Attention Retractable Frequency Fusion Transformer for Image Super Resolution

Qiang Zhu[1,*], Pengfei Li[1], Qianhui Li[1]
[1]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-

*Corresponding author

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. SwinFIR [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 $I_{LR} \in \mathbb{R}^{H\times D\times C_{in}}$, where H, D, and C_{in} are the height, width, and number of color channels. Firstly, LR image is sent to a 3×3 convolution layer to obtain shallow feature $F_0 \in \mathbb{R}^{H\times D\times C}$, where C is the dimension size of the feature. Next, the shallow feature is normalized and fed into the N residual groups to generate the deep feature. Specifically, each residual group consists of N_B 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 F_1. Then shallow feature and the refined feature are added to obtain the final constructed feature $F_R = F_0 + F_1$. Finally, we employ the pixel shuffle layer to generate the high-resolution image I_{SR} from the feature F_R.

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-

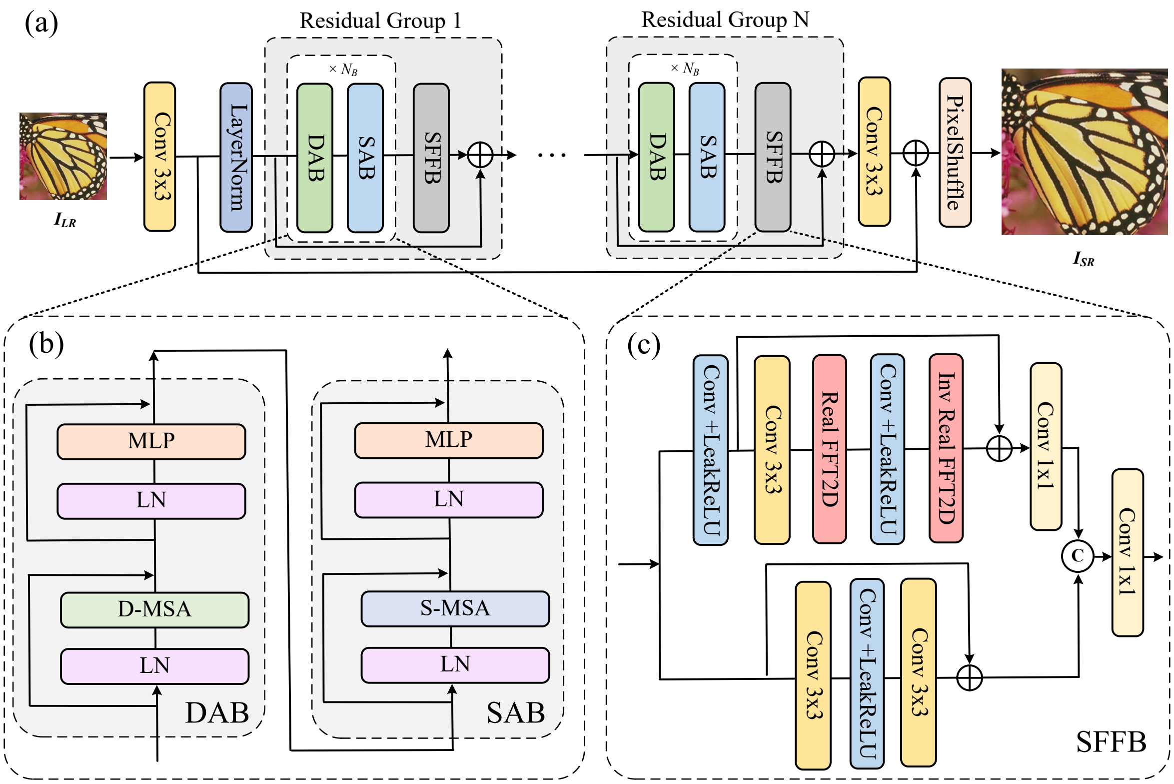

Figure 1. (a) The architecture of our proposed ARFFT for image super resolution. (b) The structure of two successive attention blocks DAB and SAB with two attention modules D-MSA and S-MSA. (c) The structure of spatial-frequency fusion block.

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 $WI \times WI$ 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_{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,

$$X_{\text{frequency}} = H_{\text{frequency}}(X). \tag{1}$$

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_{finit} for the frequency transforming,

$$X_{\text{finit}} = C_L(X), \tag{2}$$

where C_L denotes a 3×3 convolution layer with a LeakyReLU activate function. The X_{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,

$$X_{\text{frequency}} = C_1(\hat{\mathcal{F}}_T(C_L(\mathcal{F}_T(C(X_{\text{finit}}))) + X_{\text{finit}})), \tag{3}$$

where C_1 denotes a 1×1 convolution layer, $\mathcal{F}_T$ denotes a Fast Fourier Transform layer, $\hat{\mathcal{F}}_T$ denotes a 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_{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 batch 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 $Loss_1$ composed by the L_1 loss and the FFTLoss to obtain the initial SR performance

$$Loss_1 = \|I_{\text{HR}} - I_{\text{SR}}\|_1 + \alpha FFTLoss(I_{\text{HR}}, I_{\text{SR}}), \tag{6}$$

where I_{HR} is the corresponding HR image and α is the penalty factor with a value of 0.1.

After using PTS, we also adopt another loss function $Loss_2$, i.e., L_2 loss, to fine-tune our SR model for further improving the SR performance and obtain the final SR results,

$$Loss_2 = \|I_{\text{HR}} - I_{\text{SR}}\|_2. \tag{7}$$

With PTS and adopting the $Loss_2$ 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 1×1 patch as a token. All the convolution layers are equipped with 3×3 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.

Table 1. Quantitative comparisons on five SR test datasets.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	$\times$ 4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN	$\times$ 4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN	$\times$ 4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
SRFBN	$\times$ 4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
HAN	$\times$ 4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IGNN	$\times$ 4	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
CSNLN	$\times$ 4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
RFANet	$\times$ 4	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.9187
NLSA	$\times$ 4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
IPT	$\times$ 4	32.64	N/A	29.01	N/A	27.82	N/A	27.26	N/A	N/A	N/A
SwinIR	$\times$ 4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
ART	$\times$ 4	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
CAT-R	$\times$ 4	32.89	0.9044	29.13	0.7955	27.95	0.7500	27.62	0.8292	32.16	0.9269
CAT-A	$\times$ 4	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
HAT	$\times$ 4	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
Ours	$\times$ 4	33.18	0.9042	29.55	0.8012	28.14	0.7546	28.42	0.8496	33.08	0.9330

The initial learning rate is 2×10^{-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 2×10^{-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 ARFFT for 600k iterations with the initial learning rate of 1×10^{-5}. 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 1×10^{-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 β_1 = 0.9, β_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°, 180°, and 270°. 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,

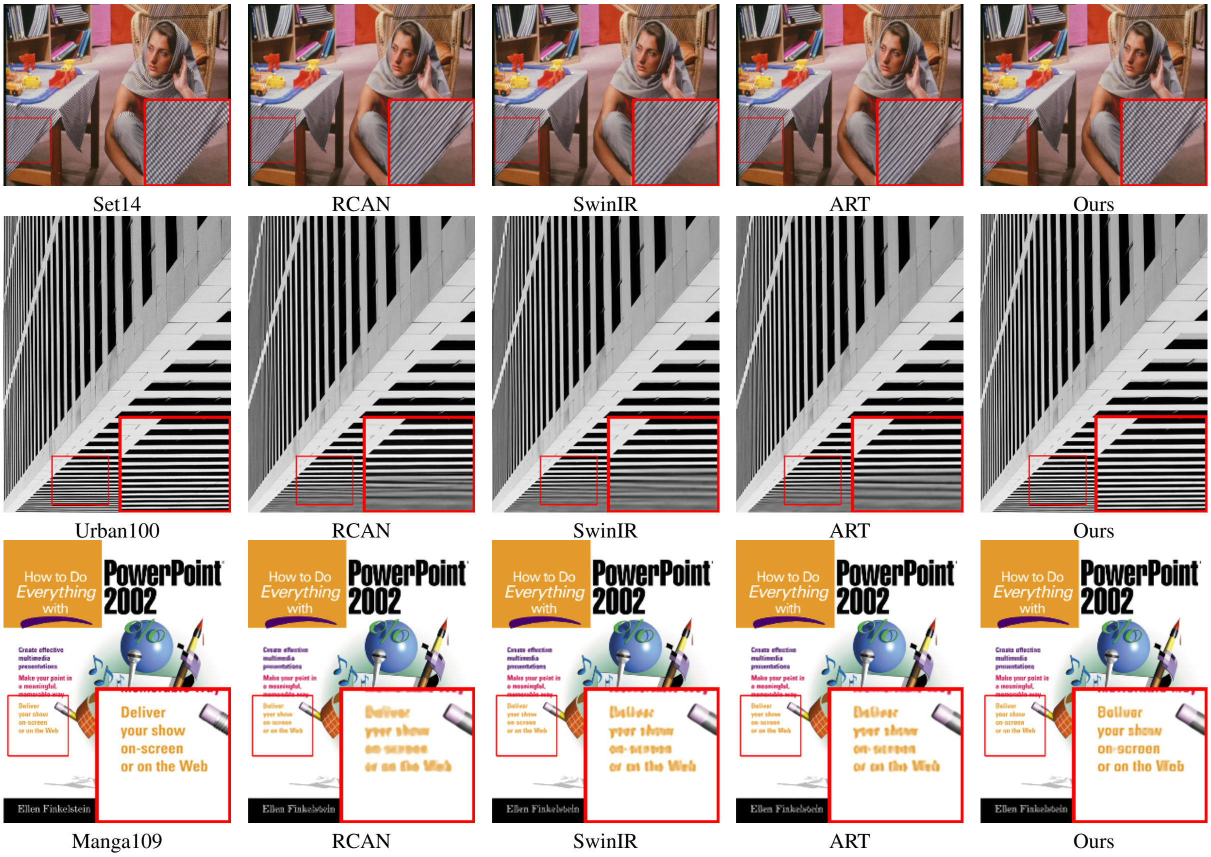

Figure 2. Visual quality comparisons of $\times$ 4 image SR on Set14, Urban100 and Manga109 test datasets.

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

	Baseline	+ SFFB	+ SFFB + PTS
Set14	29.20 / 0.7946	29.45 / 0.7998	29.55 / 0.8012
Urban100	27.88 / 0.8336	28.26 / 0.8445	28.42 / 0.8496
Manga109	32.46 / 0.9285	32.89 / 0.9312	33.08 / 0.9330

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 $\times$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

Table 3. NTIRE 2023 Challenge Results with $\times$4 SR in terms of PSNR and SSIM on validation phase and testing phase.

	Validation phase	Testing phase
PSNR	31.13	31.18
SSIM	0.85	0.86

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

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In ECCV, 2014. 1

[2] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In CVPRW, 2017. 1, 5

[3] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In ECCV, 2018. 1, 5

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In CVPR, 2021. 1, 2, 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 1

[6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In ICCVW, 2021. 1, 2, 5

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021. 1, 2

[8] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In CVPR, 2022. 1, 2, 4

[9] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In CVPR, 2022. 2

[10] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer, In ICLR, 2023. 1, 2, 3, 5, 6

[11] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In NeurIPS, 2022. 2, 5

[12] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In CVPR, 2023. 2, 5

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021. 2

[15] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. In NeurIPS, 2021. 2

[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

[17] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In ICLR, 2022.

[18] Max Ehrlich, and Larry Davis. Deep residual learning in the jpeg transform domain. In ICCV, 2019.

[19] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In CVPR, 2020.

[20] Xin Li, Xin Jin, Tao Yu, Yingxue Pang, Simeng Sun, Zhizheng Zhang, and Zhibo Chen. Learning omni-frequency region-adaptive representations for real image super-resolution. In AAAI, 2021. 2

[21] Sangwook Baek, and Chulhee, Lee. Single image super-resolution using frequency-dependent convolutional neural networks," In ICIT, 2020. 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

[23] Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, and Thomas S. Huang. D3: Deep dual domain based fast restoration of jpeg-compressed images. In CVPR. 2016. 2

[24] Jingwei Xin, Jie Li, Xinrui Jiang, Nannan Wang, Heng Huang, and Xinbo, Gao. Wavelet-based dual recursive network for image super-resolution, In TNNLS, 2022. 2

[25] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin, SwinFIR: Revisiting the SwinIR with fast fourier convolution and improved training for image super-resolution. arXiv, 2022. 2

[26] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In ICCV, 2021. 4

[27] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In CVPRW, 2017. 4

[28] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In CVPRW, 2017. 4

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

[30] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single image super-resolution based on non negative neighbor embedding. In BMVC, 2012.

[31] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In ICCS, 2010. 4

[32] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In ICCV, 2001. 4

[33] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self exemplars. In CVPR, 2015. 4

[34] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. In MTA, 2017. 4

[35] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In CVPR, 2019. 5

[36] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In CVPR, 2019. 5

[37] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In ECCV, 2020. 5

[38] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In NeurIPS, 2020. 5

[39] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Humphrey Shi. Image super resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In CVPR, 2020. 5

[40] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In CVPR, 2020. 5

[41] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In CVPR, 2021. 5