-resolution module. It contains a CAB block followed by a convolutional layer. A CAB block is applied on MPRNet generated features to emphasize the more informative features for SR and suppress the less informative ones. Since the third stage of MPRNet operates on 128

channels, a convolutional layer transforms the CAB-generated features by decreasing the number of channels from 128 to 64. It aims to achieve a trade-off between computational complexity and accuracy for the SR.

Super-resolution Module. The super-resolution module takes the transformed features as input and employs a single-scale pipeline to reconstruct the HR sharp image. The image SR problem can be treated as a process to recover the high-frequency information (like regions full of edges) as much as possible since the low-frequency information is less informative (like uniform regions). Thus, the abundant low-frequency information can be forwarded to the final HR output with relatively less processing. Inspired by RCAN [8], a residual-in-residual (RIR) structure is exploited to bypass the abundant low-frequency information via long and short skip connections and make the network focus on learning the high-frequency information. The RIR structure stacks 17 residual groups (RG), where each residual group contains 10 residual multi-scale channel attention blocks (RMS-CAB) followed by a convolution layer (see Fig. 3). After the RIR structure, the pixel-shuffling layer [29] follows to upscale the spatial resolution of the features four times. Then a convolutional layer is used to generate a colored image from 64 channels.

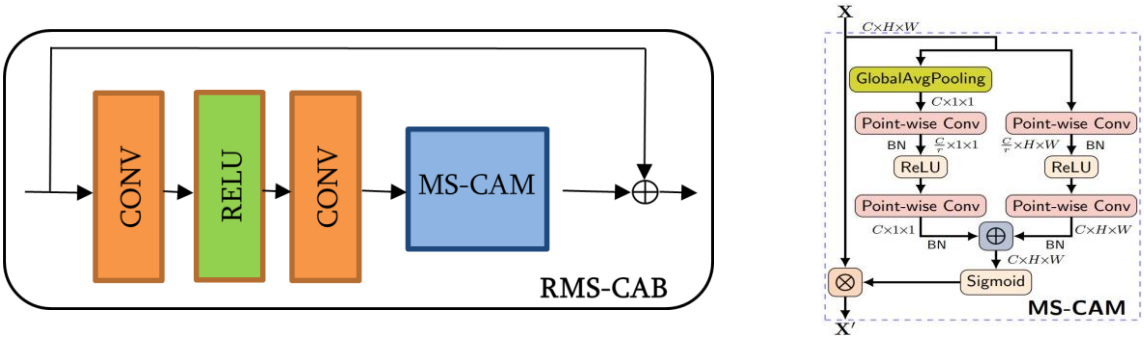

Fig. 3. The RMS-CAB block and MS-CAM [22] block. Please, zoom in for the best view.

The employed RMS-CAB block is inspired by CAB [3]. It consists of a Conv-RELU-Conv structure and MS-CAM [22] block with the residual connection. Different from CAB, it employs MS-CAM [22] block instead of SE block [28] to aggregate both the local and global feature contexts. The SE and MS-CAM blocks aim to capture the interdependencies among feature channels and reweight them by amplifying the more informative channels and suppressing the less informative ones. In contrast to the SE block, the MS-CAM block captures the information from the globally distributed large and locally distributed small objects, while the SE block ignores the signals present in small objects. The MS-CAM block employs global and local branches (see Fig. 3) to implement the channel attention in two scales. The global branch applies a global average pooling operation and calculates the mean value for each channel to capture the channel-wise feature statistics globally. Then two point-wise convolutions follow with RELU activation between them. Unlike the global branch, the local branch operates on the original spatial resolution and does not exploit the global average pooling operation. The features generated by two branches are aggregated via broadcasting addition. Then a sigmoid activation follows, and the final features are element-wise multiplicated with the input.

Training Strategy and Loss Function. The whole network can be trained jointly from scratch, but this approach causes sub-optimal results. The reason is that for the most part of the joint training, the deblurring module will generate blurry results, which will hamper the training process of the SR module. As a result, the network will not learn the exact mapping between blurry LR and sharp HR images. Since it would be beneficial for the SR module to operate on deblurred features, a simple and effective training strategy is exploited, consisting of three phases.

In the first training phase, only the deblurring module (MPRNet) is trained. The MPRNet generates LR deblurred image at each stage, denoted as ($\hat{L}_1$, $\hat{L}_2$, $\hat{L}_3$). Since, in the proposed architecture, the third stage of the MPRNet does not produce a deblurred image, a convolutional layer is added to generate the $\hat{L}_3$ image for the training phase only. The deblurring loss is defined as

$$\mathcal{L}_{\mathrm{DB}} = \sum_{\mathrm{s}=1}^{3} \mathcal{L}_{\mathrm{char}}\left(\hat{\mathrm{L}}_{\mathrm{s}}, \mathrm{L}\right) + \lambda_{edge} * \mathcal{L}_{\mathrm{edge}}(\hat{\mathrm{L}}_{\mathrm{s}}, \mathrm{L}),$$

$$\mathcal{L}_{\mathrm{char}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \sqrt{||\hat{\mathrm{L}}_s^i - \mathrm{L}^i||^2 + \varepsilon^2},$$

$$\mathcal{L}_{\mathrm{edge}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \sqrt{||\Delta\hat{\mathrm{L}}_s^i - \Delta\mathrm{L}^i||^2 + \varepsilon^2},$$

where $\mathcal{L}_{\mathrm{DB}}$ denotes the deblurring loss, N is the number of training images, and L is the bicubic downsampled version of the ground-truth HR image. As in [3], $\mathcal{L}_{\mathrm{char}}$ and $\mathcal{L}_{\mathrm{edge}}$ denote the Charbonnier loss [31] and the Edge loss [3], Δ denotes the Laplacian operator. ε and λ_{edge} were empirically set to 10^{-3} and 0.05, respectively.

In the second training phase, only the SR module is trained (including the transformation module) by employing the pre-trained frozen deblurring module. It is beneficial for the SR module since, in contrast to joint training from scratch, it starts to operate on fully processed deblurred features and does not have to deal with blurry features now. To preserve the structural details, like edges, the combination of pixel-wise and gradient losses ($\mathcal{L}_{\mathrm{pixel}}$ and $\mathcal{L}_{\mathrm{grad}}$) is employed as in [16]:

$$\mathcal{L}_{\mathrm{pixel}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \left\| \hat{\mathrm{H}}^i - \mathrm{H}^i \right\|_1,$$

$$\mathcal{L}_{\mathrm{grad}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \left\| \nabla\hat{\mathrm{H}}^i - \nabla\mathrm{H}^i \right\|_1,$$

$$\mathcal{L}_{\mathrm{SR}} = \mathcal{L}_{\mathrm{pixel}} + \lambda_{grad}\mathcal{L}_{\mathrm{grad}},$$

where $\hat{\mathrm{H}}$ and H are the reconstructed and ground-truth high-resolution images, respectively. $\mathcal{L}_{\mathrm{SR}}$ denotes the SR loss, ∇ is the image gradient operator and λ_{grad} is set to 0.1 as in [16].

In the last training phase, the deblurring module is unfrozen, and the whole network is trained jointly. The deblurring loss is employed only in the first and second stages of the MPRNet. This trick aims to force the third stage of MPRNet to pay more attention to generating the SR features during the last phase of the training. The following combination of loss functions is exploited

$$\mathcal{L} = \mathcal{L}_{\mathrm{SR}} + \alpha\mathcal{L}_{\mathrm{DB}},$$

where α was empirically set to 0.5 as in [13].

Implementation Details. The REDS dataset [32] is used to train the proposed network. It contains 24.000 training images, 3000 validation images, and 3000 testing images. The validation dataset is used for quantitative evaluations since the ground-truth images of the testing

set are not available. The size of the LR blurry and ground-truth HR images is equal to 320x180 and 1280x720, respectively.

The proposed method is trained with the three-phase training strategy as discussed above. The training patch size for the blurry LR and ground-truth HR images is 64x64 and 256x256, respectively, which are randomly cropped from the training set. Randomly horizontal and vertical flips are applied combined with rotation for data augmentation. The Adam optimizer [33] is used with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The initial learning rate and the batch size for the three phases of training are set to $(10^{-4}, 10)$, $(10^{-4}, 3)$, and $(10^{-5}, 3)$, respectively. During the training of each phase, the learning rate was gradually modified to obtain the best results. The network is trained and evaluated on a single 'NVIDIA GeForce GTX 1660 Ti with Max-Q Design' GPU.

At the test time, the TLSC mechanism [7] is applied to the whole network to address the issue of statistics distribution shift between the training and testing phases. The local window size of TLSC was empirically set to 96x96. The proposed method takes about 1.1 seconds to recover a 1280x720 size HR image from the 320x180 size LR blurry input.

4. Results

The proposed network is evaluated on the REDS validation dataset [32] both quantitatively and qualitatively. The PSNR [17] / SSIM [18] scores are calculated using the computation codes released by [16] for a fair comparison.

Table 1 summarizes the quantitative comparison results between the existing methods on the Val300 [16] dataset derived from the REDS validation set by sampling each tenth image from the validation set.

Table 1: Quantitative comparison results on the Val300 dataset [16]. * denotes the results cited from [16], # denotes the self-ensemble strategy [12].

Methods	Bicubic*	GFN*	RCAN*	[MPRNet + RCAN]*	[SRN + RCAN]*	Shoyan et al	CNLRN	*Proposed method*
PSNR	23.848	26.635	27.338	27.550	27.610	27.164	27.770 / 27.922#	*27.620 / 27.740#*
SSIM	0.6481	0.7447	0.7661	0.7740	0.7745	0.7610	0.7784 / 0.7813#	*0.7735 / 0.7760#*

As shown in Table 1, the PSNR score of the SR method RCAN [8] is at least 0.29dB less than the proposed method since the blur degradation hampers the performance of SR techniques. The existing SR methods [8-10] employ only a single-scale pipeline, which fails to extract the contextual information to remove the blur degradation. The joint method GFN [14] does not perform well since the SR branch operates on the blurry LR image. The joint method proposed by Shoyan et al. [13] generates promising results (see Fig. 1, Fig. 4), but their reconstruction module is not large enough to fully reuse the contextually and spatially enriched features extracted from MPRNet [3]. The cascaded approaches of deblurring and SR methods (MPRNet [3] + RCAN [8], SRN [2] + RCAN [8]) still obtain less PSNR/SSIM scores than the proposed method since they employ the Squeeze-and-Excitation mechanism [28]. In contrast, the proposed method exploits the benefits of MS-CAM [22] blocks to better process the relatively small structures present in the LR blurry image. The joint method CNLRN [16] obtains higher

PSNR/SSIM results than the proposed method. However, the CNLRN [16] fails to reconstruct enough sharpness for relatively small structures in the LR blurry image and generates smoothed texture details, as shown in Fig. 4. The self-ensemble strategy [12] is also employed to increase the performance of the proposed method by running the model on 8 augmented LR images followed by inverse transformation and averaging, resulting in 27.740/0.7760 dB PSNR/SSIM scores, respectively.

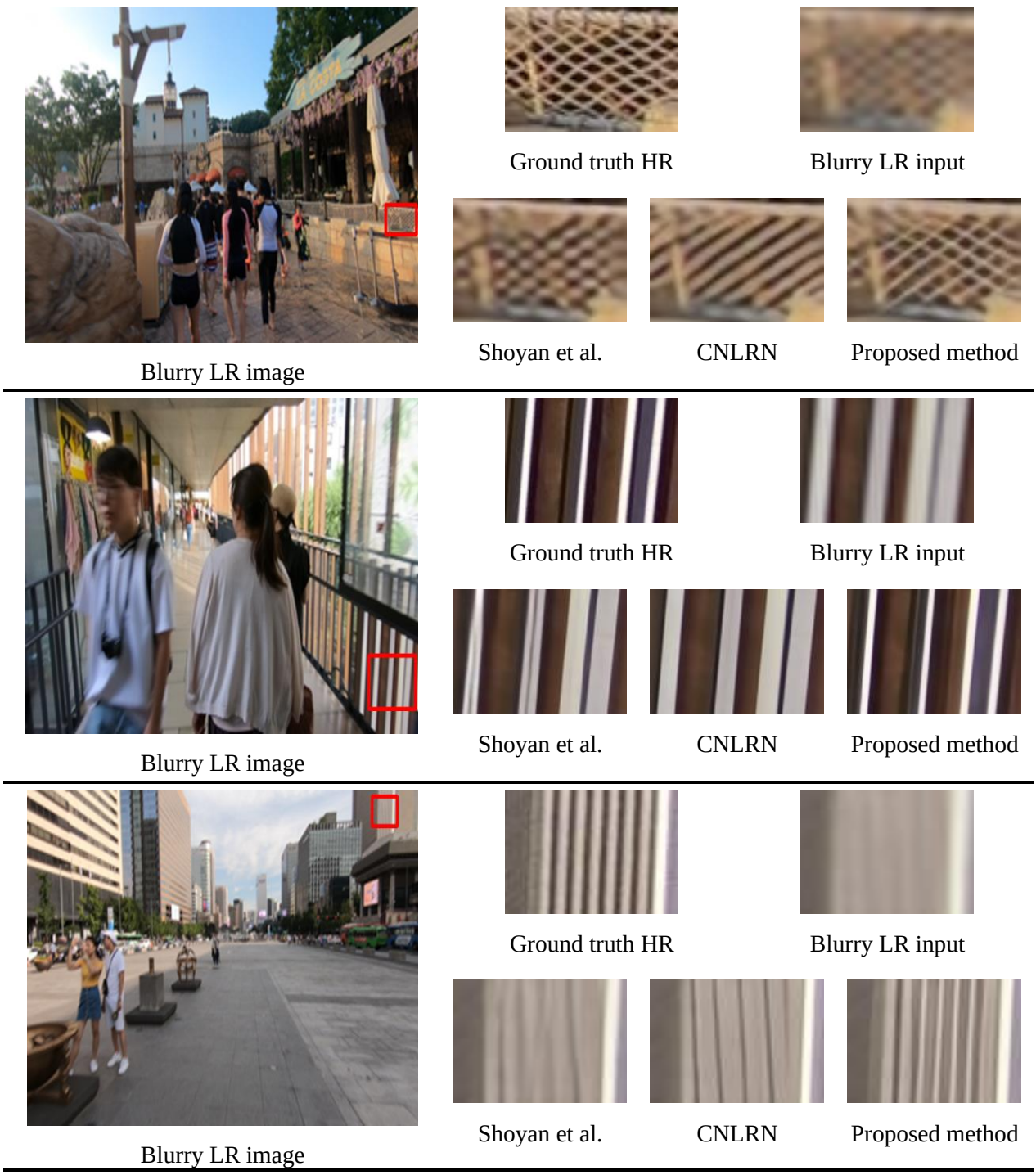

Fig. 4. Qualitative comparison results on the validation dataset. Please, zoom in for the best view.

Table 2: Quantitative comparison results on the REDS [32] validation dataset.

Methods	CNLRN	PDAN	EDPN	*Proposed method*
PSNR	27.828	27.890	28.010	*27.660*
SSIM	0.7794	0.7798	0.8203	*0.7745*
#Params	83.06M	61.00M	13.34M	*34.12M*

Table 2 summarizes the quantitative comparison results for the proposed method and three top-ranked methods in NTIRE 2021 [15] challenge (CNLRN [16], PDAN [20], EDPN [21]). The methods were evaluated on the REDS [32] validation set. The results of PDAN [20] and EDPN [21] were cited from their papers. As shown in Table 2, the proposed method obtains less PSNR/SSIM scores than the other methods, but it has fewer parameters (~34M) and, therefore, less computational complexity compared to CNLRN (~83M) and PDAN (~61M). It makes the model more applicable in scenarios when computational resources are limited. It should be mentioned that the results of PDAN are not reproducible since the source code, pre-trained model, and network implementation details are not publicly available. It also refers to EDPN since no pre-trained model is provided, while the publicly available source code is one of the authors' ablation studies.

The qualitative comparison results on the REDS validation set are demonstrated in Fig. 4. As it is shown, the CNLRN [16] and the method proposed by Shoyan et al. [13] fail to recover the small structures present in the LR blurry image. Instead, the proposed method successfully removes the complex motion blur from the relatively small structures and reconstructs fine texture details. As shown, unlike the other methods, the HR image generated by this method is closer to the ground-truth HR image.

5. Ablation Study

The ablation experiments were performed on the Val300 dataset. Table 3 summarizes the effect of different modifications in terms of PSNR [17] / SSIM [18] metrics on the RGB channel for the proposed method. The baseline model employs only the deblurring and transformation modules followed by upsampling layer [29]. As a Model-1, the SR module is added to the baseline model with 12 RGs, each containing 10 RMS-CAB blocks followed by a convolution layer.

Table 3: Ablation study on the Val300 dataset [16].

Modifications	Models			
	Baseline	**Model-1**	**Model-2**	**Model-3**
Baseline	✓	✓	✓	✓
120 RMS-CAB		✓		
170 RMS-CAB			✓	✓
TLSC				✓
PSNR	27.098	27.453	27.501	**27.620**
SSIM	0.7530	0.7696	0.7713	**0.7735**

This change brings ~0.36dB improvement in terms of PSNR. When the number of RGs is increased to 17, the PSNR score is improved from 27.453dB to 27.501dB (Model-2). Finally, the TLSC mechanism [7] is employed on the whole network (Model-3) at the test time, achieving a 27.620dB score in terms of PSNR.

The models were trained with the proposed three-phase training strategy by gradually modifying the learning rate to obtain the best results.

To show the effect of the proposed training strategy, the Model-1 is trained jointly from scratch. It achieves 27.371dB in terms of PSNR, about 0.08dB less than the same model trained with the proposed three-phase training strategy.

6. Conclusion

This paper proposes an end-to-end single branch architecture to reconstruct a sharp HR image from the given motion blurry LR image. The proposed method reuses the contextually and spatially enriched features extracted from the MPRNet [3] in the super-resolution subnetwork. The super-resolution subnetwork is designed based on MS-CAM [22] blocks. A three-phase training strategy is exploited where a pre-trained frozen deblurring module is used as a feature extractor to train the SR module. The deblurring module is unfrozen in the last phase of the training. Experiments show that, unlike the other methods, the proposed method successfully removes the complex motion blur from the relatively small structures and reconstructs fine texture details. The source code and the pre-trained model are available at https://github.com/misakshoyan/joint-motion-deblur-and-sr.

References

[1] A. Vaswani et al. “Attention Is All You Need”, *arXiv preprint arXiv:1706.03762*, 2017.

[2] X. Tao et al., “Scale-recurrent network for deep image deblurring”, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 8174–8182, 2018.

[3] S. W. Zamir et al., “Multi-Stage Progressive Image Restoration”, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, pp. 14816-14826, 2021.

[4] S. W. Zamir et al. “Restormer: Efficient Transformer for High-Resolution Image Restoration”, *arXiv preprint arXiv:2111.09881*, 2021.

[5] X. Mao et al. “Deep Residual Fourier Transformation for Single Image Deblurring”, *arXiv preprint arXiv:2111.11745*, 2021.

[6] S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 257-265, 2017.

[7] X. Chu et al. “Revisiting Global Statistics Aggregation for Improving Image Restoration”, *arXiv preprint arXiv:2112.04491*, 2021.

[8] Y. Zhang et al., “Image super-resolution using very deep residual channel attention networks”, *Proceedings of European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 294-310, 2018.

[9] T. Dai et al., “Second-Order Attention Network for Single Image Super-Resolution”, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, pp. 11057-11066, 2019.

[10] B. Niu et al., "Single image super-resolution via a holistic attention network", *Proceedings of European Conference on Computer Vision (ECCV)*, Glasgow, UK, pp. 191–207, 2020.
[11] J. Liang et al., "SwinIR: Image Restoration Using Swin Transformer", *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, Canada, pp. 1833-1844, 2021.
[12] B. Lim et al., "Enhanced deep residual networks for single image super-resolution", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, USA, pp. 1132-1140, 2017.
[13] M. Shoyan et al., "Single Image Joint Motion Deblurring and Super-Resolution", *Proceedings of 13th International Conference on Computer Science and Information Technologies (CSIT)*, Yerevan, Armenia, pp. 182-186, 2021.
[14] X. Zhang et al. "Gated fusion network for degraded image super resolution", *International Journal of Computer Vision*, vol. 128, no. 6, pp. 1699-1721, 2020.
[15] S. Nah et al., "NTIRE 2021 Challenge on Image Deblurring", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, USA, pp. 149-165, 2021.
[16] H. Bai et al, "Learning A Cascaded Non-Local Residual Network for Super-resolving Blurry Images", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, USA, pp. 223-232, 2021.
[17] Wikipedia, (2014) The peak signal-to-noise ratio, [Online]. Available: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio
[18] Z. Wang et al., "Image quality assessment: from error visibility to structural similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.
[19] R. Zhang et al., "The unreasonable effectiveness of deep features as a perceptual metric", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 586-595, 2018.
[20] S. Xi, J. Wei and W. Zhang, "Pixel-Guided Dual-Branch Attention Network for Joint Image Deblurring and Super-Resolution", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, USA, pp. 532-540, 2021.
[21] R. Xu et al., "EDPN: Enhanced Deep Pyramid Network for Blurry Image Restoration", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, USA, pp. 414-423, 2021.
[22] Y. Dai et al., "Attentional Feature Fusion", *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, USA, pp. 3559-3568, 2021.
[23] T. H. Kim, B. Ahn and K. M. Lee, "Dynamic Scene Deblurring", *Proceedings of IEEE International Conference on Computer Vision*, Sydney, Australia, pp. 3160-3167, 2013.
[24] J. Pan et al., "Blind image deblurring using dark channel prior", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 1628-1636, 2016.
[25] R. Gonzalez and R. Woods, Digital Image Processing, 4th ed., Pearson, New York, 2018.
[26] M. Bevilacqua et al. "Low-complexity single-image super-resolution based on nonnegative neighbor embedding", *Proceedings of British Machine Vision Conference (BMVC)*, Guildford, UK, paper 135, pp. 1-10, 2012.
[27] R. Timofte, V. D. Smet, and L. V. Gool. "A+: Adjusted anchored neighborhood regression for fast super-resolution", *Proceedings of Asian Conference on Computer Vision (ACCV)*. Singapore, Singapore, pp. 111-126, 2014.

[28] J. Hu, L. Shen and G. Sun, “Squeeze-and-Excitation Networks”, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132-7141, 2018.
[29] W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 1874-1883, 2016.
[30] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation”, *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, pp. 234-241, 2015.
[31] P. Charbonnier et al, “Two deterministic half-quadratic regularization algorithms for computed imaging”, *Proceedings of 1st International Conference on Image Processing*, Austin, USA, pp. 168-172, 1994.
[32] S. Nah et al., “NTIRE 2019 challenges on video deblurring and super-resolution: Dataset and study”, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, USA, pp. 1996–2005, 2019.
[33] D. Kingma and J. Ba, “Adam: A method for stochastic optimization”, *arXiv preprint arXiv:1412.6980*, 2014.

Submitted 18.05.2021, accepted 14.09.2021.

Պատկերից շարժման հետևանքով առաջացած շաղվածության հեռացում և պատկերի կետայնության բարձրացում՝ օգտագործելով շերտային ուշադրության բազմամասշտաբ մոդուլները

Միսակ Տ. Սհոյան

Հայաստանի Ազգային Պոլիտեխնիկական Համալսարան

e-mail: misakshoyan@gmail.com

Ամփոփում

Վերջին տասնամյակում խորը փաթույթային նեյրոնային ցանցերը զգալիորեն զարգացրել են պատկերի կետայնության բարձրացման մեթոդները՝ վերակառուցելով պատկերի կառուցվածքային և տարածական իրատեսական մանրամասներ։ Պատկերի կետայնության բարձրացման դասական խնդիրներում ենթադրվում է, որ ցածր կետայնության պատկերն ունի կետայնության նվազման որոշակի դեգրադացիա։ Սակայն, իրական սցենարներում, պատկերի բարդ դեգրադացիաներն անխուսափելի են և շարժման հետևանքով առաջացած շաղվածությունը պատկերի դեգրադացիայի տարածված տեսակ է, որն առաջանում է պատկերի նկարահանման գործընթացում՝ տեսախցիկի կամ տեսարանի շարժման

հետևանքով: Այս աշխատանքում առաջարկվում է լրիվ փաթույթային նեյրոնային ցանց՝ տրված ցածր կետայնությամբ և շարժման հետևանքով առաջացած շաղվածություն պարունակող պատկերներից բարձր կետայնության հստակ պատկերներ վերակառուցելու համար: Շաղվածության հեռացման ենթացանցը հիմնված է բազմափուլային պրոգրեսիվ ճարտարապետության վրա, մինչդեռ կետայնության բարձրացման ենթացանցը նախագծելու համար օգտագործվում են շերտային ուշադրության բազմամասշտաբ մոդուլները: Կիրառվում է ուսուցանման պարզ և արդյունավետ եղանակ, որտեղ նախապես ուսուցանված և սառեցված շաղվածության հեռացման մոդուլն օգտագործվում է կետայնության բարձրացման մոդուլն ուսուցանելու համար: Ուսուցման վերջին փուլում շաղվածության հեռացման մոդուլը նույնպես ուսուցանվում է: Կատարված փորձերը ցույց են տալիս, որ ի տարբերություն մյուս մեթոդների, առաջարկվող մեթոդը վերակառուցում է համեմատաբար փոքր դետալները և կառուցվածքային մանրամասները՝ միաժամանակ հաջողությամբ հեռացնելով շարժման հետևանքով առաջացած բարդ շաղվածությունը: Իրականացման կոդը և նախապես ուսուցանված մոդելը հասանելի են https://github.com/misakshoyan/joint-motion-deblur-and-sr կայքում:

Բանալի բառեր՝ Շարժման հետևանքով առաջացած շաղվածության հեռացում, կետայնության բարձրացում, շերտային ուշադրություն:

Удаление размытости вызванной движением и увеличение разрешения одного изображения с использованием многомасштабных модулей внимания канала

Мисак Т. Сгоян

Национальный Политехнический Университет Армении

e-mail: misakshoyan@gmail.com

Аннотация

За последнее десятилетие глубокие сверточные нейронные сети значительно продвинули методы увеличения разрешения одного изображения, реконструируя реалистичные текстурные и пространственные детали. В классических задачах увеличении разрешения изображения предполагается, что изображение с низким разрешением имеет определенную деградацию понижения разрешении. Однако, в реальных сценариях, сложные деградации изображения неизбежны, и размытие изображения вызванное движением является распространенным типом деградации, которое происходит в процессе съемки изображения из-за движения камеры или сцены. В этой работе предлагается полностью сверточная нейронная сеть для реконструкции четких изображений с высоким разрешением из заданных размытых изображений

вызванных движением с низким разрешением. Подсеть удаления размытости основана на многоступенчатой прогрессивной архитектуре, в то время как подсеть увеличения разрешения разработана с использованием многомасштабных модулей внимания каналов. Используется простая и эффективная стратегия обучения, в которой предварительно обученный и замороженный модуль удаления размытости используется для обучения модуля увеличения разрешения. Модуль удаления размытости размораживается на последнем этапе обучения. Эксперименты показывают, что, в отличие от других методов, предлагаемый метод реконструирует относительно небольшие структуры и текстурные детали, успешно удаляя сложное размытие вызванное движением. Код реализации и предварительно обученная модель общедоступны по адресу https://github.com/misakshoyan/joint-motion-deblur-and-sr.

Ключевые слова: Удаление размытости вызванной движением, увеличение разрешения, внимание канала.