From Local Windows to Adaptive Candidates via Individualized Exploratory: Rethinking Attention for Image Super-Resolution

Chunyu Meng[1], Wei Long[1], Shuhang Gu[1]

[1]University of Electronic Science and Technology of China
{mengchunyu88, shuhanggu}@gmail.com

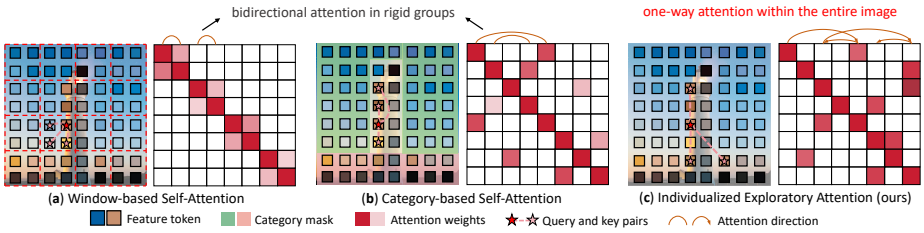

Fig. 1: (a) Window-based self-attention groups tokens according to spatial proximity and restricts attention candidates within each window, resulting in attention maps that are symmetric along the diagonal. (b) Category-based self-attention groups tokens based on coarse texture similarity, but still lacks flexibility due to predefined grouping, and similarly produces symmetric attention patterns. (c) Our proposed Individualized Exploratory Attention (IET) allows each token to adaptively and asymmetrically explore one-way neighbors, leading to inherently asymmetric attention maps.

Abstract. Single Image Super-Resolution (SISR) is a fundamental computer vision task that aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input. Transformer-based methods have achieved remarkable performance by modeling long-range dependencies in degraded images. However, their feature-intensive attention computation incurs high computational cost. To improve efficiency, most existing approaches partition images into fixed groups and restrict attention within each group. Such group-wise attention overlooks the inherent asymmetry in token similarities, thereby failing to enable flexible and token-adaptive attention computation. To address this limitation, we propose the Individualized Exploratory Transformer (IET), which introduces a novel Individualized Exploratory Attention (IEA) mechanism that allows each token to adaptively select its own content-aware and independent attention candidates. This token-adaptive and asymmetric design enables more precise information aggregation while maintaining computational efficiency. Extensive experiments on standard SR benchmarks demonstrate that IET achieves state-of-the-art performance under comparable computational complexity. The code is available at here.

1 Introduction

Image Super-Resolution (SR) aims to reconstruct clear and detailed high-resolution (HR) images from low-resolution (LR) inputs. It plays an important role in enhancing perceptual quality and providing reliable visual details for applications such as medical imaging, satellite observation, and video surveillance. However, since one LR image can correspond to multiple possible HR versions, SR remains an ill-posed and challenging task in computer vision.

Earlier SR methods mainly used Convolutional Neural Networks (CNNs) [9, 10, 14, 19, 24], which extract features through shared convolutional kernels, and have achieved promising reconstruction results. Recently, Transformer-based methods [4,23,44,47] have been increasingly applied to image reconstruction, employing a more complex self-attention mechanism that learns feature relationships adaptively instead of fixed kernels. This improvement allows Transformers to capture more flexible and diverse feature representations.

Although the self-attention mechanism in Transformers enables tokens to leverage information from others, its computational complexity grows quadratically with the number of tokens. To efficiently handle the large number of tokens in image data, most methods restrict self-attention computation so that each token interacts with only a limited subset of others. Window-based self-attention is the most common approaches used for this purpose, where an image is divided into local windows, and tokens attend only to the others within the same window [5, 26, 27]. This design focuses on local feature aggregation but limits the receptive field. To overcome this limitation, another line of research, known as category-based self-attention, defines attention calculation according to semantic content [25,43]. Specifically, these methods first categorize all tokens into several classes, then further divide each class into groups, and finally apply attention calculation within each group. Essentially, this strategy reorders the image by grouping semantically similar tokens together, allowing them to mutually assist in reconstruction. These methods consider image content when defining attention, but they still restrict computation within each group, preventing tokens from interacting with others in the same category. Moreover, the accuracy of category boundaries still needs further improvement, as it is difficult to partition tokens into sufficiently fine-grained classes. Significant variation often exists among tokens within the same category, while boundary tokens are unable to interact with more relevant tokens from other categories.

Overall, the above window-based and category-based self-attention methods both fall under grouped attention, which limits the flexibility of individual tokens as they neglect the asymmetric nature of similarity. In super-resolution, a token may need to pull information from several other tokens to recover fine structures, while those tokens do not need anything in return from the original one. If this directional relation is not modeled, information flow becomes redundant and the reconstruction is suboptimal. An ideal SR model should allow each token to flexibly search for its own one-way similar neighbors across a wide spatial range, forming individualized attention candidates that best support feature aggregation. By contrast, grouped attention imposes rigid group boundaries and

encourages symmetric aggregation within groups, which restricts the ability of a token to choose the sources it truly needs. Modeling asymmetric and individualized relations is thus crucial for restoring complex textures and preserving structures.

In this paper, we propose a novel Individualized Exploratory Attention (IEA) mechanism that connects adjacent attention blocks to progressively select content-aware and asymmetric one-way attention candidates, as shown in Fig. 1. We leverage the attention map from previous blocks to adjust the similarity relationships: if token A is similar to B, and B is similar to C in the preceding layer, it is very likely that A is similar to C in subsequent layers, and vice versa. Specifically, in the first block, each token attends only to nearby regions centered on itself. In subsequent blocks, a propagation mechanism allows connection to new long-range similar neighbors, thereby establishing more comprehensive similarity relationships. Meanwhile, a sparsification mechanism is applied to remove low-similarity tokens identified in earlier layers, maintaining an appropriate attention candidates. IEA operates as a dynamic process, starting from a local attention scope and progressively evolving into a content-aware and long-range one through layer-wise refinement. Under comparable computational budgets, our SR model can effectively identify suitable one-way attention candidates over a broader spatial range, leading to improved super-resolution quality.

The main contributions of this paper are as follows:

- We propose a Individualized Exploratory Attention (IEA) mechanism that enables content-aware and token-adaptive attention candidates selection. IEA progressively expands the attention candidates with new similar tokens while pruning low-similarity ones for efficiency.
- We develop the Individualized Exploratory Transformer (IET), which utilizes the IEA mechanism to effectively select high-quality attention candidates. As a result, IET can flexibly capture global dependencies without compromising computational efficiency.
- Extensive experiments on standard super-resolution benchmarks demonstrate that our method outperforms recent state-of-the-art approaches under comparable computational budgets, and ablation studies validate the effectiveness of each proposed component.

2 Related Work

Transformer-based SR. Over the past decade, deep learning has greatly advanced single-image super-resolution (SR). Starting from SRCNN [10], which introduced deep learning to SR with a simple three-layer CNN, numerous studies have explored various architectural enhancements to boost performance [8,13, 19, 20, 24, 31–33, 45, 46]. VDSR [19] deepened the architecture, while DRCN [20] adopted a recursive design. EDSR [24] and RDN [46] refined residual blocks, further enhancing CNN-based SR. Inspired by Transformers [36], attention mechanisms were introduced to visual tasks—Wang et al. [39] first integrated non-local attention into CNNs, demonstrating its effectiveness. Subsequent works, such as

CSNLN [32] and NLSA [31], leveraged cross-scale and sparse attention to better capture long-range dependencies while improving computational efficiency.

Subsequently, with the introduction of ViT [12] and its variants [7, 26, 38], the efficacy of pure Transformer-based models in image classification has been established. IPT [3] first introduces a pre-trained Transformer network to various image restoration tasks, demonstrating the strong representation capability of large-scale Transformers. Then, SwinIR [23] and CAT [5] employ window-based self-attention mechanisms to model local dependencies efficiently, where SwinIR utilizes shifted windows to aggregate features within a local area, and CAT further improves cross-window interaction through rectangular-window attention and axial shifting. Building on these ideas, HAT [4] enlarges the attention window and incorporates channel attention to extend the receptive field, while PFT [27] adopts sparse attention to further expand the interaction range without significantly increasing computational cost. Beyond window-based strategies, ATD [43] introduces category-based self-attention, which groups tokens with similar semantic characteristics to achieve content-aware feature aggregation. Recently, CATANet [25] further improves both the feature aggregation process and the handling of category boundaries. In addition, IPG [34] improves SR by prioritizing detail-rich pixels to enhance reconstruction. In this paper, building upon the effectiveness of the attention mechanism in image SR, we propose a IET method that lets tokens to independently explore and select most relevant neighbors from the global space through IEA, thereby enables content-aware and token-adaptive attention candidates selection with low computational cost.

Transformer with Sparse Attention. Sparse attention reduces computational complexity by restricting attention candidates. Existing methods typically partition feature maps into rigid groups: for example, SwinIR [23] groups tokens by spatial windows, ART [42] adopts dilated windows, NLSA [31] assigns tokens via hash buckets, and ATD [43] clusters them with learnable centers. In contrast, IEA enables token-adaptive attention candidates. Starting from local neighborhoods, it progressively expands the receptive field through similarity propagation, allowing each token to identify the most relevant neighbors and thereby achieving both accurate and efficient attention computation.

3 Methodology

3.1 Motivation

Self Attention. Self-attention [36], the core operation of Transformers, measures token similarity and aggregates features through attention weights. Given $Q, K, V \in \mathbb{R}^{N\times d}$, it can be expressed as

$$A_{sa} = \text{Softmax}\Big(QK^\top/\sqrt{d}\Big), \quad O_{sa} = A_{sa}V \tag{1}$$

where N and d denote the number and dimension of tokens respectively, $A_{sa} \in \mathbb{R}^{N\times N}$ denotes the attention map, and $O_{sa} \in \mathbb{R}^{N\times d}$ denotes the output. Despite its effectiveness, the quadratic computational complexity and redundant

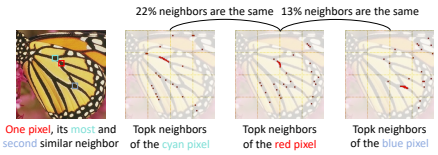

Fig. 2: Asymmetry and repetition of similarity relationships, the yellow grids represents 32×32 windows.

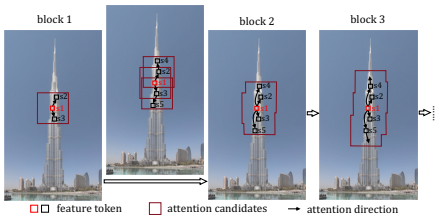

Fig. 3: The propagation mechanism.

interactions among dissimilar tokens significantly limit its efficiency and scalability. For efficiency, grouped attention is commonly adopted, which roughly clusters tokens into groups to improve super-resolution performance. However, it overlooks the asymmetric nature of similarity: token A may need to aggregate information from token B, while B does not necessarily require A. Therefore, an ideal SR model should adaptively capture such asymmetric relationships and avoid restricting each token within rigid groups.

Individualized Attention. We analyze token-wise similarity using the attention maps of global self-attention. As illustrated in Fig. 2, similar neighbors are sparsely distributed over a region far beyond local windows and do not follow fixed spatial priors. This observation is consistent with prior findings: window-based methods often yield suboptimal performance, while enlarging receptive fields via fixed dilated windows brings only limited gains. Motivated by this insight, we allow tokens to attend to any arbitrary k tokens across the entire feature map, reducing the complexity of global self-attention in Eq. 1 from $O(N^2)$ to $O(kN)$. We first introduce the concept of individualized attention, where each token maintains its own asymmetric and independent set of attention candidates. Specifically, individualized attention is implemented via an index matrix $I \in \mathbb{R}^{N\times k}$, which explicitly records the global indices of the k most similar neighbors for each token. This formulation yields a flexible attention mechanism: when I covers all tokens, it reduces to standard self-attention; when restricted to a local window, it becomes window-based self-attention. Formally, the individualized attention is proposed as in the following equation

$$A_{ia} = \text{Softmax}(\text{SMM}(Q, K, I)/\sqrt{d}), \tag{2}$$

$$O_{ia} = \text{SMM}(A_{ia}, V, I), \tag{3}$$

where SMM denotes sparse matrix multiplication, $A_{ia} \in \mathbb{R}^{N\times k}$ denotes the individualized attention map, and $O_{ia} \in \mathbb{R}^{N\times d}$ denotes the final output. In the individualized attention, the optimization of selecting attention candidates is explicitly simplified into the optimization of the index I.

Individualized Exploratory Attention. To assign token-adaptive and accurate attention candidates, we propose a propagation mechanism. We observe

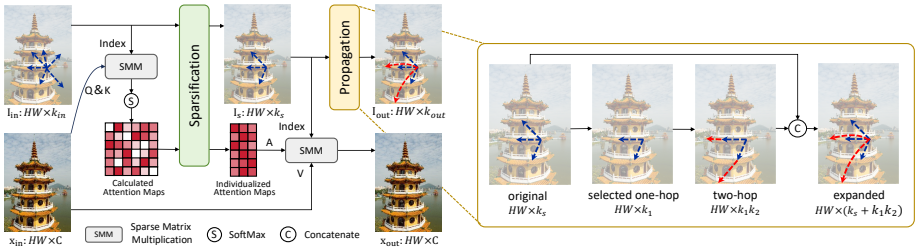

Fig. 4: The proposed individualized exploratory attention. We first apply the sparsification mechanism to prune neighbors with low similarity, and then employ the propagation mechanism to explore new two-hop neighbors.

that similar tokens tend to share overlapping neighborhoods, which enables efficient receptive field expansion by exploring higher-order connections in the similarity graph. As illustrated in Fig. 3, the initial attention candidates are confined to token-centered local neighborhoods. In subsequent blocks, the candidate set progressively propagates toward more correlated regions, resulting in a semantics-aware expansion of the receptive field and allowing each token to identify more suitable bases for reconstruction.

3.2 Individualized Exploratory Attention

In IEA, consecutive attention blocks are connected to enable progressive refinement of attention candidates. We propose a propagation mechanism that efficiently enlarges the receptive field by exploring higher-order neighbor relations. To further reduce computational cost, we introduce a sparsification strategy that prunes low-relevance connections and skips unnecessary attention computations. Specifically, the first block initializes with a locally restricted neighborhood, establishing a compact and computationally efficient foundation. In the following blocks, the propagation mechanism promotes second-hop neighbors identified in the previous layer to direct candidates, thereby broadening the attention coverage and capturing longer-range dependencies. Meanwhile, the sparsification mechanism removes connections that have been proved to exhibit low similarity, effectively enhancing the selectivity of attention computation. Through this iterative process of propagation and sparsification across layers, IEA gradually evolves into a content-aware and token-adaptive attention structure, achieving flexible feature aggregation under comparable computational budgets. The following subsections will introduce our attention candidates initialization methods, propagation and sparsification mechanisms in detail and explain how the attention candidates are optimized layer by layer.

Attention Candidates Initialization. As described above, IEA initializes a local attention scope in the first block, and progressively expands it in later blocks. A practical challenge is to ensure that every token discovers two-hop neighbors distinct from its one-hop ones. Following the principles of Expander Graphs, this property holds if each node has a unique neighborhood. To meet this requirement, we first consider to define a local attention scope centered on

each token. Although this initialization satisfies the requirement, the attention scope of adjacent tokens overlap significantly, leading to a decrease in propagation efficiency. To address this issue, we propose a dilation strategy, which performs dense attention computation within a small region while uniformly and sparsely sampling tokens from a larger area. For the distant region, one token are uniformly sampled from each $d \times d$ patch and added to the attention candidates, where d is a predefined dilation factor. This design enables tokens to capture detailed local information as well as coarse global context, thereby achieving a larger initial receptive field. The choice of d will be explored in ablation studies.

Our initialization strategy shares a superficial similarity with conventional sparse window designs in that it relies on spatial priors. However, unlike traditional window-based methods that treat such spatial grouping as the final attention structure, our design uses it only as a coarse starting point. Although this spatially driven dilation yields limited direct gains, it provides diverse initial candidates that are subsequently refined through semantics-aware propagation, forming a coarse-to-fine construction of token-adaptive attention.

Propagation and Sparsification. For the subsequent blocks, we introduce Propagation and Sparsification mechanisms to progressively refine the attention candidates. Specifically, given Q^l, $K^l \in \mathbb{R}^{N \times D}$ and the initial attention candidates $I^l_{\text{in}} \in \mathbb{R}^{N \times k^l_{in}}$, we can compute the individualized attention map $A^l_{\text{cal}} \in \mathbb{R}^{N \times k^l_{in}}$ according to the individualized attention mechanism as in Eq.2. Specifically, we first perform sparsification mechanism on A^l_{cal} to remove tokens with low attention scores, preserving only those strongly correlated with the query token. For token i, this process is achieved through the following formula:

$$\begin{cases} S_i = \text{TopK}(A^l_{\text{cal}}[i,:], k^l_{\text{s}}) \\ A^l_{\text{s}}[i,] = \{\, A^l_{\text{cal}}[i,j] \mid j \in S_i \,\} \\ I^l_{\text{s}}[i,] = \{\, I^l_{\text{in}}[i,j] \mid j \in S_i \,\} \end{cases} \tag{4}$$

where $\text{TopK}(\cdot)$ denotes the operation that selects the indices of the top-K elements with the highest values. Then, with A^l_{s}, $I^l_{\text{s}} \in \mathbb{R}^{N \times k^l_s}$, and $V^l \in \mathbb{R}^{N \times D}$, we can compute the output O^l as in Eq. 3. After that, our goal is to perform propagation mechanism on I^l_{s} and A^l_{s} to obtain more optimal candidates $I^l_{\text{out}} \in \mathbb{R}^{N \times k^l_{out}}$, which will then serve as the initial attention candidates I^{l+1}_{in} for the next block. Due to the high computational cost of exploring all two-hop neighbors, we adopt a simplified and efficient approximation strategy. For token i, we first select the top k^l_1 one-hop neighbors with the highest attention scores:

$$N^{(1)}_i = \{\, I^l_{\text{s}}[i,j] \mid j \in \text{TopK}(A^l_{\text{s}}[i,:], k^l_1) \,\} \tag{5}$$

where $N^{(1)}_i$ represents the one-hop neighbor set of the token i. Then for every selected one-hop neighbor $u \in N^{(1)}_i$, we still select its top k^l_2 one-hop neighbors:

$$N^{(1)}_u = \{\, I^l_{\text{s}}[u,v] \mid v \in \text{TopK}(A^l_{\text{s}}[u,:], k^l_2) \,\} \tag{6}$$

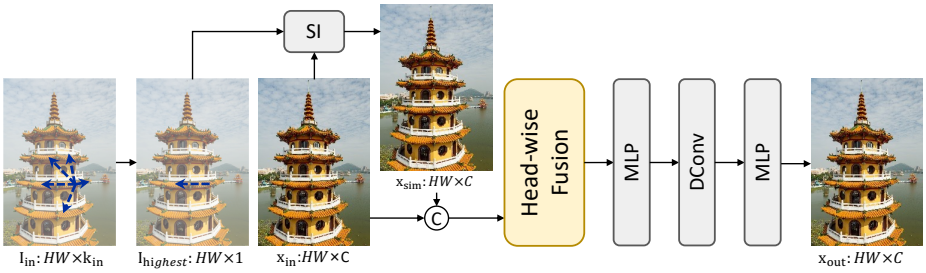

Fig. 5: The proposed Similarity-Fused Feed-Forward Network (SF-FFN), which enables cross-channel feature interaction among semantically similar tokens.

Next, we can union all the one-hop neighbors of token u to obtain the two-hop neighbors of token i, and finally merge the one- and two-hop neighbors as:

$$N_i^{(2)} = \cup_{u \in N_i^{(1)}} N_u^{(1)} \tag{7}$$

$$I_{\text{out}}^l = \text{Dedup}\left(I_{\text{s}}^l \cup N^{(2)}\right) \tag{8}$$

where $\cup(\cdot)$ denotes the union, and $\text{Dedup}(\cdot)$ denotes the removal of duplicated elements to form a unique set.

The proposed propagation mechanism operates under a locally optimal principle, enlarging the attention scope toward the directions of highest similarity at the current stage. This enables a more comprehensive exploration of potential attention candidates. Meanwhile, the sparsification mechanism prunes low-similarity neighbors, preserving the core relational structure while reducing computational complexity. The overall architecture of IEA is illustrated in Fig. 4. Notably, directly computing the sparse matrix multiplication (SMM) with dense matrix operations is highly inefficient. Thus we modify a CUDA-based sparse matrix multiplication framework [27] to suit our setting, enabling efficient individualized attention computation.

3.3 Similarity-Fused FFN

The Feed-Forward Network (FFN) is an important component of Transformer-based architectures, which is traditionally viewed as a token-wise nonlinear mapping. Then several works [22, 37, 40] introduce convolution operations into the FFN to enhance local feature fusion.

We further propose the Similarity-Fused Feed-Forward Network (SF-FFN), which exploits the refined token similarities obtained from IEA and enables the fusion of highly correlated tokens. As shown in Fig. 5, the SF-FFN selects the most similar neighbors I_{highest} for each token based on the updated neighbor indices I_{in}. It then replaces each token in the input feature map x_{in} with its most similar neighbor to obtain x_{sim}, and finally fuses the two representations. Specifically, we use a head-wise MLP and depthwise convolution as the fusion module, which introduces only a very small amount of extra computational cost.

From the perspective of the entire Transformer architecture, the SF-FFN provides an important complement to the attention module, whose capacity for cross-channel interaction is inherently limited. In particular, when IEA supplies more comprehensive and accurate similarity relationships, SF-FFN can effectively enhance information exchange among highly correlated tokens.

3.4 The Overall Network Architecture

The overall architecture of IET follows the design commonly used in recent state-of-the-art Transformer-based SR models [4,23,43]. However, we replace the standard self-attention blocks and feed-forward networks with our proposed IEA block and SF-FFN, respectively Notably, we apply relative positional encoding only in the first block, where the attention scope is still local. For the remaining layers, we utilize LePE positional encoding [11] directly instead. Moreover, a progressive attention mechanism [27] is incorporated into each block to refine and model similarity relationships with greater precision. For classical SR, the IET network consists of 8 blocks, each containing 4 attention layers. It adopts multi-head attention with 6 heads and uses a total of 240 channels. For lightweight SR, the IET-light network also consists of 8 blocks, each containing only 3 layers. It uses 3 attention heads with 54 total channels. For all models, we only perform propagation mechanism in the last layer of first four blocks, with k_1 being [22, 20, 14, 12], k_2 being [12, 10, 7, 6] respectively.

4 Experiments

4.1 Experiment Setting

We train our model on the DF2K dataset, which is constructed by merging DIV2K [35] and Flickr2K [24]. The model is optimized for 500K iterations using the Muon [18] optimizer with an initial learning rate of 2×10^{-4}. The model is initially trained on image patches of size 50×50 with a batch size of 60, and subsequently fine-tuned using 75×75 patches with a reduced batch size of 30. Moreover, we employ a MultistepLR scheduler, which reduces the learning rate by half at specified iterations [250000, 400000, 450000, 475000]. Furthermore, all computational cost measurements reported in this paper are based on outputs with a spatial resolution of 1280×640.

4.2 Ablation Study

We conduct ablation studies on the proposed IET-light model, with all models trained for 250k iterations on the DIV2K dataset at $\times 4$ scale.

Effects of Propagation and Sparsification Mechanism. In order to demonstrate the effectiveness of the key design choices in the proposed individualized exploratory transformer (IET) model, we establish five models with different component combinations and evaluate their performance on Urban100 and

Table 1: Ablation study on the effects of each component.

Propagation	Sparsification	SF-FFN	FLOPs	PSNR (dB)
			180.4G	26.17 / 30.58
	✓		61.2G	26.28 / 30.66
✓			215.9G	26.91 / 30.92
✓	✓		68.9G	26.96 / 31.05
✓	✓	✓	69.5G	**27.03** / **31.09**

Table 2: Ablation on the number of prpagation steps.

Propagation Depth	FLOPs	Set5	Urban100	Manga109
1	65.4G	32.29	26.52	30.70
2	68.5G	32.35	26.66	30.81
3	69.1G	32.48	26.83	30.97
4	69.5G	**32.55**	**27.03**	**31.09**
5	69.8G	32.53	26.96	31.05

Table 3: Ablation on dilation settings.

d for Train	1	2	3	2	2
d for Infer	1	2	3	3	4
PSNR (dB)	26.86	26.97	26.99	27.03	27.02

Table 4: Ablation on the ratio of k_1 to k_2 in propagation, evaluated on Urban100.

k_1/k_2	0.5	1	2	3
PSNR(dB)	26.98	27.01	27.03	27.00

Manga109 datasets, as shown in Table 1. The first row represents the baseline model, where all modules are disabled. This configuration is equivalent to dilated attention, meaning that all modules use the same fixed attention candidate.

The second model, which enables the sparsification mechanism, shows a marginal improvement of about 0.11 dB on Urban100 and 0.08 dB on Manga109. Skipping low-correlation attention computations not only reduces computational cost but also improves performance. This suggests that spatially initialized attention candidates contain a large number of weakly related neighbors, which are not only computationally redundant but can even degrade performance. The third model introduces the propagation mechanism, achieving a significant improvement of about 0.66 dB on Urban100 and 0.29 dB on Manga109. The propagation mechanism not only compensates for the coarse nature of the initial dilation-based expansion but also further extends the receptive field. Simply introducing direct connections to two-hop neighbors brings considerable performance improvement, highlighting the importance of relational propagation beyond local neighborhoods. The fourth model, combining the propagation and sparsification, improves PSNR by 0.79 on Urban10 and 0.47 dB on Manga109 compared with the baseline. The final model integrates SF-FFN based on the fourth one, achieving an improvement of 0.07 dB on Urban100 and 0.04 dB on Manga109. SF-FFN concatenates similar tokens along the channel dimension before projection, thereby introducing cross-channel interactions. This design effectively complements the attention mechanism, which lacks explicit cross-channel modeling capability.

Importantly, the primary performance gain comes from the propagation mechanism. While dilated initialization expands attention candidates based solely on spatial proximity, it remains a coarse expansion strategy. Its role is mainly to provide sufficiently diverse initial neighbors for subsequent relational exploration. In contrast, propagation performs semantic-level expansion by exploring neighbors of neighbors, effectively introducing meaningful long-range dependencies. This relational reasoning over two-hop connections accounts for the major improvement. Sparsification brings a modest performance gain, but its principal benefit lies in reducing computational redundancy by removing weakly correlated connections. Moreover, the SF-FFN introduces explicit cross-channel interactions to further improve the model.

Table 5: Ablation on SF-FFN placement across the final blocks. 0 block from last denotes that there is no SF-FFN.

Blocks (from last)	0	1	2	3
PSNR	26.96	27.00	27.03	27.02

Table 6: Ablation on the number of similar tokens fused in SF-FFN. 0 similar token denotes that there is no SF-FFN.

Fused Similar Tokens	0	1	2
PSNR	26.96	27.03	26.97

Effects of Propagation Depth. We apply the propagation mechanism only in the last layer of each block to ensure stability. To study how many blocks should incorporate the propagation mechanism, we evaluate models with different numbers of propagation steps applied in the earlier blocks. As shown in Tab. 2, using four propagation steps achieves the best performance. This result indicates that propagation is effective for exploring new neighbors and enriching attention candidates. Meanwhile, applying propagation too late or too frequently may introduce unstable or noisy long-range candidates, suggesting that later blocks require higher accuracy in similarity modeling and benefit less from excessive propagation.

Effects of Dilation on Attention Candidates Initialization. In IET, the attention candidate initialization uniformly samples tokens from each $d \times d$ patch, where the dilation factor d determines the spatial sampling density. Different dilations can be used during training and inference since dilation only affects attention candidate selection, not model parameters. In general, a larger d corresponds to a wider initial receptive field , allowing tokens to access more distant contextual information, while $d = 1$ restricts attention to the local neighborhood. As shown in Tab. 3, increasing d during both training and inference effectively broadens the receptive field and enhances feature aggregation, leading to improved reconstruction quality. However, beyond $d = 3$, the performance gain saturates, suggesting that excessively long-range dependencies contribute marginally to SR recovery. Moreover, during training, using $d = 2$ achieves slightly better results than $d = 3$, as a smaller dilation allows for smaller patches and thus larger batch sizes, leading to more stable optimization. This experiment also verifies that the similarity modeling in attention is weakly correlated with spatial distance, highlighting the rationality of selecting attention candidates according to image content rather than fixed spatial positions. Finally, we adopt $d = 2$ for training and $d = 3$ for inference in our IET model.

Effects of k_1/k_2 Ratio on Propagation. In our IEA, tokens first selects their k_1 most similar token as one-hop neighbors, and then collects k_2 most similar tokens for each one-hop neighbor to obtain two-hop neighbors. We further study the impact of the ratio between k_1 and k_2 on reconstruction performance. As shown in Tab. 4, the model achieves the best results when $k_1/k_2 = 2$, suggesting that appropriately enlarging the one-hop neighbor set improves stability and enhances the effectiveness of the propagation.

Effects of SF-FFN. Our network consists of 8 Transformer blocks, and we explore the effect of inserting the proposed SF-FFN at different positions. As shown in Tab. 5, placing SF-FFN in the early blocks yields limited improvement since the attention maps at early stages are relatively noisy and cannot accurately rep-

Table 7: Quantitative comparison (PSNR/SSIM) with state-of-the-art methods on **classical SR** task. The best and second best results are colored with red and blue. Results on ×3 model are presented in the supplementary material.

Method	Scale	Params	FLOPs	Set5		Set14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [24]	×2	42.6M	22.14T	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [45]	×2	15.4M	7.02T	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
HAN [33]	×2	63.6M	7.24T	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IPT [3]	×2	115M	7.38T	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [23]	×2	11.8M	3.04T	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9433	39.92	0.9797
CAT-A [5]	×2	16.5M	5.08T	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
ART [42]	×2	16.4M	7.04T	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
HAT [4]	×2	20.6M	5.81T	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
MambaIRv2-B [15]	×2	22.9M	6.27T	38.65	0.9631	34.89	0.9275	32.62	0.9053	34.49	0.9468	40.42	0.9810
IPG [34]	×2	18.1M	5.35T	38.61	0.9632	34.47	0.9270	32.60	0.9052	34.48	0.9464	40.24	0.9810
ATD [43]	×2	20.1M	6.07T	38.61	0.9629	34.95	0.9276	32.65	0.9056	34.70	0.9476	40.37	0.9810
PFT [27]	×2	19.6M	5.03T	38.68	0.9635	35.00	0.9280	32.67	0.9058	34.90	0.9490	40.49	0.9815
IET (ours)	×2	19.7M	5.02T	38.74	0.9636	35.09	0.9286	32.71	0.9064	35.07	0.9499	40.61	0.9816
EDSR [24]	×3	43.0M	9.82T	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [45]	×3	15.6M	3.12T	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
HAN [33]	×3	64.2M	3.21T	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
IPT [3]	×3	116M	3.28T	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR [23]	×3	11.9M	1.35T	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
CAT-A [5]	×3	16.6M	2.26T	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.38	0.9546
ART [42]	×3	16.6M	3.12T	35.07	0.9325	31.02	0.8541	29.51	0.8159	30.10	0.8871	35.39	0.9548
HAT [4]	×3	20.8M	2.58T	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
MambaIRv2-B [15]	×3	23.1M	2.78T	35.18	0.9334	31.12	0.8557	29.55	0.8169	30.28	0.8905	35.61	0.9556
IPG [34]	×3	18.3M	2.39T	35.10	0.9332	31.10	0.8554	29.53	0.8168	30.36	0.8901	35.53	0.9554
ATD [43]	×3	20.3M	2.69T	35.11	0.9330	31.13	0.8556	29.57	0.8176	30.46	0.8917	35.63	0.9558
PFT [27]	×3	19.8M	2.23T	35.15	0.9333	31.16	0.8561	29.58	0.8178	30.56	0.8931	35.67	0.9560
IET (ours)	×3	19.9M	2.25T	35.20	0.9337	31.23	0.8571	29.61	0.8185	30.81	0.8966	35.82	0.9566
EDSR [24]	×4	43.0M	5.54T	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [45]	×4	15.6M	1.76T	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
HAN [33]	×4	64.2M	1.81T	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IPT [3]	×4	116M	1.85T	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [23]	×4	11.9M	0.76T	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
CAT-A [5]	×4	16.6M	1.27T	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
ART [42]	×4	16.6M	1.76T	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
HAT [4]	×4	20.8M	1.45T	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
MambaIRv2-B [15]	×4	23.1M	1.57T	33.14	0.9057	29.23	0.7975	28.00	0.7511	27.89	0.8344	32.57	0.9295
IPG [34]	×4	17.0M	1.30T	33.15	0.9062	29.24	0.7973	27.99	0.7519	28.13	0.8392	32.53	0.9300
ATD [43]	×4	20.3M	1.52T	33.10	0.9058	29.24	0.7974	28.01	0.7526	28.17	0.8404	32.62	0.9306
PFT [27]	×4	19.8M	1.26T	33.15	0.9065	29.29	0.7978	28.02	0.7527	28.20	0.8412	32.63	0.9306
IET (ours)	×4	19.8M	1.26T	33.22	0.9069	29.35	0.7994	28.06	0.7536	28.43	0.8464	32.81	0.9318

resent token similarity. In contrast, inserting SF-FFN in the later blocks allows it to operate on more stable attention patterns, leading to better reconstruction performance. The best results are achieved when SF-FFN is placed after the last two blocks. we conclude that attention maps in the later blocks represent more accurate similarity relationships, making them more suitable for cross-channel fusion in SF-FFN. We further analyze how many similar tokens should be fused in the SF-FFN. As shown in Tab. 6, when each token interacts with only its most similar neighbor, the model achieves the best trade-off between performance and efficiency. Using more neighbors introduces redundant information and slightly degrades reconstruction accuracy. From these results, we draw a conclusion that in the SR task, where high-precision similarity is critical, each token benefits most from cross-channel interaction with only its single most similar neighbor.

Table 8: Quantitative comparison (PSNR/SSIM) with state-of-the-art methods on **lightweight SR** task. The best and second best results are colored with red and blue. Results on ×3 model are presented in the supplementary material.

Method	Scale	Params	FLOPs	Set5		Set14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [1]	×2	1,592K	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
IMDN [17]	×2	694K	158.8G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
LAPAR-A [21]	×2	548K	171G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
LatticeNet [28]	×2	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
SwinIR-light [23]	×2	910K	244G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
ELAN [44]	×2	582K	203G	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
SwinIR-NG [6]	×2	1181K	274.1G	38.17	0.9612	33.94	0.9205	32.31	0.9013	32.78	0.9340	39.20	0.9781
OmniSR [37]	×2	772K	194.5G	38.22	0.9613	33.98	0.9210	32.36	0.9020	33.05	0.9363	39.28	0.9784
MambaIRv2-light [15]	×2	774K	286.3G	38.26	0.9615	34.09	0.9221	32.36	0.9019	33.26	0.9378	39.35	0.9785
IPG-Tiny [34]	×2	872K	245.2G	38.27	0.9616	34.24	0.9236	32.35	0.9018	33.04	0.9359	39.31	0.9786
ATD-light [43]	×2	753K	348.6G	38.28	0.9616	34.11	0.9217	32.39	0.9023	33.27	0.9376	39.51	0.9789
PFT-light [27]	×2	776K	278.3G	38.36	0.9620	34.19	0.9232	32.43	0.9030	33.67	0.9411	39.55	0.9792
IET-light (Ours)	×2	783K	277.4G	38.44	0.9624	34.28	0.9245	32.50	0.9038	34.03	0.9435	39.75	0.9794
CARN [1]	×3	1,592K	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IMDN [17]	×3	703K	71.5G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
LAPAR-A [21]	×3	544K	114G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
LatticeNet [28]	×3	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
SwinIR-light [23]	×3	918K	111G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
ELAN [44]	×3	590K	90.1G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
SwinIR-NG [6]	×3	1190K	114.1G	34.64	0.9293	30.58	0.8471	29.24	0.8090	28.75	0.8639	34.22	0.9488
OmniSR [37]	×3	780K	88.4G	34.70	0.9294	30.57	0.8469	29.28	0.8094	28.84	0.8656	34.22	0.9487
MambaIRv2-light [15]	×2	781K	126.7G	34.71	0.9298	30.68	0.8483	29.26	0.8098	29.01	0.8689	34.41	0.9497
IPG-Tiny [34]	×2	878K	109.0G	34.64	0.9292	30.61	0.8470	29.26	0.8097	28.93	0.8666	34.30	0.9493
ATD-light [43]	×2	753K	154.7G	34.70	0.9300	30.68	0.8485	29.32	0.8109	29.16	0.8710	34.60	0.9505
PFT-light [27]	×2	776K	123.5G	34.81	0.9305	30.75	0.8493	29.33	0.8116	29.43	0.8759	34.60	0.9510
IET-light (Ours)	×3	790K	123.1G	34.90	0.9314	30.83	0.8504	29.41	0.8135	29.75	0.8808	34.89	0.9524
CARN [1]	×4	1,592K	90.9G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
IMDN [17]	×4	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
LAPAR-A [21]	×4	659K	94G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
LatticeNet [28]	×4	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
SwinIR-light [23]	×4	930K	63.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN [44]	×4	582K	54.1G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
SwinIR-NG [6]	×4	1201K	63G	32.44	0.8980	28.83	0.7870	27.73	0.7418	26.61	0.8010	31.09	0.9161
OmniSR [37]	×4	792K	50.9G	32.49	0.8988	28.78	0.7859	27.71	0.7415	26.65	0.8018	31.02	0.9151
IPG-Tiny [34]	×4	887K	61.3G	32.51	0.8987	28.85	0.7873	27.73	0.7418	26.78	0.8050	31.22	0.9176
MambaIRv2-light [15]	×2	790K	75.6G	32.51	0.8992	28.84	0.7878	27.75	0.7426	26.82	0.8079	31.24	0.9182
ATD-light [43]	×2	769K	87.1G	32.62	0.8997	28.87	0.7884	27.77	0.7439	26.97	0.8107	31.47	0.9198
PFT-light [27]	×2	792K	69.6G	32.63	0.9005	28.92	0.7891	27.79	0.7445	27.20	0.8171	31.51	0.9204
IET-light (Ours)	×4	801K	69.4G	32.70	0.9019	29.00	0.7908	27.85	0.7464	27.41	0.8227	31.73	0.9226

4.3 Comparison with State-of-the-Art Methods

We evaluate our model against various super-resolution baselines on standard benchmark datasets, including Set5 [2], Set14 [41], BSD100 [29], Urban100 [16], and Manga109 [30]. The comparison covers both traditional and recent advanced SR approaches, such as EDSR [24], RCAN [45], HAN [33], SwinIR [23], CAT [5], ART [42], HAT [4], IPG [34], ATD [43], and PFT [27]. The performance of all models is measured using PSNR and SSIM under ×2, ×3, and ×4 upscaling settings. Computational costs are evaluated at an output resolution of 1280×640.

The results shown in Table 7, indicate that with a similar number of parameters, the proposed IET model achieves significantly better performance than PFT. Notably, on the ×2 Urban100 benchmark, IET outperforms PFT by **0.17dB** and ATD by **0.37dB**. For lightweight super-resolution, we further compare our method with efficient SR networks such as CARN [1], IMDN [17], LAPAR [21], SwinIR [23], ELAN [44], and OmniSR [37]. As shown in Table 8, the proposed IET-light consistently surpasses PFT-light [27] on all benchmark datasets. Specifically, on the ×2 Urban100 dataset, IET-light exceeds PFT-light

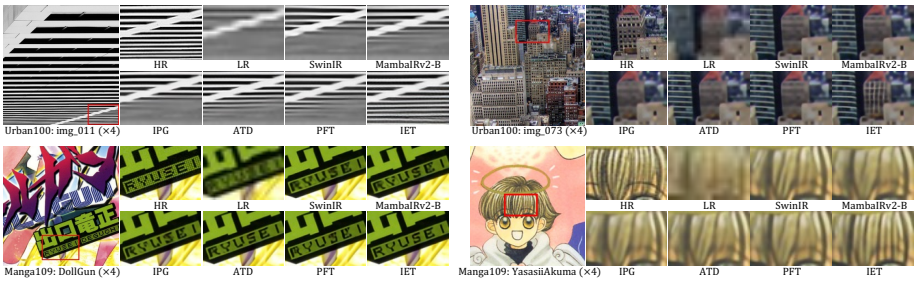

Fig. 6: Visual comparisons of **IET** and other SOTA image super-resolution methods.

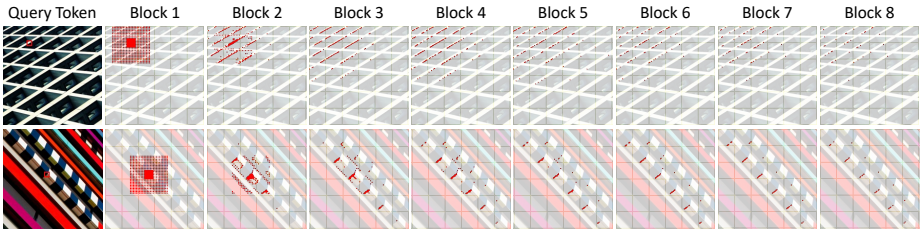

Fig. 7: Visualization of the attention candidates initialization and propagation process. The yellow grids represents 32×32 windows, which is the largest window size used among window-based methods.

by **0.36dB** and ATD-light by **0.76dB**. Moreover, IET-light outperforms SwinIR by 0.02dB on Set5 and 0.22dB on Urban100, while requiring only **9.1%** of the total computational complexity. For the ×4 SR setting, IET-light also achieves a **0.21dB** improvement over PFT-light and a **0.44dB** gain over ATD-light on Urban100. The strong performance of IET stems from the proposed Individualized Exploratory Attention, which enables content-aware and token-adaptive attention candidates, and then aggregates information among the tokens that are most appropriate for feature enhancement.

We also provide some visual examples using different methods to qualitatively verify the efficacy of IET, as shown in Fig.6, which clearly demonstrate our advantage in recovering sharp edges and clean textures.

4.4 Visualization Analysis

We further visualize the refinement process of attention candidates in Fig. 7, where yellow grids represent 32×32 windows and red points denotes attention candidates. In the first block, the attention scope is initialized as a dense local and sparse global region centered on each token. This allows every token to accurately capture nearby information while roughly perceiving distant context, providing rich initial cues for subsequent propagation. In the following blocks, IEA gradually expands the attention candidates by aggregate with new two-hop similar tokens while pruning low-similarity neighbors, maintaining efficiency with comparable computational cost.

5 Conclusion

We propose the Individualized Exploratory Transformer (IET), a content-aware self-attention framework with token-adaptive candidate selection. Its core module refines attention through propagation and sparsification, modeling asymmetric similarity while reducing redundancy. This flexible and efficient design achieves state-of-the-art results on super-resolution benchmarks and shows potential for broader vision and language tasks.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62476051) and the Sichuan Natural Science Foundation (No. 2024NSFTD0041).

References

1. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network (Mar 2018)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.l.A.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Procedings of the British Machine Vision Conference 2012 (Sep 2012). https://doi.org/10.5244/c.26.135, http://dx.doi.org/10.5244/c.26.135
3. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer (Dec 2020)
4. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22367–22377 (June 2023)
5. Chen, Z., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Cross aggregation transformer for image restoration. In: NeurIPS (2022)
6. Choi, H., Lee, J., Yang, J.: N-gram in swin transformers for efficient lightweight image super-resolution (Nov 2022)
7. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers (Dec 2021)
8. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jan 2020). https://doi.org/10.1109/cvpr.2019.01132, http://dx.doi.org/10.1109/cvpr.2019.01132
9. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV. pp. 184–199. Springer (2014)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 295–307 (Jun 2015). https://doi.org/10.1109/tpami.2015.2439281, http://dx.doi.org/10.1109/tpami.2015.2439281

11. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows (2021)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (Oct 2020)
13. Gu, S., Guo, S., Zuo, W., Chen, Y., Timofte, R., Van Gool, L., Zhang, L.: Learned dynamic guidance for depth image reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2437–2452 (2019)
14. Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., Zhang, L.: Convolutional sparse coding for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1823–1831 (2015)
15. Guo, H., Guo, Y., Zha, Y., Zhang, Y., Li, W., Dai, T., Xia, S.T., Li, Y.: Mambairv2: Attentive state space restoration. arXiv preprint arXiv:2411.15269 (2024)
16. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Oct 2015). https://doi.org/10.1109/cvpr.2015.7299156, http://dx.doi.org/10.1109/cvpr.2015.7299156
17. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th ACM International Conference on Multimedia (Oct 2019). https://doi.org/10.1145/3343031.3351084, http://dx.doi.org/10.1145/3343031.3351084
18. Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/
19. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Dec 2016). https://doi.org/10.1109/cvpr.2016.182, http://dx.doi.org/10.1109/cvpr.2016.182
20. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1637–1645 (2016)
21. Li, W., Zhou, K., Qi, L., Jiang, N., Lu, J., Jia, J.: Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond (Jan 2020)
22. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)
23. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: Image restoration using swin transformer (2021)
24. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Aug 2017). https://doi.org/10.1109/cvprw.2017.151, http://dx.doi.org/10.1109/cvprw.2017.151
25. Liu, X., Liu, J., Tang, J., Wu, G.: Catanet: Efficient content-aware token aggregation for lightweight image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17902–17912 (2025)
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
27. Long, W., Zhou, X., Zhang, L., Gu, S.: Progressive focused transformer for single image super-resolution. arXiv preprint arXiv:2503.20337 (2025)

28. Luo, X., Xie, Y., Zhang, Y., Qu, Y., Li, C., Fu, Y.: LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block, p. 272–289 (Nov 2020). https://doi.org/10.1007/978-3-030-58542-6_17, http://dx.doi.org/10.1007/978-3-030-58542-6_17
29. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 (Nov 2002). https://doi.org/10.1109/iccv.2001.937655, http://dx.doi.org/10.1109/iccv.2001.937655
30. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications p. 21811–21838 (Nov 2016). https://doi.org/10.1007/s11042-016-4020-z, http://dx.doi.org/10.1007/s11042-016-4020-z
31. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Nov 2021). https://doi.org/10.1109/cvpr46437.2021.00352, http://dx.doi.org/10.1109/cvpr46437.2021.00352
32. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
33. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single Image Super-Resolution via a Holistic Attention Network, p. 191–207 (Oct 2020). https://doi.org/10.1007/978-3-030-58610-2_12, http://dx.doi.org/10.1007/978-3-030-58610-2_12
34. Tian, Y., Chen, H., Xu, C., Wang, Y.: Image processing gnn: Breaking rigidity in super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24108–24117 (June 2024)
35. Timofte, R., Agustsson, E., Gool, L.V., Yang, M.H., Zhang, L., Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Aug 2017). https://doi.org/10.1109/cvprw.2017.149, http://dx.doi.org/10.1109/cvprw.2017.149
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
37. Wang, H., Chen, X., Ni, B., Liu, Y., jinfan, L.: Omni aggregation networks for lightweight image super-resolution. In: Conference on Computer Vision and Pattern Recognition (2023)
38. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Mar 2022). https://doi.org/10.1109/iccv48922.2021.00061, http://dx.doi.org/10.1109/iccv48922.2021.00061
39. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
40. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17683–17693 (June 2022)

41. Zeyde, R., Elad, M., Protter, M.: On Single Image Scale-Up Using Sparse-Representations, p. 711–730 (Jan 2012). https://doi.org/10.1007/978-3-642-27413-8_47, http://dx.doi.org/10.1007/978-3-642-27413-8_47
42. Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Accurate image restoration with attention retractable transformer. In: ICLR (2023)
43. Zhang, L., Li, Y., Zhou, X., Zhao, X., Gu, S.: Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2856–2865 (June 2024)
44. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: European Conference on Computer Vision. pp. 649–667. Springer (2022)
45. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image Super-Resolution Using Very Deep Residual Channel Attention Networks, p. 294–310 (Oct 2018). https://doi.org/10.1007/978-3-030-01234-2_18, http://dx.doi.org/10.1007/978-3-030-01234-2_18
46. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution (Feb 2018)
47. Zhou, Y., Li, Z., Guo, C.L., Bai, S., Cheng, M.M., Hou, Q.: Srformer: Permuted self-attention for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12791 (2023)