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

Jiadi Chen Huanhuan Long
Zhejiang Normal University, Jinhua, China
zjnucjd@163.com huanhuanlong@zjnu.edu.cn

1. Team details

Team name: ZHEstar
Team leader information:
Jiadi Chen(zjnucjd@163.com)
+86 150 2676 0436
Zhejiang Normal University,Jinhua,China
Rest of the team members: Huanhuan Long
Affiliation:
Zhejiang Normal University
User names/Entries:
jiadi-chen/LKFN
HHLzjnu
Best scoring in development phase: 26.93
Link:
https : / / github . com / ThediidehT / NTIRE2024_ESR_LKFN

2. Method details

2.1. General method description

We propose the Large Kernel Frequency-enhanced Network (LKFN). The overall architecture is in Fig.1. It is based on the powerful distillation network in BSRN [5] and LKDN [9]. We make modifications to its distillation block and replace the attention module with our more efficient and powerful frequency-enhanced pixel attention (FPA). In our proposed large kernel frequecy-enhanced block (LKFB), we replace the BSRB in BSRN with our partial large kernl block (PLKB) as shown in Fig .2. Inspired by PConv [2], PLKB first divides the input feature map into two halves in the channel dimension. One half undergoes a 5×5 depth-wise convolution (The third PLKB with a dilation rate of 3.), and the result is then channel-wise concatenated with the unprocessed other half. A 1×1 convolution is subsequently used to perform data exchange between these two parts. This approach can better preserve information brought by large receptive fields at different levels, while enjoying the lightweight effect of channel reduction. Last, a GELU [3] activation layer is applied. The output of each PLKB is concatenated after channel compression by a 1×1 convolution, and then input into our FPA module after another layer of channel compression. In the FPA module, we transform the spatial domain feature map to the frequency domain through Fourier transform, and then pass the frequency domain map through a three-layer 1×1 convolution, with an activation layer following each layer. The result is added to the initial frequency domain map via a residual connection to obtain the enhanced frequency domain attention map. Then it was transformed back to the spatial domain and multiplied by the input spatial feature map.

2.2. Training strategy

The proposed LKFN consists of 8 LKFBs and the feature channel is set to 28. The training data includes 800 images from DIV2K [1] and the first 10K images from LSDIR [4]. Follow LKDN, we use Adan Optimizer [10] in the whole process. The training process is as follows:

1. Training with a input patch size of 64×64 and a mini-batch size of 64 from scratch by minimizing the L1 loss. 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.

2. Finetuning 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.

2.3. Other details about our method

- The number of parameters is 90K, FLOPs 5.81G, GPU memory consumption 671.17M, number of activations 174.03M, average runtime on validation and test dataset is 11.54ms on RTX 4090 GPU.

- To our best knowledge, it is the first method based on a pure frequency domain attention mechanism in super-resolution task. Previous approaches based on the frequency domain (Fourier transform) mostly had poor

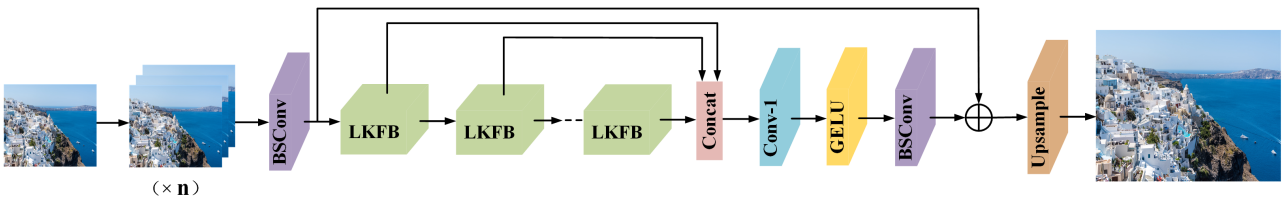

Figure 1. The architecture of large kernel frequency-enhanced network (LKFN).

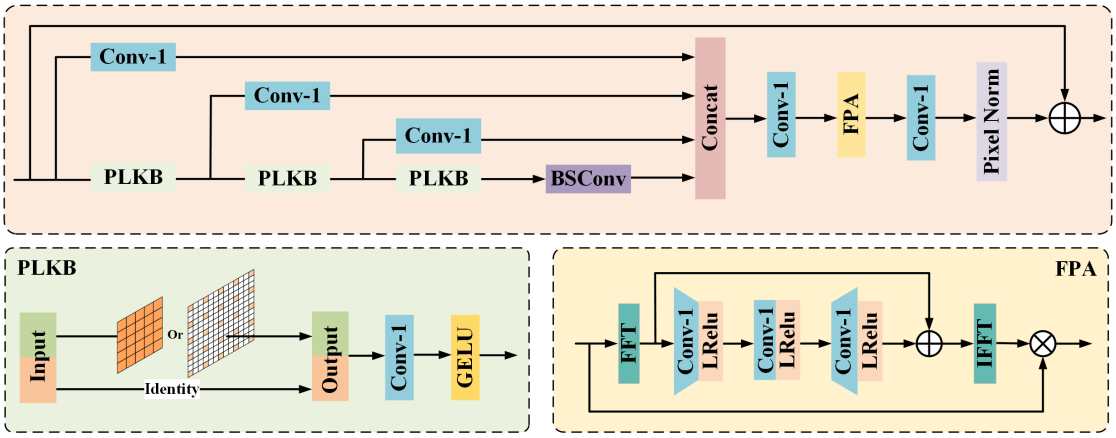

Figure 2. The large kernel frequency-enhanced block (LKFB)

performance, mainly due to improper feature map processing operations in the frequency domain that caused the phase map of the original image to be destroyed. This might not affect semantic information for high-level visual tasks [7], but for low-level vison like SR, it would lead to a significant degradation in the quality of the reconstructed image, with obvious square-shaped artifacts appearing in the image. Methods based on the Fourier transform in super-resolution are either used in the loss function [8], or integrated with traditional spatial domain methods [11], or the model parameters themselves are large enough to mask the negative effects of improper frequency domain operations. Our method reignites the possibility of studying super-resolution in the frequency domain.

- For an image with a resolution of a × b pixels, each pixel is composed of a × b different frequency components, and vice versa. Therefore, feature maps in the frequency domain have natural global properties. We believe that the multi-level fusion feature map obtained from the large kernel structure can achieve global cross-channel processing in its frequency domain, yielding global attention effects comparable to self-attention while maintaining lower complexity. It can better capture long-range pixel dependencies in low-level vison tasks.

- Comparing with LKDN [9] and MDRN [6] in the ESR competition last year, our LKFN outperforms these two methods by a large margin on the Set14 and Urban100 benchmarks in the case of $\times 2$ and $\times 3$ upscaling, and has a slight advantage in the $\times 4$ case, as shown in Tab.1. Both of these methods have attention mechanisms that obtain attention maps in the spatial domain, and their convolution kernel sizes and branch structures are fixed. Therefore, even if the model structure is optimal for $\times 4$ upscaling, the performance will be greatly reduced at different scales. However, our method is based on the frequency domain and is more flexible. It can still achieve overwhelming results in $\times 2$ and $\times 3$ cases. Moreover, our model is more lightweight with less than 300K parameters while still achieving state-of-the-art performance on various benchmarks.

3. Other details

- **Planned submission of a solution(s) description paper at NTIRE 2024 workshop.**

scale	params	Set14	Urban100
$\times 2$	291K	34.00 / 0.9207	32.92 / 0.9350
$\times 3$	299K	30.54 / 0.8453	28.73 / 0.8628
$\times 4$	309K	28.80 / 0.7862	26.60 / 0.8011

Table 1. Our LKFN performance on Set14 and Urban100

Our team's participating model comes from a yet-to-be-published paper of ours. We plan to include the content related to this ESR challenge in the paper after the competition ends, and submit it to the CVPR workshop before April 5th.

- **General comments and impressions of the NTIRE 2024 challenge.**

  I think this challenge series is well organized, and I have also followed the previous entries and related papers, which have inspired me a lot. The exciting results have promoted the development of ESR, which I think is great, and I hope this challenge can continue in the future.

- **What do you expect from a new challenge in image restoration, enhancement and manipulation?**

  Perhaps we can consider Scene Text Image Super-Resolution (STISR)? I am researching in the field of super-resolution, and perhaps in addition to general super-resolution, we can also consider some special-purpose super-resolution methods, such as STISR mentioned earlier, which can be used to enhance the effect of low-resolution input images in text recognition, in order to improve the accuracy of text recognition.

- **Other comments: encountered difficulties, fairness of the challenge, proposed subcategories, proposed eval uation method(s), etc.**

  So far everything seems fine, I don't have any suggestions.

References

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

[2] 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. 1

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

[4] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPR*, pages 1775–1787, 2023. 1

[5] 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

[6] 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

[7] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *NeurIPS*, 2021. 2

[8] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *ICCV*, 2023. 2

[9] 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, 2

[10] 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. 1

[11] 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. 2