Abstract

The extraction and proper utilization of convolution neural network (CNN) features have a significant impact on the performance of image super-resolution (SR). Although CNN features contain both the spatial and channel information, current deep techniques on SR often suffer to maximize performance due to using either the spatial or channel information. Moreover, they integrate such information within a deep or wide network rather than exploiting all the available features, eventually resulting in high computational complexity. To address these issues, we present a binarized feature fusion (BFF) structure that utilizes the extracted features from residual groups (RG) in an effective way. Each residual group (RG) consists of multiple hybrid residual attention blocks (HRAB) that effectively integrates the multiscale feature extraction module and channel attention mechanism in a single block. Furthermore, we use dilated convolutions with different dilation factors to extract multiscale features. We also propose to adopt global, short and long skip connections and residual groups (RG) structure to ease the flow of information without losing important features details. In the paper, we call this overall network architecture as hybrid residual attention network (HRAN). In the experiment, we have observed the efficacy of our method against the state-of-the-art methods for both the quantitative and qualitative comparisons.

1. Introduction

In this paper, we address the Single Image Super Resolution (SISR) problem, where the objective is to reconstruct the accurate high-resolution (HR) image from a single low-resolution (LR) image. It is known as an ill-posed problem, since there are multiple solutions available for mapping any LR image to HR images. This problem is intensified when the up-sampling factor becomes larger. Because HR images preserve much richer information than LR images, SISR techniques are popular in many practical applications, such as surveillance [43], Face Hallucination [35], Hyperspectral imaging [14], medical imaging [26] etc.

Numerous deep learning based methods have been proposed in recent years to address the SISR problem. Among them, SRCNN [3] is considered as the first attempt to come up with a deep-learning based solution with its three convolution layers. SRCNN outperformed the existing SISR approaches that typically used either multiple images with different scaling factors and/or handcrafted features. Later, Kim et al. [4] proposed an architecture named VDSR that extended the depth of CNN up to twenty layers while adding a global residual connection within the architecture. DRCN [11] also increased the depth of network through a recursive supervision and skip connection, and improved the performance. However, due to increasing depth of the networks, vanishing gradient resisted the network to be converged [7]. In the image classification domain, to solve the aforementioned problem, He et al. [7] proposed a residual block by which a network over 1000 layers was successfully trained. Inspired by its very deep architecture with residual blocks, EDSR [17] proposed much wider and deeper networks for the SISR problem using residual blocks, called EDSR and MDSR [17], respectively.

Very recently, Zhang et al. [40] proposed RCAN that utilizes a channel attention block to exploit the interdependencies across the feature channels. Moreover, Li et al. [16] proposed MSRN that improved the reconstruction performance by exploiting the information of spatial features rather than increasing the depth of CNNs. MSRN combines the features extracted from different convolution filter sizes and concatenates the outputs of all residual blocks through a hierarchical feature fusion (HFF) technique, utilizing the information of the intermediate feature maps. By doing so, MSRN achieved comparable performance against EDSR [17] although having a 7-times smaller model size. In [42], Zhang et al. proposed DCSR in which they proposed a mixed convolution block that combines dilated convolution layers and conventional convolution layers to attain larger receptive field sizes. Nonetheless, most of these CNN-based methods focused either on increasing the number of layers [10, 11, 17, 40] or on extending the width and height in a layer of CNN to achieve higher performance [16]. In this way, they put less focus on exploiting the by-product CNN features, e.g., spatial and channel information, simultaneously, and thus suffer to maximize the performance at times.

Moreover, the strong correlations between the input LR and output HR images [16] lead us to making an assumption
that, apart from the high-level features, the both low-level and mid-level features also play vital roles for reconstructing an super-resolution (SR) image. Therefore, we argue that, they should be treated precisely in this paper.

In the previous work, dense connections were used \cite{32}, which added every feature to subsequent features with residual connections. As a variant of dense connections, hybrid feature fusion (HFF) \cite{32, 41, 16} was proposed to remove the trivial residual connections and to directly concatenate all the output features from the residual blocks for the SISR problem. However, this direct feature concatenation prohibit the features from smooth feature transformation from low to high levels, resulting in resulting in improper utilization of various low-level and mid-level features. This may introduce redundancy in feature utilization, thus increasing the cost of computation complexity. In our ablation study in Section 4.1, this problem will be verified.

To solve this problem, in this paper, we propose a binarized feature fusion (BFF) scheme that combines adjacent feature maps with $1 \times 1$ convolutions which are repeatedly performed until remaining a single feature map. This allows all the features extracted from the CNN to be integrated smoothly, thus fully utilizing various features with different levels. Moreover, to efficiently extract the features, unlike previous work that used main residual blocks, we propose to use residual groups (RG) that constructs with the proposed hybrid residual attention block (HRAB). Our proposed HRAB extracts both the spatial and channel information with the notion that the both information is important in the reconstruction of high quality SR images and should be extracted simultaneously in a single module.

Moreover, compared to MSRN \cite{16} that concatenates the conventional convolution layers with different kernel sizes to enlarge receptive field sizes, proposed method concatenates dilated convolution layers with different dilation factors exploiting much larger receptive fields while significantly decreasing the number of parameters, i.e., convolution weights. Furthermore, to ease the flow of information, we introduce the short, long and global skip connections. We conduct comprehensive experiments to verify the efficacy of our method, where we observe its superiority against other state-of-the-art methods.

We summarize the overall contributions of this work as,

- We propose a BFF to transfer all the images features smoothly by the end of the network. This structure allows the network to smoothly transform the features with different levels and generate an effective feature map in the final reconstruction stage.

- We propose a hybrid residual attention block (HRAB) that considers both channel and spatial attention mechanisms to exploit the channel and spatial dependencies. The spatial attention mechanism extracts the fine spatial features with larger receptive field sizes whereas the channel attention guides in selecting the most important feature channels thus in the end, we have more discriminative features.

- Other than previous works, we employ BFF on residual groups (RG) rather than residual blocks (HRAB)

- For extracting the multiscale spatial features, we propose to use a mixed dilated convolution block with different dilation factors. Compared to the previous work in \cite{16} that used the large kernel sizes to secure large receptive fields, our proposed method can achieve a similar performance even with smaller kernel sizes. Moreover, we propose to use the dilated convolution in an effective manner to avoid the gridding problem of the conventional dilated convolution layers.

- To ease the transmission of information through out the network, we propose to adopt the global, short and long skip in our architecture.

2. Related work

There are several CNN-based SISR methods that have been proposed in the recent past. Previously, in the preprocessing step, researchers tend to use an interpolated LR image as an input that is interpolated to desired output image size which enables the network to have the same size of input and output images. In contrast, due to the additional computation complexity of interpolation, current work emphasizes to directly reconstruct HR image from LR image without interpolation.

In 2014, Dong et al. \cite{3} proposed SRCNN, the first CNN network architecture in the SR domain. It was a shallow 3 layers CNN architecture which achieved the superior performance against the previous non-CNN methods. Later, He et al. \cite{7} proposed a residual learning technique, and then Kim et al. \cite{10, 11} achieved remarkable performance with their proposed VDSR and DRCN methods. VDSR used the deep (20 layers) CNN and global residual connection whereas DRCN \cite{11} used a recursive block to increase the depth that does not require new parameters for repetitive blocks. Tai et al. \cite{28} proposed the MemNet which had memory blocks that consist of recursive and gate units. All of these methods have used the interpolated LR image as input. Due to this preprocessing, these methods add additional computation complexity along with artifacts, as also described in \cite{25}.

On the other end, the recent state-of-the-art methods directly learn the mapping from input LR image. Dong et al. \cite{4} proposed the FSRCNN, an improved version of SR-CNN, having faster training and inference time. Ledig et al. \cite{15} proposed the SRResNet, inspired from ResNet \cite{7}, to construct the deeper network. With the perceptual loss function in GAN, they proposed the SRGAN for photorealistic SR. Lim et al. \cite{17} removed the trivial modules (like batch normalization) of SRResNet, and proposed the EDSR (wider) and MDSR (deeper) that made a significant improvement in SR problem. EDSR has a large number of filters (256) whereas MDSR has a small number of filters though the depth of CNN network is increased to about 165 layers. It also won the first NTIRE SR challenge \cite{30}. It has shown that deeper networks can achieve remarkable...
performance. Consequently, Zhang et al. [40] proposed a very deep network for SR. To the extent of our knowledge, it has the largest depth in the SR domain. RCAN [40] has shown that only stacking the layers cannot improve the performance. It proposed to use the channel attention (CA) [8] mechanism to neglect the low-frequency information while selecting the valuable high-frequency feature maps. To increase the depth of the network, it proposed the residual in residual (RIR) structure. Nevertheless, RCAN [40] network is very deep and makes it difficult to use it in real-life applications due to higher inference time.

In contrast, multiscale feature extraction technique, which is less explored in SISR, has shown significant performance in object detection, [18] image segmentation, [24] and model compression [2] to achieve good tradeoffs between speed and accuracy. Li et al. proposed a multiscale residual network (MSRN) [16] having just 8 residual blocks. It used multipath convolution layers with different kernel sizes (3×3 and 5×5) to extract the multiscale spatial features. Furthermore, it proposed to use the hierarchical feature fusion (HFF) architecture to utilize the intermediate features. The intuition behind HFF architecture is to transfer the middle features at the end of the network since an increase in the depth of the network may cause the features to vanish in between the network. HFF shows comparable performance to EDSR, nevertheless its accuracy is limited. In addition, as the depth or width of a network increases, HFF also increases the computation complexity.

Therefore, we need an efficient multiscale superresolution CNN which could fully utilize the feature information as well as channel information. Considering it, we propose a hybrid residual attention network (HRAN) which combines the multiscale feature extraction along with the channel attention [8] mechanism. In this paper, we refer the multiscale feature extraction as spatial attention. Thus, the combination of the channel and spatial attention is called hybrid attention. We discuss the details of HRAN in the next section.

3. Hybrid Residual Attention Network

3.1. Network architecture

The proposed HRAN architecture is shown in Figure 1. The HRAN can be decomposed into two parts: feature extraction and reconstruction. The feature extraction is divided into two parts: shallow feature extraction and deep feature extraction. The deep feature extraction step further includes residual groups (RG) with binarized feature fusion (BFF) structure. Whereas, RG contains a sequence of hybrid residual attention blocks (HRAB) followed by 3×3 convolution. We represent the input and output of HRAN as I_{LR} and I_{SR} respectively. We aim to reconstruct the accurate HR image I_{HR} directly from LR image I_{LR}.

In the shallow feature extraction, we use two convolution layers to extract the features from input I_{LR} image.

\[ F_0 = H_{SF1}(I_{LR}), \]

Here \( H_{SF1}(\cdot) \) represents the convolution operation. \( F_0 \) is also used for global residual learning to preserve the input features. As mentioned above, we pass the \( F_0 \) for further feature extraction

\[ F_1 = H_{SF2}(F_0), \]

where \( H_{SF2}(\cdot) \) represents the convolution operation. \( F_1 \) is the output of shallow feature extraction step and will be used as input for the deep feature extraction.

\[ F_{DF} = H_{DF}(F_1) + F_0, \]
Another difference in this architecture is the feature extraction from adjacent RG blocks. First, we concatenate the adjacent RG blocks and then, we remove the redundant information from adjacent blocks using a 1x1 convolution. We repeat this procedure for all RG blocks and the resultant blocks produced through this mechanism until all the blocks integrate into a single RG block, which is convolved by a 1x1 to produce the output features. In the end, we element-wise add this output to the shallow features’ output (F0). We refer this element-wise summation as global skip connection in Figure 1.

\[ F_{i+1} = H_{RG}(F_i), \]

where \( H_{RG}(\cdot) \) represents the features extracted through single RG block whereas \( F_i \) shows the \( i^{th} \) extracted feature map. We explain the details of RG in the next section. When we extract all the features through RG blocks, then we can utilize these RG blocks with HFF architecture.

\[ M_j = H_{1 \times 1}[F_{i+1}, F_{i+2}], \]
\[ M_{j+1} = H_{1 \times 1}[F_{i+3}, F_{i+4}], \]

Here, the output of two adjacent RG blocks are channel-wise concatenated and then passed into the 1x1 convolution layer to avoid the redundant information from them. Thus, the four RG blocks produce two more blocks which are then processed in a similar manner such that \( F_{i+1} = M_j \) and \( F_{i+2} = M_{j+1} \). Thus, in the next step, \( M_j \) and \( M_{j+1} \) will act as two RG blocks. We repeat this procedure until we integrate all the RGs and resultant blocks into a single output which is further used in the input of reconstruction step.

3.3. Residual Groups (RG)

It is shown in [17] that the stacked residual blocks enhance the performances of SR but after some extent, cause crucial information loss during transmission and also makes the training slower, affecting the performance gain in the SISR [40]. Thus, rather than increasing the depth, we propose to use the residual groups (RG) (see shaded area of Figure 1) in our architecture to detect deep features. The RG consists of multiple HRABs that are followed by 1x1 convolution. We find that adding many HRABs degrade the SR performance. Thus, to preserve the information, we apply element-wise summation between the input of RG and output of 1x1 convolution and refer it as long skip connection (LSC).

The RG enables the network to remember the information through LSC whereas to detect deep features, it uses SSC within its modules, in this case, HRAB. Hence, the flow of information in RG is smoothly carried out through LSC and SSC. The details of the HRAB are mentioned in the next section.

Thus, we express the single RG block as

\[ H_{RG} = W_{RG} * H_n(H_{n-1} \cdots H_1(F_1) \cdots), \]
Here \( H_i \) represents the ‘B’ hybrid residual attention blocks (HRAB), which takes input features from previous RG block \((F_i)\) and produces the output \((F_{i+1})\). After stacking the ‘B’ HRAB modules, we apply \(3 \times 3\) convolutions with weights \(W_{RG}\). After applying LSC, the equation 10 can be rewritten as

\[
H_{RG} = W_{RG} \ast H_n \left( H_{n-1} \left( \cdots H_1 \left( F_1 \right) \cdots \right) \right) + F_1, \quad (11)
\]

The above equation represents the first RG block because it takes the shallow features \(F_1\) as input. Since, we have multiple RG blocks to extract the deep features, hence, the above equation can be generally written as

\[
H_{RG}^i = W_{RG}^i \ast H_{RG}^{i-1} + H_{RG}^{i-1} \quad \quad (12)
\]

Here \( i = 1, 2, \cdots, R \). We have ‘R’ RG blocks and each RG block uses the output of the previous block as its input except the first RG block that uses the shallow features \(F_1\) as input. Thus, for the first RG block, \(H_{RG}^1 = F_1\).

### 3.4. Hybrid Residual Attention Block (HRAB)

In this section, we propose a multiscale multipath residual attention block for the feature extraction, called hybrid residual attention block (HRAB) (see Figure 2). Our HRAB integrates both the spatial attention (SA) and channel attention (CA) mechanisms, thus, it has two separate paths for the SA and CA.

\[
H_{HRAB} (F_{i+1}) = H_{SA} (F_i) \cdot H_{CA} (F_i) \quad (13)
\]

where \( H_{SA} \) and \( H_{CA} \) denote the functions of spatial attention (SA) and channel attention (CA) respectively. Here ‘\( \cdot \)’ represents the element-wise multiplication between SA and CA functions. Unlike RCAN [40], we propose to use element-wise multiplication of the outputs of SA and CA to extract the most informative spatial features. Like RCAN [40], we also add the short skip connections (SSC) to ease the flow of information through the network.

### 3.4.1 Spatial Attention (SA)

MSRN [16] proves that multiscale features improve the performance with lesser residual blocks. In MSRN [16], authors use the multiple CNN filters with increasing kernel sizes \((3 \times 3 \text{ and } 5 \times 5)\) to extract multiscale features. The intuition behind the larger kernel size is to take advantage of large receptive fields. But, the large kernel size causes to increase the memory computation. Thus, we propose to use the dilated convolution layers with different dilation factors which can have the same receptive fields as large kernel size and memory consumption is similar to smaller kernel size. But, only stacking the dilated convolution layers produce gridding effect [36]. To avoid this problem, as illustrated in Figure 2, we propose to use the element-wise sum operation between the dilated convolutions with different factors before the concatenation operation. Suppose \(F_{i-1}\) is the input of SA then the output will be \(F_i\).

\[
S_1 = \text{LeakyReLU} \left( H_{DC1} (F_i) \right) \quad (14)
\]
\[
S_2 = \text{LeakyReLU} \left( H_{DC2} (F_i) + S_1 \right) \quad (15)
\]
\[
S = [S_1, S_2] \quad (16)
\]
\[
S_2 = \text{LeakyReLU} \left( H_{DC1} (S) \right) \quad (17)
\]
\[
S_2 = \text{LeakyReLU} \left( H_{DC2} (S) + S_1 \right) \quad (18)
\]
\[
H_{SA} (F_{i+1}) = H_{1 \times 1} \ast [S_1, S_2] \quad (19)
\]

where \( H_{DC1} \) and \( H_{DC2} \) denotes the convolution layers with dilation factors 1 and 2 respectively. First, we concatenate the output of two convolution layers to increase the channel size and at the end, we use \(1 \times 1\) convolution to reduce the channels. Thus, our input and output have the same number of channels. Our SA architecture inspires from [6] which has shown that upsampling and downsampling module within the architecture improves the accuracy in SR. For the activation unit, by following ref [12, 29], we
prefer the LeakyReLU over ReLU activation whereas we use the linear bottleneck layer as suggested in [23].

3.4.2 Channel Attention (CA)

The channel attention (CA) mechanism achieves a lot of success in image classification [8]. In SISR, RCAN [40] introduces the CA layer in the network. CA plays an important role in exploiting the interchannel dependencies because some of them have trivial information while others have the most valuable information. Therefore, we decide to use channel-wise features and incorporate the CA mechanism with SA module in our HRAB. Thus, by following [8, 40], we use the global pooling average to consider the channel-wise global information. We also experiment with global pooling variance as we thought global variance could extract more high-frequencies, in contrast, we get poor results as compared with global pooling average.

Suppose if we have C channels in the feature map \(x_1, x_2, \ldots, x_C\) then we can express each ‘c’ feature map as a single value.

\[
z_c(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j),
\]

(20)

Here, \(x_c\) is the spatial position \((i, j)\) of the feature maps.

To extract the channel-wise dependencies, we use the similar sigmoid gating mechanism as [8, 40]. Alike SA, here, we replace the ReLU with LeakyReLU activation.

\[
H_{CA}(F_{i+1}) = f(W_U LR(W_D z)),
\]

(21)

Here \(LR(\cdot)\) and \(f(\cdot)\) represent the LeakyReLU and sigmoid gating function respectively whereas \(W_D\) and \(W_U\) respectively denote the weights of downsampling and upsampling convolutions. It is noted that it is channel-wise downsampling and upsampling with reduction ratio \(r\).

Table 1. Investigation of HRAB module (with and without CA).

| Modules                              | SSIM / PSNR |
|--------------------------------------|-------------|
| Spatial attention without channel attention (SA) | 32.77 / 0.9343 |
| Spatial attention with channel attention (Hybrid Attention) | 32.95 / 0.9357 |

3.5. Implementation details

For training the HRAN network, we employ 4 RG blocks in our main architecture and in each RG block, there are 8 HRAB modules which are followed by \(3 \times 3\) convolution. For the dilated convolution layers, we use the \(3 \times 3\) convolution with dilation factor 1 and 2. We use \(C = 64\) filters in all the layers except the final layer which has 3 filters to produce a color image though our network can work for both grayscale and color images. For the channel-downscaling in CA mechanism, we set a reduction factor \(r = 4\).

4. Experimental Results

In this section, we explain the experimental analysis of our method. For this purpose, we use several public datasets that are considered as the benchmark in SISR. We provide the results of both the quantitative and qualitative experiments for the comparison of our method with several state-of-the-art networks. For the datasets, we follow the recent trends [17, 39, 41, 16, 30] and use DIV2K dataset as the training set, since it contains the high-resolution images. For testing, we choose widely used standard datasets: Set5 [1], Set14[37], BDS100 [19], Urban100 [9] and Manga109 [20]. For the degradation, we use the Bicubic Interpolation (BI).

We evaluate our results with peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [33] on luminance channel i.e., Y of the YCbCr space and we remove P-pixels from each border (P refers to upscaling factor). We provide the results for scaling factor \(2 \times, 3 \times, 4 \times, \) and \(8 \times\).

For the training settings, we follow the settings in [16]. We extract 16 LR patches randomly in each training batch with the size of \(64 \times 64\). We use ADAM optimizer with learning rate \(lr = 10^{-4}\) which decreases to half after every \(2 \times 10^5\) iterations of back-propagation. We use PyTorch framework to implement our models with GeForce RTX 2080 Ti GPU.

4.1. Ablation studies

We conduct a series of ablation studies to show the effectiveness of our model. In the first experiment, we train our model with and without CA and compare their performance with our HRAB module. For the training, we use Urban100 dataset[9] as it consists of large dataset. The results are shown in Table 1. We observe that our SA module alone achieves 32.77 dB PSNR. We also experimented on CA module alone though results were unsatisfactory. Whereas, when we combine SA with CA, i.e., our HRAB module, it achieves the 32.95 dB PSNR. This study suggests we need HRAB module containing both the spatial and channel attention for accurate SR results. We also investigate about our BFF structure using HRAB module and tested the both BFF and HFF on MSRN [16] and our proposed HRAN model to verify the effectiveness of BFF on both models. It is evident from the results that BFF structure improves the PSNR of MSRN [16] from 32.22 dB to 32.44 dB with BFF. Moreover, proposed HRAN and BFF together significantly increase the accuracy which show the

Table 2. BFF vs HFF structures. We examine the best PSNR (dB) on Urban100 (2×) with same training settings.

| Method                                           | PSNR / SSIM |
|--------------------------------------------------|-------------|
| MSRN with HFF [15]                              | 32.22 / 0.9326 |
| MSRN [15] with proposed BFF                     | 32.44 / 0.9315 |
| Our HRAN with HFF                               | 32.69 / 0.9334 |
| Our HRAN with proposed BFF                      | 32.95 / 0.9375 |
Table 3. Quantitative Comparisons of state-of-the-art methods for BI degradation model. Best, 2nd best and 3rd best results are respectively shown with Magenta, Blue, and Green colors.

| Method        | Scale | Set5  | Set14  | B100   | Urban100 | Manga100 |
|---------------|-------|-------|--------|--------|----------|----------|
| Bicubic       | ×2    | 33.66 | 0.9299 | 30.24  | 0.8688   | 29.56    | 0.8431   | 26.88 | 0.8403 | 30.80 | 0.9339 |
| SRCNN [3]     | ×2    | 36.66 | 0.9542 | 32.45  | 0.9067   | 31.36    | 0.8879   | 29.50 | 0.8946 | 35.60 | 0.9663 |
| FSRCNN [4]    | ×2    | 37.05 | 0.9560 | 32.66  | 0.9900   | 31.53    | 0.8920   | 29.88 | 0.9020 | 36.67 | 0.9710 |
| VDSR [10]     | ×2    | 37.53 | 0.9590 | 33.05  | 0.9130   | 31.90    | 0.8960   | 30.77 | 0.9140 | 37.22 | 0.9750 |
| LapSRN [12]   | ×2    | 37.52 | 0.9591 | 33.08  | 0.9130   | 31.08    | 0.8950   | 30.41 | 0.9101 | 37.27 | 0.9740 |
| MemNet [28]   | ×2    | 37.78 | 0.9579 | 33.28  | 0.9142   | 32.08    | 0.8978   | 31.31 | 0.9195 | 37.72 | 0.9740 |
| EDSR [17]     | ×2    | 38.11 | 0.9602 | 33.92  | 0.9195   | 32.32    | 0.9013   | 32.93 | 0.9351 | -/-   | -/-    |
| SRMDNF [39]   | ×2    | 37.79 | 0.9601 | 33.32  | 0.9195   | 32.05    | 0.8985   | 31.33 | 0.9204 | 38.07 | 0.9761 |
| RDN [41]      | ×2    | 38.24 | 0.9614 | 34.01  | 0.9212   | 32.34    | 0.9017   | 32.89 | 0.9353 | 39.18 | 0.9780 |
| DCSR [42]     | ×2    | 37.54 | 0.9587 | 33.14  | 0.9141   | 31.90    | 0.8959   | 30.76 | 0.9142 | -/-   | -/-    |
| MSRN [16]     | ×2    | 38.08 | 0.9605 | 33.74  | 0.9170   | 32.23    | 0.9013   | 32.22 | 0.9326 | 38.82 | 0.9868 |
| HRAN (ours)   | ×2    | 38.21 | 0.9613 | 33.85  | 0.9200   | 32.34    | 0.9016   | 32.95 | 0.9357 | 39.12 | 0.9780 |
| HRAN+ (ours)  | ×2    | 38.25 | 0.9614 | 33.99  | 0.9211   | 32.38    | 0.9020   | 33.12 | 0.9370 | 39.29 | 0.9785 |

---

effectiveness of our BFF structure.

4.2 Comparison with State-of-the-art Methods

We compare our method with 10-state-of-the-art SISR methods: SRCNN [3], FSRCNN[4], VDSR [10], LapSRN [12], MemNet [28], EDSR [17], SRMDNF [39], RDN [41], DCSR [42] and MSRN [16]. By following [17, 31], we also use self-ensemble strategy to improve the accuracy of our model at test time. We show our quantitative evaluation results in Table 3 for the scale factor of ×2, ×3, ×4, and ×8. It is evident from the results that our method outperforms most of the previous methods. Our self-ensemble model achieves the highest PSNR amongst all the models. Although RDN [41] has shown slightly better performance, from Figure 4, we observe that RDN [41] has about 22M parameters, in contrast, our HRAN model has only 7.94 M parameters though our model shows comparable performance. Instead of increasing the depth and dense connections, our HRAN model with HRAB and BFF detect the deep features without increas-
the depth of the network. Hence, this observation indicates that we can improve the network performance with HRAB and RG along with BFF without increasing the network depth. This also suggests that our network can further improve the accuracy with more HRAB’s and RG’s, though, we aim to achieve the greater accuracy by considering the memory computations.

Moreover, we present the qualitative results in Figure 3. The results of other methods are derived from [40]. In Figure 3, it can be observed from ‘img004’ image our HRAN method recovers the lattices in more details, meanwhile, other methods experience the blurring artifacts. Similar behavior is also observed in ‘Yumeiro-Cooking’ image where other methods produce blurry lines and our HRAN produces the lines similar to HR image. It shows that our model reconstructs the fine details in output SR image through extracted deep features with RGs which are then efficiently utilized by BFF.

4.3. Model Complexity Analysis

Since, we are targeting the maximum accuracy with limited memory computation, therefore our performance is best visible when we see the Table.3 along with Figure. 4. In Figure. 4, we compare our model size and its performance on Set5 [1] (×4). As we can observe that our HRAN model has fewer parameters compared to RDN [41] and EDSR [17], it still achieves the comparable performance whereas our HRAN+ outperforms the state-of-the-art methods. We have also shown analysis on much larger scale (×8) in supplementary materials. These results demonstrate the effective utilization of the features that result in performance gain in SISR.

5. Conclusions

In this paper, we propose a hybrid residual attention network (HRAN) to detect the most informative multiscale spatial features for the accurate SR image. Proposed hybrid residual attention block (HRAB) module fully utilize the high-frequency information from input features with a combination of the spatial attention (SA) and channel attention (CA). In addition, the binarized feature fusion (BFF) structure allows us to smoothly transmit all the features at the end of the network for reconstruction. Furthermore, we propose to adopt the global, short and long skip connection and residual groups (RG) to ease the flow of information. Our comprehensive experiments show the efficacy of the proposed model.
References

[1] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In British Machine Vision Conference (BMVC), pages 1–10, 2012. 6, 8
[2] C.-F. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris. Big-little net: An efficient multi-scale feature representation for visual and speech recognition. arXiv preprint arXiv:1807.03848, 2018. 3
[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2):295–307, 2016. 1, 2, 7, 8, 11, 12
[4] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In European conference on computer vision, pages 391–407. Springer, 2016. 1, 2, 4, 7, 8, 11, 12
[5] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. Proc. of ICLR, 2, 2017. 4
[6] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 5
[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1, 2
[8] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018. 3, 6
[9] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5197–5206, 2015. 6
[10] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral), June 2016. 1, 2, 4, 7, 8, 11, 12
[11] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral), June 2016. 1, 2
[12] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5, 7, 8, 11, 12
[13] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. arXiv:1710.01992, 2017. 11, 12
[14] C. Lanaras, E. Ballisavias, and K. Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In Proceedings of the IEEE International Conference on Computer Vision, pages 3586–3594, 2015. 1
[15] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681–4690, 2017. 2
[16] J. Li, F. Fang, K. Mei, and G. Zhang. Multi-scale residual network for image super-resolution. In The European Conference on Computer Vision (ECCV), September 2018. 1, 2, 3, 4, 5, 6, 7
[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017. 1, 2, 4, 6, 7, 8, 11, 12
[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2117–2125, 2017. 3
[19] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In null, page 416. IEEE, 2001. 6
[20] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yasamaki, and K. Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 76(20):21811–21838, 2017. 6
[21] T. Peleg and M. Elad. A statistical prediction model based on sparse representations for single image super-resolution. IEEE transactions on image processing, 23(6):2569–2582, 2014. 11
[22] M. S. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, pages 4491–4500, 2017. 11, 12
[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4510–4520, 2018. 6
[24] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets. Feature pyramid network for multi-class land segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018. 3
[25] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1874–1883, 2016. 2, 4
[26] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhattacharya, A. M. S. M. de Marvao, T. Dawes, D. ORegan, and D. Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 9–16. Springer, 2013. 1
[27] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 11, 12
[28] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In Proceedings of International Conference on Computer Vision, 2017. 2, 7, 8, 11, 12
[29] W. Tan, B. Yan, and B. Bare. Feature super-resolution: Make machine see more clearly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3994–4002, 2018. 5
[30] 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 Proceedings of the IEEE
[31] R. Timofte, R. Rothe, and L. Van Gool. Seven ways to improve example-based single image super resolution. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[32] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, pages 4799–4807, 2017.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[34] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In Proceedings of the IEEE International Conference on Computer Vision, pages 370–378, 2015.

[35] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. IEEE transactions on image processing, 19(11):2861–2873, 2010.

[36] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 472–480, 2017.

[37] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In International conference on curves and surfaces, pages 711–730. Springer, 2010.

[38] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3929–3938, 2017.

[39] K. Zhang, W. Zuo, and L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3262–3271, 2018.

[40] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 286–301, 2018.

[41] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In CVPR, 2018.

[42] Z. Zhang, X. Wang, and C. Jung. Dsrr: Dilated convolutions for single image super-resolution. IEEE Transactions on Image Processing, 28(4):1625–1635, 2019.

[43] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. IEEE Transactions on Image Processing, 21(1):327–340, 2012.
Supplementary Material

In this supplementary submission, we present more qualitative results with the different scaling factors. Furthermore, we also compare our method’s computation complexity with the large scaling factor.

1. Experimental Results

1.1. Model Complexity Analysis

When it comes to the large scaling factor, the reconstruction of the SR image becomes more difficult and the SR problem is intensified due to very limited information in the LR image. In this section, we compare our model computation complexity and performance on the large scaling factor (8×) in terms of a number of parameters and peak signal-to-noise ratio (PSNR) respectively in Figure 5. The results in Figure 5 shows that our HRAN and HRAN+ models outperform all the models including EDSR, and MSRN, for the scaling factor 8× with the low number of parameters.

1.2. Visual Comparisons

In this section, we compare our qualitative results with the state-of-the-art methods: SRCNN [3], SPMSR [21], FSRCNN [4], VDSR [10], IRCNN [38], SRMDNF [39], SCN [34], DRRN [27], LapSRN [12], MSLapSRN [13], Enet-PAT [22], MemNet [28], and EDSR [17].

We show the experiments’ results with the different scaling factors: 3×, 4×, and 8×. In Figure 6, we can visualize that most of the methods fail to reconstruct the fine details in ‘img_062’ and ‘img_078’ and have blurry effects. Although SRMDNF has recovered the horizontal and vertical lines but output result is more blurry. Whereas, our results have no blurry effect and have shown similar visual performance than EDSR.

For further illustrations, we also analyze our results on 8× super-resolution (SR) in Figure 7. When the scaling factor increases, we get very limited details in the LR image. From the ‘img_040’ image, what we observe that Bicubic interpolation does not recover the original patterns. Those methods (SRCNN, MemNet, and VDSR) which use interpolation as pre-scaling, lose the original structure and generate wrong patterns. Our HRAN results are more similar to EDSR, but unlike EDSR, HRAN does not produce blurry effects. Similarly, in ‘TaiyouNiSmash’ image, we observe that most of the methods could not recover the tiny lines clearly and lose the structures and the blurry effect is also evident in most of the methods.

Figure 5. Comparison of memory and performance. Results are evaluated on Set5 (×8).
Figure 6. “img_074” from Urban100 (4×): State-of-the-art results with Bicubic (BI) degradation.

Figure 7. Qualitative results for 8× SR with BI model on Manga109 and Urban100.