HMANet: Hybrid Multi-Axis Aggregation Network for Image Super-Resolution

Shu-Chuan Chu[1], Zhi-Chao Dou[1], Jeng-Shyang Pan[2,*], Shaowei Weng[3], Junbao Li[4]
[1]College of Computer Science and Engineering, Shandong University of Science and Technology
[2]School of Artificial Intelligence, Nanjing University of Information Science and Technology
[3]School of Information Engineering, Guangdong University of Technology
[4]School of Electronic and Information Engineering, Harbin Institute of Technology
scchu0803@gmail.com, douzhichao2021@163.com, jengshyangpan@gmail.com,
wswweiwei@126.com, lijunbao@hit.edu.cn

Abstract

Transformer-based methods have demonstrated excellent performance on super-resolution visual tasks, surpassing conventional convolutional neural networks. However, existing work typically restricts self-attention computation to non-overlapping windows to save computational costs. This means that Transformer-based networks can only use input information from a limited spatial range. Therefore, a novel Hybrid Multi-Axis Aggregation network (HMA) is proposed in this paper to exploit feature potential information better. HMA is constructed by stacking Residual Hybrid Transformer Blocks(RHTB) and Grid Attention Blocks(GAB). On the one side, RHTB combines channel attention and self-attention to enhance non-local feature fusion and produce more attractive visual results. Conversely, GAB is used in cross-domain information interaction to jointly model similar features and obtain a larger perceptual field. For the super-resolution task in the training phase, a novel pre-training method is designed to enhance the model representation capabilities further and validate the proposed model's effectiveness through many experiments. The experimental results show that HMA outperforms the state-of-the-art methods on the benchmark dataset. We provide code and models at https://github.com/korouuuuu/HMA.

1. Introduction

Natural images have different features, such as multi-scale pattern repetition, same-scale texture similarity, and structural similarity [44]. Deep neural networks can exploit these properties for image reconstruction. However, it cannot capture the complex dependencies between distant elements due to the limitations of CNN's fixed local receptive field and parameter sharing mechanism, thus limiting its ability to model long-range dependencies [24]. Recent research has introduced the self-attention mechanism to computer vision [19, 22]. Researchers have used the long-range dependency modeling capability and multi-scale processing advantages in the self-attention mechanism to enhance the joint modeling of different hierarchical structures in images.

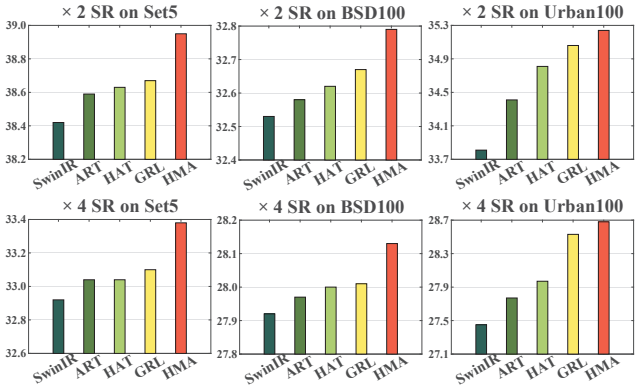

Figure 1. The performance of the proposed HMA is compared with the state-of-the-art SwinIR, ART, HAT, and GRL methods in terms of PSNR(dB). Our method outperforms the state-of-the-art methods by 0.1dB∼1.4dB.

Although Transformer-based methods have been successfully applied to image restoration tasks, there are still some things that could be improved. Existing window-based Transformer networks restrict the self-attention computation to a dense area. This strategy obviously leads to a limited receptive field and does not fully utilize the feature information from the original image. For the purpose of generating images with more realistic details, researchers consider using GAN networks or inputting the reference information to provide additional feature information [4, 11, 32]. However, the network may generate unreasonable results if the input additional feature information

does not match.

In order to overcome the above problems, we propose a hybrid multiaxial aggregation network called HMA in this paper. HMA combines channel attention and self-attention, which utilizes channel attention's global information perception capability to compensate for self-attention's shortcomings. In addition, we introduce a grid attention block to achieve the modeling across distances in images. Meanwhile, to further excite the potential performance of the model, we customize a pre-training strategy for the super-resolution task. Benefiting from these designs, as shown in Fig. 1, our proposed method can effectively improve the model performance (0.1dB$\sim$1.4dB). The main contributions of this paper are summarised as follows:

- We propose a novel Hybrid Multi-axis Aggregation network (HMA). The HMA comprises Residual Hybrid Transformer Blocks (RHTB) and Grid Attention Blocks (GAB), aiming to consider both local and global receptive fields. GAB models similar features at different image scales to achieve better reconstruction.
- We further propose a pre-training strategy for super-resolution tasks that can effectively improve the model's performance using a small training cost.
- Through a series of comprehensive experiments, our findings substantiate that HMA attains a state-of-the-art performance across various test datasets.

2. Related Works

2.1. CNN-Based SISR

CNN-based SISR methods have made significant progress in recovering image texture details. SRCNN [35] solved the super-resolution task for the first time using CNNs. Subsequently, in order to enhance the network learning ability, VDSR [14] introduced the residual learning idea, which effectively solved the problem of gradient vanishing in deep network training. In SRGAN [16], Christian Ledig et al. proposed to use generative adversarial networks to optimize the process of generating super-resolution images. The generator of SRGAN learns the mapping from low-resolution images to high-resolution images and improves the quality of the generated images by adversarial training. ESRGAN [34] introduces Residual in Residual Dense Block (RRDB) as the basic network unit and reduces the perceptual loss by using features before activation so that the images generated by EARGAN [34] have a more realistic natural texture. In addition, new network architectures are constantly being proposed by researchers to recover more realistic super-resolution image details [3, 8, 37].

2.2. Transformer-Based SISR

In recent years, Transformer-based SISR has become an emerging research direction in super-resolution, which utilizes the Transformer architecture to achieve image mapping from low to high resolution. Among them, the Swin Transformer-based SwinIR [19] model achieves the best performance beyond CNN-based on image restoration tasks. In order to further investigate the effect of pre-training on its internal representation, Chen et al. proposed a novel Hybrid Attention Transformer (HAT) [6]. The HAT introduces overlapping cross-attention blocks to enhance the interactions between neighboring windows' features, thus aggregating the cross-window information better. Our proposed HMA network learns similar feature representations through grid multiplexed self-attention and combines it with channel attention to enhance non-local feature fusion. Therefore, our method can provide additional support for image restoration through similar features in the original image.

2.3. Self-similarity based image restoration

Natural images usually have similar features in different hierarchies, and many SISR methods based on CNN have achieved remarkable results by exploring self-similarity [13, 28, 30]. In order to reduce the computational complexity, the computation of self-similarity is usually restricted to local areas. The researchers also proposed to extend the search space by geometric transformations to increase the global feature interactions [12]. In Transformer-based SISR, the computational complexity of non-local self-attention increases quadratically with the growth of image size. Recent studies have proposed using sparse global self-attention to reduce the complexity [39]. Sparse global self-attention allows more feature interactions while reducing computational complexity. The proposed GAB adopts the idea of sparse self-attention to increase global feature interactions while balancing the computational complexity. Our method allows joint modeling using similar features to generate better reconstructed images.

3. Motivation

Image self-similarity is vital in image processing, computer vision, and pattern recognition. Image self-similarity is usually characterized by multi-scale and geometric transformation invariance. Image self-similarity can be local or global. Local self-similarity means that one area of an image is similar to another, and global self-similarity means that there is self-similarity between multiple areas within the whole image. Fig. 2 shows that texture units may be repeated at regular intervals. Similarity modeling of features at different locations (*e.g.*, yellow rectangle) in the input image can provide a reference for image reconstruction in the green rectangle when recovering the features in the green rectangle. Image self-similarity has been explored with satisfactory performance in classical super-resolution algorithms.

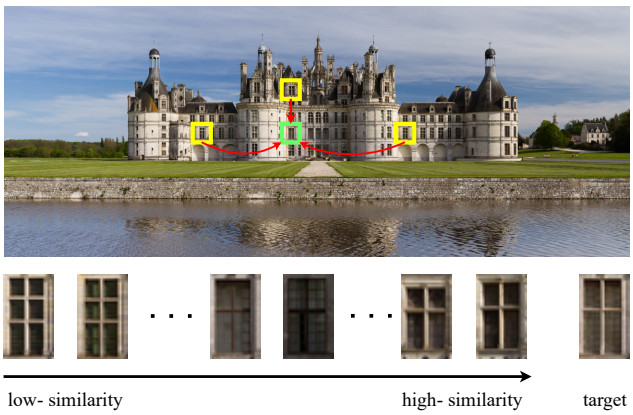

Figure 2. Example of image similarity based on non-local textures. Image from DIV2K:0830.

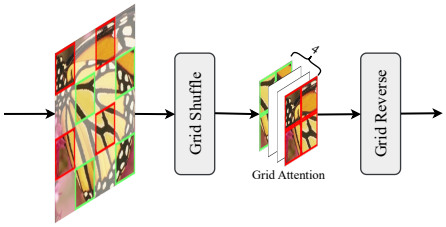

Figure 3. Grid Attention Strategies. We divide the feature map into sparse areas at specific intervals ($K = 4$) and then compute the self-attention within each set of sparse areas.

Swin Transformer [21] employs cross-window connectivity and multi-head attention mechanisms to deal with the long-range dependency modeling problem. However, Swin Transformer can only use a limited range of pixels when dealing with the SR task and cannot effectively use image self-similarity to enhance the reconstruction effect. For the purpose of increasing the range of pixels utilized by the Swin Transformer, we try to enhance the long-range dependency modeling capability of the Swin Transformer with sparse attention. As shown in Fig. 3, we suggest adding grid attention to increase the interaction between patches. The feature map is divided into K^2 groups according to the interval size K, and each group contains $\frac{H}{K} \times \frac{W}{K}$ Patches. After the grid shuffle, we can get the feature $F_G \in \mathbb{R}^{\frac{H}{K} \times \frac{W}{K} \times C}$ and compute the self-attention in each group.

Not all areas in a natural image have similarity relationships. In order to avoid the non-similar features from damaging the original features, we introduce the global feature-based interaction feature $G \in \mathbb{R}^{\frac{H}{K} \times \frac{W}{K} \times \frac{C}{2}}$ and the window-based self-attention mechanism ((S)W-MSA) to capture the similarity relationship of the whole image while modeling the similar features by Grid Multihead Self-Attention (Grid-MSA). The detailed computational procedure is described in Sec. 4.3.

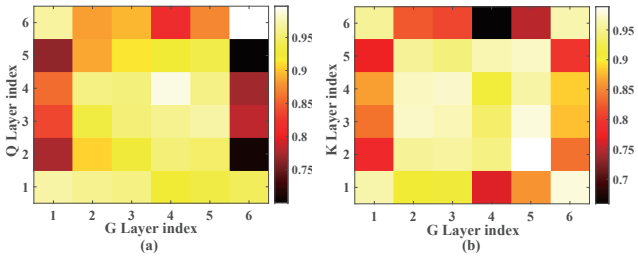

Figure 4. (a) CKA similarity between all G and Q in the $\times 2$ SR model. (b) CKA similarity between all G and K in the $\times 2$ SR model.

To make Grid-MSA work better, we must ensure the similarity between interaction features and query/key structure. Therefore, we introduce centered kernel alignment (CKA) [15] to study the similarity between features. It can be observed that the CKA similarity maps in Fig. 4 presents a diagonal structure, *i.e.*, there is a close structural similarity between the interaction features and the query/keyword in the same layer (CKA>0.9). Therefore, interaction features can be a medium for query/key interaction with global features in Grid-MSA. With the benefit of these designs, our network is able to reconstruct the image taking full advantage of the pixel information in the input image.

4. Proposed Method

As shown in Fig. 5, HMA consists of three parts: shallow feature extraction, deep feature extraction, and image reconstruction. Among them, RHTB is a stacked combination of multiple Fused Attention Blocks (FAB) and GAB. The RHTB is constructed by residual in residual structure. We will introduce these methods in detail in the following sections.

4.1. Overall Architecture

For a given low-resolution (LR) input $I_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$ (H, W, and C_{in} are the height, width, and number of input channels of the input image, respectively), we first extract the shallow features of the I_{LR} using a convolutional layer that maps the I_{LR} to high-dimensional features $F_0 \in \mathbb{R}^{H \times W \times C}$:

$$F_0 = H_{Conv}(I_{LR}), \tag{1}$$

where $H_{Conv}(\cdot)$ denotes the convolutional layer and C denotes the number of channels of the intermediate layer features. Subsequently, we input F_0 into $H_{DF}(\cdot)$, a deep feature extraction group consisting of M RHTBs and a 3×3 convolution. Each RHTB consists of a stack of N FABs, a GAB, and a convolutional layer with residual connections. Then, we fuse the deep features $F_D \in \mathbb{R}^{H \times W \times C}$ with F_0 by element-by-element summation to obtain F_{REC}. Fi-

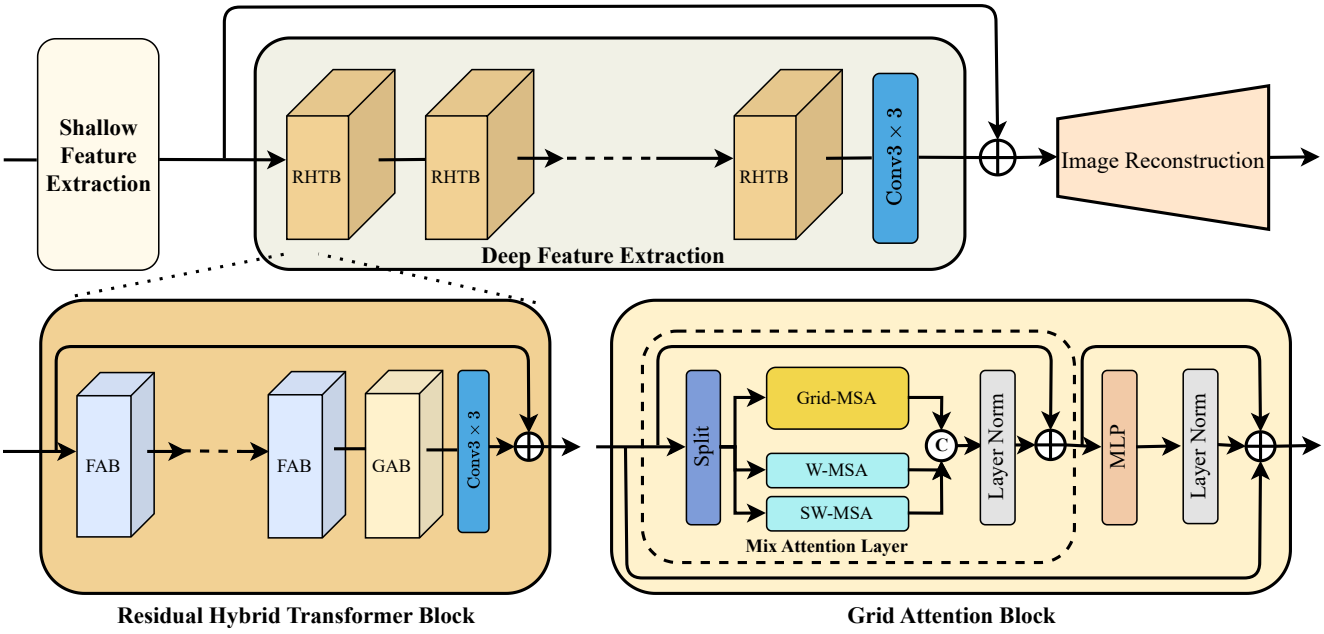

Figure 5. The overall architecture of HMA and the structure of RHTB and GAB.

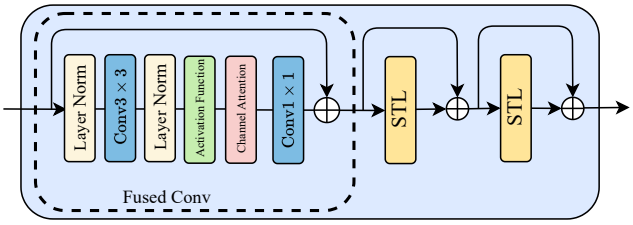

Figure 6. The architecture of FAB.

nally, we reconstruct F_{REC} into a high-resolution image I_{HR}:

$$I_{HR} = H_{REC}(H_{DF}(F_0) + F_0), \tag{2}$$

where $H_{REC}(\cdot)$ denotes the reconstruction module.

4.2. Fused Attention Block (FAB)

Many studies have shown that adding appropriate convolution in the Transformer can further improve network trainability [25, 36, 41]. Therefore, we insert a convolutional layer before the Swin Transformer Layer (STL) to enhance the network learning capability. As shown in Fig. 6, we insert the Fused Conv module ($H_{Fuse}(\cdot)$) with inverted bottlenecks and squeezed excitations before the STL to achieve enhanced global information fusion. Note that we use Layer Norm instead of Batch Norm in Fused Conv to avoid the impact on the contrast and color of the image. The computational procedure of Fused Conv is:

$$F_{Fuse} = H_{Fuse}(F_{F_{in}}) + F_{F_{in}}, \tag{3}$$

where $F_{F_{in}}$ represents the input features, and F_{Fuse} represents the features output from the Fused Conv block. Then, we add two successive STL after Fused Conv. In the STL, we follow the classical design in SWinIR, including Window-based self-attention (W-MSA) and Shifted Window-based self-attention (SW-MSA), and Layer Norm. The computation of the STL is as follows:

$$F_N = (S)W - MSA(LN(F_{W_{in}})) + F_{W_{in}}, \tag{4}$$

$$F_{out} = MLP(LN(F_N)) + F_N, \tag{5}$$

where $F_{W_{in}}$, F_N, and F_{out} indicate the input features, the intermediate features, and the output of the STL, respectively, and MLP denotes the multilayer perceptron. We split the feature map uniformly into $\frac{H \times W}{M^2}$ windows in a non-overlapping manner for efficient modeling. Each window contains M $\times$ M Patch. The self-attention of a local window is calculated as follows:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}} + B)V, \tag{6}$$

where Q, K, $V \in \mathbb{R}^{M^2 \times d}$ are obtained by the linear transformation of the given input feature $F_W \in \mathbb{R}^{M^2 \times C}$. The d and B represent the dimension and relative position encoding of the query/key, respectively.

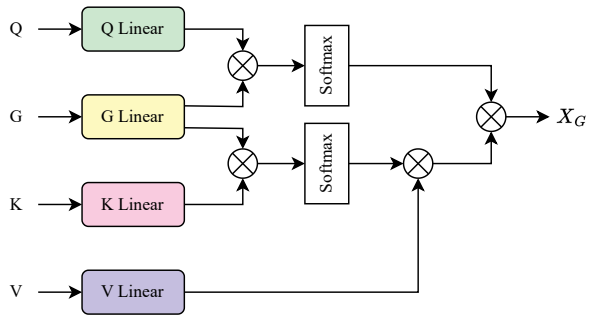

Figure 7. The computational flowchart of Grid Attention.

As shown in Fig. 6, Fused Conv expands the channel using a convolutional kernel of size 3 with a default expansion rate of 6. At the same time, a squeeze-excitation (SE) layer with a shrink rate of 0.5 is used in the channel attention layer. Finally, a convolutional kernel of size 1 is used to recover the channel.

4.3. Grid Attention Block(GAB)

We introduce GAB to model cross-area similarity for enhanced image reconstruction. The GAB consists of a Mix Attention Layer (MAL) and an MLP layer. Regarding the MAL, we first split the input feature F_{in} into two parts by channel: $F_G \in \mathbb{R}^{H\times W\times \frac{C}{2}}$ and $F_W \in \mathbb{R}^{H\times W\times \frac{C}{2}}$. Subsequently, we split F_W into two parts by channel again and input them into W-MSA and SW-MSA, respectively. Meanwhile, F_G is input into Grid-MSA.The computation process of MAL is as follows:

$$X_{W_1} = W - MSA(F_{W_1}), \tag{7}$$

$$X_{W_2} = SW - MSA(F_{W_2}), \tag{8}$$

$$X_G = Grid - MSA(F_G), \tag{9}$$

$$X_{MAL} = LN(Cat(X_{W_1}, X_{W_2}, X_G)) + F_{in}, \tag{10}$$

where X_{W_1}, X_{W_2}, and X_G are the output features of W-MSA, SW-MSA, and Grid-MSA, respectively. It should be noted that we adopt the post-norm method in GAB to enhance the network training stability. For a given input feature F_{in}, the computation process of GAB is:

$$F_M = LN(MAL(F_{in}) + F_{in}, \tag{11}$$

$$F_{out} = LN(MLP(F_M)) + F_M, \tag{12}$$

It is shown in Fig. 7 that the Q, K, and V are obtained from the input feature F_G after grid shuffle when Grid-MSA is used. $G \in \mathbb{R}^{H\times W\times \frac{C}{2}}$ is obtained from the linear transformation of the input feature F_{in} after grid shuffle. For Grid-MSA, the self-attention is calculated as follows:

$$\hat{X} = SoftMax(\frac{GK^T}{d} + B)V, \tag{13}$$

$$Attention(Q, G, \hat{X}) = SoftMax(\frac{QG^T}{d} + B)\hat{X}, \tag{14}$$

where $\hat{X}$ is the intermediate feature obtained by computing the self-attention from G, K, and V.

4.4. Pre-training strategy

Pre-training plays a crucial role in many visual tasks [1, 33]. Recent studies have shown that pre-training can also capture significant gains in low-level visual tasks. IPT [5] handles different visual tasks by sharing the Transformer module with different head and tail structures. EDT [17] improves the performance of the target task by multi-task pre-training. HAT [6] pre-trains the super-resolution task using a larger dataset directly on the same task. Instead, we propose a pre-training method more suitable for super-resolution tasks, *i.e.*, increasing the gain of pre-training by sharing model parameters among pre-trained models with different degradation levels. We first train a $\times 2$ model as the initial parameter seed when pre-training on the ImageNet dataset and then use it as the initialization parameter for the $\times 3$ model. Then, train the final $\times 2$ and $\times 4$ models using the trained $\times 3$ model as the initialization parameters of the $\times 2$ and $\times 4$ models. After the pre-training, the $\times 2$, $\times 3$, and $\times 4$ models are fine-tuned on the DF2K dataset. The proposed strategy can bring more performance improvement, although it pays an extra training cost (training a $\times 2$ model).

5. Experiments

5.1. Experimental Setup

We use DF2K dataset (DIV2K [20] dataset merged with Flicker [31] dataset) as the training set. Meanwhile, we use ImageNet [10] as the pre-training dataset. For the structure of HMA, the number of RHTB and FAB is set to 6, the window size is set to 16, the number of channels is set to 180, and the number of attentional heads is set to 6. The number of attentional heads is 3 and 2 for Grid-MSA and (S)W-MSA in GAB, respectively. We evaluate on the Set5 [2], Set14 [38], BSD100 [26], Urban100 [13], and Manga109 [27] datasets. Both PSNR and SSIM evaluations are computed on the Y channel.

5.2. Training Details

Low-resolution images are generated by down-sampling using bicubic interpolation in MATLAB. We cropped the

	Baseline			
Fused Conv	✗	✗	✓	✓
GAB	✗	✓	✗	✓
PSNR/SSIM	27.49/0.8271	28.30/0.8370	28.37/0.8375	28.42/0.8450

Table 1. Ablation study on the proposed Fused Conv and GAB.

expansion rate	2	4	6	8
PSNR	28.30	28.34	28.37	28.39

Table 2. Ablation study on expansion rate of Fused Conv.

shrink rate	2	4	6	8
PSNR	27.39	28.37	28.32	28.28

Table 3. Ablation study on shrink rate of Fused Conv.

dataset into 64×64 patches for training. Furthermore, we employed horizontal flipping and random rotation for data augmentation. The training batch size is set to 32. During pre-training with ImageNet [10], the total number of training iterations is set to 800K (1K represents 1000 iterations), the learning rate was initialized to 2×10^{-4} and halved at [300K, 500K, 650K, 700K, 750K]. We optimized the model using the Adam optimizer (with β_1=0.9 and β_2=0.99). Subsequently, we fine-tuned the model on the DF2K dataset. The total number of training iterations is set to 250K, and the initial learning rate was set to 5×10^{-6} and halved at [125K, 200K, 230K, 240K].

5.3. Ablation Study

5.3.1 Effectiveness of Fused Conv and GAB

We experimentally demonstrate the effectiveness of Fused Conv and GAB proposed in this paper. The experiments are conducted on the Urban100 [13] dataset to evaluate PSNR/SSIM. The evaluation report is presented in Tab. 1. Compared with the baseline results, the best performance is achieved when both modules are used. In contrast, the performance gains obtained when using the Fused Conv or GAB modules alone were not as good as when using them simultaneously. Although the performance of the sole use of the Fused Conv module is slightly higher than the sole use of the GAB module, the GAB module is applied for global image interaction, which can effectively improve the model SSIM value and better restore the image's texture. This means that our proposed method not only performs well on PSNR but is also excellent in restoring the image's visual effect.

5.3.2 Effects of the expansion rate and shrink rate

Tab. 2 and Tab. 3 show the effect of expansion and shrink rates on performance, respectively. The data in the table shows that the expansion rate is directly proportional to the performance, while the shrink rate is inversely proportional. Although the performance keeps increasing when the expansion rate increases, the number of parameters and the amount of computation increase quadratically. In order to balance the model performance and computation, we set the expansion rate to 6. Similarly, we set the shrink rate to 2 to get a model with as little computation as possible.

5.4. Comparison with State-of-the-Art Methods

5.4.1 Quantitative comparison

Tab. 4 shows the comparative results of our method with the state-of-the-art methods on PSNR and SSIM: EDSR [20], RCAN [40], SAN [9], IGNN [42], NLSA [29], IPT [5], SwinIR [19], ESRT [23], SRFoemer [43] EDT [17], HAT [6], HAT-L [6], and GRL [18]. In Tab. 4, it can be seen that the proposed method achieves the best performance on almost all scales on five datasets. Specifically, HMA outperforms SwinIR by 0.2dB∼1.43dB on all scales. In particular, on Urban100 [13] and MANGA109 [27] that contain a large number of repetitive textures, HMA improves by 0.98dB∼1.43dB compared to SwinIR. It is important to note that both HAT and GRL [18] introduce the channel attention in the model. However, both HAT [6] and GRL [18] perform less well than HMA, which proves the effectiveness of our proposed method.

5.4.2 Visual comparison

We provide some of the visual comparison results in Fig. 8. The comparison results are selected from the Urban100 [13] dataset: "img_011", "img_033", "img_046", "img_062", "img_067" and "img_092". In Fig. 8, PSNR and SSIM is calculated in patches marked with red boxes in the images. From the visual comparison, HMA can recover the image texture details better. Compared with other advanced methods, HMA recovers images with clearer edges. We can see many blurred areas in recovering image "img_011" and image "img_092" in other state-of-the-art methods, while HMA generates excellent visual effects. The comparison of the visual effects indicates that our proposed method also achieves a superior performance.

5.5. NTIRE 2024 Challenge

Our SR model also participated in NTIRE 2024 Image Super-Resolution (×4) [7] in the validation phase and testing phase. The respective results areshown in Tab. 5.

6. Conclusion

This study proposes a Hybrid Multi-Axis Aggregation Network (HMA) for single-image super-resolution. Our model combines Fused Convolution with self-attention

Method	Scale	Set5[2]		Set14[38]		BSD100[26]		Urban100[13]		Manga109[27]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR[20]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN[40]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN[9]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IGNN[42]	×2	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
NLSA[29]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
IPT[5]	×2	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR[19]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
ESRT[23]	×2	-	-	-	-	-	-	-	-	-	-
SRFormer[43]	×2	38.51	0.9627	34.44	0.9253	32.57	0.9046	34.09	0.9449	40.07	0.9802
EDT[17]	×2	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
HAT[6]	×2	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
GRL[18]	×2	38.67	0.9647	35.08	0.9303	32.68	0.9087	35.06	0.9505	40.67	0.9818
HMA(ours)	×2	38.79	0.9641	35.11	0.9286	32.67	0.9061	34.85	0.9493	40.73	0.9824
HAT-L†[6]	×2	38.91	0.9646	35.29	0.9293	32.74	0.9066	35.09	0.9505	41.01	0.9831
HMA†(ours)	×2	38.95	0.9649	35.33	0.9297	32.79	0.9071	35.24	0.9513	41.13	0.9836
EDSR[20]	×3	34.65	0.928	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN[40]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN[9]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
IGNN[42]	×3	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
NLSA[29]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
IPT[5]	×3	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR[19]	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
ESRT[23]	×3	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
SRFormer[43]	×3	35.02	0.9323	30.94	0.8540	29.48	0.8156	30.04	0.8865	35.26	0.9543
EDT[17]	×3	34.97	0.9316	30.89	0.8527	29.44	0.8142	29.72	0.8814	35.13	0.9534
HAT[6]	×3	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
GRL[18]	×3	-	-	-	-	-	-	-	-	-	-
HMA(ours)	×3	35.22	0.9336	31.28	0.8570	29.59	0.8682	30.65	0.8944	35.82	0.9567
HAT-L†[6]	×3	35.28	0.9345	31.47	0.8584	29.63	0.8191	30.92	0.8981	36.02	0.9576
HMA†(ours)	×3	35.35	0.9347	31.47	0.8585	29.66	0.8196	31.00	0.8984	36.10	0.9580
EDSR[20]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN[40]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN[9]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IGNN[42]	×4	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
NLSA[29]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
IPT[5]	×4	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR[19]	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
ESRT[23]	×4	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
SRFormer[43]	×4	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311	32.21	0.9271
EDT[17]	×4	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
HAT[6]	×4	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
GRL[18]	×4	33.10	0.9094	29.37	0.8058	28.01	0.7611	28.53	0.8504	32.77	0.9325
HMA(ours)	×4	33.15	0.9060	29.32	0.7996	28.05	0.7530	28.42	0.8450	32.97	0.9320
HAT-L†[6]	×4	33.30	0.9083	29.47	0.8015	28.09	0.7551	28.60	0.8498	33.09	0.9335
HMA†(ours)	×4	33.38	0.9089	29.51	0.8019	28.13	0.7562	28.69	0.8512	33.19	0.9344

Table 4. Quantitative comparison (PSNR/SSIM) with state-of-the-art methods on benchmark dataset. The top three results are marked in red, blue and green., respectively. "†" indicates that methods adopt pre-training strategy.

to better integrate different-level features during deep feature extraction. Additionally, inspired by images' inherent hierarchical structural similarity, we introduce a Grid Attention Block for modeling long-range depen-

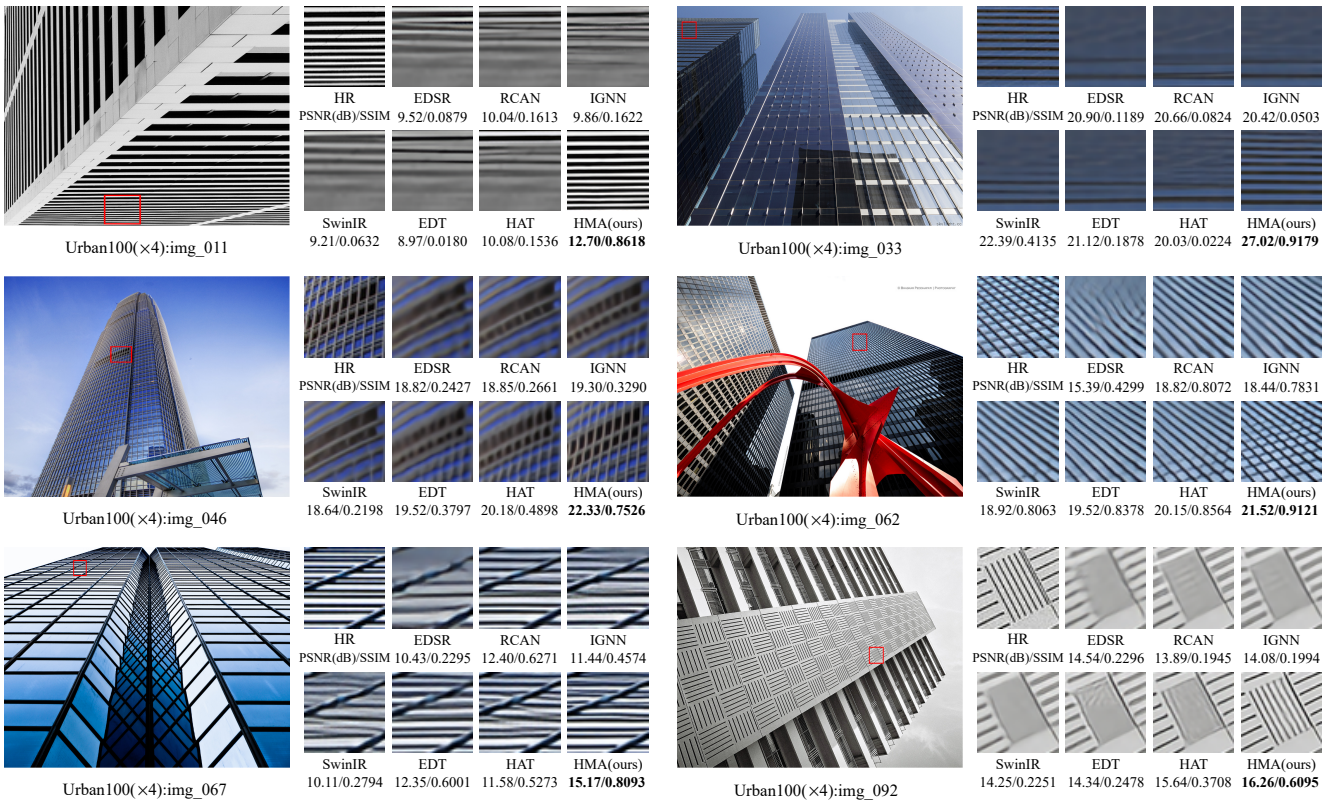

Figure 8. Visual comparison on ×4 SR. PSNR/SSIM is calculated in patches marked with red boxes in the images.

Table 5. NTIRE 2024 Challenge Results with ×4 SR in terms of PSNR and SSIM on validation phase and testing phase.

	Validation phase	Testing phase
PSNR	31.44	31.18
SSIM	0.85	0.86

dencies. The proposed network enhances multi-level structural similarity modeling by combining sparse attention with window attention. For the super-resolution task, we also designed a pre-training strategy specifically to stimulate the model's potential capabilities further. Extensive experiments demonstrate that our proposed method outperforms state-of-the-art approaches on benchmark datasets for single-image super-resolution tasks.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision – ECCV 2022*, pages 348–367, Cham, 2022. Springer Nature Switzerland. 5

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5, 7

[3] Bahri Batuhan Bilecen and Mustafa Ayazoglu. Bicubic++: Slim, slimmer, slimmest - designing an industry-grade super-resolution network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1623–1632, 2023. 2

[4] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *Computer Vision – ECCV 2022*, pages 325–342, Cham, 2022. Springer Nature Switzerland. 1

[5] 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 (CVPR)*, pages 12299–12310, 2021. 5, 6, 7

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, 2023. 2, 5, 6, 7

[7] Zheng Chen, Zongwei Wu, Eduard-Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. Ntire 2024 challenge on image super-resolution (x4): Methods and results. In *Computer Vision and Pattern Recognition Workshops*, 2024. 6

[8] Marcos V. Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *Computer Vision – ECCV 2022 Workshops*, pages 669–687, Cham, 2023. Springer Nature Switzerland. 2

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5, 6

[11] Weizhi Du and Shihao Tian. Transformer and gan-based super-resolution reconstruction network for medical images. *Tsinghua Science and Technology*, 29(1):197–206, 2024. 1

[12] Mehran Ebrahimi and Edward R. Vrscay. Solving the inverse problem of image zooming using "self-examples". In *Image Analysis and Recognition*, pages 117–130, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 2

[13] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5, 6, 7

[14] 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 (CVPR)*, 2016. 2

[15] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 3

[16] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[17] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 5, 6, 7

[18] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18278–18289, 2023. 6, 7

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, 2021. 1, 2, 6, 7

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5, 6, 7

[21] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 3

[22] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 457–466, 2022. 1

[23] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022. 6, 7

[24] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 1

[25] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *Computer Vision – ECCV 2022 Workshops*, pages 3–20, Cham, 2023. Springer Nature Switzerland. 4

[26] D. Martin, C. Fowlkes, D. Tal, and J. Malik. 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*, pages 416–423 vol.2, 2001. 5, 7

[27] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 5, 6, 7

[28] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[29] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3526, 2021. 6, 7

[30] Jian-Nan Su, Min Gan, Guang-Yong Chen, Jia-Li Yin, and C. L. Philip Chen. Global learnable attention for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8453–8465, 2023. 2

[31] 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 Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5

[32] Jingzhi Tu, Gang Mei, Zhengjing Ma, and Francesco Piccialli. Swcgan: Generative adversarial network combining swin transformer and cnn for remote sensing image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5662–5673, 2022. 1

[33] Shaoru Wang, Jin Gao, Zeming Li, Xiaoqin Zhang, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 35624–35641. PMLR, 2023. 5

[34] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 2

[35] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019. 2

[36] Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee, and Tae Hyun Kim. Enriched cnn-transformer feature aggregation networks for super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4956–4965, 2023. 4

[37] Eduard Zamfir, Marcos V. Conde, and Radu Timofte. Towards real-time 4k image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1522–1532, 2023. 2

[38] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 5, 7

[39] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022. 2

[40] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 6, 7

[41] Mo Zhao, Gang Cao, Xianglin Huang, and Lifang Yang. Hybrid transformer-cnn for real image denoising. *IEEE Signal Processing Letters*, 29:1252–1256, 2022. 4

[42] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *Advances in Neural Information Processing Systems*, pages 3499–3509. Curran Associates, Inc., 2020. 6, 7

[43] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12780–12791, 2023. 6, 7

[44] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984, 2011. 1