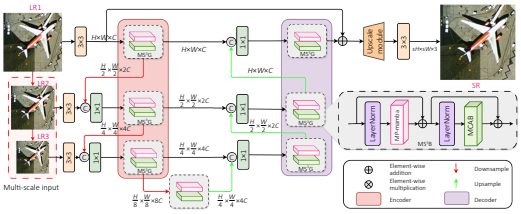
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 4 月, 第 53 卷, 第 4 期

基于多尺度增强的状态空间模型的遥感图像超分辨率重建



陈嘉诚[1], 吴　菲[1*], 屠杭垚[2], 蒋嘉伟[3], 王万良[3]

[1]嘉兴大学人工智能学院，浙江 嘉兴 314001；

[2]浙江大学计算机科学与技术学院，浙江 杭州 310015；

[3]浙江工业大学计算机科学与技术学院，浙江 杭州 310023

摘要：卷积神经网络 (Convolutional neural networks, CNN) 与视觉 Transformer (vision transformers, ViTs) 在遥感图像超分辨率任务中均展现出卓越性能。ViTs 凭借其强大的长距离依赖建模能力，通常优于传统 CNN 方法，但其计算复杂度随图像分辨率呈二次增长，严重制约了在高分辨率遥感图像重建中的实际应用。为应对这一挑战，本文提出了一种多尺度增强状态空间模型 (multi-scale augmented state space model, MS^{3}M)。与现有基于固定扫描路径的特征提取方法不同，MS^{3}M 引入了一种高效分组并行扫描策略，在维持线性计算复杂度的同时，有效建模全局与非局部特征依赖。针对遥感图像中固有的多尺度空间结构，本文进一步在状态空间模型中嵌入多感受野聚合机制，以融合不同尺度的上下文信息。此外，为增强特征表达能力，我们还提出一种高阶矩通道亲和力调制模块，用于优化局部特征表示。整个网络基于 U 型架构实现多层次特征融合，在多个公开遥感数据集上的实验结果表明，MS^{3}M 在 PSNR、SSIM 和 LPIPS 等定量指标与视觉质量上均显著优于现有先进方法，验证了所提方法的有效性与先进性。

关键词：遥感图像超分辨率；状态空间模型；分组并行扫描；多感受野聚合；高阶矩统计量

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250304 | CSTR: 32245.14.oee.2026.250304 | 中图分类号：TP751 | 文献标识码：A

陈嘉诚，吴菲，屠杭垚，等. 基于多尺度增强的状态空间模型的遥感图像超分辨率重建 [J]. 光电工程，2026，**53**(4): 250304

Chen J C, Wu F, Tu H Y, et al. Remote sensing image super-resolution reconstruction via multi-scale augmented state space model[J]. *Opto-Electron Eng*, 2026, **53**(4): 250304

1　引　言

高分辨率遥感影像 (High-resolution remote sensing image, HRRSI) 记录了丰富的地面细节，为大范围、细粒度的应用提供了重要基础，其获取对实现精细对地观测与遥感分析具有重要意义[1-2]。HRRSI 广泛应用于土地覆盖分割[3]、图像分类[4]、目标检测[5-6] 与变化检测[7] 等任务。当前，随着数据量的持续增长，对高质量遥感图像的需求日益增强。然而，成像设备 (如高空无人机、遥感卫星) 通常距地面数千米以上，所获取的目标分辨率有限。此外，大气散射与光照变化等复杂成像环境也进一步影响了图像质量。更重要的是，为缓解传输不稳定性并节约成本，遥感图像 (remote sensing image, RSI) 常经历严重压缩与下采样，造成关键信息丢失，尤其是边缘与纹理等高频细节[8]。上述因素制约了高层次语义任务的处理效果，削弱了 HRRSI 的实际应用价值。

收稿日期：2025-10-10；修回日期：2026-01-13；录用日期：2026-01-14；网络出版日期：2026-04-25

基金项目：国家自然科学基金资助项目 (62506337)；浙江省自然科学基金资助项目 (LQN25F020024); 嘉兴市公益项目 (2025CGZ007)

*通信作者：吴菲，WFMOOK@163.com

因此，从低分辨率 (low-resolution, LR) 观测中重建高分辨率图像，对于提升人类感知能力与扩展应用至关重要[9-10]。获取 HRRSI 最直接的方法是升级硬件设备以提高采集质量与传输稳定性，但该方法成本高昂、难以推广，且增大相对孔径需复杂光学系统支持。此外，高空卫星设备在发射后难以进行后续升级。相比之下，遥感图像超分辨率重建 (remote sensing single image super resolution, RSSISR) 技术提供了一种灵活且经济高效的替代方案，其旨在从 LR 图像中重建出纹理清晰、信息丰富的高分辨率 (high resolution, HR) 图像[11-12]。正因如此，RSSISR 已成为工业界与学术界的研究热点。

早期研究多依赖手工先验 (如插值法[13]、基于先验信息的方法[14]) 解决 RSSISR 中的不适定问题。然而，这类方法难以准确建模复杂退化过程，重建精度有限，且优化耗时。近年来，基于深度神经网络的方法在 RSSISR 任务中取得显著进展，性能明显优于传统方法，代表性模型包括 CNN 与 Transformer。Dong 等[15] 首次将仅含三层卷积的 SRCNN 应用于单图像超分辨率任务，取得突破。此后，基于 CNN 的模型凭借强大的局部拟合能力，几乎主导了 RSSISR 领域。例如，FeNet[16] (feature enhancement network) 引入晶格结构构建非线性特征提取模块以提升重建性能；Chen 等[17] 采用分解-融合策略，设计残差聚合与分组注意力融合模块，增强跨通道特征交互与多尺度信息感知能力；MSFFTAN[18] (multi-scale fast Fourier transform based attention network) 则通过多输入 U 形架构与局部-全局通道注意力机制，提升模型对空间尺度变化的建模能力。尽管 CNN 方法凭借卷积的局部连通性与平移不变性，能有效捕捉图像局部特征，但其感受野有限，难以聚合长程上下文与非局部特征，限制了其对遥感图像广泛空间特征的高效建模。此外，与自然图像相比，遥感图像场景更复杂、多样性更高、尺度变化更显著，方向性特征差异也更突出，进一步加大了重建难度。

近年来，源自自然语言处理 (natural language processing, NLP) 领域的 Transformer 架构[19] 受到广泛关注，并在多个高级视觉任务中展现出卓越性能。值得注与 CNN 方法相比，Transformer 中的多头自注意力机制在构建全局依赖关系方面具有显著优势，能够有效捕获并聚合遥感图像中的广泛空间特征。因此，研究者们开始探索 Transformer 在遥感图像超分辨率重建中的应用潜力。例如，Lei 等[20] 为克服传统方法在高维特征提取方面的不足，基于多阶段 Transformer 模块提出了一种新颖框架，有效提升了高维特征表征能力。随后，ESTNet[21](efficient swin transformer net) 引入分组自注意力模块，显著降低了全局注意力机制的计算开销。然而，尽管 Transformer 表现出优异的性能，现有方法仍存在一些局限性：首先，多数方法未能充分挖掘原始低分辨率图像中丰富的多尺度特征；其次，基于窗口划分的自注意力机制难以有效建模不同窗口间的关联关系；更重要的是，注意力机制的计算复杂度与图像像素数量呈二次方增长，导致其难以直接处理高分辨率遥感图像。

为解决上述问题，本文提出多尺度增强状态空间模型 (multi-scale augmented state space model, MS^{3}M)。该方法基于自然语言处理领域新兴的 Mamba 架构[22]——一种具有线性计算复杂度与高效长程依赖建模能力的状态空间模型，可替代计算复杂的多头自注意力机制。为适配二维图像的非因果特性，在二维选择性扫描基础上，提出分组并行多方向扫描机制 (图 1)，显著降低计算开销。针对状态空间模型存在的局部特征遗忘与多尺度建模能力不足问题，引入多感受野聚合模块，充分挖掘遥感图像中的多尺度上下文信息。此外，针对高维特征同质化问题，设计了高阶矩引导的通道亲和力调制模块，

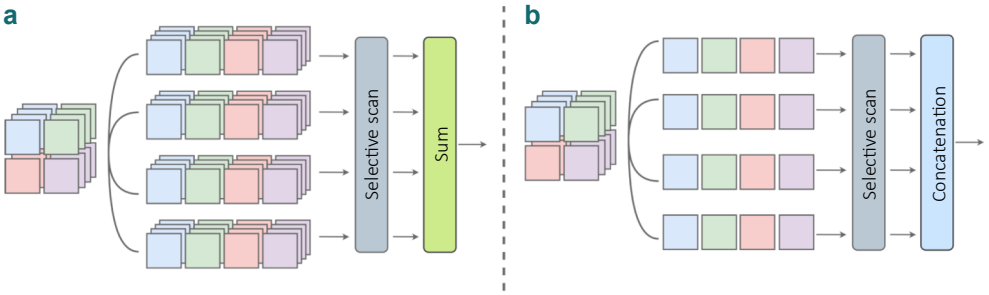


图 1 原始 Mamba 模型与本文提出的并行扫描策略对比示意图。(a) 原始扫描机制；(b) 并行扫描机制

Fig. 1 Scanning strategy in (a) the original Mamba model versus (b) the proposed parallel scanning strategy

增强特征的鲁棒性与异构性，提升对局部细节的感知能力。最终，将上述模块嵌入经典 U 型网络，构建出一个高效鲁棒的遥感图像超分辨率重建模型。

本文的主要贡献点包括以下四点：

1) 引入并行 Mamba 模块，通过分组多方向扫描策略来增强状态空间模型中的计算效率和交互性，实现更全面的空间感知和对局部和全局信息的有效建模。

2) 针对状态空间模型在结构上存在的空间尺度不敏感问题，引入了多尺度感受野聚合模块，旨在有效挖掘遥感图像中的多尺度特征，从而提升模型对多尺度信息的感知能力。

3) 提出高阶矩引导的通道亲和力调制模块，通过增强通道间交互并动态抑制特征冗余，有效减弱噪声通道的干扰，进而提升了特征聚合的效能与鲁棒性。

4) 将上述模块无缝集成于 U 形架构中，构建 MS^{3}M。在两个常用的遥感图像标准数据集上的大量实验验证表明，MS^{3}M 在超分辨率重建任务中性能明显优于当前主流方法。

2 相关工作

2.1 自然图像超分辨率重建

近年来，CNN 在 SISR 领域取得显著进展。Dong 等[15] 首次将 CNN 引入 SISR 任务，提出 SRCNN 模型，重建性能显著优于传统方法。随后，一系列基于 CNN 的超分方法相继涌现。Kim 等通过增加网络深度，构建了具有 20 层卷积的 VDSR (very deep super resolution) 模型[23]，性能明显优于浅层 SRCNN，表明适当加深网络有助于提升模型能力。在此基础上，研究者开始构建更深、更宽或结构更复杂的网络。Lim 等[24] 通过移除残差网络中的批归一化层[25] 等组件，优化结构并降低计算复杂度，提出约 50 层的增强型深度超分网络。然而，该方法对所有特征采取统一处理，且难以有效捕获长距离依赖，导致精细纹理等高频细节重构能力不足。

为平衡不同特征对重建的贡献，注意力机制被引入 CNN 超分模型，如通道注意力[26](channel-wise attention, CA) 和整体注意力[27](holistic attention, HA)。Zhang 等[28] 构建了超过 400 层的深度网络，并引入通道注意力以自适应调整通道间特征响应。然而，卷积算子存在两个固有局限：1) 感受野有限；2) 平移不变性。这限制了 CNN 建模全局依赖与动态适配输入内容的能力。为更好利用图像中的全局先验 (如自相似性)，研究者提出自相似注意力与非局部稀疏注意力模块。例如，Mei 等[29] 将非局部运算与稀疏表示结合，提出计算效率更高的非局部稀疏注意力；Chen 等[30] 则提出转置稀疏自注意力机制以增强局部特征感知能力。为进一步恢复丢失的高频信息与细节，分段卷积前馈网络也被引入模型设计。尽管这些方法在一定程度上提升了全局特征建模能力，但其密集的非局部计算仍带来较高计算复杂度和训练开销，在处理大空间分辨率遥感图像超分任务时尤为突出，严重制约了实际应用与部署。

2.2 遥感图像超分辨率重建

遥感图像在地理测绘与环境监测等领域具有重要应用价值[31]。然而，受卫星与传感器性能的限制，所获取的图像往往分辨率较低，因此需借助超分辨率处理技术以提升其细节表现与数据准确性。随着该领域逐渐受到广泛关注，近年来，学者们在遥感图像超分辨率重建方法上取得了显著进展[32]。与早期传统方法相比，基于深度学习的超分辨率方法在效率、精度和鲁棒性方面均展现出明显优势。LGCNet[33](Local-global combined network) 作为首个基于 CNN 的遥感图像超分辨率模型，通过结合局部与全局表征能力，以局部模块提取细节信息，全局模块捕获整体结构与上下文信息，从而实现图像重建。Dong 等提出了 SMSR[34]，该网络采用密集采样策略[35]，增强了图像细节与复杂特征的捕捉能力，并融合了一阶与二阶多尺度特征。近年来，注意力机制被广泛引入遥感超分辨率任务。例如，Lei 和 Shi 提出了 HSENet[36](hybrid-scale self-similarity exploitation net)，它利用 RSI 之间的单尺度和跨尺度自相似性来增强特征表示。Wang 等[37] 通过注意力机制和多层次特征融合实现了轻量化设计。Kang 等提出了 ESTNet[21](efficient swin transformer net)，包含高效的通道注意力模块和一个精心设计的分组注意力模块。DRCAN[38](Dual-resolution connected attention network) 采用了双分辨率学习结构和连接注意力机制，以充分利用不同尺度分辨率的特征信息。Li 等[39] 引入了局部化上下文感知生成分区对抗网络，通过结合全局和局部自注意力机制和双区域判别器来增强图像细节，进一步生成符合人眼视觉感知的重构遥感图像。Xiao 等[8] 则针对自注意力机制计

算开销大的问题，提出了一种基于 Top-k 筛选策略的 Transformer 模型，该模型融合了多尺度特征聚合与全局上下文注意力，进一步优化了局部与全局特征的综合利用。

2.3 状态空间模型

状态空间模型 (State space models, SSMs)[40] 起源于经典控制理论[41]，近年来被引入深度学习领域，成为状态转换建模的重要框架。其在长距离依赖建模中表现出的线性计算复杂度特性，引起了研究者的广泛关注。结构化状态空间序列模型 (Structured state-space sequence model, S4)[42] 是深度状态空间模型在远程依赖建模方面的开创性工作。随后，S5 模型[43] 在 S4 基础上引入了多输入多输出 SSM 架构和高效的并行扫描算法。Fu 等提出的 H3 模型[44] 取得了突破性进展，几乎弥合了自然语言处理中 SSMs 与 Transformers 之间的性能差距。Zhang 等[45] 基于门控激活函数的最新进展，对 S4 模型进行改进，提出了门控状态空间层，不仅提升了模型性能，还展现出对更长序列的零样本泛化能力。最近提出的 Mamba 模型[46] 通过引入选择机制和高效的硬件设计算法，构建了数据依赖型 SSM 架构，在自然语言处理任务中性能超越 Transformers，同时保持与序列长度的线性计算复杂度。这一突破性成果激发了研究者将其应用于视觉任务的兴趣。文献 [47] 通过对图像块进行前向与后向双向处理，提出了高效的视觉 Mamba 模块，显著提升了空间信息捕获能力。VMamba[48](Visual Mamba) 创新性地采用交叉扫描与交叉融合技术，有效聚合空间信息。MambaIR[49](Mamba image resolution) 集成了视觉状态空间模块与改进的 MLP 模块，并引入通道注意力机制以缓解原始 Mamba 中的局部像素遗忘和通道冗余问题。VMambaIR[50](Visual Mamba image resolution) 通过并行执行通道扫描与四向空间扫描，提出全向选择性扫描模块，以同时利用空间和通道信息。尽管现有研究显示了视觉 Mamba 在捕获图像长距离依赖与细节上下文方面的优势，但其在 RSSISR 任务中的潜力仍需深入探索，其作为该任务有效解决方案的价值也有待进一步验证。

3 方法描述

3.1 主体架构

为了探索状态空间模型在遥感图像重建任务的潜力，并解决繁重的计算开销和局部像素不感知问题，引入了 MS^{3}M 网络，一个为图像恢复量身定制的框架（具体流程如算法 1 所示）。所提模型由浅层特征提取模块、深度语义特征提取模块和图像重建模块组成，采用经典的 U 形状架构以高效感知多尺度特性，具体结构如图 2 所示。首先，给定低分辨率遥感图像$\boldsymbol{I}_{\mathrm{LR}} \in \mathbb{R}^{H\times W\times 3}$，浅层特

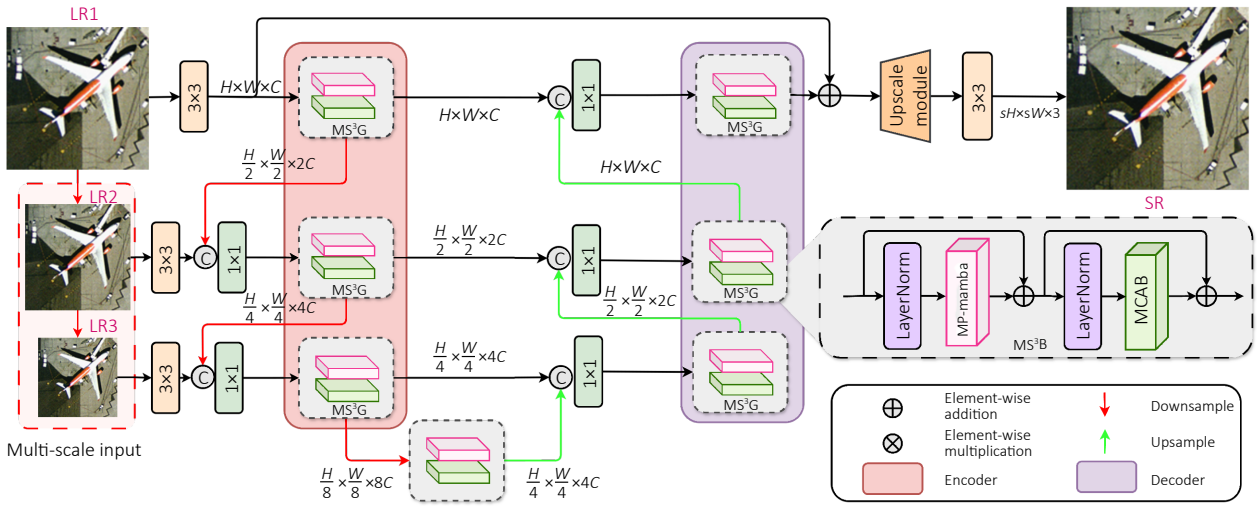


图 2 MS^{3}M 的总体架构和各模块结构

Fig. 2 Overall architecture and network structures of MS^{3}M

征提取模块使用一个3×3 卷积层从低分辨率图像中提取浅层特征 $\boldsymbol{F}_{\mathrm{s}} \in \mathbb{R}^{H\times W\times C}$，将特征从低维 RGB 空间转换到高维特征表示空间。其中，H 和 W 分别表示输入图像 $\boldsymbol{I}_{\mathrm{LR}}$的高和宽，$C$ 则表示特征通道的数量。然后浅层特征 $\boldsymbol{F}_{\mathrm{s}}$ 被输入到有三层编码器和解码器、组成的深度特征提取模块，以生成深度语义特征 $\boldsymbol{F}_{\mathrm{d}} \in \mathbb{R}^{H\times W\times C}$。在编码器和解码器的每一级中都有不同数量的 MS^{3}B 模块，以及用下采样或者上采样的卷积层。并且遵循先前的方法[18]，采用多输入机制在提升模型多尺度聚合能力的同时减轻训练难度提升模型稳定性。在此过程中，编码器在扩展通道的同时逐渐降低分辨率。此外，将低分辨率退化图像通过一个 3×3 卷积操作合并到主路径中，并进行拼接操作，再通过1×1 卷积调整通道。然后，将得到的最深特征输入 ($\boldsymbol{F}_{\mathrm{l}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$)到一个三层解码器中，通过逐步恢复特征到原来的大小以产生具有丰富潜在高频信息的特征。在此过程中，将解码器特征与编码器特征串联以辅助重建，并使用 1×1 卷积将通道数减少一半。最后利用由亚像素卷积层和 3×3 卷积层组成的重建模块对深度特征 $\boldsymbol{F}_{\mathrm{d}} \in \mathbb{R}^{H\times W\times C}$ 和 浅层特征 $\boldsymbol{F}_{\mathrm{s}} \in \mathbb{R}^{H\times W\times C}$ 进行图像重建获得输出图像 $\boldsymbol{I}_{\mathrm{SR}} \in \mathbb{R}^{sH\times sW\times 3}$。其中 sH 和 sW 分别表示输出重建遥感图像的高和宽，s 代表缩放因子。

MS^{3}B 作为 MS^{3}M 的基本组成部分，被集成在编码器、解码器、瓶颈层阶段，实现高效的并行扫描策略并融合了整体空间信息。如图 2 所示，每个 MS^{3}B 包含两个主要组件：用于空间信息混合的多尺度感知并行 Mamba 模块 (MP-Mamba) 和用于通道特征细化的高阶矩引导的通道亲和力调制模块 (MCAB)。MS^{3}B 在每个组件之前使用层标准化 (layer normalization, LN)，残差连接将输出整合到前面的输入中：

$$\boldsymbol{F}' = \boldsymbol{F} + \boldsymbol{H}_{\mathrm{MP\_Mamba}}\left(\boldsymbol{H}_{\mathrm{LN}}(\boldsymbol{F})\right), \tag{1}$$

$$\boldsymbol{F}'' = \boldsymbol{F}' + \boldsymbol{H}_{\mathrm{MCAB}}\left(\boldsymbol{H}_{\mathrm{LN}}(\boldsymbol{F}')\right), \tag{2}$$

式中：$\boldsymbol{F}$表示输入特征；$\boldsymbol{F}'$ 和 $\boldsymbol{F}''$ 分别是多尺度感知并行 Mamba 模块和高阶矩引导的通道亲和力调制模块的输出；$\boldsymbol{H}_{\mathrm{MP\_Mamba}}(\cdot)$ 和 $\boldsymbol{H}_{\mathrm{MCAB}}(\cdot)$ 分别代表多尺度感知并行 Mamba 模块和高阶矩引导的通道亲和力调制模块；$\boldsymbol{H}_{\mathrm{LN}}(\cdot)$ 是用于特征层归一化的操作算子。在该架构中，$\boldsymbol{H}_{\mathrm{MP\_Mamba}}(\cdot)$ 侧重于捕获输入特征之间的长距离空间依赖关系，而 $\boldsymbol{H}_{\mathrm{MCAB}}(\cdot)$ 则对这些特征进行细化以增强表示。

算法 1：基于多尺度增强的状态空间模型的遥感图像超分辨率重建	
输入	低分辨率遥感图像 $\boldsymbol{I}_{\mathrm{LR}} \in \mathbb{R}^{H\times W\times 3}$ 和缩放因子s
输出	超分辨率重建遥感图像 $\boldsymbol{I}_{\mathrm{SR}} \in \mathbb{R}^{sH\times sW\times 3}$
多尺度输入构建	对输入低分辨率遥感图像 $\boldsymbol{I}_{\mathrm{LR}}$通过逐级下采样获得 LR2 和 LR3，使网络同时感知不同尺度的纹理与结构信息。 $\mathrm{LR1}=\boldsymbol{I}_{\mathrm{LR}}, \mathrm{LR2}=Down_{\downarrow 2}(\mathrm{LR1}), \mathrm{LR3}=Down_{\downarrow 2}(\mathrm{LR2})$
浅层特征提取	将输入图像从RGB 空间映射为高维潜在空间。 $\boldsymbol{F}_{\mathrm{s}}^{1}=\boldsymbol{H}_{\mathrm{Conv}}(LR1), \boldsymbol{F}_{\mathrm{s}}^{2}=\boldsymbol{H}_{\mathrm{Conv}}(LR2), \boldsymbol{F}_{\mathrm{s}}^{3}=\boldsymbol{H}_{\mathrm{Conv}}(LR3)$
多尺度编码器	各尺度低分辨图像分别输入多尺度增强的状态空间模型(MS³G)编码模块，实现空间通道特征协同聚合。 $\boldsymbol{F}_{\mathrm{e}}^{1}=\boldsymbol{H}_{\mathrm{MS^3G}}^{1}\left(\boldsymbol{F}_{\mathrm{s}}^{1}\right)$ $\boldsymbol{F}_{\mathrm{e}}^{2}=\boldsymbol{H}_{\mathrm{MS^3G}}^{2}\left(\boldsymbol{H}_{\mathrm{Conv}}\left(\left[\boldsymbol{F}_{\mathrm{s}}^{2}, \mathrm{Down}_{\downarrow 2}\left(\boldsymbol{F}_{\mathrm{e}}^{1}\right)\right]\right)\right)$ $\boldsymbol{F}_{\mathrm{e}}^{3}=\boldsymbol{H}_{\mathrm{MS^3G}}^{3}\left(\boldsymbol{H}_{\mathrm{Conv}}\left(\left[\boldsymbol{F}_{\mathrm{s}}^{3}, \mathrm{Down}_{\downarrow 2}\left(\boldsymbol{F}_{\mathrm{e}}^{2}\right)\right]\right)\right)$
瓶颈层	$\boldsymbol{F}_{\mathrm{b}}=\boldsymbol{H}_{\mathrm{MS^3G}}(\boldsymbol{F}_{\mathrm{e}}^{3})$
多尺度解码器	通过卷积实现多尺度特征双向信息交互融合，以强化纹理细节及全局语义关联。解码器采用多尺度增强的状态空间模型(MS³G)解码模块以增强长程依赖建模、恢复结构纹理。 $\boldsymbol{F}_{\mathrm{d}}^{3}=\boldsymbol{H}_{\mathrm{MS^3G}}^{3}\left(\boldsymbol{H}_{\mathrm{Conv}}\left(\left[\mathrm{UP}_{\uparrow 2}(\boldsymbol{F}_{\mathrm{b}}), \boldsymbol{F}_{\mathrm{e}}^{3}\right]\right)\right)$ $\boldsymbol{F}_{\mathrm{d}}^{2}=\boldsymbol{H}_{\mathrm{MS^3G}}^{2}\left(\boldsymbol{H}_{\mathrm{Conv}}\left(\left[\mathrm{UP}_{\uparrow 2}\left(\boldsymbol{F}_{\mathrm{d}}^{3}\right), \boldsymbol{F}_{\mathrm{e}}^{2}\right]\right)\right)$ $\boldsymbol{F}_{\mathrm{d}}^{1}=\boldsymbol{H}_{\mathrm{MS^3G}}^{1}\left(\boldsymbol{H}_{\mathrm{Conv}}\left(\left[\mathrm{UP}_{\uparrow 2}\left(\boldsymbol{F}_{\mathrm{d}}^{2}\right), \boldsymbol{F}_{\mathrm{e}}^{1}\right]\right)\right)$
超分辨率重建	最终经上采样模块和卷积层生成高分辨率重建遥感图像。 $\boldsymbol{I}_{\mathrm{SR}}=H_{\mathrm{Conv}}\left(Upscale\left(\boldsymbol{F}_{\mathrm{d}}+\boldsymbol{F}_{\mathrm{s}}^{1}\right)\right)$ Return $\boldsymbol{I}_{\mathrm{SR}}$

3.2 多尺度感知并行 Mamba 模块

并行 Mamba 算子：考虑到引言中的动机，目标是设计原始 Mamba 模型的一种变体，它既具有计算效率，又能有效地建模输入序列的空间依赖关系。考虑到 Mamba 在处理通道数 C 较大的输入序列时计算效率较低，借鉴分组卷积的思想，提出了一种并行 Mamba (parallel-Mamba, P-Mamba) 操作算子 (如图 3(a) 所示)。该操作算子将输入通道划分为多个并行组，并分别对每个组应用视觉单方向选择扫描模块 (visual single direction selective scan block, VSDSSB) 模块挖掘全局空间信息。具体而言，将输入通道划分为四个并行组，每组通道大小为 $C/4$，并对每个并行组应用独立的 VSDSSB 模块。因此，所提出的并行 Mamba 算子通过将通道拆分为更小的组，有效提升了模型的计算效率。为更好地建模输入数据中的空间依赖性，四个并行分组中的每一组会按四个不同方向对输入进行扫描：从左至右、从右至左、从下到上以及从上到下，如图 3(b) 所示。设$G=4$为表示四个扫描方向的组数：从左到右、从右到左、从上到下和从下到上。由输入序列 $\boldsymbol{X}_{\mathrm{in}}$生成四个序列，分别标记为$\boldsymbol{X}_{\mathrm{LR}}$、$\boldsymbol{X}_{\mathrm{RL}}$、$\boldsymbol{X}_{\mathrm{TB}}$和 $\boldsymbol{X}_{\mathrm{BT}}$，每个序列的尺寸为 $(B,H,W,C/4)$，分别对应前述四个方向之一。随后将这些不同方向的序列展平，形成张量形状为

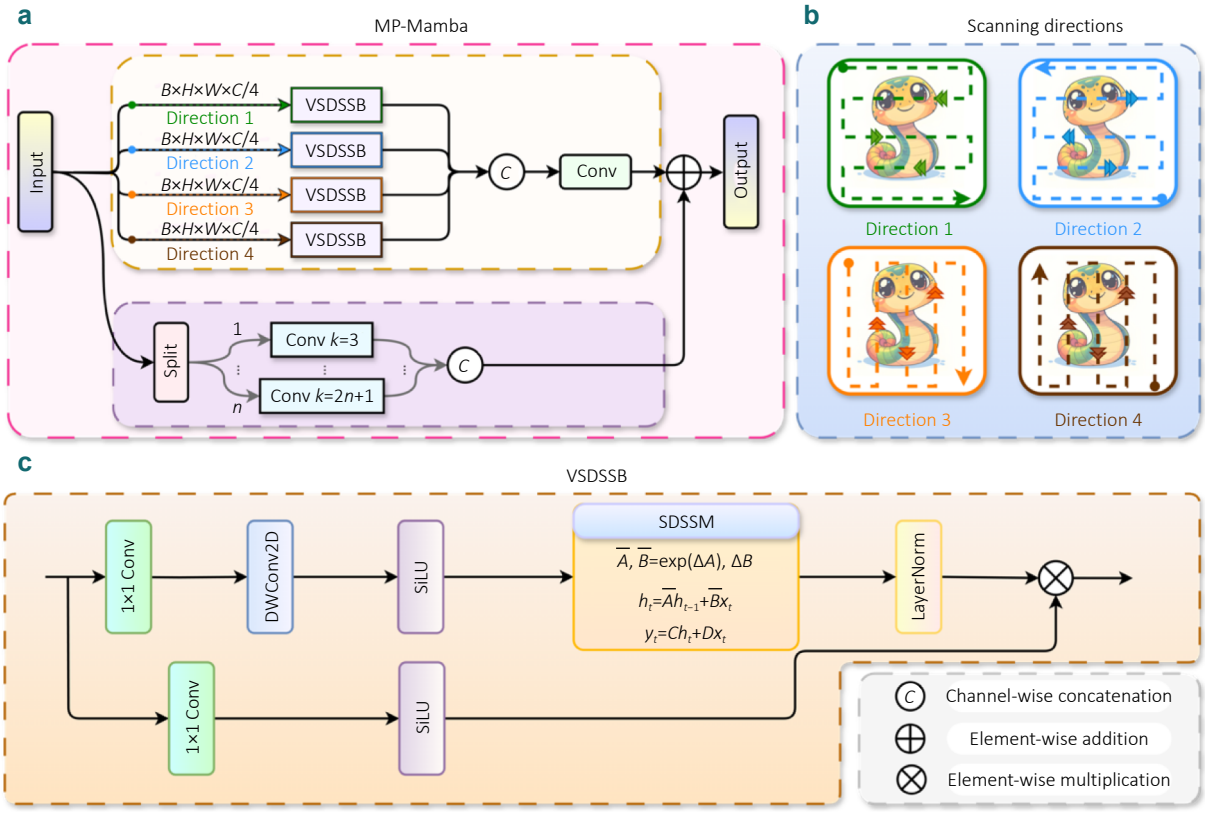


图 3 Mamba 模块。(a) 多尺度感知并行 Mamba 模块架构图；(b) Mamba 模块中的四种不同方向的扫描策略示意图；(c) 视觉单方向选择扫描模块

Fig. 3 Mamba module. (a) Architecture of the multi-scale perception parallel Mamba module; (b) Four scanning strategies of the Mamba module; (c) Structure of the unidirectional selective scanning module

$(B,N,C/4)$ 的单一序列，其中 $N=W\times H$ 表示序列中的像素数量。四个并行分组对应的参数可分别由 θ_{LR}、θ_{RL}、θ_{TB} 和 θ_{BT} 指定，每组参数代表对应方向 VSDSSB 模块的可学习参数。根据上述定义，并行 Mamba 算子的整体关系式可由如下公式所示：

$$\begin{aligned}\boldsymbol{X}_{\mathrm{P\_Mamba}} =&\boldsymbol{H}_{\mathrm{P\_Mamba}}\left(\boldsymbol{X}_{\mathrm{in}},\Theta\right)=\boldsymbol{H}_{\mathrm{Conv}}\left(\boldsymbol{H}_{\mathrm{Concat}}\left(\boldsymbol{H}_{\mathrm{VSDSSB}}\right.\right.\\&\left(\boldsymbol{X}_{\mathrm{LR}},\theta_{\mathrm{LR}}\right),\boldsymbol{H}_{\mathrm{VSDSSB}}\left(\boldsymbol{X}_{\mathrm{RL}},\theta_{\mathrm{RL}}\right),\boldsymbol{H}_{\mathrm{VSDSSB}}\\&\left.\left.\left(\boldsymbol{X}_{\mathrm{TB}},\theta_{\mathrm{TB}}\right),\boldsymbol{H}_{\mathrm{VSDSSB}}\left(\boldsymbol{X}_{\mathrm{BT}},\theta_{\mathrm{BT}}\right)\right)\right)\end{aligned}\tag{3}$$

式中：$\boldsymbol{X}_{\mathrm{in}}$ 和 $\boldsymbol{X}_{\mathrm{P\_Mamba}}$ 分别代表并行 Mamba 模块的输入和输出特征；$\boldsymbol{H}_{\mathrm{VSDSSB}}(\cdot)$ 表示视觉单方向选择扫描模块；$\boldsymbol{X}_{\mathrm{LR}}$、$\boldsymbol{X}_{\mathrm{RL}}$、$\boldsymbol{X}_{\mathrm{TB}}$ 和 $\boldsymbol{X}_{\mathrm{BT}}$ 分别表示在各自方向上扫描的输入张量；θ_{LR}、θ_{RL}、θ_{TB}、θ_{BT} 分别表示每个方向的 VSDSSB 模块的可学习参数；$\boldsymbol{H}_{\mathrm{P\_Mamba}}(\cdot)$ 是并行 Mamba 操作算子。每个并行 Mamba 算子的输出被重新调整为 $(B,H,W,C/4)$ 的形状，随后在通道维度拼接 $(\boldsymbol{H}_{\mathrm{Concat}}(\cdot))$ 并通过一个卷积层 $(\boldsymbol{H}_{\mathrm{Conv}}(\cdot))$ 进一步细化表示特征并重新恢复为 (B,H,W,C)。

多感受野聚合模块：遥感图像本质上就拥有多种不同空间尺度的图像，而且 Mamba 模型具有局部信息遗忘的本质缺陷。基于上述考虑，在并行 Mamba 算子中额外引入了多感受野聚合模块 (multi-receptive field aggregation module, MRFAM) 以弥补上述缺陷。该方法利用不同的感受野提取局部信息，实现多感受野交互。如图 3(a) 下分支所示，多感受野聚合模块利用不同的并行卷积算子处理给定输入特征 $\boldsymbol{X}_{\mathrm{in}}\in\mathbb{R}^{H\times W\times C}$，以提取多尺度局部信息并且实现多感受野交互。随后将这些特征划分为 N 部分。每个部分的特征 $\boldsymbol{X}_l^j\in\mathbb{R}^{H\times W\times\frac{C}{N}}$ 会经过具有不同卷积核尺寸的局部卷积操作。最终，将不同卷积操作的结果在通道维度进行拼接，形成最终的多尺度聚合输出特征 $\boldsymbol{X}_{\mathrm{multi\_scale}}\in\mathbb{R}^{H\times W\times C}$。上述操作可由如下公式表示：

$$\boldsymbol{X}_{\mathrm{out}}^j=\boldsymbol{H}_{\mathrm{Conv}}\left(\boldsymbol{X}_l^j,kernel=(2j+1)\right),j=1,2,\cdots N,\tag{4}$$

$$\boldsymbol{X}_{\mathrm{multi\_scale}}=\boldsymbol{H}_{\mathrm{Concat}}\left(\left[\boldsymbol{X}_{\mathrm{out}}^1,\boldsymbol{X}_{\mathrm{out}}^2,\cdots,\boldsymbol{X}_{\mathrm{out}}^N\right]\right),\tag{5}$$

式中：$\boldsymbol{X}_{\mathrm{out}}^j$ 表示卷积核大小为 $kernel=(2j+1)$ 的卷积算

子的输出；$\boldsymbol{H}_{\text{Concat}}(\cdot)$ 代表通道维度特征拼接操作。

视觉单方向扫描模块：如图 3(c) 所示，视觉单方向扫描模块通过两个平行分支 (上分支和下分支) 处理给定输入特征 $\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$。上分支通过 1×1 卷积将特征通道从原来的 C 扩展到 λC，其中 λ 表示预定义的通道膨胀因子。然后扩展的特征通过一系列连续操作：深度可分离卷积、Sigmoid 线性单元 (sigmoid linear unit, SiLU)、单方向选择扫描模块 (single direction selective scan module, SDSSM) 和 LN 层。下分支也通过 λ 进行通道膨胀，然后进行 SiLU 激活。来自两个分支的输出通过元素乘积结合，最终的 1×1 卷积线性投影操作将合并的输出降低到原始维度 C。完整的 VSDSSB 过程可以描述如下：

$$\boldsymbol{Y}=\boldsymbol{H}_{\text{LN}}\left(\boldsymbol{H}_{\text{SDSSM}}\left(\boldsymbol{H}_{\text{SiLU}}\left(\boldsymbol{H}_{\text{DWConv}}\left(\boldsymbol{H}_{\text{Conv}}\left(\boldsymbol{X}\right)\right)\right)\right)\right), \tag{6}$$

$$\boldsymbol{z}=\boldsymbol{H}_{\text{SiLU}}\left(\boldsymbol{H}_{\text{Conv}}\left(\boldsymbol{X}\right)\right), \tag{7}$$

$$\boldsymbol{x}_{\text{out}}=\boldsymbol{H}_{\text{Conv}}\left(\boldsymbol{Y}\otimes\boldsymbol{z}\right), \tag{8}$$

式中：$\boldsymbol{x}_{\text{out}}\in\mathbb{R}^{H\times W\times W}$ 表示 VSDSSB 模块的输出；⊗表示逐元素操作符；$\boldsymbol{H}_{\text{Conv}}(\cdot)$ 和 $\boldsymbol{H}_{\text{DWConv}}(\cdot)$ 分别表示 1×1 卷积和 3×3 深度可分离卷积；$\boldsymbol{H}_{\text{SiLU}}(\cdot)$ 代表 Sigmoid 线性单元激活函数；$\boldsymbol{H}_{\text{SDSSM}}(\cdot)$ 是单方向选择扫描模块操作算子。

3.3 高阶矩引导的通道亲和力调制模块

神经网络中，特征图的激活值可视为来自特定概率分布的样本。矩统计量为表征该分布提供了有效方法，其源自随机变量期望的泰勒级数展开。对于高斯分布，一阶矩 (均值) 反映集中趋势，二阶矩 (方差) 体现离散程度，三阶矩 (偏度) 表示分布不对称性等。从统计角度看，除一阶矩外，二阶矩和三阶矩在分布表征中常具关键作用，因此可用矩统计量近似表示网络中的特征图激活值。然而，仅使用一阶矩无法区分具有相同均值但不同方差 (由二阶矩体现) 的分布。低阶矩难以完整表征复杂的概率分布，而神经网络中特征图激活值的分布可能复杂多样。为全面准确地表示这类分布，需融合多阶矩信息。通过聚合多阶矩，可全面捕捉特征图激活值的分布特性，从而有效表示网络特征。基于上述分析，设计了高阶矩引导的通道亲和力调制模块 (high-order moment guided channel affinity modulation module, HMCAM) 架构。HMCAM 方法包含三个核心模块：1) 高阶矩统计量聚合；2) 高阶矩统计量融合；3) 通道重新校准。图 4 展示了 HMCAM 的整体框架。根据上述分析可知，矩统计量主要由一阶矩 $\boldsymbol{M}_1$、其他高阶矩及其组合构成。具体而言，一阶矩统计量 $\boldsymbol{M}_1=\boldsymbol{E}(\boldsymbol{X})$；而对于其他矩量 $(k\geq 2)$，其定义为 $\boldsymbol{M}_k(\boldsymbol{X})=\boldsymbol{E}\left((\boldsymbol{X}-\boldsymbol{E}(\boldsymbol{X}))^k\right)(k\geq 2)$。本文将矩统计量表示定义为其经验估计量。实际计算中仅计算与边缘分布相对应的矩统计量。具体来说，设 $\boldsymbol{X}$ 为具有概率分布 p 的有界随机样本, 则高阶聚合矩统计量 HAM_K 被定义为矩统计量表示的经验估计量，可由如下表达式表示：

$$HAM_K(p)=\alpha_1\|\varepsilon(\boldsymbol{X})\|_2+\sum_{k=2}^{K}\alpha_k\|\boldsymbol{M}_k(\boldsymbol{X})\|_2\,, \tag{9}$$

式中：$\boldsymbol{E}(\boldsymbol{X})$ 表示样本 $\boldsymbol{X}$ 的经验期望：$\boldsymbol{M}_k(\boldsymbol{X})=\boldsymbol{E}((\boldsymbol{X}-\boldsymbol{E}(\boldsymbol{X}))^k)$ 涵盖 $\boldsymbol{X}$ 的所有 k 阶中心矩；参数 K 用于限制计算的矩统计量数量；$\alpha_k\in(0,1)$ 为对应的权重参数。由于矩量阶数的上界由参数 K 限定，且边缘分布的矩统计量项存在严格随阶数增加而递减的上界，即高阶矩统计量的边际效用会随着阶数增加而衰减。因此，获取每个通道的一阶矩 $\boldsymbol{M}_1$、二阶矩 $\boldsymbol{M}_2$ 与三阶矩 $\boldsymbol{M}_3$ 作为通道表征，其输出维度为 $\mathbb{R}^{1\times 3\times C}$，从而实现分布近似精度与计算效率之间的最优平衡。高阶统计量融合方法旨在解决广泛矩统计量聚合 HAM_K 存在的维度问题，并能有效捕获通道中低阶与其他阶矩统计量的信息。具体来说，采用交叉矩统计量卷积技术，利用通道维度的一维卷积层融合多个矩量，最终生成维度为 $\mathbb{R}^{1\times 1\times C}$ 的特征 $\boldsymbol{F}_{\text{HAM}}$。通过上述矩统计量融合技术，实现了广泛的矩统计量特征 ($\boldsymbol{M}_1$、$\boldsymbol{M}_2$、$\boldsymbol{M}_3$) 自适应聚合。最后使用一个 Sigmoid 函数将上述高阶矩统计量特征转换为注意力权重来指导模型关注更重要的特征。

4 实验结果

4.1 实验设置

4.1.1 数据集说明

为评估所提方法的性能，实验选用两个遥感图像超分辨率领域广泛使用的公开数据集 (见表 1)：UCMerced LandUse[51] 与 AID[52]。所有高分辨率 (HR) 图像在 MATLAB 中通过双三次插值进行指定比例下采样，生成对应的低分辨率 (LR) 样本。

UCMerced LandUse 数据集包含 21 类典型地物场景 (如农田、飞机、棒球场、建筑等)，每类 100 张图像，共 2100 张。图像尺寸为 256 pixel×256 pixel，空间分辨率为 0.3 m/pixel。遵循该领域常见划分方式，将数据集分为训练集与测试集，并另从训练集中抽取 20% 样本作

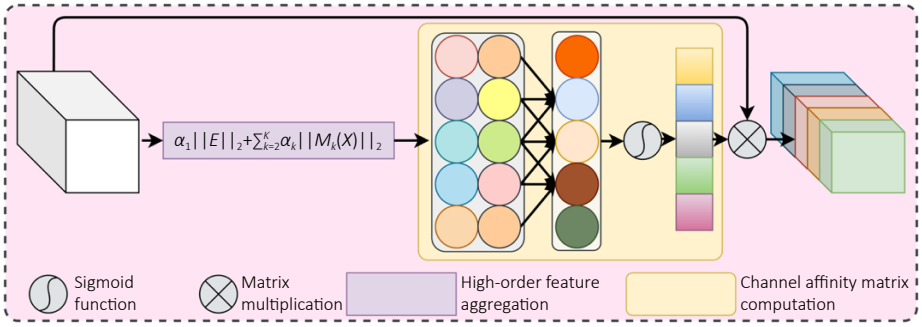


图 4 高阶矩引导的通道亲和力调制模块

Fig. 4 High-order moment guided channel affinity modulation module

表 1 实验所用数据集介绍

Table 1 Description of the experimental dataset

Dataset	每个类别数量	类别数	总图片数	空间分辨率/m	图分辨率/pixel
UCMerced LandUse	100	21	2100	0.3	256 × 256
AID	~300	30	10000	0.5	600 × 600

为验证集。

AID 为大规模航空影像数据集，涵盖来自卫星与无人机等多种平台的图像，广泛用于分类、检测与分割任务。该数据集共包含 30 个场景类别、10000 张图像，覆盖机场、建筑、车辆、住宅区等多种地物。图像尺寸为 600 pixel×600 pixel，空间分辨率为 0.5 m/pixel。参考 TransENet[20] 的划分方式，随机选取 80% 数据作为训练集，其余为测试集。此外，从每类场景中随机抽取 5 张图像构建验证集 (共 150 张)，以保持类别分布均衡。

4.1.2 评估指标

本文采用图像超分辨率领域广泛使用的峰值信噪比 (peak signal-to-noise ratio, PSNR) 和结构相似性指数 (structural similarity index measure, SSIM)[53] 作为图像重建质量的评价指标，所有超分辨率结果均在 RGB 空间上进行评估。

PSNR 作为评估图像保真度的常用指标，测量参考图像和重建图像之间的像素级差异。假设给定给定超分辨率重建图像 X_{SR} 和相对应的高分辨率真实图像 X_{HR}，PSNR 定义如下：

$$PSNR(X_{\mathrm{SR}},X_{\mathrm{HR}})=10\cdot\log_{10}\frac{max(X_{\mathrm{HR}})^2}{\|X_{\mathrm{SR}}-X_{\mathrm{HR}}\|_2^2}, \tag{10}$$

式中：$max(X_{\mathrm{HR}})$ 表示图像像素的最大可能取值；$\|X_{\mathrm{SR}}-X_{\mathrm{HR}}\|_2^2$ 表示均方误差 (mean square error, MSE)。PSNR 值越高，表明重建图像与真实图像之间的误差越小，重建质量越高。

SSIM 用于评估两幅图像在结构信息上的相似程度，包含了图像内容的多个方面，包括结构、亮度和对比度，提供了一个更全面的视觉相似性度量，其计算公式为

$$SSIM(X_{\mathrm{SR}},X_{\mathrm{HR}})=\frac{(2\cdot\mu_{\mathrm{SR}}\cdot\mu_{\mathrm{HR}}+C_1)\cdot(2\cdot\sigma_{\mathrm{SR\cdot HR}}+C_2)}{(\mu_{\mathrm{SR}}^2+\mu_{\mathrm{HR}}^2+C_1)\cdot(\sigma_{\mathrm{SR}}^2+\sigma_{\mathrm{HR}}^2+C_2)}, \tag{11}$$

式中：μ_{SR} 和 μ_{HR} 分别代表 X_{SR} 和 X_{HR} 的均值；σ_{SR} 和 σ_{HR} 分表为其标准差；$\sigma_{\mathrm{SR\cdot HR}}$ 是它们的协方差；C_1 和 C_2 为用于维持稳定的正常数。SSIM 值越高，表示重建图像在结构保持方面表现越好。SSIM 值越高，表示重建图像在结构保持方面表现越好。

此外，为进一步全面评估重建图像的感知质量，引入了学习感知图像块相似度 (learned perceptual image patch similarity, LPIPS)[54] 指标。相较于传统的像素级比较 (PSNR 等)，LPIPS 能够更好地反映人类视觉系统的感知特性。其计算公式为

$$LPIPS(X_{\mathrm{SR}},X_{\mathrm{HR}})=\sum_{l=1}^{L}w_l\cdot\|f_l(X_{\mathrm{SR}})-f_l(X_{\mathrm{HR}})\|_2^2, \tag{12}$$

式中：$f_l(X)$表示图像 X 在网络的第 l 层提取的特征；L 是网络中提取特征的层数；w_l 是每层特征差异的加权系数，可通过学习得到。本文采用预训练到 VGG 模型作为特征提取模型。LPIPS 值越低，代表重建结果在感知上越接近真实图像。最后，本文还对模型的计算效率和模型复杂度进行了分析，包括模型参数、浮点运算次数 (FLOPs) 和运行时间。其中，FLOPs 基于输入尺寸为 48 pixel×48 pixel 的图像计算得到。

4.1.3 实验环境及参数设置

为提升模型的泛化能力，本文采用了数据增强技术，包括 90° 随机旋转、随机水平翻转和垂直翻转。实验基于 PyTorch 框架实现并训练所提出的 MS^{3}M 模型，训练与测试均在一台配备 NVIDIA GeForce GTX 4090 显卡的服务器上进行。模型通过 Kaiming 初始化方法[55] 分别训练，以重建放大因子为 ×2、×3 和 ×4 的遥感图像。优化器选用 AdamW[56]，参数设置为 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$。初始学习率设为 2×10^{-4}，并每训练 400 个周期减半。训练过程中，低分辨率 (LR) 图像被随机裁剪为 48 pixel×48 pixel 的图像块作为输入，批大小 (batch size) 设置为 16，相应的高分辨率 (HR) 图像尺寸根据放大因子确定。在编解码器的每层 MS^{3}G 中均包含 3 个 MS^{3}B 模块。在 MS^{3}B 结构中，除 1×1 卷积层外，其余卷积层的通道数均设置为 48。

4.1.4 损失函数

在训练阶段，采用两种不同损失函数的加权求和作为总体训练目标以提升模型训练的稳定性和高效性。首先使用 $L1$ 损失函数来恢复 SR 图像的整体细节，目标是降低估计都恢复图像与对应真实图像之间的像素级差异：

$$\boldsymbol{L}_1(X_{\mathrm{SR}}, X_{\mathrm{HR}}) = \|X_{\mathrm{SR}} - X_{\mathrm{HR}}\|_1, \tag{13}$$

式中：X_{SR} 和 X_{HR} 分别表示重建的超分辨率图像及其对应的真实高分辨率图像。

然后，鉴于 RSISR 的主要目标是恢复丢失的高频属性，最小化频域中的偏差变得至关重要。为实现这一目标，引入基于 FFT 的频率损失函数。该损失函数通过以下公式量化傅里叶域中估计恢复图像与其对应真实图像之间的欧几里得距离：

$$\boldsymbol{L}_{FFT}(X_{\mathrm{SR}}, X_{\mathrm{HR}}) = \|FFT(X_{\mathrm{SR}}) - FFT(X_{\mathrm{HR}})\|_1, \tag{14}$$

式中：$FFT(\cdot)$ 表示快速傅里叶变换，用于将空间图像域中的特征转换到频域。

4.2 实验结果与分析

4.2.1 定量分析

为评估 MS^{3}M 算法的性能表现，选取了 9 种当前具有代表性的超分辨率重建方法作为对比基准，具体包括：Bicubic、SRCNN[15]、FSRCNN[57]、VDSR[23]、LGCNet[33]、DCM[58]、HSENet[36]、TransENet[20] 以及 ADAN[2]。在这些方法中，SRCNN、FSRCNN 与 VDSR 最初是针对自然图像设计的超分辨率重建模型；而 LGCNet、DCM、HSENet、TransENet 和 ADAN 则是近年来专门面向遥感图像提出的超分辨率重建算法，具备更好的领域适用性。实验在 UCMerced LandUse 数据集和 AID 数据集上分别进行，涵盖定量评价与视觉质量两方面的综合比较。为确保对比的公正性，所有参与比较的方法均采用官方发布的代码，并在统一的训练和测试环境下执行。实验设置包括两个数据集和三种不同的放大倍数 (×2、×3、×4)。在表 2 中展示了不同算法在 UCMerced LandUse 数据集上的重建性能，最佳结果以黑色粗体标示，次优结果则以下划线标出。从表中数据可见，MS^{3}M 在所有放大倍数下均取得了最优性能，其 PSNR 值在三种尺度上分别较次优方法 ADAN 提高了 0.06 dB、0.12 dB 和 0.05 dB；在 SSIM 指标上也分别优于 ADAN 0.004、0.003 和 0.001。相较于近期提出的其他方法，MS^{3}M 以更少的参数量获得了具有竞争力的客观指标和更优的视觉重建效果，在性能与效率之间实现了更好的平衡。

为进一步检验模型的泛化性能，还在类别更为多样、场景复杂度更高的 AID 数据集上进行了验证。如表 3 所示，MS^{3}M 在该数据集上同样表现优异。随着放大倍数的增加，MS^{3}M 与当前表现次优的 ADAN 算法之间的性能差距逐渐增大。具体来说，在 ×2 放大条件下，MS^{3}M 将 PSNR / SSIM 从 ADAN 的 36.93 / 0.9617 提升至 37.04 / 0.9617；在 ×3 放大时，由 32.96 / 0.8890 提升至 33.12 / 0.9004；在更具挑战性的 ×4 任务中，也从 29.99 / 0.8177 提升至 30.08 / 0.8203。这些结果充分说明，MS^{3}M 在不同尺度缩放任务上均保持稳定的性能优势，展现出良好的泛化能力。

为深入分析模型在不同场景类型下的重建效果，表 4 展示了各方法在 AID 数据集全部 30 类场景上、针对 ×4 缩放任务的 PSNR 表现。可以看出，MS^{3}M 在绝

表 2　在 UCMerced LandUse 数据集 (×2、×3和×4) 上的 PSNR/SSIM 结果

Table 2　PSNR/SSIM results on the UCMerced LandUse dataset (×2, ×3, and ×4)

Scale	Bicubic	SRCNN	FSRCNN	VDSR	LGCNet	DCM	HSENet	TransENet	ADAN	Ours
2	30.76/0.8789	32.84/0.9152	33.18/0.9196	33.38/0.9220	33.48/0.9235	33.65/0.9274	34.22/0.9327	35.43/0.9355	35.62/0.9717	**35.68/0.9758**
3	27.46/0.7631	28.66/0.8038	29.09/0.8167	29.28/0.8232	29.28/0.8238	29.52/0.8349	30.00/0.8420	31.03/0.8526	31.10/0.8811	**31.22/0.8845**
4	25.65/0.6725	26.78/0.7219	26.93/0.7267	26.85/0.7317	27.02/0.7333	27.22/0.7528	27.73/0.7623	28.74/0.7694	28.84/0.8003	**28.89/0.8128**

表 3 在 AID 数据集 (×2、×3 和 ×4) 上的 PSNR/SSIM 结果

Table 3 PSNR/SSIM results on the AID dataset (×2, ×3, and ×4)

Scale	Bicubic	SRCNN	FSRCNN	VDSR	LGCNet	DCM	HSENet	TransENet	ADAN	Ours
2	32.39/0.8906	34.49/0.9286	34.73/0.9331	35.05/0.9346	34.80/0.9320	35.21/0.9366	35.24/0.9368	35.28/0.9374	36.93/0.9617	**37.02/0.9623**
3	29.08/0.7863	30.55/0.8372	30.98/0.8401	31.15/0.8522	30.73/0.8417	31.31/0.8561	31.39/0.8572	31.45/0.8595	32.96/0.8889	**33.12/0.9004**
4	27.30/0.7036	28.40/0.7561	28.77/0.7729	28.99/0.7753	28.61/0.7626	29.17/0.7824	29.21/0.7850	29.38/0.7909	29.99/0.8177	**30.08/0.8203**

表 4 AID 数据集中放大因子为 ×4 时每个类别的平均 PSNR/dB

Table 4 Average PSNR for each category with an upscaling factor of ×4 on the AID dataset/dB

Class	Bicubic	SRCNN	LGCNet	VDSR	DCM	HSENet	TransENet	ADAN	Ours
Airport	27.03	28.17	28.39	28.82	28.99	29.03	29.23	29.31	29.34
Bareland	34.88	35.63	35.78	35.98	36.17	36.21	36.20	36.42	36.43
Baseball field	29.06	30.51	30.75	31.18	31.36	31.23	31.59	31.28	32.29
Beach	31.07	31.92	32.08	32.29	32.45	32.76	32.55	33.51	33.53
Bridge	28.98	30.41	30.67	31.19	31.39	31.30	31.63	30.83	30.84
Center	25.26	26.59	26.92	27.48	27.72	27.84	28.03	27.44	27.46
Church	22.15	23.41	23.68	24.12	24.29	24.39	24.51	24.62	24.67
Commercial	25.83	27.05	27.24	27.62	27.78	27.99	27.97	28.39	28.41
Dense residential	23.05	24.13	24.33	24.70	24.87	24.44	25.13	24.62	24.66
Desert	38.49	38.84	39.06	39.13	39.27	39.37	39.31	38.99	38.97
Farmland	32.30	33.48	33.77	34.20	34.42	33.90	34.58	34.19	34.23
Forest	27.39	28.15	28.20	28.36	28.47	38.31	28.56	28.37	28.35
Industrial	24.75	26.00	26.24	26.72	26.92	26.99	27.21	27.30	27.37
Meadow	32.06	32.57	32.65	32.77	32.88	32.74	32.94	33.30	33.32
Medium residential	26.09	27.37	27.63	28.06	28.25	28.11	28.45	26.94	26.97
Mountain	28.04	28.90	28.97	29.11	29.18	29.26	29.28	28.89	28.91
Park	26.23	27.25	27.37	27.69	27.82	28.23	28.01	28.11	28.12
Parking	22.33	24.01	24.40	25.21	25.74	26.17	26.40	26.01	26.09
Playground	27.27	28.72	29.04	29.62	29.92	31.18	30.30	32.00	31.98
Pond	28.94	29.85	30.00	30.26	30.39	30.40	30.53	30.33	30.35
Port	24.69	25.82	26.02	26.43	26.62	26.92	26.91	27.47	27.52
Railway station	26.31	27.55	27.76	28.19	28.38	28.47	28.61	28.42	28.47
Resort	25.98	27.12	27.32	27.71	27.88	27.99	28.08	27.66	27.69
River	29.61	30.48	30.60	30.82	30.91	30.88	31.00	30.28	30.29
School	24.91	26.13	26.34	26.78	26.94	27.51	27.22	27.52	27.59
Sparse residential	25.41	26.16	26.27	26.46	26.53	26.44	26.43	26.58	26.63
Square	26.75	28.13	28.39	28.91	29.13	29.05	29.39	28.79	28.84
Stadium	24.81	26.10	26.37	26.88	27.10	27.28	27.41	28.01	28.08
Storage tanks	24.18	25.27	25.48	25.86	26.00	26.07	26.20	26.80	26.85
Viaduct	25.86	27.03	27.26	27.74	27.93	28.12	28.21	28.01	28.11
AVG	27.30	28.40	28.61	28.99	29.17	29.21	29.38	29.99	30.08

大多数场景类别中均取得了最高精度，其主要竞争对手 ADAN 在部分类别中表现接近，但 MS^{3}M 在整体 PSNR 上仍以 0.09 dB 的优势领先，显示出其在不同场景下均具备稳健的重建性能。这一优势可归因于 MS^{3}M 在局部特征捕捉、高频细节恢复以及像素级重建激活方面的出色表现。还引入了无参考感知质量评价指标 LPIPS 进行辅助分析，其数值越低表示视觉感知质量越高。从表 5 可以看出，MS^{3}M 在感知一致性方面同样显著优于其他对比方法。具体而言，在 ×2 缩放条件下，MS^{3}M 的 LPIPS 值较 HSENet 降低了 0.0016；在×3 和×4 缩放时，较 ADAN 分别进一步降低了 0.0008 和 0.0005，这进一步验证了所提方法在提升视觉感知质量和维持人眼感知方面的有效性。如图 5 所示，通过损失函数曲线与峰值信噪比 (PSNR) 指标对本文算法与 ADAN 算法进行对比分析。图 5(a) 展示了两种模型在训练过程中的损失函数变化情况。可以观察到，本文算法的损失值在整个训练过程中均低于 ADAN 算法，且收敛速度更快，能够收敛到更优的帕累托前沿解，最终稳定在更低的水平，这表明 MS^{3}M 算法具有更优的优化效率与稳定性。性能对比结果 (图 5(b)) 进一步验证了本文算法的优越性。在 PSNR 指标上，MS^{3}M 算法自训练早期阶段即展现出优势，其 PSNR 值始终高于对比算法，并且随着训练的进行，性能差距更为显著。最终，MS^{3}M 算法达到了更高的 PSNR 峰值，这充分证明了其在提升遥感图像重建质量方面的有效性。综合而言，无论是从训练过程的收敛性还是最终的重建性能来看，本文所提出的算法均一致地优于 ADAN 算法。

4.2.2 定性分析

为全面评估 MS^{3}M 模型的视觉质量并与主流方法进行对比，本文结合视觉重建结果展开分析。图 5 至图 8 展示了不同缩放因子下的遥感图像重建效果。总体而言，MS^{3}M 在视觉上最接近真实高分辨率图像，细节还原能力显著优于对比方法。具体而言，在图 6 所示样例中，HSENet 与 TransENet 等方法的输出出现明显伪影和失真，ADAN 模型重建的田垄分界模糊，而 MS^{3}M 重建的田垄结构清晰自然，最接近真实图像。这一差异可能源于对比方法在处理复杂退化过程时未能有效抑制噪声干扰，导致错误信息聚合。MS^{3}M 通过集成高阶矩引导的通道亲和力调制模块，增强通道交互并抑制特征冗余，有效降低了噪声干扰，提升了特征聚合的鲁棒性。图 7 展示了 ×3缩放倍数下的机场跑道重建效果。对比方法难以有效恢复跑道纹理等细微结构，而 MS^{3}M 仍能重建出合理的细节信息，特别是在线性地物 (如跑道标线) 的连续性恢复方面表现突出。这归功于并行 Mamba 模块提供的全面空间感知和非局部特征聚合能力。在最具挑战性的 ×4缩放任务中 (图 8、图 9)，MS^{3}M 仍保持最佳视觉重建效果。其重建的房屋和飞机图像边缘清晰、纹理保持良好，而其他方法均出现不同程度的模糊和几何畸变，进一步验证了 MS^{3}M 在结构保持方面的优势。这得益于模型在构建全局感知能力的基础上，引入多感受野特征聚合机制，实现了全局、局部和多尺度特征的全面融合，显著提升了模型对不同尺度特征的建模能力。

4.3 消融实验

为全面评估 MS^{3}M 模型的性能，本文开展了系统的消融实验，对各模块的贡献进行深入分析。默认所有消融实验均在 AID 数据集上进行，训练周期设置为 500 代，放大因子为 ×2。

4.3.1 各模块有效性研究

为验证 MS^{3}M 模型中各模块的有效性，本文通过逐步添加子模块 (P-Mamba、MRFAM 与 HMCAM) 的方式开展消融实验。需说明的是，基线模型由参数量相当的残差模块构成。如表 6 所示，与基线模型 (模型 0) 相比，引入基于并行扫描策略的 Mamba 模块 (P-Mamba，模型 1) 使 PSNR 提升 0.1 dB，SSIM 提升 0.0004。该性能提升可归因于 P-Mamba 的全局特征建模能力与非局部感知机制，有效恢复了图像的整体结构。在 P-Mamba 基础上引入多感受野聚合模块 (MRFAM)，构成 MP-Mamba 模块 (模型 2) 后，PSNR 与 SSIM 分别进一步提升 0.06 dB

表 5　LPIPS 在尺度为 ×2、×3和 ×4 的 UCMerced LandUse 数据集上的结果

Table 5　LPIPS results on the UCMerced LandUse dataset with scaling factors of ×2, ×3, and ×4

Scale	Bicubic	SRCNN	FSRCNN	VDSR	LGCNet	DCM	HSENet	TransENet	ADAN	Ours
2	0.0721	0.0444	0.0471	0.0287	0.0293	0.0284	0.0266	0.0279	0.0256	**0.0250**
3	0.1281	0.0945	0.1062	0.0801	0.0752	0.0698	0.0654	0.0649	0.0641	**0.0633**
4	0.1650	0.1260	0.1395	0.1102	0.1093	0.1046	0.1081	0.1030	0.1022	**0.1017**

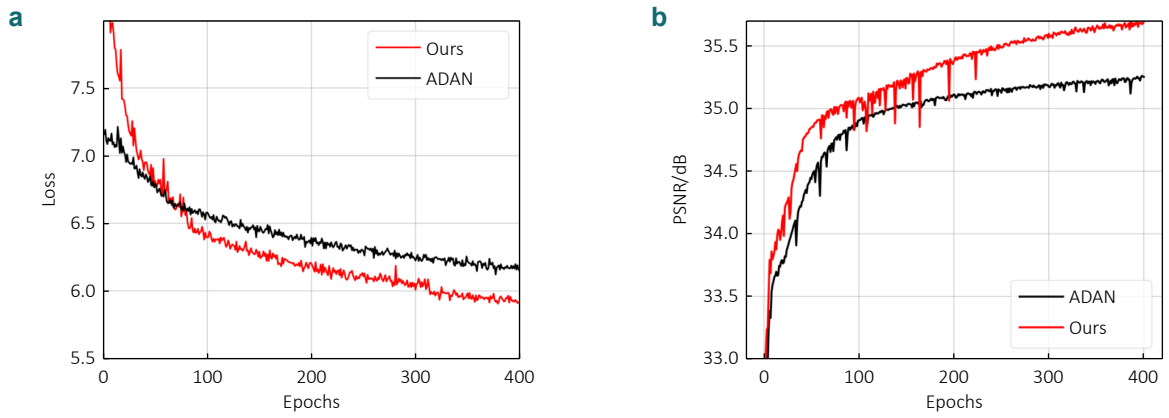


图 5 在 UCMerced Landuse 数据集上的比较结果。(a) 缩放因子为 ×3 时的损失函数分析；(b) 缩放因子为 ×2 时的 PSNR 指标分析

Fig. 5 Comparative analysis on the UCMerced Landuse dataset. (a) Loss function under a scaling factor of ×3; (b) PSNR metric under a scaling factor of ×2

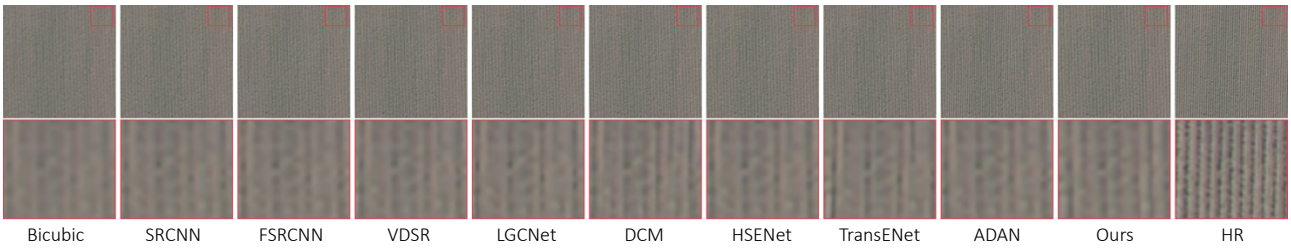


图 6 缩放因子为 ×3 时在 UCMerced Landuse 数据集上的视觉比较结果

Fig. 6 Visual comparison on the UCMerced Landuse dataset with a scaling factor of ×3

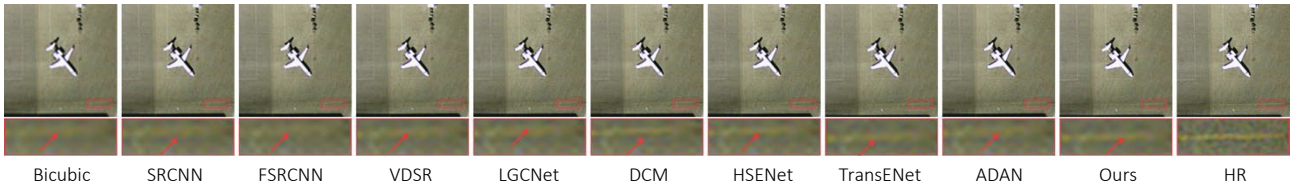


图 7 缩放因子为 ×3 时在 UCMerced Landuse 数据集上的视觉比较结果

Fig. 7 Visual comparison on the UCMerced Landuse dataset with a scaling factor of ×3

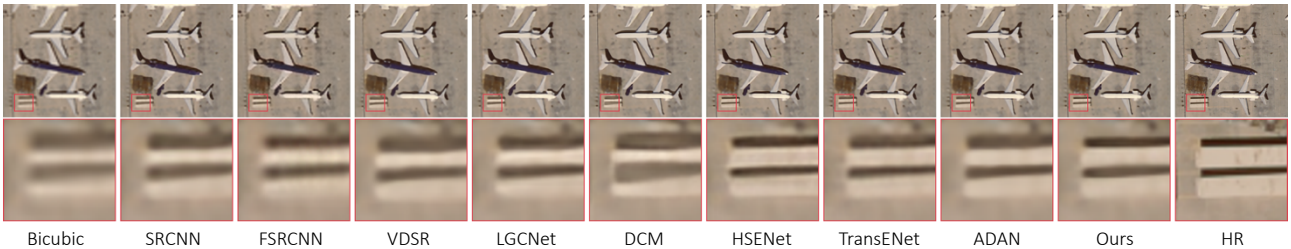


图 8 缩放因子为 ×4 时在 UCMerced Landuse 数据集上的视觉比较结果

Fig. 8 Visual comparison on the UCMerced Landuse dataset with a scaling factor of ×4

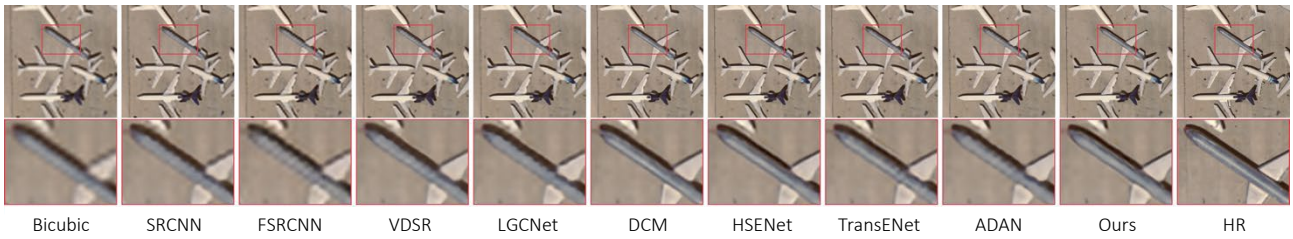


图 9 缩放因子为 ×4 时在 UCMerced Landuse 数据集上的视觉比较结果

Fig. 9 Visual comparison on the UCMerced Landuse dataset with a scaling factor of ×4

与 0.0004。MRFAM 通过提取不同尺度地物的空间变化特征，为遥感图像超分辨率重建提供了关键的多尺度上下文信息。最终，完整版 MS^{3}M 模型 (模型 3) 相较于基线模型，PSNR 显著提升 0.21 dB，SSIM 提高 0.0014。该结果表明，模型在融合多尺度空间特征的基础上，进一步通过通道维度全局信息增强了特征表征能力。实验结果表明，各模块的逐步引入带来性能的持续提升，充分证明了 P-Mamba、MRFAM 与 HMCAM 模块设计的有效性及其在超分辨率重建任务中的互补性。

4.3.2 基于并行扫描策略的 Mamba 模块 (P-Mamba) 的消融研究

为验证本文所提出的基于并行扫描策略的 Mamba 模块 (P-Mamba) 的有效性，进行了系统的消融实验，结果如表 7 所示。实验清晰地表明，本文设计的并行扫描策略在模型性能与计算效率之间取得了卓越的平衡。具体而言，与传统的单向扫描策略相比，P-Mamba 模块在模型参数量 (4.03 M vs. 4.08 M) 与计算复杂度 (5.28 G vs. 5.12 G Flops) 基本持平的情况下，在 AID 数据集上取得了显著的性能提升 (PSNR 从 36.91 dB 提升至 37.02 dB，SSIM 从 0.9602 提升至 0.9623)。这证明了并行扫描策略能够通过更丰富的扫描路径捕捉空间信息，从而生成更高质量的图像。更重要的是，与四向扫描策略相比，我们方法的优势更为突出。四向扫描虽然能带来一定的性能增益，但其计算代价极为高昂，Flops 高达 18.73 G，是本文方法 (5.28 G) 的 3.5 倍以上。而我们的 P-Mamba 模块以仅约四分之一的计算成本，在 AID 数据集上实现了与之相当甚至略优的性能 (PSNR: 37.02 vs. 37.01，SSIM: 0.9623 vs. 0.9621)，同时在 UCMerced LandUse 数据集上保持了极具竞争力的 SSIM 指标 (0.9758)。这充分说明，本文提出的并行机制成功避免了简单堆叠扫描方向带来的计算冗余，以一种更高效的方式提升特征复用效率和上下文信息融合质量，从而在保持较低复杂度的同时增强图像全局和局部信息的有效建模。因此，本文设计的 P-Mamba 模块不仅显著提升了原始状态空间

表 6 本文所提各模块结构的消融研究结果

Table 6 Ablation study results of the proposed architectural modules

	Model 0	Model 1	Model 2	Model 3 (Ours)
P-Mamba	×	√	√	√
MRFAM	×	×	√	√
HMCAM	×	×	×	√
PSNR/dB	36.81	36.91	36.97	**37.02**
SSIM	0.9609	0.9613	0.9617	**0.9623**

表 7 P-Mamba 模块的消融研究结果

Table 7 Ablation study results of the P-Mamba

Methods	# Parameters/M	Flops/G	UCMerced LandUse		AID	
			PSNR	SSIM	PSNR	SSIM
单向扫描策略	4.08	5.12	36.31	0.9713	36.91	0.9602
四向扫描策略(SSM)	4.11	18.73	36.66	0.9759	37.01	0.9621
Parallel-SSM	**4.03**	**5.28**	**35.68**	**0.9758**	**37.02**	**0.9623**

模型 (SSM) 的特征建模能力，更以一种计算高效的方式，实现了超越复杂多向扫描策略的综合性能。

4.3.3 高阶矩引导的通道亲和力调制模块的消融研究

为了进一步评估 HMCAM 模型的有效性和高效性，本文将所提的高阶矩引导的通道亲和力调制模块 (HMCAM) 与常见的重建网络中采用的通道注意力机制进行比较。表 8 (前三行数据) 展示了 HMCAM 与经典的 SENet、ECANet 和 GCT 模型在 UCMerced LandUse 和 AID 数据集上的重建性能 (PSNR 和 SSIM) 和模型效率分析 (parameters 和 Flops)。结果显示 HMCAM 模型在几乎只需要使用一半的计算资源时，取得了相当甚至更好的重建性能。具体来说，HMCAM 仅需要 GCT 模型 59% 的 Flops 计算资源，取得了比 GCT 更好的重建性能 (在 AID 数据集上取得了 0.03 dB 的领先)，进一步证明了本文所提 HMCAM 模型的优越性和有效性。此外，为了探讨不同矩统计量的有效性，实验中令 HMCAM 单独使用不同的矩统计量，如单独使用一阶矩 ($\boldsymbol{M}_1$)、二阶矩 ($\boldsymbol{M}_2$)、三阶矩 ($\boldsymbol{M}_3$)、一阶矩和二阶矩联合使用 ($\boldsymbol{M}_1+\boldsymbol{M}_2$)。本文发现，单独使用不同的矩统计量都会导致模型性能的下降，这是因为单一矩统计量只能描述特征的某些特性，不能全部感知潜在的本质属性，因此阻碍模型全面恢复重建图像。具体来说，相较于单独使用矩统计量，HMCAM 在 UCMerced LandUse 数据集上分别取得了 0.05 dB、0.06 dB、0.11 dB 的 PSNR 性能提升，这也充分证明了本文所提高阶矩统计量的有效性。此外，随着融合的不同矩统计量的增多，从而能够探索和融合不同属性的本质特征，使模型性能也逐渐增强。而且，值得注意的是，引入多种不同的矩统计量不会影响模型的参数量和计算复杂度，因此可以无缝集成在本文所提模型中。

4.4 模型复杂度分析

表 9 展示了在 UCMerced LandUse 数据集上针对 ×2 放大倍率的多种超分辨率方法的性能与计算复杂度分析结果。FLOPs 作为衡量模型计算复杂度的关键指标，反映了算法推理过程中所需的浮点运算总量，其数值直接关系到模型的实际部署效率。从结果可以看出，MS^{3}M 在 FLOPs 指标上优于多数对比模型，显示出较低的计算负担。在重建精度方面，MS^{3}M 表现突出，其 PSNR 值较 LGCNet 提升了 2.2 dB，显著优于其他参数量更大或结构更复杂的模型。这一结果表明 MS^{3}M 在计算效率与遥感图像复原质量之间实现了良好平衡。该优势主要归功于 ADAN 采用的并行多感受野感知 Mamba 机制，该机制通过并行化多尺度特征提取路径，有效增强了模型对遥感图像中不同尺度地物结构的感知能力，同时避免了传统串行或密集连接带来的计算冗余。实验结果表明，

表 8 高阶矩引导的通道亲和力调制模块的消融研究结果

Table 8 Ablation study results of the high-order moment guided channel affinity modulation module

Methods	Parameters/M	Flops/G	UCMerced LandUse		AID	
			PSNR	SSIM	PSNR	SSIM
SENet	5.26	9.38	35.2	0.9750	36.96	0.9616
ECANet	4.31	9.02	35.63	0.9751	36.96	0.9617
GCT	4.11	8.87	35.67	0.9757	36.99	0.9622
$\boldsymbol{M}_1$	4.03	5.28	35.63	0.9751	36.97	0.9617
$\boldsymbol{M}_2$	4.03	5.28	35.62	0.9749	36.94	0.9615
$\boldsymbol{M}_3$	4.03	5.28	35.57	0.9744	36.85	0.9610
$\boldsymbol{M}_1+\boldsymbol{M}_2$	4.03	5.28	35.66	0.9756	37.00	0.9620
$\boldsymbol{M}_1+\boldsymbol{M}_2+\boldsymbol{M}_3$ (HMCAM)	**4.03**	**5.28**	**35.68**	**0.9758**	**37.02**	**0.9623**

表 9 模型复杂性分析结果

Table 9 Model complexity analysis results

Methods	LGCNet	DCM	HSENet	TransENet	ADAN	MS^{3}M (Ours)
Parameters/M	0.193	2.18	5.4	37.8	4.12	**4.03**
Flops/G	7.11	7.32	10.80	9.32	7.16	**5.28**
PSNR/dB	33.48	33.65	34.22	35.43	35.62	**35.68**

该机制不仅有助于提升客观指标，还能在视觉层面产生更清晰、更自然的超分效果。综上所述，MS^{3}M 凭借其并行多感受野 Mamba 机制，以更少的计算开销实现了更优的超分辨率重建质量，较好地平衡了网络复杂度与图像复原性能之间的关系，为资源受限环境下的遥感图像处理提供了有效的解决方案。

5 总结与展望

本文提出一种多尺度感知增强状态空间模型(MS^{3}M)，用于高效实现遥感图像超分辨率重建，提升重建结果在下游高级视觉任务中的鲁棒性。该方法首先引入并行 Mamba 模块，采用分组多方向并行扫描策略，在保持线性计算复杂度的同时显著提升特征交互效率，协同捕捉局部与非局部特征；其次，针对现有 Mamba 架构在多层次特征感知中的局限，设计多尺度感受野聚合模块，有效融合不同尺度的上下文信息，增强对遥感图像中多尺度地物目标的表征能力；最后，提出基于二阶矩统计的通道亲和力调制机制，通过压缩冗余特征与抑制噪声提升特征的紧凑性与判别性，从而为实际应用提供更准确可靠的 RSIs。在多个 RSSISR 基准数据集上的实验表明，MS^{3}M 在细节恢复方面表现优异，能有效抑制伪影与噪声，显著提升 SR 图像质量。然而，当前模型计算成本仍较高，难以在资源受限的终端设备中实时部署，未来工作将聚焦于轻量化设计，以推动其在实际场景中的应用。

作者贡献

陈嘉诚：实验设计、数据分析、论文够细与撰写；吴菲：指导论文选题、论文修改、论文审核；屠杭垚，蒋嘉伟：测量实验的设计及数据整理和分析。

利益冲突

所有作者声明无利益冲突

参考文献

1. Vivone G, Deng L J, Deng S Q, et al. Deep learning in remote sensing image fusion: methods, protocols, data, and future perspectives[J]. *IEEE Geosci Remote Sens Mag*, 2025, **13**(1): 269−310.
2. Wu F, Chen J C, Yang J, et al. Remote-sensing images reconstruction based on adaptive dual-domain attention network[J]. *Opto-Electron Eng*, 2025, **52**(4): 240297.
   吴菲, 陈嘉诚, 杨俊, 等. 基于自适应双域注意力网络的遥感图像重建[J]. 光电工程, 2025, **52**(4): 240297.
3. Liu L F, Tong Z C, Cai Z C, et al. HierU-Net: a hierarchical semantic segmentation method for land cover mapping[J]. *IEEE Trans Geosci Remote Sens*, 2024, **62**: 4404614.
4. Chen K Y, Chen B W, Liu C Y, et al. RSMamba: remote sensing image classification with state space model[J]. *IEEE Geosci Remote Sens Lett*, 2024, **21**: 8002605.
5. Xiao Z J, Zhang J H, Lin B H. Feature coordination and fine-grained perception of small targets in remote sensing images[J]. *Opto-Electron Eng*, 2024, **51**(6): 240066.
   肖振久, 张杰浩, 林渤翰. 特征协同与细粒度感知的遥感图像小目标检测[J]. 光电工程, 2024, **51**(6): 240066.
6. Sharifuzzaman Sagar A S M, Chen Y, Xie Y K, et al. MSA R-CNN: a comprehensive approach to remote sensing object detection and scene understanding[J]. *Expert Syst Appl*, 2024, **241**: 122788.
7. Noman M, Fiaz M, Cholakkal H, et al. ELGC-Net: efficient local–global context aggregation for remote sensing change detection[J]. *IEEE Trans Geosci Remote Sens*, 2024, **62**: 4701611.
8. Xiao Y, Yuan Q Q, Jiang K, et al. TTST: a top-*K* token selective transformer for remote sensing image super-resolution[J]. *IEEE Trans Image Process*, 2024, **33**: 738−752.
9. Chen Y T, Xia R L, Yang K, et al. MICU: image super-resolution via multi-level information compensation and U-Net[J]. *Expert Syst Appl*, 2024, **245**: 123111.
10. Wu R Y, Yang T, Sun L C, et al. SeeSR: towards semantics-aware real-world image super-resolution[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 25456–25467. https://doi.org/10.1109/CVPR52733.2024.02405.
11. Zhang J F, Tu Y J. SwinFR: combining SwinIR and fast Fourier for super-resolution reconstruction of remote sensing images[J]. *Digital Signal Process*, 2025, **159**: 105026.
12. Elad M, Datsenko D. Example-based regularization deployed to super-resolution reconstruction of a single image[J]. *Comput J*, 2009, **52**(1): 15−30.
13. Kim K I, Kwon Y. Single-image super-resolution using sparse regression and natural image prior[J]. *IEEE Trans Pattern Anal Mach Intell*, 2010, **32**(6): 1127−1133.
14. Sun J, Sun J, Xu Z B, et al. Gradient profile prior and its applications in image super-resolution and enhancement[J]. *IEEE Trans Image Process*, 2011, **20**(6): 1529−1542.
15. Dong C, Loy C C, He K M, et al. Image super-resolution using deep convolutional networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2016, **38**(2): 295−307.
16. Wang Z Y, Li L L, Xue Y, et al. FeNet: feature enhancement network for lightweight remote-sensing image super-resolution[J]. *IEEE Trans Geosci Remote Sens*, 2022, **60**: 5622112.
17. Chen L, Liu H, Yang M H, et al. Remote sensing image super-resolution via residual aggregation and split attentional fusion network[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2021, **14**: 9546−9556.
18. Wang Z, Zhao Y W, Chen J C. Multi-scale fast Fourier transform based attention network for remote-sensing image super-resolution[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2023, **16**: 2728−2740.
19. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.
20. Lei S, Shi Z W, Mo W J. Transformer-based multistage

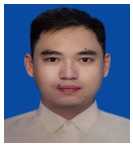
enhancement for remote sensing image super-resolution[J]. *IEEE Trans Geosci Remote Sens*, 2022, **60**: 5615611.

21. Kang X D, Duan P H, Li J E, et al. Efficient swin transformer for remote sensing image super-resolution[J]. *IEEE Trans Image Process*, 2024, **33**: 6367−6379.
22. Hatamizadeh A, Kautz J. MambaVision: a hybrid mamba-transformer vision backbone[C]//*Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025: 25261–25270. https://doi.org/10.1109/CVPR52734.2025.02352.
23. Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 1646–1654. https://doi.org/10.1109/CVPR.2016.182.
24. Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017: 1132–1140. https://doi.org/10.1109/CVPRW.2017.151.
25. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//*Proceedings of the 32nd International Conference on Machine Learning*, 2015: 448–456.
26. Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155.
27. Niu B, Wen W L, Ren W Q, et al. Single image super-resolution via a holistic attention network[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 191–207. https://doi.org/10.1007/978-3-030-58610-2_12.
28. Zhang Y L, Li K P, Li K, et al. Image super-resolution using very deep residual channel attention networks[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 294–310. https://doi.org/10.1007/978-3-030-01234-2_18.
29. Mei Y Q, Fan Y C, Zhou Y Q. Image super-resolution with non-local sparse attention[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 3516–3525. https://doi.org/10.1109/CVPR46437.2021.00352.
30. Chen G H, Xu D, He K J, et al. TSSA-Net: transposed Sparse Self-Attention-based network for image super-resolution[J]. *Eng Appl Artif Intell*, 2025, **153**: 110823.
31. Xiao Y, Yuan Q Q, Jiang K, et al. EDiffSR: an efficient diffusion probabilistic model for remote sensing image super-resolution[J]. *IEEE Trans Geosci Remote Sens*, 2024, **62**: 5601514.
32. Xiao Y, Yuan Q Q, Jiang K, et al. From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution[J]. *Inf Fusion*, 2023, **96**: 297−311.
33. Lei S, Shi Z W, Zou Z X. Super-resolution for remote sensing images via local–global combined network[J]. *IEEE Geosci Remote Sens Lett*, 2017, **14**(8): 1243−1247.
34. Dong X Y, Sun X, Jia X P, et al. Remote sensing image super-resolution using novel dense-sampling networks[J]. *IEEE Trans Geosci Remote Sens*, 2021, **59**(2): 1618−1633.
35. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
36. Lei S, Shi Z W. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution[J]. *IEEE Trans Geosci Remote Sens*, 2022, **60**: 5401410.
37. Wang H Y, Cheng S L, Li Y M, et al. Lightweight remote-sensing image super-resolution via attention-based multilevel feature fusion network[J]. *IEEE Trans Geosci Remote Sens*, 2023, **61**: 2005715.
38. Zhang X R, Li Z Y, Zhang T Y, et al. Remote sensing image super-resolution via dual-resolution network based on connected attention mechanism[J]. *IEEE Trans Geosci Remote Sens*, 2022, **60**: 5611013.
39. Li H T, Deng W H, Zhu Q Q, et al. Local-global context-aware generative dual-region adversarial networks for remote sensing scene image super-resolution[J]. *IEEE Trans Geosci Remote Sens*, 2024, **62**: 5402114.
40. Gu A, Johnson I, Goel K, et al. Combining recurrent, convolutional, and continuous-time models with linear state-space layers[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 44.
41. Kalman R E. A new approach to linear filtering and prediction problems[J]. *J Basic Eng*, 1960, **82**(1): 35−45.
42. Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces[C]//*Proceedings of the 10th International Conference on Learning Representations*, 2022: 1–27.
43. Smith J T H, Warrington A, Linderman S W. Simplified state space layers for sequence modeling[C]//*Proceedings of the 11th International Conference on Learning Representations*, 2023: 1–35.
44. Fu D Y, Dao T, Saab K K, et al. Hungry hungry hippos: towards language modeling with state space models[C]//*Proceedings of the 11th International Conference on Learning Representations*, 2023: 1–25.
45. Mehta H, Gupta A, Cutkosky A, et al. Long range language modeling via gated state spaces[C]//*Proceedings of the 11th International Conference on Learning Representations*, 2023: 1–20.
46. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces[C]//*Proceedings of the ICLR 2024*, 2024: 1–32.
47. Zhu L H, Liao B C, Zhang Q, et al. Vision mamba: efficient visual representation learning with bidirectional state space model[C]//*Proceedings of the 41st International Conference on Machine Learning*, 2024: 1–14.
48. Liu Y, Tian Y J, Zhao Y Z, et al. VMamba: visual state space model[C]//*Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024: 2373.
49. Guo H, Li J M, Dai T, et al. MambaIR: a simple baseline for image restoration with state-space model[C]//*Proceedings of the 18th European Conference on Computer Vision*, 2024: 222–241. https://doi.org/10.1007/978-3-031-72649-1_13.
50. Shi Y, Xia B, Jin X Y, et al. VmambaIR: visual state space model for image restoration[J]. *IEEE Trans Circuits Syst Video Technol*, 2025, **35**(6): 5560−5574.
51. Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification[C]//*Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010: 270–279. https://doi.org/10.1145/1869790.1869829.
52. Xia G S, Hu J W, Hu F, et al. AID: a benchmark data set for performance evaluation of aerial scene classification[J]. *IEEE Trans Geosci Remote Sens*, 2017, **55**(7): 3965−3981.
53. Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. *IEEE Trans Image*

Process, 2004, **13**(4): 600−612.

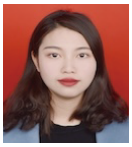
54. Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 586–595. https://doi.org/10.1109/CVPR.2018.00068.
55. He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, 2015: 1026–1034. https://doi.org/10.1109/ICCV.2015.123.
56. Zhou P, Xie X Y, Lin Z C, et al. Towards understanding convergence and generalization of AdamW[J]. *IEEE Trans Pattern Anal Mach Intell*, 2024, **46**(9): 6486−6493.
57. Dong C, Loy C C, Tang X O. Accelerating the super-resolution convolutional neural network[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 391–407. https://doi.org/10.1007/978-3-319-46475-6_25.
58. Haut J M, Paoletti M E, Fernandez-Beltran R, et al. Remote sensing single-image superresolution based on a deep compendium model[J]. *IEEE Geosci Remote Sens Lett*, 2019, **16**(9): 1432−1436.

作者简介



陈嘉诚 (1997-)，男，博士，嘉兴大学讲师，主要从事人工智能、图像处理等方面研究。

E-mail：jxdxcjc@zjxu.edu.cn



【通信作者】吴菲 (1995-)，女，博士，嘉兴大学讲师，主要从事多目标优化、图像超分辨率重建等方面研究。

E-mail：WFMOOK@163.com

Remote sensing image super-resolution reconstruction via multi-scale augmented state space model

Chen Jiacheng[1], Wu Fei[1*], Tu Hangyao[2], Jiang Jiawei[3], Wang Wanliang[3]

[1] College of Artificial Intelligence, Jiaxing University, Jiaxing, Zhejiang 314001, China;
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, 310015, China;
[3] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China

Abstract

Convolutional neural networks (CNNs) and vision transformers (ViTs) represent the two dominant paradigms in remote sensing single image super-resolution (RSSISR), each with distinct strengths. While CNNs have long been the workhorse due to their inductive biases, vision transformers have recently demonstrated superior performance in many cases, primarily attributed to their exceptional capability in modeling long-range dependencies through the self-attention mechanism. However, this advantage comes at a significant cost: the self-attention mechanism suffers from quadratic computational complexity with respect to image size. This inherent limitation becomes a critical bottleneck in RSSISR, where generating high-resolution outputs from low-resolution inputs demands extensive computation, severely restricting the practical deployment of ViTs for large-area remote sensing imagery.

To effectively overcome this fundamental challenge, we propose a novel architecture named the multi-scale augmented state space model (MS^3M). Our approach is grounded in the recent advancements of state space models (SSMs), which are renowned for their linear computational complexity and strong potential in capturing long-range interactions. Unlike existing SSM-based feature extraction methods that often rely on fixed, unidirectional scanning paths, our MS^3M introduces a grouped parallel scanning strategy. This design efficiently captures comprehensive global and non-local features without being constrained by a single scanning direction, all while rigorously maintaining linear computational complexity, thereby ensuring high efficiency.

Furthermore, acknowledging the inherent and critically important multi-scale spatial structures present in remote sensing images—from fine-grained textures of buildings to extensive patterns of farmlands—we embed a multi-receptive-field aggregation mechanism directly

Foundation item: Project supported by the National Natural Science Foundation of China (62506337), the Natural Science Foundation of Zhejiang Province (LQN25F020024), and the Science and Technology Plan Project of Jiaxing City (2025CGZ007)

*Correspondence: Wu F, E-mail: WFMOOK@163.com

into the state space model. This allows our network to seamlessly integrate contextual information across different scales, a capability essential for accurately reconstructing complex geographical objects. To further enhance the representation power of local features, we design a novel high-order moment channel affinity modulation module. This module moves beyond simple first-order statistics to optimize feature expressions, enabling more nuanced and powerful feature transformations within the network. The entire MS^{3}M framework is constructed upon a U-shaped architecture to facilitate effective multi-level feature fusion across the encoder and decoder.

We conduct extensive experiments on several public remote sensing datasets. The results demonstrate that our proposed MS^{3}M achieves state-of-the-art performance, outperforming existing leading methods in terms of objective metrics including PSNR, SSIM and LPIPS, as well as in subjective visual quality. The superior results validate the effectiveness of our architectural choices and underscore MS^{3}M's advancement as a robust and efficient solution for the challenging task of remote sensing image super-resolution.

Keywords: remote sensing image super-resolution; state space model; grouped parallel scanning; multi-receptive field aggregation; high-order moment statistics

Chen J C, Wu F, Tu H Y, et al. Remote sensing image super-resolution reconstruction via multi-scale augmented state space model[J]. *Opto-Electron Eng*, 2026, **53**(4): 250304; DOI: 10.12086/oee.2026.250304