Large Kernel Frequency-enhanced Network for Efficient Single Image Super-Resolution

Jiadi Chen Chunjiang Duanmu Huanhuan Long
Zhejiang Normal University, Jinhua, China
zjnucjd@163.com duanmu@zjnu.cn huanhuanlong@zjnu.edu.cn
https://github.com/ThediidehT/LKFN

Abstract

In recent years, there has been significant progress in efficient and lightweight image super-resolution, due in part to the design of several powerful and lightweight attention mechanisms that enhance model representation ability. However, the attention maps of most methods are obtained directly from the spatial domain, limiting their upper bound due to the locality of spatial convolutions and limited receptive fields. In this paper, we shift focus to the frequency domain, since the natural global properties of the frequency domain can address this issue. To explore attention maps from the frequency domain perspective, we investigate and correct some misconceptions in existing frequency domain feature processing methods and propose a new frequency domain attention mechanism called frequency-enhanced pixel attention (FPA). Additionally, we use large kernel convolutions and partial convolutions to improve the ability to extract deep features while maintaining a lightweight design. On the basis of these improvements, we propose a large kernel frequency-enhanced network (LKFN) with smaller model size and higher computational efficiency. It can effectively capture long-range dependencies between pixels in a whole image and achieve state-of-the-art performance in existing efficient super-resolution methods.

1. Introduction

As a low-level computer vision task, single-image super-resolution (SISR) aims to reconstruct a high resolution (HR) image from its low resolution (LR) counterpart. Since SRCNN [9] introduced deep learning to super-resolution for the first time, there has been a significant surge in the development of deep-learning-based SR models. By leveraging large amounts of data and powerful computing resources, deep learning has enabled researchers to develop increasingly sophisticated SR models that can generate high quality image from low-resolution inputs. Despite their impressive results, due to their high complexity and computational cost, traditional super-resolution networks are often difficult to use in practical applications. In this context, efficient super-resolution (ESR) networks with greatly reduced parameters and less computational complexity are gradually being introduced and developed.

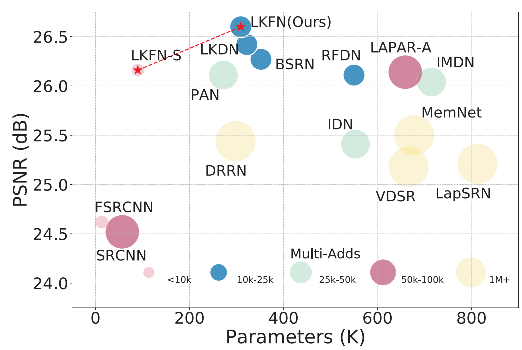

Figure 1. Comparison of model performance and complexity on Urban100 with SR(×4).

Among these ESR methods, a class of methods based on information distillation paradigm have been verified effective, which consist of three parts: feature distillation, feature condensation and feature enhancement. For the feature distillation part, IMDN [18] first introduced a progressive refinement module to reduce computational cost and achieve multi-level feature map fusion by splitting channels. RFDN [26] further introduced shallow residual block (SRB) and applied channel compression to greatly reduce the number of model parameters. By rethinking the design of SRB, BSRN [25] introduced the blueprint separable convolution (BSConv) to replace the vanilla 3×3 convolution and the GELU [16] activation function was used instead of ReLU, which achieved remarkable results. LKDN [43] used the technique of reparameterization [6, 49] to fur-

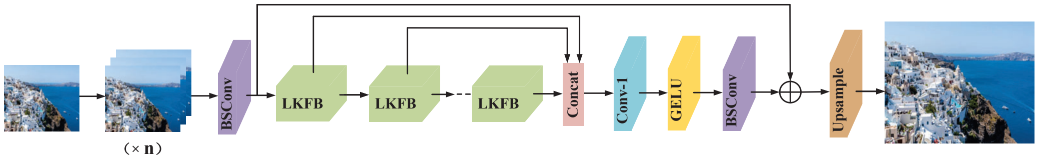

Figure 2. The architecture of Large Kernel Frequency-enhanced Network (LKFN)

ther improve the representation capability of BSConv with zero additional inference overhead. In the part of feature enhancement, all the aforementioned methods used some form of attention mechanism, from spatial attention, channel attention, pixel attention, to their combinations, such as IMDN, RFDN's contrast-aware channel attention (CCA), BSRN's enhanced spatial attention (ESA) [27] plus CCA, LKDN's large kernel attention (LKA), and MDRN's multi-level dispersion spatial attention (MDSA) plus enhanced contrast-aware channel attention (ECCA) [31]. These attention mechanisms have shown remarkable effects in preserving model accuracy while keeping the model lightweight.

In this paper, we further explore the potential of the feature distillation and feature enhancement parts as well as how to make them work better together to adapt to the super-resolution task. Keeping the information distillation framework unchanged, we propose a new spatial feature extraction block with a larger convolution kernel to replace BSConv, and a novel attention mechanism based on frequency domain image processing to realize feature enhancement. We refer to this new ESR method as the large kernel frequency-enhanced network (LKFN). Extensive experiments demonstrate that our LKFN better balances the accuracy and complexity of the model, and achieves the state-of-the-art performance among existing ESR methods (See Fig. 1). The contributions of this paper can be summarized as follows:

- We introduce larger kernel convolution to the basic feature extraction block, which provides a larger receptive field while maintaining lightweight.
- We propose a brand new attention mechanism that is completely based on frequency domain processing, which can truly achieve a global view of the whole image and is more flexible for different scales.
- The proposed LKFN achieves better super-resolution performance in a more concise and efficient manner.

2. Related Works

2.1. Exploration of Efficient Super-Resolution

Numerous approaches have been explored and achieved effective results in reducing the computational complexity of deep-learning-based super-resolution methods in various aspects. FSRCNN [10] proposed a network paradigm that places the upsampling step in the last stage, replacing the enormous computational cost incurred by SRCNN [9] which processes the upscaled input image from interpolation. The sub-pixel convolutional upsampling method proposed by ESPCN [37] has been widely adopted as an upsampling module due to its exceptional performance. DRCN [19] proposed a deep recursive convolutional network to increase the depth of the network, ensuring effectiveness while reducing the burden of too many parameters. CARN [2] used group convolutions and a cascading mechanism on residual networks to improve efficiency. ASSLN [50] proposed an aligned structured sparsity learning strategy, which successfully introduced the filter pruning technique in the SR models. DIPNet [46] integrated reparameterization, filter pruning, and knowledge distillation techniques, and won the championship in inference speed in the NTIRE 2023 efficient super-resolution challenge [24].

2.2. The Renaissance of Large Kernel Convolution

In the early days of CNN models, large convolution kernels were commonly used (e.g. AlexNet [21], SRCNN). This changed with the VGG model [38], which popularized the stacking of small convolution kernels (3×3) and became the standard for CNN architecture design. However, following the emergence of transformer-based vision models [11, 29] that emphasized the importance of global receptive field, many researchers found that using larger convolution kernels in traditional CNNs [7, 8, 15, 28, 30] can achieve comparable or even better performance than transformer-based models by reducing network depth and improving feature extraction efficiency. This trend has also influenced the design of SR models. Nevertheless, simply using large convolution kernels will lead to higher computational costs. Inspired by large kernel attention (LKA) in VAN [15], convolution kernel decomposition technique have been widely adopted to address this issue, which decomposes a large convolution kernel into three parts equivalently: a depth-wise convolution, a depth-wise dilation convolution, and a 1×1 convolution. LKASR [13] imitates the framework structure of the Transformer [41] and achieves good performance by replacing the self-attention module with LKA. VAPSR [52] uses a concise structure mainly composed of the LKA module with its attention channels amplified, it achieves excellent results with fewer parameters, proving the superiority of the LKA module. MAN [42] improves the LKA module by proposing a multi-scale large kernel attention (MLKA) that combines multiple scales in parallel and integrates it with a gated spatial attention. LKDN [43] applies the LKA to the effective information distillation framework and achieves SOTA performance. These methods all incorporate large kernel convolution into attention mechanisms to enhance feature representation. However, our proposed LKFN finds that using large kernel convolution directly in the feature extraction process can achieve significant improvement as well.

2.3. Frequency Domain Methods in CV

The Fourier transform has long been an essential tool in digital image processing. In deep learning-based vision tasks, a variety of works have attempted to incorporate it in their model design because according to the convolution theorem, point-wise update in the frequency domain globally affects all input features involved in Fourier transform. This property has a natural global attribute. FFC [5] replaces the convolution in CNNs with a local Fourier unit and performs convolutions in the frequency domain, which can complementarily address different scales. GFNet [35] proposes a global filter network that performs element-wise multiplication between frequency domain features and learnable global filters. SpectFormer [34] combines spectral and multi-headed self-attention in the original ViT [11] architecture to obtain a better representation ability. For image super-resolution, FNNSR [22] proposes a neural network design that operates entirely in the frequency domain. It takes a bicubic-upsampled image as input, transforms it into the frequency domain, and then performs element-wise multiplication using weight matrices of the same size, to achieve the effect of non-linear activation, they utilize the frequency-domain convolution. IFNNSR [45] improves FNNSR by dividing its weight matrices into four quadrants and sharing parameters, which reduces the number of parameters and improves the speed. These two methods are entirely based on the frequency domain and are too radical. Their actual performance is far behind all spatial domain methods and is only slightly better than the interpolation methods. Inspired by FFC [5], SwinFIR [48] integrates frequency-domain fast Fourier convolution with spatial domain convolution into a complementary dual-branch structure module, and embeds it into the SwinIR [29] framework, achieving impressive results. Other methods like ShuffleMixer [39] and SAFMN [40], different from directly adding frequency-domain processing modules in the model structure, they instead add frequency-domain constraints to the loss function. Our LKFN explores the combination of frequency-domain methods and attention mechanisms, proposes a frequency-enhanced pixel attention, and explains why some traditional frequency-domain operations are not suitable for super-resolution tasks.

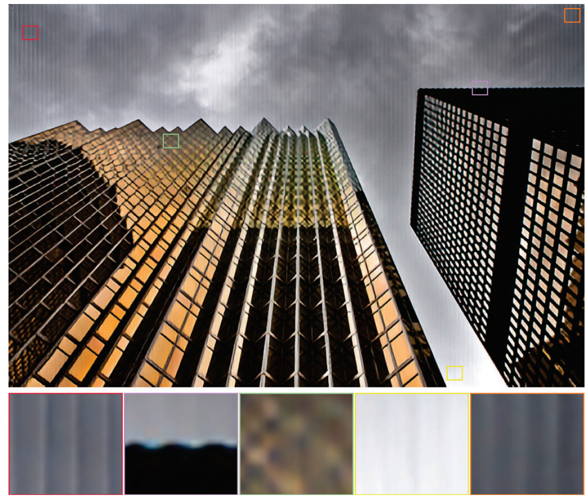

Figure 3. The super-resolved image with the operation in FFC [5].

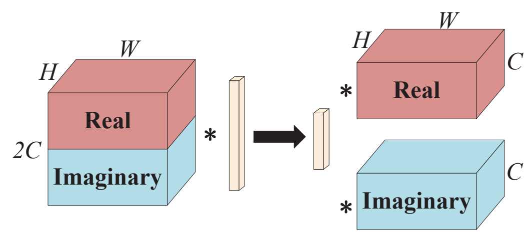

Figure 4. Improve the operations in the frequency domain.

3. Method

3.1. Rethinking Frequency-domain Operations

As we know, convolution operations in the spatial domain are equivalent to element-wise multiplication in the frequency domain. To enjoy the advantages of global view in the frequency domain, it is reasonable to use a learnable parameter matrix as a global filter in GFNet [35]. However, this method is not suitable for SR for two reasons. First, the learnable weight matrix size is fixed, that is, $C \times H \times W$, the same size as the input feature, which means it is only applicable to networks with fixed input size like image classification, object detection or semantic segmentation models while SR networks receive inputs of arbitrary resolution. Second, even if the input size can be fixed in some way, the

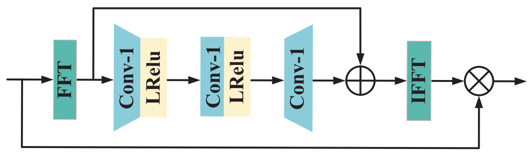

Figure 5. Architecture of frequency-enhanced pixel attention (FPA).

number of parameters in the weight matrix is large enough, greatly increasing the model size.

SwinFIR [48] explored directly replacing the vanilla convolution with the Fourier unit in the FFC [5]. However, the results did not improve the SR performance as expected, and decreased instead, which seems to contradict the theoretical advantages brought by the global view in the frequency domain. So they added spatial residual blocks to form a dual-branch structure, which improved the performance. However, it is difficult to determine how much of the effect is due to the frequency-domain processing in this design.

We studied the specific operation of the Fourier unit and found the problem. When the Fourier transform is applied to real numbers, each element of the frequency-domain feature map is a binary tuple consisting of real and imaginary parts. Since mainstream deep learning frameworks do not support direct operations on complex numbers, the Fourier unit uses a method of stacking real and imaginary parts in the channel direction and then using a 1×1 convolution to process across the doubled channels (the left part in Fig.4). Here lies the problem. This processing method causes data exchange between real and imaginary parts, which greatly destroys the phase angle, fundamentally disrupting the spatial structure and feature localization of the image. Super-resolved images obtained in this way have obvious louver-board-like artifacts, as shown in Fig.3. Considering that the FFC is designed for image classification, this method would cause a severe decrease in performance when directly applied to SR. So, we made an improvement, which is to isolate the data communication between the real and imaginary parts, see Fig.4. We use the same convolution to process the real and imaginary parts separately, which also avoids the increase in parameters caused by doubling the number of channels. The super-resolved images immediately returned to normal, and the artifacts disappeared.

3.2. Frequency-enhanced Pixel Attention

Through rethinking and improving the frequency-domain operations in SR, we can further explore the advantages brought by frequency-domain methods. Therefore, we combined frequency-domain processing with the attention mechanism and proposed the Frequency-enhanced Pixel Attention (FPA). Normally, attention maps are extracted from feature maps in the spatial domain. Due to the locality of the convolution operator, learning the correlation between pixel locations in the spatial domain can only cover a small range, which greatly reduces the effectiveness of attention mechanisms. Although using larger convolution kernels can alleviate this problem to some extent, it can not truly achieve the global attention like self-attention, while bringing larger computational costs and larger model sizes.

In our FPA, see Fig.5, we first use the fast Fourier transform $fft(\cdot)$ to convert the spatial domain feature map with shape $C \times H \times W$ into the frequency domain, obtaining a frequency-domain feature map of shape $C \times H \times \lfloor W/2 \rfloor + 1$. Since the Fourier transform of a 2D real signal is a Hermitian matrix which is conjugate symmetric, so half of the information is redundant. The frequency domain feature map is then processed by a three-layer 1×1 convolution, followed by two LeakyReLUs, and a residual connection is added with the initial frequency domain feature map. Then the pixel attention map is obtained by inverse fast Fourier transform $ifft(\cdot)$ back to the spatial domain, and multiplied by the initial input F . This process can be expressed as follows:

$$F_{attention} = ifft(fft(F) + fe(fft(F))) \tag{1}$$

$$F_{enhanced} = F_{attention} \otimes F \tag{2}$$

where $fe(\cdot)$ denotes the module of frequency-domain enhancement with the three-layer 1×1 convolution, $F_{attention}$ denotes the pixel attention map, $\otimes$ denotes element-wise product operation.

3.3. Large Kernel Frequency-enhanced Block

The specific architecture is shown in Fig.6. Inspired by LKDB in LKDN [43], we design a large kernel frequency-enhanced block (LKFB), which incorporates our powerful FPA module. On the other hand, inspired by large kernel convolutions and PConv [4], we propose the Partial Large Kernel Block (PLKB) to replace the RBSB in LKDB. In order to extract more hierarchically rich feature maps and cope with the increased parameters and computation caused by the larger convolution kernel, we further use partial convolution to reduce the channels. The finer-grained feature maps obtained in this way, combined with the global attention brought by FPA, enable the proposed LKFB to achieve comparable or even better performance in a lightweight manner. For the input F_{in}, feature distillation is performed first, the process can be expressed as

$$\begin{aligned} F_{d_1}, F_{r_1} &= D_1(F_{in}), PLKB_1(F_{in}), \\ F_{d_2}, F_{r_2} &= D_2(F_{r_1}), PLKB_2(F_{r_1}), \\ F_{d_3}, F_{r_3} &= D_3(F_{r_2}), PLKB_3(F_{r_2}), \\ F_{r_4} &= BSConv(F_{r_3}), \end{aligned} \tag{3}$$

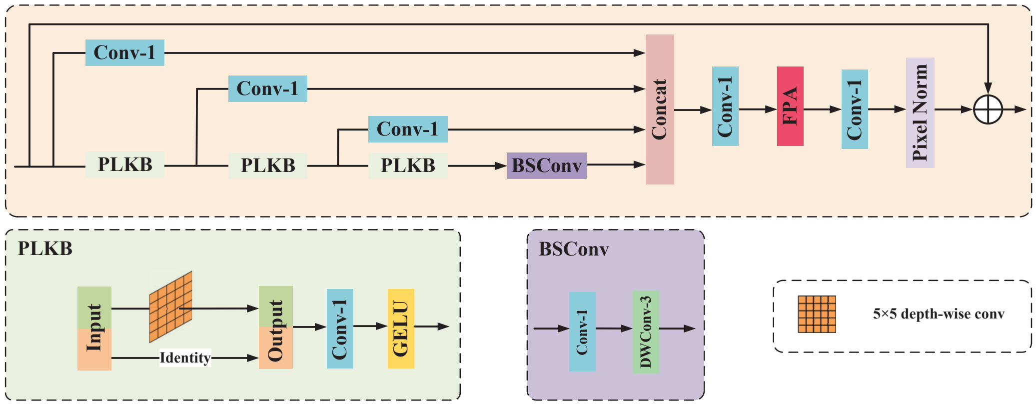

Figure 6. The architecture of Large Kernel Frequency-enhanced Block (LKFB).

$$PLKB_i(F) = Conv_{1\times1}(Conv_{DW}(F_{split_1}), F_{split_2}), \tag{4}$$

where D_i, $PLKB_i$ denotes the ith distillation(1×1 conv) and ith refinement layer using the proposed PLKB, respectively. F_{d_i}, F_{r_i} represents the ith distilled feature and ith refined feature, respectively. $BSConv$[25] is used as the last refinement layer. In the PLKBs, $Conv_{DW}$ denotes a 5×5 depth-wise convolution. $F_{split_{1,2}}$ represent the split two parts of the input feature. $(\cdot,\cdot)$ means concatenating two parts in the channel dimension. Subsequently, the distilled features from the distillation layers and the final refinement output are concatenated and fused with a 1×1 convolution:

$$F_{fused} = Conv_{1\times1}(Concat(F_{d_1}, F_{d_2}, F_{d_3}, F_{r_4})), \tag{5}$$

Next, the fused feature map undergoes image enhancement through FPA module, followed by a layer of 1×1 convolution, and finally normalized through Pixel Normalization [52]:

$$F_{enhanced} = PixelNorm(Conv_{1\times1}(FPA(F_{fused}))), \tag{6}$$

Finally, a residual connection within the block is connected with the input to enhance the learning ability of the deep model:

$$F_{out} = F_{enhanced} + F_{in}. \tag{7}$$

3.4. Network Architecture

Follow LKDN, our approach copy the original input image I_{LR} n times and stack them along the channel direction to obtain I^n_{LR}, then map it to the feature space through a 3×3 BSConv to obtain F_0:

$$F_0 = BSConv(I^n_{LR}), \tag{8}$$

Then we feed F_0 into m stacks of LKFBs to extract deep features. The output of each module in the middle is stacked together and undergoes channel compression through a 1×1 convolution and then through a GELU activation layer and a 3×3 BSConv. After that, a skip connection is used to enhance global residual learning and fuse F_0 and F_m. This process can be formulated as:

$$\begin{aligned}
&F_k = f^k_{LKFB}(\cdots f^1_{LKFB}(F_0), \cdots), 1 \le k \le m, \\
&F_{fusion} = BSConv((GELU(Concat(F_1, \cdots, F_m))), \\
&F_{df} = F_{fusion} + F_0,
\end{aligned} \tag{9}$$

where $f^k_{LKFB}(\cdot)$ denotes the kth LKFB, m is the number of used LKFBs, F_k and F_{df} represent the output feature of the kth module and the final deep feature respectively. In the final image reconstruction stage, deep feature is transformed by a vanilla 3×3 convolution to a specific number of channels, and then the super-resolved image is obtained through pixel-shuffle operation [37]:

$$I_{SR} = PixelShuffle(Conv_{3\times3}(F_{df})). \tag{10}$$

4. Experiments

4.1. Datasets and Metrics

The training dataset consists of 800 images from DIV2K [1] and first 10K images from LSDIR [23]. Our evaluation of the models is performed on commonly used benchmark datasets, including Set5 [3], Set14 [47], B100 [32], Urban100 [17], and Manga109 [33]. The training data was augmented with random horizontal flips and 90-degree rotations. The evaluation metrics used are the av-

Table 1. Ablation study on frequency-enhanced pixel attention.

Method	Params[k]	Set5	Set14	B100	Urban100	Manga109
baseline	259	37.84 / 0.9601	33.61 / 0.9178	32.14 / 0.8994	32.09 / 0.9280	38.28 / 0.9767
baseline+LKA	308	37.95 / 0.9605	33.75 / 0.9187	32.22 / 0.9003	32.41 / 0.9311	38.76 / 0.9775
baseline+MDSA+ECCA	449	37.99 / 0.9606	37.80 / 0.9193	32.22 / 0.9004	32.52 / 0.9316	38.74 / 0.9775
baseline+FPA (LKFN)	291	37.88 / 0.9603	33.78 / 0.9189	32.21 / 0.9003	32.49 / 0.9315	38.72 / 0.9772

erage peak-signal-to-noise ratio (PSNR) and the structural similarity (SSIM) on the luminance (Y) channel.

4.2. Implementation Details

LKFN consists of 8 LKFBs with the feature channel number set to 56. The mini-batch size and input patch size for each LR input are set to 64 and 64×64, respectively. We train the model using the common $\mathcal{L}_1$ loss function and the Adan optimizer [44] with default settings. The initial learning rate is set to 5×10^{-3}. The learning rate decay is following cosine annealing with T_{max} = total iterations, $\eta_{min} = 1 \times 10^{-7}$. The total number of iterations is 1000K.

A mini version of our LKFN, called LKFN-S, was designed for the NTIRE 2024 Efficient SR Challenge [36]. It consists of 8 LKFBs and the feature channel is set to 28. We set the dilation ratio of the 5×5 depth-wise convolution to 3 in the third PLKB in LKFBs. The training process includes 2 stages: (1)Training with a input patch size of 64×64 and a mini-batch size of 64 from scratch by minimizing the $\mathcal{L}_1$ loss. The learning rate schedule is the same as the standard LKFN and the total number of iterations is 1000K. (2)Fine-tuning with a input patch size of 120×120 and a mini-batch size of 64 by minimizing the MSE loss. The learning rate is set to 2×10^{-5} during this stage. The total number of iterations is 150K.

We implement all our models using PyTorch 2.0.1 and a NVIDIA GeForce RTX 4090 GPU.

4.3. Ablation Study

Effectiveness of the FPA module. To verify the effectiveness of our FPA module and compare it with other attention modules, we use LKFN without the FPA module as the baseline, and compared with the attention mechanisms of LKA and MDSA+ECCA in two SOTA models LKDN and MDRN respectively. The results are shown in Tab 1. Obviously, the performance on each benchmark of the baseline is far behind the models with attention modules. Except for Set5, the improvement brought by our FPA module is significant. We think the local features play a more important role in Set5. On the other 4 benchmarks, comparing with the MDSA+ESA method, we achieved comparable performance with only 65% parameters. With slightly fewer parameters, we exceed the performance of LKA method, demonstrating the benefits of the frequency-domain global view. The local attribution maps (LAMs) [14] and diffusion indices (DIs) [14] results are shown in Fig.8. The first three models are based on obtaining attention maps in the spatial domain, which rely more on the surrounding pixels of the target. In the LAM of LKFN, besides the red-boxed region, the pixels of the entire image have almost the same contribution with a slight red color tone. This validates that the attention map obtained from the frequency domain has a global view.

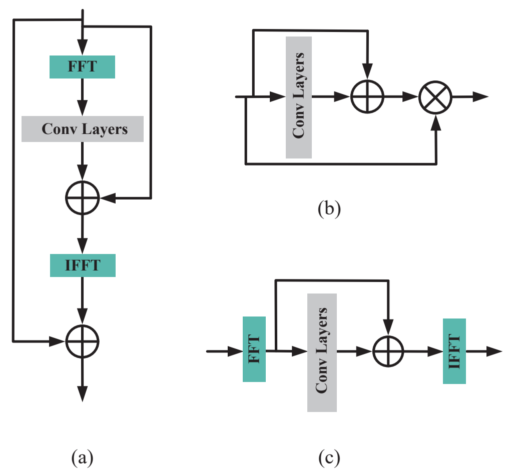

Figure 7. (a) non-attention (add) (b) spatial-enhancement (c) non-attention.

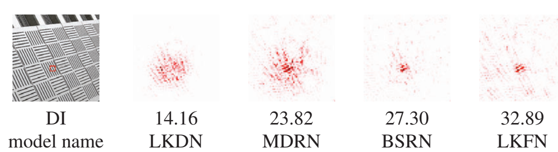

Figure 8. Comparison of LAM and DI results between LKFN and other efficient SR models. The LAM results denote the importance of each pixel in the input LR image when super-resolving the patch marked with a red box. The DI value reflects the range of involved pixels. A larger DI value means a wider range of attention.

Study of design in FPA. During the development of the FPA module, we tried other possibilities. Fig. 7(c) does not use an attention mechanism at all. Fig.7(a) does not use an attention mechanism but instead adds the feature maps. Fig.7(b) completely abandons frequency-domain process-

Table 2. Ablation study on different FPA design.

Method	Set5	Set14	B100	Urban100	Manga109
non-attention	37.86 / 0.9602	33.66 / 0.9180	32.16 / 0.8996	32.25 / 0.9292	38.60 / 0.9771
non-attention (add)	37.88 / 0.9603	33.74 / 0.9187	32.17 / 0.8998	32.33 / 0.9299	38.66 / 0.9773
spatial-enhancement	37.89 / 0.9603	33.73 / 0.9185	32.18 / 0.8998	32.21 / 0.9291	38.42 / 0.9769
standard FPA	37.88 / 0.9603	33.78 / 0.9189	32.21 / 0.9003	32.49 / 0.9315	38.72 / 0.9772

Table 3. Ablation study on PLKB.

Method	Params[k]	Set5	Set14	B100	Urban100	Manga109
BSRB	305	32.24 / 0.8960	28.67 / 0.7832	27.61 / 0.7375	26.22 / 0.7897	30.69 / 0.9106
RBSB	305	32.29 / 0.8963	28.70 / 0.7837	27.63 / 0.7377	26.27 / 0.7908	30.76 / 0.9113
PLKB	309	32.34 / 0.8963	28.71 / 0.7836	27.65 / 0.7385	26.35 / 0.7930	30.80 / 0.9119

ing and instead processes in the spatial domain. Their experimental results are shown in Tab.2. Among the three frequency-domain designs, the final FPA performs the best, followed by replacing dot product with residual connections (non-attention (add)), and the worst is the non-attention design. Each improvement in the three methods results in significant improvement, especially on Urban100. However, it's worth noting that there is almost no improvement on the Set5 dataset. On the other hand, abandoning the use of frequency-domain processing and using spatial convolution with a 1×1 kernel to obtain attention maps performs worse than the frequency-domain methods, especially on Urban100 where there are many repeated patterns (e.g. glass curtain buildings, tiled surfaces with the same pattern), and the pixel details lost due to downsampling have a high probability of being preserved in distant positions in the image. Therefore, for our frequency-domain method based on global attention, we can effectively capture such long-range dependencies. Whereas in Set5, which contains three face images and two natural images, repairing such images relies more on local details, and the small receptive nature of spatial convolutions becomes an advantage.

Effectiveness of the PLKB. To verify the advantages of using a large convolution kernel in deep feature extraction, we replace the PLKB in LKFN with the BSRB in BSRN and the RBSB in LKDN. The results are shown in Tab 3. BSRB only adds a residual connection to BSConv, and RBSB uses reparameterization to further improve deep feature extraction. In contrast, our PLKB fully utilizes the advantages of large kernel convolution while controlling the increase in parameters to only about 1%. This method improves the performance on Urban100 by 0.08dB.

4.4. Comparison with State-of-the-art Methods

We compare our LKFN with several state-of-the-art efficient super-resolution models on $2\times$, $3\times$, and $4\times$ scales, and the quantitative results are shown in Tab 4. As we just analyzed, our method stands out on Urban100. We also made some interesting findings when considering the results across different scale factors. As scale factor decreases, our method's leading advantage on Urban100 gradually increases. We believe that obtaining attention maps from the spatial domain always involves defining the kernel size, stride, dilation rate of the convolution kernels and pooling layer size (if exist) in the attention module in advance. For convenience, we usually optimize and decide the structural hyperparameters of the model only on one scale (commonly $4\times$) during the model development stage and then apply them directly to other scales. This leads to the optimal structure at one scale factor not necessarily being optimal at other scale factors. In contrast, our method uses a 1×1 convolution uniformly after Fourier transform processing, making it more adaptable and flexible in handling different scale factors. Qualitative comparisons on visual results can be found in Fig. 9, where it can be clearly observed that our method exhibits the best performance for this type of repeated pattern structure.

4.5. NTIRE 2024 Efficient SR Challenge

The aim of this challenge [36] is to devise a network that reduces one or several metrics such as runtime, parameters, and FLOPs of the baseline RLFN [20], while maintaining PSNR of around 26.90 dB on the DIV2K_LSDIR_valid dataset, and 26.99 dB on the DIV2K_LSDIR_test dataset.

Our solution, LKFN-S, for the NTIRE 2024 Efficient SR Challenge has proven to be both efficient and effective for super-resolution tasks, achieving competitive performance with just 90K parameters and 5.81G FLOPs for SR ×4. We won the 3rd place and 4th place in the Parameters sub-track and FLOPs sub-track, respectively.

5. Conclusion

In this paper, we propose the large kernel frequency-enhanced network (LKFN) that adopts the framework design of LKDN. We directly introduce large kernel convolution into the deep feature extraction module and combine it with partial convolution to better preserve information brought by large receptive fields from different levels, while effectively controlling model complexity. We also propose a frequency-domain-based pixel attention mechanism. It

Table 4. Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods, and multiply-accumulate operations is evaluated on a 1280 × 720 HQ image. The best and second-best performance are in red and blue colors, respectively.

Method	Scale	Params[K]	Multi-Adds[G]	Set5	Set14	B100	Urban100	Manga109
IMDN	×2	694	158.8	38.00 / 0.9605	33.63 / 0.9177	32.19 / 0.8996	32.17 / 0.9283	38.88 / 0.9774
PAN [51]		261	70.5	38.00 / 0.9605	33.59 / 0.9181	32.18 / 0.8997	32.01 / 0.9273	38.70 / 0.9773
RLFN [20]		527	115.4	38.07 / 0.9607	33.72 / 0.9187	32.22 / 0.9000	32.33 / 0.9299	-
FMEN [12]		748	172.0	38.10 / 0.9609	33.75 / 0.9192	32.26 / 0.9007	32.41 / 0.9311	38.95 / 0.9778
BSRN [25]		332	73.0	38.10 / 0.9610	33.74 / 0.9193	32.24 / 0.9006	32.34 / 0.9303	39.14 / 0.9782
VAPSR [52]		329	74.0	38.08 / 0.9612	33.77 / 0.9195	32.27 / 0.9011	32.45 / 0.9316	-
SAFMN [40]		228	52.0	38.00 / 0.9605	33.54 / 0.9177	32.16 / 0.8995	31.84 / 0.9256	38.71 / 0.9771
LKDN [43]		304	69.1	38.12 / 0.9611	33.90 / 0.9202	32.27 / 0.9010	32.53 / 0.9322	39.19 / 0.9784
MDRN [31]		304	65.0	38.11 / 0.9610	33.84 / 0.9205	32.32 / 0.9016	32.84 / 0.9350	39.14 / 0.9782
LKFN(ours)		291	66.6	38.06 / 0.9609	34.00 / 0.9207	32.28 / 0.9011	32.92 / 0.9350	39.12 / 0.9779
IMDN	×3	703	71.5	34.36 / 0.9270	30.32 / 0.8417	29.09 / 0.8046	28.17 / 0.8519	33.61 / 0.9445
PAN		261	39.0	34.40 / 0.9271	30.36 / 0.8423	29.11 / 0.8050	28.11 / 0.8511	33.61 / 0.9448
RFDN[26]		541	42.2	34.41 / 0.9273	30.34 / 0.8420	29.09 / 0.8050	28.21 / 0.8525	33.67 / 0.9449
FMEN		757	77.2	34.45 / 0.9275	30.40 / 0.8435	29.17 / 0.8063	28.33 / 0.8562	33.86 / 0.9462
BSRN		340	33.3	34.46 / 0.9277	30.47 / 0.8449	29.18 / 0.8068	28.39 / 0.8567	34.05 / 0.9471
VAPSR		337	33.6	34.52 / 0.9284	30.53 / 0.8452	29.19 / 0.8077	28.43 / 0.8583	-
SAFMN		233	23.0	34.34 / 0.9267	30.33 / 0.8418	29.08 / 0.8048	27.95 / 0.8474	33.52 / 0.9437
LKDN		311	31.4	34.54 / 0.9285	30.52 / 0.8455	29.21 / 0.8078	28.50 / 0.8601	34.08 / 0.9475
MDRN		311	29.6	34.58 / 0.9286	30.51 / 0.8453	29.21 / 0.8081	28.70 / 0.8627	34.07 / 0.9476
LKFN(ours)		299	30.3	34.54 / 0.9284	30.54 / 0.8452	29.19 / 0.8079	28.74 / 0.8629	34.09 / 0.9476
IMDN	×4	715	40.9	32.21 / 0.8948	28.58 / 0.7811	27.56 / 0.7353	26.04 / 0.7838	30.45 / 0.9075
PAN		272	28.2	32.13 / 0.8948	28.61 / 0.7822	27.59 / 0.7363	26.11 / 0.7854	30.51 / 0.9095
RLFN		543	29.8	32.24 / 0.8952	28.62 / 0.7813	27.60 / 0.7364	26.17 / 0.7877	-
FMEN		769	44.2	32.24 / 0.8952	28.70 / 0.7839	27.63 / 0.7379	26.28 / 0.7908	30.70 / 0.9107
BSRN		352	19.4	32.35 / 0.8966	28.73 / 0.7847	27.65 / 0.7387	26.27 / 0.7908	30.84 / 0.9123
VAPSR		342	19.5	32.38 / 0.8978	28.77 / 0.7852	27.68 / 0.7398	26.35 / 0.7941	30.89 / 0.9132
SAFMN		240	14.0	32.18 / 0.8948	28.60 / 0.7813	27.58 / 0.7359	25.97 / 0.7809	30.43 / 0.9063
LKDN		322	18.3	32.39 / 0.8979	28.79 / 0.7859	27.69 / 0.7402	26.42 / 0.7965	30.97 / 0.9140
MDRN		322	17.3	32.35 / 0.8970	28.80 / 0.7861	27.69 / 0.7404	26.60 / 0.8005	31.02 / 0.9146
LKFN(ours)		309	17.7	32.35 / 0.8971	28.80 / 0.7862	27.67 / 0.7400	26.60 / 0.8001	30.99 / 0.9140

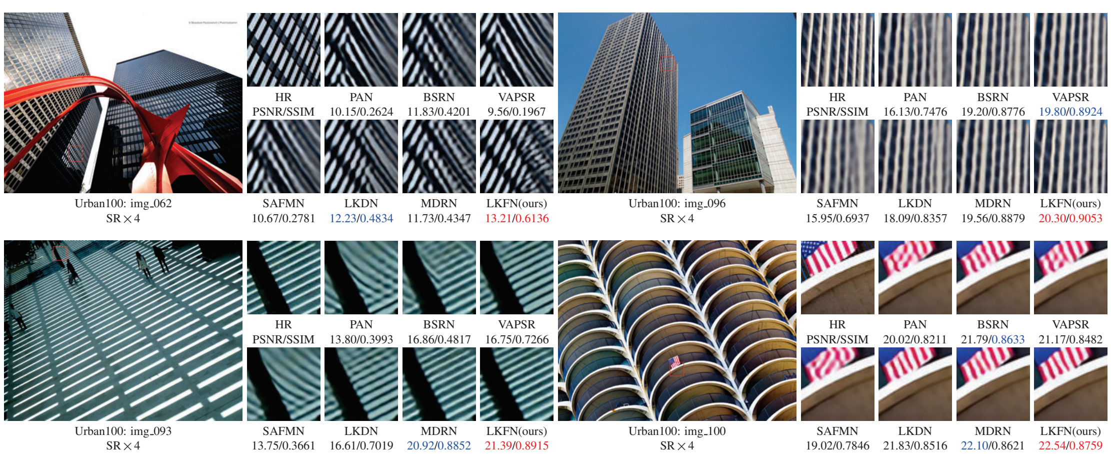

Figure 9. Visual comparisons for ×4 SR on the Urban100 dataset. The patches for comparison are marked with red boxes in the original images. PSNR/SSIM is calculated based on the patches to better reflect the performance difference, the best and second best are in red and blue respectively.

not only has a simple and compact structure but can truly achieve a global receptive field, improving the quality of attention maps. Through comparisons with other methods and rigorous analysis, our LKFN achieves SOTA in terms of parameters, Multi-Adds operations, and model performance, while achieving a balance between performance and complexity. In addition, a variant of our LKFN, LKFN-S, participated in the NTIRE 2024 efficient super-resolution challenge and won the third place in the FLOPs sub-track.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 5

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, pages 252–268, 2018. 2

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5

[4] Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don't walk: Chasing higher flops for faster neural networks. In *CVPR*, pages 12021–12031, 2023. 4

[5] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *NeurIPS*, 33:4479–4488, 2020. 3, 4

[6] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, 2021. 1

[7] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, pages 11963–11975, 2022. 2

[8] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. *arXiv preprint arXiv:2311.15599*, 2023. 2

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 1, 2

[10] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, pages 391–407. Springer, 2016. 2

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[12] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In *CVPR*, pages 853–862, 2022. 8

[13] Hao Feng, Liejun Wang, Yongming Li, and Anyu Du. Lkasr: Large kernel attention for lightweight image super-resolution. *Knowledge-Based Systems*, 252:109376, 2022. 2

[14] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, pages 9199–9208, 2021. 6

[15] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 2

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 5

[18] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, pages 2024–2032, 2019. 1

[19] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016. 2

[20] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPRW*, pages 766–776, 2022. 7, 8

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012. 2

[22] Junxuan Li, Shaodi You, and Antonio Robles-Kelly. A frequency domain neural network for fast image super-resolution. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. 3

[23] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *CVPRW*, 2023. 5

[24] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *CVPRW*, pages 1921–1959, 2023. 2

[25] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *CVPRW*, pages 833–843, 2022. 1, 5, 8

[26] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCVW*, pages 41–55, 2020. 1, 8

[27] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, pages 2359–2368, 2020. 2

[28] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022. 2

[29] 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*, pages 10012–10022, 2021. 2, 3

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 2

[31] Yanyu Mao, Nihao Zhang, Qian Wang, Bendu Bai, Wanying Bai, Haonan Fang, Peng Liu, Mingyue Li, and Shengbo Yan. Multi-level dispersion residual network for efficient image super-resolution. In *CVPRW*, pages 1660–1669, 2023. 2, 8

[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*, pages 416–423. IEEE, 2001. 5

[33] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications*, 76:21811–21838, 2017. 5

[34] Badri N Patro, Vinay P Namboodiri, and Vijay Srinivas Agneeswaran. Spectformer: Frequency and attention is what you need in a vision transformer. *arXiv preprint arXiv:2304.06446*, 2023. 3

[35] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *NeurIPS*, 2021. 3

[36] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth ntire 2024 efficient super-resolution challenge report. In *CVPRW*, 2024. 6, 7

[37] 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*, pages 1874–1883, 2016. 2, 5

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[39] Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. *NeurIPS*, 35: 17314–17326, 2022. 3

[40] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *ICCV*, 2023. 3, 8

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[42] Yan Wang, Yusen Li, Gang Wang, and Xiaoguang Liu. Multi-scale attention network for single image super-resolution. *arXiv preprint arXiv:2209.14145*, 2022. 3

[43] Chengxing Xie, Xiaoming Zhang, Linze Li, Haiteng Meng, Tianlin Zhang, Tianrui Li, and Xiaole Zhao. Large kernel distillation network for efficient single image super-resolution. In *CVPRW*, pages 1283–1292, 2023. 1, 3, 4, 8

[44] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022. 6

[45] Shengke Xue, Wenyuan Qiu, Fan Liu, and Xinyu Jin. Faster image super-resolution by improved frequency-domain neural networks. *Signal, Image and Video Processing*, 14(2): 257–265, 2020. 3

[46] Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. In *CVPRW*, pages 1692–1701, 2023. 2

[47] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 5

[48] 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 preprint arXiv:2208.11247*, 2022. 3, 4

[49] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *ACM MM*, pages 4034–4043, 2021. 1

[50] Yulun Zhang, Huan Wang, Can Qin, and Yun Fu. Aligned structured sparsity learning for efficient image super-resolution. *NeurIPS*, 34:2695–2706, 2021. 2

[51] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *ECCVW*, pages 56–72. Springer, 2020. 8

[52] Lin Zhou, Haoming Cai, Jinjin Gu, Zheyuan Li, Yingqi Liu, Xiangyu Chen, Yu Qiao, and Chao Dong. Efficient image super-resolution using vast-receptive-field attention. In *ECCVW*, pages 256–272. Springer, 2022. 3, 5, 8