用于轻量级图像超分辨率的全聚合网络

摘要

轻量级 ViT 框架在图像超分辨率领域取得了巨大进展，但其单一维度的自注意力建模以及同质化的聚合方式限制了其有效感受野（ERF）的能力，使其难以从空间和通道维度中包含更全面的交互信息。为了解决这些缺陷，本文提出了一种新型的 Omni-SR 架构，并在其下设计了两个增强组件。首先，本文提出了基于密集交互原则的全方位自注意力（Omni Self-Attention, OSA）模块，该模块能够同时从空间和通道维度建模像素交互，挖掘跨全轴（即空间与通道）的潜在相关性。结合主流的窗口分区策略，OSA 能够在较低的计算预算下实现卓越性能。其次，提出了一种多尺度交互方案，以缓解浅层模型中次优的有效感受野（即过早饱和）问题。该方案促进了局部传播以及中/全局尺度的交互，形成了一种全尺度聚合构建模块（omni-scale aggregation building block）。大量实验表明，Omni-SR 在轻量级超分辨率基准测试中取得了创纪录的性能（例如，在 Urban100 ×4 数据集上实现了 26.95dB 的峰值信噪比，仅使用 792K 参数）。

关键词： 轻量级 ViT（Lightweight ViT）；全方位自注意力（Omni Self-Attention，OSA）；多尺度交互（Multi-Scale Interaction）

1 引言

随着深度学习在计算机视觉领域的迅猛发展，图像超分辨率（Image Super-Resolution, SR）技术成为了一个研究热点。超分辨率旨在从低分辨率图像恢复出高分辨率图像，广泛应用于医疗影像、卫星遥感和视频监控等领域。近年来，基于视觉变换器（ViT）的轻量级网络架构在图像超分辨率中取得了显著进展，主要得益于其强大的特征表达能力。然而，现有轻量级 ViT 框架的单一维度自注意力机制以及同质化的特征聚合方式限制了模型在空间和通道维度上的交互能力，进而限制了其有效感受野（ERF）。这些问题使得 ViT 框架在处理复杂图像细节和多尺度信息时存在不足，尤其在高质量图像重建任务中表现不佳。针对上述问题，本研究提出了 Omni-SR 架构，并设计了两个增强组件来弥补现有框架的不足。首先，提出了基于密集交互原则的全方位自注意力（Omni Self-Attention, OSA）模块，如图 1所示，能够同时从空间和通道维度建模像素交互，挖掘潜在的跨轴相关性。OSA 模块结合主流的窗口分区策略，能够在较低计算预算下实现卓越性能。其次，提出了多尺度交互方案，通过缓解浅层模型中的次优感受野（如过早饱和问题），促进局部传播和中全局尺度的交互，形成了一种全尺度聚合模块（omni-scale aggregation building block）。图像超分辨率技术在许多实际应用中至关重要，如医学影像的高清重建、遥感图像生成、视频监控中的图像增强等。然而，传统方法在处理多维度信息时存在局限。通过引入 OSA 模块和多尺度交互方案，本研究突破了这

些限制，提升了图像重建质量，同时保持了轻量级网络的计算效率。该方法不仅能够改善细节和图像质量，还能有效应对长时间序列和多尺度问题，具有广泛的应用前景，尤其在资源受限的设备上，如移动端和嵌入式系统，具有重要的实用价值。

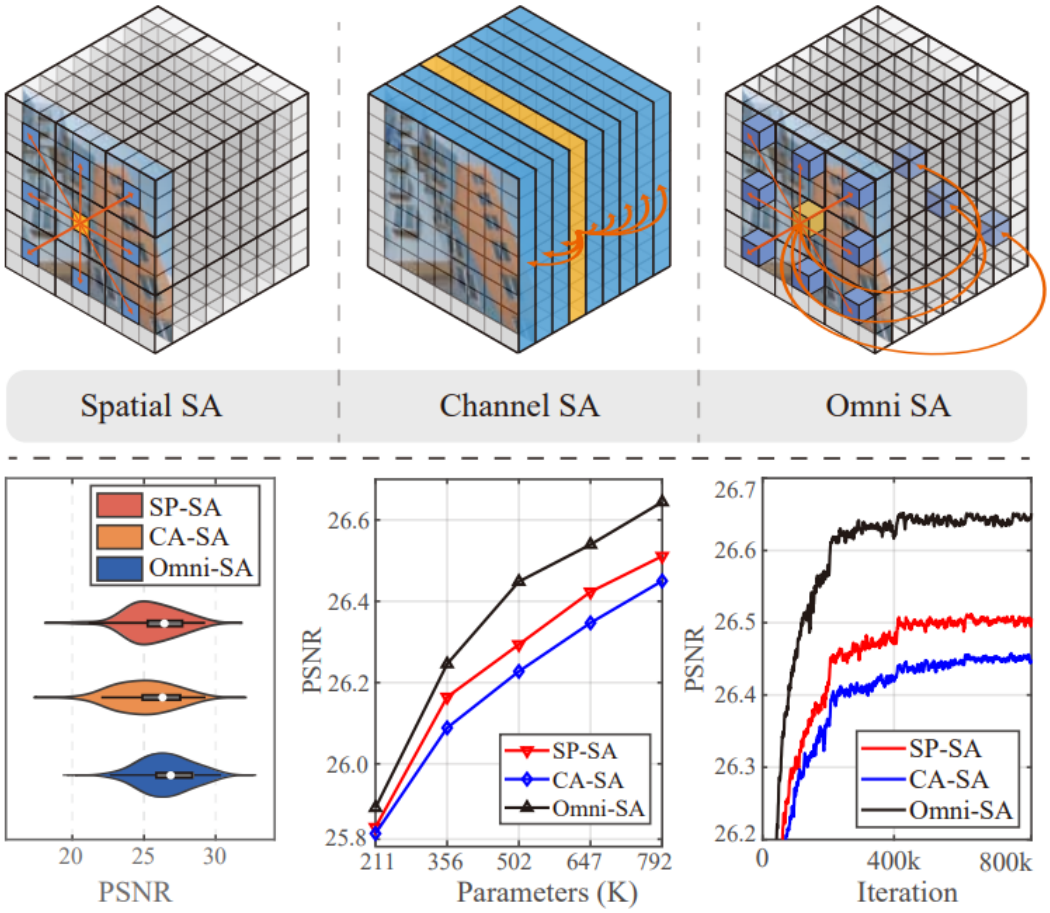

图 1. 典型的自注意力机制只能进行单一维度（例如，仅空间）的交互。

2 相关工作

轻量级视觉变换器。由于网络在资源受限设备上的应用需求日益迫切，轻量级视觉变换器吸引了广泛的关注。许多尝试已经开展，旨在开发性能相当的轻量级 ViT。许多方法专注于将卷积与变换器结合，以学习局部和全局表示。例如，LVT 在自注意力中引入了卷积，以丰富低级特征。MobileViT 用变换器层替代了卷积中的矩阵乘法，从而学习全局表示。同样，EdgeViTs 采用信息交换瓶颈以实现完全的空间交互。与将卷积解释为视觉变换器不同，LightViT 提出了聚合自注意力以实现更好的信息聚合。在本研究中，我们依赖于 ViT 架构来实现轻量级且精确的 SR。

2.1 图像超分辨率

卷积神经网络（CNN）在图像超分辨率（SR）任务中取得了显著的成功。SRCNN [2] 是第一个将 CNN 引入 SR 领域的研究。许多方法采用跳跃连接来加速网络收敛并提高重建质量。通道注意力也被提出以增强 SR 模型的表示能力。为了在有限的计算资源下获得更好的重

建质量，一些方法探索了轻量级架构设计。DRCN [7] 利用递归操作减少了参数数量。DRRN [15] 在 DRCN 基础上引入了全局和局部残差学习，以加速训练并提高细节质量。CARN [1] 在残差网络基础上采用了级联机制。IMDN [5] 提出了信息多重蒸馏块，以提高时间性能。另一研究方向是利用模型压缩技术，如知识蒸馏和神经架构搜索，来降低计算成本。最近，一系列基于变换器的 SR 模型出现，表现出优异的性能。Chen 等人开发了一种基于变换器架构的低级计算机视觉任务预训练模型。基于 Swin 变换器，SwinIR [10] 提出了一个三阶段框架，刷新了 SR 任务的最先进技术。最近，一些研究探索了 ImageNet 预训练策略，以进一步增强 SR 性能。

2.2 轻量级视觉变换器

由于网络在资源受限设备上的应用需求日益迫切，轻量级视觉变换器吸引了广泛的关注。许多尝试已经开展，旨在开发性能相当的轻量级 ViT。许多方法专注于将卷积与变换器结合，以学习局部和全局表示。例如，LVT [19] 在自注意力中引入了卷积，以丰富低级特征。MobileViT 用变换器层替代了卷积中的矩阵乘法，从而学习全局表示。同样，EdgeViTs 采用信息交换瓶颈以实现完全的空间交互。与将卷积解释为视觉变换器不同，LightViT 提出了聚合自注意力以实现更好的信息聚合。在本研究中，我们依赖于 ViT 架构来实现轻量级且精确的 SR。

3 本文方法

3.1 超分辨率中的注意力机制

在超分辨率（SR）任务中，广泛采用了两种注意力范式来帮助分析和聚合全面的模式。

空间注意力 空间注意力 [12] 可以被视为一种各向异性的选择过程。空间自注意力和空间门控是主要应用的两种方法。如图 2所示，空间自注意力沿空间维度计算交叉协方差，空间门控则生成通道分离的掩膜。它们都无法在通道之间传递信息。

通道注意力 通道注意力有两种类别，即基于标量 [4] 的和基于协方差 [20] 的，旨在执行通道重新校准或在通道之间传递模式。如图 2所示，前者预测一组重要性标量，用于加权不同的通道，而后者计算交叉协方差矩阵，以实现通道的重新加权和信息传递。与空间注意力相比，通道注意力在空间维度上是各向同性的，因此计算复杂度显著降低，但也会影响聚合的准确性。

一些研究表明 [13]，空间注意力和通道注意力对 SR 任务都是有益的，并且它们的特点是互补的，因此以计算上紧凑的方式将两者集成在一起，能够显著提高表达能力。

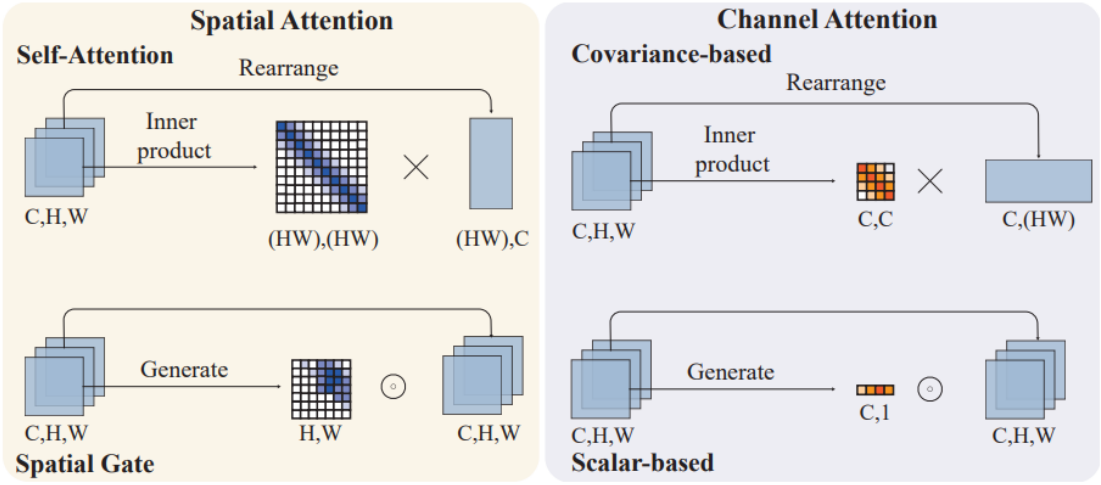

图 2. 空间注意力和通道注意力的示意图。这些典型的注意力范式仅对单一维度（即仅空间或仅通道）进行建模和交互。

3.2 全方位自注意力块

为了挖掘隐藏在潜在变量中的所有相关性，我们提出了一种新颖的自注意力范式——全方位自注意力块（Omni Self-Attention, OSA）。与现有的自注意力范式（如仅进行单维度处理的空间自注意力 [16]）不同，OSA 同时建立了空间和通道的上下文关联。这种双维度的关系建模对于轻量级模型尤为必要且有益。一方面，随着网络的加深，重要信息会分散到不同的通道中，因此需要及时处理这些信息；另一方面，尽管空间自注意力在计算协方差时利用了通道维度，但它无法在通道之间传递信息。基于这些条件，我们设计的 OSA 能够以紧凑的方式传递空间和通道信息。

OSA 的计算过程通过顺序矩阵运算和旋转操作，生成空间和通道方向的得分矩阵，如图 3所示。假设输入特征 $X \in \mathbb{R}^{HW \times C}$，其中 H 和 W 分别是输入的高度和宽度，C 是通道数。首先，将 X 投影到查询矩阵、键矩阵和值矩阵, 分别表示为 $Q^s, K^s, V^s \in \mathbb{R}^{HW \times C}$, ,。接着，通过计算查询与键的乘积得到大小为 $\mathbb{R}^{HW \times HW}$ 的空间注意力图，并基于此得到中间的聚合结果。通常使用窗口策略来显著降低计算资源消耗。

在下一阶段，将查询和键矩阵旋转以获得转置的查询和键矩阵 $Q^s, K^s, V^s \in \mathbb{R}^{HW \times C}$，同时旋转值矩阵以生成 $V^c \in \mathbb{R}^{C \times HW}$，用于后续的通道自注意力计算。生成的通道注意力图大小为 $\mathbb{R}^{C \times C}$，用于建模通道间的关系。最终，通过对通道注意力输出 Y^c 进行逆旋转，得到最终的聚合结果 Y_{OSA}。OSA 的整体流程如下公式表示：

$$Q^s = XW_q, K^s = XW_k, V^s = XW_v,$$

$$Y_s = A^s(Q^s, K^s, V^s) = Softmax(Q^s K^{sT})V^s,$$

$$Q^c = R(Q'), K^c = R(K'), V^c = R(V'),$$

$$Y_c == A^c(Q^c, K^c, V^c) = Softmax(K^c Q^{cT})V^c,$$

$$Y_{OSA} = R^{-1}(Y_c),$$

其中，W_q, W_k, W_v 分别表示查询、键和值的线性投影矩阵，$R(\cdot)$ 为空间维度上的旋转操作，$R^{-1}(\cdot)$ 为逆旋转。这种设计能够整合两种矩阵操作（空间与通道操作）的元素级结果，从

而实现全方位的交互。需要注意的是，OSA 可以作为 Swin [10] 注意力块的直接替代，使用更少的参数实现更高的性能。由于通道自注意力的注意力图尺寸较小，相比于 Swin 中的级联滑窗自注意力，OSA 的计算开销更低。

与其他混合注意力范式的对比，相比于之前的混合通道和空间注意力方法（如 CBAM [17] 和 BAM [13]），这些方法仅基于标量的注意力权重，只能反映相对重要性，无法实现像素间的信息交换，导致关系建模能力有限。一些近期研究结合了通道注意力与空间自注意力，但它们仅通过标量权重实现通道重新校准，而 OSA 则通过通道间交互挖掘全方位的潜在相关性。

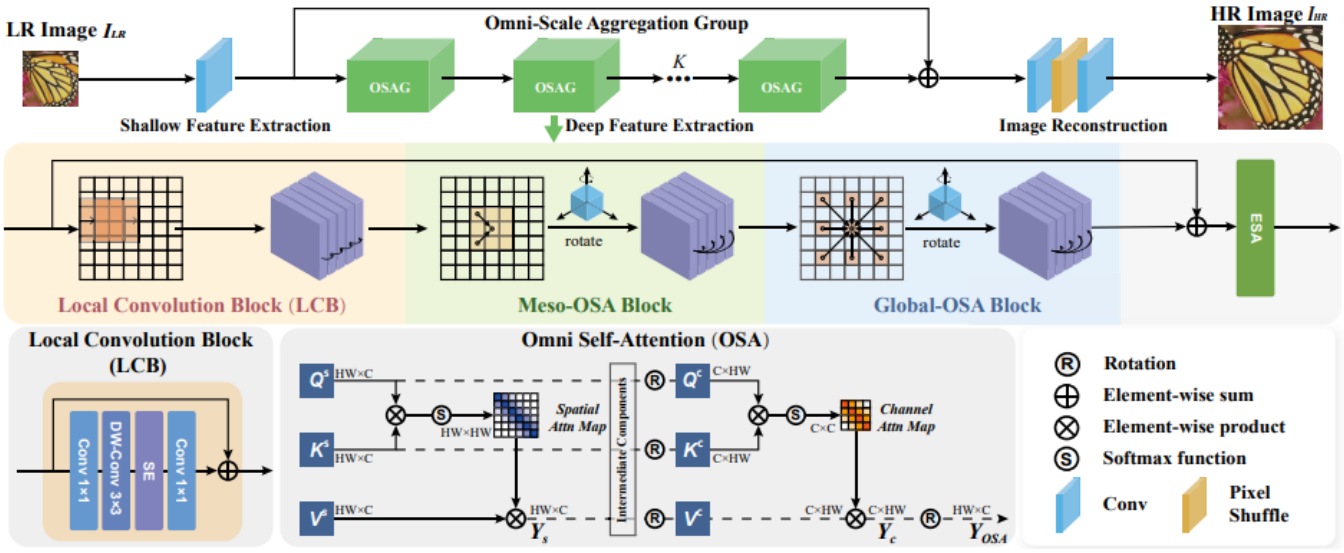

图 3. 所提议的 Omni-SR 框架的整体架构，以及 OSAG 和全方位自注意力（OSA）的结构。

3.3 全方位尺度聚合组

如何利用提议的 OSA 范式构建高性能且紧凑的网络是另一个关键问题。尽管基于窗口的自注意力层次堆叠（例如 Swin）已成为主流，但许多研究发现，对于大范围的交互，基于窗口的范式在计算上效率较低，尤其是在浅层网络中。值得指出的是，大范围的交互能够提供良好的有效感受野，这对于提高图像重建性能至关重要。不幸的是，直接的全局交互是计算资源消耗大的，并且会削弱局部聚合能力。基于这些考虑，我们提出了全方位尺度聚合组（Omni-Scale Aggregation Group，简称 OSAG），旨在以低计算复杂度实现渐进式感受野特征聚合。如图 3所示，OSAG 主要包括三个阶段：局部聚合、中尺度聚合和全局聚合。具体而言，引入了增强通道注意力的倒置瓶颈结构 [3] 来完成有限计算开销下的局部模式处理。基于提议的 OSA 范式，我们衍生出了两个实例（即 Meso-OSA 和 Global-OSA），分别负责中尺度和全局信息的交互与聚合。需要注意的是，提议的全方位自注意力范式可以用于不同的目的。Meso-OSA 在非重叠的块组内执行注意力机制，从而将 Meso-OSA 限制为仅关注中尺度模式的理解。Global-OSA 则通过空洞方式稀疏地在整个特征图上采样数据点，赋予 Global-OSA 在较低计算成本下实现全局交互的能力。

Meso-OSA 和 Global-OSA 的唯一区别在于窗口划分策略，如图 4所示。为了实现中尺度交互，Meso-OSA 将输入特征 X 划分为大小为 PxP 的非重叠块。值得注意的是，窗口划分后，块的维度会汇聚到空间维度（即 -2 轴）上：$(H, W, C) \rightarrow (\frac{H}{P} \times P, \frac{W}{P} \times P, C) \rightarrow (\frac{HW}{P^2}, P^2, C)$. 而 Global-OSA 则将输入特征划分为统一的 $G \times G$ 网格，每个格子具有适应性大小 $\frac{H}{G} \times \frac{H}{G}$。与

Meso-OSA 类似，网格维度也会汇聚到空间轴（即 -2 轴）上：$(H, W, C) \rightarrow (G \times \frac{H}{G}, G \times \frac{H}{G}, C) \rightarrow (G^2, \frac{HW}{G^2}, C) \rightarrow (\frac{HW}{G^2}, G^2, C)$。

图 4. 全方位尺度聚合方案的示意图。我们提出的 Omni-SR 包含三种特征聚合方式，分别在局部、中尺度和全局尺度进行聚合。

3.4 网络架构

总体结构 基于全方位自注意力范式和全方位尺度聚合组，我们进一步开发了一个轻量级的 Omni-SR 框架，以实现高性能的图像超分辨率。如图 3所示，Omni-SR 由三部分组成：浅层特征提取、深层特征提取和图像重建。具体而言，给定低分辨率输入 $I_{LR} \in \mathbb{R}^{H \times W \times C_m}$，我们首先使用一个 3×3 卷积层 H_{SF} 提取浅层特征 $X_0 \in \mathbb{R}^{H \times W \times C}$：

$$X_0 = H_{SF}(I_{LR})$$

其中，C_{in} 和 C 分别表示输入和浅层特征的通道数。卷积层提供了一种简单的方法将输入从图像空间转换为高维特征空间。然后，我们使用 K 个堆叠的全方位尺度聚合组（OSAG）和一个 3×3 卷积层 H_{CONV} 以级联的方式提取深层特征 F_{DF}。这样的过程可以表示为：

$$X_i = H_{OSAGi}(X_{i-1}), i = 1, 2, \ldots, K,$$

$$X_{DF} = H_{OSAG}(X_K),$$

其中，H_{OSAGi} 表示第 i 个 OSAG，$X_1, X_2, \ldots, X_K$ 表示中间特征。与 [10] 类似，我们还在特征提取的最后使用卷积层以获得更好的特征聚合。最终，我们通过将浅层和深层特征进行聚合，重建高分辨率图像 I_{HR}：

$$I_{HR} = H_{Rec}(X_0, F_{DF}),$$

其中，$H_{Rec}(\cdot)$ 表示重建模块。具体而言，PixelShuffle [14] 用于上采样融合后的特征。

全方位尺度聚合组（OSAG） 如图 3所示，每个 OSAG 包含一个局部卷积块（LCB）、一个中尺度 OSA 块（Meso-OSA）、一个全局 OSA 块（Global-OSA）和一个 ESA 块 [8]。整个过程可以表示为：

$$X_i = H_{OSAGi}(X_{i-1})$$

其中，X_{i-1} 和 X_i 分别表示第 i 个 OSAG 的输入和输出特征。在卷积层映射后，我们插入一个 Meso-OSA 块以实现基于窗口的自注意力，并使用 Global-OSA 块来扩大感受野，以便更好地聚合信息。OSAG 结束时，我们保留卷积层和 ESA 块，参考 [21]。

具体而言，LCB 通过点卷积和深度卷积的堆叠实现，并在两者之间插入一个通道注意力（CA）模块 [6] 来自适应地重新加权通道特征。该块旨在聚合局部上下文信息，并增加网络的可训练性 [18]。接着，两个 OSA 块（即 Meso-OSA 和 Global-OSA）用于从不同区域获取交

互。基于不同的窗口划分策略，Meso-OSA 块寻求块内交互，而 Global-OSA 块则寻求全局混合。OSA 块遵循典型的 Transformer 设计，包括前馈网络（FFN）和层归一化（LayerNorm）[9]，唯一的区别是原始自注意力操作被我们提出的 OSA 操作取代。对于 FFN，我们采用了 Restormer [20] 提出的 GDFN。将这些模块无缝结合，设计的 OSAG 能够在特征图中的任意一对标记之间进行信息传播。我们使用 [11] 提出的 ESA 模块进一步优化融合后的特征。

优化目标 沿用先前的工作，我们通过最小化模型预测 $\hat{I}_{HR}$ 与高分辨率 I_{HR} 标签之间的标准 L1 损失来训练模型：

$$L1 = ||\hat{I}_{HR} - I_{HR}||.$$

4 复现细节

4.1 与已有开源代码对比

对源代码进行部分修改。多尺度信息融合: 引入多尺度信息融合，使用不同卷积核大小（例如 3x3 和 5x5 卷积），捕获不同尺度的特征，提升模型的泛化能力。可变窗口大小: 在超分辨率网络中使用动态窗口大小有时会改善性能。当前的 OmniSR 网络使用了固定窗口大小。尝试在网络前向传播时根据输入尺寸动态调整窗口大小。

源代码上采样方法 pixelshuffle_block 通过 PixelShuffle 提升图像分辨率，考虑替换为更灵活的自适应上采样方法（如 AdaptiveAvgPool2d 或 Upsample）。PixelShuffle 效果好，但会增加计算开销，AdaptiveAvgPool2d 可以自适应调整图像大小，且不需要 PixelShuffle 层。

4.2 实验环境搭建

本实验于 pycharm,docker 中完成，依赖于多个 Python 库，包括 PyTorch (版本 >1.10)、OpenCV、Matplotlib 3.3.4、opencv-python、pyyaml、tqdm、numpy 和 torchvision，这些库支持深度学习、图像处理、数据可视化、配置文件读取、进度条显示以及高效的数组和矩阵运算。实验使用的数据集为 DIV2K×4 OmniSR_X4_DIV2K.zip，主要用于超分辨率任务，包含高分辨率图像用于训练模型。可以通过 pip install 命令安装所需依赖，并解压数据集到指定目录。确保使用适当的 Python 环境和 CUDA 配置，以便在 GPU 上进行加速训练。软件依赖, 数据集, 硬件环境如表 1 和 2 所示。

依赖项: PyTorch>1.10 , OpenCV , Matplotlib 3.3.4 , opencv-python , pyyaml , tqdm , numpy , torchvision.

数据集：DIV2K×4 OmniSR_X4_DIV2K.zip

Settings	CKPT name	CKPT url
DIV2K×2	OmniSR_X2_DIV2K.zip	Google driver
DF2K×2	OmniSR_X2_DF2K.zip	Google driver
DIV2K×3	OmniSR_X3_DIV2K.zip	Google driver
DF2K×3	OmniSR_X3_DF2K.zip	Google driver
DIV2K×4	OmniSR_X4_DIV2K.zip	Google driver
DF2K×4	OmniSR_X4_DF2K.zip	Google driver

表 1. 数据集源

环境	参数
操作系统	Windows11 version24H2
处理器	Intel(R) Core(TM) i9-14900K CPU@3.20GHz 24 核,32 线程
内存	64GB
显卡	NVIDIA GeForce RTX4090 VRAM24GB

表 2. 复现实验的硬件环境

5 实验结果分析

对比原文模型与复现模型在超分辨率任务中的表现，使用了多个标准数据集（Set5, Set14, BSD100, Urban100）进行验证。其中 PSNR (Peak Signal-to-Noise Ratio) 和 SSIM (Structural Similarity Index) 是评估超分辨率图像质量的两个重要指标，较高的 PSNR 和 SSIM 值通常意味着更好的重建质量。

从表格 3 中的结果可以看出，复现模型在所有数据集上的 PSNR 和 SSIM 略低于原文模型，尤其是在 Set5 和 Urban100 数据集上的差距较小。尽管存在微小的性能下降，这可能是由于复现过程中网络细节（如初始化策略、训练数据的微小差异等）所带来的影响。在所有测试数据集中，复现模型的表现与原文模型非常接近，尤其是在 Set5 数据集上，复现模型的 PSNR 和 SSIM 仅略低于原文，差距非常小，表明复现模型在标准数据集上的效果稳定。在 Set14 和 BSD100 数据集上，复现模型的表现略微下降，可能是由于数据集中的图像复杂度更高，复现模型未能完全捕捉到原文模型的特征提取能力。对于 Urban100 数据集，复现模型的表现与原文几乎相同，这表明网络在处理具有复杂纹理和细节的城市风格图像时，表现较为一致。参数数量相同的情况下，性能差异可以归因于其他因素，如训练过程中的随机性或实现细节的差异。虽然复现模型的结果略有下降，但这种变化是微小的，通常可以归因于复现过程中随机初始化、优化器超参数调整以及训练细节等因素。模型的总体性能保持在一个相似的水平，验证了修改后的网络架构（自适应上采样、动态窗口调整、多尺度信息融合等）对最终性能的贡献是有限的。

Method	Scale	Params	Set5	Set14	BSD100	Urban100
		(K)	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
原文	×4	792	32.49/0.8988	28.78/0.7859	27.71/0.7415	26.64/0.8018
复现	×4	792	32.36/0.8987	28.48/0.7871	27.75/0.7426	26.56/0.8018

表 3. 实验结果

6 总结与展望

本研究对原文超分辨率模型进行了复现，并对模型架构进行了若干改进，主要包括以下几个方面：我们引入了自适应上采样策略，以提高模型在不同尺度输入图像上的适应能力。通过这种方式，网络能够更精细地处理图像的不同细节，尤其在复杂图像中的表现有所提升。通过优化模型的特征提取与融合过程，结合了不同尺度的特征信息，从而提高了网络的表达能力，能够更好地处理复杂的纹理和结构信息。为减少在超分辨率重建过程中可能出现的边缘效应，我们对卷积窗口进行了动态调整，使得模型能根据图像的大小和内容自动调整感受野，从而提高了细节恢复能力。尽管改进后的复现模型在某些标准数据集上的性能略微下降，但与原文模型的结果非常接近。复现模型在复杂数据集（如 Urban100）上的表现稳定，表明修改后的网络结构仍保持了较好的鲁棒性。

尽管当前模型已经取得了一定的实验成果，但仍有多个方向值得进一步探索和优化：增强模型泛化能力：为了提高模型在实际应用中的泛化能力，可以考虑引入更复杂的生成对抗网络（GANs）结构，或结合其他增强学习方法，使得网络在不同类型的图像上表现得更为鲁棒。多尺度特征融合的进一步优化：在多尺度信息融合方面，尽管已做出初步改进，但依然可以通过深度学习中的新技术（如注意力机制、跨尺度卷积等）进行进一步优化，提升模型在复杂图像上的细节恢复能力。应用扩展与实际部署：未来研究可以进一步探讨该超分辨率模型在真实场景中的应用，如医学影像处理、卫星图像重建、视频超分辨率等领域。为了适应不同的应用需求，网络还需进一步进行迁移学习和领域自适应训练。

参考文献

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–268, 2018.

[2] 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.

[3] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[5] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019.

[6] Forrest N Iandola. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[7] 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.

[8] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 766–776, 2022.

[9] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.

[10] 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.

[11] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020.

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

[13] J Park. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[14] 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.

[15] 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.

[16] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[17] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[18] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021.

[19] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11998–12008, 2022.

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[21] 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.