Lightweight Image Super-Resolution with Superpixel Token Interaction

Aiping Zhang[1] Wenqi Ren[1*] Yi Liu[2] Xiaochun Cao[1]
[1]School of Cyber Science and Technology, Sun Yat-Sen University [2]Baidu Inc.
{zhangaip7@mail2, renwq3@mail, caoxiaochun@mail}.sysu.edu.cn, liuyi22@baidu.com

Abstract

Transformer-based methods have demonstrated impressive results on single-image super-resolution (SISR) task. However, self-attention mechanism is computationally expensive when applied to the entire image. As a result, current approaches divide low-resolution input images into small patches, which are processed separately and then fused to generate high-resolution images. Nevertheless, this conventional regular patch division is too coarse and lacks interpretability, resulting in artifacts and non-similar structure interference during attention operations. To address these challenges, we propose a novel super token interaction network (SPIN). Our method employs superpixels to cluster local similar pixels to form the explicable local regions and utilizes intra-superpixel attention to enable local information interaction. It is interpretable because only similar regions complement each other and dissimilar regions are excluded. Moreover, we design a superpixel cross-attention module to facilitate information propagation via the surrogation of superpixels. Extensive experiments demonstrate that the proposed SPIN model performs favorably against the state-of-the-art SR methods in terms of accuracy and lightweight. Code is available at `https://github.com/ArcticHare105/SPIN`.

1. Introduction

Single image super-resolution (SISR) is a crucial task in computer vision that aims to enhance the resolution and visual quality of low-resolution (LR) images. The goal of SISR is to generate a high-resolution (HR) image from a given LR image, which can be particularly useful in applications where high-quality images are necessary, such as medical imaging, surveillance, and digital photography.

Since the pioneering work of Dong et al. [5], numerous neural networks have been developed to tackle the challenge of reconstructing high-quality images from low-resolution inputs. Some of the CNN-based methods use deeper and more complex architectures to achieve better performance. However, these methods come with a trade-off of increased computational resources and higher cost, which can limit their application scenarios.

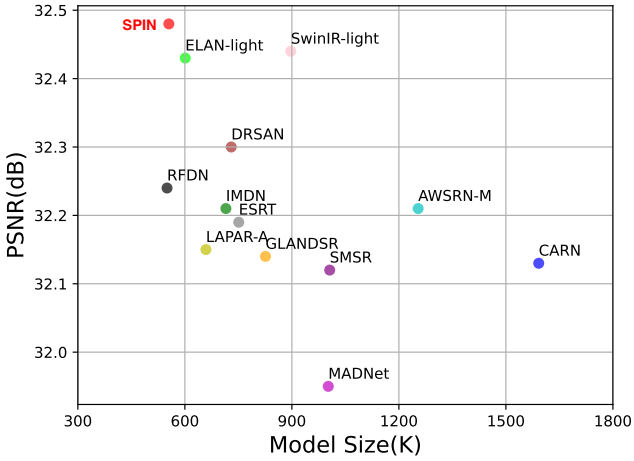

Figure 1. PSNR and model parameters for $\times 4$ super-resolution on Set5. We compare our SPIN with state-of-the-art **lightweight** Transformer-based and CNN-based models, including SwinIR-light [20], ESRT [26], ELAN-light[43], and IMDN [12], *etc.*

Attention mechanism [37] has been proven to have significant effects on both high-level vision tasks and low-level fields, including super-resolution (SR). Attention mechanisms allow the network to selectively focus on relevant regions of the input, which can improve the quality of the SR output. Capitalized on attention mechanisms, transformers have been applied to SR tasks such as SwinIR [20] and ESRT [26]. These models highlight the importance of global feature extraction abilities in SISR. Furthermore, to improve the efficiency, ELAN [43] proposes a group-wise self-attention module and shared the weights when calculating the association of patches. However, the attention mechanism has high computational complexity and memory consumption, which requires dividing large images into small patches for separate processing. While this strategy enhances the efficiency of transformer-based models, it results in some problems. Dividing patches based on a fixed shape leads to the splitting up of continuous structures, which hin-

*Corresponding author.

ders the use of similar information in other areas to enhance image details. Moreover, the local attention mechanism applied within each patch involves irrelevant regions in computation, leading to undesirable inferences.

To address these issues, we propose a novel approach that integrates local and global attention mechanisms with fine superpixel partition. We start with CNN-based shallow feature extraction on the pixels of the input image and perform local clustering to group adjacent pixels into superpixels. We then obtain local regions by clustering superpixels based on similarity and perform local feature extraction on them separately. Unlike the previous approaches [20, 43] using fixed shape patch division, which was only used for improving parallel computation efficiency, our strategy for region division is more interpretable, allowing for more flexible and adaptive division of the input image, and preventing the splitting up of continuous structures. We then introduce Superpixel Cross Attention module to enable information interaction in the long-range via the surrogation of superpixels. Furthermore, we design an Intra-Superpixel Attention (ISPA) mechanism applied to the pixels of superpixels, extending the original attention operation only in the regular image area. This ensures that local attention mechanism information interactions occur in similar areas, eliminating interference and irrelevant computation. These two proposed attention mechanisms interlace with each other and cooperate in local and global feature extraction. As shown in Fig. 1, the proposed SPIN has a good trade-off between PSNR and model size.

Our contributions are summarized below:

(a) We present a novel super-resolution model that combines superpixel clustering with the transformer structure, resulting in a more interpretable framework.

(b) We propose Intra-Superpixel Attention (ISPA) and Superpixel Cross Attention (SPCA) modules that operate within and between superpixels, enabling computation in irregular areas while maintaining the ability to capture long-range dependencies.

(c) The experiments demonstrate that the proposed method achieves better SR reconstruction performance compared to state-of-the-art lightweight SR methods.

2. Related Work

2.1. Deep Networks for Super-Resolution

With the recent advancements in deep learning, neural network-based methods have become the mainstream solutions for single image super-resolution (SR). SRCNN [5] uses a three-layer CNN network to reconstruct a high-resolution (HR) image from its bicubically downsampled low-resolution (LR) image. To further improve accuracy, recent CNN-based methods have employed more complex and effective structures. For example, Kim et al. [15] apply a deep CNN-based architecture with residual learning to improve SR accuracy.

Attention mechanisms have also been introduced in SR to extract the most important and informative features. For instance, Zhang et al. [44] use a channel attention mechanism, while Hu et al. [10] combine spatial attention with channel attention in SR. Furthermore, inspired by the success of ViT [6] in high-level vision tasks, Chen et al. [4] introduces Transformer into SR, but it required large amount of parameters. To reduce the model size, SwinIR [20] applies the Swin Transformer [24] framework to SR by dividing the entire image into small windows with a fixed size of 8×8 and shifting the windows when applying multi-head attention mechanisms. While these above methods have been effective in extracting informative features, they may require a large number of parameters.

2.2. Lightweight Super-Resolution Methods

Lightweight is a critical consideration for deep SR models, and many approaches have been proposed to improve their efficiency. For example, FSRCNN [5] and ESPCN [34] utilize a post-upsampling technique to reduce the computational burden, while CARN [1] uses group convolutions and a cascading mechanism to improve efficiency but impaired the performance. IMDN [12] applies the three-step-distillation to extract features and a slice operation to divide the extracted features, yet brings inflexibility. LatticeNet [28] introduces lattice blocks with low calculation complexity. BSRN [19] designs a depth-wise separable convolution to reduce model complexity and utilizes attention mechanisms to improve the SR reconstruction performance. Meanwhile, Lightweight Transformer-based SR approaches are proposed to reduce the model complexity, e.g., by reducing the calculated tokens by using window-based attention [20] and adopting shifted convolution and group-wise self-attention [43]. Although these approaches are lightweight and efficient, the quality of SR reconstruction still keeps room for improvement.

2.3. Pixel Clustering for Image Processing

Pixel clustering is a well-studied task in image processing, and recent advancements in deep learning methods have shown significant progress in this area. One common approach to pixel clustering is to use CNNs to generate pixel-level embeddings that group similar pixels together. For example, Liu et al. [23] develop a deep affinity network that learns pixel-wise affinities to cluster pixels. Similarly, Sun et al. [35] propose a network that learns pixel-level representations to cluster image patches.

In addition to using CNNs to generate pixel-level embeddings, clustering algorithms can also be applied to CNN features to group similar pixels into clusters. Jégou et al. [14] introduce a one-shot clustering method that uses CNN

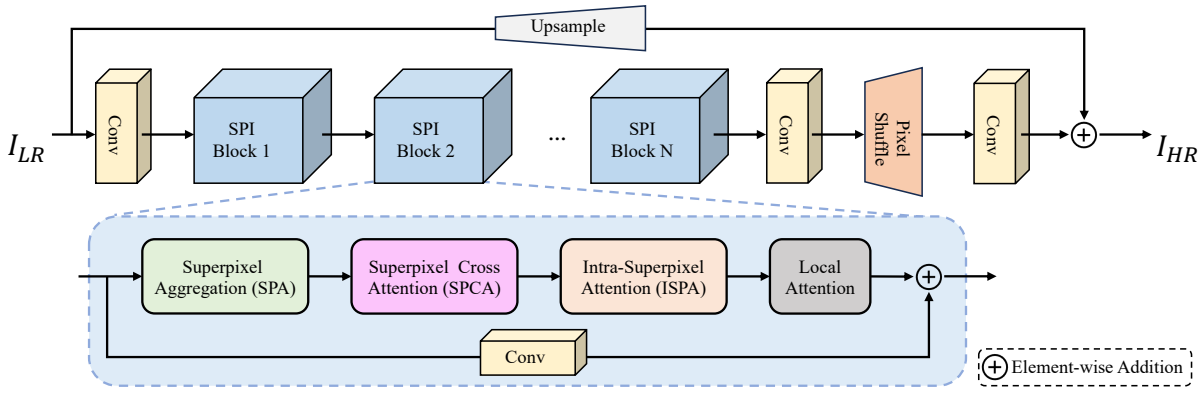

Figure 2. Configuration of the proposed method. The proposed network is mainly composed of the proposed Super-Pixel Interaction (SPI) blocks, which consist of four components: Superpixel Aggregation (SPA), Superpixel Cross Attention (SPCA), Intra-Superpixel Attention (ISPA) and local attention. The SPA module is responsible for aggregating information from the superpixels in the input image, the SPCA module captures the interactions between pixels via the surrogation of superpixels, while ISPA module captures the interactions among pixels within each superpixel. Local attention is adopted to enhance the interaction within local regions.

features to generate initial clusters, which are then further refined using a clustering algorithm. Li et al. [17] propose a weakly supervised clustering method that uses CNN features and a sparse labeling scheme to cluster pixels into object regions. These approaches leverage the power of both CNNs and clustering algorithms, enabling more accurate and efficient pixel clustering in image processing tasks.

Recently, there has been a growing interest in using graph convolutional networks (GCNs) for pixel clustering. GCNs are capable of modeling the dependencies between pixels in an image by constructing a graph representation of the image, where each pixel is a node, and the edges represent the relationships between pixels. This enables GCNs to capture more complex and non-local interactions between pixels, compared to traditional CNNs. For example, Zeng et al. [40] propose a GCN-based framework for hyperspectral image classification that uses two clustering strategies to exploit multi-hop correlations. The first clustering strategy groups similar pixels based on their spectral similarity, while the second clustering strategy groups pixels based on their spatial adjacency.

Although pixel clustering has demonstrated promising results in various image processing tasks, it has not been effectively applied in super-resolution applications.

3. Proposed Method

The architecture of the proposed model is shown in Fig. 2, which mainly consists of the proposed Super-Pixel Interaction (SPI) blocks. Before SPI blocks, we utilize an encoder, which is a 3×3 convolution, to embed the low-resolution image $\mathbf{I}_{LR}$ to a high-dimensional feature space. Given the encoder, we can get the shallow feature $\mathbf{x}_{emb}$ as:

$$\mathbf{x}_{emb} = f_{\text{encoder}}(\mathbf{I}_{LR}), \tag{1}$$

where f_{encoder} denotes the encoder of the proposed model.

Then, we stack K SPI blocks on top of the encoder to extract deeper features that contain both rich low-level and high-level information of the input image. Each SPI block includes four components: Superpixel Aggregation (SPA), Superpixel Cross Attention (SPCA), Intra-Superpixel Attention (ISPA), and local Attention.

The input feature of each block is first aggregated into superpixels via the SPA module. Then, the ISPA module captures the dependencies and interactions of pixels within each superpixel, while the SPCA module captures the dependencies and interactions between long-range pixels. In order to enhance the interaction between pixels within local regions, we utilize a local attention module, which uses window-based attention [24, 20, 21], after the ISPA and the SPCA module. We use overlapped patches to strengthen feature interaction. Formally, for the i-th SPI block, the whole process can be formulated as:

$$\begin{aligned} \mathbf{s}_i &= f_{\text{SPA}}(\mathbf{x}_{i-1}), \\ \mathbf{x}_i &= \mathbf{x}_{i-1} + f_{\text{local}}(f_{\text{ISPA}}(f_{\text{SPCA}}(\mathbf{x}_{i-1}, \mathbf{s}_i))) \end{aligned} \tag{2}$$

where $\mathbf{s}_i$ denotes the features of superpixels in the i-th SPI block, $f(\cdot)$ denotes the function of each individual component. Following previous works, the residual connection is used to ease the whole training process.

After the K SPI blocks, we adopt 3×3 convolutional layers and the pixel-shuffle operation [34] to obtain the global residual information, which is added to the upsampled image of $\mathbf{I}_{LR}$ for resolving the high-resolution image $\mathbf{I}_{SR}$.

3.1. The SPA Module

Different from previous methods that divide the input image or feature into regular patches, we propose to partition the input feature into superpixels. Compared with the

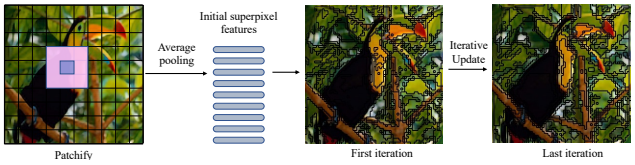

Figure 3. The Superpixel Aggregation (SPA) module of our method, which initializes the superpixel by average pooling and then updates them in an iterative way.

regular patches that may easily crop connected regions into different patches, the superpixel-wise partition can perceptually group similar pixels together, which can depict more precise boundaries, reducing the risk of generating blurry and inaccurate boundaries.

Specifically, in the process of superpixel aggregation, we utilize the soft k-means-based superpixel algorithm in SSN [13]. Given the visual tokens $\mathbf{x} \in \mathbf{R}^{N\times C}$ (where $N = H \times W$ is the number of visual tokens), each token $\mathbf{x}(i) \in \mathbf{R}^C$ is assumed to belong to one of M superpixels $\mathbf{s} \in \mathbf{R}^{M\times C}$, making it necessary to compute the association between visual tokens and superpixel tokens.

Formally, the process of superpixel aggregation is an Expectation-Maximization-like process, which contains total T iterations. Firstly, as shown in Fig. 3, we sample initial super tokens $\mathbf{s}^0$ by averaging tokens in regular grids, called Patchify. Suppose the grid size is $H_s \times W_s$, then the number of super tokens is $M = \frac{H}{H_s} \times \frac{W}{W_s}$. For the t-th iteration, we calculate the association map as:

$$\mathbf{A}^t(ij) = e^{-\|\mathbf{x}(i)-\mathbf{s}^{t-1}(j)\|_2^2}, \tag{3}$$

where $\mathbf{A}^t \in \mathbf{R}^{N\times M}$ is the association map and $\mathbf{A}^t(ij)$ is the value at the i-th row and the j-th column. Note that, superpixel aggregation only calculates the association map from each token to surrounding superpixels, which guarantees the locality of superpixels, making it also efficient in terms of both computation and memory [13].

After that, we can obtain the superpixels $\mathbf{s}^t$ as the weighted sum of visual tokens, defined as:

$$\mathbf{s}^t(j) = \frac{1}{\mathbf{z}^t(j)} \sum\nolimits_i \mathbf{A}^t(ij)\mathbf{x}(i), \tag{4}$$

where $\mathbf{z}^t(j) = \sum_i \mathbf{A}^t(ij)$ denotes the normalization term along the column. After T iterations, we can obtain the final association map $\mathbf{A}^T$. For simplicity, we omit the superscript in the following sections.

3.2. The SPCA Module

Since the superpixels capture only the locality and interconnection of pixels in local regions, which may lack the capacity of capturing long-range dependencies for super-resolution. Here, we utilize the self-attention paradigm [37] to enhance long-range communication via the surrogation

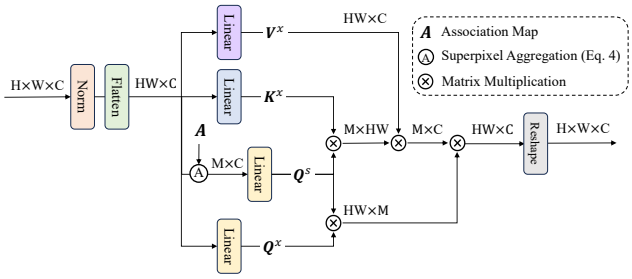

Figure 4. The proposed Superpixel Cross Attention (SPCA) module. We first propagate information from pixels to superpixels and then distribute the aggregated information to pixels by cross-attention mechanism.

of superpixels, which can help to make use of the complementarity between features to produce high-quality super-resolution images. Since pixel features are highly similar to the belonging superpixel features, making superpixels a promising surrogation to propagate information between pixels as much as possible.

As shown in Fig. 4, given the superpixel features $\mathbf{s} \in \mathbf{R}^{M\times C}$, where M denotes the number of superpixels, and the flattened pixel features $\mathbf{x} \in \mathbf{R}^{HW\times C}$. We employ the attention mechanism [37] to first propagate the pixel information to superpixels. Specifically, we use linear projections to calculate the *query*: $\mathbf{Q}^s \in \mathbf{R}^{M\times D}$, *key*: $\mathbf{K}^x \in \mathbf{R}^{HW\times D}$, and *value*: $\mathbf{V}^x \in \mathbf{R}^{HW\times C}$ as:

$$\mathbf{Q}^s = \mathbf{s}\mathbf{W}_q^s, \quad \mathbf{K}^x = \mathbf{x}\mathbf{W}_k^x, \quad \mathbf{V}^x = \mathbf{x}\mathbf{W}_v^x \tag{5}$$

where $\mathbf{W}_q^s \in \mathbf{R}^{C\times D}$, $\mathbf{W}_k^s \in \mathbf{R}^{C\times D}$, $\mathbf{W}_v^s \in \mathbf{R}^{C\times C}$ are weight matrices according to query, key and value, respectively. The output can be obtained by first calculating the similarity between the query and key and using it as the weights to aggregate the value, which can be formulated as:

$$\mathbf{s}_u = \mathrm{softmax}(\mathbf{Q}^s(\mathbf{K}^x)^T/\sqrt{D})\mathbf{V}^x, \tag{6}$$

where $\sqrt{D}$ is a scaling factor to avoid vanishing gradients, $\mathbf{s}_u$ is the updated superpixel features. Note that, unlike superpixel aggregation, this process does not take neighbor restrictions into account, ensuring the propagation of long-range information.

Once information has been propagated from pixels to superpixels, it becomes necessary to distribute the aggregated information back to pixels, so as to achieve the information propagation between pixels. Here, we further employ the attention mechanism. Specifically, we utilize another weight matrix $\mathbf{W}_q^x$ to obtain the query from pixel features. To reduce the number of parameters, we directly use the superpixel features $\mathbf{Q}^s$ as the key, and the updated superpixel features as the value, and utilize cross attention to map the updated superpixel features back to the pixel level. Similar to the Transformer block [37], we also adopt the Feed

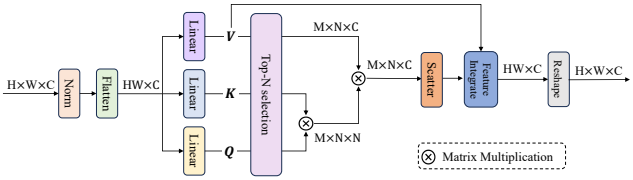

Figure 5. The proposed Intra-Superpixel Attention (ISPA) module. We select top-N pixels which are most similar to each superpixel for intra-superpixel attention. Feature integration is adopted to integrate those "ignored" pixels.

Forward Network (FFN) after the above process. Our FFN contains a layer normalization [2] layer, after which we utilize feature gating [33] to modulate the input feature and channel attention [9] to extract global information. After that, two fully-connected layers and GELU [8] activation function are used.

3.3. The ISPA Module

Given the association map, an intuitive way to improve the quality of super-resolution images is to utilize the complementarity of similar pixels within the same superpixel. To achieve this, we need to obtain the corresponding pixels of each superpixel. However, different superpixels may contain different numbers of pixels, which makes it difficult to conduct parallel processing and also result in unexpected memory consumption, because there are always some superpixels that include a large number of pixels.

To address this issue, as shown in Fig. 5, we resort to the association map $\mathbf{A}^T$ and select top-N pixels which are most similar to each superpixel. Suppose the affiliated pixels of one superpixel is $\mathbf{f} = \{\mathbf{x}(i)\}_N \in \mathbf{R}^{N\times C}$, where N denotes the number of selected pixels. We follow the standard self-attention mechanism [37], *i.e.*, Eq. 5 and Eq. 6, to conduct intra-superpixel attention, which includes weight matrices $\mathbf{W}_q^f$, $\mathbf{W}_k^f$ and $\mathbf{W}_v^f$ for query, key and value projection. After the intra-superpixel interaction, we disperse the refined pixel features back to their respective positions within the image, utilizing the indices generated during the top-N selection process.

The top-N selection may lead to some "ignored" pixels, *i.e.*, those pixels are not included in any superpixels. For those "ignored" pixels, we utilize the value projection $\mathbf{W}_v^f$ to project them to obtain the updated features, which are then integrated with those pixels that are updated by intra-superpixel interaction. Similar to the SPCA module, we adopt the same FFN after the ISPA module.

4. Experiments

In this section, we describe in detail the ablation experiments for each module and the performance of our method for different-scale super-resolution tasks.

4.1. Datasets

We use DIV2K [36] as the training set, which is a high-definition dataset including images of various natural scenes. This dataset includes 900 high-resolution images, the first 800 images are used for training, and the last 100 images for validation. Following RCAN [44], the LR samples are generated using a double triple downsampling method. In addition, we evaluate our method on five commonly used benchmarks including Set5 [3], Set14 [41], BSDS100 [29], Urban100 [11], and Manga109 [30].

4.2. Implementation Details

During training, the initial learning rate is set to 5e-4, and the training procedure stops after 1000 epochs. The optimizer used is the Adam optimizer with β_1 of 0.9 and β_2 of 0.999. To train the models, we employ randomly rotating 90°, 180°, 270°, and horizontal flip for data augmentation. In the final model, the output channel is set to 40 for all blocks. We set the number of the SPI block to 8 and employ distinct initial patches for superpixel aggregation across various SPI blocks, spanning from 12 to 24.

For evaluations, we mainly use the commonly used evaluation metrics, including peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). We follow RCAN [44] to measure the metrics on the Y channel after converting RGB to YCbCr format.

4.3. Comparison with Light-weight Models

We compare our model with state-of-the-art light-weight SR models, including CNN-based models of CARN [1], IMDN [12], LatticeNet [28], *etc.*, and transformer-based models of ESRT [26] and SwinIR [20] and ELAN [43].

Quantitative comparison. The quantitative metrics of different methods are reported in Table 1. We can observe that the transformer-based models [20, 26, 43] consistently outperform those CNN-based methods [1, 12, 38, 16, 18, 22, 42] in terms of PSNR and SSIM, by leveraging the long-range similarity between image patches. However, they always divide the image into regular patches, which may break the object, boundaries, *etc.* in the input image.

In contrast, our method leverages superpixels to enable interpretable and continuous region division for Transformer. We obtain the best or the second-best PSNR/SSIM scores on all five benchmark datasets and on all three scales. Moreover, the number of parameters is smaller than those of existing transformer-based methods.

Qualitative comparison. Fig. 6 displays visual comparisons for scale factor ×4 on Urban100, BSDS100, and Set14 datasets. The results indicate that the proposed SPIN can effectively restore textures that have been largely damaged, as long as corresponding non-local information is available in the LR images. In contrast, deep SISR models that lack

Table 1. Average PSNR/SSIM comparison with other advance CNN-based and Transformer-based SISR models. The best and the second-best results are highlighted and underlined, respectively.

Methods	Scale	Params	Set5	Set14	BSDS100	Urban100	Manga109
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
CARN [1]	×2	1592K	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
IMDN [12]		694K	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
AWSRN-M [38]		1063K	38.04/0.9605	33.66/0.9181	32.21/ 0.9000	32.23/0.9294	38.66/0.9772
MADNet [16]		878K	37.85/0.9600	33.38/0.9161	32.04/0.8979	31.62/0.9233	-
LAPAR-A [18]		548K	38.01/0.9605	33.62/0.9183	32.19/0.8999	32.10/0.9283	38.67/0.9772
RFDN [22]		534K	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278	38.88/0.9773
GLADSR [42]		812K	37.99/0.9608	33.63/0.9179	32.16/0.8996	32.16/0.9283	-
LatticeNet+ [28]		756K	38.15/0.9610	33.78/0.9193	32.25/0.9004	32.29/0.9291	-
SMSR [39]		985K	38.00/0.9601	33.64/0.9179	32.17/0.8990	32.19/0.9284	38.76/0.9771
DRSAN [32]		690K	38.11/0.9609	33.64/0.9185	32.21/0.9005	32.35/0.9304	-
LatticeNet-CL [27]		756K	38.09/0.9608	33.70/0.9188	32.21/0.9000	32.29/0.9291	-
SwinIR-light [20]		878K	38.14/0.9611	33.86/0.9206	**32.31**/0.9012	32.76/**0.9340**	39.12/0.9783
ESRT [26]		677K	38.03/0.9600	33.75/0.9184	32.25/0.9001	32.58/0.9318	39.12/0.9774
ELAN-light [43]		582K	38.17/0.9611	**33.94**/0.9207	32.30/0.9012	32.76/**0.9340**	39.11/0.9782
SPIN (Ours)		497K	**38.20/0.9615**	33.90/**0.9215**	**32.31/0.9015**	**32.79/0.9340**	**39.18/0.9784**
CARN [1]	×3	1592K	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.43/0.9427
IMDN [12]		703K	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
AWSRN-M [38]		1143K	34.42/0.9275	30.32/0.8419	29.13/0.8059	28.26/0.8545	33.64/0.9450
MADNet [16]		930K	34.16/0.9253	30.21/0.8398	28.98/0.8023	27.77/0.8439	-
LAPAR-A [18]		594K	34.36/0.9267	30.34/0.8421	29.11/0.8054	28.15/0.8523	33.51/0.9441
RFDN [22]		541K	34.41/0.9273	30.34/0.8420	29.09/0.8050	28.21/0.8525	33.67/0.9449
GLADSR [42]		821K	34.41/0.9272	30.37/0.8418	29.08/0.8050	28.24/0.8537	-
LatticeNet+ [28]		765K	34.53/0.9281	30.39/0.8424	29.15/0.8059	28.33/0.8538	-
SMSR [39]		993K	34.40/0.9270	30.33/0.8412	29.10/0.8050	28.25/0.8536	33.68/0.9445
DRSAN [32]		740K	34.50/0.9278	30.39/0.8437	29.13/0.8065	28.35/0.8566	-
LatticeNet-CL [27]		765K	34.46/0.9275	30.37/0.8422	29.12/0.8054	28.23/0.8525	-
SwinIR-light [20]		886K	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624	33.98/0.9478
ESRT [26]		770K	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574	33.95/0.9455
ELAN-light [43]		590K	34.64/0.9288	30.55/0.8463	29.21/0.8081	28.69/0.8624	34.00/0.9478
SPIN (Ours)		569K	**34.65/0.9293**	**30.57/0.8464**	**29.23/0.8089**	**28.71/0.8627**	**34.24/0.9489**
CARN [1]	×4	1592K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.42/0.9070
IMDN [12]		715K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
AWSRN-M [38]		1254K	32.21/0.8954	28.65/0.7832	27.60/0.7368	26.15/0.7884	30.56/0.9093
MADNet [16]		1002K	31.95/0.8917	28.44/0.7780	27.47/0.7327	25.76/0.7746	-
LAPAR-A [18]		659K	32.15/0.8944	28.61/0.7818	27.61/0.7366	26.14/0.7871	30.42/0.9074
RFDN [22]		550K	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
GLADSR [42]		826K	32.14/0.8940	28.62/0.7813	27.59/0.7361	26.12/0.7851	-
LatticeNet+ [28]		777K	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873	-
SMSR [39]		1006K	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	30.54/0.9085
DRSAN [32]		730K	32.30/0.8954	28.66/0.7838	27.61/0.7381	26.26/0.7920	-
LatticeNet-CL [27]		777K	32.30/0.8958	28.65/0.7822	27.59/0.7365	26.19/0.7855	-
SwinIR-light [20]		897K	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
ESRT [26]		751K	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100
ELAN-light [43]		601K	32.43/0.8975	28.78/0.7858	27.69/0.7406	26.54/0.7982	30.92/0.9150
SPIN (Ours)		555K	**32.48/0.8983**	**28.80/0.7862**	**27.70/0.7415**	**26.55/0.7998**	**30.98/0.9156**

non-local attention are unable to reconstruct damaged textures accurately. For example, when comparing the reconstruction results for image 'B100/148026', it is evident that our model produces results that are very close to the HR, whereas other competitive SISR models without non-local attention such as CARN [1] and IMDN [12] are not suited for recovering such severely damaged regions.

Additionally, when compared with other attention-based deep SISR methods like ESRT [26], SwinIR-light [20] and ELAN-light [43], our SPIN model still maintains superior reconstruction quality. Besides, for the image 'Urban100/img020', even without much textural information, our method can also accurately recover the damaged image.

5. Ablation Study

We further conduct ablation studies to better understand and evaluate each component in the proposed SPIN. For fair

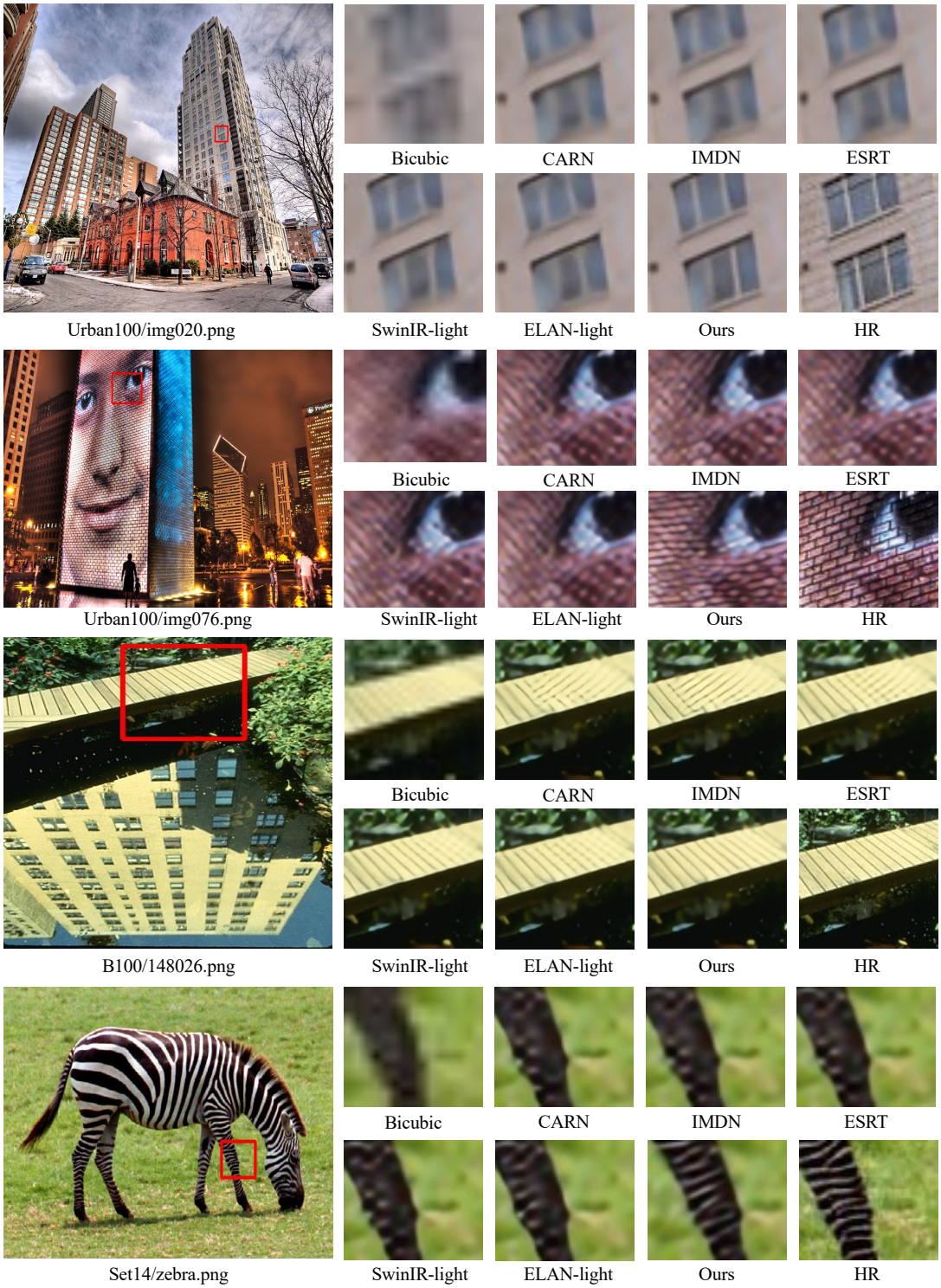

Figure 6. Qualitative comparison of state-of-the-art classic and lightweight Transformer-based SR models for $\times 4$ upscaling task. The Ours(SPIN) can restore more accurate and sharper details than the other models.

comparisons with the designed baselines, we implement all experiments based on $\times 4$ SPIN and train them under the same setting. The experimental results in Table 2 are measured on DIV2K-val [36] and Manga109 [31] datasets.

Effectiveness of ISPA and SPCA. The ISPA module and the SPCA module perform an important role in our method

Table 2. Average PSNR/SSIM comparison of different settings of our model under $\times 4$ setting.

Methods	Set5	Set14	BSDS100	Urban100	Manga109
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Only ISPA	32.36/0.8967	28.69/0.7828	27.62/0.7380	26.30/0.7919	30.63/0.9110
Only SPCA	32.31/0.8961	28.68/0.7831	27.63/0.7385	26.34/0.7928	30.62/0.9111
Parallel	32.45/0.8975	28.80/0.7851	27.69/0.7401	26.52/0.7972	30.96/0.9143
SVQ	32.30/0.8962	28.71/0.7831	27.63/0.7380	26.28/0.7909	30.59/0.9104
Patch	32.40/0.8972	28.70/0.7840	27.64/0.7395	26.34/0.7944	30.70/0.9123
Full	**32.48/0.8983**	**28.80/0.7862**	**27.70/0.7415**	**26.55/0.7998**	**30.98/0.9156**

to capture the long-range and short-range information for recovering damaged images. To evaluate the effectiveness of the proposed modules, we show their performance in Table 2. Specifically, we evaluate the setting of using only the ISPA module or the SPCA module. To guarantee a fair comparison with our base setting, we use the 12 blocks to ensure a similar number of parameters with our final model. It is evident that utilizing only one module results in a decline in performance, as it lacks either short-range or long-range information. Moreover, we also try to adopt the parallel setting, where the inter- and intra-interaction are performed in a parallel way in the block. As we can see, this way obtains slightly inferior performance to our final setting, but still largely outperforms the above two settings, demonstrating the necessity of simultaneously capturing long-range and short-range information. In our final setting, we choose the sequential implementation due to its better performance.

Pixel Aggregation. To evaluate the effectiveness of our proposed pixel aggregation, we validate two other pixel aggregation strategies in Table 2. The first strategy is to use the process of soft vector quantization (SVQ) or the Gaussian Mixture Model (GMM). Specifically, soft vector quantization is similar to our superpixel aggregation, but without the restriction of only calculating similarity among neighbor pixels. Therefore, the affinities between pixels and superpixels are non-local.

As shown, this strategy actually obtains slightly poor performance than ours. The reason may be that soft vector quantization places equal emphasis on all pixels in the image, without considering the spatial relationships between them, which could lead to ambiguousness between regions. As stated in LAM [7], the important pixels of each pixel in the input LR image with respect to the SR image are usually located in the neighborhood. Furthermore, we also try to utilize only the local attention module in our method, which actually resembles Swin Transformer [25] but without window-shift. The result is predictably inferior to ours since the regular patches usually destroy the structural information of the object, boundaries, *etc.*, demonstrating the effectiveness of our method.

Depth of the network. We evaluate the influence of the depth of the proposed network. We mainly change the number of blocks of the network, from 4 blocks to 10 blocks.

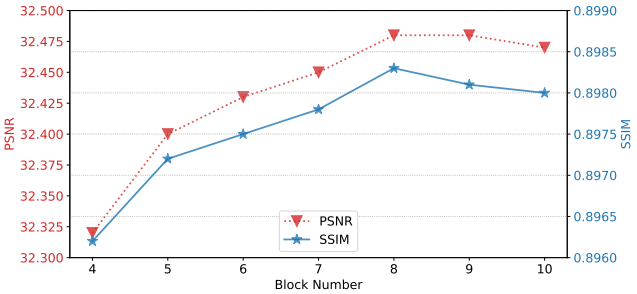

Figure 7. Performance (PSNR and SSIM) of SPIN under $\times 4$ setting with different numbers of blocks on the Set5 dataset.

As shown in Fig. 7, with the number of blocks increasing, the performance of our network is also improved. However, when the number of blocks is larger than 8, the performance starts to drop. We think the reason may be that the over-parameterized network overfits the training data, leading to poor generalization ability on other benchmarks.

6. Conclusion

In this paper, we have proposed a novel approach called the Super Token Interaction Network (SPIN), which leverages superpixels to group local similar pixels into interpretable local regions. Our method employs intra-superpixel attention to facilitate local information interaction within irregular local superpixel areas, while the superpixel cross-attention module facilitates long-range information interaction via the surrogation of superpixels. Extensive experiments demonstrate that SPIN outperforms state-of-the-art super-resolution methods in terms of accuracy and lightweight. In addition, the proposed method offers a promising solution to the challenge of processing entire images with interpretable region division.

Acknowledgement

This work is supported by the National Key R&D Program of China under Grant 2022YFB3103504, National Natural Science Foundation of China (No. 62025604, 62172409, 62261160653), Shenzhen Science and Technology Program (No. 20220016, RCYX20221008092849068).

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *European Conference on Computer Vision*, pages 252–268, 2018.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10, 2012.

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.

[6] 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.

[7] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021.

[8] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units. *arXiv preprint arXiv:1606.08415*, 2016.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[10] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3911–3927, 2019.

[11] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015.

[12] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the ACM International Conference on Multimedia*, pages 2024–2032, 2019.

[13] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *European Conference on Computer Vision*, pages 352–368, 2018.

[14] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[16] Rushi Lan, Long Sun, Zhenbing Liu, Huimin Lu, Cheng Pang, and Xiaonan Luo. Madnet: a fast and lightweight network for single-image super resolution. *IEEE Transactions on Cybernetics*, 51(3):1443–1453, 2020.

[17] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016.

[18] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *Advances in Neural Information Processing Systems*, 33:20343–20355, 2020.

[19] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–843, 2022.

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision*, pages 1833–1844, 2021.

[21] Jie Liu, Chao Chen, Jie Tang, and Gangshan Wu. From coarse to fine: Hierarchical pixel integration for lightweight image super-resolution. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 1666–1674, 2023.

[22] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *European Conference on Computer Vision*, pages 41–55, 2020.

[23] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *European Conference on Computer Vision*, pages 686–703, 2018.

[24] 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 *IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.

[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*, 2021.

[26] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 457–466, 2022.

[27] Xiaotong Luo, Yanyun Qu, Yuan Xie, Yulun Zhang, Cuihua Li, and Yun Fu. Lattice network for lightweight image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[28] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 272–289, 2020.

[29] 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 *IEEE International Conference on Computer Vision*, pages 416–423, 2001.

[30] 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(20):21811–21838, 2017.

[31] Yusuke Matsui, Kota Ito, Yuji Aramaki, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *arXiv preprint arXiv:1510.04389*, 2015.

[32] Karam Park, Jae Woong Soh, and Nam Ik Cho. Dynamic residual self-attention network for lightweight single image super-resolution. *IEEE Transactions on Multimedia*, 2021.

[33] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 11908–11915, 2020.

[34] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[35] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015.

[36] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[38] Chaofeng Wang, Zheng Li, and Jun Shi. Lightweight image super-resolution with adaptive weighted learning network. *arXiv preprint arXiv:1904.02358*, 2019.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4917–4926, 2021.

[40] Hao Zeng, Qingjie Liu, Mingming Zhang, Xiaoqing Han, and Yunhong Wang. Semi-supervised hyperspectral image classification with graph clustering convolutional networks. *arXiv preprint arXiv:2012.10932*, 2020.

[41] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proceedings of the International Conference on Curves and Surfaces (ICCS)*, pages 711–730, 2010.

[42] Xinyan Zhang, Peng Gao, Sunxiangyu Liu, Kongya Zhao, Guitao Li, Liuguo Yin, and Chang Wen Chen. Accurate and efficient image super-resolution via global-local adjusting dense network. *IEEE Transactions on Multimedia*, 23:1924–1937, 2021.

[43] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, pages 649–667. Springer, 2022.

[44] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018.