计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010091] [CSTR: 32024.14.csa.010091]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

MTSR: Mamba-Transformer 协同增强的轻量化图像超分辨率模型①

庞梦鑫, 董智红, 曹　鹏, 洪京平, 张鸣赟

(北京印刷学院 信息工程学院, 北京 102600)
通信作者: 董智红, E-mail: dongzhihong@bigc.edu.cn

摘　要: 为解决现有轻量化图像超分辨率模型在平衡全局感受野、局部特征提取能力与计算效率方面的不足, 并针对 Mamba 架构在上下文建模中存在的跨 token 交互能力不足问题, 提出一种名为 MTSR 的高效 Mamba-Transformer 协同网络. 首先, 构建一种混合协同架构, 通过合理配比 Mamba 与 Transformer 模块, 利用 Transformer 卓越的跨 token 交互能力弥补纯 Mamba 模型在上下文建模方面的缺陷, 实现了长程依赖建模与计算效率的有效平衡. 其次, 设计一种深度卷积注意力前馈网络, 用以替代传统的多层感知机. 此网络能够显著增强局部细节特征的提取能力和通道间的信息交互, 从而减少重建过程中的像素级信息损失, 从而更充分地发挥 Mamba 模块的性能潜力. 最后, 提出一个三重深度可分离浅层细化模块, 该模块专注于高效捕获并增强图像的浅层特征, 为后续的非线性映射提供更丰富的原始纹理信息. 在 5 个公开基准数据集上的大量实验结果表明, 所提 MTSR 模型相较于当前的轻量化 SOTA 模型 SRFormer-light 和 MambaIR-light, 峰值信噪比 (PSNR) 分别获得了高达 0.31 dB 和 0.38 dB 的性能增益, 同时保持了 Mamba 高效推理速度的优势. 实验结果表明, 该方法为轻量化图像超分辨率领域提供了一种兼具高性能与高效率的有效解决方案.

关键词: 超分辨率; 轻量化; 状态空间模型; 自注意力; Mamba; Transformer

引用格式: 庞梦鑫,董智红,曹鹏,洪京平,张鸣赟.MTSR: Mamba-Transformer 协同增强的轻量化图像超分辨率模型.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10091.html

MTSR: Mamba-Transformer Collaboration Enhanced Lightweight Image Super-resolution Model

PANG Meng-Xin, DONG Zhi-Hong, CAO Peng, HONG Jing-Ping, ZHANG Ming-Yun

(School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China)

Abstract: This study proposes an efficient Mamba-Transformer collaborated network, termed MTSR, to address the challenges of balancing the global receptive field, local feature extraction, and computational efficiency in lightweight image super-resolution models, while tackling the insufficient cross-token interaction capability in Mamba-based context modeling. First, a hybrid synergy architecture is constructed by integrating Mamba and Transformer modules in an optimal ratio. This design leverages the Transformer's superior cross-token interaction ability to compensate for the contextual modeling deficiencies in purely Mamba-based models. Second, a novel deep convolutional attention feed-forward network is designed to replace the traditional multi-layer perceptron. This network significantly enhances the capability for local detail extraction and effective inter-channel communication, thus reducing pixel-level information loss and amplifying the performance potential of the Mamba modules. Finally, a triple depthwise-separable shallow refinement block is introduced to efficiently capture and enhance the shallow-level features of an image, thus providing richer

① 基金项目: 北京市科技计划 (Z241100007624008); 北京印刷学院信息与通信工程一级学科博士点培育项目 (21090525004); 北京印刷学院科研平台建设项目 (KYCPT202509)
收稿时间: 2025-09-01; 修改时间: 2025-09-22; 采用时间: 2025-10-09; csa 在线出版时间: 2026-01-08

original texture information for subsequent nonlinear mapping. Extensive experiments on five public benchmark datasets demonstrate that the proposed MTSR model achieves peak signal-to-noise ratio (PSNR) gains of up to 0.31 and 0.38 dB over the state-of-the-art lightweight models SRFormer-light and MambaIR-light, respectively, while maintaining a highly competitive inference speed. This study validates that the proposed method provides an effective solution for the field of lightweight image super-resolution, combining both high performance and high efficiency.

Key words: super-resolution; lightweight; state space model; self-attention; Mamba; Transformer

1 引言

图像超分辨率 (super-resolution, SR) 技术旨在从对应的低分辨率 (low-resolution, LR) 图像中重建出高分辨率 (high-resolution, HR) 图像, 以增强图像的细节纹理与视觉质量. 作为计算机视觉 (computer vision, CV) 领域中一个极具挑战性的底层视觉问题[1,2], 该技术已在医疗成像、卫星遥感和视频监控等多个领域获得了广泛应用. 传统的 SR 技术主要依赖于插值或稀疏表示模型, 然而, 这些方法在处理复杂或真实世界场景时, 其性能往往难以令人满意.

近年来, 深度学习的兴起极大地推动了 SR 性能的提升, 其中卷积神经网络 (convolutional neural network, CNN) 的应用尤为突出. 2014 年, Dong 等人[3]将 CNN 结构首次引入图像复原领域, 提出了超分辨率卷积神经网络 SRCNN, 该模型采用一个简洁的 3 层卷积结构学习从 LR 到 HR 的端到端映射. 这项工作凸显了 CNN 在捕捉精细图像细节方面的强大表征能力, 并迅速引起了学术界的广泛关注. 早期的 CNN 模型设计紧凑, 通过逐层叠加卷积来提取并聚合局部特征, 相较于传统方法展现了更强的学习能力和鲁棒性. 为进一步提升模型性能, 研究人员后续引入了一系列关键的改进策略, 包括残差学习[4]、跳跃连接[5]、密集连接[6]以及注意力机制[7]等, 这些架构改进显著增强了全局上下文信息的整合能力, 并有效缓解了梯度消失与信息衰减等问题. 与此同时, 2014 年, Goodfellow 等人[8]提出的生成对抗网络 (generative adversarial network, GAN) 进一步拓宽了 SR 的研究边界. 基于 GAN 的方法通过对抗性训练, 显著改善了重建图像的感知真实感, 尤其在恢复高频细节方面提供了全新的视角与技术路径.

2017 年, Vaswani 等人[9]提出了 Transformer 架构, 利用其核心的自注意力机制高效建模长程依赖关系. 随后, Dosovitskiy 等人[10]提出的 Vision Transformer (ViT) 成功将这一架构应用于计算机视觉任务, 通过将图像划分为块序列, 实现了卓越的图像识别性能. 受此启发[11], 研究者开始将 Transformer 引入图像超分辨率领域, 以克服 CNN 模型在长程依赖建模上的不足. 其中, Liang 等人[12]提出的 SwinIR 是代表性工作之一, 它采用基于移位窗口的自注意力机制, 将计算限制在局部窗口内, 有效降低了标准 Transformer 的二次方计算复杂度, 同时通过窗口移位实现了跨窗口的信息交互, 在多个图像复原任务中取得了领先性能. 类似地, Zhou 等人[13]提出的 SRFormer 则通过探索置换自注意力机制, 进一步优化了特征聚合方式. 这些基于 Transformer 的模型充分利用了自注意力的全局上下文建模能力, 在恢复图像的结构信息和复杂纹理方面展现出显著优势. 此外, 近期扩散模型 (diffusion model, DM) 因其在图像生成合成任务中强大的能力而备受瞩目, 部分研究已探索将其应用于 SR 领域并取得了富有前景的结果[14].

尽管深度学习为 SR 领域带来了巨大进步, 但现有方法仍存在显著局限. CNN 因其固有的局部卷积操作, 难以有效建模长程依赖, 从而限制了对图像整体结构的恢复能力. Transformer 虽通过全局感受野解决了这一问题, 但其计算与内存复杂度随输入序列长度呈二次方增长, 这在处理大规模图像输入时造成了严重的性能瓶颈, 并且其在捕获细粒度局部特征方面会表现出一定的局限性.

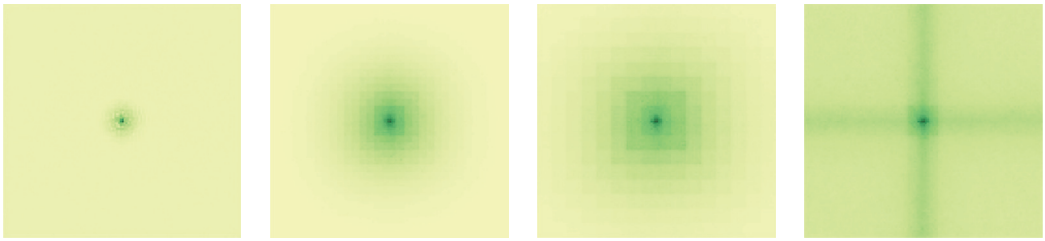
2023 年, Gu 等人[15]提出了基于状态空间模型 (state space model, SSM) 的 Mamba 架构, 为高效建模长序列和捕获长程依赖提供了新的解决方案. 该架构通过引入一种选择性扫描机制, 实现了与序列长度呈线性相关的计算复杂度, 成功规避了 Transformer 自注意力机制的二次方开销. 这一高效率特性使得 Mamba 模块能够在不显著增加计算成本的前提下, 有效捕获远距离像素间的相关性. 很快, Mamba 架构便被应用于图像复原领域. Shi 等人[16]提出的 VmambaIR 和 Guo 等人[17]提出的 MambaIR 是早期的探索性工作. 它们通过将二

维图像展平为一维序列并采用多向扫描策略, 将 Mamba 应用于图像 SR 任务, 初步验证了 SSM 在平衡性能与效率方面的潜力. 这些模型能够在不显著增加计算成本的前提下, 有效捕获远距离像素间的相关性, 对于 SR 任务中的纹理重建至关重要.

然而, 近期研究表明, 将这些新兴架构应用于图像 SR 任务时, 新的问题也随之浮现. 尽管 Transformer 和 Mamba 都能有效建模全局依赖, 但二者机理不同, 各有侧重. Transformer 的自注意力机制擅长进行全局、并行的跨 token 信息交互, 但计算成本高昂. Mamba 通过循环式的状态更新实现长程依赖建模, 推理效率极高, 但在需要灵活、动态的 token 间信息聚合与复用的复杂场景中, 其上下文交互能力相对受限. 在以 Transformer 架构为主体的 SR 模型中直接用 Mamba 模块替代其中的自注意力层, 其性能并不理想. 这暗示了 Mamba 模块与 Transformer 架构中传统的多层感知机 (multi-layer perceptron, MLP) 前馈网络之间可能存在结构上的不兼容性[17]. 此外, 一项针对 Mamba 语言模型的全面实证研究发现[18], 在训练预算有限的情况下, 纯 Mamba 架构在处理需要强上下文感知与灵活信息复用的任务时表现欠佳. 其部分原因在于 SSM 层促进跨 token 信息交互的能力有限, 而这正是自注意力机制的优势所在. 为缓解该问题, 该研究提出将 Transformer 组件与 Mamba 框架结合的混合模型, 旨在保留 Mamba 计算效率的同时增强上下文建模能力. 但是, 一个纯 Mamba 主干网络是否会在 CV 任务中遭遇类似限制, 仍是一个有待探究的问题.

为克服上述局限性, 本文提出了一种融合 Mamba 与 Transformer 的轻量化图像 SR 网络, 命名为 MTSR[19], 即 Mamba-Transformer 超分辨率 (Mamba-Transformer super-resolution). 提出一种深度卷积注意力 (deep convolutional attention, DCA) 模块, 作为一个改进的前馈网络单元, 它旨在与 Mamba 模块无缝集成, 在保持与 Transformer 结构兼容性的同时, 显著增强 Mamba 捕获局部特征的能力. 设计了一个混合注意力状态空间块 (hybrid attention state space block, HASSB). 在该模块中, Mamba 与移位窗口多头自注意力 (shifted window multi-head self-attention, SW-MHSA) 协同工作, 并共享 DCA 作为其前馈网络. 实验评估证明, 这种混合设计显著提升了重建性能. 提出了一种三重深度可分离浅层细化 (triple depthwise-separable shallow refinement, TDSR) 模块, 以多尺度、多层次的方式提取浅层特征, 从而丰富网络处理精细细节的能力.

此外, 如图 1 所示的有效感受野 (effective receptive field, ERF) 分析表明, 与传统的 CNN 模型乃至 Transformer 模型相比, MTSR 获得了更大的全局感受野, 这表明模型能更好地激活远距离像素并建模长程依赖. 同时, 如图 2 所示在不同数据集的推理时间评估证实, 本方法在实现卓越重建质量的同时, 保留了 Mamba 架构固有的高速推理优势.



(a) A^2F (b) SwinIR (c) SRFormer (d) MTSR

图 1 部分模型的有效感受野

全面的消融研究表明, 单独使用 Mamba 作为 SR 任务的主干网络, 会因跨 token 信息流受限而导致性能瓶颈. 通过策略性地引入少量 Transformer 组件, 可以大幅提升 SR 图像的质量.

2 预备知识

SSM 是一种用于描述动态系统的数学框架, 其核心功能是将一个输入信号 $x(t)\in\mathbb{R}^L$ 映射到一个输出信号 $y(t)\in\mathbb{R}^L$. 该模型通过引入隐状态变量 $h(t)\in\mathbb{R}^N$ 来捕捉系统的动态特性. 其数学表达通常被构建为线性常微分方程 (ordinary differential equation, ODE) 的形式:

$$h'(t)=Ah(t)+Bx(t) \tag{1}$$

$$y(t)=Ch(t)+Dx(t) \tag{2}$$

其中, N表示状态向量$h(t)$的维度, 代表系统的内部状态数量. 矩阵$A \in \mathbb{R}^{N\times N}$、$B \in \mathbb{R}^{N\times L}$、$C \in \mathbb{R}^{L\times N}$和$D \in \mathbb{R}^{L\times L}$分别对应状态矩阵、输入映射矩阵、输出矩阵以及输入到输出的直接传输矩阵.

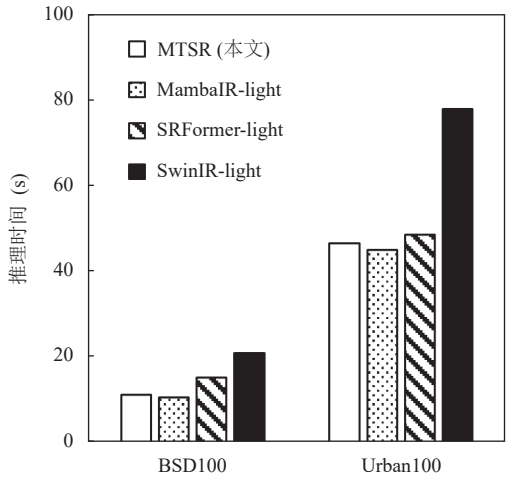


图 2 不同模型在单个 NVIDIA GTX 2080Ti 上的推理时间

然而, 直接在计算中实现连续时间模型存在困难, 因此需进行离散化处理. 一种常用的方法是零阶保持器 (zero-order hold, ZOH), 该方法通过设定一个时间步长Δ, 将连续参数A和B转换为离散形式的对应参数$\bar{A}$和$\bar{B}$. ZOH 的核心思想是假设在每个离散的采样间隔Δ内,输入信号$x(t)$是一个常数,即$x(t) = x_k$对于$t \in [(k-1)\Delta, k\Delta)$. 基于此假设, 连续状态方程式 (1) 作为一个线性常微分方程, 其在t_0时刻的初值$h(t_0)$到t时刻的解为:

$$h(t) = \mathrm{e}^{A(t-t_0)}h(t_0) + \int_{t_0}^{t} \mathrm{e}^{A(t-T)}Bx(\tau)\mathrm{d}\tau$$

将此解应用于一个采样间隔, 令$t_0 = (k-1)\Delta$且$t = k\Delta$. 在此区间内, $x(\tau)$被视为常数x_k, 因此可以得到离散状态$h_k = h(k\Delta)$的递推关系:

$$h_k = \mathrm{e}^{A\Delta}h_{k-1} + \left(\int_0^{\Delta} \mathrm{e}^{As}\mathrm{d}s\right)Bx_k$$

通过求解上式中的积分项, 便可得到离散化后的状态矩阵$\bar{A}$和输入映射矩阵$\bar{B}$. 这种离散化方法在 ZOH 假设下是精确的, 它为 SSM 在数字计算平台上的高效实现奠定了理论基础.

离散化后的模型表达如下:

$$h'_t = \bar{A}h_{t-1} + \bar{B}x_t \tag{3}$$

$$y_t = Ch_t + Dx_t \tag{4}$$

离散化后的参数$\bar{A}$和$\bar{B}$计算如下:

$$\bar{A} = \mathrm{e}^{\Delta A} \tag{5}$$

$$\bar{B} = (\Delta A)^{-1}\left(\mathrm{e}^{\Delta A} - I\right)B \tag{6}$$

通过利用矩阵指数, 该公式能够在固定采样率下支持稳定且高效的数值计算. 这种递归动态可以被整合到深度学习架构中, 通过循环神经网络 (recurrent neural network, RNN) 实现.

此外, 通过矩阵幂级数将式 (3) 和式 (4) 进行迭代展开 (假设初始状态$h_0 = 0$), 可以观察到:

$$\begin{cases} h_1 = \bar{B}x_1 \\ h_2 = \bar{A}h_1 + \bar{B}x_2 = \bar{A}\bar{B}x_1 + \bar{B}x_2 \\ h_3 = \bar{A}h_2 + \bar{B}x_3 = \bar{A}^2\bar{B}x_1 + \bar{A}\bar{B}x_2 + \bar{B}x_3 \\ \cdots \end{cases}$$

将此展开式代入式 (4), 任意时刻k的输出y_k都可以表示为对所有历史输入$x_1, \cdots, x_k$的加权求和, 其权重系数恰好构成了卷积核K的各个元素. 因此, SSM 模型可以被理解为一个具有特定结构 (由矩阵$\bar{A}$, $\bar{B}$, C 定义) 的深度卷积网络:

$$K \triangleq \left[C\bar{B}, C\bar{A}\bar{B}, \cdots, C\bar{A}^{L-1}\bar{B}\right] \tag{7}$$

$$y = x \otimes K \tag{8}$$

其中, $\otimes$表示卷积操作, K 代表结构化的卷积核, L 是输入序列的长度.

这种从递归到卷积的等价转换为利用快速傅里叶变换等并行计算技术来加速模型训练提供了可能, 也是 Mamba 等模型能够高效处理长序列的关键.

SSM 领域的最新进展催生了 Mamba 这一新颖架构. 它通过引入与输入相关的动态参数B、C和Δ, 实现了特征的自适应调整. 这一改进在保留传统 SSM 长序列建模能力的同时, 提升了模型的灵活性与上下文响应能力. Mamba 的设计从 S4 模型中汲取灵感, 并扩展了一种递归机制, 从而能够在图像复原任务中保留长期依赖性并激活更广阔的空间上下文. 此外, Mamba 的并行扫描算法提供了与卷积操作类似的显著计算优势, 支持快速且高效的训练. 这些特性使 Mamba 成为解决如图像 SR 等高复杂度任务的强大方案.

3 本文方法

本文提出的 MTSR 网络架构主要由 4 个核心部分组成: 浅层特征提取 (shallow feature extraction, SFE)、深层特征提取 (deep feature extraction, DFE)、TDSR 以及图像重建 (IR), 其整体结构如图 3 所示.

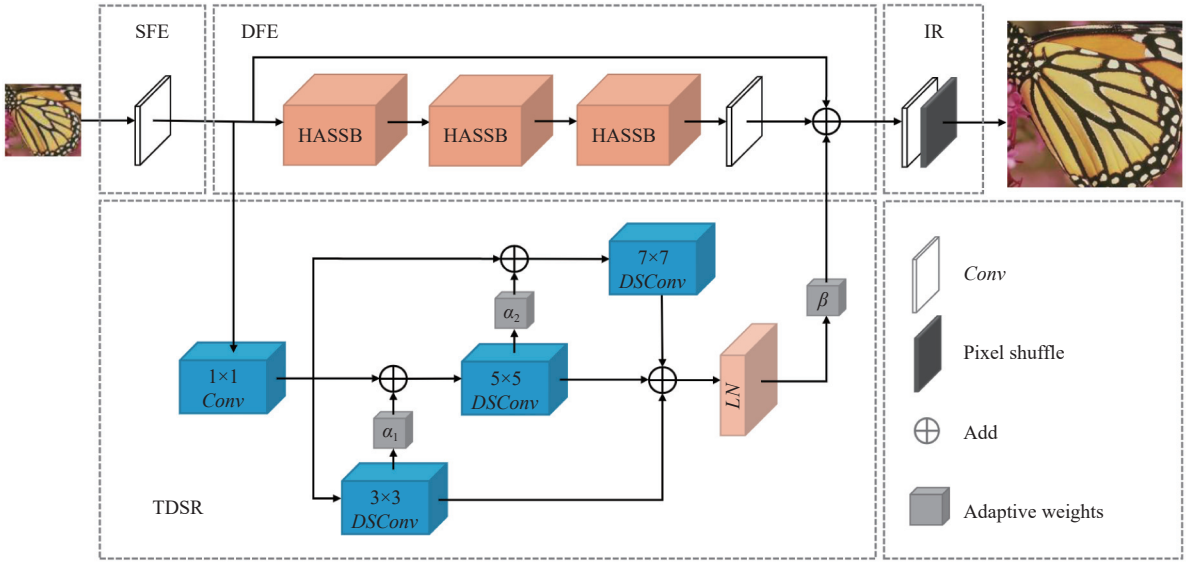


图 3 本文提出的 MTSR 的总体结构

对于一个给定的低分辨率输入图像$I_{LR} \in \mathbb{R}^{H\times W\times C_{in}}$, 模型首先通过 SFE 模块提取其初步特征$F_0 \in \mathbb{R}^{H\times W\times C}$, 该过程可表示为:

$$F_0 = H_{SFE}(I_{LR}) \tag{9}$$

其中, H_{SFE}代表浅层特征提取函数, 由一个单独的3×3卷积层实现. 提取到的特征F_0随后被送入 TDSR 模块进行精炼, 以获得增强后的浅层特征F_{SF}:

$$F_{SF} = H_{TDSR}(F_0) \tag{10}$$

DFE 模块由K个串联的 HASSB 构成, 其设计旨在捕获图像中复杂的高频结构. 该过程定义如下:

$$F_i = H_{HASSB_i}(F_{i-1}),\ i = 1,2,\cdots,K \tag{11}$$

其中, H_{HASSB_i}代表第i个 HASSB 模块的映射函数. 在最后一个 HASSB 模块的输出F_K之后, 应用一个额外的3×3卷积层, 生成最终的深层特征F_{DF}:

$$F_{DF} = Conv(F_K) \tag{12}$$

最后, 模型通过一个残差连接聚合浅层特征F_0、精炼浅层特征F_{SF}以及深层特征F_{DF}, 得到最终的融合特征F_{OUT}. 该特征被送入 IR 模块, 以生成最终的高分辨率输出图像:

$$F_{OUT} = F_0 + F_{SF} + F_{DF} \tag{13}$$

$$I_{HR} = H_{IR}(F_{OUT}) \tag{14}$$

3.1 三重深度可分离浅层细化模块 (TDSR)

以往的 SR 模型通常采用单个3×3卷积来提取浅层特征. 然而, 近期的研究表明, 丰富浅层特征的表征能力能够显著提升超分辨率图像的重建质量[20]. 受此启发, 本文提出了 TDSR 模块, 其结构如图 3 所示. 该模块利用具有不同感受野的卷积核, 对浅层特征进行多尺度、分层次的提取.

具体而言, TDSR 首先对输入特征图F_0应用一个1×1的逐点卷积, 以增强初始的通道间交互. 随后, 并行地采用 3 个核大小分别为3×3、5×5和7×7的深度可分离卷积, 以捕获多尺度的浅层特征. 虽然更大的卷积核能提供更广阔的感受野, 但也会引入更多的参数和计算开销. 为了平衡性能与效率, 本文采纳了深度可分离卷积的方案[21].

这 3 个深度可分离卷积层被组织在不同的分支中, 每个分支的输出经过逐级加权聚合, 生成中间特征FL_1、FL_2和FL_3. 然后将这些输出相加, 并通过层归一化 (layer normalization, *LN*) 进行处理, 以确保特征分布的一致性. 最后, 通过一个可学习的自适应权重β进行缩放, 得到 TDSR 最终的浅层特征输出F_{SF}:

$$FS_0 = Conv_{1\times1}(F_0) \tag{15}$$

$$FL_1 = DSConv_{3\times3}(FS_0) \tag{16}$$

$$FL_2 = DSConv_{5\times5}(FS_0 + \alpha_1 \cdot FL_1) \tag{17}$$

$$FL_3 = DSConv_{7\times7}(FS_0 + \alpha_2 \cdot FL_2) \tag{18}$$

$$F_{SF} = \beta \cdot LN(FL_1 + FL_2 + FL_3) \tag{19}$$

其中, FS_0表示对F_0应用1×1卷积后的输出, α_1和α_2是可学习的自适应权重. *Conv*代表标准卷积操作, 而*DSConv*代表深度可分离卷积.

3.2 混合注意力状态空间块 (HASSB)

本文所提出的 HASSB 模块的结构如图 4 所示. 先前在自然语言处理领域的研究表明, 仅依赖 Mamba 的模型可能存在上下文学习能力受限的问题[18], 研究者将其归因于 Mamba 在跨 token 信息交互方面的能力不足. 鉴于上下文信息在图像 SR 中对于精确恢复精细视觉细节同样至关重要, 此类限制可能会严重影响 SR 性能. 为解决此问题, 本文提出的 HASSB 模块融合了基于 Mamba 构建的 *VSSL* (vision state space layer) 和基于 Transformer 架构的 *MHSAL* (multi-head self-attention layer) 组件. 这种混合设计旨在利用 Transformer 强大的跨 token 信息交互能力, 来弥补 Mamba 的不足. 后续实验结果证明, 在相似参数量下, 这种混合结构优于以往纯 Mamba 或纯 Transformer 的 SR 网络.

每个 HASSB 模块包含N个由 *VSSL* 或 *MHSAL* 块组成的层. 对于一个给定的特征输入F_i, 每一层通过 *VSSL* 或 *MHSAL* 模块对其进行处理, 生成中间特征F_{i+1}, 然后将其送入下一层进行进一步的表征学习:

$$F_{i+1}=Oper(F_i),\ i=0,1,2,\cdots,N,\ Oper\in\{VSSL,MHSAL\} \tag{20}$$

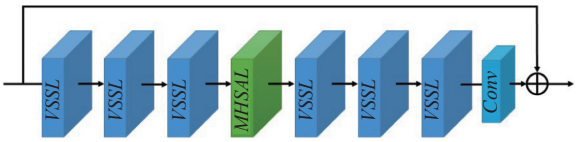
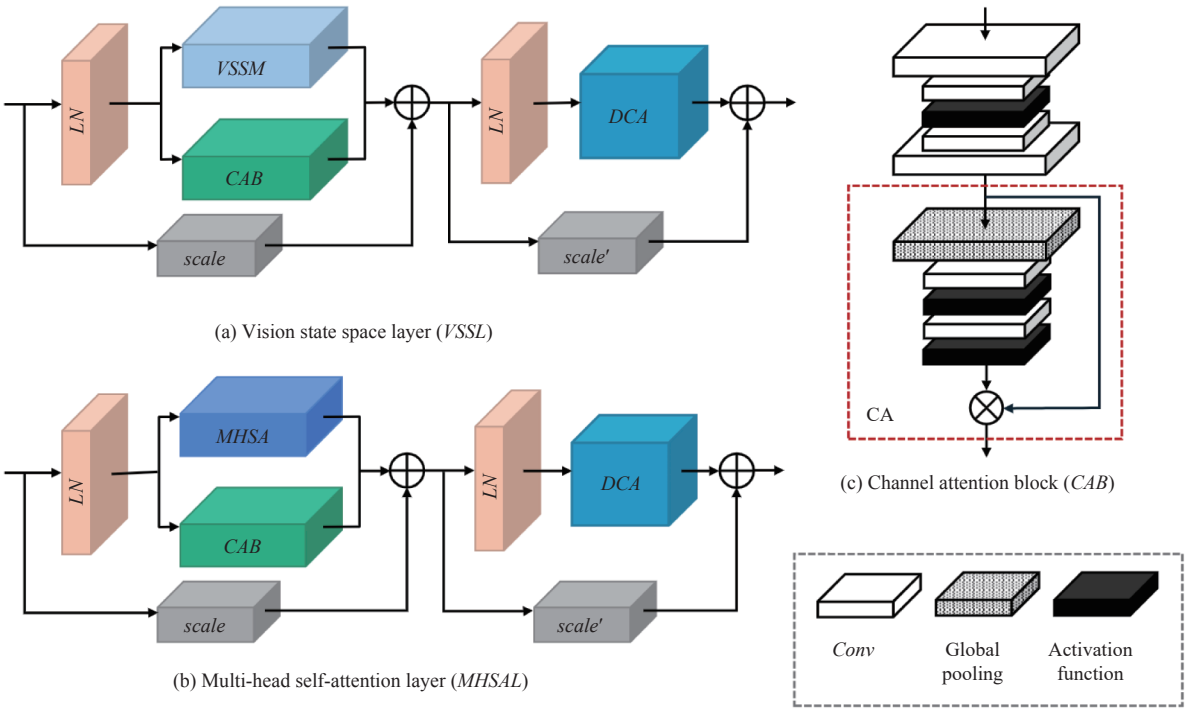


图 4 混合注意力状态空间块 (HASSB) 结构

如图 5(a) 所示, 对于给定的输入F_i, *VSSL* 首先应用 *LN* 对每个样本的激活值进行归一化. 归一化后的输出被传递到视觉状态空间模块 (vision state space module, *VSSM*), 该模块负责建模长程依赖关系. 与此同时, 一个通道注意力块 (channel attention block, *CAB*) 以 0.01 的固定权重并行操作, 用以强调与任务相关的通道, 从而增强网络关注判别性特征的能力[22]. 在残差连接中引入可学习的缩放因子$scale\in\mathbb{R}^C$, 以提升模型的稳定性和收敛速度. 该过程由式 (21) 所示:

$$M_i=VSSM(LN(F_i))+0.01\cdot CAB(LN(F_i))+scale\cdot F_i \tag{21}$$



(a) Vision state space layer (*VSSL*)

(b) Multi-head self-attention layer (*MHSAL*)

(c) Channel attention block (*CAB*)

图 5 *VSSL*、*MHSAL* 和 *CAB* 的结构

得到的M_i随后经过另一个 *LN* 层进行归一化, 并输入到本文提出的 *DCA* 模块中, 该模块旨在增强局部细节表征并促进跨通道信息交互. 在残差连接中使用了另一个可学习的缩放参数$scale'\in\mathbb{R}^C$, 如式 (22) 所示:

$$F_{i+1} = DCA(LN(M_i)) + scale' \cdot M_i \tag{22}$$

如图 5(b), *MHSAL* 块将 *VSSL* 流程中的 *VSSM* 替换为 *MHSA* 机制, 而架构的其余部分保持不变. 其相应的计算过程由式 (23) 定义:

$$M_i = MHSA(LN(F_i)) + 0.01 \cdot CAB(LN(F_i)) + scale \cdot F_i \tag{23}$$

如图 5(c) 所示, 本文进一步引入了一个轻量化的 *CAB* 模块, 以提高模型对关键特征的敏感度. 特征图经归一化后, 首先通过一个1×1卷积来降低通道维度. 随后采用一个3×3的分组深度卷积以降低计算成本, 接着是 GELU 激活函数和另一个1×1卷积, 以恢复原始通道数. 之后, 使用一个扩张率为 2 的扩张深度卷积来扩大有效感受野. 最后, 应用一个标准的通道注意力 (channel attention, CA) 模块对特征进行自适应调制[23].

3.3 视觉状态空间模块 (*VSSM*)

为了满足图像处理的需求, 同时保持线性时间复杂度, 本文引入了 *VSSM*[17]. *VSSM* 建立在离散化的状态空间建模之上, 将输入的特征图解释为随时间演化的序列, 其结构如图 6 所示.

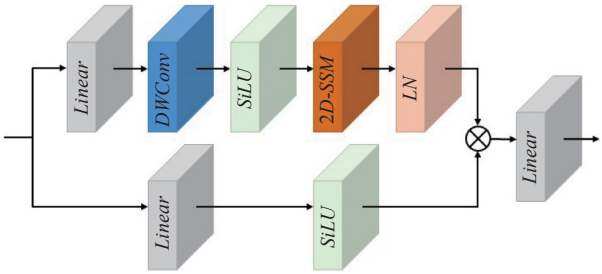


图 6 视觉状态空间模块的结构

为了更好地使基于 Mamba 的架构适应二维视觉数据, *VSSM* 采用了多向展平策略, 而非单一方向的扫描. 这种方法使得每个像素能够与其更广泛的邻域进行交互, 从而增强了上下文建模能力. 处理后的特征图计算如下:

$$X_1 = LN(2D\text{-}SSM(SiLU(DWConv(Linear(X))))) \tag{24}$$

$$X_2 = SiLU(Linear(X)) \tag{25}$$

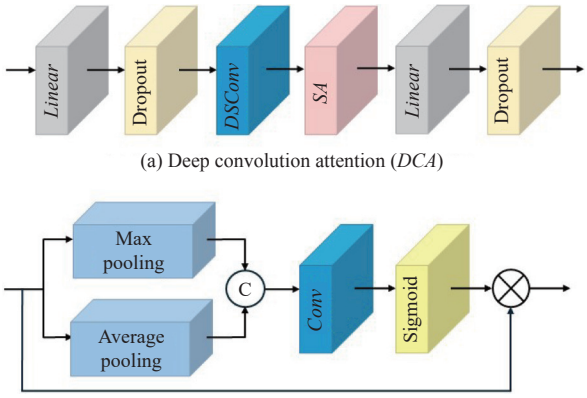
$$X_{out} = Linear(X_1 \odot X_2) \tag{26}$$

其中, *DWConv*指代深度卷积, *SiLU*表示激活函数, ⊙代表逐元素的哈达玛积. 该设计协同地结合了通过通道和空间操作实现的局部增强能力, 以及由二维状态空间扫描实现的全局建模能力. 因此, *VSSM* 特别适用于图像复原和其他高分辨率视觉任务.

3.4 深度卷积注意力前馈网络 (*DCA*)

近期研究表明, 在基于 Mamba 的架构中采用传统的 MLP 作为前馈网络, 通常会导致性能欠佳. 为解决这一局限, 本文提出了一种专为图像 SR 任务设计、并为匹配 Mamba 特性而量身定制的新型前馈网络模块 *DCA*.

如图 7(a) 所示, 对于一个给定的、经 *LN* 归一化的输入特征向量$L_{in} \in \mathbb{R}^C$, 一个全连接层首先将其投影到更高维度的空间, 从而丰富其表征能力. 随后是一个 Dropout 层, 通过随机失活部分神经元来减轻过拟合. 接着, 应用一个由深度卷积 (*DWConv*) 和逐点卷积 (*PWConv*) 构成的深度可分离卷积 (*DSConv*) 来捕获局部特征同时增强通道间的交互. 之后, 采用一个轻量化的空间注意力 (spatial attention, *SA*) 机制来进一步提炼局部表征并加强细粒度的空间建模. 第 2 个线性层再将特征投影回原始的通道维度C, 最后应用一个 Dropout 层产生输出L_{out}.



(a) Deep convolution attention (*DCA*)

(b) Spatial attention (*SA*)

图 7 本文提出的深度卷积注意力前馈网络结构

若忽略 Dropout 层, 该模块的前向计算可表示为:

$$L_{out} = Linear(DSConv(SA(Linear(L_{in})))) \tag{27}$$

如图 7(b) 所示, *SA* 模块在通道维度上同时执行最大池化和平均池化. 两个输出被拼接起来, 并通过一个不含偏置的7×7卷积层, 以保持特征的一致性. 然后, 一个 Sigmoid 激活函数生成空间注意力图, 该图通过逐元素相乘的方式应用于原始特征图, 从而增强空间上显著的区域. 整个 *SA* 过程可表示为:

$$X_{concat} = [\max(X, dim = 1), mean(X, dim = 1)] \tag{28}$$

$$X_{out} = X \odot \sigma(Conv_{7\times7}(X_{concat})) \tag{29}$$

其中, X表示输入特征图, $[\cdot,\cdot]$代表拼接操作, σ是 Sigmoid 激活函数, $\odot$代表逐元素乘法.

消融实验证实, *DCA* 模块显著改善了 SR 任务中的局部特征建模能力和整体性能, 其表现优于 Mamba 中传统的基于 MLP 的前馈网络设计. 此外, *DCA* 也能有效地集成到 *MHSAL* 中, 展现出强大的兼容性和灵活性.

3.5 图像重建 (IR)

近年来, SR 模型中的 IR 模块通常采用由卷积层后接像素重组 (pixel shuffle) 操作进行上采样的设计, 如图 3 所示. 然而, 随着图像超分辨率倍数的增加, 与像素重组配对的卷积层会导致模型参数量的大幅增长. 为在轻量化网络设计的约束下解决此问题, 本文进一步将 IR 模块中标准的3×3卷积缩减为更紧凑的1×1卷积.

针对重建性能的量化评估, 本文采用峰值信噪比 PSNR (peak signal-to-noise ratio) 和结构相似性指数 (structural similarity index measure, SSIM) 这两个指标, 计算均在 YCbCr 颜色空间的 Y 通道上进行. PSNR 通过量化重建图像与真实图像之间的差异来衡量像素级的保真度, 更高的 PSNR 值表示失真度更低. SSIM 则通过考量亮度、对比度和结构模式来评估感知相似度, 更高的 SSIM 分数意味着重建图像更好地保留了参考图像的感知质量. 结合使用 PSNR 和 SSIM 能够更全面地评估模型的重建能力. 如表 1 所示, 经过 50 个训练周期后的对比实验表明, 这一调整在重建性能损失可忽略的同时, 有效降低了参数开销.

表 1 IR 模块中不同卷积核大小对评价指标的影响 (50 epochs)

卷积核	参数量 (k)	BSD100 PSNR (dB)/SSIM	Urban100 PSNR (dB)/SSIM	Manga109 PSNR (dB)/SSIM
1×1	920	27.74/0.7425	26.69/0.8038	31.24/0.9169
3×3	943	27.74/0.7427	26.70/0.8040	31.23/0.9167

3.6 损失函数

为训练所提出的 MTSR 网络, 本文采用平均绝对误差 (mean absolute error, MAE), 即通常所说的L_1损失函数, 作为优化目标. 该损失函数衡量重建的 HR 图像与对应的真实图像之间的逐像素差异. 具体而言, 两幅图像首先被展平为向量, 然后计算预测图像I_{pred}和真实图像I_{true}之间的L_1范数, 如式 (30) 所定义:

$$L_1 = \frac{\left\| I_{\text{pred}} - I_{\text{true}} \right\|_1}{H \times W \times C} \tag{30}$$

其中, I_{pred}表示生成的 HR 图像, I_{true}表示真实图像. 符号$\|\cdot\|_1$代表L_1范数, 而H、W和C分别对应图像的高度、宽度和通道数.

4 实验及结果分析

4.1 数据集与实验设置

在训练阶段, 本文的 MTSR 模型使用了 DIV2K 数据集[24], 该数据集包含 1 000 张覆盖广泛场景的高质量图像. 其划分为 800 张训练图像、100 张验证图像和 100 张测试图像. 为了全面评估模型的性能, 本文采用 5 个被广泛使用的基准测试数据集: Set5[25]、Set14[26]、BSD100[27]、Urban100[28]和 Manga109[29].

在训练过程中, 输入的原始 LR 图像被裁剪为64×64的图像块. 本文遵循了标准的实践, 采用水平翻转和随机旋转的方式进行数据增强. 初始学习率设置为2×10^{-4}, 并在预设的训练节点处减半. 为了加速训练进程, 本文使用×2 缩放因子的超分辨率模型的权重来初始化×3 缩放因子和×4 缩放因子的模型, 并相应地减半初始学习率和总训练迭代次数[30]. 模型优化采用了 Adam 优化器[31], 其中β_1=0.9, β_2=0.99. 所有训练实验均在配置有两块 NVIDIA GeForce RTX 4090 GPU 的 Ubuntu 22.04 系统上完成.

4.2 模型性能对比

表 2 展示了本文提出的 MTSR 网络与多种代表性图像 SR 方法的详细性能对比, 其中包括经典方法以及近期的 SOTA 轻量化方法, 如 CARN[32]、A^2F[33]、Swin-IR-light[12]、MambaIR-light[17]和 SRFormer-light[13].

值得注意的是, 在参数量相当的情况下, 对于×3 缩放因子的超分辨率任务, MTSR 在 Manga109 数据集上的 PSNR 值相较于基于 Transformer 的当前 SOTA 模型 SRFormer-light 和基于 Mamba 的 MambaIR-light, 分别取得了 0.31 dB 和 0.38 dB 的提升. 在×4 缩放因子的超分辨率任务中, MTSR 在 Manga109 数据集上的性能更比 MambaIR-light 高了 0.44 dB. 此外, 得益于对 IR 组件的优化设计, 与其他模型相比, MTSR 的参数量随着超分辨率倍数的增加呈现出更缓慢的增长趋势. 这些结果表明, MTSR 不仅在同等复杂度的轻量化 SR 模型中达到了 SOTA 性能, 而且在更高的超分辨率倍数下展现出更为显著的性能优势.

本文中所有报告的浮点运算次数 (FLOPs) 均是基于将 LR 图像超分辨率重建至1280×720目标分辨率下计算的.

表 2 不同 SR 方法在×2、×3、×4 缩放因子下的参数量、FLOPs、PSNR 和 SSIM

模型	缩放因子	参数量 (k)	FLOPs (G)	Set5	Set14	BSD100	Urban100	Manga109
				PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
CARN	×2	1 592	222.8	37.76/0.959 0	33.52/0.916 6	32.09/0.897 8	31.92/0.925 6	38.36/0.976 5
A^2F		1 363	306.1	38.09/0.960 7	33.78/0.919 2	32.23/0.900 2	32.46/0.931 3	38.95/0.977 2
SwinIR-light		910	244.4	38.14/0.961 1	33.86/0.920 6	32.31/0.901 2	32.76/0.934 0	39.12/0.978 3
MambaIR-light		905	334.2	38.13/0.961 0	33.95/0.920 8	32.31/0.901 3	32.85/0.934 9	39.20/0.978 2
SRFormer-light		853	236.3	38.23/0.961 3	33.94/**0.920 9**	32.36/0.901 9	32.91/0.935 3	39.28/0.978 5
MTSR (本文)		918	297.4	**38.25/0.961 5**	**33.97/0.920 9**	**32.37/0.902 1**	**33.10/0.936 4**	**39.44/0.978 7**
CARN	×3	1 592	118.8	34.29/0.925 5	30.29/0.840 7	29.06/0.803 4	28.06/0.849 3	33.50/0.944 0
A^2F		1 367	136.3	34.54/0.928 3	30.41/0.843 6	29.14/0.806 2	28.40/0.857 4	33.83/0.946 3
SwinIR-light		918	110.8	34.62/0.928 9	30.54/0.846 3	29.20/0.808 2	28.66/0.862 4	33.98/0.947 8
MambaIR-light		913	148.5	34.63/0.928 8	30.54/0.845 9	29.23/0.808 4	28.70/0.863 1	34.12/0.947 9
SRFormer-light		861	105.4	34.67/0.929 6	30.57/0.846 9	29.26/0.809 9	28.81/0.865 5	34.19/0.948 9
MTSR (本文)		919	131.0	**34.74/0.930 1**	**30.67/0.848 4**	**29.32/0.810 9**	**29.02/0.868 8**	**34.50/0.950 2**
CARN	×4	1 592	90.9	32.13/0.893 7	28.60/0.780 6	27.58/0.734 9	26.07/0.783 7	30.47/0.908 4
A^2F		1 374	77.2	32.32/0.896 4	28.67/0.783 9	27.62/0.737 9	26.32/0.793 1	30.72/0.911 5
SwinIR-light		930	63.6	32.44/0.897 6	28.77/0.785 8	27.69/0.740 6	26.47/0.798 0	30.92/0.915 1
MambaIR-light		924	84.6	32.42/0.897 7	28.74/0.784 7	27.68/0.740 0	26.52/0.798 3	30.94/0.913 5
SRFormer-light		873	62.8	32.51/0.898 8	28.82/0.787 2	27.73/0.742 2	26.67/0.803 2	31.17/0.916 5
MTSR (本文)		920	77.7	**32.57/0.900 0**	**28.90/0.789 2**	**27.77/0.743 8**	**26.80/0.807 0**	**31.38/0.918 6**

4.3 可视化质量对比

如图 8 所示, 在 Urban100 数据集上, 对 MTSR 与几种 SOTA 轻量化图像 SR 方法在×4 缩放因子下的重建质量进行了可视化比较. 在 img004 中, MTSR 更优地保留了 LR 输入中的原始结构形状, 而其他方法则产生了可见的扭曲和形变. 在 img024 和 img067 中, MTSR 在恢复精细纹理方面表现卓越, 生成了更清晰的图案和更锐利的边缘细节, 而对比方法更容易产生模糊或涂抹的纹理. 对于包含复杂纹理图案的 img092, 一些基线模型出现了明显的伪影或结构性错误, 而 MTSR 则提供了更准确且视觉上更连贯的结果.

总而言之, 与其他方法相比, MTSR 能够从 LR 输入中稳定地重建出更高质量的 HR 图像.

4.4 消融实验

为验证模型各组件的有效性, 本文比较了不同前馈网络模块的性能, 并评估了多种网络架构在 SR 任务中的表现. 所有消融实验均在相同的网络深度、训练参数和训练周期下进行, 以确保公平对比.

为了评估所提出的 TDSR 模块在×4 缩放因子下的实际效果, 本文进行了消融研究, 其结果如表 3 所示. 结果表明, TDSR 通过其多尺度、多层次的特征提取机制, 有效地捕获了图像中更多样化的低频信息和边缘轮廓, 为后续的 IR 模块提供了更高质量的输入, 从而改善了最终的 SR 效果. 这证明了对浅层特征进行精细化处理对轻量化 SR 模型而言是一项高收益的策略.

为验证所提出的 *DCA* 前馈模块的有效性, 本文在×4 超分辨率设置下进行了对比实验, 将其与使用传统 MLP 前馈网络的基线模型进行了比较. 如表 4 所示, 结果表明与不含前馈网络以及使用 MLP 作为前馈网络的变体相比, 集成 *DCA* 模块后, 模型性能分别获得了高达 0.29 dB 和 0.16 dB 的提升. 此外, *DCA* 在所有测试数据集上均一致地取得了最佳性能, 证实了简单的 MLP 缺乏对图像局部空间信息的感知能力, 其全连接形式的变换无法有效捕获像素邻域内的精细结构. 而 *DCA* 通过引入深度可分离卷积和空间注意力, 显式地将局部上下文归纳偏置融入特征变换中, 与 Mamba 的全局建模能力形成了互补, 从而实现了更优的整体性能.

为了进一步分析 *DCA* 的性能增益来源, 本文通过逐一移除其子模块的方式进行了额外的消融分析, 结果如表 5 所示. 结果表明, 单独移除任一组件均会导致明显的性能下降, *DCA* 的优越性并非来自单一组件, 而是源于深度卷积提供的局部上下文建模能力与空间注意力带来的特征动态聚焦能力之间的协同作用.

为了探究单独使用 Mamba 在图像 SR 任务中的潜在局限性, 并验证引入 Transformer 组件能否有效解决这些问题, 本文进行了深入的实验, 将提出的 MTSR 与一个纯 Mamba 模型 (其中所有的 *MHSAL* 模块均被替

换为 *VSSL* 模块) 进行比较. 表 6 的实验结果显示, MTSR 的性能比其纯 Mamba 对应版本高出多达 0.28 dB, 并在所有基准测试数据集上均获得更高的分数. 这些数据充分表明在混合 SR 框架中同时集成 Transformer 和 Mamba 组件的优越性.

本文还研究了在 Mamba-Transformer 架构中改变 Transformer 模块比例对重建性能的影响, 结果如表 7 和图 9 所示.

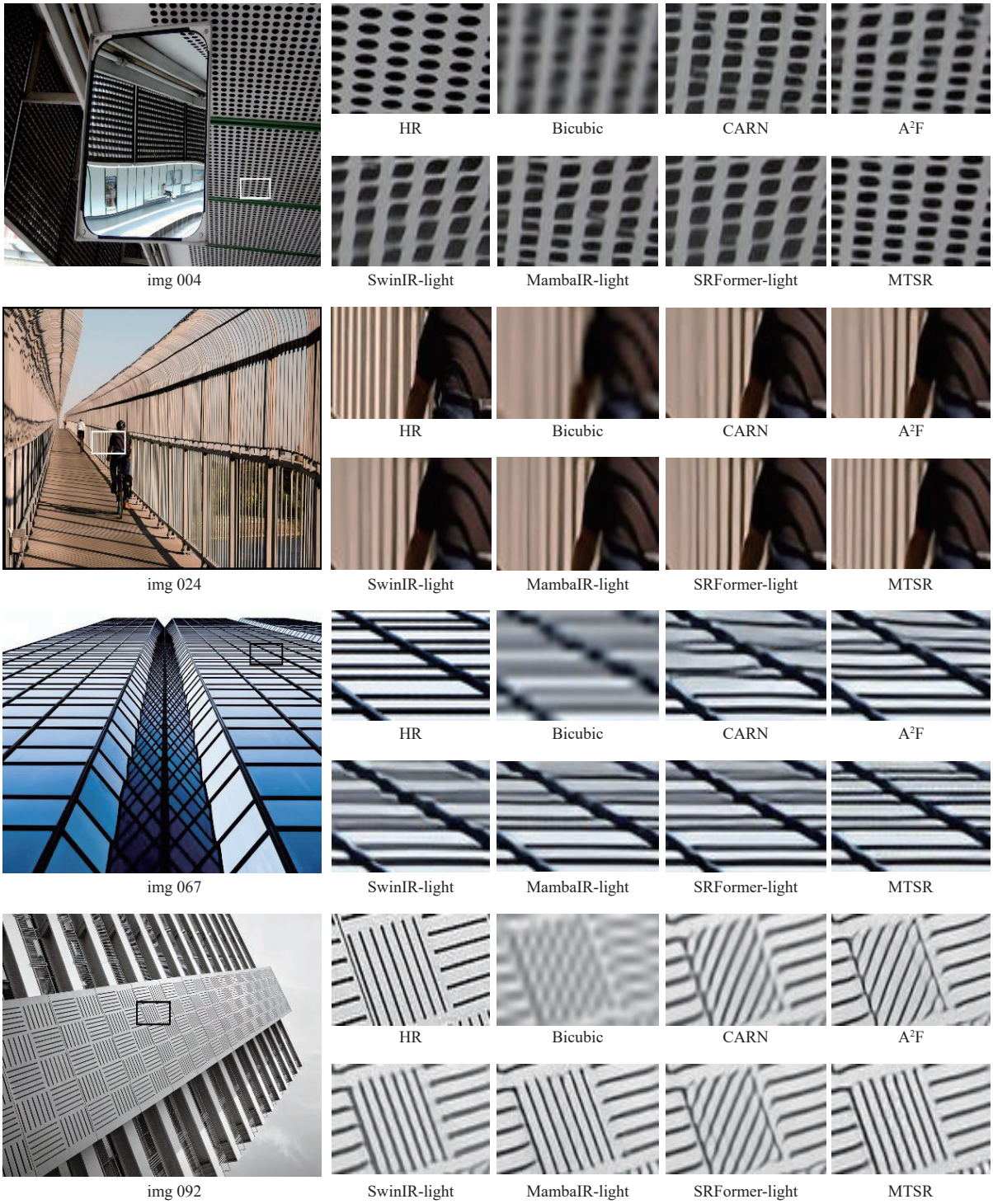


图 8 本文提出的 MTSR 和最近的 SOTA 轻量化图像 SR 方法在 Urban100 数据集上的视觉重建质量比较 (×4)

表 3 MTSR 包含和不包含 TDSR 在×4 时 SR 的比较

模型	BSD100	Urban100	Manga109
	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
w/o TDSR	27.75/0.7431	26.73/0.8054	31.30/0.9174
MTSR	27.77/0.7438	26.80/0.8070	31.38/0.9186

表 4 MTSR 使用不同 FFN 设置在×4 时 SR 的比较

模型	BSD100	Urban100	Manga109
	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
w/o FFN	27.70/0.7429	26.51/0.8051	31.12/0.9179
w/ MLP	27.73/0.7434	26.64/0.8062	31.19/0.9180
w/ *DCA*	27.77/0.7438	26.80/0.8070	31.38/0.9186

表 5 *DCA* 模块逐一移除子模块后在×4 时 SR 的比较

模型	BSD100	Urban100	Manga109
	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
w/o *DWConv*	27.72/0.7419	26.58/0.8013	31.16/0.9161
w/o *PWConv*	27.72/0.7419	26.59/0.8016	31.17/0.9159
w/o SA	27.75/0.7429	26.70/0.8048	31.28/0.9175
MTSR	27.77/0.7438	26.80/0.8070	31.38/0.9186

表 6 MTSR 包含和不包含 *MHSAL* 在×4 时 SR 的比较

模型	BSD100	Urban100	Manga109
	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
w/o *MHSAL*	27.76/0.7430	26.52/0.8050	31.16/0.9181
MTSR	27.77/0.7438	26.80/0.8070	31.38/0.9186

表 7 MTSR 使用不同 *MHSAL* 模块比例在×4 时 SR 的比较

模块比例	BSD100	Urban100	Manga109
(%)	PSNR (dB)/SSIM	PSNR (dB)/SSIM	PSNR (dB)/SSIM
0	27.76/0.7430	26.52/0.8050	31.16/0.9181
14.3 (本文)	27.77/0.7438	26.80/0.8070	31.38/0.9186
42.9	27.74/0.7424	26.63/0.8024	31.23/0.9170
71.4	27.70/0.7413	26.49/0.7989	31.00/0.9149
100	27.62/0.7386	26.20/0.7902	30.63/0.9107

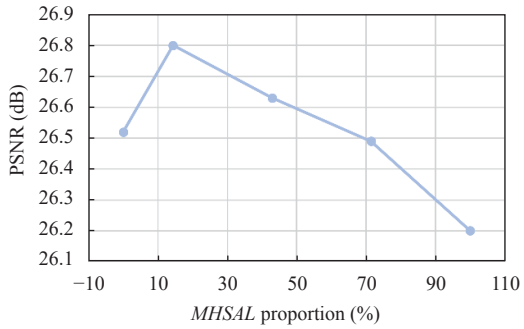


图 9 在 Urban100 数据集上, 评估×4 缩放因子下不同 *MHSAL* 模块比例对 SR 任务的 PSNR 性能影响

通过综合分析表 6、表 7 和图 9, 可以清晰地看到, 与完全由 Mamba 或 Transformer 单一模块构建的模型相比, 这种混合架构提供了更显著的优势. 并且当 Transformer 模块的比例相对较低时 (约 14.3%), 模型达到了最佳性能, 而随着该比例的增加, 性能反而逐渐下降. 基于这些发现, 本文得出结论: 在基于 Mamba 的 SR 模型中, 恰当地集成少量 Transformer 模块, 可以有效弥补其在 token 级交互方面的不足, 从而显著提升整体性能.

5 结论

本文针对图像超分辨率任务, 提出一种名为 MTSR 的新型轻量化模型. 本模型的核心工作在于设计了一种混合注意力状态空间块 (HASSB), 其采用 Mamba 与 Transformer 的混合架构, 既保留了 Mamba 架构固有的高效推理能力, 又充分利用了 Transformer 卓越的跨 token 信息交互优势. 为解决在 Mamba 模型中直接采用 MLP 前馈网络所导致的性能欠佳问题, 本文进一步提出了一种深度卷积注意力 (*DCA*) 前馈模块, 该模块显著增强了模型对局部特征的表征能力. 值得一提的是, *DCA* 模块同样能与基于 Transformer 的框架完全兼容, 展现了良好的设计灵活性. 此外, 为了更有效地利用浅层特征, 本文还设计了三重深度可分离浅层细化模块 (TDSR), 该模块通过采用多尺度深度可分离卷积, 实现了高效的多尺度浅层特征提取.

充分的定量和定性实验结果均表明, MTSR 的综合性能超越了当前的 SOTA 轻量化超分辨率模型. 未来的研究工作可进一步探索 MTSR 模型在更大规模数据集上的可扩展性, 及其在其他底层视觉任务中的适应潜力.

参考文献

1 Li JH, Pei ZH, Li WJ, *et al*. A systematic survey of deep learning-based single-image super-resolution. ACM Computing Surveys, 2024, 56(10): 249. [doi: 10.1145/3659100]

2 Lepcha DC, Goyal B, Dogra A, *et al*. Image super-resolution: A comprehensive review, recent trends, challenges and applications. Information Fusion, 2023, 91: 230–260. [doi: 10.1016/j.inffus.2022.10.007]

3 Dong C, Loy CC, He KM, *et al*. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(2): 295–307. [doi: 10.1109/TPAMI.2015.2439281]

4 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE

Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

5 Kim J, Lee JK, Lee KM. Deeply-recursive convolutional network for image super-resolution. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1637–1645.

6 Zhang YL, Tian YP, Kong Y, *et al*. Residual dense network for image super-resolution. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2472–2481. [doi: 10.1109/CVPR.2018.00262]

7 Dai T, Cai JR, Zhang YB, *et al*. Second-order attention network for single image super-resolution. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11057–11066. [doi: 10.1109/CVPR.2019.01132]

8 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

10 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021. 1–21.

11 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

12 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. Montreal: IEEE, 2021. 1833–1844. [doi: 10.1109/ICCVW54120.2021.00210]

13 Zhou YP, Li Z, Guo CL, *et al*. SRFormerV2: Taking a closer look at permuted self-attention for image super-resolution. arXiv:2303.09735, 2023.

14 Xia B, Zhang YL, Wang SY, *et al*. DiffIR: Efficient diffusion model for image restoration. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 13049–13059.

15 Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752, 2023.

16 Shi Y, Xia B, Jin XY, *et al*. VmambaIR: Visual state space model for image restoration. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(6): 5560–5574. [doi: 10.1109/TCSVT.2025.3530090]

17 Guo H, Li J, Dai T, *et al*. MambaIR: A simple baseline for image restoration with state-space model. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 222–241.

18 Waleffe R, Byeon W, Riach D, *et al*. An empirical study of mamba-based language models. arXiv:2406.07887, 2024.

19 Gendy G, He GH, Sabor N. Lightweight image super-resolution based on deep learning: State-of-the-art and future directions. Information Fusion, 2023, 94: 284–310. [doi: 10.1016/j.inffus.2023.01.024]

20 Chen YT, Xia RL, Yang K, *et al*. MFFN: Image super-resolution via multi-level features fusion network. The Visual Computer, 2024, 40(2): 489–504. [doi: 10.1007/s00371-023-02795-0]

21 Ding XH, Zhang XY, Han JG, *et al*. Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11953–11965. [doi: 10.1109/CVPR52688.2022.01166]

22 Chen XY, Wang XT, Zhou JT, *et al*. Activating more pixels in image super-resolution Transformer. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22367–22377.

23 Zhang YL, Li KP, Li K, *et al*. Image super-resolution using very deep residual channel attention networks. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 294–310. [doi: 10.1007/978-3-030-01234-2_18.

24 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. [doi: 10.1109/CVPRW.2017.149]

25 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. Guildford: BMVC, 2012. 135.1–135.10. [doi: 10.5244/C.26.135]

26 Zeyde R, Elad M, Protter M. On single image scale-up using

sparse-representations. Proceedings of the 7th International Conference on Curves and Surfaces. Avignone: Springer, 2012. 711–730. [doi: 10.1007/978-3-642-27413-8_47]

27 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. [doi: 10.1109/ICCV.2001.937655]

28 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. Boston: IEEE, 2015. 5197–5206. [doi: 10.1109/CVPR.2015.7299156]

29 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]

30 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. [doi: 10.1109/CVPRW.2017.151]

31 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015. 1–15.

32 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.

33 Wang XH, Wang Q, Zhao YZ, *et al*. Lightweight single-image super-resolution network with attentive auxiliary feature learning. Proceedings of the 15th Asian Conference on Computer Vision. Kyoto: Springer, 2020. 268–285.

(校对责编: 张重毅)