ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230435

Efficient Information Modulation Network for Image Super-Resolution

Xiao Liu[a], Xiangyu Liao[a], Xiuya Shi[a], Linbo Qing[a] and Chao Ren[a;*]

[a]College of Electronics and Information Engineering, Sichuan University

Abstract. Recent researches have shown that the success of Transformers comes from their macro-level framework and advanced components, not just their self-attention (SA) mechanism. Comparable results can be obtained by replacing SA with spatial pooling, shifting, MLP, fourier transform and constant matrix, all of which have spatial information encoding capability like SA. In light of these findings, this work focuses on combining efficient spatial information encoding technology with superior macro architectures in Transformers. We rethink spatial convolution to achieve more efficient encoding of spatial features and dynamic modulation value representations by convolutional modulation techniques. The large-kernel convolution and Hadamard product are utilizated in the proposed Multi-orders Long-range convolutional modulation (MOLRCM) layer to imitate the implementation of SA. Moreover, MOLRCM layer also achieve long-range correlations and self-adaptation behavior, similar to SA, with linear complexity. On the other hand, we also address the sub-optimality of vanilla feed-forward networks (FFN) by introducing spatial awareness and locality, improving feature diversity, and regulating information flow between layers in the proposed Spatial Awareness Dynamic Feature Flow Modulation (SADFFM) layer. Experiment results show that our proposed efficient information modulation network (EIMN) performs better both quantitatively and qualitatively. Codes and supplementary materials link: https://github.com/liux520/EIMN.

1 Introduction

Single Image Super-Resolution (SISR), as a crucial low-level computer vision task, aims to reconstruct high-resolution (HR) clear images from their low-resolution (LR) counterparts. It is proved to be a significant part in image transformation tasks and has become increasingly important in both industry and academia fields. Over time, SISR has evolved from interpolation-based, reconstruction-based, learning-based to deep learning algorithms gradually.

Recently, the Transformer has attracted significant interest in the computer vision community, thanks to its powerful representation capabilities. However, several works have found that the excellence of Transformer comes from the macro-level framework and advanced components rather than its self-attention (SA) mechanism to some extent [38, 36, 12]. Surprisingly, comparable results on multiple mainstream computer vision tasks can still be obtained by replacing SA with spatial pooling [38], spatial shifting [36], spatial MLP [33, 32, 34, 13, 23], fourier transform [31, 17] and constant matrix [12], all of which have spatial information encoding capability similar to SA. This observation raises question about the true source of Transformer's superiority and highlights the importance of macro-level architectures and advanced training setups and components. These explorations encourage more researchers to rethink the design of Transformer framework by combining the superior macro architectures in Transformer with efficient spatial information encoding technology.

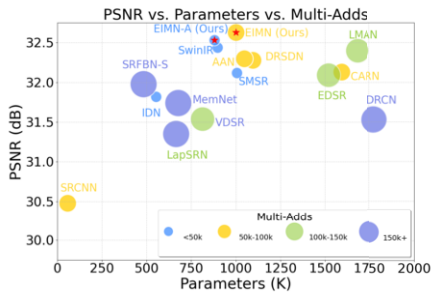

Figure 1: Trade-off between performance and model complexity on Set5 $\times 4$ dataset. Multi-Adds are calculated on 1280×720 HR images.

Following this motivation, our mainly purpose is to design a new spatial information encoder for encoding spatial features efficiently by introducing large kernel convolution operation and convolutional modulation technology to realize long-range correlations and self-adaptive behavior like SA, within the powerful Transformer macro framework. The basic idea is to replace the SA with the multi-order Long-range convolutional modulation (MOLRCM) layer and integrate the re-weighting process into the large kernel convolutional modulation technology for enhancing spatial information encoding ability, where extracted long-range and multi-order convolutional features act as weight matrices to self-adaptively modulate the value features. Our method is fully convolutional, enabling linear computational complexity rather than quadratically, which makes it more suitable for vision tasks, such as SISR. Further, building on MOLRCM and Spatial Awareness Dynamic Feature Flow Modulation (SADFFM), we propose the efficient information modulation network (EIMN) for SISR.

Our contributions can be summarized as follows: **(1)** We present a novel approach, named EIMN, to achieve efficient SISR that leverages the potential of large kernel ConvNets and advanced Transformer macro framework. Specifically, we introduce a Transformer-style ConvNet by replacing the SA and FFN with the proposed MOLRCM and DFFM layers for enhancing spatial- and channel-wise information encoding ability. **(2)** MOLRCM and SADFFM

* Corresponding Author. Email: chaoren@scu.edu.cn

modules are designed based on the analysis of the generation process of SA and the sub-optimality of vanilla FFN. The former utilizes large kernel convolution modulation technology to encode long-range and multi-order spatial information as a weight matrix, and self-adaptively realibrates value features. The latter introduces spatial awareness and locality, improves feature diversity, and dynamically regulates the flow of information between layers compared to vanilla FFN. **(3)** The experiments on five popular benchmark datasets demonstrate that our method performs better than other recently advanced transformer-based methods both in quantitative and qualitative results.

2 Related Work

2.1 Efficient SISR

Both ConvNets and Transformers have gained high performance in the SISR field. Although the results of SISR networks have been enhanced with the increase in depth and width, large architectures tend to be slow and power hungry. These excellent models require massive computational resources, memory and battery and not suitable for edge devices. Therefore, in order to overcome these issues, a number of efficient networks have been proposed to balance reconstruction performance and model complexity. A2N [3] consisted of a non-attention branch and a coupling attention branch, which fuse by the weights generated by the dynamic attention module. LMAN [35] exploited group convolution to extract and fuse multi-scale features before a channel attention layer to obtain discriminative features. SMSR [37] explored the sparsity in SISR to improve inference efficiency. DRSDN [4] proposed a plug-and-play NAS method to explore diverse architectures for SISR. These remarkable works have promoted the development of efficient SISR. However, our work rethinks the design of efficient SISR by introducing large kernel convolution and convolutional modulation techniques, thus making better utilization of spatial convolution, which remains a hot research topic.

2.2 Transformers

IPT was the first attempt to introduce Transformers-based backbone into the low-level image restoration, which is pre-trained with multi-task manner on ImageNet and finetuned to the desired task. Since then, it has been difficult to apply the SA mechanism to the SISR due to its quadratic computational cost. Therefore, there exist a host of works aiming to decrease complexity to make Transformer more suitable for vision tasks. *e.g.*, Swin Transformer[25] and SwinIR [19] limited SA calculation in non-overlapped local windows instead of global and introduced shift operation to perform cross-window interaction, which significantly reduces the computational complexity on HR feature map while capturing local context. Recently, it has been observed that the superiority of Transformer does not derive from the SA mechanism, but from its macro framework as well as advanced components [38, 36, 33, 32, 34, 13, 23]. Therefore, based on the macro framework and advanced components of Transformer architecture, our work effectively uses large kernel convolution to generate a modulation matrix to self-adaptively realibrate value features. This approach avoids the quadratic complexity of SA, enabling powerful inductive biases and highly parallel optimization of ConvNets.

2.3 Large Kernel Convolutional Modulation

The convolutional modulation technology can be viewed as an adaptive selection process that adaptively emphasizes important regions and suppresses irrelevant regions by mining the underlying relevance of feature representations sufficiently. The SA mechanism uses similarity score matrices to recalibrate the value representations, which is similar to the modulation technique. Although SA provides an intrinsic benefit in capturing long-range pixel inter-relationships, the 2D structure images are flattened into 1D sequences for inter-relationship learning, which not only compromises the integrity of the 2D images but also impairs image-specific neighborhood relationships. Recent studies have explored the advantages of incorporating large kernel convolutions, which can facilitate the construction of inter-pixel correlations by gathering responses from a larger region. VAN [11] proposed to decompose a large kernel convolution operation to capture long-range relationships. SegNeXt [10] decomposed large kernel convolutions to depth-wise strip convolutions for obtaining multi-scale context information and extracting strip-like features such as human and telephone poles. RepLKNet exhibited superior scalability for ConvNets with large kernels (up to 31×31). SLaK [24] studied extremely large kernels from the perspective of sparsity and proposed a pure CNN architecture equipped with sparse factorized 51×51 kernels which performs better than state-of-the-art (SOTA) hierarchical Transformers and modern ConvNet architectures. ConvNeXt [26] modernized a standard ResNet towards the design of a vision transformer and presented a powerful ConvNet with 7×7 depth-wise convolution.

3 Methodology

3.1 Overall Architecture

Recent investigations have demonstrated that the excellence of Transformers mainly results from the macro framework and advanced components [38, 36, 12]. Consequently, our structure is based on Transformer framework and simulates the execution of resource-consuming SA to propose MOLRCM layer with linear complexity. The pipeline of our method is shown in Figure 2, which consists of: (1) feature extraction, (2) nonlinear mapping, and (3) image reconstruction. The input and output of the model are illustrated as $\boldsymbol{I_{LR}}$ and $\boldsymbol{I_{SR}}$.

Feature extraction: Coarse features are extracted from low-resolution image I_{LR} by a 3×3 convolution layer.

$$\boldsymbol{F_{shallow}} = f_{ext}(\boldsymbol{I_{LR}}) \tag{1}$$

where f_{ext} indicates the convolution operation for coarse features extraction, $F_{shallow}$ is the extracted coarsed shallow features.

Nonlinear mapping: The extracted shallow features are flowed to a stack of EIMB for refining the feature mappings. Each of these blocks can be decomposed into two components: MOLRCM layer $MOLRCM(\cdot)$ and SADFFM layer $SADFFM(\cdot)$,

$$\boldsymbol{F_{deep}} = f^{n}_{EIMB}(f^{n-1}_{EIMB}(\cdots f^{0}_{EIMB}(\boldsymbol{F_{shallow}})\cdots)) \tag{2}$$

$$\boldsymbol{X}^{'} = \boldsymbol{X} + MOLRCM(Norm(\boldsymbol{X})) \tag{3}$$

$$\boldsymbol{Y} = \boldsymbol{X}^{'} + SADFFM(Norm(\boldsymbol{X}^{'})) \tag{4}$$

where f_{EIMB} and $\boldsymbol{F_{deep}}$ indicate the building block and output feature mappings, $Norm(\cdot)$ represents a normalization operation, *e.g.*, LayerNorm (LN), BatchNorm (BN). Notably, the proposed EIMB in our work shares similar macro architecture with Transformer: $LN \longrightarrow SA \longrightarrow LN \longrightarrow FFN$. The difference is that we replace SA and FFN with the proposed MOLRCM layer and DFFM layer for enhancing spatial and channel information encoding capabilities, respectively.

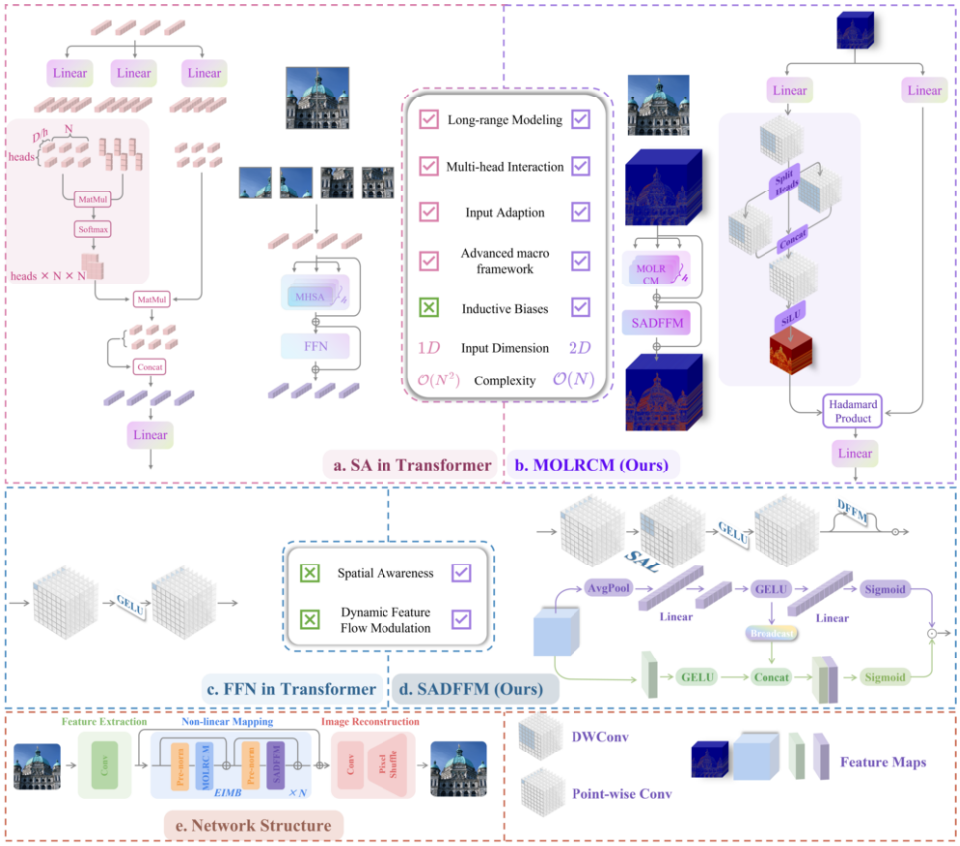

Figure 2: (a) and (b): The comparison of the SA and our MOLRCM. MOLRCM reconsiders spatial convolution to achieve efficient modeling of spatial features, enabling share similar advantages like SA. (c) and (d): The comparison of the FFN and ours SADFFM. SADFFM address the sub-optimality of FFN by introducing spatial awareness and locality, improving feature diversity, and regulating information flow between layers. (e) The architecture of the proposed PMSDSEN. It mainly consists of three parts: feature extraction, non-linear mapping and reconstruction.

Image reconstruction: Refined features are delivered to reconstruction layer for upsampling to the HR size. It is denoted as:

$$\boldsymbol{I_{SR}} = f_{rec}(\boldsymbol{F_{shallow}} + \boldsymbol{F_{deep}}) \tag{5}$$

where f_{rec} indicates the reconstruction module including a 3 × 3 convolution layer and a sub-pixel layer.

3.2 MOLRCM

We first analyze the execution process of SA in detail and then propose the MOLRCM layer to imitate and replace this process in a linear complexity manner. The main technique employed in the MOLRCM layer is multi-order long-range convolutional modulation operation which effectively integrates the modeling of large-range spatial relationships and the features re-weighting process.

3.2.1 The Limitations of SA

★ **Details of SA.** Firstly, let's revisit the vanilla SA. In SA, we suppose that the input token is $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, where N and d represent the number of tokens and dimensionality, respectively.

- **step 1:** The input token $\boldsymbol{X}$ is linearly embeded to queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ by weight matrices $\boldsymbol{W}_q \in \mathbb{R}^{d\times d_q}$, $\boldsymbol{W}_k \in \mathbb{R}^{d\times d_k}$ and $\boldsymbol{W}_v \in \mathbb{R}^{d\times d_v}$, such that $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q$, $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k$ and $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v$ respectively. $\boldsymbol{Q}$ and $\boldsymbol{K}$ are used to compute attention scores $A(\boldsymbol{Q}, \boldsymbol{K})$ which determine the degree of attention that one is supposed to give other tokens when encoding the token in current position, where, $A(\boldsymbol{Q}, \boldsymbol{K}) = \boldsymbol{Q} \cdot \boldsymbol{K}^\top$.
- **step 2:** Normalization attention scores for train gradient stability and position encoding is also applied to incorporate the order of sequences. Here, processed scores are translated into the probabilities by softmax function. Finally, the output of SA could be obtained by multiplying the attention weights $A(\boldsymbol{Q}, \boldsymbol{K})$ with values $\boldsymbol{V}$, hence, the original SA operation Attention(·) is defined as:

$$\begin{aligned} Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= A(\boldsymbol{Q}, \boldsymbol{K}) \cdot \boldsymbol{V} \\ &= Softmax(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^\top}{\sqrt{d_k}} + B) \cdot \boldsymbol{V} \end{aligned} \tag{6}$$

★ **Sub-optimality of SA.** By analyzing the execution process of SA, we consider that it has the following shortcomings:

(1) Quadratically complexity: The SA involves extensive interactions between query-key, followed by the aggregation of query-value. This process results in a quadratic growth of computational complexity with respect to sequence length N, as expressed by the $\mathcal{O}(2N^2D)$. As a result, SA can be computationally expensive and memory-intensive, especially for large inputs [19, 39].

(2) Damage in spatial structure: The process of converting 2D structure images into 1D squences for pairwise similarities calculation in SA leads to a damage of 2D image integrity and image-specific neighborhood relationships, which contains the resolution, sharpness, and color accuracy, as well as other distorted or modified of the images in an unintended way [11, 10].

(3) Weak modeling in local information: The SA operates on a sequence of inputs, where the attention weights are calculated based

on the pairwise interactions between all the elements in the sequence. Compared to convolutional operations, SA is indeed better at processing global information but weaker at handling local information. However, in the field of SISR, both local and global information are important to help models better understand and reconstruct HR images. For instance, local information can help the model better understand the structure and features of the image, and global information can help the model better understand the semantic and contextual relationships of the image [19, 39].

3.2.2 The Proposed MOLRCM layer

■ **Motivations and solutions.** The purpose of our proposed MOLRCM layer based on the spatial convolutional modulation technology is to solve the problem of SA as discussed above while preserving its merits including long-range spatial relationships and dependencies modeling and input content adaptation. Specifically, we rethink spatial convolution for taking advantage of convolution to achieve more efficient encoding of spatial features and realize dynamic modulation value representations by convolutional modulation techniques. The advantages of our method are discussed below:

(1) It is well-known that convolution operations can effectively capture local features because they have local connections with shared weights [39, 19, 25]. In addition, we also introduce large kernel convolution to encode long-range relationship dependencies to enable long-range relationship modeling similar to SA [11, 10, 6].

(2) The complexity of convolution is linear because the operation involves sliding a fixed-size kernel over the input feature and performing a dot product operation at each position. The complexity of convolution operation can be expressed as $\mathcal{O}(2K^2C^2HW)$, where K, C, H and W denote kernel size, the number of input/output channel, the height and width of feature map. Since the kernel size is usually small compared to the image size, the computational complexity is linear with respect to the input size.

(3) Since convolution processes two-dimensional images, it uses two-dimensional convolution operation that preserves local information in the input feature and enable to exploit the spatial relationships between neighboring pixels. This process can preserve the two-dimensional structure in the input image.

■ **Details of MOLRCM layer.** On the basis of the fundamental observation that SA is calculated by a matrix dot-product operation between query and key to generate attention scores, and then the recalibrated value representations are obtained via a weighted sum of all other positions. We propose that these two procedures can also be mimicked using large-kernel convolution and hadamard product. Specifically, large kernel can aggregate spatial contextual information from a larger region. Therefore, we employ depth-wise and depth-wise dilated convolutions with large kernel size k × k (k > 3) to generate context scores in a long-range and multi-order manner because of fewer parameters and computational burden, followed by hadamard product between the output of large-kernel convolutions and the value representations to re-weight value representations. In ConvNets, we prefer to describe these processes as convolutional modulation technology. By doing so, the proposed MOLRCM layer enables a flexible and effective modulation of the feature representation, promoting the modeling of complex image patterns with high adaptability and representational power. Specific details are as follows: As shown in Fig. 2b, given the input $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$, the latent space features $\boldsymbol{Q} \in \mathbb{R}^{H\times W\times C_q}$ and $\boldsymbol{V} \in \mathbb{R}^{H\times W\times C_v}$ can be generated by two linear project layers represented by matrices $\boldsymbol{W}_q \in \mathbb{R}^{C\times C_q}$ and $\boldsymbol{W}_v \in \mathbb{R}^{C\times C_v}$, respectively. Then, the generation of the attention weight matrices for the proposed convolution modulation operation consists of the following four steps:

- **step 1:** Regionality perception. In order to effectively extract features from coarse input features that typically exhibit local structure and spatial redundancy, it is necessary to introduce structural induction bias. To achieve this, a single depth-wise convolution (DW-Conv) is employed, utilizing a kernel size of 5×5 to extract more valuable features from the initial coarse features. This process can be expressed mathematically as a transformation of the input feature $\boldsymbol{F_{in}}$, allowing for subsequent feature refinement.

$$\boldsymbol{F_{region}} = \text{DW-Conv}(\boldsymbol{F_{in}}) \tag{7}$$

- **step 2:** Multi-order large-range contextual information extraction. In order to effectively capture large-range [7] and multi-order [30] contextual information for visual tasks, it is important to consider both local and global features. To achieve this, three parallel branches utilizing depth-wise dilation convolution (DW-D-Conv) are employed to implement multi-order interactions. Specifically, the low-order feature $\boldsymbol{F_{region}}$ is divided into three parts along the channel dimension, denoted as $\boldsymbol{F_l} \in \mathbb{R}^{H\times W\times C_l}$, $\boldsymbol{F_m} \in \mathbb{R}^{H\times W\times C_m}$, and $\boldsymbol{F_h} \in \mathbb{R}^{H\times W\times C_h}$, where $C_l + C_m + C_h = C$. Subsequently, DW-D-Conv$_{5\times5,d=2}$ and DW-D-Conv$_{7\times7,d=3}$ are applied to the features $\boldsymbol{F_l}$ and $\boldsymbol{F_h}$, with 5×5 and 7×7 denoting different kernel sizes and $d \in \{2, 3\}$ representing dilation ratios. Finally, the responses from multiple branches are concatenated to extract large-range and multi-order context information. This process can be represented mathematically using the following equation:

$$\begin{aligned}\boldsymbol{F_{multi}} = Concat(&\text{DW-D-Conv}_{5\times5,d=2}(\boldsymbol{F_l}), \boldsymbol{F_m}, \\ &\text{DW-D-Conv}_{7\times7,d=3}(\boldsymbol{F_h}))\end{aligned} \tag{8}$$

- **step 3:** Feature integration. The extracted high-quality large-range and multi-order representations are delivered into the last projection layer for two purposes: (1) cross-channel information integration and (2) estimating the importance of each point and generating attention weight. This is achieved using a standard convolution operation with kernel size 1×1.

$$\boldsymbol{F_{integration}} = Conv_{1\times1}(\boldsymbol{F_{multi}}) \tag{9}$$

- **step 4:** Gating activation. Following the design paradigm of common attention techniques [14], a gating mechanism is implemented to capture long-range spatial statistical characteristics from the aggregated information. For this purpose, we employ Sigmoid Linear Unit (SiLU) gating, which is an advanced version of sigmoid and known for its self-stabilizing property. The gating function is defined as: $x \cdot Sigmoid(x)$. Finally, the last spatial convolution modulation weight can be generated, as described in following formula:

$$A(\boldsymbol{Q}) = SiLU(\boldsymbol{F_{integration}}) \tag{10}$$

At last, we obtain finally self-adaptively modulation output by Hadamard product between $A(\boldsymbol{Q})$ and values $\boldsymbol{V}$ and the implementation process of MOLRCM layer can be expressed as:

$$MOLRCM(\boldsymbol{Q}, \boldsymbol{V}) = A(\boldsymbol{Q}) \odot \boldsymbol{V} \tag{11}$$

3.2.3 Analysis and Conclusion

In summary, our MOLRCM layer presents comparable benefits to SA employed in Transformer models, while requiring only linear complexity.

(1) **Large-range spatial information modeling:** The incorporation of large kernel size convolution, such as 7×7 and 9×9, enables the aggregation of information from a larger spatial area in an efficient manner, facilitating the learning of distance spatial relationships and dependencies modeling. This approach bears similarity to the SwinIR [19] and Swin Transformer [25] methods, which perform spatial information encoding within a local window typically of size 7×7. Furthermore, our methodology introduces a potent inductive bias and obviates the need for positional embedding due to the unique advantages inherent to the convolution operation. In summary, our method enables effective long-range spatial relationships modeling and multi-order features interaction similar to that achieved by the Transformer method.

(2) **Input self-adaptation:** Following the modeling of long-range and multi-order feature relationships through large kernel convolution modulation technology, the extracted convolution features generated is utilized as an attention weight matrix to self-adaptive recalibrate value representations similar to SA in the Transformer. By mining the underlying relevance of its own feature representations sufficiently, high-scoring positions will be given adequatlye focus, and insignificant positions will be suppressed to the extent. This process enables the identification of crucial features and facilitates their effective utilization in subsequent processing steps. Through this mechanism, our method is able to adaptively compute representations that capture complex relationships from input elements.

(3) **Linear complexity:** Our method utilizes pure convolution operation to generate attention maps instead of SA. Following many classic ConvNets, this design choice makes our method fully convolutional, resulting in linear computational complexity with the input size. Compared SA, our method is more suitable for resource-limited devices and HR input images.

3.3 SADFFM

3.3.1 The Limitations of FFN

★ **Details of FFN.** As shown in the Figure 2(c), the FFN, the only non-linear unit in the vanilla Transformer, plays an important effect in channel aggregation. According to the observation, just two linear transform layers with channel expand or squeeze ratio r are used to implement channel aggregation in mainstream methods. The first layer expands the channel dimensions from C to rC, and the second layer projects high-dimensional features from rC back to C. A non-linear activation function is inserted between these two linear layers. These steps can be expressed as a formula:

$$FFN(\boldsymbol{X}) = f_{proj2}(\sigma(f_{proj1}(\boldsymbol{X}))) \tag{12}$$

where f_{proj1} and f_{proj2} indicate two linear transform layers. σ represents non-linear activation function, such as Gaussian Error Linear Unit (GELU).

★ **Sub-optimality of FFN.** We argue that the vanilla FFN suffers from sub-optimality in two key aspects. First, the FFN lacks locality awareness, meaning that it fails to explicitly model local patterns in the input features. As a result, information aggregation may only occur at individual positions, leading to a lack of feature interaction between adjacent pixels or regions. This limitation can be particularly problematic for tasks that require the modeling of spatial relationships, such as SISR. To overcome this limitation, incorporating spatial awareness layers in the FFN has been shown to be an effective solution since they can learn local features and spatial relationships. Second, the FFN suffers from channel redundancy when equipped with a large number of channels in the intermediate layers. This occurs when different channels within a layer carry similar or redundant information, leading to increased computational cost without improving performance.

3.3.2 The Proposed SADFFM layer

■ **Motivations and solutions.** For the above two issues, we design a new channel information encoder named SADFFM layer including spatial awareness layer (SAL) and dynamic feature flow modulation (DFFM) layer layer for introducing spatial awareness and locality, improving feature diversity, and dynamically regulating the flow of information between layers, as shown in Figure 2(d).

■ **Details of SADFFM layer.** For the above two issues, we design a new channel information encoder named SADFFM layer, as shown in Fig. 2d.

(1) Spatial awareness: The FFN lacks the ability to model local patterns and spatial relationships, which can be important for SISR. The inverted residual block (IRB) employs a depth-wise convolution between two linear transform layers. This allows for local information to be aggregated between nearby pixels on each channel. The proposed DFFM layer adopts the IRB's design paradigm by replacing the point-wise convolutional layers in the vanilla FFN with a combination of depthwise separable convolutions and excitation-and-squeeze modules, which reduces the number of parameters and computation while still capturing local patterns and structures.

(2) Dynamic feature flow modulation: We introduce a novel approach to explicitly model channel and spatial feature relationships in order to reduce channels redundancy in our SADFFM layer, which is named DFFM. As depicted in Fig. 2d, our modulation mechanism involves two branches, one for channel and the other for spatial feature relationship modeling. To model inter-channel relationships, we apply a global average pooling operation across spatial dimensions on the input feature $\boldsymbol{F} \in \mathbb{R}^{H \times W \times C}$ to obtain global context representations $\boldsymbol{F_c} \in \mathbb{R}^{1 \times 1 \times C}$, which are used to aggregate inter-channel relationships. The global context $\boldsymbol{F_c}$ then passes through squeeze and expansion layers with ratio r and a non-linear activation function to obtain global inter-channel relationships. Finally, sigmoid gating is used to compute the modulation output $\hat{\boldsymbol{A}}_{\boldsymbol{c}} \in \mathbb{R}^{1 \times 1 \times C}$. To model spatial feature relationships, we first squeeze the channel dimensions with ratio r by a linear layer followed by a non-linear activation function to obtain $\boldsymbol{F_s} \in \mathbb{R}^{H \times W \times \frac{C}{r}}$. Then, global representations $\boldsymbol{F_c}$ are broadcasted to the spatial branch for channel and spatial information fusion. The concatenated features in the channel and spatial levels, denoted as $\boldsymbol{F_{sc}} \in \mathbb{R}^{H \times W \times \frac{2C}{r}}$, are then flowed through a linear transform layer for further dimension compression, followed by a sigmoid gating. This results in $\hat{\boldsymbol{A}}_{\boldsymbol{s}} \in \mathbb{R}^{H \times W \times 1}$. Finally, the last DFFM is obtained by element-wise product: $\hat{\boldsymbol{A}_{\boldsymbol{c \cdot s}}} = \hat{\boldsymbol{A}}_{\boldsymbol{c}} \cdot \hat{\boldsymbol{A}}_{\boldsymbol{s}}$. Ultimately, the proposed SADFFM layer can be described as:

$$SADFFM(\boldsymbol{X}) = DFFM(f_2(\sigma(SAL(f_1(\boldsymbol{X}))))) \tag{13}$$

where f_1 and f_2 mean linear layer, respectively. σ indicates the non-linear activation function GELU. Notably, the proposed SADFFM layer not only realize the modulation adaptability in the spatial dimension but also in the channel dimension. These modifications enhance the efficiency and performance of the network in modeling both long-range dependencies and local information.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. Following previous works, the

Table 1: Quantitative comparison with SOTA methods on five popular benchmark datasets. **Red** text indicates the best and **blue** text indicates the second best PSNR/SSIM results, respectively. 'Multi-Adds' is calculated with a 1280 × 720 HR image.

Method	Scale	#Params(K)	Multi-Adds(G)	Set5	Set14	BSDS100	Urban100	Manga109
EDSR-baseline [20]	×2	1370	316	37.99/0.9604	33.57/0.9175	32.16/0.8994	31.98/0.9272	38.54/0.9769
SRFBN-S [18]	×2	282	574.4	37.78/0.9597	33.35/0.9156	32.00/0.8970	31.41/0.9207	38.06/0.9757
SMSR [37]	×2	985	131.6	38.00/0.9601	33.64/0.9179	32.17/0.8990	32.19/0.9284	38.76/0.9771
A2N [3]	×2	1036	247.5	38.06/0.9608	33.75/0.9194	32.22/0.9002	32.43/0.9311	38.87/0.9769
LMAN [35]	×2	1531	347.1	38.08/0.9608	33.80/0.9023	32.22/0.9001	32.42/0.9302	38.92/0.9772
SwinIR [19]	×2	878	195.6	38.14/0.9611	33.86/0.9206	32.31/0.9012	32.76/0.9340	39.12/0.9783
B-GSCN 10 [22]	×2	1490	343	38.04/0.9606	33.64/0.9182	32.19/0.8999	32.19/0.9293	38.64/0.9771
DRSDN [4]	×2	1055	243.1	38.06/0.9607	33.65/0.9189	32.23/0.9003	32.40/0.9308	-
FPNet [8]	×2	1615	-	38.13/0.9619	33.83/0.9198	32.29/0.9018	32.04/0.9278	-
PILN [29]	×2	580	-	38.08/0.9607	33.72/0.9181	32.23/0.9003	32.38/0.9306	38.92/0.9771
NGswin [5]	×2	998	140.4	38.05/0.9610	33.79/0.9199	32.27/0.9008	32.53/0.9324	38.97/0.9777
EIMN-A(Ours)	×2	860	186.3	**38.26/0.9619**	**34.12/0.9222**	**32.40/0.9034**	**33.15/0.9373**	**39.48/0.9788**
EIMN(Ours)	×2	981	212.7	**38.26/0.9620**	**34.14/0.9227**	**32.41/0.9034**	**33.23/0.9381**	**39.42/0.9786**
EDSR-baseline [20]	×3	1555	160	34.37/0.9270	30.28/0.8417	29.09/0.8052	28.15/0.8527	33.45/0.9439
SRFBN-S [18]	×3	375	686.4	34.20/0.9255	30.10/0.8372	28.96/0.8010	27.66/0.8415	33.02/0.9404
SMSR [37]	×3	993	67.8	34.40/0.9270	30.33/0.8412	29.10/0.8050	28.25/0.8536	33.68/0.9445
A2N [3]	×3	1036	1175	34.47/0.9279	30.44/0.8437	29.14/0.8059	28.41/0.8570	33.78/0.9458
LMAN [35]	×3	1718	173.8	34.56/0.9286	30.46/0.8439	29.17/0.8067	28.47/0.8576	34.00/0.9470
SwinIR [19]	×3	886	872	34.60/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624	33.98/0.9097
B-GSCN 10 [22]	×3	1510	154	34.30/0.9271	30.35/0.8425	29.11/0.8035	28.20/0.8535	33.54/0.9445
DRSDN [4]	×3	1071	109.8	34.48/0.9282	30.41/0.8445	29.17/0.8072	28.45/0.8589	-
FPNet [8]	×3	1615	-	34.48/0.9285	30.53/0.8454	29.20/0.8086	28.19/0.8534	-
PILN [29]	×3	588	-	34.39/0.9269	30.34/0.8415	29.08/0.8048	28.09/0.8500	33.68/0.9446
NGswin [5]	×3	1007	66.6	34.52/0.9282	30.53/0.8456	29.19/0.8078	28.52/0.8603	33.89/0.9470
EIMN-A(Ours)	×3	868	83.58	**34.70/0.9299**	**30.65/0.8481**	**29.31/0.8121**	**28.87/0.8660**	**34.45/0.9492**
EIMN(Ours)	×3	990	95.2	**34.76/0.9304**	**30.70/0.8490**	**29.33/0.8127**	**29.05/0.8698**	**34.60/0.9502**
EDSR-baseline [20]	×4	1518	114	32.09/0.8938	28.58/0.7813	27.57/0.7357	26.04/0.7849	30.35/0.9067
SRFBN-S [18]	×4	483	852.9	31.98/0.8923	28.45/0.7779	27.44/0.7313	25.71/0.7719	29.91/0.9008
SMSR [37]	×4	1006	41.6	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	30.54/0.9085
A2N [3]	×4	1047	72.4	3230/0.8966	28.71/0.7842	27.61/0.7374	26.27/0.7920	30.67/0.9110
LMAN [35]	×4	1673	122.0	32.40/0.8974	28.72/0.7842	27.66/0.7388	26.36/0.7934	30.84/0.9129
SwinIR [19]	×4	897	49.6	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
B-GSCN 10 [22]	×4	1530	88	32.18/0.8950	28.60/0.7821	27.59/0.7364	26.12/0.7872	30.50/0.9080
DRSDN [4]	×4	1095	63.1	32.28/0.8962	28.64/0.7836	27.64/0.7388	26.30/0.7933	-
FPNet [8]	×4	1615	-	32.32/0.8962	28.78/0.7856	27.66/0.7394	26.09/0.7850	-
PILN [29]	×4	600	-	32.22/0.8949	28.62/0.7813	27.59/0.7365	26.19/0.7878	30.54/0.9086
NGswin [5]	×4	1019	36.4	32.33/0.8963	28.78/0.7859	27.66/0.7396	26.45/0.7963	30.80/0.9128
EIMN-A(Ours)	×4	880	47.78	**32.53/0.8993**	**28.89/0.7882**	**27.79/0.7447**	**26.68/0.8027**	**31.22/0.9148**
EIMN(Ours)	×4	1002	54.37	**32.63/0.9008**	**28.94/0.7897**	**27.82/0.7458**	**26.88/0.8084**	**31.52/0.9183**

DF2K dataset containing 3450 images is utilized as the training images, including 2650 images from Flick2K[21] and 800 images from DIV2K[1]. During testing, five standard benchmark datasets: Set5[2], Set14[40], BSD100[27], Urban100[15] and Manga109[28] are used to evaluate our method. We evaluate the average peak-signal-to-noise ratio (PSNR) and the structural similarity (SSIM) on the luminance (Y) channel of YCbCr color space. More implementation details are described in the Table 2.

Table 2: Hyper-parameters of the training process.

Training Config	Settings
Dataset	DF2K (Flick2K [21]+DIV2K [1])
Random rotation	(90°, 180°, 270°)
Random flipping	Horizontal
Patch size	64×64
Batch size	16
Optimizer	Adam [16]
Base learning rate	$5e^{-4}$
Optimizer mementum	β_1=0.9, β_2=0.999
Weight decay	$1e^{-4}$
Learning rate schedule	Cosine decay
Learning rate bound	$1e^{-7}$
Loss function	L_1

4.2 Comparison with SOTA Methods

Quantitative Results. In Table 1, we compare the proposed method with recent SOTA efficiently SISR approaches for upscale factor of ×2, ×3 and ×4. Notably, SwinIR [19] and NGswin [5] is the recently advanced transformer-based method. Obviously, our approach achieves the best performance with comparable parameters and Multi-Adds. Specifically, we obtain 0.1∼0.47dB improvement on five benchamrk datasets respectively compared the second-best approach SwinIR [19] with lower complexity, which indicates replacing the SA with the proposed MOLRCM layer based on the multi-order large-range convolutional modulation technology lead to better results.

Qualitative Results. In Figure 3, we display the ×4 SR results visualization. For the images "img 062", our method reconstructs the clearest lattice, stripe and text patterns with the minimal blurry effects and artifacts compared to other methods, which validates the usefulness and effectiveness of our method. Take the image "img 062" as an example, only our method generates stripes with accurate direction and minimal blurry, while the other methods produce incorrect stripes and a large range blurring effects.

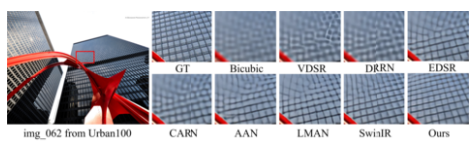

Figure 3: Qualitative comparison of SOTA methods on Urban100 (×4).

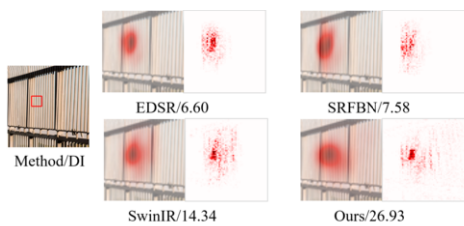

Figure 4: Results of Local Attribution Maps. A more widely red area and higher DI represent a larger range pixels utilization.

LAM Results. In Figure 4, we analyze local attribution maps (LAM) [9] results between AAN [3], EDSR [20], LMAN [35] and our method to investigate pixels utilization range in the input image when reconstructing the selected area. Diffusion index (DI), an eval-

Table 3: Ablation study on the subordinate components of the proposed SADFFM layer. The best performance is in red colors. Where, CA and SA denote general channel and spatial attention, DFFM denotes the proposed dynamic feature flow modulation, SAL denotes the propoesd spatial awareness layer.

Model	Componets				#Params(K)	Set5	Set14	BSDS100	Urban100	Manga109
	SAL	CA	SA	DFFM		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
FFN	✗	✗	✗	✗	896	32.00/0.8927	28.52/0.7795	27.54/0.7359	25.77/0.7729	30.14/0.9014
FFN+SAL	✔	✗	✗	✗	937	32.31/0.8972	28.75/0.7850	27.69/0.7413	26.28/0.7901	30.76/0.9097
FFN+SAL+CA	✔	✔	✗	✗	945	32.42/0.8985	28.84/0.7867	27.74/0.7427	26.46/0.7954	31.03/0.9132
FFN+SAL+SA	✔	✗	✔	✗	953	32.42/0.8986	28.84/0.7866	27.74/0.7427	26.46/0.7953	31.03/0.9133
SADFFM (Ours)	✔	✗	✗	✔	1002	32.63/0.9008	28.94/0.7897	27.82/0.7458	26.88/0.8084	31.52/0.9183

Table 4: Ablation study on the different layer sequences of the convolutional modulation technology within the proposed MOLRCM layer. The best performance is in red colors. where, k_1-k_2-k_3 represents the kernel size of $DWConv$-$DWDConv$-$DWDConv$ squence in the convolutional modulation module.

Layers Sequence	#Params(K)	Set5	Set14	BSDS100	Urban100	Manga109
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
3-3-5	968	32.42/0.8981	28.77/0.7861	27.70/0.7416	26.48/0.7967	31.00/0.9129
5-5-7(Ours)	1002	32.63/0.9008	28.94/0.7897	27.82/0.7458	26.88/0.8084	31.52/0.9183
7-7-9	1053	32.60/0.9002	28.91/0.7890	27.79/0.7450	26.82/0.8069	31.41/0.9177

uation metric tool, reflects the ability of model to extract feature and utilizate effective information. As shown in Fig. 5, our method uses larger range pixel information to reconstruct area drawed with a red box, which demonstrates our method attains a larger receptive field by an efficient large kernel size convolutional modulation manner.

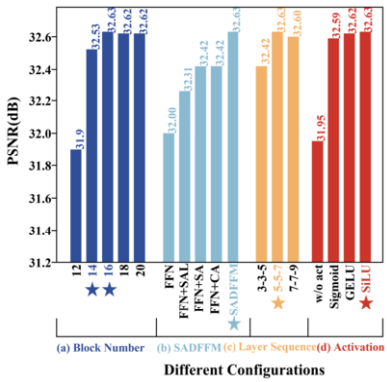

Figure 5: The general overview of the relationship between different configurations of ablation experiments and model performance: (1) Architecture configuration, which involves varying the number of building blocks in (a). (2) Subordinate components in the proposed SADFFM layer, such as spatial awareness (SAL) introduction and channel redundancy decrease technologies (SA, CA and DFFM) in (b). (3) Subordinate components in the MOLRCM layer based on the convolutional modulation technology, which include different choices of layer sequences and activation functions in (c) and (d). Notably, the asterisk indicates our method.

4.3 Ablation Study on Micro Design

In this section, we conduct ablation studies on some micro designs involved in our final proposed model. The micro designs consist of three parts: architecture configuration, subordinate components in MOLRCM layer based on the convolutional modulation technology, subordinate components in the SADFFM layer. We show their effects on the final performance. The general overview of the relationship between all configurations in ablation experiments and model performance can be observed in Figure 5, and the specific data are shown in Tables 3 and 4.

Architecture configuration. We first perform ablation experiments on the number of EIMBs in the nonlinear mapping part to search for the better balance between model complexity and performance, as shown in Figure 5(a). Experiment results indicate that the performance improvements with the increase in the number of stacked blocks until the highest value is reached at 16 blocks. Further increasing the number of blocks would lead to a slight decrease in network performance. Therefore, considering both model complexity and performance, we set the number of blocks to 16.

Subordinate components in the SADFFM layer. We conduct a detailed study on the impact of each component in the SADFFM layer, as presented in Figure 5(b). To compare the effectiveness of the proposed SADFFM layer, we also include the results of original FFN, which is frequently used in Transformer-style model. Remarkably, both DFFM and SAL achieve performance enhancements by a large margin, demonstrating the effectiveness of two modules. The specific comparison results on the five benchmarks are shown in the Table 3.

Subordinate components in MOLRCM layer. Finally, we also conduct a detailed research on the impact of each component in MOLRCM layer based on the convolutional modulation technology, as shown in Figure 5(c) and (d), including the choice of layer sequences and activation functions. Experiment results demonstrate the efficiency of our approach and the significant performance gains achieved through the carefully designed convolutional modulation module which consists of 5-5-7 layer squence for extracting large-range and multi-order contextual information and SiLU activation function that preserves the mean and variance of the input data and improves the learning process. The specific comparison results on the five benchmarks are shown in the Tables 4.

5 Conclusions

The primary focus of this work, named EIMN, is new information encoding techniques design for efficiently encoding spatial- and channel-wise features. For the design of MOLRCM layer, the extracted multi-order long-range convolutional features are utilized as weight matrices to self-adaptively re-calibrate the value representations, realizing efficient large-range spatial relationships modeling and multi-heads features interaction similar to SA with linear complexity rather than quadratically. For the design of SADFFM layer, we also address the sub-optimality of vanilla FFN in two aspects: locality absence and channel redundancy by introducing spatial awareness and locality, improving feature diversity, and dynamically regulating the flow of information between layers. The experiment results evaluated on five popular benchmarks demonstrate that our method performs better both quantitatively and qualitatively.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62171304 and the Cooperation Science and Technology Project of Sichuan University and Dazhou City under Grant 2022CDDZ-09.

References

[1] Eirikur Agustsson and Radu Timofte, 'Ntire 2017 challenge on single image super-resolution: Dataset and study', in *CVPRW*, pp. 126–135, (2017).

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

[3] Haoyu Chen, Jinjin Gu, and Zhi Zhang, 'Attention in attention network for image super-resolution', *arXiv preprint arXiv:2104.09497*, (2021).

[4] Guoan Cheng, Ai Matsune, Hao Du, XinZhi Liu, and Shu Zhan, 'Exploring more diverse network architectures for single image super-resolution', *KBS*, **235**, 107648, (2022).

[5] Haram Choi, Jeongmin Lee, and Jihoon Yang, 'N-gram in swin transformers for efficient lightweight image super-resolution', in *CVPR*, pp. 2071–2081, (2023).

[6] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding, 'Scaling up your kernels to 31x31: Revisiting large kernel design in cnns', in *CVPR*, pp. 11963–11975, (2022).

[7] 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).

[8] Alireza Esmaeilzehi, M Omair Ahmad, and MNS Swamy, 'Fpnet: A deep light-weight interpretable neural network using forward prediction filtering for efficient single image super resolution', *IEEE TCAS-II*, **69**(3), 1937–1941, (2021).

[9] Jinjin Gu and Chao Dong, 'Interpreting super-resolution networks with local attribution maps', in *CVPR*, pp. 9199–9208, (2021).

[10] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu, 'Segnext: Rethinking convolutional attention design for semantic segmentation', *arXiv preprint arXiv:2209.08575*, (2022).

[11] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu, 'Visual attention network', *arXiv preprint arXiv:2202.09741*, (2022).

[12] Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A Smith, and Roy Schwartz, 'How much does attention actually attend? questioning the importance of attention in pretrained transformers', *arXiv preprint arXiv:2211.03495*, (2022).

[13] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng, 'Vision permutator: A permutable mlp-like architecture for visual recognition', *IEEE TPAMI*, **45**(1), 1328–1334, (2022).

[14] Jie Hu, Li Shen, and Gang Sun, 'Squeeze-and-excitation networks', in *CVPR*, pp. 7132–7141, (2018).

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, 'Single image super-resolution from transformed self-exemplars', in *CVPR*, pp. 5197–5206, (2015).

[16] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[17] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon, 'Fnet: Mixing tokens with fourier transforms', *arXiv preprint arXiv:2105.03824*, (2021).

[18] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu, 'Feedback network for image super-resolution', in *CVPR*, pp. 3867–3876, (2019).

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, 'Swinir: Image restoration using swin transformer', in *ICCV*, pp. 1833–1844, (2021).

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, 'Enhanced deep residual networks for single image super-resolution', in *CVPRW*, pp. 136–144, (2017).

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, 'Enhanced deep residual networks for single image super-resolution', in *CVPRW*, pp. 136–144, (2017).

[22] Cong Liu and Pengcheng Lei, 'An efficient group skip-connecting network for image super-resolution', *KBS*, **222**, 107017, (2021).

[23] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le, 'Pay attention to mlps', *NeurIPS*, **34**, 9204–9215, (2021).

[24] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, 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).

[25] 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*, pp. 10012–10022, (2021).

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, 'A convnet for the 2020s', in *CVPR*, pp. 11976–11986, (2022).

[27] 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*, volume 2, pp. 416–423. IEEE, (2001).

[28] 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).

[29] Jiayi Qin, Lihui Chen, Seunggil Jeon, and Xiaomin Yang, 'Progressive interaction-learning network for lightweight single-image super-resolution in industrial applications', *IEEE TII*, **19**(2), 2183–2191, (2023).

[30] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser-Lam Lim, and Jiwen Lu, 'Hornet: Efficient high-order spatial interactions with recursive gated convolutions', *NeurIPS*, (2022).

[31] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou, 'Global filter networks for image classification', *NeurIPS*, **34**, 980–993, (2021).

[32] Long Sun, Jinshan Pan, and Jinhui Tang, 'Shufflemixer: An efficient convnet for image super-resolution', in *NeurIPS*, (2022).

[33] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al., 'Mlp-mixer: An all-mlp architecture for vision', *NeurIPS*, **34**, 24261–24272, (2021).

[34] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al., 'Resmlp: Feedforward networks for image classification with data-efficient training', *IEEE TPAMI*, (2022).

[35] Jin Wan, Hui Yin, Zhihao Liu, Aixin Chong, and Yanting Liu, 'Lightweight image super-resolution by multi-scale aggregation', *IEEE TBC*, **67**(2), 372–382, (2021).

[36] Guangting Wang, Yucheng Zhao, Chuanxin Tang, Chong Luo, and Wenjun Zeng, 'When shift operation meets vision transformer: An extremely simple alternative to attention mechanism', in *AAAI*, volume 36, pp. 2423–2430, (2022).

[37] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo, 'Exploring sparsity in image super-resolution for efficient inference', in *CVPR*, pp. 4917–4926, (2021).

[38] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan, 'Metaformer is actually what you need for vision', in *CVPR*, pp. 10819–10829, (2022).

[39] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, 'Restormer: Efficient transformer for high-resolution image restoration', in *CVPR*, pp. 5728–5739, (2022).

[40] 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*, pp. 711–730. Springer, (2012).