计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):143−151 [doi: 10.15888/j.cnki.csa.009829] [CSTR: 32024.14.csa.009829]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于混合泛化 Transformer 的轻量化图像超分辨率重建①

刘俊辰, 张文波, 杨大为

(沈阳理工大学 信息科学与工程学院, 沈阳 110159)
通信作者: 杨大为, E-mail: dwyang@sylu.edu.cn

摘　要: 基于 Transformer 方法凭借自注意力机制在图像超分辨率重建领域中展现出卓越的性能, 然而自注意力机制也带来了非常高的计算成本, 针对此问题提出一种基于混合泛化 Transformer 的轻量化图像超分辨率重建模型. 该模型建立在 SwinIR 网络架构的基础上, 首先, 采用矩形窗口自注意机制 (RWSA), 利用不同头部的水平和垂直矩形窗口代替传统的正方形窗口模式, 整合跨越不同窗口的特征. 其次, 引用递归泛化自注意力机制 (RGSA) 将输入特征递归地聚合到具有代表性的特征映射中, 然后利用交叉注意力来提取全局信息, 同时将 RWSA 和 RGSA 交替结合, 以更有效地利用全局上下文信息. 最后, 为了激活更多的像素以获得更好的恢复, 使用通道注意力机制和自注意力机制并联地对输入图像进行特征提取. 在 5 种基准数据集的测试结果表明, 该模型在保持模型参数轻量化的同时取得了更好的重建性能.

关键词: 超分辨率重建; 轻量化; 通道注意力; 矩形窗口自注意力; 递归泛化自注意力

引用格式: 刘俊辰,张文波,杨大为.基于混合泛化 Transformer 的轻量化图像超分辨率重建.计算机系统应用,2025,34(3):143–151. http://www.c-s-a.org.cn/1003-3254/9829.html

Lightweight Image Super-resolution Reconstruction Based on Hybrid Generalization Transformer

LIU Jun-Chen, ZHANG Wen-Bo, YANG Da-Wei

(School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China)

Abstract: Transformer method, relying on a self-attention mechanism, exhibits remarkable performance in the field of image super-resolution reconstruction. Nevertheless, the self-attention mechanism also brings about a very high computational cost. To address this issue, a lightweight image super-resolution reconstruction model based on a hybrid generalized Transformer is proposed. This model is built based on the SwinIR network architecture. Firstly, the rectangular window self-attention (RWSA) mechanism is adopted. It utilizes horizontal and vertical rectangular windows with different heads to replace the traditional square window pattern, integrating features across different windows. Secondly, the recursive generalized self-attention (RGSA) mechanism is introduced to recursively aggregate input features into representative feature maps, followed by the application of cross-attention to extract global information. Meanwhile, RWSA and RGSA are alternately combined to make more effective use of global context information. Finally, to activate more pixels for better recovery, the channel attention mechanism and self-attention mechanism are used in parallel to extract features from the input image. Test results of five benchmark datasets show that this model achieves better reconstruction performance while keeping the model parameters lightweight.

Key words: super-resolution reconstruction; lightweight; channel attention; rectangular window self-attention; recursive generalized self-attention

① 基金项目: 辽宁省自然科学基金面上项目 (2022-MS-276)
收稿时间: 2024-07-30; 修改时间: 2024-09-19; 采用时间: 2024-11-08; csa 在线出版时间: 2025-01-21
CNKI 网络首发时间: 2025-01-22

1 引言

超分辨率重建算法的目标就是将低分辨率 (low resolution, LR) 图像或视频转换为高分辨率 (high resolution, HR) 图像或视频. 该领域的研究不仅是简单地放大图像, 而是通过恢复丢失的细节和纹理, 使图像在视觉上更加清晰和锐利. 超分辨率重建技术在许多领域都有广泛的应用, 包括数字图像处理、视频增强、医学影像和监控系统等[1].

随着深度学习技术的飞速发展, 它已成为超分辨率重建领域的主流方法[2]. 2014 年, Dong 等人[3]首次将深度学习引入该领域之中, 提出的 SRCNN 算法在图像重建质量上实现了显著提升. 随后, Lim 等人[4]通过引入残差模块和堆叠结构提出了 EDSR 算法, 进一步推动了性能提升. Zhang 等人[5]结合了注意力机制与残差学习提出了 RCAN 算法, 实现了更为卓越的视觉效果. Zhou 等人[6]提出了 IDNN 算法, 利用图像内部纹理跨尺度多次复现的特性使图像内部高清纹理信息得到充分利用.

近年来, Transformer 算法在多个图像处理领域取得了突破性进展, 其中也包括超分辨率重建领域. Chen 等人[7]提出 IPT, 一种 ImageNet 数据集上预训练的图像处理 Transformer. Liang 等人[8]基于 Swin Transformer 结构提出了 SwinIR 模型, 通过引入窗口内的局部自注意力机制及窗口间的转置连接, 更有效地捕捉了图像的局部细节与全局上下文信息. 而 Zhang 等人[9]则开创性地提出了高效长距离网络 ELAN, 同时提升了模型的效率与性能. Chen 等人[10]提出了 HAT, 通过利用 CNN 和 Transformer 的互补优势并聚合跨窗口信息获得了更好的效果. Ray 等人[11]提出了 CFAT, 在 HAT 的基础上提出一种非重叠三角形窗口技术, 从而使注意力机制能够在更多的图像像素上被激活.

当前, 该领域的算法倾向于构建更复杂且深层的网络架构, 以追求更高的重建性能. 然而, 这也导致了计算复杂度的显著提升. 为了平衡模型大小和重建性能, 轻量化的网络结构设计成为研究热点. Ahn 等人[12]提出的 CARN 通过级联和递归策略实现了高效特征提取. 随后, Hui 等人[13]相继提出的 IDN 和 IMDN 则引入了信息蒸馏结构, 通过复用特征来优化资源利用. 受 Lattice 滤波器启发的 Luo 等人[14]提出 LatticeNet, 实现了轻量级且效果出色的网络. Li 等人[15]则提出了 LAPAR 网络将图像超分辨率任务转化为基于多个预定义过滤器的线性回归任务, 利用像素自适应回归实现高效的重建. Lu 等人[16]提出了 ESRT 网络模型, 通过增强相似块的特征表达能力和长期依赖性, 验证了 Transformer 在轻量级超分辨率任务中的有效性.

尽管基于 Transformer 的方法凭借其有效捕获远程依赖关系的能力, 在图像超分辨率重建领域展现了巨大潜力. 然而, 其内置的自注意力机制导致计算复杂度显著上升, 训练过程中需要大量计算资源与存储空间, 进而延长了训练周期并增加了成本. 与此同时, 当前主流的轻量化网络结构设计虽成功降低了计算成本, 却也不可避免地牺牲了一定程度的模型性能, 尤其是在重建边缘细节丰富的图像时面临挑战. 因此, 如何在保持低计算负荷的同时, 提升模型重建图像的细节丰富度, 成为本文研究的重要方向.

针对上述问题, 本文提出了一种基于混合泛化 Transformer 的轻量化图像超分辨率重建算法. 具体而言, 首先在自注意力机制中引入矩形窗口代替传统的正方形窗口, 此设计能够在不增加计算开销的前提下, 通过跨窗口特征聚合来扩大感受野. 随后, 我们设计了递归泛化自注意力机制, 以更低的复杂度同时有效地提取全局信息. 同时, 将递归泛化自注意力机制与局部自注意力交替结合, 以捕获图像中的多尺度特征. 最后, 在模型中结合了通道注意力和自注意力, 旨在激活更多像素信息从而优化重建图像的质量.

2 整体网络架构

基于 SwinIR 网络的基础框架, 我们设计了一种基于混合泛化 Transformer 的轻量化图像超分辨率重建算法网络. 如图 1 所示, 该网络整体分为 3 个核心模块: 浅层特征提取模块、深层特征提取模块和上采样重建模块[17].

首先, 我们将输入图像 $I_{\mathrm{LR}} \in R^{H\times W\times C_{\mathrm{in}}}$ 输入浅层特征提取模块得到浅层特征 $F_0 \in R^{H\times W\times C}$, 其中 H、W、C_{in}、C 分别代表图像的高度、宽度、输入通道数和特征数. 其中浅层特征提取模块仅有由一个 3×3 卷积层组成[18], 其主要作用是将输入数据从原始的低维空间有效映射至高维特征空间. 这一过程能够更细致地捕捉输入数据中的细节与特性, 为后续处理或分析奠定更为全面且丰富的信息基础.

随后, 将提取到的浅层特征 F_0 通过深层特征提取

模块进一步地提取到图像深层特征$F_k \in R^{H\times W\times C}$. 其中该模块由$n_1$个残差组 (RG) 和一个 3×3 卷积层组成, 这里的卷积层用于聚合之前从 RG 中提取的特征. 经过每个 RG 后得到对应的输出$F_1, F_2, \cdots, F_{n_1}$, 其中$F_{n_1}$表示经过最后一个 RG 得到的特征, F_k表示为F_{n_1}经过 3×3 卷积层得到的深层特征.

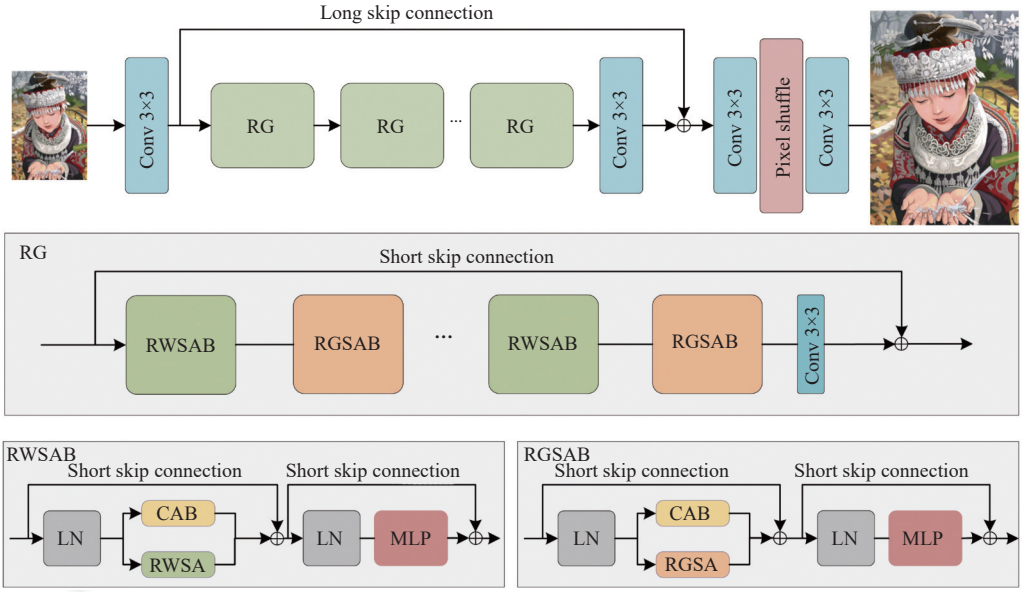

图 1 整体网络结构图

其中每个RG由n_2个 Transformer 块和一个 3×3 卷积层组成, 并采用残差连接来保证训练的稳定性. 每个 Transformer 块都由两层归一化 (LN)、自注意机制和多层感知器 (MLP) 组成. 该网络中的自注意机制分为两种类型: 矩形窗口自注意力 (RWSA) 和递归泛化自注意 (RGSA), 这两种类型的自注意力机制交替排列, 并在每个自注意机制并联一个通道注意力模块 (CAB). 为了平衡通道注意力模块和自注意机制的影响, 我们引入了一个缩放因子α其应用于通道注意力模块的输出权重上. 其中每 Transformer 块中具体的计算过程如下:

$$X_{\text{int}} = SA(LN(X_{\text{in}})) + \alpha CAB(LN(X_{\text{in}})) + X_{\text{in}} \quad (1)$$

$$X_{\text{out}} = MLP(LN(X_{\text{int}})) + X_{\text{int}} \quad (2)$$

其中, X_{in}, X_{int}和X_{out}分别为输入特征, 中间特征和输出特征, SA代表自注意机制, 其中包括 RWSA 或者 RGSA 这两种类型.

最后, 通过残差连接将浅层特征F_0和深层特征F_k进行融合, 在提升分辨率的同时保留图像的原始纹理和边缘信息. 随后由上采样重建模块处理生成 HR 图像$I_{\text{HR}} \in R^{H\times W\times C_{\text{out}}}$, 其中$C_{\text{out}}$为输出通道数, 本算法中上采样重建模块由“3×3 卷积+亚像素卷积[19]+3×3 卷积”构成, 使用亚像素卷积将深度特征上采样到与高分辨率输出相同的尺度, 并且在亚像素卷积之前和之后都有一个 3×3 卷积层来聚合并优化这些特征.

2.1 矩形窗口自注意力

本文引用一种新颖的窗口注意力机制 RWSA, 如图 2 所示, 该机制的核心优势在于其采用了非标准矩形窗口布局, 这与传统的正方形窗口模式形成鲜明对比. 通过独立调整矩形窗口的高度 (sh) 和宽度 (sw), 我们能够灵活地捕捉不同方向上的特征依赖关系.

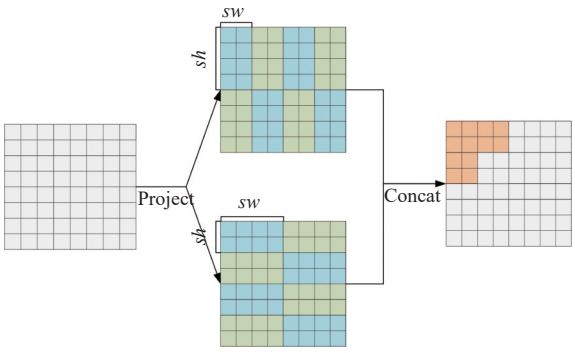

图 2 矩形窗口自注意力

为优化矩形窗口结构的利用效率, 我们将其细化为水平窗口 ($sh < sw$) 和垂直窗口 ($sh > sw$) 两种类型,

分别专注于捕捉水平方向和垂直方向上的特征变化, 从而显著增强了特征的精细化提取能力. 在模型的具体实现中, 这两种窗口并行部署于不同注意力头, 使得模型能够并行处理多个方向上的特征信息, 并借助注意力机制动态调整各方向信息的权重, 从而实现对输入数据的全面而精细的解析. 最终, 将两部分的输出沿着通道维度进行拼接. 这种设计使得矩形窗口能够在有限计算资源下, 更高效地捕捉每个像素在水平和垂直方向上的独特特征.

2.2 递归泛化自我注意力

尽管 Transformer 的自注意力机制能够有效地捕获全局信息, 但其计算复杂度却随着图像尺寸的增大而呈现二次方增长, 导致计算成本急剧上升. 为了应对这一挑战, 我们提出使用 RGSA, 如图 3 所示. RGSA 通过递归泛化模块将任意分辨率的图像特征整合成一系列具有代表性的特征图. 这些特征图有效地聚合了整个图像的关键特征信息, 为图像分析提供了全局视角. 随后, 通过计算输入特征与这些特征图之间的交叉注意力, 使得输入图像中的每个特征标记都能够获得全局感受野.

对于输入特征$X_{\text{in}} = R^{H\times W\times C}$, 我们首先通过递归地重复使用$T = \log_{s_r}(H/h)$次的单次深度卷积来压缩特征的空间大小, 以获得粗略的聚合图$Y = R^{h\times w\times c}$, 其中, h为特征图尺寸常数设置为 4, $w = W\times(h/H)$, s_r是卷积步长. 假设$W \leqslant H$, 那么$w \leqslant h$. 然后, 通过进一步的 3×3 深度卷积和 1×1 逐点卷积来细化这些聚合图, 以生成具有代表性的特征图$X_r \in R^{h\times w\times C_r}$. 递归泛化模块的公式表示为:

$$Y = W_r^T(X_{\text{in}}) = W_r(W_r(\cdots(W_r(X_{\text{in}})))) \tag{3}$$

$$X_r = W_p W_d(Y) \tag{4}$$

其中, W_r是具有s_r步长的深度卷积, W_d为 3×3 深度卷积, W_p为 1×1 的逐点卷积. 此外, 1×1 的逐点卷积用于将通道数从C扩展到$C_r = C\times c_r$, 其中c_r是调整因子. 通过递归泛化模块可以聚合输入图像特征的全局信息. 同时, 递归设计使得算法能灵活处理不同尺寸的输入, 通过动态调整递归次数T来适应图像超分辨率中的不同需求.

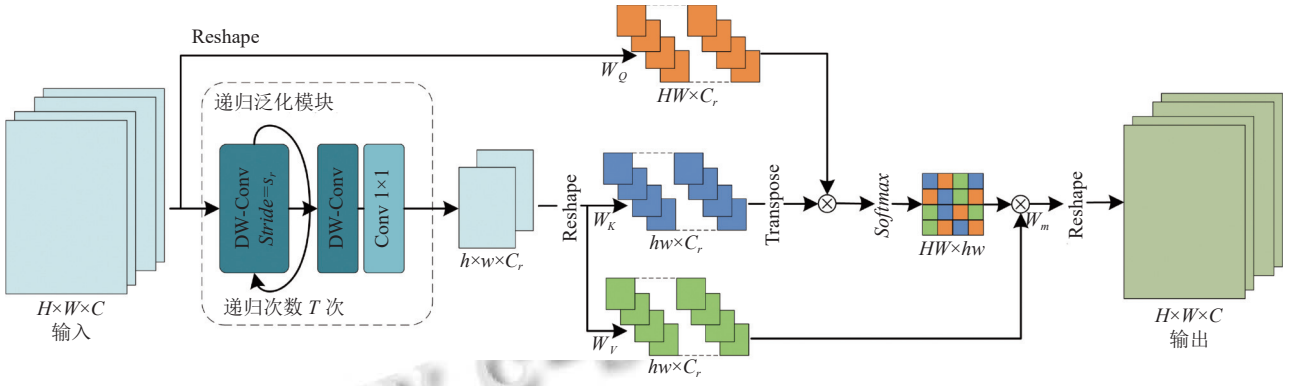

图 3 递归泛化自我注意力模块

交叉注意力机制的核心在于通过特定的变换与计算, 实现输入特征与代表图之间的相关性捕捉. 首先, 将输入特征X_{in}进行重塑, 并映射为查询$Q \in R^{HW\times C_r}$向量. 这一步骤的目的是将原始特征转换为适用于注意力机制计算的形式. 与此同时, 将代表图X_r重塑映射为键$K \in R^{hw\times C_r}$向量和值$V \in R^{hw\times C}$向量, 键和值的映射过程旨在提取图中各个节点或子结构的信息, 以供后续的注意力计算使用. 接下来, 通过计算查询与键的点积, 得到一个注意力矩阵$A \in R^{HW\times hw}$. 这个矩阵反映了输入特征中每个元素与代表图中每个节点或子结构之间的相关性强度. 最后, 根据注意力矩阵对值进行加权求和, 得到交叉注意力的输出. 这个输出综合了输入特征和代表图的信息, 反映了两者之间的相关性. 总的来说, 整个交叉注意力过程可以定义为:

$$Q = W_Q X_{\text{in}},\ K = W_K X_r,\ V = W_V X_r \tag{5}$$

$$A = Softmax\left(QK^T/\sqrt{C_r}\right) \tag{6}$$

$$Cross\text{-}Attention(X_{\text{in}}, X_r) = W_m(A\cdot V) \tag{7}$$

其中, $W_Q \in R^{C\times C_r}$、$W_K \in R^{C_r\times C_r}$和$W_V \in R^{C_r\times C}$为关键的可学习参数矩阵. $W_m \in R^{C\times C}$是一个形状为$R^{C\times C}$的

特征融合投影矩阵, 用于交叉注意力的计算. 与标准的自注意力机制相似, 此投影矩阵允许我们执行通道分割, 将通道分为多个头, 并在这些头上并行执行注意力操作. 在多头注意力架构中, 我们将输入特征的通道分割成多个子空间, 每个子空间都独立地计算注意力权重. 这种方法不仅增强了模型关注不同信息类型的能力, 还实现了信息的并行处理, 从而提高了计算效率.

通过应用交叉注意力机制, 输入特征能够与代表图进行有效的交互, 并捕捉它们之间的相关性. 最后, 我们通过重塑交叉注意力的结果来获得输出特征 $X_{\text{out}} \in R^{H\times W\times C}$. 通过结合递归泛化模块和交叉注意力, RGSA 在保持较低计算成本的同时, 有效地捕获了全局空间信息. 使得模型能够更全面地理解输入数据.

2.3 通道注意力模块

如图 4 所示, 基于残差通道注意力模块 (RCAB) 的基础上设计出了 CAB. 由于不需要构建更深的网络深度, 我们移除了 RCAB 中的残差连接, 并巧妙地融合了卷积运算与通道注意力 (CA) 机制, 以优化特征提取与增强的性能. 为了提升模型的表达能力和训练稳定性, 我们在两个卷积层中使用 *GELU* 激活函数代替传统的 *ReLU* 激活函数. 此外, 我们还实施了针对性的通道压缩策略, 通过引入压缩因子β来减少卷积层的通道数, β取值为 3.

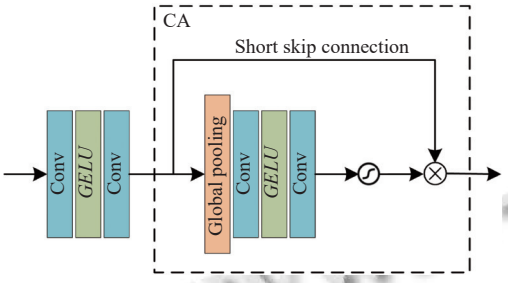

图 4 通道注意力模块

具体来说, 对于通道数为C的输入特征, 首个卷积层将通道数压缩至C/β, 随后第 2 个卷积层再将特征通道数扩展回C. 这种压缩与扩展策略不仅有效降低了模型的复杂度, 还显著提升了处理效率, 确保了模型在保持计算效率的同时, 能够捕获到足够的特征信息. 为了进一步提升模型对特征通道的敏感性, 我们引入了 CA, 能够自适应地调整不同通道的重要性, 从而实现对关键特征的增强和对冗余信息的抑制. 通过这样的设计, CAB 模块在保证计算成本可控的前提下, 实现了更为精准、高效的特征提取与增强.

3 实验分析

3.1 参数设置

训练环境配置为 Ubuntu 20.04 操作系统, 搭载 Intel Core i5-12400KF CPU, 以及配备 16 GB 显存的 NVIDIA GeForce RTX 4060 Ti GPU. Python 版本为 3.8.19, PyTorch 版本为 2.0.0, CUDA 版本为 11.8.

在训练过程中, 我们采用了 Adam 优化器, 其中 $\beta_1 = 0.9$和$\beta_2 = 0.99$, 并将损失函数设定为$L1$. 训练轮次 (epochs) 设置为 500k, 初始学习速率设置为2×10^{-4}, 在训练轮次达到[250k, 400k, 450k, 475k]时, 将学习率减半, 批量处理的大小 (batch_size) 设置为 8.

在轻量化网络模型设计中, 我们将每个残差组数n_1设为 3, 残差组中的 Transformer 块数n_2设为 6, 其中矩形窗口自注意力 (RWSAB) 和递归泛化自我注意力模块 (RGSAB) 这两种模块交替排列, 数量各为 3. 矩形窗口大小设置为 8×32, 模型的通道维度设为 48, 注意力机制中的头数设为 6, MLP 的扩展比设为 2. 对于 RGSA, 卷积步长s_r和调整因子c_r分别设置为 4 和 0.5, 以优化其在网络中的表现.

3.2 数据集与评价指标

选择 DIV2K[20] (包含 800 张图像) 和 Flickr2K[21] (包含 2 650 张图像) 作为训练数据集, 低分辨率图像通过双三次插值下采样生成. 对输入图像进行随机裁剪至 64×64 大小, 并应用随机水平翻转和旋转等增强技术以增加训练样本的多样性[22]. 实验在 5 个基准数据集上进行测试: Set5[23]、Set14[24]、B100[25]、Urban100[26]和 Manga109[27], 同时考虑了×2、×3 和×4 这 3 种不同的放大因子.

为了评估我们的方法, 我们在 YCbCr 空间的 Y 通道上使用了峰值信噪比 (PSNR) 和结构相似性 (SSIM)[28]. 其中, PSNR 值越高表明重建图像与原始图像之间的误差越小, SSIM 值越接近 1 则表明重建图像在结构上与原始图像越相似[29]. 此外, 我们还进行了主观评估, 基于直观的视觉效果来判断图像的清晰度与细节保留程度, 以补充客观指标的不足.

3.3 实验结果对比分析

我们将所提出的轻量化网络模型与近年来流行的轻量级图像超分辨率网络进行了对比, 其中包括 EDSR-baseline、CARN、IMDN、LatticeNet、LAPAR-A 以及基于 Transformer 的 ESRT、ELAN-light 和 SwinIR-

light. 如表 1 所示, 最佳结果以加粗字体标出, 次优结果则以下划线表示.

从表 1 中数据可以明显看出, 我们提出的方法在大多数情况下均取得了卓越的性能表现. 尤其是在放大因子为×3 和×4 时, 5 个基准测试数据集上的 PSNR 和 SSIM 指标均达到了最佳水平.

表 1 不同方法的 PSNR 和 SSIM 的对比

放大因子	方法	复杂度	Set5	Set14	B100	Urban100	Manga109
			PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
×2	EDSR-baseline	1370k	37.99/0.9604	33.57/0.9175	32.16/0.8994	31.98/0.9272	38.54/0.9769
	CARN	1592k	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
	IMDN	692k	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
	LatticeNet	756k	38.06/0.9607	33.70/0.9187	32.20/0.8999	32.25/0.9288	—
	LAPAR-A	548k	38.01/0.9605	33.62/0.9183	32.19/0.8999	32.10/0.9283	38.67/0.9772
	ESRT	677k	38.03/0.9600	33.75/0.9184	32.25/0.9001	32.58/0.9318	39.12/0.9774
	ELAN-light	582k	38.17/0.9611	**33.94/0.9207**	32.30/0.9012	**32.76/0.9340**	39.11/0.9782
	SwinIR-light	878k	38.14/0.9611	33.86/0.9206	**32.31**/0.9012	**32.76/0.9340**	39.12/0.9783
	本文算法	822k	**38.21/0.9613**	33.91/0.9205	32.30/**0.9015**	32.65/0.9331	**39.25/0.9784**
×3	EDSR-baseline	1555k	34.37/0.9270	30.28/0.8417	29.09/0.8052	28.15/0.8527	33.45/0.9439
	CARN	1592k	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.50/0.9440
	IMDN	703k	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
	LatticeNet	765k	34.53/0.9281	30.39/0.8424	29.15/0.8059	28.33/0.8538	—
	LAPAR-A	544k	34.36/0.9267	30.34/0.8421	29.11/0.8054	28.15/0.8523	33.51/0.9441
	ESRT	770k	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574	33.95/0.9455
	ELAN-light	590k	34.61/0.9288	30.55/0.8463	29.21/0.8081	28.69/0.8624	34.00/0.9478
	SwinIR-light	886k	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624	33.98/0.9478
	本文算法	906k	**34.69/0.9293**	**30.60/0.8471**	**29.25/0.8097**	**28.73/0.8643**	**34.35/0.9490**
×4	EDSR-baseline	1518k	32.09/0.8938	28.58/0.7813	27.57/0.7357	26.04/0.7849	30.35/0.9067
	CARN	1592k	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
	IMDN	715k	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
	LatticeNet	777k	32.18/0.8943	28.61/0.7812	27.57/0.7355	26.14/0.7844	—
	LAPAR-A	659k	32.15/0.8944	28.61/0.7818	27.61/0.7366	26.14/0.7871	30.42/0.9074
	ESRT	751k	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100
	ELAN-light	601k	32.43/0.8975	28.78/0.7858	27.69/0.7406	26.54/0.7982	30.92/0.9150
	SwinIR-light	897k	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
	本文算法	869k	**32.50/0.8985**	**28.85/0.7875**	**27.73/0.7421**	**26.61/0.8023**	**31.16/0.9162**

相较于基线模型 SwinIR-light, 本文方法在 Manga109 数据集上展现出显著优势. 针对尺度因子×3 和×4 的 PSNR 增益分别高达 0.37 dB 与 0.24 dB, 这突显了我们模型在激活复杂像素特征及捕获更多的细节信息的卓越能力, 尤其适用于 Manga109 这类风格独特、细节丰富的图像. 同时, 在 Urban100 数据集上, 我们的方法也表现出色, 实现了尺度因子×3 和×4 下 SSIM 增益分别为 0.0019 和 0.0043 的提升. Urban100 以其丰富的重复纹理结构著称, 这一成果表明, 我们模型通过跨越不同窗口的特征与递归自注意力机制, 有效捕获了全局信息, 对重建此类图像具有关键作用.

然而, 在 Set5、Set14 和 B100 这 3 个数据集上, 尽管我们方法有所改进, 但提升幅度不及 Urban100 和 Manga109 显著. 主要因为这些数据集图像内容的多样性, 涵盖了人物、自然风景、动物等多种类型, 增加了模型处理复杂图像变化的难度. 因此, 面对多样化的图像内容, 我们的模型在泛化能力上仍有提升空间.

3.4 视觉效果对比分析

在图 5 中, 我们分别展示了在各个数据集上不同的放大因子的可视化比较结果. 实验结果表明, 多数技术难以精确恢复图像的纹理, 并经常伴随有模糊伪影的出现. 相比之下, 我们提出的方法显著提升了图像细节清晰度, 并更有效地缓解了模糊伪影问题, 恢复更多的图像细节. 例如在 UchuKigekiM774 图片中, 针对嘴边皱纹的黑色与脸部红色部分, 本文方法重建图像的颜色更加鲜明, 确保了色彩的准确性和丰富性, 有效避免了色彩失真或淡化的问题. 在 zebra 图片的斑马腿部纹路部分, 相较于其他算法可能导致重建的图像细节模糊与噪声过多问题, 本文方法重建的图像层次更加清晰, 显著减少了边缘模糊现象, 增强了图像的视觉细

节与真实感. 在 148 026 图片和 img076 图片中, 本文方法更清晰地勾勒出图像中的轮廓, 成功恢复了更多细腻的纹理细节, 使图像看起来更加接近原 HR 图像. 这些视觉比较结果清楚地表明, 我们的方法能够更加有效的提取全局信息并且激活更多的像素点来重建高质量的图像, 从而进一步证明了该方法的有效性.

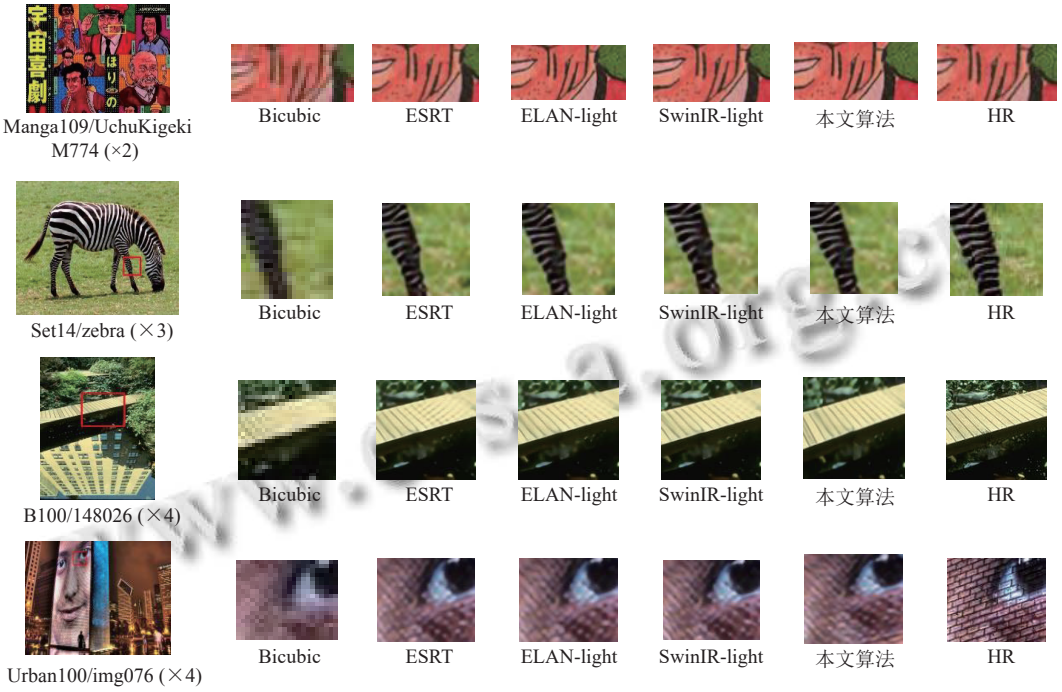

图 5 视觉效果对比分析

3.5 消融实验

为了验证所使用模块 (即 RWSA、RGSA 和 CAB) 的有效性, 设计并实施了不同模块组合的消融实验. 如表 2 所示, 将 RWSA 和 RGSA 分别替代 SwimIR 中原有的所有自注意力机制模块, 发现 RGSA 在维持与 RWSA 相近性能水平的同时, 显著降低了模型的计算复杂度, 证明了 RGSA 保持较低计算成本的同时能够有效地捕获全局信息. 进一步地, 我们创新性地尝试将 RWSA 与 RGSA 交替排列于网络模型中, 这一改动不仅巧妙地平衡了模型复杂度与性能之间的关系, 还在多个数据集上的性能指标均有所提升, 初步验证了这两种模块间互补的有效性. 最后, 我们在每个自注意力层上并联一个 CAB, 通过其独特的卷积注意力机制, 增强了模型对图像特征的提取能力, 尽管这一改动增加了模型的复杂度, 但模型性能却实现了显著飞跃. 从而充分验证了这 3 种模块组合的有效性.

表 2 不同模块组合下对×2 模型的性能影响

方法	复杂度	Set5	Set14	B100	Urban100	Manga109
		PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
RWSA	612k	38.10/0.9609	33.73/0.9192	32.26/0.9008	32.32/0.9303	39.04/0.9780
RGSA	553k	38.08/0.9608	33.70/0.9190	32.26/0.9010	32.33/0.9307	39.01/0.9780
RWSA+RGSA	567k	38.11/0.9610	33.72/0.9192	32.26/0.9010	32.48/0.9318	39.07/0.9780
RWSA+RGSA+CAB	822k	38.21/0.9613	33.91/0.9205	32.30/0.9015	32.65/0.9331	39.25/0.9784

对于 CAB 采用α来控制模块提取特征的权重, 以进行特征融合, α越大, 表示 CAB 提取的特征权重越大. 如表 3 所示, 分别对α取值为 0、1、0.1 和 0.01 在放大因子为×2 的 Set5 测试数据集进行对比实验, 发现α的取值并不影响模型的复杂度, 当α取值为 0.01 时模型性能最好.

表 3 α取值对×2 模型的 PSNR 性能影响 (dB)

α	0	1	0.1	0.01
PSNR	38.11	38.15	38.17	38.21

实验对比了不同窗口尺寸 (如 4×16、8×32、16×64) 下的模型表现, 如表 4 所示, 发现适度增加窗口尺寸能够显著提升模型性能.

表 4 不同窗口尺寸对×2 模型的性能影响

窗口尺寸大小	Set5	Set14	B100	Urban100	Manga109
	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
4×16	38.13/0.9609	33.74/0.9193	32.25/0.9007	32.33/0.9301	39.08/0.9780
8×32	38.21/0.9613	33.91/0.9205	32.30/0.9015	32.65/0.9331	39.25/0.9784
16×64	35.16/0.9611	33.81/0.9201	32.28/0.9013	32.57/0.9325	39.17/0.9782

这一提升主要归功于增大的注意力窗口所带来的更广阔感受野, 显著增强了模型捕捉细节信息及详尽局部上下文的能力. 然而, 当窗口尺寸过度增大至 16×64 时, 却出现了性能反转的现象. 这主要是因为, 在给定序列长度受限的条件下, 过大的窗口会导致窗口间重叠区域的显著减少, 从而削弱了模型对全局信息的捕捉能力. 同时, 过大的窗口尺寸也会带来相应的时间复杂度增加, 计算成本上升. 因此, 为了平衡模型性能和计算成本, 选择窗口尺寸为 8×32 的模型综合效果最佳.

4 总结

本文提出了一种基于混合泛化 Transformer 的轻量化图像超分辨率重建算法. 该算法基于 SwinIR 网络的基础上, 引入了矩形窗口注意力机制, 通过并行地执行水平窗口和垂直窗口的注意力操作, 在不增加额外计算成本的前提下, 有效扩大了感受野. 同时, 该算法创新性地将递归泛化自注意力机制与矩形窗口注意力机制交替结合, 能够保持低计算成本的同时, 对全局空间信息进行建模. 最后, 将通道注意力机制与所有自注意力机制并联, 充分利用它们的互补优势, 以激活更多的输入像素, 从而获得更好的图像恢复效果. 在 5 个基准数据集上的实验结果表明, 该算法在精度和视觉质量方面均优于其他算法. 具体而言, 该算法在保持低计算负荷的同时, 显著提升了模型重建图像的细节丰富度. 在轻量化超分辨率重建方面取得了一定成效, 但在细节捕捉方面, 距离真实图片仍有明显不足. 后续工作中, 我们将继续深入探索并尝试创新的改进策略, 以进一步优化和完善现有模型, 在保持甚至降低模型复杂度的同时, 实现更高精度的细节重建.

参考文献

1 张瑾, 李佳莹, 李晓阳, 等. 基于 SRGAN 的图像超分辨率重建. 电脑知识与技术, 2024, 20(1): 14–17.

2 吴丽君, 蔡周威, 陈志聪. 基于改进残差特征蒸馏的轻量级超分辨率网络. 计算机与现代化, 2022(11): 89–94. [doi: 10.3969/j.issn.1006-2475.2022.11.013]

3 Dong C, Loy CC, He KM, *et al*. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295–307. [doi: 10.1109/TPAMI.2015.2439281]

4 Lim B, Son S, Kim H, *et al*. Enhanced deep residual networks for single image super-resolution. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1132–1140.

5 Zhang YL, Li KP, Li K, *et al*. Image super-resolution using very deep residual channel attention networks. Proceeding of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 294–310.

6 Zhou SC, Zhang JW, Zuo WM, *et al*. Cross-scale internal graph neural network for image super-resolution. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 295.

7 Chen HT, Wang YH, Guo TY, *et al*. Pre-trained image processing Transformer. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12294–12305.

8 Liang JY, Cao JZ, Sun GL, *et al*. SwinIR: Image restoration using Swin Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 1833–1844.

9 Zhang XD, Zeng H, Guo S, *et al*. Efficient long-range attention network for image super-resolution. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 649–667.

10 Chen XY, Wang XT, Zhang WL, *et al*. HAT: Hybrid attention Transformer for image restoration. arXiv:2309.05239, 2023.

11 Ray A, Kumar G, Kolekar MH. CFAT: Unleashing triangular Windows for image super-resolution. arXiv:2403.16143, 2024.

12 Ahn N, Kang B, Sohn KA. Fast, accurate, and lightweight super-resolution with cascading residual network. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 256–272.

13 Hui Z, Gao XB, Yang YC, *et al*. Lightweight image super-

resolution with information multi-distillation network. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 2024–2032.

14 Luo XT, Xie Y, Zhang YL, *et al*. LatticeNet: Towards lightweight image super-resolution with lattice block. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 272–289.

15 Li WB, Zhou K, Qi L, *et al*. LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2020. 1708.

16 Lu ZS, Li JC, Liu H, *et al*. Transformer for single image super-resolution. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 456–465.

17 朱凯, 李理, 张彤, 等. 视觉 Transformer 在低级视觉领域的研究综述. 计算机工程与应用, 2024, 60(4): 39–56. [doi: 10.3778/j.issn.1002-8331.2304-0139]

18 王鑫, 余磊. 多尺度双注意力的图像超分辨率重建方法. 计算机与现代化, 2024(8): 77–87. [doi: 10.3969/j.issn.1006-2475.2024.08.013]

19 Shi WZ, Caballero J, Huszár F, *et al*. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1874–1883.

20 Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1122–1131.

21 Timofte R, Agustsson E, van Gool L, *et al*. NTIRE 2017 challenge on single image super-resolution: Methods and results. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1110–1121.

22 徐雯捷, 宋慧慧, 袁晓彤, 等. 轻量级注意力特征选择循环网络的超分重建. 中国图象图形学报, 2021, 26(12): 2826–2835. [doi: 10.11834/jig.200555]

23 Bevilacqua M, Roumy A, Guillemot C, *et al*. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. Proceedings of the 2012 British Machine Vision Conference. Surrey: BMVA Press, 2012. 1–10.

24 Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations. Proceedings of the 7th International Conference on Curves and Surfaces. Avignon: Springer, 2012. 711–730.

25 Martin D, Fowlkes C, Tal D, *et al*. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver: IEEE, 2001. 416–423.

26 Huang JB, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 5197–5206.

27 Matsui Y, Ito K, Aramaki Y, *et al*. Sketch-based manga retrieval using Manga109 dataset. Multimedia Tools and Applications, 2017, 76(20): 21811–21838. [doi: 10.1007/s11042-016-4020-z]

28 徐国明, 王杰, 马健, 等. 基于双重注意力残差网络的偏振图像超分辨率重建. 光子学报, 2022, 51(4): 0410001.

29 刘婉春, 景明利, 王子昭, 等. 基于 Transformer 和双残差网络的图像去模糊算法研究. 信息技术与信息化, 2023(1): 217–220. [doi: 10.3969/j.issn.1672-9528.2023.01.051]

(校对责编: 张重毅)