FreqFormer: Frequency-aware Transformer for Lightweight Image Super-resolution

Tao Dai[1,2] , **Jianping Wang**[1] , **Hang Guo**[3] , **Jinmin Li**[3] , **Jinbao Wang**[1,2,*] , **Zexuan Zhu**[1,2,*]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University
{daitao, wangjb, zhuzx}@szu.edu.cn, wangjianping2022@email.szu.edu.cn

Abstract

Transformer-based models have been widely and successfully used in various low-vision visual tasks, and have achieved remarkable performance in single image super-resolution (SR). Despite the significant progress in SR, Transformer-based SR methods (*e.g.*, SwinIR) still suffer from the problems of heavy computation cost and low-frequency preference, while ignoring the reconstruction of rich high-frequency information, hence hindering the representational power of Transformers. To address these issues, in this paper, we propose a novel Frequency-aware Transformer (FreqFormer) for lightweight image SR. Specifically, a Frequency Division Module (FDM) is first introduced to separately handle high- and low-frequency information in a divide-and-conquer manner. Moreover, we present Frequency-aware Transformer Block (FTB) to extracting both spatial frequency attention and channel transposed attention to recover high-frequency details. Extensive experimental results on public datasets demonstrate the superiority of our FreqFormer over state-of-the-art SR methods in terms of both quantitative metrics and visual quality. Code and models are available at https://github.com/JPWang-CS/FreqFormer.

1 Introduction

Single image super-resolution (SR), aiming at reconstructing high-resolution (HR) images from their low-resolution (LR) counterpart, has received much attention in computer vision and has a variety of potential applications [Wang *et al.*, 2020; Zhang *et al.*, 2023; Cui *et al.*, 2024; Li *et al.*, 2023b; Guo *et al.*, 2024], such as medical imaging and video surveillance [Ren *et al.*, 2019]. Recently, various CNN-based [Dai *et al.*, 2023] and Transformer-based SR methods [Dai *et al.*, 2019; Liang *et al.*, 2021] have been developed and achieved superior performance in image super-resolution.

In particular, vision Transformer with window self-attention has received much attention due to its superior performance in image restoration tasks. Among them, SwinIR [Liang *et al.*, 2021] obtains remarkable performance by adopting self-attention on a local window. ELAN [Zhang *et al.*, 2022] improves SwinIR by computing self-attention with different window sizes. Recently, SRFormer [Zhou *et al.*, 2023] develops permuted self-attention with fewer parameters and computations.

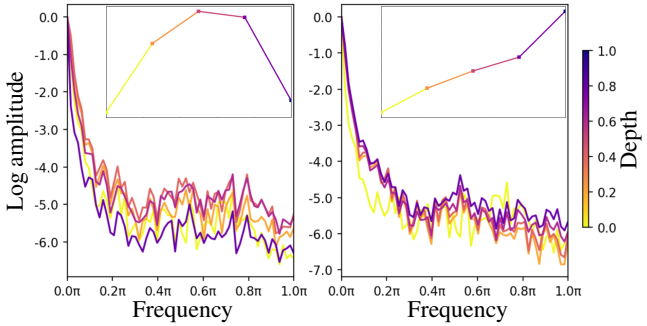


Figure 1: The power spectrum comparison of intermediate feature maps produced by SwinIR [Liang *et al.*, 2021](left) and our FreqFormer(right). SwinIR shows a significant amplitude drop in the high-frequency phase (*e.g.*, phase π) , while ours shows an opposite amplitude increase in the high-frequency phase.The thumbnail illustrates the transformation trend of feature maps as the model depth increases. Lines in darker colors correspond to features from deeper layers.

Despite the great success of Transformer-based SR methods, these methods usually contains several drawbacks. **First**, these methods are usually computational intensive, due to the huge model parameters, thus hindering the practical use of SR models in real applications. For example, the advanced SwinIR contains more than 11M parameters. **Second**, it is found that Transformer-based SR methods prefer to retain low-frequency information (see Fig.1) while ignoring the reconstruction of rich high-frequency information, hence hindering the representational power of Transformers. As shown in Fig.1, we compute the power spectrum of the feature maps of different layer depths produced from SwinIR. We can see that SwinIR shows a significant amplitude drop in the high-frequency phase as the layer goes deep, which indicates that the vanilla attention prefers to focus on low-frequency information while neglecting high-frequency information which is important to SR. Thus, these observations raise a natural question: *How to reduce model size while restoring high-*

*Corresponding author: Jinbao Wang and Zexuan Zhu

frequency image details for Transformer-based SR methods?

Inspired by the above observations, we propose FreqFormer, a Frequency-aware Transformer for lightweight image super-resolution. As shown in Fig. 2, our FreqFormer mainly contains Frequency Division Module (FDM) and Frequency-aware Transformer Block (FTB), which integrates spatial, channel, and frequency information to reconstruct the high-frequency representation. Specifically, FDM is used to separately handle high- and low-frequency information, and recover high-frequency details. Then, Frequency-aware Cascade Attention (FCA) in FTB is introduced to mix the high-frequency information from the shallow feature extraction with the channel information, and perform high-frequency recovery of the shallow extracted information. To obtain better feature representation, we further incorporate spatial frequency attention and channel transposed attention in FCA. In this way, our FreqFormer can capture spatial, frequency, and channel contexts, facilitating inter-block feature aggregation across dimensions and compensating for the loss of long-distance detail information in attention. Moreover, for the further fusion of feature representations, we design a Dual Frequency Aggregation Feed-Forward Network (DFFN), introducing a frequency gate in the middle of the fully connected layer to aggregate frequency information. Overall, our FreqFormer enhances high-frequency features within attention blocks, supplements high-frequency details externally, and achieves a comprehensive enhancement of global information from attention mechanism outputs.

Our main contributions are summarized as follows:

- We design a novel Frequency-aware Transformer, FreqFormer, which integrates spatial, channel, and frequency information to reconstruct the high-frequency representation in vanilla transformers.
- We propose the Frequency Division Module (FDM) to process high- and low-frequency information in a divide-and-conquer manner, and introduce Frequency-aware Transformer Block (FTB) to accomplish high-frequency restoration and multi-feature aggregation.
- Extensive experiments demonstrate that our FreqFormer outperforms existing state-of-the-art methods while maintaining low computational complexity and model size.

2 Related Work

2.1 CNN-based SR Methods

Since the introduction of CNNs by SRCNN [Dong *et al.*, 2016a] into the field of image super-resolution, significant success has been achieved. SRCNN is a groundbreaking work that not only utilizes CNNs but also surpasses traditional methods. Subsequently, numerous works and methods [Tai *et al.*, 2017; Kim *et al.*, 2016b] have delved deeper into network layers, exploring improved performance and structures. Some methods [Lim *et al.*, 2017; Zhang *et al.*, 2018; Dong *et al.*, 2016a; Tai *et al.*, 2017] employ abundant skip connections to accelerate network convergence and enhance reconstruction quality. For expediting SR inference, FSRCNN [Dong *et al.*, 2016b] extracts features at the LR scale and performs upsampling operations at the network's end. This pixel-shuffling [Sun *et al.*, 2023; Shi *et al.*, 2016] upsampling framework has been widely adopted in subsequent models. To augment the representational capacity of SR models, some models have introduced channel attention [Zhang *et al.*, 2018]. Recently, some works [Chen *et al.*, 2023] have explored the aggregation of spatial and channel features to further enhance SR performance.

2.2 Vision Transformer Based Methods

In recent years, Transformers have demonstrated significant potential in both natural language processing and computer vision tasks, such as image classification [Dosovitskiy *et al.*, 2021; Liu *et al.*, 2021; Li *et al.*, 2021] and segmentation [Wang *et al.*, 2021; Wu *et al.*, 2020]. Notably, the Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] has demonstrated the advantages of Transformers in feature encoding over CNNs. While ViT excels in modeling long-range dependencies in features, a body of work has established that CNNs and Transformers can be complementary. Due to their impressive performance, Transformers have been introduced for low-level visual tasks, including image SR [Chen *et al.*, 2024; Wang *et al.*, 2022]. For example, SwinIR performs self-attention on a local window for image restoration, while SRformer [Zhou *et al.*, 2023] develops permuted self-attention to enlarge receptive field.

Unlike previous works that extract features in spatial domain, we propose a novel frequency-aware Transformer, called FreqFormer, which integrates spatial, channel, and frequency information to reconstruct the high-frequency representation.

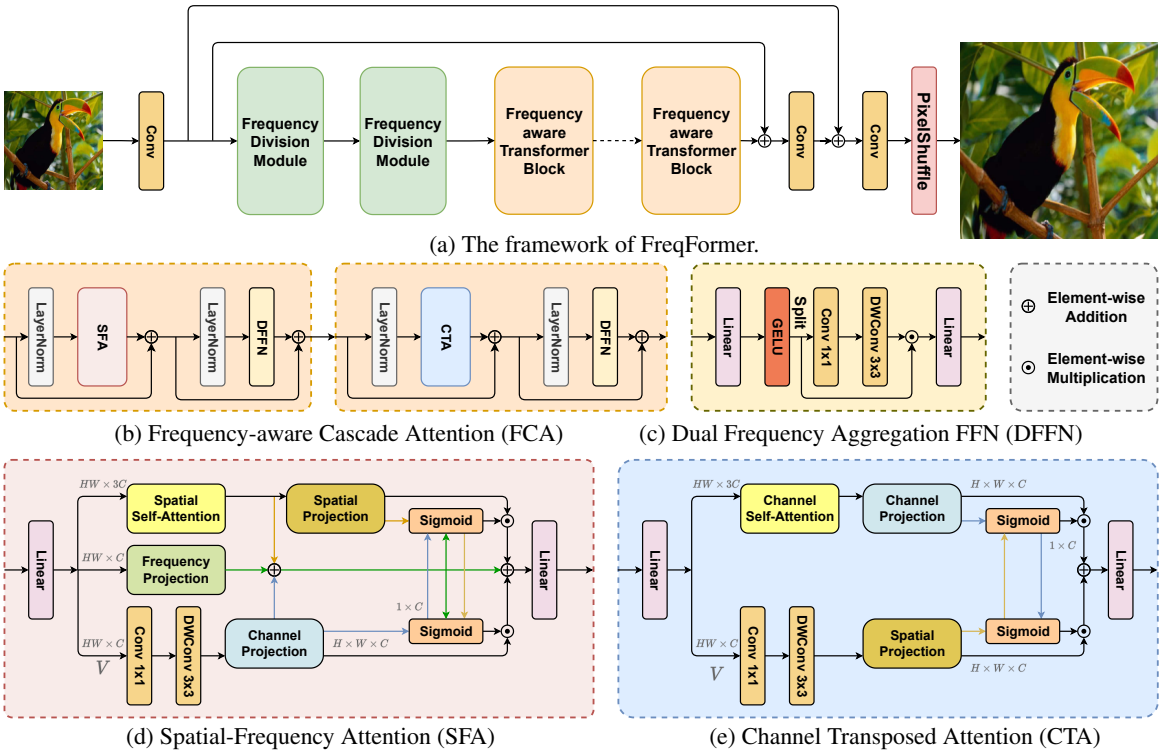
3 Methodology

As shown in Fig. 2a, our proposed FreqFormer comprises three components: shallow feature extraction, deep feature extraction, and high-resolution image reconstruction. Formally, given a LR image I_{LR}, a 3×3 convolution layer is used to extract the shallow feature $F_S \in \mathbb{R}^{H\times W\times C}$. Then F_S will go through the proposed Frequency Division Module (FDM) to process high- and low- frequency features in a divide-and-conquer manner. After that, several stacked Frequency-aware Transformer Block (FTB) is employed followed by several convolutions and a skip connection to obtain the deep features F_D. Finally, Pixel-Shuffle is used to up-sample the resolution of F_D and generate the super-resolved images I_{SR}. More details are described in the following sections.

3.1 Frequency Division Module

As shown in the previous work [Park and Kim, 2022], high-frequency information in the original image would be easily lost when extracting feature in spatial domain. For this reason, we develop Frequency Division Module to separately handle high- and low- frequency in the F_S, and recover the high-frequency details before entering deeper layers.

Inspired by the work [Patro and Agneeswaran, 2023], we first use Dual-Tree Complex Wavelet Transform (DTCWT) to separate the high- and low-frequency information of F_S. Specifically, the real part of the DTCWT results can be formulated as



(a) The framework of FreqFormer.

(b) Frequency-aware Cascade Attention (FCA)

(c) Dual Frequency Aggregation FFN (DFFN)

(d) Spatial-Frequency Attention (SFA)

(e) Channel Transposed Attention (CTA)

Figure 2: The overall architecture of our Frequency-aware Transformer (FreqFormer), which mainly consists of Frequency Division Module (FDM) and Frequency-aware Transformer Block (FTB).

$$
\begin{aligned}
X_F(u,v) &= X_\phi(u,v) + X_\psi(u,v), \\
X_\phi(u,v) &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} c_{M,h,w}\phi_{M,h,w}, \\
X_\psi(u,v) &= \sum_{m=0}^{M-1}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{k=0}^{6} d^k_{m,h,w}\Psi^k_{m,h,w},
\end{aligned}
\tag{1}
$$

where X_ϕ denotes the low-frequency scaling component and X_ψ denotes the high-frequency wavelet component. M refers to resolution, and k refers to directional selectivity.

Similarly, we can also obtain the imaginary part of the DTCWT and finally obtain L_{in} and H_{in}. $L_{in} \in \mathbb{R}^{C\times H\times W}$ represents the low-frequency component of the image. We then use 1×1 and depth-wise convolution to capture both global low-frequency information and local detailed information, resulting in $L_{out} \in \mathbb{R}^{C\times H\times W}$.

However, obtaining useful features from the high-frequency components $H_{in} \in \mathbb{R}^{k\times C\times H\times W\times 2}$ is a significant challenge. This is because the presence of multiple angles increases the dimensions by a factor of k, and due to the inclusion of real and imaginary values, the actual parameter count increases by $2k$ times. We aggregate channel and high-frequency information using fully connected layers. For computational simplification, we divide channels into C_b blocks, each with c_d channels, *i.e.*, $H_{in} \in \mathbb{R}^{2\times k\times H\times W\times C_b\times C_d}$. Along the last dimension, two linear transformations of high-frequency information provide fine-grained details such as edges, patterns, and small features. After that, we perform the inverse transform to bring the spectrum back to the physical domain, obtaining $F_{FDM} \in \mathbb{R}^{H\times W\times C}$. Additionally, we employ residual connections to transfer the high-frequency information extracted by F_{FDM} through DTCWT to different depths of the network.

3.2 Frequency-aware Transformer Block

As shown in previous work [Park and Kim, 2022], self-attention can be considered as a low-pass filter, and thus cannot work well in reconstructing high frequency details. For this reason, we propose the Frequency-aware Transformer Block (FTB) with Frequency-aware Cascade Attention (FCA) as well as Dual Frequency Aggregation Feed-Forward Network (DFFN) to help reconstruct the high-frequency details. As seen in Fig. see Fig. 2b, FCA mainly consists of two cascaded attention modules, i.e., spatial-channel attention and the channel transposed attention.

Spatial-Frequency Attention. To give the high-frequency components more focus, we propose the Spatial-Frequency Attention (SFA) shown in Fig. 2d, which combines high-frequency and channel information into self-attention to enhance recovery of high-frequency details.

In detail, we first partition the input into rectangular windows of dimensions $\frac{H\times W}{wh\times ww}$, where wh and ww denote the height and width of the rectangular window. For the i-th rectangular window feature $X^i \in \mathbb{R}^{(wh\times ww)\times C}$, we set the

queries, keys, and values as Q^i_{SF}, K^i_{SF}, and V^i_{SF}, generated by linear projection matrices without bias. The self-attention computation process is expressed as

$$Y^i_{SF\text{-}A} = \text{SoftMax}(Q^i_{SF}(K^i_{SF})^T/\sqrt{d} + B) \cdot V^i_{SF}, \quad (2)$$

where B represents relative position encoding. Additionally, following the Swin Transformer [Liu *et al.*, 2021], we use a shifting window operation to capture more spatial information. Finally, reshaping and concatenating all $Y^i_{SF\text{-}A}$ results in the output spatial self-attention feature $Y_{SF\text{-}A}$.

To extract spatial, frequency, and channel features, we perform multi-feature extraction using $Y_{SF\text{-}A}$, X, and V_{SF}. This involves Spatial-Projection (SP), Channel-Projection (CP), and Frequency-Projection (FP) modules. Initially, we employ a series of 1×1 convolutions and depth-wise convolution on V_{SF} for global and local feature fusion, obtaining preliminary channel features Y_C. Subsequently, we perform multi-dimensional mapping on the features separately for spatial, channel, and frequency dimensions as

$$\begin{aligned} Y_S &= f_{SP}(Y_{SF\text{-}A}), \\ Y_F &= f_{FP}(Linear(X)), \\ Y_C, Y_{C2} &= f_{CP}(Y_C), \end{aligned} \quad (3)$$

where Y_S represents the output after spatial mapping. We extract spatial information from $Y_{SF\text{-}A}$ using a set of 1×1 convolutions and depth-wise convolution, followed by channel-wise division. To learn more complex spatial feature representations, we introduce non-linear relationships for one-half of the channels. Finally, aggregation along the channel dimension completes the spatial feature mapping.

On the other hand, Y_F represents the extraction and transformation of frequency features. We initially map X to the frequency feature space using a fully connected layer. Global high-frequency information is extracted through a 1×1 convolution , followed by 3×3 max-pooling and a series of 1×1 convolutions and activation layers to aggregate global and local high-frequency information, resulting in the high-frequency enhanced Y_F. We simulate a 19×19 convolution using depth-wise dilation convolution (DWD-Conv) to increase the receptive field, achieving significant information extraction within channels. This convolution combination considers both spatial and channel information, greatly reducing parameter count.

Two types of outputs are provided, Y_{C1} and Y_{C2}. The former further enhances Y_C, while the latter focuses on feature extraction and mapping in the channel dimension. Next, we aggregate high-frequency and low-frequency information. In addition to fusing high-frequency information, we aim to preserve low-frequency information as much as possible. Therefore, no additional fusion is applied to the self-attention output $Y_{SF\text{-}A}$. For the high-frequency feature Y_F, we add $Y_{SF\text{-}A}$ and Y_{C2} to it to ensure that low-frequency information still dominates. For the extracted channel information Y_{C2} and frequency information Y_F, to reduce computational complexity, we pass their global information to $Y_{SF\text{-}A}$ through average pooling layers.

Finally, we perform feature fusion through cross-branch weighting. The feature fusion process is as follows:

$$\begin{aligned} Y_{SF} =& Y_{SF\text{-}A} \cdot f(\text{AvgPool}(Y_F \cdot Y_{C2})) + \\ & Y_{C1} \cdot f(Y^F \cdot Y^S) + Y_F, \end{aligned} \quad (4)$$

where $f(\cdot)$ represents the sigmoid function. Through the cross-branch weight fusion approach, we attempt to integrate the missing channel and frequency information in self-attention and the globally focused low-frequency information.

Channel Transposed Attention. Channel Transposed Attention (CTA) adopts a different strategy compared to SFA, conducting self-attention computation along the channel dimension. We employ a similar approach by dividing the channels into multiple heads, applying channel attention as illustrated in Fig. 2e. Let the input be Y_{SF}, the calculation of queries, keys, and values (Q_C, K_C, and V_C) in the self-attention process is expressed as

$$Y_{C\text{-}A} = \text{SoftMax}\left((Q_C)^T K_C/\alpha\right) \cdot V_C, \quad (5)$$

where α is a learnable temperature parameter used to adjust the dot product. Finally, we obtain the channel attention feature Y_{CA} by reshaping the connections of different attention heads. Unlike SFA, to reduce computational overhead, we only aggregate spatial and channel features in channel attention. However, for other feature dimensions, we use the same projection approach to obtain Y_{C1}, Y_{C2}, and Y_S. We then proceed with information extraction and cross-weight transfer as

$$Y_{CA} = Y_{C1} \cdot f(Y_S) + DW\text{-}Conv(V_C) \cdot f(Y_{C2}), \quad (6)$$

where $f(\cdot)$ represents the sigmoid function. Finally, we weight the channel attention output with spatial features, parallelly cross-weighting the extraction of channel features with spatial information. This complementary approach aggregates spatial and channel information.

Dual Frequency Aggregation Feed-Forward Network. To better integrate the advantages of Transformer and CNN for dual-frequency aggregation, we present the Dual Frequency Aggregation Feed-Forward Network (DFFN) (see Fig. 2(c)). Specifically, the input Y_{SF} or Y_{CA} is first projected onto X_{in} through a fully connected layer, followed by a GELU activation function. Then, a frequency gate fg is used to extract information in the frequency domain. The frequency gate is a submodule comprising two convolution layers that separate the input feature into low-frequency and high-frequency information, respectively. Then the low-frequency information remains unchanged, while the high-frequency information undergoes a series of 1×1 convolutions and depth-wise convolution($DW\text{-}Conv$) to enhance details as

$$X_{fg} = X_{in} \cdot DW\text{-}Conv(Conv_{1\times1}(X_{in})), \quad (7)$$

X_{fg} represents the multiplication of the information from the two parts, resulting in a feature map that combines both low and high-frequency information. Finally, another fully connected layer maps the fused feature map back to the original

Method	Years	Scale	Params (K)	Set5 PSNR / SSIM	Set14 PSNR / SSIM	BSD100 PSNR / SSIM	Urban100 PSNR / SSIM	Manga109 PSNR / SSIM
VDSR	CVPR16		666	36.66 / 0.9542	33.05 / 0.9127	31.90 / 0.8960	30.76 / 0.9140	37.22 / 0.9750
EDSR	CVPRW17		1,555	37.99 / 0.9604	33.57 / 0.9175	32.16 / 0.8994	31.98 / 0.9272	38.54 / 0.9769
CARN	ECCV18		1,592	37.76 / 0.9590	33.52 / 0.9166	32.09 / 0.8978	31.92 / 0.9256	38.36 / 0.9765
IMDN	MM19		694	38.00 / 0.9605	33.63 / 0.9177	32.19 / 0.8996	32.17 / 0.9283	38.88 / 0.9774
LatticeNet	ECCV20		756	38.15 / 0.9610	33.78 / 0.9193	32.25 / 0.9005	32.43 / 0.9302	- / -
ESRT	CVPRW22		777	38.03 / 0.9600	33.75 / 0.9184	32.25 / 0.9001	32.58 / 0.9318	39.12 / 0.9774
SwinIR	ICCVW21	$\times 2$	878	38.14 / 0.9611	33.86 / 0.9206	32.31 / 0.9012	32.76 / 0.9340	39.12 / 0.9783
SwinIR-NG	CVPR23		1181	38.17 / 0.9612	33.94 / 0.9205	32.31 / 0.9013	32.78 / 0.9340	39.20 / 0.9781
CRAFT	ICCV23		737	38.23 / 0.9615	33.92 / 0.9211	32.33 / 0.9016	32.86 / 0.9343	39.39 / 0.9786
SRFormer-light	ICCV23		853	38.23 / 0.9613	33.94 / 0.9209	32.36 / 0.9019	32.91 / 0.9353	39.28 / 0.9785
FreqFormer	Ours		870	**38.31** / **0.9616**	**34.12** / **0.9220**	**32.41** / **0.9026**	**33.25** / **0.9374**	**39.65** / **0.9792**
VDSR	CVPR16		666	33.66 / 0.9213	29.77 / 0.8314	28.82 / 0.7976	27.14 / 0.8279	32.01 / 0.9340
EDSR	CVPRW17		1,555	34.37 / 0.9270	30.28 / 0.8417	29.09 / 0.8052	28.15 / 0.8527	33.45 / 0.9439
CARN	ECCV18		1,592	34.29 / 0.9255	30.29 / 0.8407	29.06 / 0.8034	28.06 / 0.8493	33.50 / 0.9440
IMDN	MM19		703	34.36 / 0.9270	30.32 / 0.8417	29.09 / 0.8046	28.17 / 0.8519	33.61 / 0.9445
LatticeNet	ECCV20		765	34.53 / 0.9281	30.39 / 0.8424	29.15 / 0.8059	28.33 / 0.8538	- / -
ESRT	CVPRW22		770	34.42 / 0.9268	30.43 / 0.8433	29.15 / 0.8063	28.46 / 0.8574	33.95 / 0.9455
SwinIR	ICCVW21	$\times 3$	886	34.62 / 0.9289	30.54 / 0.8463	29.20 / 0.8082	28.66 / 0.8624	33.98 / 0.9478
SwinIR-NG	CVPR23		1190	34.64 / 0.9293	30.58 / 0.8471	29,24 / 0.8090	28.71 / 0.8627	34.24 / 0.9489
CRAFT	ICCV23		744	34.71 / 0.9295	30.61 / 0.8469	29.24 / 0.8093	28.77 / 0.8635	34.29 / 0.9491
SRFormer-light	ICCV23		861	34.67 / 0.9296	30.57 / 0.8469	29.26 / 0.8099	28.81 / 0.8655	34.19 / 0.9489
FreqFormer	Ours		878	**34.86** / **0.9307**	**30.71** / **0.8488**	**29.35** / **0.8116**	**29.15** / **0.8710**	**34.80** / **0.9513**
VDSR	CVPR16		666	31.35 / 0.8838	28.01 / 0.7674	27.29 / 0.7251	25.18 / 0.7524	28.83 / 0.8870
EDSR	CVPRW17		1,518	32.09 / 0.8938	28.58 / 0.7813	27.57 / 0.7357	26.04 / 0.7849	30.35 / 0.9067
CARN	ECCV18		1,592	32.13 / 0.8937	28.60 / 0.7806	27.58 / 0.7349	26.07 / 0.7837	30.47 / 0.9084
IMDN	MM19		715	32.21 / 0.8948	28.58 / 0.7811	27.56 / 0.7353	26.04 / 0.7838	30.45 / 0.9075
LatticeNet	ECCV20		777	32.30 / 0.8962	28.68 / 0.7830	27.62 / 0.7367	26.25 / 0.7873	- / -
ESRT	CVPRW22		751	32.19 / 0.8947	28.69 / 0.7833	27.69 / 0.7379	26.39 / 0.7962	30.75 / 0.9100
SwinIR	ICCVW21	$\times 4$	897	32.44 / 0.8976	28.77 / 0.7858	27.69 / 0.7406	26.47 / 0.7980	30.92 / 0.9151
SwinIR-NG	CVPR23		1201	32.44 / 0.8980	28.83 / 0.7870	27.73 / 0.7418	26.61 / 0.8010	31.09 / 0.9161
CRAFT	ICCV23		753	32.52 / 0.8989	28.85 / 0.7872	27.72 / 0.7418	26.56 / 0.7995	31.18 / 0.9168
SRFormer-light	ICCV23		873	32.51 / 0.8988	28.82 / 0.7872	27.73 / 0.7422	26.67 / 0.8032	31.17 / 0.9165
FreqFormer	Ours		889	**32.69** / **0.9007**	**28.95** / **0.7898**	**27.79** / **0.7444**	**26.84** / **0.8093**	**31.59** / **0.9201**

Table 1: Quantitative comparison (PSNR/SSIM) for **lightweight image SR** with state-of-the-art methods on **benchmark datasets**. The best and second-best results are marked in red and blue colors.

feature space X_{DFFN}. Through this structure, DFFN can extract frequency domain information while preserving spatial domain information.

At the end of DFFN, to compensate for the loss of long-distance detail information resulting from the transition from DFB to different layers, we incorporate a series of 1×1 convolutions and depth-wise convolution at the end of each FTB group, based on the depth. This allows us to recover high-frequency details after the self-attention mechanism using the High-Frequency Recovery Block (HFRB).

3.3 Loss Function

Following the previous work, we use the L_1 loss to minimize the distance between the model predictions I_{SR} and the ground truth I_{HR}, expressed as

$$\mathcal{L}_1 = ||I_{SR} - I_{HR}||_1 \tag{8}$$

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We followed the methodologies established in previous studies to train and test our model. Specifically, experiments were conducted for upscaling factors of $\times 2$, $\times 3$, and $\times 4$. Two training datasets, DIV2K [Lim *et al.*, 2017] and Flickr2K [Radu Timofte and Zhang, 2017], were used for model training. Additionally, five benchmark testing datasets—Set5 [Bevilacqua *et al.*, 2012], Set14 [Zeyde *et al.*, 2010], B100 [D. *et al.*, 2002], Urban100 [Huang *et al.*, 2015], and Manga109 [Matsui *et al.*, 2016]—were used to evaluate the model. Image quality was evaluated using PSNR and SSIM metrics, computed on the Y-channel (brightness) in the YCbCr color space. The low-resolution (LR) images were created by downscaling the corresponding high-resolution (HR) images using bicubic interpolation.

Implementation Details. In our training setup, the model was configured with a patch size of 64×64, and a batch size of 32. The training process comprised 500,000 iterations, with an initial learning rate of 2×10^{-4}. The learning rate was halved at specific milestones: [250K, 400K, 450K, 475K]. Data augmentation techniques, including random horizontal flipping, and rotations at 90°, 180°, and 270°, were applied to the training set. For optimization, the Adam optimizer was employed with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to minimize the $\mathcal{L}_1$ loss. Additionally, the model was trained using the PyTorch toolkit on 4 NVIDIA 3090 GPUs.

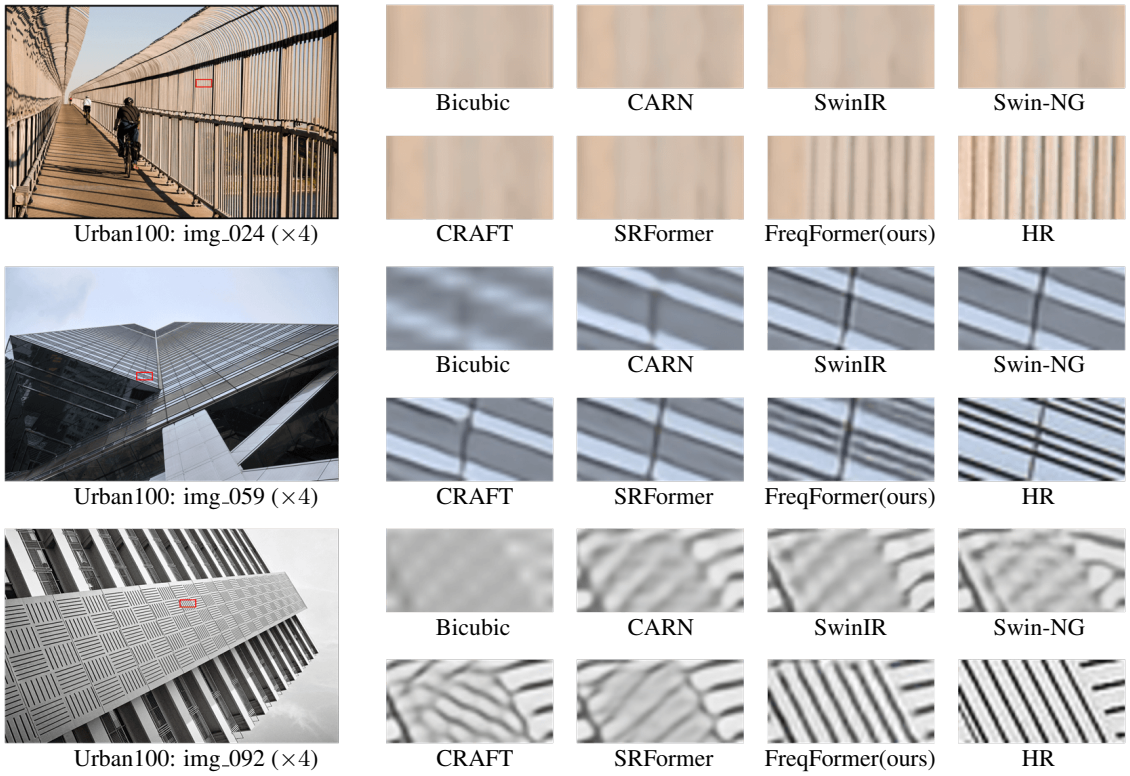


Figure 3: Qualitative comparison for **lightweight image SR**(×4) task in some challenging cases.

FDM	SFA	CTA	Params (K)	FLOPs (G)	PSNR (dB)
	✓		856	55.2	38.26
		✓	719	46.5	38.16
	✓	✓	788	50.8	38.27
✓	✓	✓	870	57.4	38.29

(a) Ablation study of alternate strategy.

Spatial-Projection	Frequency-Projection	Channel-Projection	Params (K)	PSNR (dB)
✓			778	38.26
	✓		719	38.25
		✓	769	38.26

(b) Ablation study of Projection, w/o FDM.

Table 2: Ablation studies. The models are trained on DIV2K and Flickr2K, and tested on Set5(×2)

4.2 Comparison With State-of-the-Arts

We compare the effectiveness of the proposed FreqFormer with several state-of-the-art Single Image Super-Resolution (SISR) methods, including VDSR [Kim *et al.*, 2016a], EDSR [Lim *et al.*, 2017], IMDN [Hui *et al.*, 2019], LatticeNet [Luo *et al.*, 2020], ESRT [Lu *et al.*, 2022], SwinIR [Liang *et al.*, 2021] , SwinIR-NG [Choi *et al.*, 2023], CRAFT [Li *et al.*, 2023a], and SRFormer [Zhou *et al.*, 2023].

Quantitative Results. The quantitative results on five benchmark datasets are presented in Table 1. The FreqFormer, proposed in this study, outperforms all comparative methods on the benchmark datasets, showcasing significant advantages across all test cases. Compared to traditional CNN-based methods, such as EDSR, the FreqFormer exhibits notable performance improvements on the Manga109 dataset at scaling factors of ×2, ×3, and ×4, achieving improvements of 1.11 dB, 1.35 dB, and 1.24 dB, respectively. Notably, when compared to other Transformer architectures with similar parameter counts, like SwinIR and ESRT, FreqFormer consistently achieves superior performance. On Urban100 and Manga109 datasets at scaling factor of ×2, our model also demonstrates performance gains of 0.34 dB and 0.26 dB. These quantitative results collectively underscore the effectiveness and necessity of the self-attention modules in restoring high-frequency information.

Qualitative Comparison. In Fig. 3, we present qualitative comparisons at ×4 magnification. For distant high-frequency details, our model demonstrates state-of-the-art results. In img_024, we are able to restore high-frequency details that other models cannot, while alternative methods may result in blurriness or artifacts in these complex areas. Similar observations can be found in img_059 and img_092, suggesting that our approach effectively reduces artifacts, preserves more structures, and captures surprising details.

4.3 Real-world Image Super-Resolution

In Fig. 4, we present the visual results generated by different methods on real-world low-resolution images. In the first row of visual comparisons, the wall texture created by our FreqFormer appears more natural, clear, and accurate, eliminating

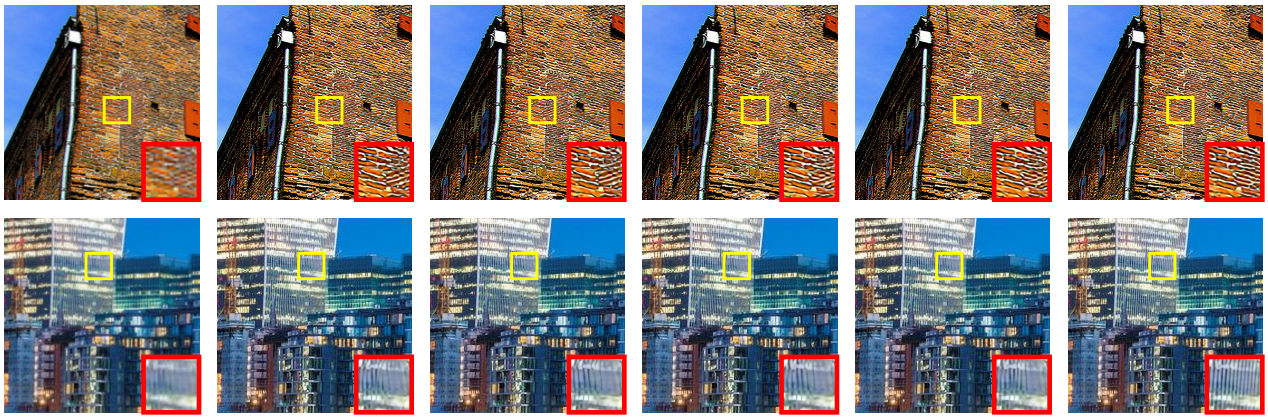


Figure 4: Qualitative comparisons with recent state-of-the-art methods on the **real-world image SR**($\times 4$) task.

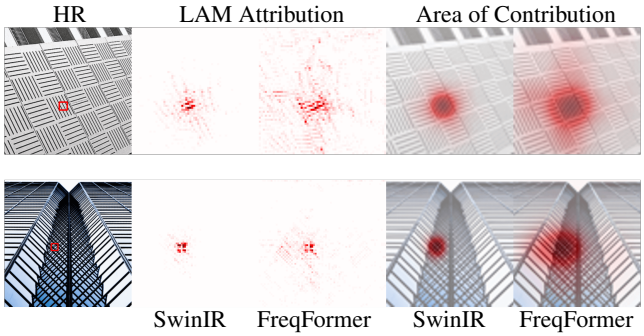


Figure 5: In challenging reconstruction tasks such as SwinIR [Liang *et al.*, 2021] (left) and FreqFormer (right), we can observe that FreqFormer achieves superior results across a wide range of pixels.

the edge distortions caused by traditional Transformers. In the second row, the glass curtain produced by FreqFormer is sharper and more cohesive. It is evident that our approach can produce and restore sharper edges, especially noticeable in visual effects, demonstrating the significant potential applicability of our method in real-world scenarios.

4.4 LAM Visualization

As seen in Fig. 5, the pixels restored by FreqFormer extend across almost the entire image, in contrast to SwinIR, which only aggregates within a limited range. These results indicate that our approach can activate more pixels to reconstruct low-resolution input images. These observations confirm the effectiveness of our method and demonstrate the superiority of our approach from an interpretability perspective.

4.5 Ablation Study

For a fair comparison, we trained models on the DIV2K dataset and the Flickr2K dataset, and tested them on the Set5 dataset in ablation studies. All models share the same implementation details as FreqFormer, except for the iteration count, which is set to 300K. Additionally, we set the output size to $3\times256\times256$ for calculating FLOPs.

Alternate Strategy. To showcase the efficacy of FDM, SFA, and CTA, we carried out several sets of experiments, as depicted in the structure shown in Fig. 2a. Firstly, we compared the performance of DFFN with and without FDM, as indicated in the first and second rows of the table. We replaced all attention blocks with SFA and CTA, respectively. The third row represents the concatenation of SFA and CTA, alternately used in the FCA. By comparison, we can observe that the model using only SFA performs significantly better than the one using only CTA. However, the best performance (38.27 dB) is achieved when alternating between CFA and CTA. Furthermore, the combined use of CFA and CTA results in a decrease in Params by 68K and FLOPs by 4.4G compared to using SFA alone. This suggests that CTA serves as a complementary block to SFA and plays a crucial role in image restoration. We also performed an ablation study on FDM, and the results indicate that combining high-frequency information and channel information after shallow feature extraction effectively restores high-frequency details.

Different Projection. We verify the effectiveness of projection parts in Table 2b. Firstly, we conducted an ablation study on Spatial-Projection, providing a PSNR of 38.26dB at 30K iterations. In the second row, we tested the most critical ablation experiment on Frequency-Projection. Despite having 719K parameters, the model achieved only 38.25dB PSNR, indicating the effectiveness and necessity of restoring high-frequency information after the self-attention mechanism. The last row, Channel-Projection, demonstrates that feature restoration solely in spatial dimensions is insufficient. To make better use of information for image restoration, channel information needs to be incorporated.

5 Conclusion

In this paper, we propose FreqFormer, a novel Hybrid Attention Transformer. Our model integrates spatial, frequency, and channel information, activating more pixels for reconstruction through multi-branch aggregation. Additionally, by blending frequency and channel information, along with various frequency aggregations, it effectively restores intricate details in images, such as edges, patterns, and small features. Extensive experiments have demonstrated that our FreqFormer outperforms several state-of-the-art methods in various image super resolution tasks.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China, under Grant 2022YFF1202104, National Natural Science Foundation of China, under Grant (62302309,12326619), Shenzhen Science and Technology Program (Grant No.JCYJ20220818101014030), Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No.SZUBDSC-OF2024-23).

References

[Bevilacqua *et al.*, 2012] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Alberi Morel. Low-complexity single image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012.

[Chen *et al.*, 2023] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *CVPR*, 2023.

[Chen *et al.*, 2024] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. In *ICLR*, 2024.

[Choi *et al.*, 2023] Haram Choi, Jeongmin Lee, and Jihoon Yang. N-gram in swin transformers for efficient lightweight image super-resolution. In *CVPR*, 2023.

[Cui *et al.*, 2024] Y. Cui, W. Ren, X. Cao, and A. Knoll. Image restoration via frequency selection. *IEEE TPAMI*, 2024.

[D. *et al.*, 2002] Martin D., Fowlkes C., Tal D., and Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2002.

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

[Dai *et al.*, 2023] Tao Dai, Mengxi Ya, Jinmin Li, Xinyi Zhang, Shu-Tao Xia, and Zexuan Zhu. Cfgn: A lightweight context feature guided network for image super-resolution. *IEEE TETCI*, 2023.

[Dong *et al.*, 2016a] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 2016.

[Dong *et al.*, 2016b] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, 2016.

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

[Guo *et al.*, 2024] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *arXiv*, 2024.

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

[Hui *et al.*, 2019] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM*, MM, 2019.

[Kim *et al.*, 2016a] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[Kim *et al.*, 2016b] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016.

[Li *et al.*, 2021] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. In *CVPR*, 2021.

[Li *et al.*, 2023a] Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In *ICCV*, 2023.

[Li *et al.*, 2023b] Jinmin Li, Tao Dai, Mingyan Zhu, Bin Chen, Zhi Wang, and Shu-Tao Xia. Fsr: A general frequency-oriented framework to accelerate image super-resolution networks. In *AAAI*, 2023.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR*, 2021.

[Lim *et al.*, 2017] B. Lim, S. Son, H. Kim, S. Nah, and K. Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017.

[Liu *et al.*, 2021] 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.

[Lu *et al.*, 2022] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *CVPR*, 2022.

[Luo *et al.*, 2020] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *ECCV*, 2020.

[Matsui *et al.*, 2016] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *IEEE MTA*, 2016.

[Park and Kim, 2022] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2022.

[Patro and Agneeswaran, 2023] Badri N. Patro and Vijay Srinivas Agneeswaran. Scattering vision transformer: Spectral mixing matters. In *NeurIPS*, 2023.

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

[Ren *et al.*, 2019] Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, and Tong Xin. Face video deblurring using a 3d facial prior. In *ICCV*, 2019.

[Shi *et al.*, 2016] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[Sun *et al.*, 2023] Bin Sun, Yulun Zhang, Songyao Jiang, and Yun Fu. Hybrid pixel-unshuffled network for lightweight image super-resolution. In *AAAI*, 2023.

[Tai *et al.*, 2017] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, 2017.

[Wang *et al.*, 2020] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE TPAMI*, 2020.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

[Wang *et al.*, 2022] 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.

[Wu *et al.*, 2020] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. In *CVPR*, 2020.

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

[Zhang *et al.*, 2018] 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.

[Zhang *et al.*, 2022] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, 2022.

[Zhang *et al.*, 2023] Aiping Zhang, Wenqi Ren, Yi Liu, and Xiaochun Cao. Lightweight image super-resolution with superpixel token interaction. In *ICCV*, 2023.

[Zhou *et al.*, 2023] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*, 2023.