Attention Retractable Frequency Fusion Transformer for Image Super Resolution

Qiang Zhu
Pengfei Li, Qianhui Li
School of Information and Communication Engineering, University of Electronic Science and Technology of China
zhuqiang@std.uestc.edu.cn, lipengfei_uestc@163.com, 2020010902028@std.uestc.edu.cn

Abstract
Transformer-based image super-resolution (SR) has offered promising performance gains over the convolutional neural network-based one due to the adoption of parameter-independent global interactions. However, the existing Transformer-based methods are limited to obtaining enough global information due to the use of self-attention within non-overlapping windows, which restricts the receptive fields. To address this issue, we construct an effective image SR model based on the attention retractable frequency Transformer with the proposed spatial-frequency fusion block. In our method, the spatial-frequency fusion block is designed to strengthen the representation ability of the Transformer and extend the receptive field to the whole image to improve the quality of SR results. Furthermore, a progressive training strategy is proposed to use image patches with different sizes to train our SR model to further improve the SR performance. The experimental results demonstrate that our proposed method outperforms the state-of-the-art methods over various benchmark datasets, both objectively and subjectively.

1. Introduction
Image super-resolution (SR) aims to compose the high-resolution (HR) image from the low-resolution (LR) counterpart. Recently, the convolutional neural network (CNN) has been investigated to design various image SR models [1–3]. Super resolution CNN (SRCNN) [1] firstly introduced CNN into image SR. Then, several methods utilized residual learning, e.g., Enhanced deep residual networks (EDSR) [2], and attention mechanism, e.g., residual channel attention networks (RCAN) [3], to compose very deep networks for image SR. These CNN-based networks have achieved remarkable performance. However, due to adopted parameter-dependent receptive field scaling and content-independent local interactions of convolutions, CNN is limited to model the long-range dependencies [6].

To break this limitation, some Transformer-based image SR networks were proposed [4, 6, 8, 10] by modeling the long-range dependencies to improve SR performance. For example, the image processing Transformer (IPT) [4] was designed to be pre-trained on ImageNet [5] to maximally excavate the performance of the Transformer so as to achieve high SR performance. SwinIR [6] was proposed based on the Swin Transformer [7] to significantly improve the SR performance. In addition, an attention retractable Transformer (ART) [10] was developed based on SwinIR with an attention retractable module and achieved state-of-the-art results on the image SR task.

Although the Transformer-based image SR methods achieve impressive performance, they still suffer from a defect. For example, IPT [4] uses dense attention with short token sequences from a dense area of the image, which causes a restricted receptive field. In addition, SwinIR [6] adopts the window-based and local attention strategy to construct a model, which restricts employing large receptive fields to capture global information. ART [10] noted this defect and design the attention retractable module based on sparse attention. But the accessible receptive field of ART also is limited due to only using the four as interval size in sparse attention block in SR task while the larger interval size easily causes worse performance.

To solve the problem of ART, we design a spatial-frequency fusion block (SFFB) based on Fast Fourier Transform (FFT) to enlarge the receptive field in the frequency domain, which accordingly composes our proposed attention retractable frequency fusion Transformer (ARFFT) for image SR. The architecture of our ARFFT is illustrated in Fig. 1. It is developed based on ART in which two self-attention blocks are adopted. The first block is the dense attention block (DAB) and the second block is the sparse attention block (SAB). With these two blocks, both the local and the non-local receptive fields are captured. To extend the receptive field to the whole image, we design the spatial-
frequency fusion block (SFFB) for ART, targeting better SR performance. In addition, to further improve the SR performance of our model, we proposed a progressive training strategy to use different-size patches to progressively train our SR model to achieve promising SR results.

2. Related work

2.1. Vision Transformer

The application of Transformer to machine translation [13] has achieved impressive performance. In addition, Transformer has also been applied to the computer vision task. For example, ViT [14] was proposed using Transformer to project large image patches into token sequences to achieve image recognition task. Glance and Gaze Transformer [15] was proposed to design the Glance and Gaze branches to efficiently model both long-range dependencies and local context for some high-level vision tasks. Multi-axis vision Transformer [16] was developed using the multi-axis attention based on blocked local and dilated global attention to achieve the SOTA performance on image classification.

In addition to the high-level vision tasks, Transformer was also applied to the low-level vision tasks [4, 6, 8–12]. For instance, IPT [4] was designed using a pre-trained Transformer to achieve high SR performance. SwinIR [6] was proposed based on the Swin Transformer [7] to achieve a strong image restoration baseline. Restormer [8] was developed by making several key blocks based on the Transformer structure such that it can capture long-range pixel interactions. UFormer [9] introduced a novel locally-enhanced window Transformer block to significantly reduce the computational complexity of the high-resolution feature. Besides, a learnable multi-scale restoration modulator was proposed in UFormer to adjust features in multiple layers of the decoder so as to have a high capability for capturing both local and global dependencies for image restoration task. In addition, an attention retractable Transformer (ART) [10] was developed using an attention retractable module to enlarge the receptive field for improving SR performance. Cross aggregation Transformer (CAT) [11] designed a rectangle-window self-attention to aggregate features to obtain a large receptive field. Besides, CAT developed a locality complementary module to realize the coupling of global and local information for improving image restoration performance. Hybrid attention Transformer (HAT) [12] combined both channel attention and window-based self-attention to utilize global statistics and strong local fitting capability. Moreover, an overlapping cross-attention module was designed to better aggregate the cross-window information for enhancing the interaction of features. HAT was constructed with these attentions and achieved state-of-the-art results on the image SR task.

2.2. Frequency Learning

Lot of works were studied based on frequency domain in low-level restoration tasks [20–25]. Some of these methods [20–22] studied to decompose features into different frequency bands by multi-branch CNN to enhance the details. Typically, omni-frequency region-adaptive network [20] used multi-branch CNN to separate different frequency components and enhances these features with the proposed frequency enhancement unit. Frequency-dependent convolutional neural networks [21] divided the input images into three sub-frequency groups and trained the convolutional neural network for each sub-frequency group. The final SR image was constructed by combining the multi-SR images from multiple networks. Besides, frequency aggregation network [22] extracted different frequencies of the LR image and pass them to a channel attention-grouped residual dense network individually to output corresponding features. Then aggregating these residual dense features adaptively to recover the HR image with enhanced details and textures. The other methods [23–25] transformed images into frequency domain. For example, D⁢³ [23] designs a dual-domain restoration network to remove artifacts of compressed images. Wavelet-based dual recursive network [24] was proposed to decompose the LR image into a series of wavelet coefficients and predicted the corresponding series of HR wavelet coefficients using networks so as to construct the final HR image. SwinIR [25] extends SwinIR by replacing fast Fourier convolution to explore the image-wide receptive field for improving the SR performance.

3. Proposed Method

The architecture of our proposed ARFFT is illustrated in Fig. 1 (a). Given an LR image ⁎∈, where ⁢, ⁢, and ⁢ are the height, width, and number of color channels. Firstly, LR image is sent to a 3×3 convolution layer to obtain shallow feature ⁢ where ⁢ is the dimension size of the feature. Next, the shallow feature is normalized and fed into the ⁢ residual groups to generate the deep feature. Specifically, each residual group consists of ⁢ the combination block of DAB, SAB, and a SFFB. After that, the deep feature passes through another 3×3 convolution layer to get refined feature ⁢ Then shallow feature and the refined feature are added to obtain the final constructed feature ⁢ from the feature ⁢.

3.1. Retractable Attention

We apply two attention strategies, i.e., the dense multi-head self-attention module (D-MSA) and the dense multi-head self-attention module (S-MSA), to design two self-
attention blocks, i.e., DAB and SAB. The structure is illustrated in Fig. 1(b).

In DAB, D-MSA helps each token to interact with a smaller number of tokens from the neighborhood position of a non-overlapping $W \times W$ window. Meanwhile, in SAB, S-MSA allows each token to interact with the same number of tokens as D-MSA, but which are from sparse positions of a $W I \times W I$ window, where $I$ is interval size. ART [10] demonstrates that the application of these two blocks enables our model to capture local and non-local receptive fields simultaneously. The successive attention blocks are applied to provide interactions for both local dense tokens and non-local sparse tokens. However, increasing interval size $I$ is limited. In ART, the increased interval size easily causes worse SR performance, which impacts the model to access a larger receptive field for improving the SR performance.

### 3.2. Spatial-frequency Fusion Block

To explore the larger receptive field, we design SFFB to strengthen the representation ability of the Transformer and extend the receptive field to the whole image to improve the SR performance. As shown in Fig. 1(c), the SFFB network consists of two primary branches: a frequency branch and a spatial branch. We send input feature $X$ into these two branches to generate $X_{\text{frequency}}$ and $X_{\text{spatial}}$ respectively. We will respectively introduce two branches as follow.

In frequency branch, a frequency branch network $H_{\text{frequency}}$ is designed to obtain the frequency enhanced feature,

$$X_{\text{frequency}} = H_{\text{frequency}}(X).$$

The frequency branch network is illustrated in Fig. 1(c). Specifically, The $X$ is firstly refined using a convolution layer to obtain the initial feature $X_{\text{finit}}$ for the frequency transforming,

$$X_{\text{finit}} = C_L(X),$$

where $C_L$ denotes a 3×3 convolution layer with a LeakyReLU activate function. The $X_{\text{finit}}$ is transformed into the frequency domain using the 2-D Fast Fourier Transform (FFT) to extract the global information for generating high-quality frequency features. The inverse 2-D FFT operation is performed to transform the frequency feature into the spatial feature,

$$X_{\text{frequency}} = C_1(\hat{F}_T(C_L(F_T(C(X_{\text{finit}})) + X_{\text{finit}}))).$$
where \( C_1 \) denotes a \( 1 \times 1 \) convolution layer, \( F_T \) denotes a Fast Fourier Transform layer, \( \bar{F}_T \) denotes an inverse Fast Fourier Transform layer.

Besides, spatial information also needs to be explored. We use convolution layers and activate functions to construct the spatial branch to increase the expressiveness of the feature for obtaining the refined spatial feature. The \( X_{\text{spatial}} \) is represented as

\[
X_{\text{spatial}} = C(C_L(C(X))) + X. \tag{4}
\]

Based on the frequency branch and spatial branch, the output of the SFFB is denoted as

\[
X_{SFFB} = C_1([X_{\text{frequency}}, X_{\text{spatial}}]), \tag{5}
\]

where \([\cdot]\) denotes a concatenation operation.

### 3.3. Progressive Training Strategy

In general, the SR model is trained with only a patch size to achieve the highest performance on the validation set will be selected as the final one. However, in the test phase, the whole image is fed into the SR model to generate SR results. The inconsistent patch size in the training stage and test stage easily causes the final SR performance to decrease. We propose a novel progressive training strategy (PTS) used based on multi-training stages to improve SR performance. Specifically, the progressive training strategy utilized multi-training stages to gradually obtain the final SR results. Our SR model is trained with different patch sizes of training datasets in different training stages. The model of the previous stage is utilized to initiate the current model. We set three training stages and our SR model is gradually trained using the patch size of 48, 64, and 84, respectively so as to obtain the improved SR performance.

Different from Restormer [8], we use the PTS to train our SR model with the fixed batch size and fixed patch size at each stage, while Restormer only uses one stage to gradually reduce the patch size and increase the patch size to obtain the final SR model. Besides, in Restormer, the update points for changing the patch size and batch size pairs is difficult to set for specific SR model. The inaccurate update points can easily cause missing the best SR model to affect the final SR performance. Our PTS avoids this problem, the best model is selected in each training stage for initialization of the next training stage so as to effectively obtain the final SR results.

### 3.4. Loss Function

In addition to the structure of our network, the loss function also determines whether the model can achieve good results. In low-level visual tasks, such as denoising and deblurring, the \( L_1 \), \( L_2 \), and perceptual adversarial loss functions are often used to optimize neural networks. Recently, the Fast Fourier Transform loss (FFTLoss) [26] is proposed to focus on the frequency information of restoration results during the training network so as to get better performance in super-resolution tasks. In our method, we adopt the \( L_1 \) loss, the \( L_2 \) loss, and the FFTLoss [26] to train our proposed image SR model targeting high-quality results.

In each training stage of PTS, we firstly use the basic loss function \( \text{Loss}_1 \) composed by the \( L_1 \) loss and the FFTLoss to obtain the initial SR performance

\[
\text{Loss}_1 = \| I_{HR} - I_{SR} \|_1 + \alpha \text{FFTLoss}(I_{HR}, I_{SR}), \tag{6}
\]

where \( I_{HR} \) is the corresponding HR image and \( \alpha \) is the penalty factor with a value of 0.1.

After using PTS, we also adopt another loss function \( \text{Loss}_2 \), i.e., \( L_2 \) loss, to fine-tune our SR model for further improving the SR performance and obtain the final SR results,

\[
\text{Loss}_2 = \| I_{HR} - I_{SR} \|_2. \tag{7}
\]

With PTS and adopting the \( \text{Loss}_2 \) loss function, our model achieves state-of-the-art SR performance.

### 4. Experiments

#### 4.1. Datasets

We train our proposed ARFFT with a large combination training dataset consisting of DIV2K [27], Flicker2K [28] and LSDIR [29]. Additionally, we use Bicubic downsampling to obtain the low-resolution inputs using 4 scale factor downsampling operation. DIV2K includes 800 training images and Flicker2K includes 2650 training images. Besides, LSDIR is a new large-scale dataset containing 84991 high-quality training images, 1000 validation images, and 1000 test images to fully exploited the information of datasets. To evaluate our model performance, we perform validation on Image Super-Resolution benchmark datasets Set5 [31], Set14 [32], BSD100 [32], Urban100 [33] and Manga109 [34] for our SR task.

#### 4.2. Implementation details

For the network settings, we set the number of Residual Group and the number of the combination block \( N_B \) are 6 and 12. The non-overlapping window size \( W \), the interval size of S-MSA, and the number of attention heads in D-MSA/S-MSA are set as 12, 4, and 6. The channel dimension is set as 180 for most layers. In practice, we treat 1x1 patch as a token. All the convolution layers are equipped with \( 3 \times 3 \) kernel, 1-length stride, and 1-length padding, so the height and width of feature map remain unchanged.

Our ARFFT is trained using progressive training strategy to gradually improve SR performance. Specifically, in the first training stage, we use the batch size and patch size pair \([32,48]\) to train our initial SR results for 600k iterations.
The initial learning rate is $2 \times 10^{-4}$ and is reduced by half as the training iteration reaches 200k, 400k, 500k, where 1k means one thousand. In the second training stage, we adjust the batch size and patch size pair as $[16, 64]$ and initial learning rate as $2 \times 10^{-5}$ to train our ARFFT for improving SR performance, the number of iterations and adjustment of the learning rate is the same as the first training stage. In the third training stage, we use the batch size and patch size pair $[8, 84]$ to train our ARFFT for improving SR performance. The learning rate is reduced by half as the training iteration reaches 200k, 350k, 450k. Moreover, we fine-tune our SR model keeping the batch size and patch size pair $[8, 84]$ and using the $L_2$ loss for 10k iterations with a small learning rate of $1 \times 10^{-6}$. Except for the first training stage, the best model of the previous stage is utilized to initiate the training of the current stage. ADAM optimizer is utilized to optimize our SR model in all the training process with $\beta_1 = 0.9, \beta_2 = 0.999$, and zero weight decay. We also use the data augmentation on the training data through the horizontal flip and random rotation of $90^\circ, 180^\circ, 270^\circ$. Our proposed model is implemented with PyTorch and trained with 4 NVIDIA RTX 3090 GPUs. The evaluation experimental results with in terms of PSNR and SSIM values on the Y channel of images transformed to YCbCr space.

### 4.3. Quantitative Results

We adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) for performance evaluation. Besides, we compare our proposed model with the state-of-the-art SR methods, including CNN-based approaches (EDSR [2], RCAN [3], SAN [35], SRFBN [36], HAN [37], IGNN [38], CSNLN [39], RFANet [40], NLSA [41]) and Transformer-based SR methods (IPT [4], SwinIR [6], ART [10], CAT-R [11], CAT-A [11], HAT [12]). The PSNR and SSIM results of our model for $\times 4$ image SR are presented in Table 1. As one can see from Table 1, our ARFFT achieves the best performance on all five benchmark datasets. Compared with the existing Transformer-based state-of-the-art methods, i.e, SwinIR, ART, CAT-R, CAT-A, HAT, our SR model obtains significant performance gain for $\times 4$ SR. Especially, our ARFFT achieves 0.32dB in terms of PSNR gain on Set14, 0.45dB in terms of PSNR gain on Urban100, and 0.60dB in terms of PSNR gain on Manga109 comparing the competitive method HAT. It benefits from our spatial-frequency fusion block, progressively training strategy, and larger datasets for training enabling our SR model to have stronger representation ability. These results demonstrate that our ARFFT is a stronger Transformer-based deep image SR network.

### 4.4. Qualitative Results

We provide some challenging examples for visual comparison ($\times 4$) on three test datasets in Fig. 2. Compared with representative CNN-based methods, i.e., RCAN, and representative Transformer-based methods, i.e., SwinIR and ART, we can see that our ARFFT is able to restore more detailed edges and textures. Specifically, the periodic texture of the tablecloth is clearly restored by our ARFFT, but the restored results of ART and SwinIR only focus on the simple texture due to the restricted receptive field so as to restore poor visual results. The parallel stripes with small intervals on Urban100 are failed to restored using SwinIR and ART, but the result of our ARFFT is very clear. Besides,
Figure 2. Visual quality comparisons of ×4 image SR on Set14, Urban100 and Manga109 test datasets.

Table 2. Validity of SFFB and PTS with ×4 SR in terms of PSNR and SSIM on Set14, Urban100 and Manga109.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>+ SFFB</th>
<th>+ SFFB + PTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set14</td>
<td>29.20 / 0.7946</td>
<td>29.45 / 0.7998</td>
<td>29.55 / 0.8012</td>
</tr>
<tr>
<td>Urban100</td>
<td>27.88 / 0.8336</td>
<td>28.26 / 0.8445</td>
<td>28.42 / 0.8496</td>
</tr>
<tr>
<td>Manga109</td>
<td>32.46 / 0.9285</td>
<td>32.89 / 0.9312</td>
<td>33.08 / 0.9330</td>
</tr>
</tbody>
</table>

our ARFFT also has the stronger ability to restore the blurring words on the Manga109 dataset. Compared with other methods, ARFFT obtains visually pleasing results by introducing the spatial-frequency fusion block to restore more details. It indicates that our ARFFT performs outstanding visual results for image SR.

4.5. Ablation Study

In this section, we demonstrate the importance of our method in our SR model. We train our models for ×4 image SR based on the same combination training dataset (Mentioned in 4.1) for ablation experiments. The results are evaluated on the Set14, Urban100 and Manga109 benchmark datasets. Specifically, we employ ART [10] as our baseline model. Based on the baseline model, we add the SFFB in residual groups to construct the Base_SFFB to verify the effectiveness of SFFB. The SFFB can improve by 0.38 dB and 0.43dB in terms of PSNR gain on the Urban100 and Manga109 datasets compared with the baseline. Furthermore, we employ the SFFB and proposed progressive training strategy to construct our ARFFT further improve the SR performance, which achieves a significant gain of 0.16 dB and 0.19dB on the Urban100 and Manga109 datasets in SR performance comparing the Base_SFFB. Overall, with the SFFB and PTS, our model attains a captivating performance gain of 0.54dB and 0.62dB in terms of PSNR over the baseline on the Urban100 and Manga109 datasets.

4.6. NTIRE 2023 Challenge

It consists of DIV2K, Flickr2K, and LSDIR three datasets for NTIRE 2023 Image Super-Resolution (x4)
Challenge. Specifically, in the DIV2K, the training data includes 800 high and low-resolution image pairs. The validation data includes 100 low-resolution images used for generating super-resolution corresponding images, the high-resolution images will be released when the final phase of the challenge starts. The test data includes 100 diverse images used to generate low-resolution corresponding images. Our SR model also participated in this Challenge in the validation phase and testing phase. The respective results are shown in Table 3.

5. Conclusion

In this paper, we propose an attention retractable frequency fusion Transformer (ARFFT) for image super-resolution. Due to the restricted receptive field of ART, we proposed a spatial-frequency fusion block (SFFB) to further enlarge the receptive field to improve the quality of constructed SR results. Additionally, the progressing training strategy (PTS) is proposed to gradually obtain better SR performance. With the SFFB and PTS, our model outperforms other existing state-of-the-art methods for super-resolution task.

References


[16] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In ECCV, 2022. 2


[22] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, and Zhibo Chen. FAN: Frequency aggregation network for real image super-resolution. In ECCV, 2020. 2


[29] https://data.vision.ee.ethz.ch/yawli/index.html 4


