Feedback Pyramid Attention Networks for Single Image Super-Resolution

Huapeng Wu, Jie Gui, Senior Member, IEEE, Jun Zhang, Member, IEEE, James T. Kwok, Fellow, IEEE, and Zhihui Wei, Member, IEEE

Abstract—Recently, convolutional neural network (CNN) based image super-resolution (SR) methods have achieved significant performance improvement. However, most CNN-based methods mainly focus on feed-forward architecture design and neglect to explore the feedback mechanism, which usually exists in the human visual system. In this paper, we propose feedback pyramid attention networks (FPAN) to fully exploit the mutual dependencies of features. Specifically, a novel feedback connection structure is developed to enhance low-level feature expression with high-level information. In our method, the output of each layer in the first stage is also used as the input of the corresponding layer in the next state to re-update the previous low-level filters. Moreover, we introduce a pyramid non-local structure to model global contextual information in different scales and improve the discriminative representation of the network. Extensive experimental results on various datasets demonstrate the superiority of our FPAN in comparison with the state-of-the-art SR methods.

Index Terms—Super-resolution, feedback mechanism, pyramid non-local structure.

I. INTRODUCTION

SINGLE image super-resolution (SISR) [1] is a class of techniques that infers a high resolution (HR) image from its corresponding low resolution (LR) image. SR is an ill-posed inverse problem since each LR patch has to be mapped onto multiple HR patches. To solve this problem, researchers have proposed a multitude of learning-based methods to learn the mapping function from LR patches to their HR counterparts [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15].

Due to the strong learning ability, deep neural network [16] based methods have been proposed to learn the nonlinear mapping between LR and HR image pairs. Dong et al. [4] firstly presented a super-resolution convolutional neural network (SRCNN) to learn an end-to-end nonlinear mapping from LR to its HR counterpart. Kim et al. proposed a very deep super-resolution (VDSR) network [5] by using residual learning with up to 20 convolutional layers, achieving substantial improvement over the SRCNN method. In order to reduce the parameters of the model, Kim et al. further proposed Deeply-Recurrent Convolutional Network (DRCN) [6] via utilizing recursive-supervision and achieved similar results to VDSR. Tai et al. introduced Deep Recursive Residual Network (DRRN) [7], which utilized local and global residual learning to train a deeper model and showed favorable performances against DRCN with fewer parameters. Based on Laplacian pyramid framework, Lai et al. [8] proposed the Laplacian Pyramid Super-Resolution Network (LapSRN) to reconstruct HR images progressively. Ledig et al. [10] applied residual net (ResNet) architecture to construct a deeper network (SRRResNet) for image SR. Enhanced Deep Super-Resolution (EDSR) [11] proposed by Lee et al. employed a simplified ResNet [17] architecture by removing the normalization layers in the SRResNet and won the championship of the NTIRE2017 super-resolution challenge [18]. To further improve the performance of the networks, Residual Dense Network (RDN) [19] and Channel-wise and Spatial Feature Modulation Network (CSFM) [20] utilize the residual and the dense skip connections [21] in their networks to fuse multi-level features for the SR reconstruction.

Although these feedforward methods achieve good results, the feedback mechanism [22], [23], [24], [25] has not been fully exploited in deep learning-based super-resolution methods. In cognition theory, feedback connections that link the cortical visual areas can transfer response signals from high-level cortical areas to the current low-level neuron [23], [24]. Inspired by this mechanism, feedback connections have been used in many deep neural networks [26], [27], [28]. The feedback connections can bring the semantic information from the higher layers back to refine the information of the previous layers in a top-down manner. Recently, Haris et al. [29] introduced the feedback mechanism into an SR module to update feature errors iteratively. Liu et al. [30] proposed the new feedback version Hierarchical Back Projection Network (HBPN), applying the new up-sampling and down-sampling back projection block to return the LR and HR residual errors. To further improve the performance, Liu et al. [31] introduced the attention mechanism to the feedback block to...
In this paper, we propose a novel simple feedback connection layers and feedback skip connection layers, which can increase the training difficulty of the network. Therefore, in this paper, we propose a novel simple feedback connection structure to promote the flow of information through the network.

**Attention mechanism.** In recent years, attention mechanisms have been successfully applied to a large number of tasks via rescaling the feature response to concentrate on the important components of the inputs [35], [36], [37], [38], [39], [40], [41], [42]. Hu et al. [35] introduced the channel-wise attention mechanism, which proposed a squeeze-and-excitation block (SENet) to rescale channel-wise feature responses for image classification and achieved excellent results. Later, Zhang et al. [36] directly integrated the squeeze-and-excitation block into the residual architecture to construct a residual channel attention network (RCAN) and achieved significant performance gains for image SR. Inspired by classic non-local mean, Wang et al. [43] introduced a non-local neural network that firstly incorporated non-local operations in deep convolutional neural networks for image recognition and video classification. In the non-local block, the spatial attention mask is obtained by computing the similarity matrix between each query position and all positions in the input feature map. Then they adaptively modulate feature representations at a position by a weighted sum of the features at all positions based on the obtained spatial attention mask. Considering that the non-local network (NL) calculates the correlation on the entire feature map, which will increase the computational complexity dramatically when the spatial dimensionality is large. To resolve this problem, Liu et al. [44] introduced a non-local recurrent network (NLRN) with the limited neighborhood for image restoration. Dai et al. [37] adopted region-level non-local enhancement operations in second-order attention network (SAN). However, SAN [37] and NLRN [44] still suffer from a huge computational burden, which greatly hinders the use of non-local blocks. GCNet proposed by Cao et al. [45] surprisingly found that the attention maps of different query positions are almost the same. By simplifying the non-local block, GCNet has the advantage of the non-local block with capturing the long-range dependencies and the SENet with the lightweight property.

In this paper, we propose a novel feedback pyramid attention network (termed FPAN) for SISR, which is shown in Fig. 1. In the FPAN, we can not only enhance the representative ability of the proposed network but also improve the flow of information across the network by using a feedback connection structure. As a building block, feedback pyramid attention block (FPAB) is constructed by multiple forward skip connection layers and feedback skip connection layers, which can generate powerful high-level representations (illustrated in Fig. 2). The principle of our feedback scheme is that the output of each layer in the first stage is used as the input of the corresponding layer in the second stage to refine low-level representations. In each block, we further put forward to Laplacian pyramid attention to study non-local correlations at different scales, which is beneficial to enhance the discrimination of the network (shown in Fig. 2 (b)). The corresponding experiments show that our proposed method achieved favorable results in comparison with the state-of-the-art SR methods.

In summary, the main contributions of our method are as follows.

1. We introduce a feedback pyramid attention network (FPAN) for SISR. Our FPAN can adaptively refine feature representations and capture richer context information for image SR.

2. We propose a novel feedback connection mechanism, which not only efficiently improves the information flow, but also enhances high-level feature representation via multiple skip connections. Besides, the proposed pyramid non-local block can capture long-distance spatial contextual information at multiple scales, which efficiently and adaptively recalibrates feature presentations. In comparison with state-of-the-art methods, our method shows superiority in terms of model size and performance.

The rest of this paper is organized as follows. In Section II, we analyze the proposed FPAN in detail. Experimental results...
Fig. 2. (a) The illustration of our feedback pyramid attention block (FPAB). (b) The details of the pyramid non-local block.

$\otimes$ denotes matrix multiplication.

$\oplus$ denotes broadcast element-wise addition.

with analysis are presented in Section III to demonstrate the efficiency of our method. Finally, conclusions are provided in Section IV.

II. PROPOSED METHOD

In this section, we will introduce the proposed feedback pyramid attention networks in detail.

A. Network Architecture

The overall framework of our FPAN is illustrated in Fig. 1, which is composed of two parts: 1) feature extraction part, and 2) reconstruction part. In this paper, $I^{LR}$ and $I^{SR}$ represent the input LR image and output HR image of our FPAN. We firstly utilize one convolution layer to extract the initial features $F_0$ from the input LR.

$$F_0 = H_0 \left( I^{LR} \right), \quad (1)$$

where $H_0$ denotes a convolution operation. $F_0$ is then fed into a series of feedback pyramid attention blocks (FPABs) to refine feature maps. Supposing $G$ FPABs are used for further feature extraction, that is

$$F_G = M_G \left( M_{G-1} \left( \cdots M_1 \left( F_0 \right) \cdots \right) \right), \quad (2)$$

where $F_0$ is the output of the first convolution layer $H_0$, $M_g \left( g = 1, 2, \cdots, G \right)$ denotes the $d$-th FPAB operation.

After acquiring a set of feature maps $\left[ F_G, F_{G-1}, \cdots, F_1 \right]$, we apply global feature fusion and global skip connection:

$$F_{GF} = F_0 + H_{hff} \left[ F_G, F_{G-1}, \cdots, F_1 \right], \quad (3)$$

where $H_{hff}$ is composed of hierarchical feature fusion (1 $\times$ 1 convolution filter) and last feature extraction function (3 $\times$ 3 convolution filter), $F_{GF}$ is the output of the feature extraction part. Subsequently, we utilize a sub-pixel convolution layer [46] followed by a convolution layer for
reconstructing HR image.

\[ I^{SR} = H_{\text{recon}} (F_G) = H_{\text{FPAN}} (I^{LR}), \]

where \( H_{\text{recon}} \) is the upscaling and reconstruction operation, \( H_{\text{FPAN}} \) represents the function of our FPAN.

The FPAN is optimized with the \( L_1 \) loss function [36], [37], [38] by minimizing the difference between the reconstructed image \( I^{SR} \) and the ground truth \( I^{HR} \). Given a training dataset \( \{I_i^{LR}, I_i^{HR}\}_{i=1}^{N} \), where \( N \) represents the number of the LR inputs and the corresponding HR inputs. The loss function is formulated as

\[ L(\theta) = \frac{1}{N} \sum_{i=1}^{N} ||H_{\text{FPAN}} (I_i^{LR}) - I_i^{HR}||_1, \]

where \( \theta \) denotes the parameter set of the FPAN. In the next subsection, we will introduce the proposed feedback pyramid attention block in detail.

### B. Feedback Pyramid Attention Block (FPAB)

The structure of the feedback pyramid attention block (FPAB) is shown in Fig. 2, the module mainly contains three parts: feedback connection structure, pyramid non-local block and local residual learning. Let \( F_{g-1} \) and \( F_g \) denote the input and output of the \( g \)-th FPAB and \( F_g \) can be obtained by

\[ F_g = F_{g-1} + f_{\text{PN}} (f_{\text{FC}} (F_{g-1})), \]

where \( f_{\text{FC}} \) and \( f_{\text{PN}} \) are the proposed feedback connection structure and pyramid non-local block.

#### 1) Feedback Connection Structure

The feedback mechanism is known as the efficient iterative procedure to refine low-level representations using high-level semantic information. Unlike the clique structure [33] that each layer is both the input and output of another one in the same block, we propose a novel feedback connection mechanism, where the output of each layer except for the input layer in the first stage is used as the input of the corresponding layer in the second stage in the same block. As illustrated in Fig. 2 (a), the proposed feedback structure contains two stages. At the beginning, the input layer \( F_{g-1} \) initializes all layers by simple feedforward skip connections. The outputs of two adjacent convolutional layers are fed to the subsequent layer in the first stage. In the second stage, the output feature \( X_1 \) of the first stage is fed to the second stage as the new input to update the previous layers by building a skip connection between the corresponding convolutional layers of the two stages. Specifically, the proposed feedback connection mechanism can be presented as follows.

The initial stage is

\[ X_0 = \sigma (W_0 \ast F_{g-1}). \]

The first stage is

\[ X_i^1 = \sigma (W_i^1 \ast [X_i^{i-1}, X_i^{i-2}]), \]

and the second stage is

\[ X_i^2 = \sigma (W_i^2 \ast [X_i^{i-1}, X_i^{i-1}]), \]

where \( X_i^j \) denotes the feature maps of the \( i \)-th \((i \geq 1)\) convolutional layer in the \( j \)-th \((j = 1, 2)\) stage. \( X_i^{1} = F_{g-1} \). \( X_i^{1} = X_0 \). \( X_i^{2} = X_1 \) is the output of the first stage. \( \ast \) is the convolutional operator. \( \sigma \) represents the activation function (ReLU) [47]. \([\cdot, \cdot]\) refers to the concatenation operation. Then, the output of each stage is fed to the subsequent layers to enhance the information propagation in the network.

#### 2) Pyramid Non-Local Block

SR aims to restore high frequency components of images. However, most SR methods recover image details blindly and lack abilities to identify high frequency regions. Thus, we consider if we can enhance the network sensitivity to high frequency information, thereby further improving the representational ability of the network, which is critical for SR. On the other hand, the convolution operation focuses only on local regions and cannot obtain long-dependencies at a time. Given this, in this paper, we design a pyramid non-local block illustrated in Fig. 2 (b), which can capture rich contextual information in the spatial dimensions and improve the discriminative ability of the model.

##### a) Non-local Module

The non-local module [43] can learn long-range relationships among any spatial positions and has been widely applied in visual tasks. To further improve the representation ability of the network, we will introduce a non-local module into the network to capture rich contextual information and enlarge the receptive field to the entire feature map. Let \( X = \{x_i\}_{i=1}^{N} \) denotes the input of the non-local module, which has \( C \) feature maps with size of \( H \times W \). For the convenience of the calculation, we reshape \( X \) to \( X \in \mathbb{R}^{C \times N} \), i.e., \( N = H \times W \) and the spatial attention operation is defined as

\[ S = f (\theta (X)^T, \varphi (X)), \]

where \( S \in \mathbb{R}^{N \times N} \), \( \theta (\cdot) \) and \( \varphi (\cdot) \) are two linear feature embeddings (e.g. \( 1 \times 1 \) convolution operations). \( f (\cdot, \cdot) \) is the similarity function that measures the correlations between any two positions in the embedding feature map. Specifically, most papers adopt softmax as a measure function:

\[ s_{ij} = f (\theta (x_i)^T, \varphi (x_j)) = \frac{\exp (\theta (x_i) \ast \varphi (x_j))}{\sum_{j=1}^{N} \exp (\theta (x_i) \ast \varphi (x_j))}, \]

where \( s_{ij} \) represents the similarity between the \( j \)-th location \( x_j \) and the \( i \)-th location \( x_i \) in the spatial dimension. Similarly, we also consider a linear embedding operation for input \( X \): \( g (X) = W_g X \) with the learned parameters \( W_g \), which can be implemented as \( 1 \times 1 \) convolution and we reshape it to \( \mathbb{R}^{C \times N} \). Then, the final output \( y_i \) at location \( i \) as follows:

\[ y_i = W_y \ast \sum_{j=1}^{N} (s_{ij} \ast g (x_j)) + x_i, \]

where \( j \) is the index that enumerates all possible positions, capturing long-range contextual correlations across the whole feature map. \( W_y \) is the learned weight parameter (initialized as zero) and the residual connection structure “+x_i” guarantee the flow of the original information.
However, this non-local operation requires a lot of computational costs, especially when the spatial dimensionality is large. A global context block (GC) was presented in [45]. The GC block simplifies the computational complexity of non-local module with SE block [35], while enhancing the discriminative presentation of the network based on an independent attention map. The GC block mainly contains three steps, 1) adopt global attention pooling to implement context modeling, 2) transform operation to achieve channel-wise dependencies, 3) feature fusion to aggregate the global context feature. The global context block is defined as

\[
y_i = x_i + W_{c2}ReLU \left( LN \left( \sum_{j=1}^{N} \frac{e^{W_{c1}j}}{\sum_{m=1}^{N} e^{W_{c1}m}} x_j \right) \right)
\]

(13)

where \(W_{c}, W_{c1}\), and \(W_{c2}\) denote three linear feature embeddings (e.g. \(1 \times 1\) convolutional operations), \(\alpha_j = \frac{e^{W_{c1}j}}{\sum_{m=1}^{N} e^{W_{c1}m}}\) represents the weight for context modeling, \(LN (\cdot)\) is layer normalization [48], \(\delta (\cdot) = W_{c2}ReLU (LN (W_{c1} (\cdot)))\) is the feature transform, and “+” denotes the feature fusion.

b) Pyramid Non-local Mechanism: Inspired by the above non-local mechanism, we design a pyramid non-local structure to exploit the relationship between the features at different scales. Although the proposed pyramid non-local structure is somewhat similar to channel attention, channel attention uses the average pooling operation to aggregate the global contextual information. In the proposed pyramid non-local structure, as illustrated in Fig. 2 (b), similar to the non-local structure, in each scale, we adopt \(1 \times 1\) convolution to learn a spatial attention map independent of query position. After performing softmax and reshaping, the spatial attention map is multiplied by the corresponding input feature to get a feature vector with global information \((C \times 1 \times 1)\). Then, the feature vectors at different scales are concatenated together to learn the channel-wise interdependencies via adopting the channel-wise rescaling strategy. Finally, multi-scale global context features are added to the features of all positions to achieve feature fusion. In comparison with the previous non-local network, the proposed pyramid non-local structure can not only greatly reduce the computational complexity by simplifying the calculation of spatial attention mask but also capture global contextual information at different scales, which is beneficial to further improve the discriminative representation of the network. Therefore, our method can achieve an effective combination of non-local structure and channel-wise dependencies. The pyramid non-local structure can be formulated as

\[
\begin{aligned}
Y &= X + PNLB (X) \\
PNLB (X) &= \delta (Concat (PA_1, PA_2, PA_4)) \\
PA_i &= \sum_{j=1}^{N} a^i_j x_j \\
a^i &= soft \max (W^i_k x^i)
\end{aligned}
\]

(14)

where \(PNLB (\cdot)\) denotes pyramid non-local operation, \(i\) is the scale parameter of the Laplacian pyramid, \(PA_i\) represents the context modeling module at the \(i\)-th scale. \(Concat (\cdot)\) is a feature concatenation operation. The lightweight of the PNLB allows it to be plugged into multiple layers to better capture the long-range dependency with only a slight computational burden.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets and Evaluation Metrics

In the experiment, we adopt 800 high-quality images from DIV2K dataset [18] as the training set. Meanwhile, we use five standard benchmark datasets for testing, including Set5 [49], Set14 [50], Bsd100 [51], Urban100 [52], and Manga109 [53] with 5, 14, 100, 100 and 109 images for 3 upscaling factors, respectively. We carry out three experiments with three different degradation models, including bicubic (BI), blur-downscale (BD), and downscale-noise (DN) degradation models. The parameter settings of all degradation models are consistent with RDN [19]. The SR images are first converted to YCbCr color space and evaluated by PSNR and SSIM [54] on Y channel (i.e., luminance).

B. Implementation Details

Data augmentation is performed on above 800 training images, which includes random rotating 90°, 180°, 270°, and horizontal flipping. In each min-batch, 16 LR color patches with the size of \(48 \times 48\) are randomly extracted as inputs. Our model is optimized with Adam [55], where the parameters are set to \(\beta_1 = 0.9, \beta_2 = 0.999\), and \(\varepsilon = 10^{-8}\). The learning rate is initialized to \(10^{-4}\) and reduced to half every 200 epochs. We implement our model by utilizing PyTorch [56] with an Nvidia 2080Ti GPU.

In the proposed network, we set all convolution layers to have 64 filters and the kernel size is \(3 \times 3\) except for those \(1 \times 1\) convolutional layers in the pyramid non-local unit and feature fusion part. We adopt zero-padding strategy to keep the feature size fixed. In PNLB, we use \(6 \times 6\) convolutional layer with two striding and two padding to perform down-sampling operations. In addition, the reduction ratio of the
C. Ablation Analysis

In this section, we investigate the effects of different components, including feedforward dense skip connections, feedback skip connections and pyramid attention. For a fair comparison, the baseline and its variants share the same experimental configuration and are trained from scratch.

To demonstrate the effectiveness of the proposed feedback connection mechanism, we first compare the feedback network (P2) with its feedforward counterpart (P0, P1). As shown in Table I, the baseline P0 (without the feedforward dense skip connections, the feedback skip connections, and attention mechanism) performance is relatively low, its PSNR only reaches 32.33 dB on Set5 (×4). Then, P1 that adds the feedforward dense skip connections obtain 32.35 dB. After integrating the feedback skip connections, the performance can be improved from 32.35 dB to 32.43 dB. In Fig. 3, we visualize the average feature maps from the first convolution layer (Conv 1) without/with feedback structure. Each average feature map is the mean of the output feature in the channel dimension. In comparison with the network without feedback structure, the features derived from the feedback network contain richer texture structure details and show stronger feature representation ability. Specifically, in the lower part of Fig. 3, we notice that the feedback network can accurately learn the shape of the nostrils. This is due to that the feedback structure can not only further extract high-frequency information similar to a high-pass filter, but also increase the size of the receptive field to produce better representations. Meanwhile, as discussed above, we further show the effect of attention. It can be seen that the network with attention (P3 and P4) performs better than those without attention (from P0 to P2). These observations indicate that the attention mechanism can adaptively exploit more important information for SR. Specifically, when we compare the global context block [45] (P3, S = 1) and pyramid non-local block (P4, S = 1, 2, 4), we can see that our pyramid attention mechanism can achieve better performance than the network with the global context block (i.e., 32.55 dB v.s. 32.49 dB). In addition, we also show some other quantitative comparison results, including the number of parameters and the number of floating-point operations (FLOPs), which are illustrated in Fig. 4. FLOPs is computed with 512 × 512 HR image at ×4 scaling factor. It can be found that although more pyramid scales can improve PSNR, the number of parameters and FLOPs will also increase. It is worth noting that two scales (S = 1, 2) yield worse results than the case with one scale (S = 1) for PSNR performance on Set5, but it is different on other data sets. We think that this problem may be caused by different image properties. Besides, considering that the three-
| Methods          | Scale/Datasets /Para (M) | Set5  | Set14  | Bsd100 | Urban100 | Manga109 |
|------------------|-------------------------|-------|--------|--------|----------|----------|
|                  |                         | PSNR  | SSIM   | PSNR   | SSIM     | PSNR     | SSIM     | PSNR   | SSIM     |
| Bicubic          |                         | 33.66 | 0.9299 | 30.24  | 0.8688  | 29.56    | 0.8431  | 26.88  | 0.8403  |
| SRCNN [4]        |                         | 36.66 | 0.9542 | 32.45  | 0.9067  | 31.36    | 0.8879  | 29.50  | 0.8946  |
| VDSR [5]         |                         | 37.53 | 0.9590 | 33.05  | 0.9130  | 31.90    | 0.8960  | 30.77  | 0.9140  |
| DRDCN [6]        |                         | 37.63 | 0.9588 | 33.04  | 0.9118  | 31.85    | 0.8942  | 30.75  | 0.9133  |
| LapSRN [8]       |                         | 37.52 | 0.9590 | 33.01  | 0.9130  | 31.87    | 0.8945  | 30.77  | 0.9130  |
| EDSR [11]        |                         | 38.11 | 0.9602 | 33.92  | 0.9195  | 32.32    | 0.9013  | 32.93  | 0.9351  |
| DBPN [29]        |                         | 38.09 | 0.9600 | 33.85  | 0.9190  | 32.27    | 0.9000  | 32.55  | 0.9324  |
| SRJFBN [32]      |                         | 38.11 | 0.9609 | 33.82  | 0.9196  | 32.29    | 0.9010  | 32.62  | 0.9328  |
| RCAN(2018) [36]  | DIV2K/15.44             | 38.27 | 0.9614 | 34.12  | 0.9216  | 32.34    | 0.9027  | 33.34  | 0.9384  |
| RCAN+(2018) [36] | DIV2K/15.44             | 38.33 | 0.9617 | 34.23  | 0.9225  | 32.41    | 0.9031  | 33.54  | 0.9399  |
| SAN(2019) [37]   | DIV2K/15.67             | 38.31 | 0.9620 | 34.07  | 0.9213  | 32.42    | 0.9028  | 33.10  | 0.9370  |
| SAN+(2019) [37]  | DIV2K/15.67             | 38.35 | 0.9619 | 34.44  | 0.9244  | 32.50    | 0.9038  | 33.73  | 0.9416  |
| DRLN(2020) [14]  | DIV2K+Flicker2K /34.43  | 38.27 | 0.9616 | 34.28  | 0.9234  | 32.44    | 0.9028  | 33.73  | 0.9390  |
| DRLN+(2020) [14] | DIV2K+Flicker2K /34.43  | 38.34 | 0.9619 | 34.43  | 0.9247  | 32.47    | 0.9032  | 33.54  | 0.9402  |
| HAN(2020) [40]   | DIV2K/63.61             | 38.27 | 0.9614 | 34.16  | 0.9217  | 32.41    | 0.9027  | 33.35  | 0.9385  |
| HAN+(2020) [40]  | DIV2K/63.61             | 38.33 | 0.9617 | 34.24  | 0.9224  | 32.45    | 0.9030  | 33.53  | 0.9398  |
| FPAN (ours)      | DIV2K/14.02             | 38.19 | 0.9612 | 33.88  | 0.9210  | 32.30    | 0.9012  | 32.72  | 0.9393  |
| FPAN+(ours)      | DIV2K/14.02             | 38.25 | 0.9614 | 33.93  | 0.9210  | 32.35    | 0.9017  | 32.94  | 0.9357  |

scale ($S = 1, 2, 4$) model obtains a significant performance improvement over the previous two models. Although the four-scale model ($S = 1, 2, 4, 8$) is also increasing, the growth gate becomes small and the parameters and FLOPs are still increasing as $S$ increases. Therefore, to balance the speed and accuracy, we finally set the pyramid scale $S$ to 1, 2, and 4 in our method. These comparisons consistently demonstrate the superiority of our proposed FPAN.
D. Results With Bicubic Degradation (BI)

To verify the effectiveness of our FPAN, we compare our FPAN with some other state-of-the-art SR methods: Bicubic, SRCNN [4], VDSR [5], DRCN [6], LapSRN [8], EDSR [11], DBPN [29], SRFBN [32], RCAN [36], SAN [37], DRLN [14], HAN [40] and RFN [15]. As in [11], [19], and [36], we also adopt self-ensemble strategy to further improve our FPAN and denote the self-ensembled FPAN as FPAN+. Table II shows all the quantitative results for various scale factors. In comparison
with these methods, our method is comparable to these recently proposed state-of-the-art SR methods (including the self-ensemble methods). Compared with the classic SR feedback networks (DBPN [29] and SRFBN [32]), our method employs fewer training images (DIV2K + Flicker2K + ImageNet v.s. DIV2K + Flicker2K v.s. DIV2K) and achieves competitive results, while only needs the 20.8% and 14.6% FLOPs of DBPN [29] and SRFBN [32], respectively. Among these methods, DRLN and RFN contain much more training images at the cost of more training time. Although our performance is relatively low, our FPAN has fewer parameters than other SR methods. Specifically, the parameters of DRLN and HAN are several times more than our FPAN (34.58 M v.s. 64.20 M v.s. 14.16 M). The above results imply that our FPAN can achieve a comparable tradeoff between performance and model complexity.

### Table III

**Quantitative Results With BD and DN Degradation Modules. Average PSNR/SSIM Comparison Results for Scaling Factor ×3**

| Methods     | Model | Set5  | Set14  | BSD100 | Urban100 | Manga109 |
|-------------|-------|-------|--------|--------|----------|----------|
|             | PSNR  | SSIM  | PSNR  | SSIM  | PSNR    | SSIM    |
| Bicubic     |       |       |        |        |         |          |
| BD          | 28.78 | 0.8308| 26.38  | 0.7271| 23.52   | 0.6862  |
| DN          | 24.01 | 0.5369| 22.87  | 0.4724| 21.63   | 0.4687  |
| SRCNN [4]   |       |       |        |        |         |          |
| BD          | 32.05 | 0.8944| 28.80  | 0.8074| 28.13   | 0.7736  |
| DN          | 25.01 | 0.6950| 23.78  | 0.5898| 23.76   | 0.5558  |
| FSRCNN [57] |       |       |        |        |         |          |
| BD          | 26.23 | 0.8124| 24.44  | 0.7106| 24.86   | 0.6832  |
| DN          | 24.18 | 0.6932| 23.62  | 0.5856| 23.61   | 0.5556  |
| VDSR [5]    |       |       |        |        |         |          |
| BD          | 33.25 | 0.9150| 29.46  | 0.8244| 28.57   | 0.7989  |
| DN          | 25.20 | 0.7183| 24.00  | 0.6112| 24.00   | 0.5749  |
| IRCNN_G [58]|       |       |        |        |         |          |
| BD          | 33.38 | 0.9182| 29.63  | 0.8281| 28.65   | 0.7922  |
| DN          | 25.70 | 0.7379| 24.45  | 0.6305| 24.28   | 0.5900  |
| IRCNN_C [58]|       |       |        |        |         |          |
| BD          | 33.17 | 0.9157| 29.55  | 0.8271| 28.49   | 0.7886  |
| DN          | 27.48 | 0.7925| 25.92  | 0.6932| 25.55   | 0.6481  |
| RDN [19]    |       |       |        |        |         |          |
| BD          | 34.58 | 0.9280| 30.53  | 0.8447| 29.23   | 0.8079  |
| DN          | 28.47 | 0.8151| 26.60  | 0.7101| 25.93   | 0.6373  |
| RDN+ [19]   |       |       |        |        |         |          |
| BD          | 34.70 | 0.9289| 30.64  | 0.8463| 29.30   | 0.8093  |
| DN          | 28.35 | 0.8173| 26.67  | 0.7117| 25.97   | 0.6587  |
| SRFBN [32]  |       |       |        |        |         |          |
| BD          | 34.66 | 0.9283| 30.48  | 0.8439| 29.21   | 0.8069  |
| DN          | 28.53 | 0.8182| 26.60  | 0.7144| 25.95   | 0.6603  |
| SRFBN+ [32] |       |       |        |        |         |          |
| BD          | 34.77 | 0.9290| 30.64  | 0.8458| 29.28   | 0.8080  |
| DN          | 28.59 | 0.8198| 26.67  | 0.7159| 25.99   | 0.6636  |
| FPAN (ours) |       |       |        |        |         |          |
| BD          | 34.62 | 0.9286| 30.60  | 0.8455| 29.27   | 0.8095  |
| DN          | 28.53 | 0.8182| 26.63  | 0.7113| 25.95   | 0.6603  |
| FPAN+ (ours)|       |       |        |        |         |          |
| BD          | 34.72 | 0.9292| 30.71  | 0.8474| 29.34   | 0.8100  |
| DN          | 28.62 | 0.8193| 26.69  | 0.7128| 25.98   | 0.6612  |

Fig. 7. Visual quality for ×3 SR with BD and DN models on Set5 dataset [49]. The best results are highlighted.
To further illustrate the superiority of our FPAN, in Fig. 5, we show visual results with an upscaling factor $\times 4$ for these images from Urban100 dataset. From that, we can see that our method yields the best visual results among all existing compared methods. Using “image_092” as an example, we can see that the Bicubic interpolation generates heavy blurring artifacts along the edges and visually displeasing textures. Later, some developed methods can recover some structure, but still suffer from the wrong texture direction and fail to recover more accurate image details. However, the proposed FPAN alleviates distorted lines and produces more faithful textures and structures. In Fig. 6, we notice that other methods suffer from blurring artifacts and fail to restore the clear “H” letter, but the proposed FPAN can generate a more accurate “H” letter in the SR image. The reason is that FPAN can use high-level information to refine the low-level feature information to achieve self-correction. In addition, the pyramid attention mechanism is used to further enhance the discriminative representation of the network, thus generating a better SR result.

E. Results With BD and DN Degradations

Following [19] and [32], we also show SR results with BD and DN degradation models. We compare our FPAN with some state-of-the-art methods: SRCNN [4], FSRCNN [57], VDSR [5], IRCNN_G [58], IRCNN_C [58], RDN [19], SRFBN [32]. As shown in Table III, compared with other methods, our FPAN and FPAN+ can generate competitive results on almost all quantitative results with scaling factor $\times 3$. In particular, for the BD degradation model, our method achieves better performance than other state-of-the-art methods even without self-ensemble.

In Fig. 7, we show visual comparisons with BD and DN degradation models. The recovered results from other methods are blurry. In contrast, our FPAN and FPAN+ can alleviate the distortion and produce more faithful details in SR results. It also verifies that our FPAN has the ability to capture the important discriminative features for BD and DN degradation models.

F. Model Size Comparison

In Fig. 8, we illustrate the comparisons about model size and performance on Set 5 with scale factor $\times 2$. Our FPAN has much fewer parameters than EDSR but obtains higher performance. In comparison with DBPN and SRFBN, our FPAN and FPAN+ obtain better results with a relatively large model. It implies that our FPAN has a good tradeoff between model size and performance.

IV. Conclusion

In this paper, we propose a deep feedback pyramid attention network (FPAN) for image SR. Specifically, the proposed feedback connection structure allows FPAN to not only refine low-level representation with high-level information but also effectively boost the flow of information as well as the feature reuse. In addition, we introduce a pyramid non-local block to capture the long-distance dependencies in multiple scales and make the network concentrate on learning important high-frequency information. Extensive evaluations on SR with BI, BD and DN models show the superior performance of our FPAN in terms of quantitative and visual results.

REFERENCES

[1] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, “Learning low-level vision,” Int. J. Comput. Vis., vol. 40, no. 1, pp. 25–47, 2000.
[2] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in Proc. Asian Conf. Comput. Vis. Singapore: Springer, 2014, pp. 111–126.
[3] H. Wu, J. Zhang, and Z. Wei, “High resolution similarity directed adjusted anchored neighborhood regression for single image super-resolution,” IEEE Access, vol. 6, pp. 25240–25247, 2018.
[4] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2015.
[5] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1646–1654.
[6] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1637–1645.
[7] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3147–3155.
[8] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep Laplacian pyramid networks for fast and accurate super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 624–632.
[9] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, “MDCN: Multi-scale dense cross network for image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 7, pp. 2547–2561, Jul. 2021.
[10] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 4681–4690.
[11] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 136–144.
[12] J. Zhang et al., “A two-stage attentive network for single image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1020–1033, Mar. 2022.
[13] X. Zhu, K. Guo, S. Ren, B. Hu, M. Hu, and H. Fang, “Lightweight image super-resolution with expectation-maximization attention mechanism,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1273–1284, Mar. 2022.
[14] S. Anwar and N. Barnes, “Densely residual Laplacian super-resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1192–1204, Mar. 2022.
[18] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, “NTIRE 2017 challenge on single image super-resolution: Methods and results,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[15] Z. Hui, J. Li, X. Gao, and X. Wang, “Progressive perception-oriented network for single image super-resolution,” Inf. Sci., vol. 546, pp. 769–786, Feb. 2021.

[16] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, p. 436, May 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[29] M. Haris, G. Shakhnarovich, and N. Ukita, “Deep back-projection network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2472–2481.

[30] Y. Hu, J. Li, Y. Huang, and X. Gao, “Channel-wise and spatial feature modulation network for single image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 11, pp. 3911–3927, Nov. 2020.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[32] D. J. Felleman and D. C. Van Essen, “Distributed hierarchical processing in the primate cerebral cortex,” Cerebral Cortex, vol. 1, no. 1, pp. 1–47, Feb. 1991.

[33] C. D. Gilbert and A. M. Sigman, “Brain states: Top-down influences in sensory processing,” Neuron, vol. 54, no. 5, pp. 677–696, 2007.

[34] J. Hupé, A. James, B. Payne, S. Lomber, P. Girard, and J. Bullier, “Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons,” Nature, vol. 394, no. 6695, pp. 784–787, 1998.

[35] V. A. F. Lamme and P. R. Roelfsema, “The distinct modes of vision offered by feedforward and recurrent processing,” Trends Neurosciences, vol. 23, no. 11, pp. 571–579, Nov. 2000.

[36] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, “Human pose estimation with iterative error feedback,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4733–4742.

[37] A. Shrivastava and A. Gupta, “Contextual priming and feedback for faster R-CNN,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 330–348.

[38] A. R. Zamir et al., “Feedback networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1308–1317.

[39] M. Haris, G. Shakhnarovich, and N. Ukit, “Deep back-projection networks for super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1664–1673.

[40] Z.-S. Liu, L.-W. Wang, C.-T. Li, and W.-C. Siu, “Hierarchical pre-projection network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 2041–2050.

[41] Z.-S. Liu, L.-W. Wang, C.-T. Li, W.-C. Siu, and Y.-L. Chan, “Image super-resolution via attention based back projection networks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3517–3525.

[42] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, “Feedback network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3867–3876.

[43] Y. Yang, Z. Zhong, T. Shen, and Z. Lin, “Convolutional neural networks with alternately updated clique,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2413–2422.

[44] Z. Zhong, T. Shen, Y. Yang, Z. Lin, and C. Zhang, “Joint sub-bands learning with clique structures for wavelet domain super-resolution,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 165–175.

[45] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[46] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 286–301.

[47] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11065–11074.

[48] H. Wu et al., “Multi-grained attention networks for single image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 2, pp. 512–522, Feb. 2021.
Jie Gui (Senior Member, IEEE) received the B.S. degree in computer science from Hohai University, Nanjing, China, in 2004, the M.S. degree in computer applied technology from the Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China, in 2007, and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, in 2010. He is currently a Professor with the School of Cyber Science and Engineering, Southeast University. He has published more than 60 papers in international journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, KDD, and ACM MM. He is the Area Chair, a Senior PC Member, or a PC Member of many conferences, such as NeurIPS and ICML.

Jun Zhang (Member, IEEE) received the B.Sc. and M.Sc. degrees in computational mathematics from Wuhan University, Wuhan, China, in 1999 and 2002, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, Nanjing, China, in 2010. He is currently a Professor with the Nanjing University of Science and Technology. His current research interests include mathematical modeling and algorithms for inverse problems in image processing and machine learning, and the variational regularization methods and learning-based methods in image denoising, restoration, and super-resolution.

James T. Kwok (Fellow, IEEE) received the Ph.D. degree in computer science from The Hong Kong University of Science and Technology, Hong Kong, in 1996. He is currently a Professor with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology. His current research interests include kernel methods, machine learning, pattern recognition, and artificial neural networks. He received the IEEE Outstanding Paper Award in 2004 and the Second Class Award in Natural Sciences from the Ministry of Education, China, in 2008. He has been the program co-chair for a number of international conferences and served as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2006 to 2012. He is currently an Associate Editor of Neurocomputing.

Zhihui Wei (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from South East University, Nanjing, China, in 1983, 1986, and 2003, respectively. He is currently a Professor and a Doctoral Supervisor with the Nanjing University of Science and Technology. His current research interests include mathematical image modeling, multi-scale analysis, video and image coding and compressing, watermarking and steganography, speech and audio processing, mathematical methods, and machine learning techniques in data science, such as multiscale geometrical analysis, manifold learning and regularization, optimization method for big data processing, and deep learning.