Dual Aggregation Transformer for Image Super-Resolution

Zheng Chen[1], Yulun Zhang[2]*, Jinjin Gu[3,4], Linghe Kong[1]*, Xiaokang Yang[1], Fisher Yu[2]

[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]The University of Sydney, [4]Shanghai AI Laboratory

Abstract

Transformer has recently gained considerable popularity in low-level vision tasks, including image super-resolution (SR). These networks utilize self-attention along different dimensions, spatial or channel, and achieve impressive performance. This inspires us to combine the two dimensions in Transformer for a more powerful representation capability. Based on the above idea, we propose a novel Transformer model, Dual Aggregation Transformer (DAT), for image SR. Our DAT aggregates features across spatial and channel dimensions, in the inter-block and intra-block dual manner. Specifically, we alternately apply spatial and channel self-attention in consecutive Transformer blocks. The alternate strategy enables DAT to capture the global context and realize inter-block feature aggregation. Furthermore, we propose the adaptive interaction module (AIM) and the spatial-gate feed-forward network (SGFN) to achieve intra-block feature aggregation. AIM complements two self-attention mechanisms from corresponding dimensions. Meanwhile, SGFN introduces additional non-linear spatial information in the feed-forward network. Extensive experiments show that our DAT surpasses current methods. Code and models are obtainable at https://github.com/zhengchen1999/DAT.

1. Introduction

Single image super-resolution (SR) is a traditional low-level vision task that focuses on recovering a high-resolution (HR) image from a low-resolution (LR) counterpart. As an ill-posed problem with multiple potential solutions for a given LR input, various approaches have emerged to tackle this challenge in recent years. Many of these methods utilize convolutional neural networks (CNNs) [12, 47, 10, 29]. However, the convolution adopts a local mechanism, which hinders the establishment of global dependencies and restricts the performance of the model.

Recently, Transformer proposed in natural language processing (NLP) has performed notably in multiple high-level vision tasks [13, 39, 24, 11, 7]. The core of Transformer
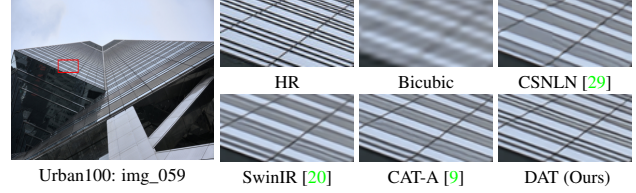
---

*Corresponding authors: Yulun Zhang, yulun100@gmail.com; Linghe Kong, linghe.kong@sjtu.edu.cn

Figure 1: Visual comparison (×4) on Urban100. CSNLN, SwinIR, and CAT-A suffer from blurring artifacts.

is the self-attention (SA) mechanism, which is capable of establishing global dependencies. This property alleviates the limitations of CNN-based algorithms. Considering the potential of Transformer, some researchers attempt to apply Transformer to low-level tasks [20, 44, 42, 9], including image SR. They explore efficient usages of Transformer on high-resolution images from different perspectives to mitigate the high complexity of global self-attention [13]. For the spatial aspect, some methods [20, 46, 9] apply local spatial windows to limit the scope of self-attention. For the channel aspect, the "transposed" attention [44] is proposed, which calculates self-attention along the channel dimension rather than the spatial dimension. These methods all exhibit remarkable results due to the strong modeling ability in their respective dimensions. Spatial window self-attention (SW-SA) is able to model fine-grained spatial relationships between pixels. Channel-wise self-attention (CW-SA) can model relationships among feature maps, thus exploiting global image information. Generally, both extracting spatial information and capturing channel context are crucial to the performance of Transformer in image SR.

Motivated by the aforementioned findings, we propose the Dual Aggregation Transformer (DAT) for image SR. Our DAT aggregates spatial and channel features via the inter-block and intra-block dual way to obtain powerful representation capability. Specifically, we alternately apply spatial window and channel-wise self-attention in successive dual aggregation Transformer blocks (DATBs). Through this alternate strategy, our DAT can capture both spatial and channel context and realize inter-block feature aggregation between different dimensions. Moreover, the two self-attention mechanisms complement each other. Spatial window self-attention enriches the spatial expression of each feature map, helping to model channel depen-

dencies. Channel-wise self-attention provides the global information between features for spatial self-attention, expanding the receptive field of window attention.

Meanwhile, since self-attention mechanisms focus on modeling global information, we incorporate convolution to self-attention in parallel, to complement Transformer with the locality. To enhance the fusion of the two branches and aggregate both spatial and channel information within a single self-attention module, we propose the adaptive interaction module (AIM). It consists of two interaction operations, spatial-interaction (S-I) and channel-interaction (C-I), which act between two branches to exchange information. Through S-I and C-I, the AIM adaptively re-weight the feature maps of two branches from the spatial or channel dimension, according to different self-attention mechanisms. Besides, with AIM, we design two new self-attention mechanisms, adaptive spatial self-attention (AS-SA) and adaptive channel self-attention (AC-SA), based on the spatial window and channel-wise self-attention, respectively.

Furthermore, another component of the Transformer block, the feed-forward network (FFN) [38], extracts features through fully-connected layers. It ignores modeling spatial information. In addition, the redundant information between channels obstructs further advances in feature representation learning. To cope with these issues, we design the spatial-gate feed-forward network (SGFN), which introduces the spatial-gate (SG) module between two fully-connected layers of FFN. The SG module is a simple gating mechanism (depth-wise convolution and element-wise multiplication). The input feature of SG is partitioned into two segments along the channel dimension for convolution and multiplicative bypass. Our SG module can complement FFN with additional non-linear spatial information and relieve channel redundancy. In general, based on AIM and SGFN, DAT can realize intra-block feature aggregation.

Overall, with the above three designs, our DAT can aggregate spatial and channel information through the inter-block and intra-block dual way to achieve strong feature expressions. Consequently, as displayed in Fig. 1, our DAT achieves superior visual results against recent state-of-the-art SR methods. Our contributions are three-fold:

- We design a new image SR model, dual aggregation Transformer (DAT). Our DAT aggregates spatial and channel features in the inter-block and intra-block dual manner to obtain powerful representation ability.

- We alternately adopt spatial and channel self-attention, realizing inter-block spatial and channel feature aggregation. Moreover, we propose AIM and SGFN to achieve intra-block feature aggregation.

- We conduct extensive experiments to demonstrate that our DAT outperforms state-of-the-art methods, while retaining lower complexity and model size.

2. Related Work

Image Super-Resolution. Deep CNN-based approaches exhibit significant efficacy in the field of image SR. SR-CNN [12] is the pioneering work, which first utilizes CNN and outperforms traditional approaches. Following this attempt, substantial dedication has been invested in deepening the layer of the network for better performance. For instance, RCAN [47] designs residual in residual structure [16] and builds a 400+ layers model. Besides, attention mechanisms [48, 30, 29, 3] in terms of spatial or channel dimensions are adopted to achieve further improvement in modeling ability. However, it is still hard for the majority of CNN-based methods to effectively model global dependencies in both spatial and channel dimensions.

Vision Transformer. Transformer demonstrates remarkable performance in high-level vision tasks [13, 39, 37]. A series of Transformer-based methods are proposed to improve the efficiency and effectiveness of Transformer for high-level tasks. Swin Transformer [24] applies local windows to limit the attention scope and shift operations to increase the window interaction. DaViT [11] proposes dual self-attention to capture global context with linear complexity. Due to the remarkable performance of Transformer, researchers have been exploring the utilization of Transformer in low-level vision [42, 4, 8, 46]. SwinIR [20] utilizes spatial window self-attention and the shift operation, following the design of Swin Transformer. Restormer [44] operates self-attention along channel dimensions and applies the U-Net architecture [32]. These methods remarkably outperform CNN-based methods. It reveals that both spatial and channel information are important for performance.

Feature Aggregation. Several works have attempted to aggregate features among different dimensions in multiple vision tasks [43, 48, 35] for performance improvement. In CNN, researchers apply attention mechanisms on both spatial and channel dimensions to enhance feature expressions, such as SCA-CNN [6] and DANet [14]. In Transformer [13], the spatial self-attention models long-range dependencies between pixels. Some researchers explore introducing channel attention in Transformer [49, 7] to aggregate spatial and channel information. It effectively boosts the modeling ability of Transformer. In our work, we alternately utilize spatial and channel self-attention to achieve inter-block feature aggregation. Moreover, we propose AIM and SGFN to obtain intra-block feature aggregation.

3. Method

In this section, we begin by introducing the architecture of dual aggregation Transformer (DAT). Subsequently, we elaborate on the core component of DAT: Dual Aggregation Transformer Block (DATB). Finally, we analyze dual feature aggregation across spatial and channel dimensions.
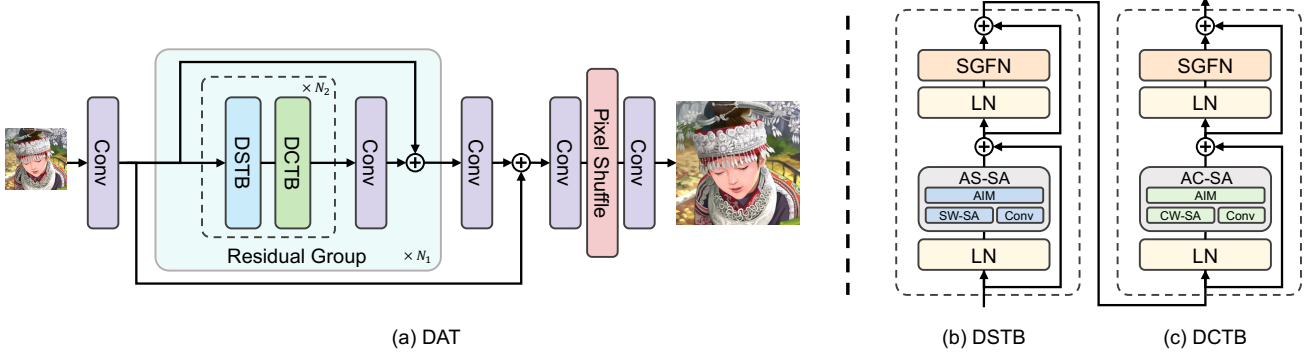
Figure 2: The network architecture of our method. (a) Dual aggregation Transformer (DAT). (b) Dual spatial Transformer block (DSTB). (c) Dual channel Transformer block (DCTB). DSTB and DCTB are two consecutive dual aggregation Transformer blocks (DATBs).

3.1. Architecture

The overall network of the proposed DAT comprises three modules: shallow feature extraction, deep feature extraction, and image reconstruction, as illustrated in Fig. 2. Initially, given a low-resolution (LR) input image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$, we employ a convolution layer to process it and generate the shallow feature $F_S \in \mathbb{R}^{H \times W \times C}$. Notations H and W denote the height and width of the input image, while C represents the number of feature channels.

Subsequently, the shallow feature F_S undergoes processing within the deep feature extraction module to acquire the deep feature $F_D \in \mathbb{R}^{H \times W \times C}$. The module is stacked by multiple residual groups (RGs), with the total number N_1. Meanwhile, to ensure training stability, a residual strategy is employed in the module. Each RG contains N_2 pairs of dual aggregation Transformer blocks (DATBs). As depicted in Fig. 2, each DATB pair contains two Transformer blocks, utilizing spatial and channel self-attention, respectively. A convolution layer is introduced at the end of RG to refine features extracted from Transformer blocks. Besides, for each RG, the residual connection is employed.

Finally, we reconstruct the high-resolution (HR) output image $I_{HR} \in \mathbb{R}^{H_{out} \times W_{out} \times 3}$ through the reconstruction module, where H_{out} is the height of the output image, and W_{out} denotes image width. In this module, the deep feature F_D is upsampled through the pixel shuffle method [33]. And convolution layers are employed to aggregate features before and after the upsampling operation.

3.2. Dual Aggregation Transformer Block

The dual aggregation Transformer block (DATB) is the core component of our proposed method. There are two kinds of DATB: dual spatial Transformer block (DSTB) and dual channel Transformer block (DCTB), as depicted in Fig. 2. DSTB and DCTB are based on spatial window self-attention and channel-wise self-attention, respectively. By alternately organizing DSTB and DCTB, DAT can realize inter-block feature aggregation between spatial and channel dimensions. Moreover, the adaptive interaction module (AIM) and the spatial-gate feed-forward network (SGFN) are proposed to achieve intra-block feature aggregation. Next, We describe the details below.

Spatial Window Self-Attention. The spatial window self-attention (SW-SA) computes attention within windows. As displayed in Fig. 3(a), given the input $X \in \mathbb{R}^{H \times W \times C}$, we generate $query$, key, and $value$ matrices (denoted as Q, K, and V, respectively) through linear projection, where all matrices are in $\mathbb{R}^{H \times W \times C}$ space. The process is defined as

$$Q = XW_Q, K = XW_K, V = XW_V, \qquad (1)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C \times C}$ are linear projections with biases omitted. Subsequently, we partition Q, K, and V into non-overlapping windows, and flat each window, which contains N_w pixels. We denote the reshaped projection matrices as Q_s, K_s, and V_s (all sizes are $\mathbb{R}^{\frac{HW}{N_w} \times N_w \times C}$). Then, we split them into h heads: $Q_s = [Q_s^1, \ldots, Q_s^h]$, $K_s = [K_s^1, \ldots, K_s^h]$, and $V_s = [V_s^1, \ldots, V_s^h]$. The dimension of each head is $d = \frac{C}{h}$. The illustration in Fig. 3(a) is the situation with $h = 1$, where certain details are omitted for simplicity. The output Y_s^i for the i-th head is defined as

$$Y_s^i = \text{softmax}(Q_s^i (K_s^i)^T / \sqrt{d} + D) \cdot V_s^i, \qquad (2)$$

where D denotes the relative position encoding [40]. Finally, we obtain the feature $Y_s \in \mathbb{R}^{H \times W \times C}$ by reshaping and concatenating all Y_s^i. The process is formulated as

$$Y_s = \text{concat}(Y_s^1, \ldots, Y_s^h),$$
$$\text{SW-SA}(X) = Y_s W_p, \qquad (3)$$

where $W_p \in \mathbb{R}^{C \times C}$ is the linear projection to fuse all features. Moreover, following the design of Swin Transformer [24], we employ shift window operations by default to capture more spatial information.

Channel-Wise Self-Attention. The self-attention mechanism in the channel-wise self-attention (CW-SA) is performed along the channel dimension. Following previous works [44, 1], we divide channels into heads and apply attention per head separately. As described in Fig. 3(b), given input X, we apply linear projection to generate $query$,
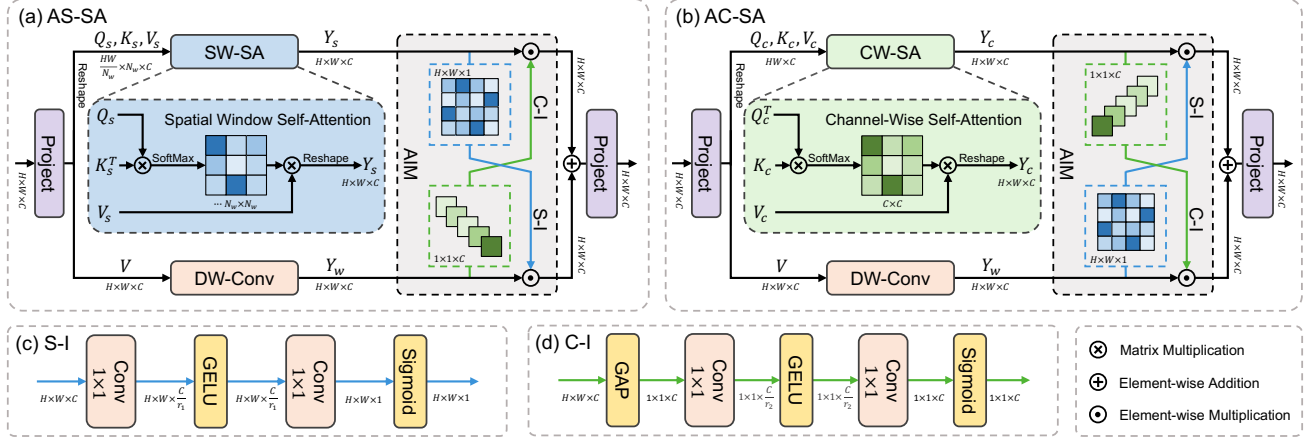
Figure 3: Illustration of adaptive interaction module (AIM). (a) Adaptive spatial self-attention (AS-SA), SW-SA equipped with AIM. (b) Adaptive channel self-attention (AC-SA), CW-SA equipped with AIM. (c) Spatial-interaction (S-I). (d) Channel-interaction (C-I).

key, and $value$ matrices, and reshape all of them to size $\mathbb{R}^{HW \times C}$. We denote the reshaped matrices as Q_c, K_c, and V_c. Same as the operation in SW-SA, we divide the projection vector into h heads. Note that Fig. 3(b) also depicts the case of $h=1$ for simplicity. Then the channel self-attention process of i-th head can be calculated as

$$Y_c^i = V_c^i \cdot \mathrm{softmax}((Q_c^i)^T K_c^i / \alpha), \qquad (4)$$

where $Y_c^i \in \mathbb{R}^{HW \times d}$ is the output for the i-th head, and α is a learnable temperature parameter to adjust the inner products before the softmax function. Finally, we get the attention feature $Y_c \in \mathbb{R}^{H \times W \times C}$ by concatenating and reshaping all Y_c^i. The process definition is the same as Eq. (3).

Adaptive Interaction Module. Since that self-attention focuses on capturing global features, we incorporate a convolution branch parallel to the self-attention module to introduce locality into Transformer. However, simply adding the convolution branch cannot effectively couple global and local features. Moreover, although alternate execution of SW-SA and CW-SA can capture both spatial and channel features, information of different dimensions still cannot be effectively utilized within a single self-attention.

To overcome these issues, we propose the adaptive interaction module (AIM), which acts between two branches, shown in Fig. 3. It adaptively re-weights features of two branches from the spatial or channel dimension, according to the kind of self-attention mechanism. Therefore, the two branch features can be better fused. Also, both spatial and channel information can be aggregated in a single attention module. Based on AIM, we design two new self-attention mechanisms, named adaptive spatial self-attention (AS-SA) and adaptive channel self-attention (AC-SA).

Firstly, we operate the parallel depth-wise convolution (DW-Conv) on $value$ of self-attention (V, defined in Eq. (1)), to establish the direct connection between self-attention and convolution. We denote the convolution out-

put as $Y_w \in \mathbb{R}^{H \times W \times C}$. **Then**, we introduce the AIM to adaptively adjust two features. Specifically, the AIM is based on attention mechanisms [17], including two interaction operations: spatial-interaction (S-I) and channel-interaction (C-I). Given two input features, $A \in \mathbb{R}^{H \times W \times C}$ and $B \in \mathbb{R}^{H \times W \times C}$, spatial-interaction calculates the spatial attention map (denoted as S-Map, size is $\mathbb{R}^{H \times W \times 1}$) of one input (here is B). Channel-interaction infers the channel attention map (denoted as C-Map, size is $\mathbb{R}^{1 \times 1 \times C}$). The operations are illustrated in Fig. 3(c, d), calculated as

$$\begin{aligned} \text{S-Map}(B) &= f(W_2 \sigma(W_1 B)), \\ \text{C-Map}(B) &= f(W_4 \sigma(W_3 H_{GP}(B))), \end{aligned} \qquad (5)$$

where H_{GP} denotes the global average pooling, $f(\cdot)$ is the sigmoid function, and $\sigma(\cdot)$ represents the GELU function. $W_{(\cdot)}$ indicates the weight of the point-wise convolution for downscaling or upscaling channel dimensions. The reduction ratios of W_1 and W_2 are r_1, $\frac{C}{r_1}$, respectively. W_3 has a reduction ratio r_2, and W_4 has an increasing ratio r_2. Subsequently, the attention map is applied to another input (here is A), enabling the interaction. The process is formulated as

$$\begin{aligned} \text{S-I}(A, B) &= A \odot \text{S-Map}(B), \\ \text{C-I}(A, B) &= A \odot \text{C-Map}(B), \end{aligned} \qquad (6)$$

where $\odot$ denotes the element-wise multiplication. **Finally**, with AIM, we design two new self-attention mechanisms, AS-SA and AC-SA, based on SW-SA and CW-SA, respectively. As depicted in Fig. 3(a, b), for SW-SA, we introduce channel-spatial interaction between the two branches. For CW-SA, we apply spatial-channel interaction. Given the input $X \in \mathbb{R}^{H \times W \times C}$, the process is defined as

$$\begin{aligned} \text{AS-SA}(X) &= (\text{C-I}(Y_s, Y_w) + \text{S-I}(Y_w, Y_s))W_p, \\ \text{AC-SA}(X) &= (\text{S-I}(Y_c, Y_w) + \text{C-I}(Y_w, Y_c))W_p, \end{aligned} \qquad (7)$$

where Y_s, Y_c, and Y_w are the outputs of SW-SA, CW-SA, and DW-Conv defined above. W_p is the projection matrix the same as Eq. (3). Besides, we collectively refer to AC-SA and AS-SA as adaptive self-attention (A-SA) for simplicity.
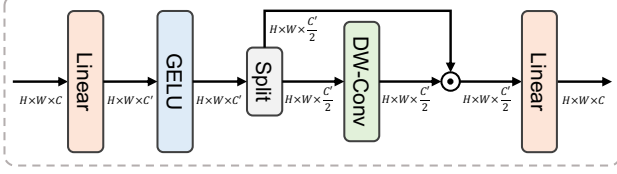
Figure 4: Illustration of spatial-gate feed-forward network.

With AIM, our proposed AS-SA and AC-SA have two advantages over SW-SA and CW-SA. ***Firstly, better coupling of local (convolution) and global (attention).*** Convolution aggregates locality information in the neighbourhood, while self-attention models long-range dependencies. However, considering the feature misalignment between the two branches, simple addition is not convincing enough. Through adaptive interaction, the outputs of the two branches can be adaptively adjusted to fit each other, thus achieving better feature fusion. ***Secondly, stronger modeling ability.*** For AS-SA, the complementary clues improve its channel-wise modeling ability, through channel-interaction. For AC-SA, the representation capability is boosted by additional spatial knowledge, through spatial-interaction. Furthermore, through adaptive interaction, global information can flow from self-attention to the convolution branch. It enhances the output of convolution.

Spatial-Gate Feed-Forward Network. The feed-forward network (FFN) [38] has a non-linear activation and two linear projection layers to extract features. However, it ignores modeling spatial information. Besides, the redundant information in channels hinders feature expression competence. To overcome the above limitations, we propose the spatial-gate feed-forward network (SGFN), introducing spatial-gate (SG) to FFN. As shown in Fig. 4, our SG module is a simple gate mechanism, consisting of depth-wise convolution and element-wise multiplication. Along the channel dimension, we divide the feature map into two parts for convolutional and multiplicative bypass. Overall, given the input $\hat{X}\in\mathbb{R}^{H\times W\times C}$, SGFN is computed as

$$\hat{X}' = \sigma(W_p^1 \hat{X}), \quad [\hat{X}_1', \hat{X}_2'] = \hat{X}',$$
$$\text{SGFN}(\hat{X}) = W_p^2(\hat{X}_1' \odot (W_d \hat{X}_2')), \tag{8}$$

where W_p^1 and W_p^2 represent linear projections, σ denotess the GELU function, and W_d is the learnable parameters of the depth-wise convolution. Both $\hat{X}_1'$ and $\hat{X}_2'$ are in $\mathbb{R}^{H\times W\times \frac{C'}{2}}$ space, where C' denotes the hidden dimension in SGFN. Compared with FFN, our SGFN is able to capture non-linear spatial information and ease the channel redundancy of fully-connected layers. Moreover, different from previous works [22, 5, 36], our SG module utilizes depth-wise convolution to maintain computational efficiency.

Dual Aggregation Transformer Block. Our dual aggregation Transformer block (DATB) is equipped with the adaptive self-attention (A-SA) and the spatial-gate feed-forward network (SGFN). Given the input $X_{l-1}\in\mathbb{R}^{H\times W\times C}$ of the

l-th block, the block is defined as

$$X_l' = \text{A-SA}(\text{LN}(X_{l-1})) + X_{l-1},$$
$$X_l = \text{SGFN}(\text{LN}(X_l')) + X_l', \tag{9}$$

where X_l is the output features, and $\text{LN}(\cdot)$ is the LayerNorm layer. Since A-SA includes AS-SA and AC-SA, there are two types of DATB, dual spatial Transformer block (DSTB) and dual channel Transformer block (DCTB). DSTB applies AS-SA, while DCTB adopts AC-SA.

3.3. Dual Feature Aggregation

Our DAT is capable of aggregating the spatial and channel features through the inter-block and intra-block dual manner to obtain powerful feature representations.

Inter-block Aggregation. DAT alternately adopts DSTB and DCTB to capture features in both dimensions, and make use of their complementary advantages. Specifically, DSTB models long-range spatial context, enhancing the spatial expression of each feature map. Meanwhile, DCTB can better build channel dependencies. DCTB models global channel context, which in turn helps DSTB to capture spatial features and also enlarge the receptive field. Consequently, both spatial and channel information flow between consecutive Transformer blocks and thus can be aggregated.

Intra-block Aggregation. AIM can complement spatial window self-attention with channel knowledge, and enhance channel-wise self-attention from the spatial dimension. Moreover, SGFN is able to introduce additional nonlinear spatial information into FFN that only models channel relationships. Therefore, DAT can aggregate spatial and channel features in each Transformer block.

4. Experiments

4.1. Experimental Settings

Implementation Details. We build two variants of DAT with different complexity, called DAT-S and DAT. For DAT-S, there are 6 residual groups (RGs), and each RG contains 3 pairs of dual aggregation Transformer blocks (DATBs) (3 DSTBs and 3 DCTBs). The attention head number, channel dimension, and channel expansion factor in SGFN are set as 6, 180, and 2 for both DSTB and DCTB. For all DSTBs, we set the window size as 8×16. For DAT, we enlarge the channel expansion factor to 4 and the window size to 8×32. Other settings remain the same as DAT-S.

Data and Evaluation. We follow most previous works [15, 20] to train and test our models. Specifically, we apply two datasets: DIV2K [34] and Flickr2K [21], for training, and five benchmark datasets: Set5 [2], Set14 [45], B100 [26], Urban100 [18], and Manga109 [27], for testing. We carry out experiments under upscaling factors: ×2, ×3, and ×4. LR images are generated from HR images by bicubic degradation. The evaluation of SR results is performed using two metrics: PSNR and SSIM [41], which are calculated on the Y channel (*i.e.*, luminance) of the YCbCr space.

CW-SA	SW-SA	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
✓		16.38	274.54	32.80	0.9340
	✓	16.40	282.76	33.20	0.9379
✓	✓	16.39	278.15	33.34	0.9388

(a) Ablation study of alternate strategy.

Baseline	DW-Conv	AIM	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
✓			16.39	278.15	33.34	0.9388
✓	✓		16.47	279.31	33.41	0.9392
✓	✓	✓	16.84	280.61	33.52	0.9400

(b) Ablation study of AIM.

Model	SA→Conv	Conv→SA	AIM
Params (M)	16.65	16.65	16.84
FLOPs (G)	279.96	279.96	280.61
PSNR (dB)	33.43	33.47	33.52
SSIM	0.9401	0.9397	0.9400

(c) Further ablation study of AIM.

Model	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
FFN	16.84	280.61	33.52	0.9400
SGFN w/o Conv	14.50	242.39	33.44	0.9390
SGFN w/o Split	17.15	286.55	33.53	0.9404
SGFN	14.66	245.36	33.57	0.9405

(d) Ablation study of SGFN.

Model	DCTB	DSTB	DAT
Params (M)	14.65	14.67	14.66
FLOPs (G)	241.75	248.97	245.36
PSNR (dB)	33.26	33.43	33.57
SSIM	0.9376	0.9391	0.9405

(e) Ablation study of different blocks.

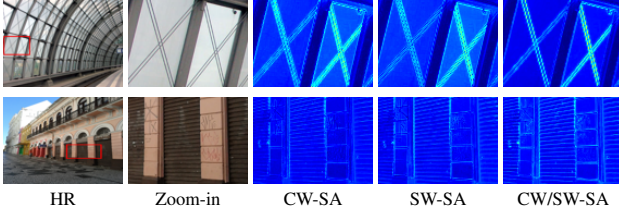Table 1: Ablation studies. The models are trained on DIV2K and Flickr2K, and tested on Urban100 ($\times$2)

Figure 5: Visualization of different attention strategies.

HR Zoom-in CW-SA SW-SA CW/SW-SA

Training Settings. We train models with patch size 64×64 and batch size 32. The training iterations are 500K. We optimize models by minimizing the L_1 loss through Adam optimizer [19] (β_1=0.9 and β_2=0.99). We initially set the learning rate as 2×10^{-4}, and half it at milestones: [250K,400K,450K,475K]. Furthermore, during training, we randomly utilize rotation of $90°$, $180°$, and $270°$ and horizontal flips to augment the data. Our model is implemented based on PyTorch [31] with 4 A100 GPUs.

4.2. Ablation Study

We train models on the dataset DIV2K [34] and Flickr2K [21] and test them on Urban100 [18] in the ablation study. For a fair comparison, all models have the same implementation details (*e.g.*, residual group number) as DAT. The iterations are 300K. Besides, we set the output size as 3×256×256 to compute FLOPs.

Alternate Strategy. To investigate the effect of the strategy for alternating using SW-SA and CW-SA, we carry out several experiments, and list results in Table 1a. The first and second rows of the table mean we replace all attention modules in DAT with CW-SA or SW-SA, where SW-SA adopts the 8×8 window size. The third row represents alternately applying two SA in consecutive Transformer blocks in DAT. Moreover, all models apply the regular FFN [38] and do not adopt AIM in SA. Comparing the three models, we can observe that the model utilizing SW-SA outperforms the model using CW-SA. Furthermore, alternately applying two SA can get the best performance of 33.34 dB. It indicates that exploiting both channel and spatial information is crucial to accurate image restoration.

Additionally, we visualize the last feature maps before upsampling of models with different attention strategies in Fig. 5. CW-SA, SW-SA, and CW/SW-SA correspond to the models in the first, second, and third rows of Table 1a, respectively. We observe that alternately utilizing two self-attention can obtain sharper textures and edges than the other two models. It further demonstrates that the alternate strategy can effectively enhance the expression of features.

Adaptive Interaction Module. We verify the effectiveness of the adaptive interaction module (AIM). **Firstly**, in Table 1b, we conduct a break-down ablation to investigate the impact of our AIM. The baseline is the model in the third row of Table 1a, which yields 33.34 dB. Then we introduce a parallel depth-wise convolution (DW-Conv) to self-attention (both SW-SA and CW-SA). The model obtains a 0.07 dB gain over baseline. Finally, we apply the AIM to aggregate two branches and advance the PSNR from 33.41 to 33.52 dB. It proves that our AIM can effectively improve Transformer performance. **Secondly**, we further analyze the adaptive interaction between the two branches in Table 1c. Specifically, our AIM consists of two direction interactions: from SA to Conv (denoted as SA→Conv), and from Conv to SA (denoted as Conv→SA). We conduct experiments on three cases: only SA→Conv, only Conv→SA, and complete two directions (namely, AIM). The model adopting Conv→SA outperforms the model using SA→Conv by 0.04 dB. It means aggregating information to self-attention has a greater impact on performance. And applying the complete AIM gets the best performance. These results align with the analysis in Section 3.2.

Spatial-Gate Feed-Forward Network. To illustrate the impact of the spatial-gate feed-forward network (SGFN), we carry out an ablation study in Table 1d. We compare models using regular FFN [38], SGFN without depth-wise convolution (denoted as SGFN w/o Conv), SGFN without split channel operation (denoted as SGFN w/o Split), and our proposed SGFN. **Firstly**, compared with FFN, utilizing SGFN can effectively reduce the parameters (2.18M) and FLOPs (32.25G) while improving the performance. **Secondly**, the performance is severely degraded when we re-

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [21]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [47]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [10]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
RFANet [23]	×2	38.26	0.9615	34.16	0.9220	32.41	0.9026	33.33	0.9389	39.44	0.9783
HAN [30]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
CSNLN [29]	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
NLSA [28]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
ELAN [46]	×2	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
DFSA [25]	×2	38.38	0.9620	34.33	0.9232	32.50	0.9036	33.66	0.9412	39.98	0.9798
SwinIR [20]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
CAT-A [9]	×2	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
DAT-S (ours)	×2	38.54	0.9627	34.60	0.9258	32.57	0.9047	34.12	0.9444	40.17	0.9804
DAT (ours)	×2	38.58	0.9629	34.81	0.9272	32.61	0.9051	34.37	0.9458	40.33	0.9807
DAT+ (ours)	×2	38.63	0.9631	34.86	0.9274	32.63	0.9053	34.47	0.9465	40.43	0.9809
EDSR [21]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [47]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [10]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
RFANet [23]	×3	34.79	0.9300	30.67	0.8487	29.34	0.8115	29.15	0.8720	34.59	0.9506
HAN [30]	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
CSNLN [29]	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
NLSA [28]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
ELAN [46]	×3	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
DFSA [25]	×3	34.92	0.9312	30.83	0.8507	29.42	0.8128	29.44	0.8761	35.07	0.9525
SwinIR [20]	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
CAT-A [9]	×3	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.38	0.9546
DAT-S (ours)	×3	35.12	0.9327	31.04	0.8543	29.51	0.8157	29.98	0.8846	35.41	0.9546
DAT (ours)	×3	35.16	0.9331	31.11	0.8550	29.55	0.8169	30.18	0.8886	35.59	0.9554
DAT+ (ours)	×3	35.19	0.9334	31.17	0.8558	29.58	0.8173	30.30	0.8902	35.72	0.9559
EDSR [21]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [47]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [10]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
RFANet [23]	×4	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.918
HAN [30]	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
CSNLN [29]	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
NLSA [28]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
ELAN [46]	×4	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
DFSA [25]	×4	32.79	0.9019	29.06	0.7922	27.87	0.7458	27.17	0.8163	31.88	0.9266
SwinIR [20]	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
CAT-A [9]	×4	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
DAT-S (ours)	×4	33.00	0.9047	29.20	0.7962	27.97	0.7502	27.68	0.8300	32.33	0.9278
DAT (ours)	×4	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343	32.51	0.9291
DAT+ (ours)	×4	33.15	0.9062	29.29	0.7983	28.03	0.7518	27.99	0.8365	32.67	0.9301

Table 2: Quantitative comparison with state-of-the-art methods. The best and second-best results are coloured red and blue.

move the depth-wise convolution in SGFN. It reveals the significance of spatial information. **Thirdly**, after removing the split operation in SGFN, the PSNR value slightly drops, while the model size and complexity increase a lot. It proves that the information redundancy in channel features impairs the performance of models.

Different Blocks. From the above analyses, we display the effect of each proposed component. We further compare our proposed Transformer blocks, DCTB and DSTB, in Table 1e. The DCTB and DSTB represent that we replace all Transformer blocks in DAT with DCTB or DSTB. We can discover that the models using single-type blocks have suboptimal performance. The model adopting DSTB performs better than the model using DCTB, aligning with the results presented in Table 1a. Moreover, both DSTB and DCTB outperform corresponding CW-SA and SW-SA.

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

We compare our two models, DAT-S and DAT, with the current 11 state-of-the-art image SR methods: EDSR [21], RCAN [47], SAN [10], RFANet [23], HAN [30], CSNLN [29], NLSA [28], ELAN [46], DFSA [25], SwinIR [20] and CAT-A [9]. Consistent with prior studies [47, 20], we employ a self-ensemble strategy during testing, denoted by the symbol "+". Table 2 presents quantitative comparisons, while Fig. 6 provides visual comparisons.

Quantitative Results. Table 2 shows results for image SR on factors: ×2, ×3, and ×4. With self-ensemble, our DAT+ outperforms the compared methods on all benchmark datasets with three factors. Meanwhile, DAT performs better than previous methods, except for the PSNR value on the Urban100 dataset (×4) compared with CAT-A. Specifically, compared with SwinIR and CAT-A, our DAT achieves significant gains on the Manga109 dataset (×2), yielding 0.41 dB and 0.23 dB improvements, respectively. Besides, the small vision model, DAT-S, also achieves comparable or better performance compared to previous methods. All these quantitative results indicate that aggregating spatial and channel information from inter-block and intra-block can effectively improve image reconstruction quality.
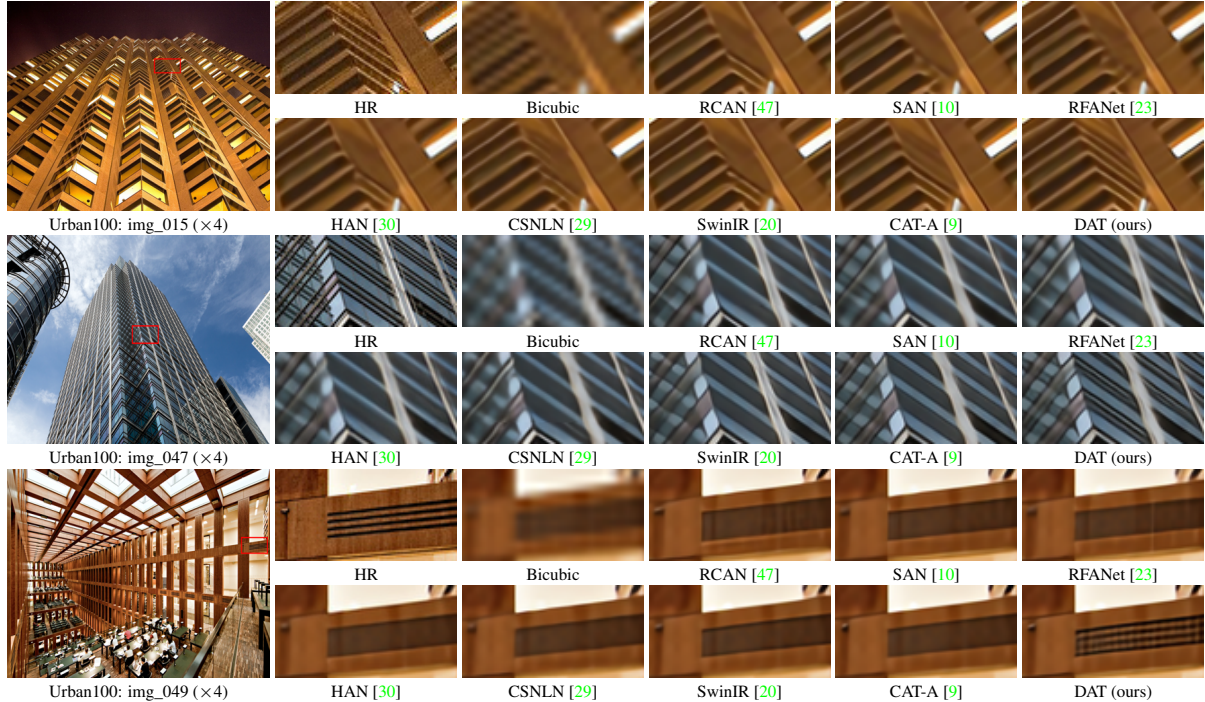
Figure 6: Visual comparison for image SR (×4) in some challenging cases.

Method	EDSR [21]	RCAN [47]	SwinIR [20]	CAT-A [9]	DAT-S (ours)	DAT (ours)	DAT-2 (ours)
Params (M)	43.09	15.59	11.90	16.60	11.21	14.80	11.21
FLOPs (G)	823.34	261.01	215.32	360.67	203.34	275.75	216.93
Urban100	26.64	26.82	27.45	27.89	27.68	27.87	27.86
Manga109	31.02	31.22	32.03	32.39	32.33	32.51	32.41

Table 3: Model complexity comparisons (×4). PSNR (dB) on Urban100 and Manga109, FLOPs, and Params are reported.

Visual Results. We show visual comparisons (×4) in Fig. 6. In some challenging scenarios, the previous methods may suffer blurring artifacts, distortions, or inaccurate texture restoration. In contradistinction, our method effectively mitigates artifacts, preserving more structures and finer details. For instance, in img_015, most compared methods hardly recover details and generate undesired artifacts. However, our DAT can restore the correct structures with clear textures. We can find similar observations in img_047 and img_049. This is mainly because our method has a more powerful representation ability by extracting complex features from different dimensions.

4.4. Model Size Analyses

We further compare our method with several image SR methods in terms of computational complexity (*e.g.*, FLOPs), parameter numbers, and performance at ×4 scale in Table 3. We set the output size as 3×512×512 to compute FLOPs and evaluate performance with PSNR tested on Urban100 and Manga109. Compared with CAT-A [9], our DAT has comparable or better performance with less computational complexity and model size. Besides, DAT-S

obtains excellent performance with lower FLOPs and parameters than SwinIR [20]. Moreover, to further reveal the better trade-off between model size and performance of our method, we introduce an additional variant model, DAT-2, which is detailed in the supplementary material.

5. Conclusion

In this paper, we propose the dual aggregation Transformer (DAT), a new Transformer model for image SR. Our DAT aggregates spatial and channel features in the inter-block and intra-block dual manner, for powerful representation competence. Specifically, successive Transformer blocks alternately apply spatial window and channel-wise self-attention. DAT can model global dependencies through this alternate strategy and achieve inter-block feature aggregation among spatial and channel dimensions. Furthermore, we propose the adaptive interaction module (AIM) and the spatial-gate feed-forward network (SGFN) to enhance each block and realize intra-block feature aggregation between two dimensions. AIM strengthens the modeling ability of two self-attention mechanisms from corresponding dimensions. Meanwhile, SGFN complements the feed-forward network with non-linear spatial information. Extensive experiments indicate that DAT outperforms previous methods.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bo-janowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Na-talia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *NeurIPS*, 2021. 3

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

[3] Haoyu Chen, Jinjin Gu, and Zhi Zhang. Attention in at-tention network for image super-resolution. *arXiv preprint arXiv:2104.09497*, 2021. 2

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yip-ing Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 2

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 5

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 2

[7] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *CVPR*, 2022. 1, 2

[8] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. *arXiv preprint arXiv:2303.06373*, 2023. 2

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022. 1, 7, 8

[10] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 1, 7, 8

[11] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transform-ers. In *ECCV*, 2022. 1, 2

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1, 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth 16x16 words: Trans-formers for image recognition at scale. In *ICLR*, 2021. 1, 2

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[15] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation net-works. In *CVPR*, 2018. 4

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 5, 6

[19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration us-ing swin transformer. In *ICCVW*, 2021. 1, 2, 5, 7, 8

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 5, 6, 7, 8

[22] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. In *NeurIPS*, 2021. 5

[23] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 7, 8

[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 *ICCV*, 2021. 1, 2, 3

[25] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, 2021. 7

[26] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 5

[27] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Mul-timedia Tools and Applications*, 2017. 5

[28] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, 2021. 7

[29] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaus-tive self-exemplars mining. In *CVPR*, 2020. 1, 2, 7, 8

[30] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 2, 7, 8

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Al-ban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[33] 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 *CVPR*, 2016. 3

[34] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 5, 6

[35] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *NeurIPS*, 2021. 2

[36] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, 2022. 5

[37] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 5, 6

[39] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 1, 2

[40] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *ICLR*, 2022. 3

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[42] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 1, 2

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 2

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1, 2, 3

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proc. 7th Int. Conf. Curves Surf.*, 2010. 5

[46] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, 2022. 1, 2, 7

[47] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 1, 2, 7, 8

[48] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 2

[49] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *ICML*, 2022. 2