Neural Computing and Applications (2022) 34:10707–10717
https://doi.org/10.1007/s00521-022-07006-w

ORIGINAL ARTICLE

Single image super-resolution via deep progressive multi-scale fusion networks

Yue Que[1] · Hyo Jong Lee[2]

Received: 25 February 2021 / Accepted: 30 January 2022 / Published online: 13 April 2022

Abstract
Deep convolutional neural network-based single image super-resolution (SR) models typically process either upsampled full-resolution or original low-resolution features, which suffer from context lack and spatially imprecision, respectively. To solve this, we propose a novel progressive SR network to preserve spatial precision through the original resolution and to receive rich contextual information from low-to-high resolution representations. Our proposed progressive, selective scale fusion network includes four key points: (a) parallel multi-scale convolution branches to extract multi-scale features, (b) information exchange across the multi-scale branches, (c) attention mechanism-based multi-scale feature fusion, and (d) gradual aggregation of multi-scale streams from low-to-high resolutions. The proposed method learns hierarchical features that aggregate contextual information from different resolution streams while maintaining high-resolution spatial details. Both quantitative and qualitative experiments on benchmark and real-world datasets show that our method offers a favorable performance against state-of-the-art methods for SR tasks with different scaling factors.

Keywords Image super-resolution · Multi-scale feature · Deep learning · Convolutional neural networks

1 Introduction

Recently, single image-based super-resolution (SR) has become an important task that aims at learning a nonlinear mapping to reconstruct a clear, high-resolution image from a downgraded low-resolution image. Image SR is a fundamental task being widely used in various computer vision applications, e.g., security and surveillance imaging, medical imaging, image recognition, and remote sensing. However, SR is very challenging due to the ill-posed inverse procedure caused by irreversible image degradation. To address this problem, numerous single image SR algorithms have been developed, which can be divided into three categories: interpolation-based methods, reconstruction-based methods, and learning-based methods [1, 2].

Interpolation-based methods are the simplest solution that calculates values of the interpolated pixels through neighboring pixels, so the computational complexity is relatively low [3, 4]. However, high-resolution images generated by these methods often suffer from edge halos and artifacts [5]. Model-based reconstruction focuses on designing degradation models to reconstruct high-resolution images [6]. This type of method mainly involves edge prior [7, 8], gradient prior [9], non-local means priors [10, 11], and sparsity priors [12–14]. To better represent image features, some methods combine multi-priors to improve the SR performance. However, these methods tend to blur images, thus leading to details loss and over-smoothness [15]. In contrast, the learning-based methods obtain a high-resolution image by learning the mapping between every pair of high-resolution and low-resolution images. This kind of method exhibit better performance than traditional ones by effectively eliminating the edge halo and artifacts. Although numerous kinds of approaches have been introduced to address the ill-posed inverse problem in single image SR, unacceptable artifacts in high-resolution images remain a common problem.

✉ Hyo Jong Lee
hlee@jbnu.ac.kr

[1] School of Information Engineering, East China Jiaotong University, Nanchang 330013, China

[2] Division of Computer Science and Engineering, CAIIT, Jeonbuk National University, Jeonju 54896, South Korea

In recent years, deep learning brings a boost to single image SR tasks. Dong et al. [17] firstly used a 3-layer convolutional neural network (SRCNN) to reconstruct a high-resolution image by learning the nonlinear mapping between the high-resolution and low-resolution images. However, the shallow network brought limited improvements. Kim et al. [18] presented a 20-layer deep convolutional neural network (VDSR) with global residual connections to improve the performance. Shi et al. [19] developed a contextualized multi-task convolutional network to super-resolved low-resolution images, which enhances the structural details by taking the global boundary context and the residual context as complementary information. However, the main drawback of these deep models is that the number of parameters increases with the network depth. Some interesting approaches have been proposed to tackle this limitation. Kim et al. [20] presented an image SR method based on a deeply recursive convolutional network (DRCN), which involved weight sharing across recursive layers. Moreover, Lai et al. [21] introduced a Laplacian pyramid image SR network (LapSRN) based on a CNN cascade which gradually reconstructed the sub-band residuals of high-resolution images. Since the input of the network was not upsampled, the computational efficiency was improved. Similar solutions have been adopted by EDSR [22], RDB [23], RCAN [24], and IDN [25], and they used upsampling operations at the end of the whole network to generate the final restored high-resolution images. EDSR [22] improves the residual SR network by removing the batch normalization layers and constructing a deeper and wider network. RCAN [24] proposed a residual in residual structure and adopted a channel attention mechanism to build a deeper network, which improves the network representation ability and reduces the training difficulty. Although these deep learning-based methods have achieved outstanding performance in objective evaluation, most of them tend to build deeper and more complex network structures, which makes training more difficult. In addition, most models simply stack the building modules in a chained fashion without capturing features across modules.

Overall, figuring out how to perform upsampling (i.e., generating high-resolution output from low-resolution input) is a key problem in image SR. Most of the existing CNN-based SR methods typically follow two architecture designs according to the employed upsampling operations and their locations in the model. The first is pre-upsampling SR [17, 18, 20], in which a low-resolution image is first interpolated to obtain a "coarse" high-resolution image and then transformed into a "refined" image through an end-to-end CNN mapping. The intuition is that it may be easier to conduct the first stage using traditional methods than to learn a direct mapping from a low-dimensional space to a high-dimensional space; thus, the CNN only needs to learn how to refine the coarse image, which is simpler. Moreover, since we avoid transposed convolutions here, checkerboard artifacts may be circumvented. The downsides, however, are the predefined upsampling methods may amplify noise and cause blurring, and these networks extract less contextual information due to their limited receptive field. The second typical architecture design is post-upsampling SR [22–25]. In this architecture, the low-resolution images are directly passed to CNNs and upsampling is always performed last using a learnable layer. Since CNN-based feature extraction is performed in the lower dimensional space (before upsampling), the computational complexity is more reduced than pre-upsampling methods. Furthermore, by using a learnable upsampling layer, the model is flexible and can be trained end-to-end. Although these approaches make full use of deep learning technology to increase resolution automatically, the fine spatial details can be lost as the layer goes deeper, making it harder to recover them in the last upsampling stage. Compared to multi-frame SR reconstruction, which deals with multiple low-resolution images or video frames with similar backgrounds, single image SR reconstruction cannot make use of information between different frames. Hence, making full use of the hierarchical information in a single low-resolution image is key to further improving performance.

To fully exploit the multi-scale features of the original images, this work proposes a novel progressive and selective scale fusion network for single image super-resolution. The proposed network not only leverages original-scale representations but also gradually expands them to representations of the target scale. This study aims to leverage the representation capabilities from all low-to-high-resolution parallel convolutions. In contrast with most existing methods that simply use post-upscaling or pre-upscaling, our network processes multi-resolution features by applying parallel convolution branches that provide us with a more precise and context-rich feature representation. In addition, different from existing methods that process each scale separately, the information across parallel streams can be efficiently exchanged with repeated multi-resolutions by designing the selective scale fusion module. At the same time, a new selective scale fusion mechanism is used to exchange fine-to-coarse and coarse-to-fine information on each stream. Rather than simply concatenating or weighting the features from multi-scale streams, the proposed fusion method uses self-attention to dynamically select useful scale sets from each branch representation. Moreover, our fusion units combine features from the different receptive fields while retaining their unique complementary features.

In summary, the main contributions are threefold:

1. We connect low-to-high scale convolution branches in parallel rather than implementing post-upscaling or pre-upscaling. Our method not only utilizes the features at the original resolution but also fully exploits multi-scale features from low resolution to high resolution so that the learned representation is more abundant and accurate.
2. We propose a novel selective scale fusion method to repeat aggregate multi-scale features that adaptively combines different receptive fields and accurately retains the input feature information at each spatial scale. As a result, the representations at different resolutions complement each other.
3. Extensive experiments are conducted on both simulation and real image benchmark datasets. The experimental results demonstrate that the proposed method leads to better performance than a series of state-of-the-art approaches in terms of both visual and objective quality.

The rest of the paper is organized as follows. Section 2 discusses related works for single-image SR. Section 3 introduces the proposed framework for learning disentangled representations. In Sect. 4, we present the experimental procedures and results. Finally, we provide a brief conclusion in Sect. 5.

2 Related work

2.1 Deep feature extraction for image super-resolution

Various feature extraction modules based on CNN for different vision tasks have been introduced to offer improved performance, including residual block [26], dense block [27], and inception block [28]. Haris et al. [29] used the inception module to extract multiple features from low-resolution images for SR. Tong et al. [30] used dense skip connections to effectively aggregate low-level features and high-level features to improve the image SR performance and to alleviate the vanishing gradient problem of very deep networks. Zhang et al. [23] introduced a residual dense block to exploit rich local features by densely connected convolutional layers. Although the residual dense block fully used residual learning and dense connections, both adopted convolution kernels in a fixed size, which limited the extraction of image multi-scale features. Furthermore, the computational complexity of the dense blocks increases heavily as the number of dense connections increases. To address such problems, Li et al. [31] presented a multi-scale residual block (MSRB) to fully extract multi-scale image features with different sizes of convolution kernels. In addition, these local multi-scale features with global features solved the problem of features disappearing. Furthermore, a 1×1 convolution layer at the end of each MSRB was utilized to fuse global features and to reduce the computational complexity. However, two MSRB bypasses of the same depth make this method difficult to fully utilize shallow and deep local image features. Qin et al. [32] proposed a multi-scale feature fusion residual block with multiple intertwined paths, which provides a more accurate representation of the local features. However, the extracted features were not spatial multi-resolution features because the sizes of both the feature map and the convolutional kernel were the same in different paths. Based on this, a novel selective scale fusion network for single image SR is proposed in our work.

2.2 Multi-scale feature fusion

Multi-scale features have attracted extensive attention in both traditional feature extraction [33] and deep learning [34] to enhance the performance of computer vision tasks. The ability of feature extractors to represent context at different scales is a fundamental requirement to generate multi-scale features for visual tasks. CNNs can learn multi-scale features adaptively, from coarse to fine, to a certain extent through a set of convolution operators [35]. How to design a more efficient structure to make full use of multi-scale features is the key to further improving the performance of CNN-based single image SR. A simple approach is to input multi-scale feature maps into multiple networks and fuse the output response maps [36]. The Hourglass [37] net produces features across all scales and pools them down to a very low resolution, then upsamples and combines features across multiple scales. Similarly, U-Net [38] and SegNet [39] gradually combine high-resolution features from the contracting path with those from the upsampled expanding path through skip connections. Chen et al. [40] introduced a cascaded pyramid network consisting of global-net and refine-net. The former aggregates low-to-high level features gradually, and the latter aggregates the low-to-high level features processed by the convolution operations. PSPNet [41] and DeepLabV2/3 [42] combine the pyramid features generated by the pyramid pooling and atrous spatial pyramid pooling. The proposed multi-scale feature fusion module outputs different-resolution representations rather than single resolution features and is repeated several times inspired by Sun et al. [43].

3 Proposed method

3.1 Overall pipeline

This work aims to conduct end-to-end learning for mapping from an original low-resolution image to a high-resolution image by constructing a selective scale fusion network. The overall network structure is shown in Fig. 1, and it contains parallel multi-scale convolutions, repeated multi-scale fusion modules, and an output head demonstrated in Fig. 3. We use the $\times$ 8 scale as an example; the whole network structure can be divided into four stages. The network progressively adds low-to-high resolution branches, forms new stages, and combines multi-scale branches in parallel. Therefore, the scales for the parallel branches of a later stage are composed of the scales from the previous stage and an additional higher one. The first stage consists of input low-resolution convolutions, and the second, third, and fourth stages are formed by repeating modularized multi-resolution modules. Specifically, the first stage includes two residual units, each of which consists of two ResNet blocks. In addition, each ResNet block consists of three consecutive operations, a 3×3 convolution, rectified linear units (ReLU) [44], and another 3×3 convolution. Thus, the four stages contain four parallel convolution streams, in which the resolution of the feature maps progressively increases, and accordingly, the number of channels is halved. The fusion unit crosses parallel subnets so that each subnet repeatedly receives information from the other parallel subnets. Therefore, the scales for parallel branches of a later stage consist of scales from the previous stage, and an extra higher one. For an s stage network, the resolution of the output is 2^{s-1} times the resolution of the first branch. The multiple parallel convolutions in our network generate a spatially precise output by maintaining original resolution representations while receiving rich contextual information from different resolutions. Moreover, the proposed network allows information exchange across parallel streams to consolidate the high-resolution features with the help of low-resolution features, and vice versa.

3.2 Selective scale feature fusion

The fusion unit exchanges information across multi-scale feature representations. The mechanism of the adaptive regulation of the sensory domain can be generated by multi-scale features (at the same layer). Then, the features are aggregated, selected, and integrated into the CNNs. A simple concatenation or summation is commonly used for feature fusion. However, these options provide limited representation ability for the deep network [45, 56]. To this end, a self-attention-based nonlinear process is introduced in our work to fuse multiple-resolution features, which is referred to as a selective scale feature fusion. A fusion example with three scales is shown in Fig. 2; the selective scale feature fusion module dynamically adjusts receptive fields including aggregation and selection steps. The aggregation step obtains global feature representations by combining information from multi-scale branches, while the selection step utilizes the obtained global feature representations to recalibrate feature maps and their aggregations for different streams. Furthermore, a detailed example is provided for two steps of the case with three branches, but it can be easily extended to more branches. For the aggregation step, the selective scale fusion module acquires the inputs of three parallel convolution branches carrying the information of different scales, F_1, F_2, and F_3. First, the module sums these multi-scale features F in the dimension of $H \times W \times C$ and then applies the global average pooling on F to obtain channel-wise statistics. Next, a convolutional layer is used to produce a compact feature representation z by reducing the number of feature channels. Finally, three parallel convolutional layers (one

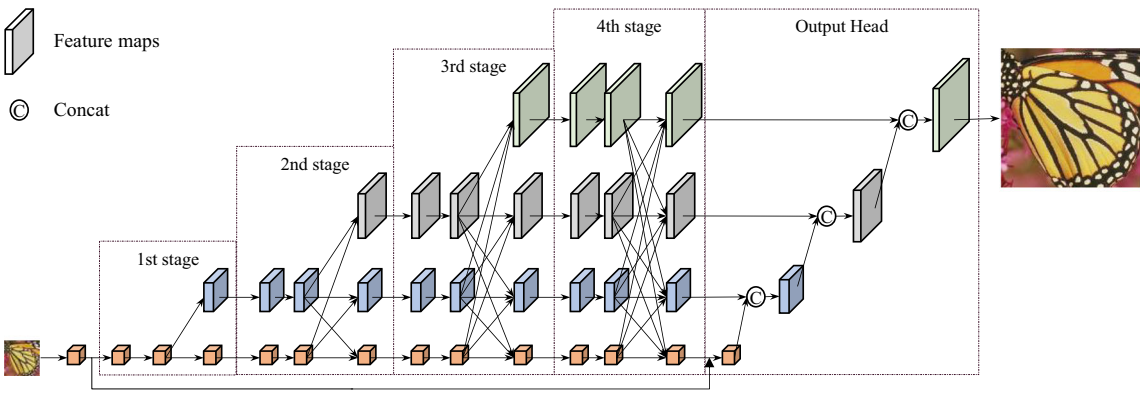
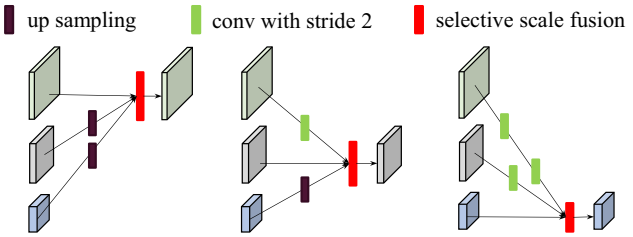
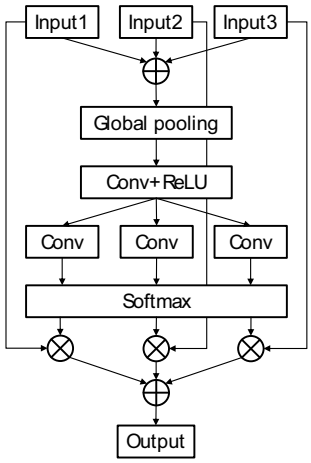


Fig. 1 The architecture of our proposed selective scale fusion network for image super-resolution. We use the $\times$ 8 scale as an example. There are four stages and an output head. The first stage consists of low-resolution convolutions. The second (third, fourth) stage repeats two-resolution (three-resolution, four-resolution) blocks. The detail is explained in Sect. 3.1



(a) Fusing the information for different scale.



(b) Selective scale fusion mechanism.

Fig. 2 Schematic for selective scale feature fusion with 3 inputs. It operates on features from multiple convolutional branches, and performs fusion based on self-attention

for each scale branch) process the feature vector z and generate three feature representations in the dimension of $1 \times 1 \times C$. The selection step uses SoftMax to generate attention activation w_1, w_2, and w_3, which are then used to adaptively recalibrate the multi-scale feature maps F1, F2, and F3, respectively. The final procedure of the multi-scale feature recalibration and fusion is defined as $F = w_1 \bullet F_1 + w_2 \cdot F_2 + w_3 \cdot F_3$.

3.3 Output head

Global residual learning is applied to generate the original resolution feature maps before executing the output module. While the local multi-scale features are further adaptively involved in global feature learning. The deep features are obtained after global residual learning, and the proposed model takes full advantage of all the different scale features before global residual learning. We explore three types of output heads, which are termed V1, V2, and V3, respectively. As shown in Fig. 3a, the output of V1 is the representation only from the high-resolution branch. V2 scales the low-resolution representation with bilinear upsampling, concatenates the four representations in series, and then mixes the four representations by a 1×1 convolution. This is illustrated in Fig. 3b. V3 is the proposed representation method, and it upsamples the low-resolution representation with pixel-shuffle outputting 2^{8-k} channels, which will be concatenated to a $k+1$th low-resolution representation. This is illustrated in Fig. 3c, and $k = \{1, 2, 3\}$.

3.4 Implementation details

The main body of the proposed network includes four stages with four parallel convolution branches. The scaling resolutions are 1, 2, 4, and 8. The first stage includes four residual units, and the second, third, and fourth stages contain 2, 2, 1 modularized blocks, respectively. The numbers of output channels in the convolution branches for the four resolutions are C, C/2, C/4, and C/8, from a low to a high resolution. It is worth noting that the network includes 2, 3, and 4 stages depending on the scaling factor 2, 4, and 8, respectively. Except for the convolutional kernel size in the selective scale fusion module being set to 1×1, the kernel size for all other convolutional layers is set to 3×3. For the convolutional layer with a kernel size of 3×3, we pad zeros to each side of the input to keep the feature size fixed. Since the network outputs high-resolution images in color, the output channel number of the final convolutional layer is 3.

The mean square error (MSE) function and the L_2 loss function are the most widely used loss functions in deep learning-based SR image tasks. Although optimization based on MSE or L_2 can achieve relatively higher objective evaluation indexes, such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [46], they tend to produce over-smooth textures, resulting in blurred visual effects [59]. The multi-scale structural similarity index (MS-SSIM) is more insensitive to changes in brightness or shifts in colors. However, it can better preserve high-frequency information compared to the other loss functions. The L_1 norm can better preserve the color and luminance characteristics, but it does not generate quite the same contrast as MS-SSIM [47]. Hence, the comprehensive loss function is obtained by combining them as follows:

$$\text{Loss} = (1 - \text{MS}) + \alpha \cdot L_1 \tag{1}$$

where MS and L_1 are the content loss that evaluates the MS-SSIM and the 1-norm distance between the recovered image and the reference image, and α is the coefficient to balance different loss terms.

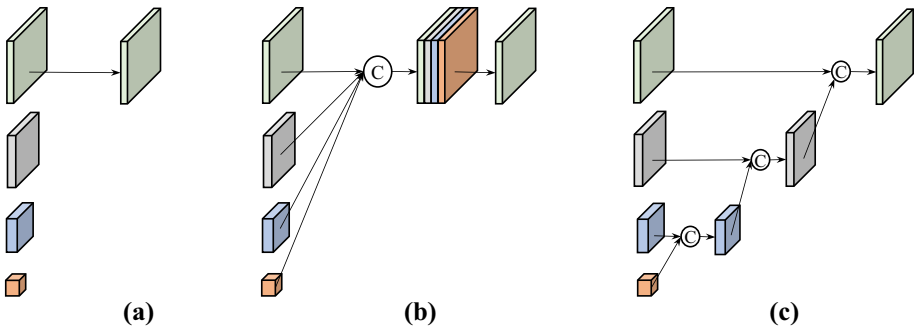


Fig. 3 Various kinds of output heads. **a** V1: only outputting the representation from the high-resolution convolution stream. **b** V2: concatenating the (upsampled) representations that are from all the resolutions. **c** V3: progressively upsampling and concatenating the output representations that are from all the resolutions

4 Experimental results

This section evaluates the proposed network architecture on the benchmark image SR datasets. It first describes the experimental datasets and settings and then presents a comparison with state-of-the-art methods. Finally, some ablation studies are also presented.

4.1 Experimental datasets and settings

Recently, Timofte et al. [48] introduced a large dataset (DIV2K) for single image SR studies. DIV2K contains 1000 high-quality images (2K resolution), in which 800 are used for training, 100 are used for verification and the rest 100 for testing. Flickr2K [22] is another large high-quality dataset collected on Flickr, and it contains 2,650 2K resolution images covering different scenes including people, animals, landscapes, and more. Both datasets are widely used in image SR studies, and some methods use them together for training to help the model generate more natural high-resolution images. To this end, we follow this training scheme. Furthermore, we train our models in RGB channels and augment the training dataset with random horizontal flips and 90-degree rotations. We evaluate the trained models on four widely used benchmark datasets: Set5 [49], Set14 [16], B100 [50], and Urban100 [51] for single image SR.

To fully prove the effectiveness of our proposed model, we adopt bicubic down-sampling by using the MATLAB function 'imresize' to simulate low-resolution images with a scaling factor of 2, 4, and 8, respectively. We randomly crop 16 low-resolution RGB patches with a size of 40×40 as inputs in each training batch. For optimization, this work uses Adam [52] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All models are implemented using the PyTorch and are trained using an NVIDIA Titan Xp GPU.

4.2 Comparison with state-of-the-art methods

This subsection compares the proposed model with state-of-the-art single image SR algorithms including SRCNN [17], FSRCNN [53], VDSR [18], DRCN [20], LapSRN [21], SRMDNF [55], CARN [54], MSRN [31], SEAN [57] and SRFBN [58]. For a fair comparison, we use the image SR results publicly provided by authors for existing methods. In addition, the image SR results are evaluated with PSNR and SSIM on the Y channel (i.e., luminance) of a transformed YCbCr space. A comparison results including the proposed algorithm and 12 state-of-the-art algorithms for $\times$ 2, $\times$ 4, and $\times$ 8 scaling SR are shown in Table 1. Noticing that there is a big gap between the results of the basic bicubic interpolation method and other CNN-based methods, which means that the bicubic cannot produce any extra details. All these CNN-based methods use a well-designed network to learn the mapping function between low-resolution and high-resolution images directly and all of them achieve great improvement. In general, the proposed method can achieve superior results compared with all the other methods including the extremely competitive SRFBN. Almost all the quantitative results of our method are the best. Specifically, for scale $\times$ 2, the best results are achieved on Set5, Set14, and Urban100, respectively. Our PSNR is only 0.01 dB less than SRFBN on the B100 dataset but we achieve the best SSIM, which indicates our model can better recover visible structures. For scale $\times$ 4 and $\times$ 8, the proposed method outperforms the others on all datasets. When compared with the multi-scale residual network (MSRN), the proposed method outperforms it on all datasets with all scaling factors. This shows better effectiveness in our selective scale feature fusion method over the multi-scale residual block in MSRN. Furthermore, a modified version of the MSRN is used for soft-edge extraction for the image in SEAN. Compared with SEAN, our method also achieves better results on all scaling factors. This further demonstrates the importance of the image multi-scale features on a different scale-SR problem.

To illustrate the visual quality of our proposed method against other state-of-the-art methods, we show several SR results on the B100 and Urban100 with the $\times$ 4 scaling factor among different methods in Fig. 4. Noting that these examples contain rich structured contents, which is a

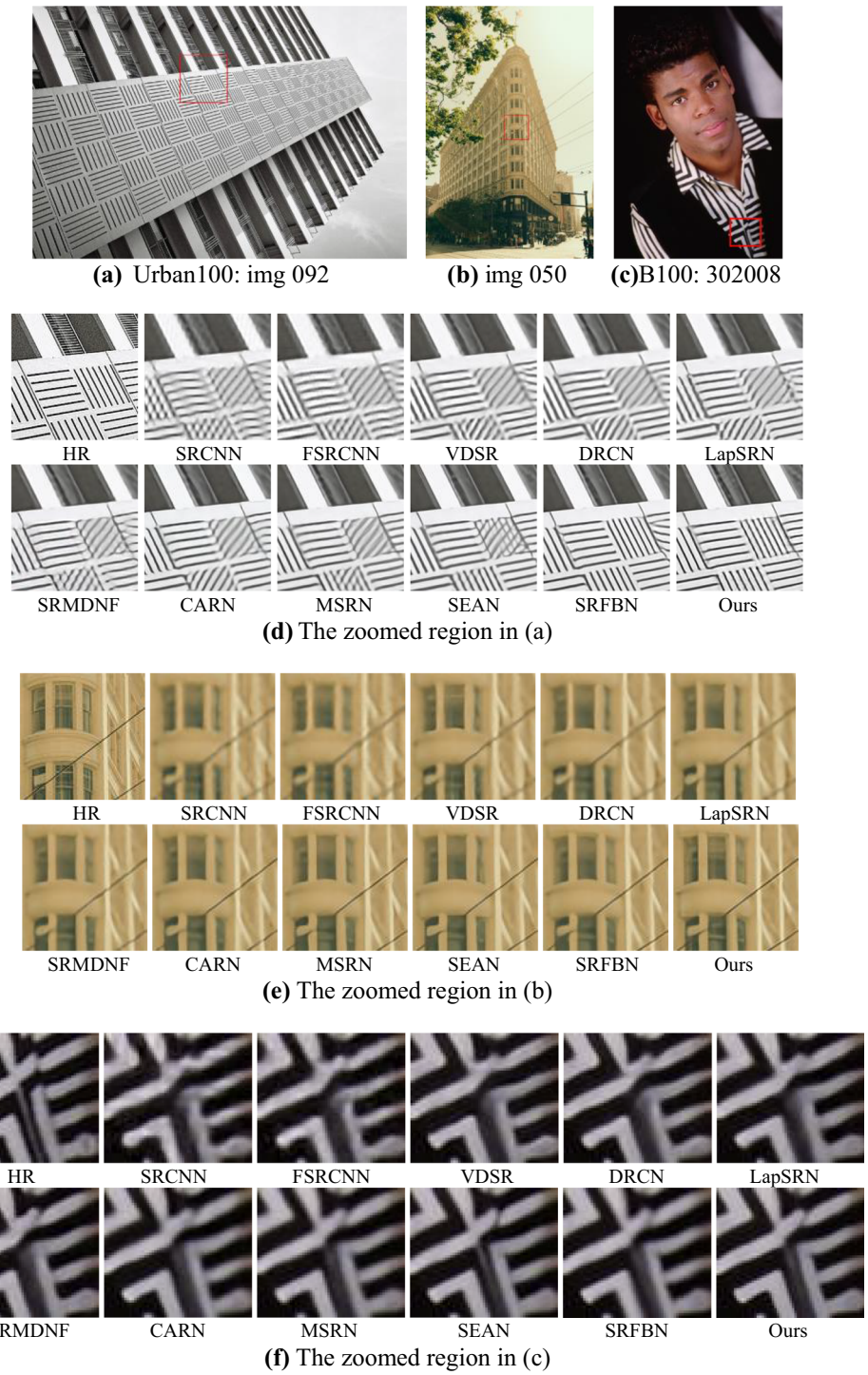
Table 1 The PSNR and SSIM results of different methods on Set5, Set14, B100, and Urban100 with down-sampling factor × 2, × 4, and × 8

Method	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM
Bicubic	× 2	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403
SRCNN	× 2	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946
FSRCNN	× 2	37.05/0.9560	32.66/0.9090	31.53/0.8920	29.88/0.9020
VDSR	× 2	37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140
DRCN	× 2	37.63/0.9584	33.06/0.9108	31.85/0.8947	30.76/0.9147
LapSRN	× 2	37.52/0.9591	33.08/0.9130	31.08/0.8950	30.41/0.9101
SRMDNF	× 2	37.79/0.9601	33.32/0.9159	32.05/0.8985	31.33/0.9204
CARN	× 2	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256
MSRN	× 2	38.08/0.9605	33.74/0.9170	32.23/0.9002	32.22/0.9326
SEAN	× 2	38.08/0.9609	33.75/0.9190	32.27/0.9008	32.50/0.9318
SRFBN	× 2	38.11/0.9609	33.82/0.9196	**32.29**/0.9010	32.62/0.9328
Ours	× 2	**38.19/0.9614**	**33.92/0.9216**	32.28/**0.9026**	**32.65/0.9339**
Bicubic	× 4	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577
SRCNN	× 4	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221
FSRCNN	× 4	30.72/0.8660	27.61/0.7550	26.98/0.7150	24.62/0.7280
VDSR	× 4	31.35/0.8830	28.02/0.7680	27.29/0.0726	25.18/0.7540
DRCN	× 4	31.56/0.8810	28.15/0.7627	27.24/0.7150	25.15/0.7530
LapSRN	× 4	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560
SRMDNF	× 4	31.96/0.8925	28.35/0.7787	27.49/0.7337	25.68/0.7731
CARN	× 4	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837
MSRN	× 4	32.07/0.8903	28.60/0.7751	27.52/0.7273	26.04/0.7896
SEAN	× 4	32.33/0.8970	28.72/0.7855	27.65/0.7388	26.32/0.7942
SRFBN	× 4	32.47/0.8983	28.81/0.7868	27.72/0.7409	26.60/0.8015
Ours	× 4	**32.54/0.9006**	**28.85/0.7931**	**27.73/0.7489**	**26.68/0.8081**
Bicubic	× 8	24.40/0.6580	23.10/0.5660	23.67/0.5480	20.74/0.5160
SRCNN	× 8	25.33/0.6900	23.76/0.5910	24.13/0.5660	21.29/0.5440
FSRCNN	× 8	20.13/0.5520	19.75/0.4820	24.21/0.5680	21.32/0.5380
VDSR	× 8	25.93/0.7240	24.26/0.6140	24.49/0.5830	21.70/0.5710
DRCN	× 8	25.93/0.6743	24.25/0.5510	24.49/0.5168	21.71/0.5289
LapSRN	× 8	26.15/0.7380	24.35/0.6200	24.54/0.5860	21.81/0.5810
MSRN	× 8	26.59/0.7254	24.88/0.5961	24.70/0.5410	22.37/0.5977
Ours	× 8	**26.85/0.7521**	**25.01/0.6321**	**24.93/0.5996**	**22.65/0.6031**

relatively difficult case for image SR tasks, our model can recover sharper lines, clearer contours, and richer details. In general, the proposed method can generate more convincing SR results. Moreover, we can observe that most of the compared algorithms have difficulty in reconstructing the lost details in the low-resolution image. Specifically, we can see from image 'img092' that some unpleasant artifacts are produced during the degradation process. All compared methods cannot handle this problem well, since the texture directions obtained by these methods are all inaccurate to a varying degree. In contrast, we can see that our model can alleviate the artifact effects and reconstruct accurate and clear contents. The proposed model makes full use of multi-scale information to obtain more faithful SR results. The proposed model makes full use of multi-scale information to obtain more faithful SR results, which can be further verified through the SSIM scores of our model since the SSIM focuses on the visible structures in the image.

4.3 Experiment on real-world images

To further present the performance of the proposed model, some additional experiments are implemented using historical images with JPEG compression artifacts. We compare five existing methods: SRCNN [17], FSRCNN [53], SRMDNF [55], MSRN [31], and SEAN [57]. As shown in Fig. 5, the image from the historical dataset contains the



(a) Urban100: img 092 **(b)** img 050 **(c)**B100: 302008

(d) The zoomed region in (a)

(e) The zoomed region in (b)

(f) The zoomed region in (c)

Fig. 4 Image super-resolution results with scaling factors × 4

letters "NOV" and "CEWV." The first four compared methods suffered from blurring and unpleasant artifacts. The results of SEAN are cleaner than those four methods, but it produces some errors. By contrast, the proposed method recovers clearer and more natural super-resolved results. Except for the MSRN method, most of the compared models do not make full use of multi-scale features to enhance the representation capability of the network. Although the MSRN uses a two-bypass network with a different convolution kernel to extract multi-scale image features, its feature fusion method is a simple concatenation operation. In addition, the proposed method fuses hierarchical features from both multi-scale features and shallow feature extraction layers for the final representation. These comparison results demonstrate the benefits of learning multi-scale features from the input image, which

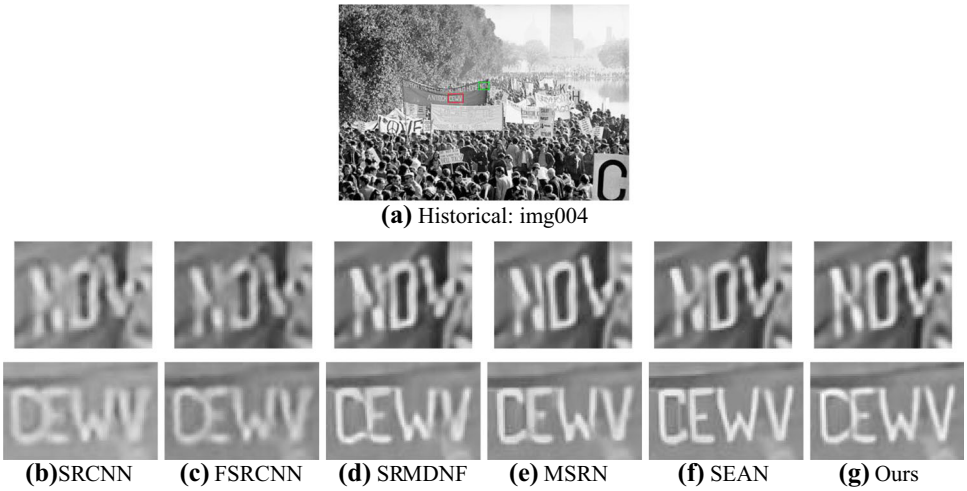


Fig. 5 Visual results on real-world images with scaling factor × 4

(a) Historical: img004

(b)SRCNN **(c)** FSRCNN **(d)** SRMDNF **(e)** MSRN **(f)** SEAN **(g)** Ours

enables our network to perform robustly across various scene types.

4.4 Ablation studies

This subsection performs an ablation study for the components in the proposed network over the super-resolution task with a × 4 scale factor. It first studies how the representation module affects the image SR performance by using the estimation of the PSNR and SSIM. Our network outputs different response features, from low-to-high resolutions. The output of V1 is the representation only from the high-resolution branch, and the output of V2 is concatenated and is mixed with different representations in series. While the V3 is our proposed representation method, we gradually upsample and concatenate different feature representations from low-to-high resolutions. Table 2 shows that the proposed representation method brings improved performance. Next, we analyzed our feature fusion strategy in Table 3, and the results indicate that the proposed selective scale fusion network achieves superior performance compared to summation and concatenation.

Table 2 Ablation study on different representation modules

Representations	V1	V2	V3 (ours)
Set5	32.45/0.8988	32.48/0.8990	**32.54/0.9006**
Set14	28.79/0.7924	28.82/0.7925	**28.85/0.7931**
B100	27.69/0.7479	27.71/0.7482	**27.73/0.7489**
Urban100	26.49/0.8027	26.59/0.8052	**26.68/0.8081**

Table 3 Ablation study on different feature aggregations

Fusion Methods	Summation	Concatenation	Ours
Set5	32.42/0.8979	32.48/0.8989	**32.54/0.9006**
Set14	28.80/0.7920	28.83/0.7924	**28.85/0.7931**
B100	27.70/0.7473	27.72/0.7485	**27.73/0.7489**
Urban100	26.59/0.8051	26.63/0.8062	**26.68/0.8081**

4.5 Analysis and discussion

The above extensive experiments on benchmark datasets demonstrate that the proposed approach substantially outperforms the state-of-the-art methods in terms of the quantitative metrics (PSNR and SSIM) and visual quality. Besides, the comparative advantages of our model become more appealing, surpassing these competing methods on the real-world test sample. For a single image SR problem, input and output images are highly correlated. It is crucial to fully exploit the features of the input image and transfer them to the end of the network for reconstruction. Most deep learning-based image SR methods either maintain the original resolution features along with the network structure or use a pre-upscaling operation to process the target resolution. The former helps preserve accurate spatial details, and the latter provides a better representation of contextual information. However, most existing methods only exhibit one of the above advantages although combining the two aspects can maximize useful information from the original image. In contrast, our model can gradually aggregate this hierarchical information to form more representative features. The success of the proposed method stems from two aspects: (1) gradually upsampling by adding low-to-high resolution subnetworks one by one to form more stages, and connection of the multi-resolution subnetworks in parallel; and (2) nonlinear multi-resolution

feature fusion repeatedly using a self-attention mechanism and rendering reliable high-resolution features. Overall, an ambidextrous design of progressive multi-scale fusion network not only inherits the superiority of post-upscaling structure but also utilizes the complementary set of multi-scale branches. Our network provides better-contextualized features, and the proposed repeated multi-scale fusion dynamically combines variable receptive fields.

5 Conclusion

In this paper, we propose a new network structure for single image SR, in which the base stream was designed for original-scale processing and a series of complementary parallel streams to explore rich multi-scale information. Furthermore, a new selective scale fusion mechanism is proposed to learn the relationships between the features across multi-scale branches and to adaptively fuse the multi-scale features. The proposed feature fusion strategy can preserve the original image details while dynamically adjusting the receptive field. The effectiveness of our method is demonstrated through consistently better results compared to state-of-the-art image SR tasks on several benchmark datasets.

Acknowledgements This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2019R1D1A3A03103736, in part by the Natural Science Foundation of Jiangxi Province under Grant 20202BABL212007, and in part by the Project of the Education Department of Jiangxi Province under Grant GJJ210628.

Declarations

Conflict of interest The authors declare no conflict of interest. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work entitled "Single Image Super-Resolution via Deep Progressive Multi-scale Fusion Networks."

References

1. Huang S, Sun J, Yang Y, Fang Y, Lin P, Que Y (2018) Robust single-image super-resolution based on adaptive edge-preserving smoothing regularization. IEEE Trans Image Process 27(6):2650–2663
2. Li F, Bai H, Zhao Y (2020) filternet: adaptive information filtering network for accurate and fast image super-resolution. IEEE Trans Circuit Syst Video Technol 30(6):1511–1523
3. Wang H, Gao X, Zhang K, Li J (2016) Single-image super-resolution using active-sampling gaussian process regression. IEEE Trans Image Process 25(2):935–948
4. Zhu Z, Guo F, Yu H, Chen C (2014) Fast single image super-resolution via self-example learning and sparse representation. IEEE Trans Multimed 16(8):2178–2190
5. Park SC, Park MK, Kang MG (2003) Super-resolution image reconstruction: a technical overview. IEEE Signal Process Mag 20(3):21–36
6. Yue L, Shen H, Li J, Yuanc Q, Zhang H, Zhang L (2016) Image superresolution: the techniques, applications, and future. Signal Process 128:389–408
7. Tai YW, Liu S, Brown MS, Lin S (2010). Super resolution using edge prior and single image detail synthesis. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, USA, pp 2400–2407
8. Wang L, Xiang S, Meng G, Wu H-Y, Pan C (2013) Edge-directed single-image super-resolution via adaptive gradient magnitude selfinterpolation. IEEE Trans Circuits Syst Video Technol 23(8):1289–1299
9. Sun J, Sun J, Xu Z, Shum H-Y (2011) Gradient profile prior and its applications in image super-resolution and enhancement. IEEE Trans Image Process 20(6):1529–1542
10. Protter M, Elad M, Takeda H, Milanfar P (2009) Generalizing the nonlocal-means to super-resolution reconstruction. IEEE Trans Image Process 18(1):36–51
11. Mairal J, Bach F, Ponce J, Sapiro G, Zisserman A (2009) Non-local sparse models for image restoration. In: 2009 IEEE 12th international conference on computer vision, IEEE, USA, pp 2272–2279
12. Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. IEEE Trans Image Process 19(11):2861–2873
13. Dong W, Zhang L, Shi G, Wu X (2011) Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. IEEE Trans Image Process 20(7):1838–1857
14. Peleg T, Elad M (2014) A statistical prediction model based on sparse representations for single image super-resolution. IEEE Trans Image Process 23(6):2569–2582
15. Ren C, He X, Nguyen TQ (2017) Single image super-resolution via adaptive high-dimensional non-local total variation and adaptive geometric feature. IEEE Trans Image Process 26(1):90–106
16. Zeyde R, Elad M, Protter M (2010) On single image scale-up using sparse-representations. In: International conference on curves and surfaces, Springer, Berlin, Heidelberg, pp 711–730
17. Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell 38(2):295–307
18. Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1646–1654
19. Shi Y, Wang K, Chen C, Xu L, Lin L (2017) Structure-preserving image super-resolution via contextualized multitask learning. IEEE Trans Multimed 19(12):2804–2815
20. Kim J, Lee JK, Lee KM (2016) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1637–1645
21. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 624–632
22. Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 136–144
23. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2472–2481

24. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
25. Hui Z, Wang X, Gao X (2018) Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 723–731
26. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
27. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
28. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
29. Haris M, Widyanto MR, Nobuhara H (2017) Inception learning super-resolution. Appl Opt 56(22):6043–6048
30. Tong T, Li G, Liu X, Gao,Q (2017) Image super-resolution using dense skip connections. In: Proceedings of the IEEE international conference on computer vision, pp 4799–4807
31. Li J, Fang F, Mei K, Zhang G (2018) Multi-scale residual network for image super-resolution. In: Proceedings of the European conference on computer vision (ECCV), pp 517–532
32. Qin J, Huang Y, Wen W (2020) Multi-scale feature fusion residual network for single image super-resolution. Neurocomputing 379:334–342
33. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
34. Chen Y, Li J, Xiao H, Jin X, Yan S, Feng J (2017) Dual path networks. In: Advances in neural information processing systems, 30
35. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P (2019) Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662
36. Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C (2015) Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 648–656
37. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision, Springer, Cham, Denmark, pp 483–499
38. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, Cham, Denmark, pp 234–241
39. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
40. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J (2018) Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7103–7112
41. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2881–2890
42. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
43. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5693–5703
44. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp 315–323. JMLR workshop and conference proceedings
45. Li X, Wang W, Hu X, Yang J (2019) Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 510–519
46. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
47. Zhao H et al (2017) Loss functions for image restoration with neural networks. IEEE Trans Comput Imag 3(1):47–57
48. Timofte R, Agustsson E, Van Gool L, Yang MH, Zhang L (2017) Ntire 2017 challenge on single image super-resolution: methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 114–125
49. Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML (2012) Low-complexity single-image super-resolution based on non-negative neighbor embedding. In: Proceedings of the British machine vision conference
50. Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings eighth IEEE International conference on computer vision. ICCV 2001, vol 2, IEEE, pp 416–423
51. Huang JB, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5197–5206
52. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. In: Proceedings of the 3rd international conference for learning representations
53. Dong C, Loy CC, Tang X (2016). Accelerating the super-resolution convolutional neural network. In: European conference on computer vision, Springer, Cham, Denmark, pp 391–407
54. Ahn N, Kang B, Sohn KA (2018) Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV), pp 252–268
55. Zhang K, Zuo W, Zhang L (2018) Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3262–3271
56. Zamir SW et al (2020) Learning enriched features for real image restoration and enhancement. In: European conference on computer vision, Springer, Cham, Denmark, pp 492–511
57. Fang F, Li J, Zeng T (2020) Soft-edge assisted network for single image super-resolution. IEEE Trans Image Process 29:4656–4668
58. Li Z, Yang J, Liu Z, Yang X, Jeon G, Wu W (2019) Feedback network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3867–3876
59. Que Y, Lee HJ (2021) Residual dense U-Net for abnormal exposure restoration from single images. IET Image Process 15(4):115–126