# Activating More Pixels in Image Super-Resolution Transformer

# 激活图像超分辨率变换器中的更多像素

Xiangyu Chen Xintao Wang Jiantao Zhou Yu Qiao Chao Dong

向宇 Chen 新涛 Wang 建滔 Zhou 裘宇 Qiao 董超 Dong

State Key Laboratory of Internet of Things for Smart City, University of Macau

澳门大学智能城市物联网国家重点实验室

Shenzhen Key Lab of Computer Vision and Pattern Recognition,

深圳计算机视觉与模式识别重点实验室，

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

深圳先进技术研究院，中国科学院

Shanghai Artificial Intelligence Laboratory ARC Lab, Tencent PCG

上海人工智能实验室 ARC 实验室，腾讯 PCG

https://github.com/XPixelGroup/HAT

Abstract

摘要

Transformer-based methods have shown impressive performance in low-level vision tasks, such as image super-resolution. However, we find that these networks can only utilize a limited spatial range of input information through attribution analysis. This implies that the potential of Transformer is still not fully exploited in existing networks. In order to activate more input pixels for better reconstruction, we propose a novel Hybrid Attention Transformer (HAT). It combines both channel attention and window-based self-attention schemes, thus making use of their complementary advantages of being able to utilize global statistics and strong local fitting capability. Moreover, to better aggregate the cross-window information, we introduce an overlapping cross-attention module to enhance the interaction between neighboring window features. In the training stage, we additionally adopt a same-task pre-training strategy to exploit the potential of the model for further improvement. Extensive experiments show the effectiveness of the proposed modules, and we further scale up the model to demonstrate that the performance of this task can be greatly improved. Our overall method significantly outperforms the state-of-the-art methods by more than .

基于变换器的算法在低层次视觉任务中，如图像超分辨率，已经展现出令人印象深刻的表现。然而，通过归属分析，我们发现这些网络只能利用输入信息的一个有限空间范围。这意味着在现有网络中，变换器的潜力仍未被完全挖掘。为了更好地重建，激活更多输入像素，我们提出了一个新颖的混合注意力变换器（HAT）。它结合了通道注意力和基于窗口的自注意力机制，从而利用了它们能够利用全局统计信息和强大的局部拟合能力的互补优势。此外，为了更好地聚合跨窗口信息，我们引入了一个重叠的交叉注意力模块，以增强相邻窗口特征之间的交互。在训练阶段，我们还额外采用了一种同任务预训练策略，以进一步挖掘模型的潜力。广泛的实验证明了所提出模块的有效性，我们还进一步扩展了模型规模，以证明这项任务的表现可以得到显著改善。我们的整体方法比现有最佳方法超出 以上。

# 1. Introduction

# 1. 引言

Single image super-resolution (SR) is a classic problem in computer vision and image processing. It aims to reconstruct a high-resolution image from a given low-resolution input. Since deep learning has been successfully applied to the SR task [10], numerous methods based on the convolutional neural network (CNN) have been proposed and almost dominate this field in the past few years. Recently, due to the success in natural language processing, Transformer [53] has attracted the attention of the computer vision community. After making rapid progress on high-level vision tasks , Transformer-based methods are also developed for low-level vision tasks , as well as for SR . Especially, a newly designed network, SwinIR [31], obtains a breakthrough improvement in this task.

单幅图像超分辨率（SR）是计算机视觉和图像处理中的经典问题。它的目标是根据给定的低分辨率输入重建高分辨率图像。自从深度学习成功应用于SR任务以来[10]，基于卷积神经网络（CNN）的许多方法被提出 ，并且在过去几年几乎主导了这一领域。最近，由于在自然语言处理上的成功，Transformer [53] 引起了计算机视觉社区的注意。在高级视觉任务上取得快速进展后 ，基于Transformer的方法也被开发用于低级视觉任务 ，以及用于SR 。特别是，一个新设计的网络，SwinIR [31]，在这个任务上取得了突破性的改进。

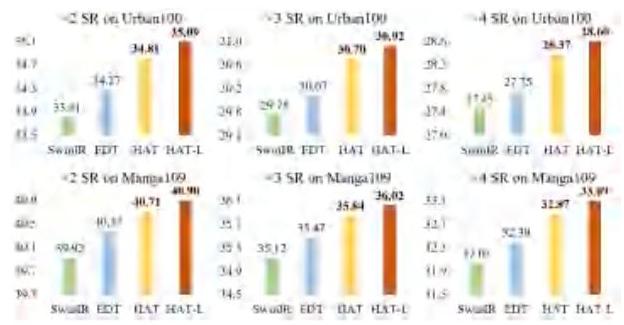


Figure 1. Performance comparison on PSNR(dB) of the proposed HAT with the state-of-the-art methods SwinIR [31] and EDT [27]. HAT-L represents a larger variant of HAT. Our approach can surpass the state-of-the-art methods by .

图1. 提出的HAT与最先进方法SwinIR [31] 和EDT [27] 在PSNR（分贝）上的性能比较。HAT-L代表HAT的一个较大变体。我们的方法可以超越当前最先进的方法 。

Despite the success, "why Transformer is better than CNN" remains a mystery. An intuitive explanation is that this kind of network can benefit from the self-attention mechanism and utilize long-range information. Thus, we employ the attribution analysis method LAM [15] to examine the involved range of utilized information for reconstruction in SwinIR. Interestingly, we find that SwinIR does NOT exploit more input pixels than CNN-based methods (e.g., RCAN [68]) in super-resolution, as shown in Fig. 2. Besides, although SwinIR obtains higher quantitative performance on average, it produces inferior results to RCAN in some samples, due to the limited range of utilized information. These phenomena illustrate that Transformer has a stronger ability to model local information, but the range of its utilized information needs to be expanded. In addition, we also find that blocking artifacts would appear in the intermediate features of SwinIR, as depicted in Fig. 3. It demonstrates that the shift window mechanism cannot perfectly realize cross-window information interaction.

尽管取得了成功，但“为什么Transformer比CNN更好”仍然是一个谜。一种直观的解释是，这类网络可以从自注意力机制中受益，并利用长距离信息。因此，我们采用归因分析方法LAM [15]来检查SwinIR在重建中利用的信息范围。有趣的是，我们发现SwinIR在超分辨率中并没有比基于CNN的方法（例如RCAN [68]）利用更多的输入像素，如图2所示。此外，尽管SwinIR在平均定量性能上取得了更高的成绩，但由于利用的信息范围有限，在一些样本中，它产生的结果不如RCAN。这些现象说明Transformer具有更强的建模局部信息的能力，但其利用信息的范围需要扩大。此外，我们还发现SwinIR的中间特征中会出现阻塞伪影，如图3所示。这表明位移窗口机制无法完美实现跨窗口信息交互。

Corresponding author.

对应作者。

To address the above-mentioned limitations and further develop the potential of Transformer for SR, we propose a Hybrid Attention Transformer, namely HAT. Our HAT combines channel attention and self-attention schemes, in order to take advantage of the former’s capability in using global information and the powerful representative ability of the latter. Besides, we introduce an overlapping cross-attention module to achieve more direct interaction of adjacent window features. Benefiting from these designs, our model can activate more pixels for reconstruction and thus obtains significant performance improvement.

为了解决上述局限性并进一步开发Transformer在SR方面的潜力，我们提出了一个混合注意力Transformer，即HAT。我们的HAT结合了通道注意力和自注意力方案，以利用前者在利用全局信息方面的能力和后者的强大表征能力。此外，我们引入了一个重叠交叉注意力模块，以实现相邻窗口特征更直接的交互。得益于这些设计，我们的模型可以激活更多像素进行重建，从而获得了显著的性能提升。

Since Transformers do not have an inductive bias like CNNs, large-scale data pre-training is important to unlock the potential of such models. In this work, we provide an effective same-task pre-training strategy. Different from IPT [6] using multiple restoration tasks for pre-training and EDT [27] using multiple degradation levels for pre-training, we directly perform pre-training using large-scale dataset on the same task. We believe that large-scale data is what really matters for pre-training, and experimental results also show the superiority of our strategy. Equipped with the above designs, HAT can surpass the state-of-the-art methods by a huge margin , as shown in Fig. 1.

由于变换器不像卷积神经网络（CNNs）那样具有归纳偏置，大规模数据预训练对于释放此类模型的潜力至关重要。在这项工作中，我们提供了一种有效的同任务预训练策略。与使用多种恢复任务进行预训练的IPT [6]以及使用多种退化级别进行预训练的EDT [27]不同，我们直接在相同任务的大规模数据集上进行预训练。我们认为大规模数据是预训练中真正重要的因素，实验结果也显示了我们的策略的优越性。配备了上述设计，HAT能够在图1所示的性能上大大超越现有最佳方法 。

Contributions: 1) We design a novel Hybrid Attention Transformer (HAT) that combines self-attention, channel attention and a new overlapping cross-attention to activate more pixels for better reconstruction. 2) We propose an effective same-task pre-training strategy to further exploit the potential of SR Transformer and show the importance of large-scale data pre-training for the task. 3) Our method achieves state-of-the-art performance. By further scaling up HAT to build a big model, we greatly extend the performance upper bound of the SR task.

贡献：1）我们设计了一种新颖的混合注意力变换器（HAT），它结合了自注意力、通道注意力和一种新的重叠交叉注意力，以激活更多像素以实现更好的重建。2）我们提出了一种有效的同任务预训练策略，以进一步挖掘SR变换器的潜力，并展示了大规模数据预训练对任务的重要性。3）我们的方法达到了现有最佳性能。通过进一步扩大HAT构建大型模型，我们极大地提高了SR任务性能的上限。

# 2. Related Work

# 2. 相关工作

# 2.1. Deep Networks for Image SR

# 2.1. 用于图像超分辨率的深度网络

Since SRCNN [10] first introduces deep convolution neural networks (CNNs) to the image SR task and obtains superior performance over conventional SR methods, numerous deep networks , have been proposed for SR to further improve the reconstruction quality. For instance, many methods apply more elaborate convolution module designs, such as residual block and dense block , to enhance the model representation ability. Several works explore more different frameworks like recursive neural network and graph neural network [72]. To improve perceptual quality, introduce adversarial learning to generate more realistic results. By using attention mechanism, achieve further improvement in terms of reconstruction fidelity. Recently, a series of Transformer-based networks are proposed and constantly refresh the state-of-the-art of SR task, showing the powerful representation ability of Transformer.

自SRCNN [10] 首次将深度卷积神经网络（CNNs）引入图像超分辨率（SR）任务，并在传统SR方法上获得优越性能以来，许多深度网络 ， 被提出用于SR以进一步提高重建质量。例如，许多方法应用了更精细的卷积模块设计，如残差块 和密集块 ，以增强模型的表征能力。一些工作探索了更多不同的框架，如递归神经网络 和图神经网络 [72]。为了提高感知质量， 引入对抗性学习以生成更真实的结果。通过使用注意力机制， 在重建保真度方面取得了进一步的改进。最近，一系列基于Transformer的网络 被提出，并不断刷新SR任务的最先进水平，展示了Transformer强大的表征能力。

To better understand the working mechanisms of SR networks, several works are proposed to analyze and interpret the SR networks. LAM [15] adopts the integral gradient method to explore which input pixels contribute most to the final performance. DDR [37] reveals the deep semantic representations in SR networks based on deep feature dimensionality reduction and visualization. FAIG [62] aims to find discriminative filters for specific degradations in blind SR. RDSR [23] introduces channel saliency map to demonstrate that Dropout can help prevent co-adapting for real-SR networks. SRGA [38] aims to evaluate the generalization ability of SR methods. In this work, we exploit LAM [15] to analyse and understand the behavior of SR networks.

为了更好地理解SR网络的工作机制，已经提出了几项工作来分析和解释SR网络。LAM [15] 采用积分梯度方法来探索哪些输入像素对最终性能的贡献最大。DDR [37] 通过深度特征降维和可视化揭示了SR网络中的深层语义表征。FAIG [62] 旨在为盲SR中的特定退化找到判别性滤波器。RDSR [23] 引入通道显著性图来证明Dropout有助于防止真实SR网络中的共适应。SRGA [38] 旨在评估SR方法的泛化能力。在这项工作中，我们利用LAM [15] 来分析和理解SR网络的行为。

# 2.2. Vision Transformer

# 2.2. 视觉Transformer

Recently, Transformer [53] has attracted the attention of computer vision community due to its success in the field of natural language processing. A series of Transformer-based methods have been developed for high-level vision tasks, including image classification , object detection , segmentation , etc. Although vision Transformer has shown its superiority on modeling long-range dependency , there are still many works demonstrating that the convolution can help Transformer achieve better visual representation . Due to the impressive performance, Transformer has also been introduced for low-level vision tasks . Specifically, IPT [6] develops a ViT-style network and introduces multi-task pre-training for image processing. SwinIR [31] proposes an image restoration Transformer based on [39]. VRT [30] introduces Transformer-based networks to video restoration. EDT [27] adopts self-attention mechanism and multi-related-task pre-training strategy to further refresh the state-of-the-art of SR. However, existing works still cannot fully exploit the potential of Transformer, while our method can activate more input pixels for better reconstruction.

最近，Transformer [53] 由于在自然语言处理领域的成功而引起了计算机视觉社区的注意。一系列基于Transformer的方法 已经被开发用于高级视觉任务，包括图像分类 、目标检测 、分割 等。尽管视觉Transformer在建模长距离依赖性 上表现出了其优越性，但仍有大量工作证明卷积可以帮助Transformer实现更好的视觉表征 。由于出色的性能，Transformer也被引入到低级视觉任务 中。具体来说，IPT [6] 开发了一个基于ViT风格的网络，并引入了多任务预训练用于图像处理。SwinIR [31] 提出了一个基于 [39] 的图像修复Transformer。VRT [30] 引入了基于Transformer的网络进行视频修复。EDT [27] 采用自注意力机制和多相关任务预训练策略，进一步刷新了超分辨率（SR）的最先进水平。然而，现有工作仍无法完全发掘Transformer的潜力，而我们的方法可以激活更多输入像素以实现更好的重建。

# 3. Methodology

# 3. 方法论

# 3.1. Motivation

# 3.1. 研究动机

Swin Transformer [39] has already presented excellent performance in image super-resolution [31]. Then we are eager to know what makes it work better than CNN-based methods. To reveal its working mechanisms, we resort to a diagnostic tool - LAM [15], which is an attribution method designed for SR. With LAM, we could tell which input pixels contribute most to the selected region. As shown in Fig. 2, the red marked points are informative pixels that contribute to the reconstruction. Intuitively, the more information is utilized, the better performance can be ob-

Swin Transformer [39] 在图像超分辨率 [31] 上已经展示了卓越的性能。然后我们迫切地想知道是什么让它比基于CNN的方法工作得更好。为了揭示其工作机制，我们求助于一个诊断工具 - LAM [15]，这是一种为超分辨率设计的归因方法。通过LAM，我们可以判断哪些输入像素对选定区域贡献最大。如图2所示，红色标记的点是为重建做出贡献的信息像素。直观上，利用的信息越多，性能可能越好。

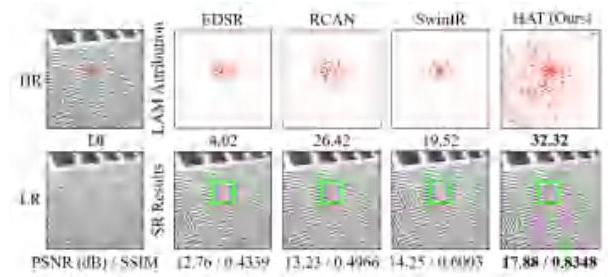


Figure 2. LAM [15] results for different networks. The LAM attribution reflects the importance of each pixel in the input LR image when reconstructing the patch marked with a box. Diffusion index (DI) [15] reflects the range of involved pixels. A higher DI represents a wider range of utilized pixels. The results indicate that SwinIR utilizes less information compared to RCAN, while HAT uses the most pixels for reconstruction.

图 2。不同网络的 LAM [15] 结果。LAM 归因反映了在重建用方框标记的贴图时，输入 LR 图像中每个像素的重要性。扩散指数 (DI) [15] 反映了参与像素的范围。DI 值越高，表示使用了更广泛的像素。结果显示，与 RCAN 相比，SwinIR 使用的信息较少，而 HAT 用于重建的像素最多。

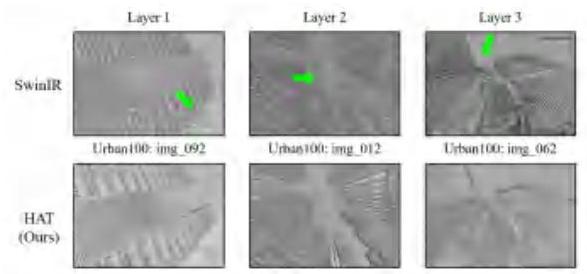


Figure 3. The blocking artifacts appear in the intermediate features of SwinIR [31]. "Layer N " represents the intermediate features after the layer (i.e., RSTB in SwinIR and RHAG in HAT.)

图 3。SwinIR [31] 的中间特征中出现了块状伪影。"第 N 层" 表示在 层之后（即 SwinIR 中的 RSTB 和 HAT 中的 RHAG）的中间特征。

tained. This is true for CNN-based methods, as comparing RCAN [68] and EDSR [32]. However, for the Transformer-based method - SwinIR, its LAM does not show a larger range than RCAN. This is in contradiction with our common sense, but could also provide us with additional insights. First, it implies that SwinIR has a much stronger mapping ability than CNN, and thus could use less information to achieve better performance. Second, SwinIR may restore wrong textures due to the limited range of utilized pixels, and we think it can be further improved if it could exploit more input pixels. Therefore, we aim to design a network that can take advantage of similar self-attention while activating more pixels for reconstruction. As depicted in Fig. 2, our HAT can see pixels almost all over the image and restore correct and clear textures.

这种现象对于基于 CNN 的方法来说确实如此，比如比较 RCAN [68] 和 EDSR [32]。然而，对于基于 Transformer 的方法 - SwinIR，其 LAM 并没有显示出比 RCAN 更大的范围。这与我们的常识相矛盾，但也可能为我们提供额外的洞见。首先，这意味着 SwinIR 比 CNN 具有更强的映射能力，因此可以使用更少的信息来实现更好的性能。其次，由于使用的像素范围有限，SwinIR 可能会恢复错误的纹理，我们认为如果它能利用更多的输入像素，则可以进一步改进。因此，我们旨在设计一个网络，能够利用类似的自注意力机制同时为重建激活更多的像素。如图 2 所示，我们的 HAT 几乎可以看到图像上的所有像素，并恢复正确且清晰的纹理。

Besides, we can observe obvious blocking artifacts in the intermediate features of SwinIR, as shown in Fig. 3. These artifacts are caused by the window partition mechanism, which suggests that the shifted window mechanism is inefficient to build the cross-window connection. Some works for high-level vision tasks also point out that enhancing the connection among windows can improve the window-based self-attention methods. Thus, we strengthen cross-window information interactions when designing our approach and the blocking artifacts in the intermediate features obtained by HAT are significantly alleviated.

此外，我们可以在图3所示的SwinIR中间特征中观察到明显的阻塞伪影。这些伪影是由窗口划分机制引起的，这表明平移窗口机制在建立跨窗口连接方面效率不高。一些针对高级视觉任务的研究 也指出，增强窗口之间的连接可以提高基于窗口的自注意力方法。因此，在设计我们的方法时，我们加强了跨窗口信息交互，并且通过HAT获得的中间特征中的阻塞伪影得到了显著缓解。

# 3.2. Network Architecture

# 3.2. 网络架构

# 3.2.1 The Overall Structure

# 3.2.1 整体结构

As shown in Fig. 4, the overall network consists of three parts, including shallow feature extraction, deep feature extraction and image reconstruction. The architecture design is widely used in previous works [31, 68]. Specifically, for a given low-resolution (LR) input , we first exploit one convolution layer to extract the shallow feature , where and denote the channel number of the input and the intermediate feature. Then, a series of residual hybrid attention groups (RHAG) and one convolution layer are utilized to perform the deep feature extraction. After that, we add a global residual connection to fuse shallow features and deep features , and then reconstruct the high-resolution result via a reconstruction module. As depicted in Fig. 4, each RHAG contains several hybrid attention blocks (HAB), an overlapping cross-attention block (OCAB) and a convolution layer with a residual connection. For the reconstruction module, the pixel-shuffle method [47] is adopted to up-sample the fused feature. We simply use loss to optimize the network parameters.

如图4所示，整个网络由三部分组成，包括浅层特征提取、深层特征提取和图像重建。这种架构设计在之前的工作 [31, 68] 中得到了广泛应用。具体来说，对于给定的低分辨率（LR）输入 ，我们首先利用一个卷积层提取浅层特征 ，其中 和 分别表示输入和中间特征的通道数。然后，使用一系列残差混合注意力组（RHAG）和一个 卷积层 进行深层特征提取。之后，我们添加了一个全局残差连接来融合浅层特征 和深层特征 ，并通过一个重建模块重建高分辨率结果。如图4所示，每个RHAG包含几个混合注意力块（HAB）、一个重叠交叉注意力块（OCAB）和一个带有残差连接的 卷积层。对于重建模块，我们采用了像素洗牌方法 [47] 来上采样融合特征。我们简单地使用 损失来优化网络参数。

# 3.2.2 Hybrid Attention Block (HAB)

# 3.2.2 混合注意力块（HAB）

As shown in Fig. 2, more pixels are activated when channel attention is adopted, as global information is involved to calculate the channel attention weights. Besides, many works illustrate that convolution can help Transformer get better visual representation or achieve easier optimization . Therefore, we incorporate a channel attention-based convolution block into the standard Transformer block to enhance the representation ability of the network. As demonstrated in Fig. 4, a channel attention block (CAB) is inserted into the standard Swin Transformer block after the first LayerNorm (LN) layer in parallel with the window-based multi-head self-attention (W-MSA) module. Note that shifted window-based self-attention (SW-MSA) is adopted at intervals in consecutive HABs similar to . To avoid the possible conflict of CAB and MSA on optimization and visual representation, a small constant is multiplied to the output of CAB. For a given input feature , the whole process of HAB is computed as

如图2所示，当采用通道注意力时，会有更多的像素被激活，因为全局信息被用于计算通道注意力权重。此外，许多研究说明卷积可以帮助Transformer获得更好的视觉表现或实现更易于优化的效果 。因此，我们将基于通道注意力的卷积块整合到标准Transformer块中，以增强网络的表征能力。如图4所示，在标准Swin Transformer块中的第一个层归一化（LN）层之后，并行于基于窗口的多头自注意力（W-MSA）模块，插入了一个通道注意力块（CAB）。注意，在连续的HAB之间间隔地采用了移位窗口基于自注意力（SW-MSA） 。为了避免CAB和MSA在优化和视觉表征上可能出现的冲突，将一个小的常数 乘以CAB的输出。对于给定的输入特征 ，HAB的整个过程计算如下：

where and denote the intermediate features. represents the output of HAB. Especially, we treat each pixel as a token for embedding (i.e., set patch size as 1 for patch embedding following [31]). MLP denotes a multilayer perceptron. For calculation of the self-attention module, given an input feature of size , it is first partitioned into local windows of size , then self-attention is calculated inside each window. For a local window feature , the query, key and value matrices are computed by linear mappings as and . Then the window-based self-attention is formulated as

其中 和 表示中间特征。 表示HAB的输出。特别地，我们将每个像素视为一个嵌入令牌（即，按照[31]将贴片大小设置为1进行贴片嵌入）。MLP表示多层感知器。在自注意力模块的计算中，给定一个大小为 的输入特征，它首先被划分为 个大小为 的局部窗口，然后在每个窗口内计算自注意力。对于一个局部窗口特征 ，查询、键和值矩阵通过线性映射计算为 和 。然后，基于窗口的自注意力被表述为：



Figure 4. The overall architecture of HAT and the structure of RHAG and HAB.

图4. HAT的整体架构以及RHAG和HAB的结构。

where represents the dimension of query/key. denotes the relative position encoding and is calculated as [53]. Note that we use a large window size to compute self-attention, since we find it significantly enlarges the range of used pixels, as depicted in Sec.4.2. Besides, to build the connections between neighboring non-overlapping windows, we also utilize the shifted window partitioning approach [39] and set the shift size to half of the window size.

其中 表示查询/键的维度。 表示相对位置编码，计算方式如 [53] 所示。注意，我们使用较大的窗口尺寸来计算自注意力，因为我们发现这显著增加了使用的像素范围，如图 Sec.4.2 所示。此外，为了建立相邻非重叠窗口之间的连接，我们还采用了移位窗口划分方法 [39]，并将移位大小设置为窗口大小的一半。

A CAB consists of two standard convolution layers with a GELU activation [17] and a channel attention (CA) module, as shown in Fig. 4. Since the Transformer-based structure often requires a large number of channels for token embedding, directly using convolutions with constant width incurs a large computation cost. Thus, we compress the channel numbers of the two convolution layers by a constant . For an input feature with channels, the channel number of the output feature after the first convolution layer is squeezed to , then the feature is expanded to channels through the second layer. Next, a standard CA module [68] is exploited to adaptively rescale channel-wise features.

CAB 由两个标准卷积层和一个 GELU 激活函数 [17] 以及一个通道注意力（CA）模块组成，如图 4 所示。由于基于 Transformer 的结构通常需要大量的通道进行标记嵌入，直接使用固定宽度的卷积会导致计算成本很大。因此，我们通过一个常数 来压缩两个卷积层的通道数。对于一个具有 通道的输入特征，经过第一个卷积层后输出特征的通道数被压缩到 ，然后特征通过第二层扩展到 通道。接下来，利用一个标准的 CA 模块 [68] 来自适应地调整通道特征。

# 3.2.3 Overlapping Cross-Attention Block (OCAB)

# 3.2.3 重叠交叉注意力块（OCAB）

We introduce OCAB to directly establish cross-window connections and enhance the representative ability for the window self-attention. Our OCAB consists of an overlapping cross-attention (OCA) layer and an MLP layer similar to the standard Swin Transformer block [39]. But for OCA, as depicted in Fig. 5, we use different window sizes to partition the projected features. Specifically, for the of the input feature is partitioned into non-overlapping windows of size , while are unfolded to overlapping windows of size . It is calculated as

我们介绍了OCAB，用以直接建立跨窗口连接并增强窗口自注意力的表示能力。我们的OCAB包括一个重叠交叉注意（OCA）层和一个类似于标准Swin Transformer块[39]的MLP层。但对于OCA，如图5所示，我们使用不同的窗口大小来划分投影特征。具体来说，对于输入特征 的 被划分为 个大小为 的非重叠窗口，而 被展开为 个大小为 的重叠窗口。其计算方式如下：

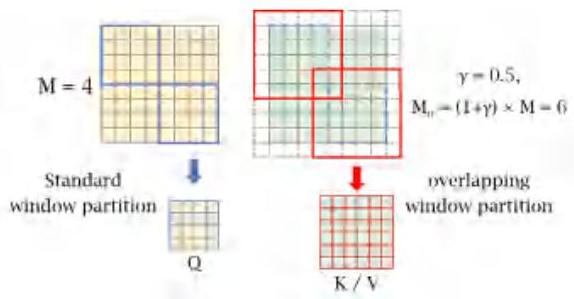


Figure 5. The overlapping window partition for OCA.

图5. OCA的重叠窗口划分。

where is a constant to control the overlapping size. To better understand this operation, the standard window partition can be considered as a sliding partition with the kernel size and the stride both equal to the window size . In contrast, the overlapping window partition can be viewed as a sliding partition with the kernel size equal to , while the stride is equal to . Zero-padding with size is used to ensure the size consistency of overlapping windows. The attention matrix is calculated as Equ. 2, and the relative position bias is also adopted. Unlike WSA whose query, key and value are calculated from the same window feature, OCA computes key/value from a larger field where more useful information can be utilized for the query. Note that although Multi-resolution Overlapped Attention (MOA) module in [44] performs similar overlapping window partition, our OCA is fundamentally different from MOA, since MOA calculates global attention using window features as tokens while OCA computes cross-attention inside each window feature using pixel token.

其中 是一个用于控制重叠大小的常数。为了更好地理解这个操作，可以将标准窗口划分视为核大小和步长均等于窗口大小的滑动划分。相比之下，重叠窗口划分可以看作是核大小等于 ，而步长等于 的滑动划分。使用大小为 的零填充以确保重叠窗口的大小一致性。注意力矩阵按等式2计算，并且也采用了相对位置偏差 。与WSA从相同的窗口特征计算查询、键和值不同，OCA从更大的字段中计算键/值，其中可以用于查询的更多信息。请注意，尽管文献[44]中的多分辨率重叠注意力（MOA）模块执行了类似的重叠窗口划分，但我们的OCA与MOA在本质上不同，因为MOA使用窗口特征作为标记来计算全局注意力，而OCA在每个窗口特征内部使用像素标记来计算交叉注意力。

# 3.3. The Same-task Pre-training

# 3.3. 同任务预训练

Pre-training is proven effective on many high-level vision tasks . Recent works also demonstrate that pre-training is beneficial to low-level vision tasks. IPT [6] emphasizes the use of various low-level tasks, such as denoising, deraining, super-resolution and etc., while EDT [27] utilizes different degradation levels of a specific task to do pre-training. These works focus on investigating the effect of multi-task pre-training for a target task. In contrast, we directly perform pre-training on a larger-scale dataset (i.e., ImageNet [9]) based on the same task, showing that the effectiveness of pre-training depends more on the scale and diversity of data. For example, when we want to train a model for , we first train a model on ImageNet, then fine-tune it on the specific dataset, such as DF2K. The proposed strategy, namely same-task pretraining, is simpler while bringing more performance improvements. It is worth mentioning that sufficient training iterations for pre-training and an appropriate small learning rate for fine-tuning are very important for the effectiveness of the pre-training strategy. We think that it is because Transformer requires more data and iterations to learn general knowledge for the task, but needs a small learning rate for fine-tuning to avoid overfitting to the specific dataset.

预训练在许多高级视觉任务上已被证明是有效的 。近期的工作 也表明预训练对低级视觉任务有益。IPT [6] 强调了使用各种低级任务，如去噪、去雨、超分辨率等，而 EDT [27] 则利用特定任务的不同的退化级别来进行预训练。这些工作专注于研究多任务预训练对目标任务的影响。相比之下，我们直接在更大规模的数据库（即 ImageNet [9]）上基于同一任务进行预训练，表明预训练的有效性更多地依赖于数据的规模和多样性。例如，当我们想要训练一个用于 的模型时，我们首先在 ImageNet 上训练一个 模型，然后在其特定数据集上进行微调，如 DF2K。所提出的策略，即同任务预训练，更简单且带来了更多的性能提升。值得一提的是，预训练的足够训练迭代次数和微调的适当小学习率对于预训练策略的有效性非常重要。我们认为这是因为 Transformer 需要更多的数据和迭代来学习任务的通用知识，但在微调时需要一个小的学习率以避免对特定数据集过拟合。

# 4. Experiments

# 4. 实验

# 4.1. Experimental Setup

# 4.1. 实验设置

We use DF2K (DIV2K [33]+Flicker2K [49]) dataset as the training dataset, since we find that using only DIV2K will lead to overfitting. When utilizing pre-training, we adopt ImageNet [9] following [6, 27]. For the structure of HAT, we keep the depth and width the same as SwinIR. Specifically, the RHAG number and HAB number are both set to 6 . The channel number is set to 180 . The attention head number and window size are set to 6 and 16 for both (S)W-MSA and OCA. For the hyper-parameters of the proposed modules, we set the weighting factor in HAB , the squeeze factor between two convolutions in CAB , and the overlapping ratio of OCA as 0.01,3 and 0.5 . For the large variant HAT-L, we directly double the depth of HAT by increasing the RHAG number from 6 to 12 . We also provide a small version HAT-S with fewer parameters and similar computation to SwinIR. In HAT-S, the channel number is set to 144 and the depth-wise convolution is used in CAB. Five benchmark datasets including Set5 [2], Set14 [66], BSD100 [40], Urban100 [19] and Manga109 [41] are used to evaluate the methods. For the quantitative metrics, PSNR and SSIM (calculated on the channel) are reported. More training details can refer to the supp. file.

我们使用DF2K（DIV2K [33]+Flicker2K [49]）数据集作为训练数据集，因为我们发现仅使用DIV2K会导致过拟合。在采用预训练时，我们遵循[6, 27]的方法，使用ImageNet [9]进行预训练。对于HAT的结构，我们保持与SwinIR相同的深度和宽度。具体来说，RHAG的数量和HAB的数量都设置为6。通道数设置为180。对于(S)W-MSA和OCA，注意力头的数量和窗口大小都设置为6和16。对于所提出模块的超参数，我们设置了HAB中的权重因子 ，CAB中两个卷积之间的squeeze因子 ，以及OCA的重叠比例 分别为0.01、3和0.5。对于大型变体HAT-L，我们通过将RHAG的数量从6增加到12，直接将HAT的深度加倍。我们还提供了一个小型版本HAT-S，参数更少，计算量与SwinIR相似。在HAT-S中，通道数设置为144，并在CAB中使用深度可分离卷积。使用包括Set5 [2]、Set14 [66]、BSD100 [40]、Urban100 [19]和Manga109 [41]在内的五个基准数据集来评估方法。对于定量指标，报告了PSNR和SSIM（在 通道上计算）。更多训练细节可参考附录文件。

Table 1. Quantitative comparison on PSNR(dB) of different window sizes.

表1. 不同窗口大小的PSNR（分贝）定量比较。

| Size | Set5 | Set14 | BSD100 | Urban100 | Manga109 |
| --- | --- | --- | --- | --- | --- |
| (8,8) | 32.88 | 29.09 | 27.92 | 27.45 | 32.03 |
| (16,16) | 32.97 | 29.12 | 27.95 | 27.81 | 32.15 |

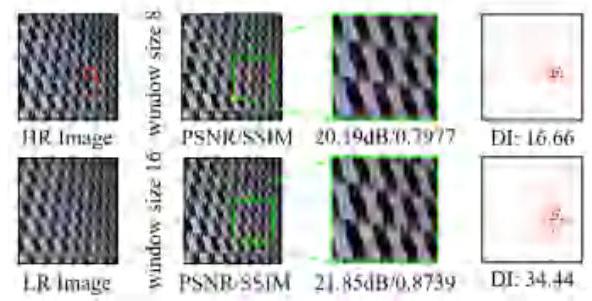


Figure 6. Qualitative comparison of different window sizes.

图6. 不同窗口大小的定性比较。

# 4.2. Effects of different window sizes

# 4.2. 不同窗口大小的影响

As discussed in Sec. 3.1, activating more input pixels for SR tends to achieve better performance. Enlarging window size for the window-based self-attention is an intuitive way to realize the goal. In [27], the authors investigate the effects of different window sizes. However, they conduct experiments based on the shifted cross local attention and only explore the window size up to . We further explore how the window size of self-attention influences the representation ability. To eliminate the influence of our newly-introduced blocks, we conduct the following experiments directly on SwinIR. As shown in Tab. 1, the model with a large window size of obtains better performance, especially on the Urban100. We also provide the qualitative comparison in Fig. 6. For the red marked patch, the model with window size of 16 utilizes much more input pixels than the model with window size of 8 . The quantitative performance of the reconstructed results also demonstrates the effectiveness of large window size. Based on this conclusion, we directly use window size 16 as our default setting.

如第3.1节所述，为超分辨率激活更多输入像素倾向于实现更好的性能。对于基于窗口的自注意力，增大窗口尺寸是一种直观的实现目标的方式。在文献[27]中，作者研究了不同窗口尺寸的影响。然而，他们基于移位交叉局部注意力进行实验，并且仅探索到 的窗口尺寸。我们进一步探讨自注意力的窗口尺寸如何影响表征能力。为了消除我们新引入块的影响，我们在SwinIR上直接进行以下实验。如表1所示，具有较大窗口尺寸 的模型在Urban100上获得了更好的性能。我们还在图6中提供了定性比较。对于红色标记的块，窗口尺寸为16的模型比窗口尺寸为8的模型利用了更多的输入像素。重建结果的定量性能也证明了较大窗口尺寸的有效性。基于这一结论，我们直接将窗口尺寸16作为默认设置。

# 4.3. Ablation Study

# 4.3. 剥离研究

Effectiveness of OCAB and CAB. We conduct experiments to demonstrate the effectiveness of the proposed CAB and OCAB. The quantitative performance reported on the Urban100 dataset for is shown in Tab. 2. Compared with the baseline results, both OCAB and CAB bring the performance gain of . Benefiting from the

OCAB和CAB的有效性。我们进行了实验以证明所提出CAB和OCAB的有效性。在Urban100数据集上报告的 的定量性能如表2所示。与基线结果相比，OCAB和CAB都带来了 的性能提升。得益于

Table 2. Ablation study on the proposed OCAB and CAB.

表2. 对所提出OCAB和CAB的剥离研究。

|  | Baseline | | | |
| --- | --- | --- | --- | --- |
| OCAB | X |  | X |  |
| CAB | X | X |  |  |
| PSNR |  |  |  |  |

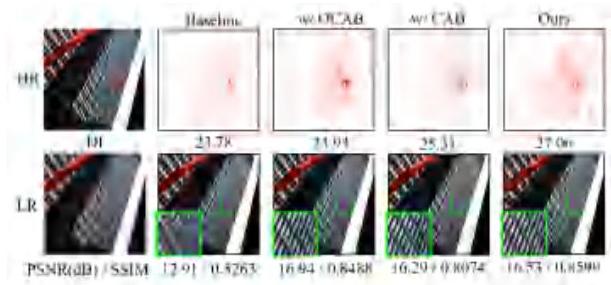


Figure 7. Ablation study on the proposed OCAB and CAB.

图7. 对所提出OCAB和CAB的剥离研究。

two modules, the model obtains a further performance improvement of . We also provide qualitative comparison to further illustrate the influence of OCAB and CAB, as presented in Fig. 7. We can observe that the model with OCAB has a larger scope of the utilized pixels and generate better-reconstructed results. When CAB is adopted, the used pixels even expand to almost the full image. Moreover, the result of our method with OCAB and CAB obtains the highest DI [15], which means our method utilizes the most input pixels. Although it obtains a little lower performance than w/OCAB, our method gets the highest SSIM and reconstructs the clearest textures.

两个模块，模型在 上获得了进一步的性能提升。我们还提供了定性的比较，以进一步说明OCAB和CAB的影响，如图7所示。我们可以观察到，使用OCAB的模型利用的像素范围更广，并生成了更好的重建结果。当采用CAB时，使用的像素甚至扩展到几乎整个图像。此外，我们的方法使用OCAB和CAB的结果获得了最高的DI [15]，这意味着我们的方法利用了最多的输入像素。尽管它的性能略低于使用OCAB，但我们的方法获得了最高的SSIM，并重建了最清晰的纹理。

Effects of different designs of CAB. We conduct experiments to explore the effects of different designs of CAB. First, we investigate the influence of channel attention. As shown in Tab. 3, the model using CA achieves a performance gain of compared to the model without CA. It demonstrates the effectiveness of the channel attention in our network. We also conduct experiments to explore the effects of the weighting factor of CAB. As presented in the manuscript Sec. 3.2.2, is used to control the weight of CAB features for feature fusion. A larger means a larger weight of features extracted by CAB and represents CAB is not used. As shown in Tab. 4, the model with of 0.01 obtains the best performance. It indicates that CAB and self-attention may have potential issue in optimization, while a small weighting factor for the CAB branch can suppress this issue for the better combination.

CAB不同设计的效应。我们进行了实验来探索CAB不同设计的影响。首先，我们研究了通道注意力的作用。如表3所示，使用CA的模型与没有CA的模型相比，性能提高了 。这证明了通道注意力在我们的网络中的有效性。我们还进行了实验来探索CAB权重因子 的影响。如手稿第3.2.2节所示， 用于控制CAB特征在特征融合中的权重。一个较大的 意味着CAB提取的特征权重较大， 表示没有使用CAB。如表4所示， 为0.01的模型获得了最佳性能。这表明CAB和自注意力在优化上可能存在潜在问题，而CAB分支的小权重因子可以抑制这个问题，以实现更好的组合。

Effects of the overlapping ratio. In OCAB, we set a constant to control the overlapping size for the overlapping cross-attention. To explore the effects of different overlapping ratios, we set a group of from 0 to 0.75 to examine the performance change, as shown in Tab. 5. Note that means a standard Transformer block. It can be found that the model with performs best. In contrast, when is set to 0.25 or 0.75, the model has no obvious performance gain or even has a performance drop. It illustrates that inappropriate overlapping size cannot benefit the interaction of neighboring windows.

重叠比的影响。在OCAB中，我们设置了一个常数 来控制重叠注意力的重叠大小。为了探索不同重叠比的影响，我们设置了一组从0到0.75的 来检查性能变化，如表5所示。注意 表示标准Transformer块。可以发现，带有 的模型表现最佳。相比之下，当 设置为0.25或0.75时，模型没有明显的性能提升，甚至性能下降。这说明不适当的重叠大小不能促进相邻窗口的交互。

Table 3. Effects of the channel attention (CA) module in CAB.

表3。CAB中通道注意力（CA）模块的影响。

| Structure | w/o CA | w/ CA |
| --- | --- | --- |
| PSNR / SSIM | 27.92dB / 0.8362 | 27.97dB / 0.8367 |

Table 4. Effects of the weighting factor in CAB.

表4。CAB中加权因子 的影响。

|  | 0 | 1 | 0.1 | 0.01 |
| --- | --- | --- | --- | --- |
| PSNR |  |  |  |  |

Table 5. Ablation study on the overlapping ratio of OCAB.

表5。OCAB的重叠比消融研究。

|  | 0 | 0.25 | 0.5 | 0.75 |
| --- | --- | --- | --- | --- |
| PSNR |  |  |  |  |

# 4.4. Comparison with State-of-the-Art Methods

# 4.4. 与现有先进方法的比较

Quantitative results. Tab. 6 shows the quantitative comparison of our approach and the state-of-the-art methods: EDSR [32], RCAN [68], SAN [8], IGNN [72], HAN [43], NLSN [42], RCAN-it [34], as well as approaches using ImageNet pre-training, i.e., IPT [6] and EDT [27]. We can see that our method outperforms the other methods significantly on all benchmark datasets. Concretely, HAT surpasses SwinIR by on Urban100 and on Manga109. When compared with the approaches using pre-training, HAT also has large performance gains of more than against EDT on Urban100 for all three scales. Besides, HAT with pre-training outperforms SwinIR by a huge margin of up to on Urban100 for SR. Moreover, the large model HAT-L can even bring further improvement and greatly expands the performance upper bound of this task. HAT-S with fewer parameters and similar computation can also significantly outperforms the state-of-the-art method SwinIR. (Detailed computational complexity comparison can be found in the supp. file.) Note that the performance gaps are much larger on Urban100, as it contains more structured and self-repeated patterns that can provide more useful pixels for reconstruction when the utilized range of information is enlarged. All these results show the effectiveness of our method.

定量结果。表6展示了我们的方法与现有先进方法的定量比较：EDSR [32]，RCAN [68]，SAN [8]，IGNN [72]，HAN [43]，NLSN [42]，RCAN-it [34]，以及使用ImageNet预训练的方法，即IPT [6]和EDT [27]。我们可以看到，我们的方法在所有基准数据集上都显著优于其他方法。具体来说，HAT在Urban100上超过SwinIR ，在Manga109上超过 。与使用预训练的方法相比，HAT在Urban100上对所有三个尺度的EDT也有超过 的性能提升。此外，经过预训练的HAT在Urban100上的 超分辨率中，以高达 的巨大差距超过SwinIR。而且，大型模型HAT-L甚至可以带来进一步的改进，极大地扩展了这项任务的性能上限。参数更少、计算量相似的HAT-S也可以显著超过现有最佳方法SwinIR。（详细的计算复杂度比较可以在补充文件中找到。）值得注意的是，在Urban100上的性能差距更大，因为它包含更多结构化和自我重复的模式，当利用的信息范围增大时，可以为重建提供更多有用的像素。所有这些结果都显示了我们的方法的有效性。

Visual comparison. We provide the visual comparison in Fig. 8. For the images "img\_002", "img\_011", "img\_030", "img\_044" and "img\_073" in Urban100, HAT successfully recovers the clear lattice content. In contrast, the other approaches all suffer from severe blurry effects. We can also observe similar behaviors on "PrayerHaNemurenai" in Manga109. When recovering the characters, HAT obtains significantly clearer textures than other methods. The visual results also demonstrate the superiority of our approach.

视觉比较。我们在图8中提供了视觉比较。在Urban100数据集中的图像 "img\_002"、"img\_011"、"img\_030"、"img\_044" 和 "img\_073" 中，HAT成功恢复了清晰的格子内容。相比之下，其他方法都存在严重的模糊效果。我们还可以在Manga109中的 "PrayerHaNemurenai" 上观察到类似的行为。在恢复字符时，HAT比其他方法获得了明显更清晰的纹理。视觉结果也证明了我们方法的优势。

# 4.5. Study on the pre-training strategy

# 4.5. 预训练策略研究

In Tab. 6, we can see that HAT can benefit greatly from the pre-training strategy, by comparing the performance of HAT and . To show the superiority of the proposed same-task pre-training, we also apply the multi-related-task pre-training [27] to HAT for comparison using full Ima-geNet, under the same training settings as [27]. As depicted as Tab. 7, the same-task pre-training performs better, not only in the pre-training stage but also in the fine-tuning process. From this perspective, multi-task pre-training probably impairs the restoration performance of the network on a specific degradation, while the same-task pre-training can maximize the performance gain brought by large-scale data. To further investigate the influences of our pre-training strategy for different networks, we apply our pre-training to four networks: SRResNet (1.5M), RRDBNet (16.7M), SwinIR (11.9M) and HAT (20.8M), as shown in Fig. 9. First, we can see that all four networks can benefit from pretraining, showing the effectiveness of the proposed same-task pre-training strategy. Second, for the same type of network (i.e., CNN or Transformer), the larger the network capacity, the more performance gain from pre-training. Third, although with less parameters, SwinIR obtains greater performance improvement from the pre-training compared to RRDBNet. It suggests that Transformer needs more data to exploit the potential of the model. Finally, HAT obtains the largest gain from pre-training, indicating the necessity of the pre-training strategy for such large models. Equipped with big models and large-scale data, we show the performance upper bound of this task is significantly extended.

在表6中，我们可以看到HAT从预训练策略中受益匪浅，通过比较HAT和 的性能。为了展示所提出同类任务预训练的优势，我们还将多相关任务预训练 [27] 应用于HAT进行比较，使用完整的ImageNet，并采用与 [27] 相同的训练设置。如表7所示，同类任务预训练不仅在预训练阶段表现更好，而且在微调过程中也是如此。从这个角度看，多任务预训练可能会损害网络在特定退化上的恢复性能，而同类任务预训练可以最大化大规模数据带来的性能增益。为了进一步研究我们的预训练策略对不同网络的影响，我们将预训练应用于四种网络：SRResNet（1.5M）、RRDBNet（16.7M）、SwinIR（11.9M）和HAT（20.8M），如图9所示。首先，我们可以看到所有四种网络都能从预训练中受益，显示了所提出同类任务预训练策略的有效性。其次，对于同一类型的网络（即CNN或Transformer），网络容量越大，从预训练中获得性能增益越多。第三，尽管参数较少，但SwinIR从预训练中获得的性能提升比RRDBNet更大。这表明Transformer需要更多的数据来挖掘模型的潜力。最后，HAT从预训练中获得了最大的增益，表明对于这样的大型模型，预训练策略的必要性。借助大型模型和大规模数据，我们展示了这个任务的性能上限得到了显著扩展。

Table 6. Quantitative comparison with state-of-the-art methods on benchmark datasets. The top three results are marked in red, blue and green. " " indicates that methods adopt pre-training strategy on ImageNet.

表6. 与最先进方法在基准数据集上的定量比较。前三个结果分别用红色、蓝色和绿色标记。" "表示方法在ImageNet上采用了预训练策略。

| Method | Scale | Training Dataset | Set5 | | Set14 | | BSD100 | | Urban100 | | Manga109 | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| EDSR |  | DIV2K | 38.11 | 0.9602 | 33.92 | 0.9195 | 32.32 | 0.9013 | 32.93 | 0.9351 | 39.10 | 0.9773 |
| RCAN |  | DIV2K | 38.27 | 0.9614 | 34.12 | 0.9216 | 32.41 | 0.9027 | 33.34 | 0.9384 | 39.44 | 0.9786 |
| SAN |  | DIV2K | 38.31 | 0.9620 | 34.07 | 0.9213 | 32.42 | 0.9028 | 33.10 | 0.9370 | 39.32 | 0.9792 |
| IGNN |  | DIV2K | 38.24 | 0.9613 | 34.07 | 0.9217 | 32.41 | 0.9025 | 33.23 | 0.9383 | 39.35 | 0.9786 |
| HAN |  | DIV2K | 38.27 | 0.9614 | 34.16 | 0.9217 | 32.41 | 0.9027 | 33.35 | 0.9385 | 39.46 | 0.9785 |
| NLSN |  | DIV2K | 38.34 | 0.9618 | 34.08 | 0.9231 | 32.43 | 0.9027 | 33.42 | 0.9394 | 39.59 | 0.9789 |
| RCAN-it |  | DF2K | 38.37 | 0.9620 | 34.49 | 0.9250 | 32.48 | 0.9034 | 33.62 | 0.9410 | 39.88 | 0.9799 |
| SwinIR |  | DF2K | 38.42 | 0.9623 | 34.46 | 0.9250 | 32.53 | 0.9041 | 33.81 | 0.9427 | 39.92 | 0.9797 |
| EDT |  | DF2K | 38.45 | 0.9624 | 34.57 | 0.9258 | 32.52 | 0.9041 | 33.80 | 0.9425 | 39.93 | 0.9800 |
| HAT-S (ours) |  | DF2K | 38.58 | 0.9628 | 34.70 | 0.9261 | 32.59 | 0.9050 | 34.31 | 0.9459 | 40.14 | 0.9805 |
| HAT (ours) |  | DF2K | 38.63 | 0.9630 | 34.86 | 0.9274 | 32.62 | 0.9053 | 34.45 | 0.9466 | 40.26 | 0.9809 |
|  |  | ImageNet | 38.37 |  | 34.43 | - | 32.48 | - | 33.76 |  |  | - |
|  |  | DF2K | 38.63 | 0.9632 | 34.80 | 0.9273 | 32.62 | 0.9052 | 34.27 | 0.9456 | 40.37 | 0.9811 |
| (ours) |  | DF2K | 38.73 | 0.9637 | 35.13 | 0.9282 | 32.69 | 0.9060 | 34.81 | 0.9489 | 40.71 | 0.9819 |
| - (ours) |  | DF2K | 38.91 | 0.9646 | 35.29 | 0.9293 | 32.74 | 0.9066 | 35.09 | 0.9505 | 41.01 | 0.9831 |
| EDSR |  | DIV2K | 34.65 | 0.9280 | 30.52 | 0.8462 | 29.25 | 0.8093 | 28.80 | 0.8653 | 34.17 | 0.9476 |
| RCAN |  | DIV2K | 34.74 | 0.9299 | 30.65 | 0.8482 | 29.32 | 0.8111 | 29.09 | 0.8702 | 34.44 | 0.9499 |
| SAN |  | DIV2K | 34.75 | 0.9300 | 30.59 | 0.8476 | 29.33 | 0.8112 | 28.93 | 0.8671 | 34.30 | 0.9494 |
| IGNN |  | DIV2K | 34.72 | 0.9298 | 30.66 | 0.8484 | 29.31 | 0.8105 | 29.03 | 0.8696 | 34.39 | 0.9496 |
| HAN |  | DIV2K | 34.75 | 0.9299 | 30.67 | 0.8483 | 29.32 | 0.8110 | 29.10 | 0.8705 | 34.48 | 0.9500 |
| NLSN |  | DIV2K | 34.85 | 0.9306 | 30.70 | 0.8485 | 29.34 | 0.8117 | 29.25 | 0.8726 | 34.57 | 0.9508 |
| RCAN-it |  | DF2K | 34.86 | 0.9308 | 30.76 | 0.8505 | 29.39 | 0.8125 | 29.38 | 0.8755 | 34.92 | 0.9520 |
| SwinIR |  | DF2K | 34.97 | 0.9318 | 30.93 | 0.8534 | 29.46 | 0.8145 | 29.75 | 0.8826 | 35.12 | 0.9537 |
| EDT |  | DF2K | 34.97 | 0.9316 | 30.89 | 0.8527 | 29.44 | 0.8142 | 29.72 | 0.8814 | 35.13 | 0.9534 |
| HAT-S (ours) |  | DF2K | 35.01 | 0.9325 | 31.05 | 0.8550 | 29.50 | 0.8158 | 30.15 | 0.8879 | 35.40 | 0.9547 |
| HAT (ours) |  | DF2K | 35.07 | 0.9329 | 31.08 | 0.8555 | 29.54 | 0.8167 | 30.23 | 0.8896 | 35.53 | 0.9552 |
|  |  | ImageNet | 34.81 | - | 30.85 | - | 29.38 | - | 29.49 | - | - | - |
|  |  | DF2K | 35.13 | 0.9328 | 31.09 | 0.8553 | 29.53 | 0.8165 | 30.07 | 0.8863 | 35.47 | 0.9550 |
| (ours) |  | DF2K | 35.16 | 0.9335 | 31.33 | 0.8576 | 29.59 | 0.8177 | 30.70 | 0.8949 | 35.84 | 0.9567 |
| HAT-L (ours) |  | DF2K | 35.28 | 0.9345 | 31.47 | 0.8584 | 29.63 | 0.8191 | 30.92 | 0.8981 | 36.02 | 0.9576 |
| EDSR |  | DIV2K | 32.46 | 0.8968 | 28.80 | 0.7876 | 27.71 | 0.7420 | 26.64 | 0.8033 | 31.02 | 0.9148 |
| RCAN |  | DIV2K | 32.63 | 0.9002 | 28.87 | 0.7889 | 27.77 | 0.7436 | 26.82 | 0.8087 | 31.22 | 0.9173 |
| SAN |  | DIV2K | 32.64 | 0.9003 | 28.92 | 0.7888 | 27.78 | 0.7436 | 26.79 | 0.8068 | 31.18 | 0.9169 |
| IGNN |  | DIV2K | 32.57 | 0.8998 | 28.85 | 0.7891 | 27.77 | 0.7434 | 26.84 | 0.8090 | 31.28 | 0.9182 |
| HAN |  | DIV2K | 32.64 | 0.9002 | 28.90 | 0.7890 | 27.80 | 0.7442 | 26.85 | 0.8094 | 31.42 | 0.9177 |
| NLSN |  | DIV2K | 32.59 | 0.9000 | 28.87 | 0.7891 | 27.78 | 0.7444 | 26.96 | 0.8109 | 31.27 | 0.9184 |
| RRDB |  | DF2K | 32.73 | 0.9011 | 28.99 | 0.7917 | 27.85 | 0.7455 | 27.03 | 0.8153 | 31.66 | 0.9196 |
| RCAN-it |  | DF2K | 32.69 | 0.9007 | 28.99 | 0.7922 | 27.87 | 0.7459 | 27.16 | 0.8168 | 31.78 | 0.9217 |
| SwinIR |  | DF2K | 32.92 | 0.9044 | 29.09 | 0.7950 | 27.92 | 0.7489 | 27.45 | 0.8254 | 32.03 | 0.9260 |
| EDT |  | DF2K | 32.82 | 0.9031 | 29.09 | 0.7939 | 27.91 | 0.7483 | 27.46 | 0.8246 | 32.05 | 0.9254 |
| HAT-S (ours) |  | DF2K | 32.92 | 0.9047 | 29.15 | 0.7958 | 27.97 | 0.7505 | 27.87 | 0.8346 | 32.35 | 0.9283 |
| HAT (ours) |  | DF2K | 33.04 | 0.9056 | 29.23 | 0.7973 | 28.00 | 0.7517 | 27.97 | 0.8368 | 32.48 | 0.9292 |
|  |  | ImageNet | 32.64 | - | 29.01 | - | 27.82 |  | 27.26 | - | - |  |
|  |  | DF2K | 33.06 | 0.9055 | 29.23 | 0.7971 | 27.99 | 0.7510 | 27.75 | 0.8317 | 32.39 | 0.9283 |
| HAT (ours) |  | DF2K | 33.18 | 0.9073 | 29.38 | 0.8001 | 28.05 | 0.7534 | 28.37 | 0.8447 | 32.87 | 0.9319 |
| HAT-L (ours) |  | DF2K | 33.30 | 0.9083 | 29.47 | 0.8015 | 28.09 | 0.7551 | 28.60 | 0.8498 | 33.09 | 0.9335 |

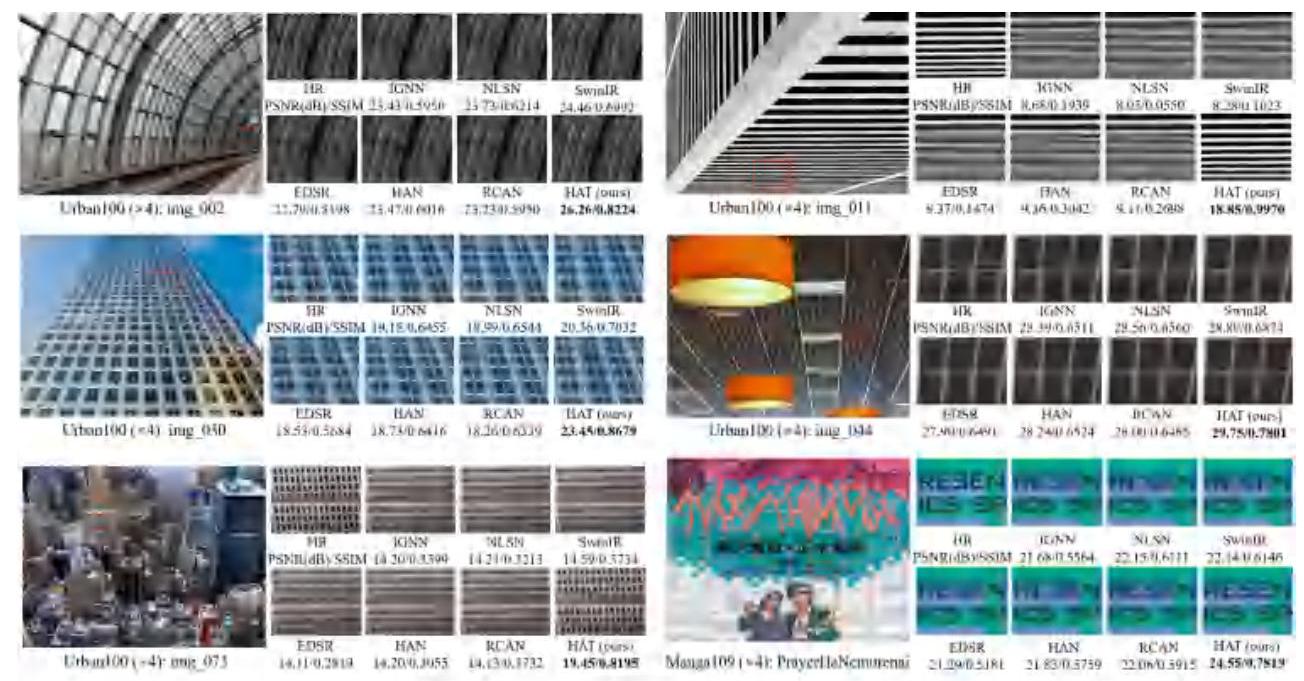


Figure 8. Visual comparison on . The patches for comparison are marked with red boxes in the original images. PSNR/SSIM is calculated based on the patches to better reflect the performance difference.

图8. 在 上的视觉比较。比较的斑块在原始图像中用红色框标记。基于斑块计算PSNR/SSIM，以更好地反映性能差异。

# 5. Conclusion

# 5. 结论

In this paper, we propose a novel Hybrid Attention Transformer, HAT, for single image super-resolution. Our model combines channel attention and self-attention to activate more pixels for high-resolution reconstruction. Besides, we propose an overlapping cross-attention module to enhance the interaction of cross-window information. Moreover, we introduce a same-task pre-training strategy to further exploit the potential of HAT. Extensive experiments show the effectiveness of the proposed modules and the pretraining strategy. Our approach significantly outperforms the state-of-the-art methods quantitatively and qualitatively.

在本文中，我们提出了一种新颖的混合注意力变换器，HAT，用于单幅图像超分辨率。我们的模型结合了通道注意力和自注意力，以激活更多像素用于高分辨率重建。此外，我们提出了一种重叠交叉注意力模块，以增强跨窗口信息的交互。此外，我们引入了一种同任务预训练策略，以进一步挖掘HAT的潜力。大量实验证明了所提出模块和预训练策略的有效性。我们的方法在定量和定性上均显著优于现有最先进方法。

Table 7. Quantitative results on of HAT using two kinds of pre-training strategies on under the same training setting. The full ImageNet dataset is adopted to perform pretraining and DF2K dataset is used for fine-tuning.

表7. 在相同训练设置下，HAT使用两种预训练策略在 上的定量结果。完整ImageNet数据集用于预训练，DF2K数据集用于微调。

| Strategy | Stage | Set5 | Set14 | Urban100 |
| --- | --- | --- | --- | --- |
| Multi-related-task pre-training | pre-training | 32.94 | 29.17 | 28.05 |
| fine-tuning | 33.06 | 29.33 | 28.21 |
| Same-task pre-training(ours) | pre-training | 33.02 | 29.20 | 28.11 |
| fine-tuning | 33.07 | 29.34 | 28.28 |

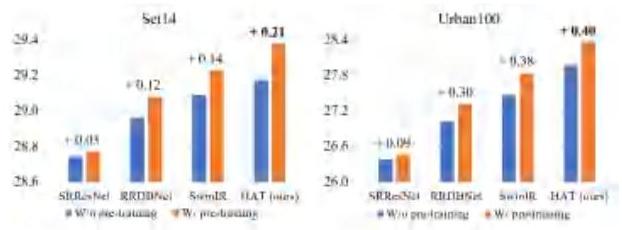


Figure 9. Quantitative comparison on PSNR(dB) of four different networks without and with the same-task pre-training on .

图9. 在 上，四种不同网络在有和无同任务预训练情况下的PSNR(dB)定量比较。

Acknowledgement. This work was supported in part by Macau Science and Technology Development Fund under SKLIOTSC-2021-2023, 0072/2020/AMJ, 0022/2022/A1; in part by Alibaba Innovative Research Program; in part by the National Natural Science Foundation of China under Grant (61971476, 62276251), the Joint Lab of CAS-HK; and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2020356).

致谢。这项工作得到了澳门科学技术发展基金的部分资助，项目编号为SKLIOTSC-2021-2023, 0072/2020/AMJ, 0022/2022/A1；部分得到了阿里巴巴创新研究计划的资助；部分得到了中国国家自然科学基金（编号61971476, 62276251）的资助，以及中国科学院-香港联合实验室的资助；还有部分得到了中国科学院青年创新促进会的资助（编号2020356）。

# References

# 参考文献

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021. 5

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

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xi-aopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021. 2

[4] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer, 2021. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213-229. Springer, 2020. 2

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12299-12310, 2021. 1, 2, 5, 6

[7] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haib-ing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems, 34, 2021. 2

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248-255, 2009. 5

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In European conference on computer vision, pages 184-199. Springer, 2014. 1, 2

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2):295-307, 2015. 1, 2

[12] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In European conference on computer vision, pages 391-407. Springer, 2016. 1, 2

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12124-12134, 2022. 2, 3

[14] 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: Transformers for image recognition at scale, 2020. 1, 2, 5

[15] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9199-9208, 2021. 1, 2, 3, 6

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16000- 16009, 2022. 5

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016. 4

[18] Gao Huang, Yulin Wang, Kangchen Lv, Haojun Jiang, Wen-hui Huang, Pengfei Qi, and Shiji Song. Glance and focus networks for dynamic visual recognition, 2022. 2

[19] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5197-5206, 2015. 5

[20] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer, 2021. 2, 3

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1646-1654, 2016. 2

[22] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1637-1645, 2016. 2

[23] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6002-6012, 2022. 2

[24] Xiangtao Kong, Hengyuan Zhao, Yu Qiao, and Chao Dong. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12016-12025, June 2021. 1

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681-4690, 2017. 2

[26] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2022. 2, 3

[27] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision, 2021. 1, 2, 5, 6, 7

[28] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers, 2021. 2

[29] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 833-843, June 2022. 1

[30] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer, 2022. 2

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1833-1844, 2021. 1, 2, 3

[32] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 136-144, 2017. 1, 2, 3, 6

[33] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 136-144, 2017. 5

[34] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. Revisiting rcan: Improved training for image super-resolution, 2022. 6

[35] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. Advances in neural information processing systems, 31, 2018. 2

[36] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. International journal of computer vision, 128(2):261-318, 2020. 2

[37] Yihao Liu, Anran Liu, Jinjin Gu, Zhipeng Zhang, Wenhao Wu, Yu Qiao, and Chao Dong. Discovering" semantics" in super-resolution networks, 2021. 2

[38] Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of super-resolution networks. arXiv preprint arXiv:2205.07019, 2022. 2

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012-10022, 2021. 1, 2, 3, 4

[40] 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 Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, volume 2, pages 416-423. IEEE, 2001. 5

[41] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 76(20):21811-21838, 2017. 5

[42] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3517-3526, 2021. 2, 6

[43] 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 European conference on computer vision, pages 191-207. Springer, 2020. 2, 6

[44] Krushi Patel, Andres M Bur, Fengjun Li, and Guanghui Wang. Aggregating global features into local vision transformer, 2022. 2, 3, 5

[45] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems, 34, 2021. 2

[46] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Studying stand-alone self-attention in vision models. 2019. 2

[47] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1874-1883, 2016. 2, 3

[48] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3147-3155, 2017. 2

[49] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 114-125, 2017. 5

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347-10357. PMLR, 2021. 2

[51] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. CVPR, 2022. 2

[52] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12894-12904, 2021. 2

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 1, 2, 4

[54] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 568-578, 2021. 1, 2

[55] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1905-1914, 2021. 2

[56] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops, pages 0-0, 2018. 2

[57] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17683-17693, 2022. 1, 2

[58] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 2

[59] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22-31, 2021. 2, 3

[60] Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. Pale transformer: A general vision transformer backbone with pale-shaped attention. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2731-2739, 2022. 2, 3

[61] Tete Xiao, Piotr Dollar, Mannat Singh, Eric Mintun, Trevor Darrell, and Ross Girshick. Early convolutions help transformers see better. Advances in Neural Information Processing Systems, 34, 2021. 2, 3

[62] Liangbin Xie, Xintao Wang, Chao Dong, Zhongang Qi, and Ying Shan. Finding discriminative filters for specific degradations in blind super-resolution. Advances in Neural Information Processing Systems, 34, 2021. 2

[63] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Feng-wei , and Wu. Incorporating convolution designs into visual transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 579-588, 2021. 2, 3

[64] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. Advances in Neural Information Processing Systems, 34:7281-7293, 2021. 2

[65] Syed Waqas Zamir, Aditya Arora, Salman Khan, Mu-nawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image

restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5728- 5739, 2022. 1, 2

[66] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In International conference on curves and surfaces, pages 711-730. Springer, 2010. 5

[67] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3096- 3105, 2019. 2

[68] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European conference on computer vision (ECCV), pages 286-301, 2018. 1, 2, 3, 4, 6

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

[70] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2472-2481, 2018. 1, 2

[71] Yucheng Zhao, Guangting Wang, Chuanxin Tang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. A battle of network structures: An empirical study of cnn, transformer, and mlp, 2021. 3

[72] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. Advances in neural information processing systems, 33:3499-3509, 2020. 2, 6

# Activating More Pixels in Image Super-Resolution Transformer Supplementary Material

# 激活图像超分辨率变换器中的更多像素 补充材料

State Key Laboratory of Internet of Things for Smart City, University of Macau

智慧城市物联网国家重点实验室，澳门大学

Shenzhen Key Lab of Computer Vision and Pattern Recognition,

深圳计算机视觉与模式识别关键实验室，

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

深圳先进技术研究院，中国科学院

Shanghai Artificial Intelligence Laboratory ARC Lab, Tencent PCG

上海人工智能实验室 ARC 实验室，腾讯PCG

https://github.com/XPixelGroup/HAT

# A. Training Details

# A. 训练细节

We use DF2K (DIV2K+Flicker2K) with 3450 images as the training dataset when training from scratch. The low-resolution images are generated from the ground truth images by the "bicubic" down-sampling in MATLAB. We set the input patch size to and use random rotation and horizontally flipping for data augmentation. The mini-batch size is set to 32 and total training iterations are set to . The learning rate is initialized as and reduced by half at . For , we initialize the model with pre-trained weights and halve the iterations for each learning rate decay as well as total iterations. We adopt Adam optimizer with and to train the model. For the same-task pretraining, the full ImageNet dataset with 1.28 million images is first exploited to pre-train the model for iterations. The initial learning rate is also set to but reduced by half at . Then, we adopt DF2K dataset to fine-tune the pre-trained model. For fine-tuning, we set the initial learning rate to 1e-5 and halve it at for total training iterations.

当从头开始训练时，我们使用包含3450张图像的DF2K（DIV2K+Flicker2K）作为训练数据集。低分辨率图像是通过MATLAB中的“双三次”下采样从真实图像生成的。我们将输入补丁大小设置为 并使用随机旋转和水平翻转进行数据增强。最小批次大小设置为32，总训练迭代次数设置为 。学习率初始化为 并在 时减半。对于 ，我们使用预训练的 权重初始化模型，并将每个学习率衰减的迭代次数以及总迭代次数减半。我们采用带有 和 的Adam优化器来训练模型。对于同任务预训练，首先使用包含128万张图像的全ImageNet数据集进行模型预训练 迭代。初始学习率也设置为 但在 时减半。然后，我们采用DF2K数据集对预训练模型进行微调。对于微调，我们将初始学习率设置为1e-5，并在 时减半，总共进行 次训练迭代。

# B. Analysis of Model Complexity

# B. 模型复杂性分析

We conduct experiments to analyze the computational complexity of our method from three aspects: window size for calculation of self-attention, overlapping cross-attention block (OCAB) and channel attention block (CAB). We also compare our method with the Transformer-based method SwinIR. The SR performance on Urban 100 are reported and the number of Multiply-Add operations is counted at the input size of . Note that pre-training techniques (including pre-training) are NOT used for all the models in this section. The experimental setup is completely fair.

我们进行实验以从三个方面分析我们方法的计算复杂性：自注意力计算中的窗口大小、重叠交叉注意力块（OCAB）和通道注意力块（CAB）。我们还与基于Transformer的方法SwinIR进行了比较。 在Urban 100上的SR性能报告以及输入大小为 时的乘加操作数已统计。注意，本节中所有模型均未使用预训练技术（包括 预训练）。实验设置是完全公平的。

First, we use the standard Swin Transformer block as the backbone to explore the influence on different window sizes. As shown in Tab. 8, enlarging window size can bring a large performance gain with a little increase in parameters and increase in Multi-Adds.

首先，我们使用标准的Swin Transformer块作为主干网络，以探索不同窗口大小的影响。如表8所示，增大窗口大小可以带来较大的性能提升 ，同时参数数量略有增加和 乘加操作的增多。

Table 8. Model complexity comparison of window sizes.

表8。不同窗口大小的模型复杂性比较。

| window size | #Params. | #Multi-Adds. | PSNR |
| --- | --- | --- | --- |
| (8, 8) | 11.9M | 53.6G |  |
| (16, 16) | 12.1M | 63.8G |  |

Table 9. Model complexity comparison of OCAB and CAB.

表9。OCAB和CAB的模型复杂性比较。

| Method | #Params. | #Multi-Adds. | PSNR |
| --- | --- | --- | --- |
| Baseline | 12.1M | 63.8G |  |
| w/ OCAB | 13.7M | 74.7G |  |
| w/ CAB | 19.2M | 92.8G |  |
| Ours | 20.8M | 103.7G |  |

Table 10. Model complexity comparison of CAB sizes.

表10。不同CAB大小的模型复杂性比较。

| in CAB | #Params. | #Multi-Adds. | PSNR |
| --- | --- | --- | --- |
| 1 | 33.2M | 150.1G |  |
| 2 | 22.7M | 107.1G | 27.92dB |
| 3 (default) | 19.2M | 92.8G | 27.91dB |
| 6 | 15.7M | 78.5G |  |
| w/o CAB | 12.1M | 63.8G | 27.81dB |

Table 11. Model complexity comparison of SwinIR and HAT.

表11。SwinIR和HAT的模型复杂性比较。

| Method | #Params. | #Multi-Adds. | PSNR |
| --- | --- | --- | --- |
| SwinIR | 11.9M | 53.6G | 27.45dB |
| HAT-S (ours) | 9.6M | 54.9G |  |
| SwinIR-L1 | 24.0M | 104.4G | 27.53dB |
| SwinIR-L2 | 23.1M | 102.4G |  |
| HAT (ours) | 20.8M | 103.7G |  |

Then, we use window size 16 as the baseline to investigate the computational complexity of the proposed OCAB and CAB. As illustrated in Tab. 9, our OCAB obtains a performance gain with a limited increase of parameters and Multi-Adds. It demonstrates that the effectiveness and efficiency of the proposed OCAB. Besides, adding CAB to the baseline model also achieves better performance.

然后，我们以窗口大小16为基准，研究提出的OCAB和CAB的计算复杂性。如表9所示，我们的OCAB在参数和乘加操作数量有限增加的情况下获得了性能提升，这证明了所提出OCAB的有效性和效率。此外，将CAB添加到基准模型中也实现了更好的性能。

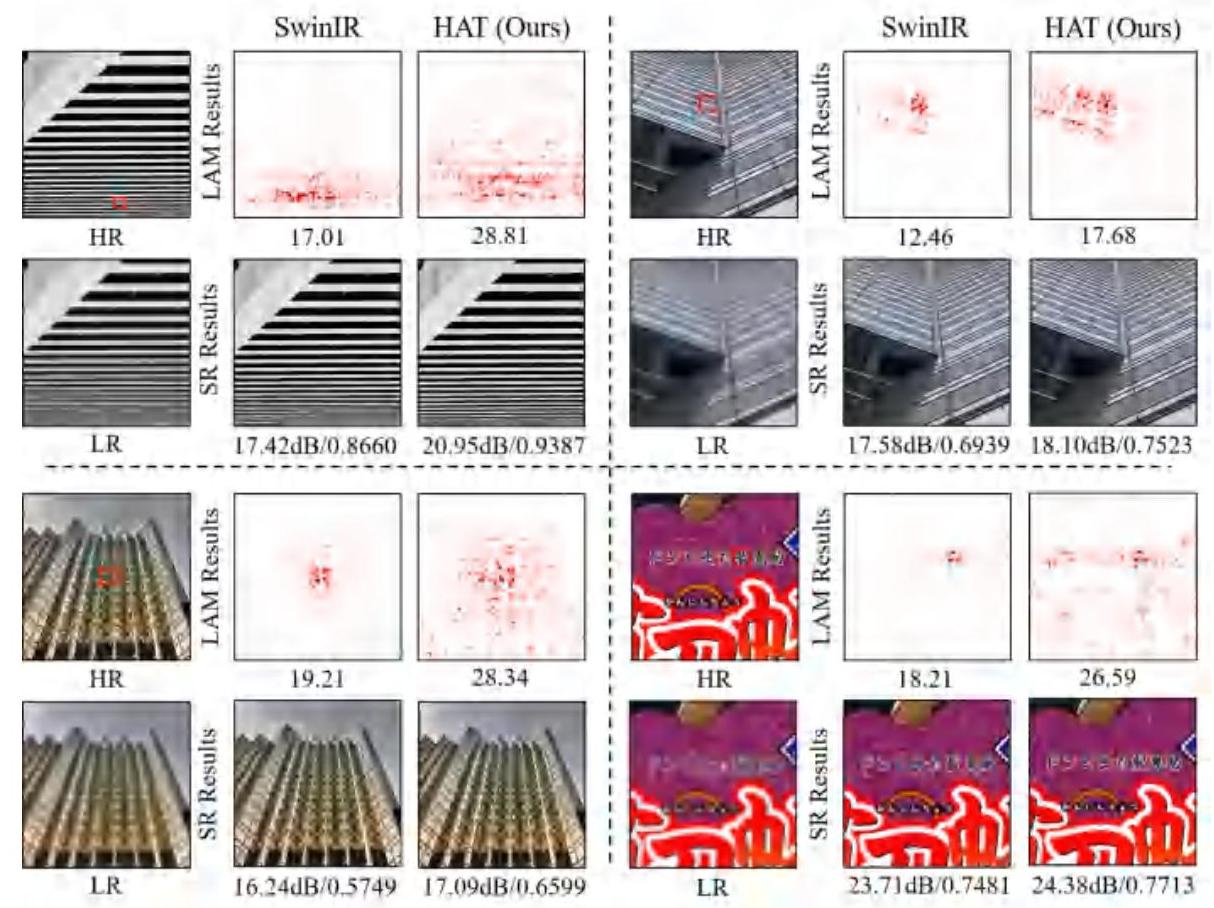


Figure 10. Comparison of LAM results between SwinIR and HAT.

图10。SwinIR和HAT的LAM结果比较。

Since CAB seems to be computationally expensive, we further explore the influence on CAB sizes by modulating the squeeze factor (mentioned in Sec.3.2.2 in the main paper). As shown in Tab. 10, adding a small CAB whose equals 6 can bring considerable performance improvement. When we continuously reduce , the performance increases but with larger model sizes. To balance the performance and computations, we set to 3 as the default setting.

由于 CAB 似乎计算成本较高，我们进一步探讨了通过调节挤压因子 （在主论文的第3.2.2节中提到）对 CAB 大小的影响。如表10所示，添加一个 等于6的小型 CAB 可以带来显著的性能提升。当我们连续减小 时，性能提高但模型大小增大。为了平衡性能和计算，我们将 设为3作为默认设置。

Furthermore, we compare HAT and SwinIR with the similar numbers of parameters and Multi-Adds in two settings, as presented in Tab. 11. 1) We compare HAT-S with the original version of SwinIR. With less parameters and comparable computations, HAT-S significantly outperforms SwinIR. 2) We enlarge SwinIR by increasing the width and depth of SwinIR to achieve similar computations as HAT, denoted as SwinIR-L1 and SwinIR-L2. HAT achieves the best performance at the lowest computational cost.

此外，我们在两种设置下，如表11所示，比较了具有相似参数数量和 Multi-Adds 的 HAT 和 SwinIR。1）我们比较了 HAT-S 与 SwinIR 的原始版本。在参数更少且计算相当的情况下，HAT-S 显著优于 SwinIR。2）我们通过增加 SwinIR 的宽度和深度来扩大 SwinIR，以达到与 HAT 相似的计算量，分别表示为 SwinIR-L1 和 SwinIR-L2。HAT 在最低的计算成本下实现了最佳性能。

Overall, we find that enlarging the window size for the calculation of self-attention is a very cost-effective way to improve the Transformer model. Moreover, the proposed OCAB can bring an obvious performance gain with limited increase of computations. Although CAB is not as efficient as above two schemes, it can also bring stable and considerable performance improvement. Benefiting from the three designs, HAT can substantially outperforms the state-of-the-art method SwinIR with comparable computations.

总的来说，我们发现增大计算自注意力时的窗口大小是提高 Transformer 模型性能的一种非常有效的方法。此外，提出的 OCAB 在计算量有限增加的情况下可以带来明显的性能提升。尽管 CAB 的效率不如上述两种方案，但它也可以带来稳定且显著的性能改进。得益于这三种设计，HAT 可以在计算相当的情况下显著优于最先进的方法 SwinIR。

# C. More Visual Comparisons with LAM

# C. 与 LAM 的更多视觉比较

We provide more visual comparisons with LAM results to compare SwinIR and our HAT. The red points in LAM results represent the used pixels for reconstructing the patch marked with a red box in the HR image, and Diffusion Index (DI) is computed to reflect the range of involved pixels. The more pixels are utilized to recover the specific input patch, the wider the distribution of red points is in LAM and the higher DI is. As shown in Fig. 10, the LAM attribution of HAT expands to the almost full image, while that of SwinIR only gathers in a limited range. For the quantitative metric, HAT also obtains a much higher DI value than SwinIR. All these results demonstrate that our method activates more pixels to reconstruct the low-resolution input image. As a result, SR results generated by our method have higher PSNR/SSIM and better visual quality.

我们提供了与 LAM 结果的更多视觉对比，以比较 SwinIR 和我们的 HAT。LAM 结果中的红色点代表用于重建 HR 图像中用红色框标记的贴图的像素，并且计算了扩散指数（DI）以反映涉及像素的范围。用于恢复特定输入贴图的像素越多，LAM 中红色点的分布就越广，DI 就越高。如图 10 所示，HAT 的 LAM 归因几乎扩展到整张图像，而 SwinIR 的归因仅集中在有限范围内。对于定量指标，HAT 也比 SwinIR 获得了更高的 DI 值。所有这些结果都表明，我们的方法激活了更多像素来重建低分辨率输入图像。因此，由我们的方法生成的 SR 结果具有更高的 PSNR/SSIM 和更好的视觉质量。