图像超分辨率双重聚合Transformer

郑晨[1], 张宇伦[2*], 金金Gu[3,4], 孔令鹤[1*], 杨晓康[1], 鱼飞[2]l上海交通大学，[2]苏黎世联邦理工学院，[3]悉尼大学，[4]上海人工智能实验室

摘要

*Transformer*最近在低级视觉任务中获得了相当大的流行度，包括图像超分辨率（*SR*）。这些网络利用不同维度的自注意力机制，空间或通道，并取得了显著的性能。这启发我们将两个维度结合起来，以获得更强大的表示能力。基于上述思想，我们提出了一种新的*Transformer*模型，双重聚合*Transformer*（*DAT*），用于图像*SR*。我们的*DAT*在空间和通道维度上以块间和块内双重方式聚合特征。具体来说，我们在连续的*Transformer*模块中交替应用空间和通道自注意力机制。这种替代策略使*DAT*能够捕获全局上下文并实现块间特征聚合。此外，我们提出了自适应交互模块（*AIM*）和空间门控前馈网络（*SGFN*）以实现块内特征聚合。*AIM*补充了来自相应维度的两种自注意力机制。同时，*SGFN*在前馈网络中引入了额外的非线性空间信息。大量实验表明，我们的*DAT*优于当前方法。代码和模型可在
https://github.com/zhengchen1999/DAT。

1. 简介

单图像超分辨率（SR）是一项传统的低级视觉任务，它专注于从低分辨率（LR）的对应物中恢复高分辨率（HR）图像。由于对于给定的LR输入存在多个潜在解，这是一个不适定问题，近年来出现了各种方法来应对这一挑战。许多这些方法利用了卷积神经网络（CNNs）[12, 47, 10, 29]。然而，卷积采用局部机制，这阻碍了全局依赖的建立，并限制了模型的性能。

最近，Transformer在自然语言处理（NLP）领域提出的模型在多个高级视觉任务中表现突出 [13, 39, 24, 11, 7]。Transformer的核心

*通讯作者：张宇伦，yulun100@gmail.com；孔令鹤，linghe.kong@sjtu.edu.cn

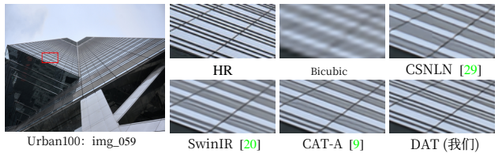


图1：Urban100上的视觉比较 (×4)。CSNLN、SwinIR 和 CAT-A 出现模糊伪影。

自注意力（SA）机制，能够建立全局依赖关系。这一特性缓解了基于CNN的算法的局限性。考虑到Transformer的潜力，一些研究人员尝试将其应用于低级任务 [20, 44, 42, 9],，包括图像超分辨率。他们从不同角度探索了Transformer在处理高分辨率图像时的有效用法，以减轻全局自注意力 [13]的高复杂度。在空间方面，一些方法 [20, 46, 9]应用局部空间窗口来限制自注意力的范围。在通道方面，提出了“转置”注意力 [44]，该注意力沿通道维度而非空间维度计算自注意力。这些方法均因其各自维度上的强大建模能力而取得了显著成果。空间窗口自注意力（SW-SA）能够建模像素之间的细粒度空间关系。通道自注意力（CW-SA）可以建模特征图之间的关系，从而利用全局图像信息。通常，提取空间信息和捕捉通道上下文对于Transformer在图像超分辨率中的性能至关重要。

基于上述发现，我们提出了用于图像超分辨率的双重聚合Transformer（DAT）。我们的DAT通过块间和块内双重方式聚合空间和通道特征，以获得强大的表示能力。具体而言，我们在连续的双聚合Transformer块（DATBs）中交替应用空间窗口自注意力和通道自注意力。通过这种替代策略，我们的DAT能够捕获空间和通道上下文，并在不同维度之间实现块间特征聚合。此外，两种自注意力机制相互补充。空间窗口自注意力丰富了每个特征图的空间表达，有助于建模通道依赖性。

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

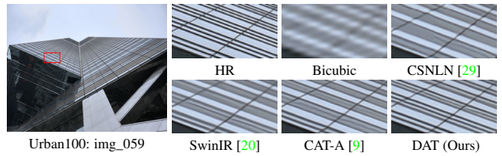


Figure 1: Visual comparison (×4) on Urban100. CSNLN, SwinIR, and CAT-A suffer from blurring artifacts.

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-

通道自注意力为空间自注意力提供了特征之间的全局信息，扩展了窗口注意力的感受野。

与此同时，由于自注意力机制专注于建模全局信息，我们将卷积与自注意力并行结合，以补充Transformer的局部性。为了增强两个分支的融合，并在单个自注意力模块内聚合空间和通道信息，我们提出了自适应交互模块（AIM）。它由两个交互操作组成，空间交互（S-I）和通道交互（C-I），它们在两个分支之间作用以交换信息。通过S-I和C-I，AIM根据不同的自注意力机制，自适应地重新加权两个分支的空间或通道维度的特征图。此外，借助AIM，我们基于空间窗口和通道自注意力，分别设计了两种新的自注意力机制：自适应空间自注意力（AS-SA）和自适应通道自注意力（AC-SA）。

此外，Transformer块中的另一个组件，前馈网络（FFN） [38], 通过全连接层提取特征。它忽略了建模空间信息。此外，通道之间的冗余信息阻碍了特征表示学习的进一步进展。为了解决这些问题，我们设计了空间门控前馈网络（SGFN），它在FFN的两个全连接层之间引入了空间门控（SG）模块。SG模块是一个简单的门控机制（深度卷积和逐元素乘法）。SG的输入特征沿通道维度分为两个段进行卷积和乘法旁路。我们的SG模块可以补充FFN，提供额外的非线性空间信息，并缓解通道冗余。一般来说，基于AIM和SGFN，DAT可以实现块内特征聚合。

总体而言，通过上述三种设计，我们的DAT可以通过块间和块内双重方式聚合空间和通道信息，以实现强大的特征表达。因此，如图1所示，我们的DAT在最近的最先进SR方法中取得了优异的视觉结果。我们的贡献有三方面：

- 我们设计了一种新的图像超分辨率模型，双重聚合Transformer（DAT）。我们的DAT以块间和块内双重方式聚合空间和通道特征，以获得强大的表示能力。
- 我们交替采用空间和通道自注意力，实现块间空间和通道特征聚合。此外，我们提出了AIM和SGFN以实现块内特征聚合。
- 我们进行了广泛的实验，以证明我们的DAT在性能上优于当前最先进方法，同时保持了较低的复杂性和模型大小。

2. 相关工作

图像超分辨率。基于深度CNN的方法在图像超分辨率领域表现出显著的效能。SR-CNN [12] 是开创性工作，首次利用CNN并优于传统方法。在此尝试之后，人们投入了大量精力加深网络层数以获得更好的性能。例如，RCAN [47] 设计了残差在残差结构 [16]并构建了一个 400+层模型。此外，注意力机制 [48, 30, 29, 3] 在空间或通道维度上被采用，以进一步提高建模能力。然而，大多数基于CNN的方法仍然难以有效地建模空间和通道维度的全局依赖关系。

视觉Transformer。Transformer在高层次视觉任务中表现出色 [13, 39, 37]。一系列基于Transformer的方法被提出以提高Transformer在高层次任务中的效率和效果。Swin Transformer [24] 应用局部窗口来限制注意力范围，并通过移位操作增加窗口交互。DaViT [11]提出了双重自注意力，以线性复杂度捕获全局上下文。由于Transformer的出色性能，研究人员一直在探索Transformer在低层次视觉 [42, 4, 8, 46]中的应用。SwinIR [20]利用空间窗口自注意力和移位操作，遵循Swin Transformer的设计。Restormer [44] 沿通道维度进行自注意力，并应用U-Net架构 [32]。这些方法显著优于基于CNN的方法。这表明空间和通道信息对性能都很重要。

特征聚合。多项工作尝试在多个视觉任务中聚合不同维度的特征 [43, 48, 35] 以提升性能。在 CNN 中，研究人员在空间和通道维度上应用注意力机制来增强特征表达，例如 SCA-CNN [6] 和 DANet [14]。在 Transformer [13], 中，空间自注意力模型用于提取像素之间的长距离依赖关系。一些研究人员探索在 Transformer [49, 7] 中引入通道注意力来聚合空间和通道信息。这有效地提升了 Transformer 的建模能力。在我们的工作中，我们交替使用空间和通道自注意力来实现块间特征聚合。此外，我们提出了 AIM 和 SGFN 来获得块内特征聚合。

3. 方法

在本节中，我们首先介绍了双重聚合Transformer（DAT）的架构。随后，我们详细阐述了DAT的核心组件：双聚合Transformer块（DATB）。最后，我们分析了跨空间和通道维度的双特征聚合。

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

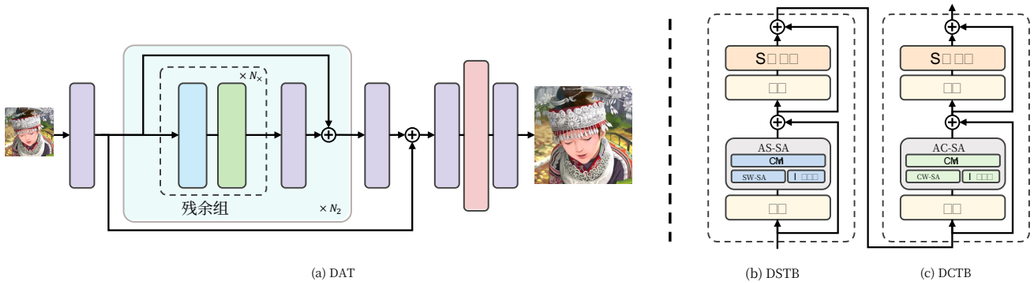
Image Super-Resolution. Deep CNN-based approaches exhibit significant efficacy in the field of image SR. SR-CNN [12] is the pioneering work, which first utilizes CNN and outperforms traditional approaches. Following this attempt, substantial dedication has been invested in deepening the layer of the 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.



(a) DAT (b) DSTB (c) DCTB

图2：我们方法的网络架构。(a) 双重聚合Transformer (DAT)。(b) 双空间Transformer模块 (DSTB)。(c) 双通道Transformer模块 (DCTB)。DSTB和DCTB是两个连续的双重聚合Transformer模块 (DATBs)。

3.1. 架构

所提出的DAT的整体网络由三个模块组成：浅层特征提取、深层特征提取和图像重建，如图2所示。最初，给定一个低分辨率（LR）输入图像$I_{LR}\in\mathbb{R}^{H\times W\times 3}$，我们采用卷积层对其进行处理并生成浅层特征$F_S\in\mathbb{R}^{H\times W\times C}$。符号 H 和 W 表示输入图像的高度和宽度，而 C 表示特征通道数。

随后，浅层特征 F_S 在深层特征提取模块内进行处理以获取深层特征 $F_D\in\mathbb{R}^{H\times W\times C}$。该模块由多个残差组 (RGs) 堆叠而成，总数为 N_1。同时，为确保训练稳定性，该模块采用了残差策略。每个RG包含 N_2 对双重聚合Transformer模块 (DATBs)。如图2所示，每个DATB对包含两个Transformer模块，分别利用空间和通道自注意力。在RG的末端引入卷积层以细化从Transformer模块提取的特征。此外，对于每个RG，采用了残差连接。

最后，我们通过重建模块重构高分辨率（HR）输出图像 $I_{HR}\in\mathbb{R}^{H_{out}\times W_{out}\times 3}$，其中 H_{out} 是输出图像的高度，而W_{out} 表示图像宽度。在这个模块中，深度特征 F_D 通过像素重排方法 [33]进行上采样。并且，在上下采样操作前后使用卷积层来聚合特征。

3.2. 双聚合Transformer模块

双重聚合Transformer模块（DATB）是我们提出的方法的核心组件。DATB有两种类型：双空间Transformer模块（DSTB）和双通道Transformer模块（DCTB），如图2所示。DSTB和DCTB分别基于空间窗口自注意力和通道自注意力。通过交替组织DSTB和DCTB，DAT可以在空间和通道维度之间实现块间特征聚合

此外，提出了自适应交互模块（AIM）和空间门控前馈网络（SGFN）以实现块内特征聚合。接下来，我们将详细描述。

空间窗口自注意力。 空间窗口自注意力（SW-SA）在窗口内计算注意力。如图 3(a)所示，给定输入 $X\in\mathbb{R}^{H\times W\times C}$，我们通过线性投影生成 $query$、key 和 $value$ 矩阵（分别表示为 Q、K和 V），其中所有矩阵都在 $\mathbb{R}^{H\times W\times C}$ 空间中。该过程定义为

$$Q=XW_Q, K=XW_K, V=XW_V, \tag{1}$$

其中 W_Q、W_K、$W_V\in\mathbb{R}^{C\times C}$ 是省略了偏置的线性投影。随后，我们将 Q、K 和 V 分割成非重叠窗口，并将每个窗口展平，其中包含 N_w 个像素。我们将重塑的投影矩阵表示为 Q_s、K_s和 V_s（所有尺寸均为 $\mathbb{R}^{\frac{HW}{N_w}\times N_w\times C}$）。然后，我们将它们分成 h 个头：Q_s=$[Q_s^1,\dots,Q_s^h]$、K_s=$[K_s^1,\dots,K_s^h]$和 V_s=$[V_s^1,\dots,V_s^h]$。每个头的维度是 d=$\frac{C}{h}$。图3(a)中的插图是 h=1的情况，为了简洁起见，省略了某些细节。第 i 个头的输出 Y_s^i 定义为

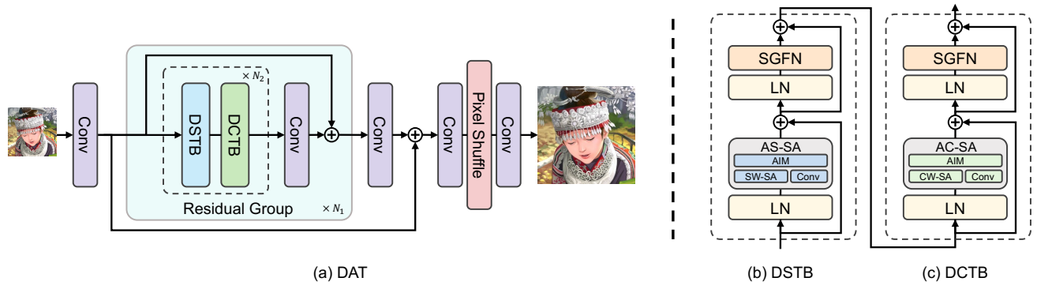
$$Y_s^i=\mathrm{softmax}(Q_s^i(K_s^i)^T/\sqrt{d}+D)\cdot V_s^i, \tag{2}$$

哪里 D 表示相对位置编码 [40]。最后，我们通过重塑和连接所有 $Y_s\in\mathbb{R}^{H\times W\times C}$ 来获得特征 Y_s^i。该过程被公式化为

$$\begin{aligned} Y_s&=\mathrm{concat}(Y_s^1,\dots,Y_s^h),\\ \text{SW-SA}(X)&=Y_sW_p, \end{aligned} \tag{3}$$

其中 $W_p\in\mathbb{R}^{C\times C}$ 是用于融合所有特征的线性投影。此外，遵循 Swin Transformer [24], 的设计，我们默认采用移位窗口操作以捕获更多空间信息。

通道自注意力。 通道自注意力（CW-SA）中的自注意力机制沿通道维度执行。遵循先前工作 [44, 1], 我们将通道分为头，并分别对每个头应用注意力。如图3(b)所示，给定输入 X，我们应用线性投影生成 $query$,



(a) DAT (b) DSTB (c) DCTB

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

$$Q=XW_Q, K=XW_K, V=XW_V, \tag{1}$$

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=$[Q_s^1,\dots,Q_s^h]$, K_s=$[K_s^1,\dots,K_s^h]$, and V_s=$[V_s^1,\dots,V_s^h]$. 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

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

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

$$\begin{aligned} Y_s&=\mathrm{concat}(Y_s^1,\dots,Y_s^h),\\ \text{SW-SA}(X)&=Y_sW_p, \end{aligned} \tag{3}$$

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 $query$,

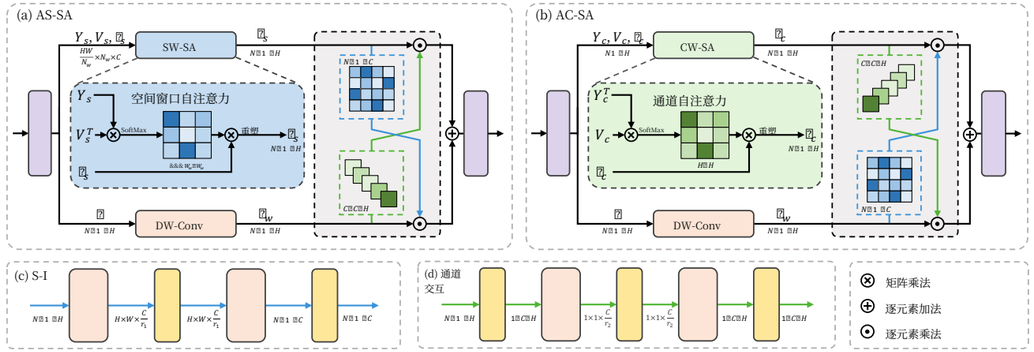


图3：自适应交互模块（AIM）的示意图。(a) 自适应空间自注意力（AS-SA），配备AIM的SW-SA。(b) 自适应通道自注意力（AC-SA），配备AIM的CW-SA。(c) 空间交互（S-I）。(d) 通道交互（C-I）。

key，和 $value$ 矩阵，并将它们重塑为大小$\mathbb{R}^{HW\times C}$。我们将重塑的矩阵表示为 Q_c, K_c，和V_c。与SW-SA中的操作相同，我们将投影向量分为 h 头。请注意，图3(b) 也描绘了 h=1的简化情况。然后第 i个头的通道自注意力过程可以计算为

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

其中 $Y_c^i \in \mathbb{R}^{HW\times d}$ 是第 i个头的输出，而 α 是一个可学习的温度参数，用于调整softmax函数之前的内积。最后，我们通过连接和重塑所有 $Y_c \in \mathbb{R}^{H\times W\times C}$ 得到注意力特征Y_c^i。该过程定义与公式(3)相同。

自适应交互模块。由于自注意力专注于捕获全局特征，我们在自注意力模块旁边加入一个卷积分支，以将局部性引入Transformer。然而，仅仅添加卷积分支并不能有效地耦合全局和局部特征。此外，尽管SW-SA和CW-SA的交替执行可以捕获空间和通道特征，但不同维度的信息仍然无法在单个自注意力内部有效利用。

为了克服这些问题，我们提出了自适应交互模块（AIM），它位于两个分支之间，如图3所示。它根据自注意力机制的类型，自适应地重新加权两个分支的空间或通道维度的特征。因此，两个分支的特征可以更好地融合。此外，空间和通道信息都可以在一个注意力模块中聚合。基于AIM，我们设计了两种新的自注意力机制，分别称为自适应空间自注意力（AS-SA）和自适应通道自注意力（AC-SA）。

首先，我们对 $value$ 的深度卷积（DW-Conv）进行并行操作（V，如公式（1）所示），以建立自注意力和卷积之间的直接连接。我们记卷积输出为输出为 $Y_w \in \mathbb{R}^{H\times W\times C}$。**然后**，我们引入AIM以自适应地调整两个特征。具体来说，AIM基于注意力机制[17]，，包括两种交互操作：空间交互（S-I）和通道交互（C-I）。给定两个输入特征 $A\in\mathbb{R}^{H\times W\times C}$和 $B\in\mathbb{R}^{H\times W\times C}$，空间交互计算一个输入（此处为 B）的空间注意力图（表示为S-Map，大小为 $\mathbb{R}^{H\times W\times 1}$）。通道交互推断通道注意力图（表示为C-Map，大小为 $\mathbb{R}^{1\times 1\times C}$）。这些操作如图3(c, d)所示，计算方法为

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

其中 H_{GP} 表示全局平均池化，$f(\cdot)$是sigmoid函数，$\sigma(\cdot)$表示GELU函数。$W_{(\cdot)}$ 表示用于缩小或放大通道维度的逐点卷积的权重。W_1 和 W_2 的缩减率分别为 r_1 和 $\frac{C}{r_1}$。W_3 的缩减率为 r_2，W_4的增量为 r_2。随后，注意力图应用于另一个输入（此处为 A），以实现交互。该过程表示为

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

在哪里 $\odot$ 表示逐元素乘法。**最后**，基于SW-SA和CW-SA，我们分别设计了两种新的自注意力机制，AS-SA和AC-SA。如图3(a, b)所示，对于SW-SA，我们在两个分支之间引入了通道-空间交互。对于CW-SA，我们应用了空间-通道交互。给定输入 $X\in\mathbb{R}^{H\times W\times C}$，过程定义为

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

哪里 Y_s, Y_c，以及 Y_w 是前面定义的 SW-SA、CW-SA 和 DW-Conv 的输出。W_p 是投影矩阵与 Eq.(3) 相同吗。此外，我们统称为 AC-SA 和 AS-SA 为自适应自注意力（A-SA），以简化。

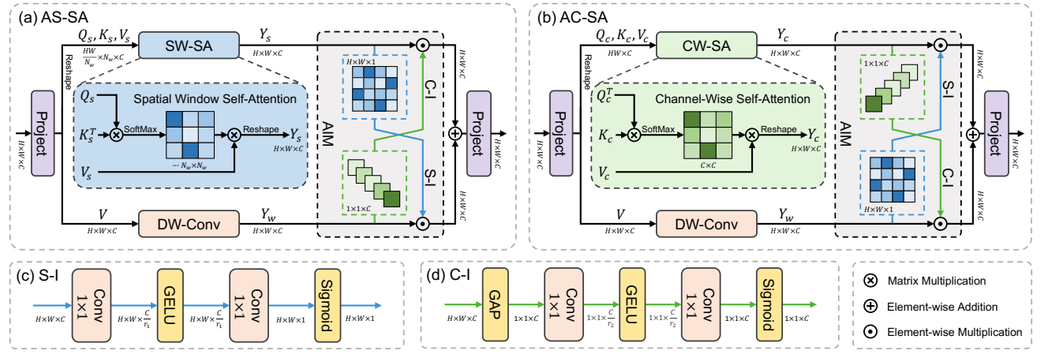


Figure 3: Illustration of adaptive interaction module (AIM). (a) Adaptive spatial self-attention (AS-SA), SW-SA equipped with AIM. (b) Adaptive channel self-attention (AC-SA), CW-SA equipped with AIM. (c) Spatial-interaction (S-I). (d) Channel-interaction (C-I).

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

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

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 output 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

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

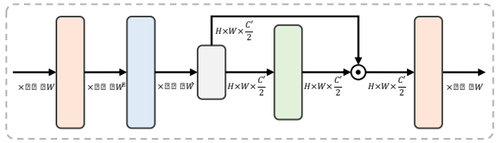
where H_{GP} denotes the global average pooling, $f(\cdot)$ is the sigmoid function, and $\sigma(\cdot)$ represents the GELU function. $W_{(\cdot)}$ indicates the weight of the point-wise convolution for downscaling or upscaling channel dimensions. The reduction ratios of W_1 and W_2 are r_1, $\frac{C}{r_1}$, 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

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

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

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

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.



F图 4：空间门控前馈网络示意图

工作。

通过 AIM，我们提出的 AS-SA 和 AC-SA 相较于 SW-SA 和 CW-SA 具有两个优势。首先，局部（卷积）和全局（注意力）的更好耦合。卷积聚合邻域内的局部信息，而自注意力建模长距离依赖关系。然而，考虑到两个分支之间的特征错位，简单的加法并不足以令人信服。通过自适应交互，两个分支的输出可以自适应地调整以彼此适应，从而实现更好的特征融合。其次，更强的建模能力。对于 AS-SA，通过通道交互，互补线索提高了其通道建模能力。对于 AC-SA，通过空间交互，额外的空间知识增强了其表示能力。此外，通过自适应交互，全局信息可以从自注意力流向卷积分支。它增强了卷积的输出。

空间门控前馈网络。 前馈网络 (FFN) [38] 具有非线性激活和两个线性投影层来提取特征。然而，它忽略了建模空间信息。此外，通道中的冗余信息阻碍了特征表达能力。为了克服上述限制，我们提出了空间门控前馈网络 (SGFN)，将空间门控 (SG) 引入 FFN。如图4所示，我们的 SG 模块是一个简单的门控机制，由深度卷积和逐元素乘法组成。在通道维度上，我们将特征图分为两部分进行卷积和乘法旁路。总体而言，给定输入 $\hat{X}\in\mathbb{R}^{H\times W\times C}$，SGFN 的计算如下：

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

其中 W_p^1 和 W_p^2 表示线性投影， σ 表示 GELU 函数，而 W_d 是深度卷积的可学习参数$\hat{}$ $\hat{}$。两者 X_1' 和 X_2' 都在$\mathbb{R}^{H\times W\times \frac{C'}{2}}$空间中，其中 C'表示 SGFN 中的隐藏维度。与 FFN 相比，我们的 SGFN 能够捕获非线性空间信息并简化全连接层的通道冗余。此外，与之前的工作 [22, 5, 36], 不同，我们的 SG 模块利用深度卷积来保持计算效率。

双聚合Transformer模块。我们的双聚合Transformer模块（DATB）配备了自适应自注意力（A-SA）和空间门控前馈网络（SGFN）。给定输入 $X_{l-1}\in\mathbb{R}^{H\times W\times C}$ 的 l-th模块，该模块定义为

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

其中 X_l 是输出特征，LN(·) 是 Layer- Norm 层。由于 A-SA 包含 AS-SA 和 AC-SA，因此 DATB 有两种类型，即双空间 Transformer 块 (DSTB) 和双通道 Transformer 块 (DCTB)。DSTB 应用 AS-SA，而 DCTB 采取 AC-SA。

3.3. 双特征聚合

我们的 DAT 能够通过块间和块内双重方式聚合空间和通道特征，以获得强大的特征表示。

块间聚合。 DAT 交替采用 DSTB 和 DCTB 来捕获两个维度的特征，并利用它们的互补优势。具体来说，DSTB 建模长程空间上下文，增强每个特征图的空间表达。同时，DCTB 能更好地构建通道依赖关系。DCTB 建模全局通道上下文，这反过来又有助于 DSTB 捕获空间特征，并扩大感受野。因此，空间和通道信息在连续的 Transformer 模块之间流动，从而可以被聚合。

块内聚合。 AIM 可以通过通道知识补充空间窗口自注意力，并从空间维度增强通道自注意力。此外，SGFN 能够将额外的非线性空间信息引入 FFN，而 FFN 仅建模通道关系。因此，DAT 可以在每个 Transformer 模块中聚合空间和通道特征。

4. 实验

4.1. 实验设置

实现细节。 我们构建了两个复杂度不同的 DAT 变体，称为 DAT-S 和 DAT。对于 DAT-S，有 6 个残差组 (RGs)，每个 RG 包含 3 对双聚合 Transformer 模块 (DATBs)（3 个 DSTBs 和 3 个 DCTBs）。在 SGFN 中，注意力头数、通道维度和通道扩展因子在 DSTB 和 DCTB 中均设置为 6、180 和 2。对于所有 DSTBs，我们将窗口大小设置为 8×16。对于 DAT，我们将通道扩展因子扩大到 4，窗口大小设置为 8×32。其他设置与 DAT-S 相同。

数据与评估。 我们遵循大多数先前工作 [15,20] 来训练和测试我们的模型。具体来说，我们应用了两个数据集：DIV2K [34] 和Flickr2K [21], 用于训练，以及五个基准数据集：Set5 [2], Set14 [45], B100 [26],Urban100 [18], 和 Manga109 [27], 用于测试。我们在上采样因子： ×2, ×3和 ×4下进行实验。LR图像通过双三次退化从HR图像生成。SR结果的评估使用两个指标：PSNR和SSIM [41], , 这些指标在YCbCr空间的Y通道（即，亮度）上计算。

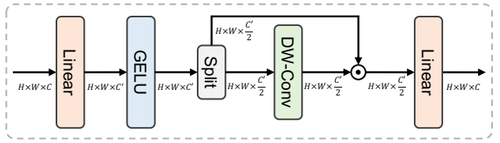


Figure 4: Illustration of spatial-gate feed-forward network.

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

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

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

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

where X_l is the output features, and LN(·) is the Layer-Norm 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 non-linear 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	参数（M） I			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) 替代策略的消融实验。

基线	DW-Conv	AIM	参数（M） F			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) AIM的消融实验。

模型	SA→Conv (		AIM
参数（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) 进一步的 AIM 消融实验。

模型	参数（M） Fl			
FFN	16.84	280.61	33.52	0.9400
SGFN 无卷积	14.50	242.39	33.44	0.9390
SGFN 无分割	17.15	286.55	33.53	0.9404
SGFN	14.66	245.36	33.57	0.9405

(d) SGFN的消融实验。

模型	DCTB	DSTB	DAT
参数（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) 不同模块的消融实验。

表1：消融研究。模型在DIV2K和Flickr2K上训练，并在Urban100 (×2) 上测试

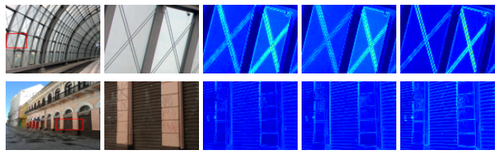


图5：不同注意力策略的可视化。

训练设置。我们使用块大小64×64和批量大小32训练模型。训练迭代次数为500K。我们通过Adam优化器最小化 L_1 损失 [19] (β_1=0.9 和 β_2=0.99)来优化模型。我们最初将学习率设置为 2×10^{-4}，并在里程碑处减半：[250K,400K,450K,475K]。此外，在训练期间，我们随机使用旋转 90°, 180°, 和270° 以及水平翻转来增强数据。我们的模型基于PyTorch [31]实现，并使用4 A100 GPU。

4.2. 消融实验

我们在 DIV2K [34]和 Flickr2K [21]数据集上训练模型，并在 Urban100 [18]上进行消融实验。为了公平比较，所有模型都具有与 DAT 相同的实现细节（例如，残余组数量）。迭代次数为 300K。此外，我们将输出大小设置为 3×256×256 以计算 FLOPs。

替代策略。为了研究交替使用SW-SA和CW-SA策略的效果，我们进行了多次实验，并将结果列于表 1a。表的第一行和第二行表示我们将DAT中的所有注意力模块替换为CW-SA或SW-SA，其中SW-SA采用 8×8 窗口大小。第三行表示在DAT中连续的Transformer模块中交替应用两种SA。此外，所有模型都应用常规FFN [38]，并且SA中不采用AIM。比较这三个模型，我们可以观察到使用SW-SA的模型性能优于使用CW-SA的模型。此外，交替应用两种SA可以获得33.34 dB的最佳性能。这表明利用通道和空间信息对于精确的图像恢复至关重要。

此外，我们在图 5中可视化了不同注意力策略模型在上采样前的最后一个特征图。CW-SA、SW-SA和CW/SW-SA分别对应表 1a的第一行、第二行和第三行中的模型。我们观察到交替利用两种自注意力可以获得比其他两个模型更清晰的纹理和边缘。这进一步证明了替代策略可以有效地增强特征的表示。

自适应交互模块。我们验证了自适应交互模块（AIM）的有效性。**首先**，在表 1b中，我们进行分解消融实验，以研究我们的 AIM 的影响。基线是表1a中第三行的模型，其 PSNR 为 33.34 dB。然后我们将并行深度卷积（DW-Conv）引入自注意力（包括 SW-SA 和 CW-SA）。该模型在基线基础上获得了 0.07 dB 的提升。最后，我们将 AIM 应用于两个分支的聚合，并将 PSNR 从 33.41 提升至 33.52 dB。这证明我们的 AIM 可以有效提升 Transformer 的性能。**其次**，我们进一步分析表 1c 中两个分支的自适应交互。具体而言，我们的 AIM 包含两个方向的交互：从自注意力到卷积（记为 SA→Conv），以及从卷积到自注意力（记为 Conv→SA）。我们进行了三种情况的实验：仅 SA→Conv、仅 Conv→SA，以及完整两个方向（即 AIM）。采用 Conv→SA 的模型比使用 SA→Conv 的模型性能提升 0.04 dB。这意味着将信息聚合到自注意力对性能的影响更大。应用完整的 AIM 获得了最佳性能。这些结果与第 3.2节的分析一致。

空间门控前馈网络。为了说明空间门控前馈网络（SGFN）的影响，我们在表1d中进行了消融实验。我们比较了使用常规FFN [38], SGFN无深度卷积（记为SGFN 无卷积）、SGFN无分通道操作（记为SGFN无分割）以及我们提出的SGFN。**首先**，与FFN相比，使用SGFN可以有效减少参数（2.18M）和FLOPs（32.25G）同时提高性能。**其次**，当我们重新

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 (×2)

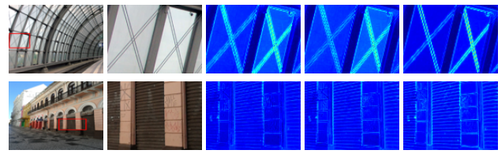


Figure 5: Visualization of different attention strategies.

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-

方法	尺度	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 (我们)	×2	38.54	0.9627	34.60	0.9258	32.57	0.9047	34.12	0.9444	40.17	0.9804
DAT (我们)	×2	38.58	0.9629	34.81	0.9272	32.61	0.9051	34.37	0.9458	40.33	0.9807
DAT+ (我们)	×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 (我们)	×3	35.12	0.9327	31.04	0.8543	29.51	0.8157	29.98	0.8846	35.41	0.9546
DAT (我们)	×3	35.16	0.9331	31.11	0.8550	29.55	0.8169	30.18	0.8886	35.59	0.9554
DAT+ (我们)	×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 (我们)	×4	33.00	0.9047	29.20	0.7962	27.97	0.7502	27.68	0.8300	32.33	0.9278
DAT (我们)	×4	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343	32.51	0.9291
DAT+ (我们)	×4	33.15	0.9062	29.29	0.7983	28.03	0.7518	27.99	0.8365	32.67	0.9301

表2： 与当前最先进方法进行定量比较。最佳和次优结果分别用 红色 和 蓝色标出。

移动SGFN中的深度卷积时。这揭示了空间信息的重要性。 **第**三，在SGFN中移除分通道操作后，PSNR值略有下降，而模型大小和复杂度增加很多。这证明通道特征中的信息冗余提升了模型的性能。

不同的模块。从上述分析中，我们展示了每个所提出组件的效果。我们进一步在表1e中比较我们提出的Transformer模块，DCTB和DSTB。DCTB和DSTB表示我们将DAT中的所有Transformer模块替换为DCTB或DSTB。我们可以发现，使用单一类型模块的模型具有次优性能。采用DSTB的模型比使用DCTB的模型表现更好，这与表1a中展示的结果一致。此外，DSTB和DCTB均优于相应的CW-SA和SW-SA。

4.3. 与当前最先进方法比较

我们将我们的两个模型DAT-S和DAT与当前的11种最先进的图像超分辨率方法进行比较：EDSR [21], RCAN [47], SAN [10], RFANet [23], HAN [30], CSNLN [29], NLSA [28], ELAN [46], DFSA [25], SwinIR [20] 和CAT-A [9]。与先前的研究一致 [47, 20], , 我们在测试过程中采用自集成策略，用符号“+”表示。表2 展示了定量比较，而图6 提供了视觉比较。

定量结果。 表 2 展示了在因素： ×2, ×3, 和 ×4上的图像超分辨率结果。通过自集成，我们的DAT+在所有包含三个因素的基准数据集上都优于对比方法。同时，DAT的性能优于先前方法，除了在Urban100数据集上的PSNR值（×4）与CAT-A相比。具体来说，与SwinIR和CAT-A相比，我们的DAT在Manga109数据集上（×2）取得了显著提升，分别提高了0.41 dB和0.23 dB。此外，小型视觉模型DAT-S的性能也达到了先前方法的水平或更好。所有这些定量结果都表明，从块间和块内聚合空间和通道信息可以有效提高图像重建质量。

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 sub-optimal 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.

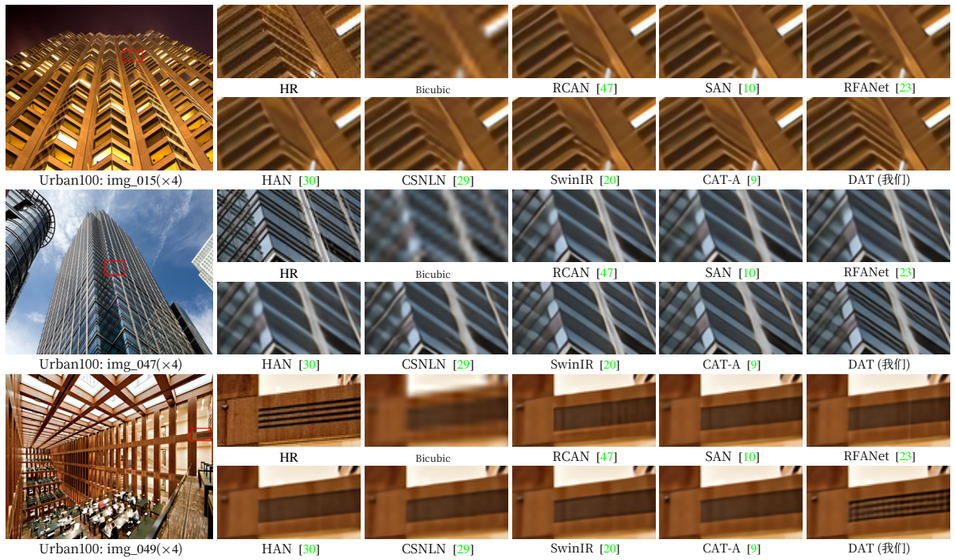


图6：在某些具有挑战性的情况下，图像超分辨率（×4）的可视化比较。

方法	EDSR [21]	[47]	[20]	[9]	DAT (我们)	DAT (我们)	DAT-2 (我们)
参数（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

表3：模型复杂度比较（×4）。在Urban100和Manga109上的PSNR（dB）、FLOPs和Params均有报告。

视觉结果。我们在图 6中展示了视觉对比（×4）。在一些具有挑战性的场景中，先前方法可能会出现模糊伪影、扭曲或不准确的纹理恢复。相反，我们的方法有效减轻了伪影，保留了更多结构细节。例如，在img_015中，大多数对比方法几乎无法恢复细节并产生不希望的伪影。然而，我们的DAT可以正确恢复结构并呈现清晰的纹理。我们在img_047 和img_049中也观察到了类似的现象。这主要是因为我们的方法通过从不同维度提取复杂特征，具有更强的表示能力。

4.4. 模型大小分析

我们进一步在计算复杂度（例如，FLOPs）、参数数量以及在 ×4 表3中<style id='14'>尺度的性能方面，将我们的方法与几种图像超分辨率方法进行了比较。我们将输出大小设置为 3×512×512 以计算FLOPs，并在Urban100和Manga109上使用PSNR评估性能。与CAT-A [9],相比，我们的DAT具有相当或更好的性能，同时计算复杂度和模型大小更小。此外，DAT-S在FLOPs和参数数量方面优于SwinIR [20]，并获得了优异的性能。此外，为了进一步揭示我们方法在模型大小和性能之间的更好权衡，我们引入了一个额外的变体模型DAT-2，其详细信息在补充材料中。

5. 结论

在本文中，我们提出了双聚合Transformer（DAT），这是一种新的图像超分辨率Transformer模型。我们的DAT以块间和块内双重方式聚合空间和通道特征，以实现强大的表示能力。具体来说，连续的Transformer模块交替应用空间窗口和通道自注意力。DAT可以通过这种替代策略对全局依赖进行建模，并在空间和通道维度之间实现块间特征聚合。此外，我们提出了自适应交互模块（AIM）和空间门控前馈网络（SGFN）来增强每个模块，并在两个维度之间实现块内特征聚合。AIM增强了来自相应维度的两种自注意力机制的建模能力。同时，SGFN为前馈网络补充了非线性空间信息。大量实验表明，DAT优于先前方法。

致谢。这项工作部分由国家自然科学基金资助（62141220、61972253、U1908212、U19B2035）、上海市科学技术重大专项（2021SHZDZX0102）和华为技术有限公司（芬兰）项目支持。

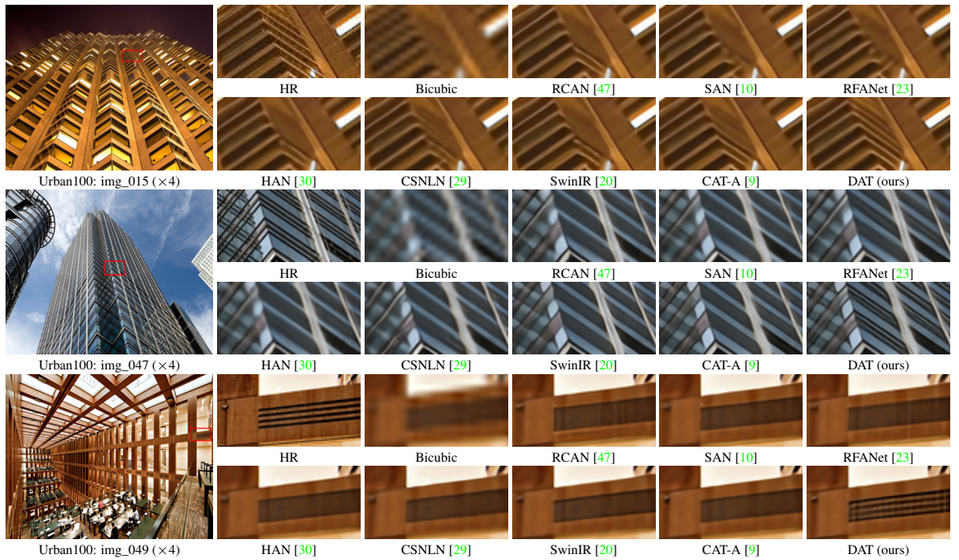


Figure 6: Visual comparison for image SR (×4) in some challenging cases.

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.

Acknowledgments. This work is supported in part by NSFC grant (62141220, 61972253, U1908212, U19B2035), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and Huawei Technologies Oy (Finland) Project.

参考文献

[1] 阿劳埃丁·阿里、雨果·图维隆、玛蒂尔德·卡隆、皮奥特·博亚诺夫斯基、马蒂斯·杜泽、阿曼德·茹林、伊万·拉普捷夫、娜塔莉亚·涅沃娃、加布里埃尔·辛纳维、雅各布·韦伯克等。Xcit：交叉协方差图像变换器。发表于 NeurIPS，2021。3[2] 马科·贝维拉卡、阿莉娜·鲁米、克里斯汀·吉勒莫和玛丽·莱娜·阿尔贝里-莫雷尔。基于非负邻域嵌入的低复杂度单图像超分辨率。发表于 BMVC，2012。5[3] 浩宇·陈、谷金金和智·张。用于图像超分辨率注意力网络中的注意力。arXiv 预印本 arXiv:2104.09497，2021。2[4] 陈浩、王云鹤、郭天宇、徐昌、邓一波、刘振华、马思伟、徐春静、徐超和高文。预训练图像处理变换器。发表于 CVPR，2021。2[5] 亮宇·陈、楚晓杰、张祥宇和建·孙。图像恢复的简单基线。发表于 ECCV，2022。5[6] 龙·陈、汉旺·张、俊·肖、力强·聂、建·邵、伟·刘和达特·桑·蔡。SCA-CNN：卷积网络中空间和通道注意力的图像描述。发表于 CVPR，2017。2[7] 强·陈、启曼·吴、建·王、庆浩·胡、涛·胡、尔瑞·丁、建·程和景东·王。Mixformer：跨窗口和维度混合特征。发表于 CVPR，2022。1, 2[8] 郑晨、张宇伦、谷金金、孔令鹤和杨晓康。图像超分辨率的递归泛化变换器。arXiv 预印本 arXiv:2303 06373，2023 2 。 .[9] 郑晨、张宇伦、谷金金、永兵·张、孔令鹤和辛·袁。图像恢复的交叉聚合变换器。发表于 NeurIPS，2022。1, 7, 8[10] 涛·戴、建瑞·蔡、永兵·张、舒涛·夏和雷·张。单图像超分辨率的二阶注意力网络。发表于 CVPR，2019。1, 7, 8[11] 明宇·丁、斌·肖、诺埃尔·科德拉、平·罗、景东·王和陆·袁。DaViT：双重注意力视觉变换器。发表于 ECCV，2022。1, 2[12] 超·董、陈·长·刘、凯明·何和晓鸥·唐。用于图像超分辨率的深度卷积网络。发表于 ECCV，2014。1, 2[13] 阿列克谢·多索维茨基、卢卡斯·贝耶、亚历山大·科列斯尼科夫、迪尔克·魏森博恩、晓华·翟、托马斯·翁特瑟宁、莫斯塔法·德赫加尼、马蒂亚斯·明德纳、乔治·海格尔德、西尔维恩·盖利等。一幅图像值16x16个词：用于图像识别的尺度变换器。发表于 ICLR，2021。1, 2[14] 俊·傅、静·刘、海杰·田、永·李、永军·鲍、智伟·方和韩庆·陆。场景分割的双重注意力网络。发表于 CVPR，2019。2[15] 穆罕默德·哈里斯、格雷格·沙赫纳罗维奇和野村道·宇田。用于超分辨率的深度反投影网络。发表于 CVPR，2018。5[16] 凯明·何、张祥宇、少庆·任和建·孙。用于图像识别的深度残差学习。发表于 CVPR，2016。2

[17] 解胡, 李沈, 和 孙刚. Squeeze-and-excitation networks. In CVPR, 2018. 4[18] 黄嘉斌, 阿比什克·辛格, 和 纳雷德拉·阿胡贾. Single image super-resolution from transformed self-exemplars. In CVPR, 2015. 5, 6[19] 迪德里克·金马 和 吉米·巴. Adam: 随机优化的方法. In ICLR, 2015. 6[20] 梁景云, 曹继章, 孙国雷, 张凯, 卢克·范·古尔, 和 拉杜·蒂莫夫特. Swinir: 使用Swin Transformer的图像恢复. In ICCVW, 2021. 1, 2, 5, 7, 8[21] Bee Lim, Son Sanghyun, Heewon Kim, Nah Seungjun, 和 Lee Kyoung Mu. 用于单图像超分辨率的增强深度残差网络. In CVPRW, 2017. 5, 6, 7, 8[22] 刘汉晓, 戴子航, David So, 和 Le Quoc V. 关注mlps. In NeurIPS, 2021. 5[23] 刘杰, 张文杰, 唐宇婷, 唐杰, 和 吴刚山. 用于图像超分辨率的残差特征聚合网络. In CVPR, 2020. 7, 8[24] 刘泽, 林宇彤, 曹越, 胡汉, 魏奕轩, 张铮, Stephen Lin, 和 郭百宁. Swin t ransformer: 使用移位窗口的分层视觉Transformer. In ICCV, 2021. 1, 2, 3[25] Salma Abdel Magid, 张宇伦, 魏东来, Jang Won-Dong, Lin Zudi, Fu Yun, 和 Hanspeter Pfister. 用于图像超分辨率的动态高通滤波和多光谱注意力. In ICCV, 2021. 7[26] David Martin, Charless Fowlkes, Doron Tal, 和 Jitendra Malik. 一个包含人类分割自然图像的数据库及其在评估分割算法和测量生态统计中的应用. In ICCV, 2001. 5[27] 松井佑介, 伊藤绫太, Aramaki Yuji, 藤本朝马, 小川寿, 山崎俊彦, 和 Aizawa Kiyoharu. 使用Manga109数据集的基于草图的漫画检索. Multimedia Tools and Applications, 2017. 5[28] 梅一群, 范宇辰, 和 周宇倩. 具有非局部稀疏注意力的图像超分辨率. In CVPR, 2021. 7[29]梅一群, 范宇辰, 周宇倩, 黄立超, 黄托马斯, 和 石汉弗莱. 具有跨尺度非局部注意力和彻底自示例挖掘的图像超分辨率. In CVPR, 2020. 1, 2, 7, 8[30] 牛本, 温伟利, 任文奇, 张向德, 杨连平, 王淑珍, 张凯浩, 曹晓春, 和 沈海峰. 通过整体注意力网络的单图像超分辨率. In ECCV, 2020. 2, 7, 8[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Lin Zeming, Desmaison Alban, Luca Antiga, 和 Adam Lerer. pytorch中的自动微分. 2017. 6[32] Olaf Ronneberger, Philipp Fischer, 和 Thomas Brox. U-net: 用于生物医学图像分割的卷积网络. In MICCAI, 2015. 2[33] 石文哲, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, 和 王泽瀚. 使用高效亚像素卷积神经网络的实时单图像和视频超分辨率. In CVPR, 2016. 3

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia 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 attention network for image super-resolution. *arXiv preprint arXiv:2104.09497*, 2021. 2

[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 *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 transformers. 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, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers 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 networks. 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 using 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. *Multimedia 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 exhaustive 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, Alban 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] 拉杜·蒂莫夫特、埃里克·阿格斯特森、卢克·范·古尔、明-玄·杨、雷·张、Bee Lim、韩·张、Heewon Kim、Seungjun Nah、Kyoung Mu Lee 等人。Ntire 2017 单图像超分辨率挑战：方法与结果。在 CVPRW，2017。5，6 [35] 伊利亚·奥·托尔斯蒂金、尼尔·豪尔斯比、亚历山大·科列斯尼科夫、卢卡斯·贝耶、晓华·翟、托马斯·翁特瑟宁、杰西卡·杨、安德烈亚斯·施泰纳、丹尼尔·凯瑟斯、雅各布·乌斯克雷特 等人。Mlp-mixer：用于视觉的全-mlp架构。在 NeurIPS，2021。2[36] 郑中·图、霍赛因·塔莱比、韩·张、冯·杨、佩扬·米兰法尔、艾伦·博维克、银晓·李。Maxim：用于图像处理的Multi-axis mlp。在 CVPR，2022。5[37] 郑中·图、霍赛因·塔莱比、韩·张、冯·杨、佩扬·米兰法尔、艾伦·博维克、银晓·李。Maxvit：Multi-axis vision transformer。在 ECCV，2022。2[38] 阿希什·瓦桑维、诺姆·沙泽尔、尼基·帕尔马、雅各布·乌斯克雷特、Llion Jones、Aidan N Gomez、Łukasz Kaiser、和 Illia Polosukhin。Attention is all you need。在 NeurIPS，2017。2，5，6[39] 文海·王、恩泽·谢、翔·李、邓平·范、凯涛·宋、丁·梁、通·卢、平·罗、玲·肖。Pyramid vision transformer：一个基于卷积的多功能主干网络，用于密集预测。在 ICCV，2021。1，2[40] 文晓·王、卢·姚、龙·陈、斌斌·林、邓·蔡、晓飞·何、伟·刘。Crossformer：一个基于跨尺度注意力的多功能视觉变压器。在 ICLR，2022。3[41] 周·王、艾伦·C·博维克、Hamid R Sheikh、和 Eero P Simoncelli。图像质量评估：从错误可见性以结构相似性为目标。TIP，2004. 5[42] 振东·王，晓东·存，建民·鲍，文刚·周，建庄·刘，和厚强·李。Uformer：一种通用的U形Transformer用于图像恢复。在CVPR，2022。1，2[43] 桑根·吴，钟灿·朴，俊英·李，和因·斯科恩。Cbam：卷积块注意力模块。在ECCV，2018。2[44] 赛义德·瓦卡斯·扎米尔，阿迪亚·阿罗拉，萨拉曼·汗，穆纳瓦尔·海亚特，法哈德·沙巴兹·汗，和明轩·杨。Restormer：用于高分辨率图像恢复的高效Transformer。在CVPR，2022。1，2，3[45] 罗曼·泽德，迈克尔·埃拉德，和马坦·普罗特。使用稀疏表示的单图像缩放。在第7届国际曲线曲面会议论文集，2010。5[46]欣欣东·张，辉·曾，石·郭，和雷·张。高效长距离注意力网络用于图像超分辨率。在ECCV，2022。1，2，7 [47] 张宇伦，坤鹏·李，凯·李，立晨·王，斌能·钟，和Fu Yun。使用非常深的残差通道注意力网络的图像超分辨率。在ECCV，2018。1，2，7，8[48] 张宇伦，坤鹏·李，凯·李，斌能·钟，和Fu Yun。用于图像恢复的残差非局部注意力网络。在ICLR，2019。2[49] 大权·周，智定·余，恩泽·谢，肖超伟，Aniashree Anandkumar，嘉师·冯，和何塞·M·阿尔瓦雷斯。理解视觉Transformer中的鲁棒性。在ICML，2022。2

[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