软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007569] [CSTR: 32375.14.jos.007569]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

用于图像去噪的特征空间上下文 Transformer*

胡雨轩[1], 张师超[2,3]

[1](中南大学 计算机学院, 湖南 长沙 410083)

[2](教育区块链与智能技术教育部重点实验室 (广西师范大学), 广西 桂林 541004)

[3](广西多源信息挖掘与安全重点实验室 (广西师范大学), 广西 桂林 541004)

通信作者: 张师超, E-mail: zhangsc@mailbox.gxnu.edu.cn

摘　要: 图像去噪是计算机视觉中的基础任务, 其关键在于利用有效先验知识恢复噪声污染下的细节信息. 针对传统卷积神经网络因固定权重与局部感受野限制而存在的性能瓶颈, 以及 Transformer 在全局建模时面临的高计算复杂度问题, 提出一种基于特征空间上下文的 Transformer 去噪方法 FSCformer. 该方法设计高效感受野模块, 通过动态捕获多尺度上下文信息, 在增强空间感知能力的同时显著降低计算开销; 采用卷积注意力模块, 将局部特征提取与全局依赖建模有机结合, 提高模型在复杂噪声环境下的鲁棒性; 提出跨特征融合机制, 通过多尺度特征的精细化交互增强图像细节保留能力. 大量实验结果表明, 该方法在去噪精度与计算效率之间实现了良好平衡, 并在多个基准数据集上优于现有的多种图像去噪模型.

关键词: 图像去噪; Transformer; 大感受野; 特征融合

中图法分类号: TP18

中文引用格式: 胡雨轩, 张师超. 用于图像去噪的特征空间上下文Transformer. 软件学报. http://www.jos.org.cn/1000-9825/7569.htm

英文引用格式: Hu YX, Zhang SC. Feature Spatial Contextual Transformer for Image Denoising. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7569.htm

Feature Spatial Contextual Transformer for Image Denoising

HU Yu-Xuan[1], ZHANG Shi-Chao[2,3]

[1](School of Computer Science and Engineering, Central South University, Changsha 410083, China)

[2](Key Laboratory of Education Blockchain and Intelligent Technology (Guangxi Normal University), Ministry of Education, Guilin 541004, China)

[3](Guangxi Key Lab of Multi-source Information Mining and Security (Guangxi Normal University), Guilin 541004, China)

Abstract: Image denoising, a fundamental task in computer vision, relies on leveraging effective prior knowledge to restore detailed information corrupted by noise. This study proposes a feature spatial contextual Transformer for image denoising FSCformer to address the performance bottlenecks of traditional convolutional neural network (CNN) caused by fixed weights and the limitation of local receptive fields, and the high computational complexity of Transformer for global modeling. In this method, an efficient receptive field module is introduced to dynamically capture multi-scale contextual information, enhancing spatial perception while significantly reducing computational overhead. Furthermore, a convolutional attention module is adopted to integrate local feature extraction with global dependency modeling, improving the robustness of the model especially in complex scenes. Additionally, a cross-feature fusion mechanism is proposed to preserve image details by promoting the fine-grained interaction of multi-scale features. Extensive experiments demonstrate that the proposed method achieves a favorable trade-off between denoising accuracy and computational efficiency, outperforming several existing image denoising models across multiple benchmark datasets.

Key words: image denoising; Transformer; large receptive field; feature fusion

* 基金项目: 国家自然科学基金重点项目 (61836016); 广西多源信息挖掘与安全重点实验室开放基金 (MIMS24-02)
收稿时间: 2025-05-05; 修改时间: 2025-07-16, 2025-09-22, 2025-10-09; 采用时间: 2025-10-27; jos 在线出版时间: 2026-03-25

1 引 言

图像去噪是计算机视觉中的基础问题之一, 其核心目标是从受噪声污染的退化图像中恢复出高质量的视觉内容, 为后续诸如图像分类、目标检测等高层次视觉任务提供清晰、可靠的输入基础. 该任务面临的挑战在于, 噪声与图像细节在局部区域内往往具有高度相似的特性, 仅依靠局部信息难以实现有效分离. 这种病态性问题的本质要求必须引入有效的图像先验知识以引导模型完成准确的图像恢复. 因此, 如何设计强大且高效的去噪方法始终是该领域内研究的重点与难点.

传统的图像去噪方法包括基于滤波和优化的算法, 尽管在早期研究中取得了一定进展, 但通常依赖手工设计的特征提取器和较强的模型假设, 导致在复杂噪声环境下的泛化能力和鲁棒性往往不足. 近年来, 卷积神经网络 (convolutional neural network, CNN) 凭借其局部连接和权重共享等固有归纳偏差优势[1]在图像去噪任务上展现出卓越性能, 逐渐成为主流方案. 具体而言, Zhang 等人[2]将具备灵活架构的 CNN 引入图像去噪领域, 通过堆叠卷积层、批归一化层和 ReLU 激活函数等模块显著提高了去噪效果. 深度迭代上下采样卷积神经网络 (deep iterative down-up convolutional neural network, DIDN)[3]进一步通过反复下采样和上采样特征图的方式扩展感受野并提升了去噪性能. FRRN[4]则通过选择具有代表性的特征表示并抑制空间与通道维度上的冗余, 实现了高效去噪.

尽管 CNN 在图像去噪中取得了显著成果, 但其固有的局限性也不容忽视[5]: (1) 在推理过程中, CNN 通常使用固定、非动态的卷积核权重, 限制了其对输入数据多样性的灵活适应能力; (2) CNN 所采用的卷积操作受限于其局部感受野, 因此在建模全局像素之间的长程依赖关系时存在困难, 导致其在处理包含复杂噪声的图像时难以捕捉关键的全局信息. 为了克服这些局限性, 研究者们开始将目光投向基于 Transformer 的架构. Transformer[6]最早应用于自然语言处理领域, 其核心的自注意力机制 (self-attention) 能够在全局范围内捕捉元素间的长程依赖关系, 因此在图像处理中同样能有效建模像素间的远程相互作用, 从而增强去噪性能. 此外, Transformer 通过动态学习的权重提高了处理图像时的灵活性, 能够自适应不同类型的噪声模式. 然而, 将 Transformer 直接应用于图像去噪仍然面临严峻挑战, 其中最突出的问题是其高昂的计算复杂度. 标准的自注意力机制需要计算每个像素与其他所有像素之间的相似度, 导致计算量随图像尺寸的增大而呈二次方增长, 极大地限制了 Transformer 在大规模图像去噪中的应用[7]. 为了突破这一计算瓶颈, 研究者提出构建卷积与 Transformer 相结合的混合架构, 利用卷积操作高效提取局部特征, 同时借助 Transformer 机制建模全局上下文, 有效兼顾性能和效率, 尝试降低整体计算开销. 例如, Zhao 等人[8]提出了 CNN-Transformer 混合架构, 采用先卷积后 Transformer 的串行设计, 融合了局部细节和全局语义, 有效地提高了去噪性能. 综上, 降低计算开销成为 Transformer 应用于图像去噪领域亟待解决的关键问题之一. 另一方面, 图像去噪不仅需要移除噪声, 还要求恢复图像中边缘和纹理等细节信息. 单纯依赖全局自注意力机制虽能捕获长程依赖, 但可能会平均化局部特征, 导致细节模糊; 而纯局部处理又无法充分利用图像的整体结构信息. 因此, 设计能够合理平衡局部特征提取与全局上下文建模, 并实现高效多尺度特征融合的机制, 是提升去噪性能与实用性的关键. 针对上述问题, 本文提出了一种用于图像去噪的特征空间上下文 Transformer (feature spatial contextual Transformer, FSCformer), 其主要贡献总结如下.

(1) 通过高效感受野模块的分解卷积和注意力机制提取局部和全局特征, 使模型能够以较低的计算代价捕捉不同位置的上下文信息, 增强了模型的空间感知能力.

(2) 采用卷积注意力模块将卷积的局部性先验与注意力机制的全局交互能力融合, 在保持高效局部特征提取的同时提升全局信息建模的精度与鲁棒性.

(3) 提出的跨特征融合机制通过精细化的多尺度特征交互充分促进了编码器与解码器路径间的信息流动, 有效保留了图像细节, 并增强了对噪声的抑制能力.

2 相关工作

2.1 基于 Transformer 的图像去噪方法

Transformer[6]凭借其卓越的全局依赖建模能力已在计算机视觉任务中得到广泛应用, 涵盖高层次任务[9−13]到

下游任务[7,14,15]. Vision Transformer (ViT)[16]作为开创性工作, 首次将图像分割为补丁序列, 并通过多头自注意力机制在局部窗口中建模图片块间相互关系. 这一思路为后续研究奠定了基础, 一系列研究[7,17−19]将 ViT 架构应用于图像去噪任务, 相较此前主流 CNN 方法[2−4,20]展现出更为优越的去噪性能和适应性. 具体而言, Restormer[21]借助深度卷积头和转置注意力机制有效建模全局上下文, 并引入门控模块控制特征流动. 为了进一步增强局部窗口间的直接交互, 交叉聚合 Transformer (cross aggregation Transformer, CAT)[22]采用水平和垂直矩形窗口的并行注意力机制扩展了注意力区域, 增强了窗口间交互, 显著提升了模型的表达能力. DSTrans[23]采用双流注意力机制, 在光谱和空间维度上分别捕捉全局和局部特征, 在细粒度特征处理上表现突出. 然而, 这些方法普遍伴随着较高的计算复杂度, 限制了其在实际场景中的应用. 为了进一步挖掘 Transformer 在去噪任务中的潜力, 研究者们提出多种改进方案. Swin Transformer[24]通过层级化滑动窗口化注意力机制实现了跨窗口交互, 在大幅降低计算成本的同时确保特征提取的完整性和准确性. 此外, Uformer[14]结合了 U-Net 和 Transformer 的优势, 在多尺度特征中实现细粒度的特征融合, 显著提升了去噪效果. 另一个值得关注的研究是 IPT (image processing Transformer)[15], 该方法通过引入大规模预训练模型进一步增强了对复杂图像噪声的处理能力, 但其高昂的计算成本和庞大的模型规模限制了其在资源受限环境中的实际应用. EfficientFormer[25]则专注于轻量化设计, 提出了舍弃冗余计算的稀疏注意力机制, 在保持较高去噪性能的前提下显著提升了模型的运行效率. 现有 CNN-Transformer 混合架构通过结合全局建模与局部归纳偏差为图像复原任务提供了更强的先验, 但它们仍存在一些局限性的权衡. 例如, Uformer[14]基于固定层级的多尺度结构, 其跳跃连接在信息压缩后可能面临特征对齐的挑战. Restormer[21]的门控 FFN 虽能有效筛选全局与局部信息, 但其引入的深度卷积与门控机制也带来了额外的计算开销. 而 CAT[22]通过矩形窗口灵活地建模了长程空间依赖, 但缺乏对通道维度权重的显式建模, 可能限制其恢复复杂纹理的能力. 受此启发, FSCformer 通过水平/垂直大核分解结合组移位操作实现动态稀疏频域解耦, 有效分离低频噪声与高频细节, 提供方向性感知以缓解局部噪声主导问题, 从而提升上下文捕获的自适应能力

2.2 大卷积核

ViT 凭借其全局感受野优势在计算机视觉领域中占据了重要地位. 与此同时, 研究者们也重新审视了 CNN 架构的潜力, 其中采用大卷积核扩展感受野成为提升 CNN 性能的重要方向之一. 在探索 Transformer 与 CNN 之间的关系及其混合架构后, Liu 等人[26]率先将大卷积核和深度卷积结合, 在不依赖自注意力机制的情况下实现了高效特征提取, 并将此范式成功应用于 ResNet[27]架构, 为传统 CNN 的发展带来了新的思路. 随后, Guo 等人[28]在此基础上将大卷积核分解为空间局部卷积、空间长程卷积和通道卷积这 3 个模块, 在通道和空间维度上实现了对特征的自适应提取, 为后续研究奠定了基础. 后续研究从多个方向推进了大卷积核技术的发展. 在分解策略方面, RepLKNet[29]借助特征重参化技术将卷积核大小扩展到大规模的同时保持了模型的高效性; 在稀疏优化方面, SLaK[30]通过特征分解和稀疏训练策略, 进一步扩大了卷积核尺寸. PeLK[31]则在稀疏矩阵优化的基础上将卷积核扩展到更大尺寸. HKCNN[32]采用异构大卷积核组合, 通过捕捉特征间的长程关系提升了模型的鲁棒性.

研究者们进一步探索将大核卷积嵌入 Transformer 块中模拟自注意力机制, 同时增强特征提取的动态性. HorNet[33]使用大卷积核捕捉长程依赖并通过门控机制模拟 Transformer 的注意力动态性, 实现了高效的全局建模. Hou 等人[34]提出了分层卷积 Transformer 架构, 使用大核深度卷积增强空间混合并结合通道注意力实现对局部和全局特征的自适应融合. TCNet[35]通过双分支大核 CNN-Transformer 混合架构与动态注意力融合, 显著提升了恶性甲状腺结节分割精度. 而 InternImage[36]则采用改进的可变形卷积作为核心算子, 借鉴 Transformer 的架构构建纯 CNN, 通过动态偏移和自适应加权捕捉长程依赖与非局部上下文信息, 显著提升了 CNN 在大型视觉基础模型上的性能. Conv2Former[37]使用共享大卷积核跨网络扩展感受野并结合动态内核捕捉输入相关权重, 同时引入 Flash Attention 优化窗口大小, 显著降低了内存使用和延迟. 这些工作共同展示了 Transformer 与大卷积核结合的趋势: 通过大核扩展感受野和动态机制优化全局建模, 从而平衡效率和性能. 然而现有方法仍存在改进空间, RepLKNet[29]的固定分解策略在降低复杂度的同时可能忽略了方向异质性, 导致边缘细节丢失. SLaK[30]的稀疏机制虽优化计算效率但可能影响局部细节恢复. ConvNeXt[26]的大核堆叠方案虽高效却缺乏动态适应能力. 针对这些局限, 本文通

过大核分解辅以额外小核的设计协调全局上下文与局部纹理提取, 同时通过组移位动态调整空间依赖关系, 并结合通道注意力机制聚焦有效信息. 这种综合性设计在保持硬件友好性的同时有效提升了噪声抑制与细节保持能力.

2.3 特征融合

特征融合机制是将来自不同来源或者模态的信息整合生成统一的特征表示, 从而增强模型对数据的理解与预测能力. He 等人[27]提出的跳跃连接将低层特征与高层特征结合, 显著提升了图像分割任务的性能. FPN[38]引入了多尺度特征融合的概念, 通过特征金字塔网络将不同网络层的特征图结合, 增强了不同粒度信息的表示能力. Su 等人[39]利用可学习的语义偏移场对齐多级特征图并进行融合, 有效解决了特征错位问题. PSPNet[40]采用空间注意力机制, 在特征融合过程中有选择性地突出重要的空间区域, 提升了语义分割的精度. 类似地, Fu 等人[41]提出了密集注意力网络, 通过加权注意力机制融合多层次特征更有效地处理空间与上下文信息. 殷崭祚等人[42]通过自注意力机制自适应地捕捉深层语义信息, 并对这些深层语义特征进行细粒度融合, 提高了检索任务中的性能. 在多模态融合方面, Kazakos 等人[43]通过集成颜色和空间深度信息以提高人体姿态估计精度. 而孙海峰等人[44]进一步提出了双模态信息量的动态门控融合模块, 根据输入深度稀疏分布动态生成融合权重, 成功实现了性能和速度的平衡. 此外, 自适应特征融合的引入在提高模型效率方面也展现了较大优势. Li 等人[45]设计了动态特征融合模型, 通过自适应权重机制实现了更加灵活且精确的特征表示. 受到上述研究启发, 本文提出了一种跨特征融合机制, 该机制结合了卷积的局部感知优势和注意力机制的全局建模能力, 自适应实现编码器-解码器间特征的高效融合, 从而促进深层语义信息与浅层空间细节的跨特征交互, 提高了去噪器的性能和鲁棒性.

3 本文方法

FSCformer 针对图像去噪的病态性和细节恢复需求, 设计了协同工作的核心模块: 高效感受野模块负责鲁棒特征提取基础, 而双阶段融合模块, 即卷积注意力模块与跨特征融合块, 则优化全局与局部信息的融合. 如图 1 所示, FSCformer 的整体架构与 DnCNN[2]一致, 包含 3 个主要部分: 浅层特征提取、深层特征提取以及图像重建. 浅层特征提取部分包含一个 3×3 卷积层从含噪图像 $I_n \in \mathbb{R}^{H\times W\times C_{\text{in}}}$ 中提取浅层特征 $F_0 \in \mathbb{R}^{H\times W\times C}$. 其中, H 和 W 为图像空间尺寸, 而 C_{in} 为输入通道数 (灰度图像为 1, 彩色图像为 3), 中间特征通道数 C 统一设置为 64. 深层特征提取部分由高效感受野模块 (efficient receptive field block, ERFB) 和双阶段融合模块 (dual-stage fusion module, DSFM) 级联构成, 负责从 F_0 中提取深层特征 $F_D \in \mathbb{R}^{H\times W\times C}$. 图像重建部分通过一个 3×3 卷积层结合全局残差连接, 将 F_D 重建为干净图像 $I_c \in \mathbb{R}^{H\times W\times C_{\text{out}}}$. 其中输出通道数 C_{out} 与输入保持一致 (灰度图为 1, 彩色图为 3). FSCformer 的整体架构可以通过公式 (1) 进行描述:

$$\begin{aligned} I_c &= I_n - FSCformer(I_n) \\ &= I_n - DIR(DFE(SFE(I_n))) \\ &= I_n - Conv_{3\times3}(DSFM(ERFB(Conv_{3\times3}(I_n)))) \end{aligned} \tag{1}$$

其中, $FSCformer(\cdot)$ 在数学上表示所提出的 FSCformer 模型, DIR 表示图像重建部分, DFE 表示深层特征提取部分, SFE 表示浅层特征提取部分, $Conv_{3\times3}$ 表示 3×3 卷积层. 此外, – 表示残差操作 (等同于图 1 中的⊕符号).

3.1 损失函数

本文采用均方误差 (mean square error, MSE) 损失作为学习目标以更新 FSCformer 的参数. 给定训练样本对 $\left\{I_{\text{gt}}^i, I_n^i\right\}(1 \leqslant i \leqslant N)$ 被用来训练去噪器 FSCformer, 并通过最小化损失函数 $\mathcal{L}_{\text{MSE}}$ 学习最优模型. 其中, N 表示训练样本的数量, I_{gt}^i 和 I_n^i 分别代表训练数据集中第 i 张真实图像 (ground truth, GT) 及其对应的含噪图像. 损失函数 $\mathcal{L}_{\text{MSE}}$ 的数学表达式如公式 (2) 所示:

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|FSCformer\left(I_n^i\right) - I_{\text{gt}}^i\right\|^2 \tag{2}$$

其中, θ 是 $FSCformer(\cdot)$ 的所有参数. 此 Adam[46]优化器被用于优化该过程中的参数更新, 且 $\beta_1 = 0.9$, $\beta_2 = 0.99$. 该过程可通过公式 (3) 表达:

$$\begin{cases}\hat{\theta}=\arg\min\limits_{\theta}\mathcal{L}_{\text{FSCformer}}\\ \theta\leftarrow\theta-\eta\cdot\nabla_{\theta}\dfrac{1}{B}(\mathcal{L}_{\text{FSCformer}})\end{cases}\tag{3}$$

其中, η 表示参数更新过程中的学习率, B 是训练数据集中小批量的样本数量.

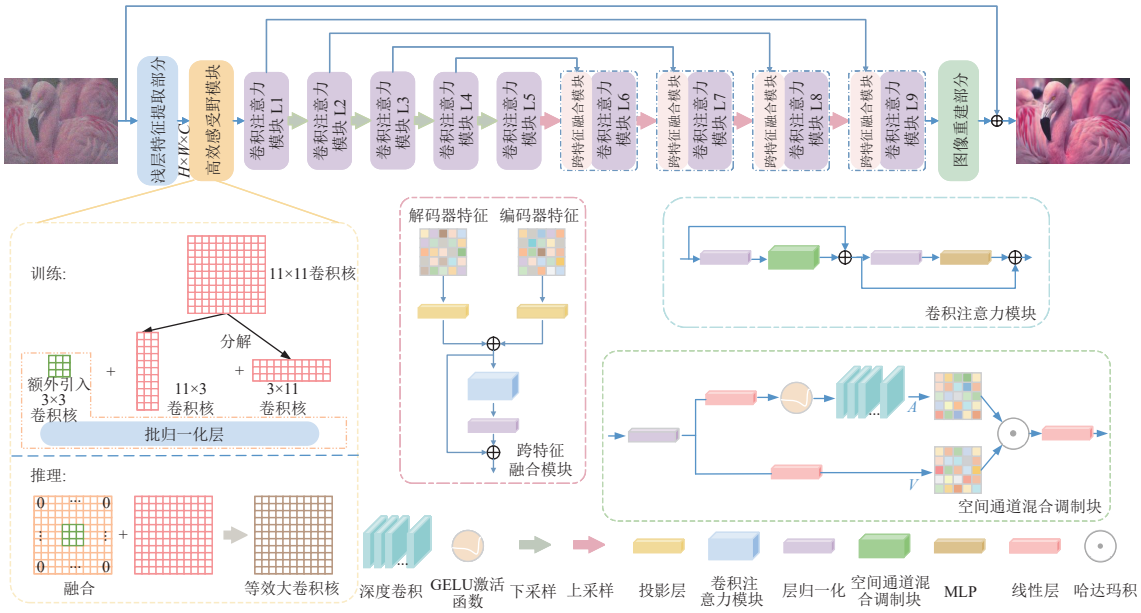


图 1 FSCformer 整体网络架构图

3.2 高效感受野模块

在含噪图像中, 噪声和图像细节在局部区域可能非常相似, 仅凭局部信息难以区分两者. 较大的感受野能够整合广泛上下文信息为模型提供全局结构理解, 从而辅助准确分离真实细节与噪声. 传统小卷积核方法虽擅长局部特征提取, 但受限于有限的感受野, 难以保留对高保真去噪至关重要的空间感知能力. 受文献 [32] 启发, 本文在深层特征提取部分设计大感受野模块以提升去噪效果. 然而, 直接增大卷积核尺寸会显著增加参数量和计算复杂度, 导致训练困难. 此外, 图像中的噪声通常遍布全频带, 而具有方向性的边缘和纹理等关键细节主要分布于高频分量, 大范围平滑区域则对应低频分量. 传统单一尺度卷积核的固定频域响应难以自适应处理这种频带混杂的信号.

为解决这些问题, 本文引入分解式大卷积核策略 [47], 将标准大卷积核分解为水平和垂直方向的组合, 并辅以小卷积核. 水平与垂直方向的卷积大核分别专注于捕捉对应方向的长程上下文特征, 形成互补的方向性频域感知; 同时引入的 3×3 卷积核专门负责提取局部高频细节, 实现了从全局结构到局部细节的完整特征覆盖. 在扩大感受野的同时, 本模块通过通道注意力机制解决可能引入的通道冗余问题. 该机制通过全局平均池化操作对空间信息进行压缩和聚合, 并利用全连接层和 Sigmoid 函数构成的瓶颈结构动态生成通道权重, 从而抑制噪声通道并增强关键特征通道. 该综合策略在显著扩大感受野的同时有效控制了参数量和计算复杂度, 实现了性能与效率的平衡.

在训练阶段, ERFB 将原始大卷积核 $K_{\text{large}}\in\mathbb{R}^{C\times C\times k\times k}$ 分解为水平方向卷积核 $K_{\text{horiz}}\in\mathbb{R}^{C\times C\times k\times s}$ 和垂直方向卷积核 $K_{\text{vert}}\in\mathbb{R}^{C\times C\times s\times k}$, 其分解关系满足 $K_{\text{large}}\approx K_{\text{horiz}}*K_{\text{vert}}$. 其中, $*$ 表示卷积操作的外积近似, 在本文中 k 的值被设置为 11. 假设输入特征图为 $F_D\in\mathbb{R}^{H\times W\times C}$ (亦记作 X), 则卷积部分的输出特征图 F_L (亦记为 Y') 的数学表达式如公式 (4) 所示:

$$F_L=\underbrace{Conv_{k\times s}(F_D,K_{\text{horiz}})}_{\text{水平卷积}}+\underbrace{Conv_{s\times k}(F_D,K_{\text{vert}})}_{\text{垂直卷积}}+\underbrace{Conv_{s\times s}(F_D,K_{\text{small}})}_{\text{小核卷积}}\tag{4}$$

其中, $K_{\text{small}}\in\mathbb{R}^{C\times C\times s\times s}$ 为 $s=3$ 的小卷积核, 用于增强局部特征提取, 补充大卷积核可能忽略的细节. 随后进行批量

归一化 (batch normalization, BN) 处理, 得到输出特征图 Y:

$$Y = \gamma \cdot \frac{Y' - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{5}$$

其中, μ 和 σ^2 分别为特征图的均值和方差, γ 和 β 是可学习的缩放和偏置参数.

为进一步提升空间上下文建模效率, 该模块利用基于组的空间移位操作将输入特征图 X 分为 G 组, 对每组特征图进行独立的卷积操作, 其空间移位操作可表述为:

$$Y_g = Conv_{k\times k}(X_g),\ g = 1,2,\ldots,G \tag{6}$$

其中, X_g 和 Y_g 分别是第 g 组的输入、输出特征图. 随后, 通过对每组卷积结果进行重排模拟移位效果, 并将重排后的特征图相加, 得到卷积层的输出 Y'. 该操作通过动态计算填充量, 并利用分组卷积模拟不同方向的移位效果, 并结合特定窗口滑动方式, 有效捕捉特征图在不同空间位置的上下文信息, 从而增强模型空间感知能力. 同时, 该操作通过减少冗余区域的计算, 显著提升了计算效率, 体现了稀疏卷积的典型优势. 与 Restormer[21]等采用标准自注意力机制或传统卷积操作相比, ERFB 在网络浅层以线性复杂度实现了有效的长距离依赖建模, 为深层特征提取提供了更具表征能力的上下文基础. 该过程可通过公式 (7) 表示:

$$\begin{cases} Y_{\text{shifted}} = Rearrange(Y_g) \\ Y' = \sum_{g=1}^{G} Y_{\text{shifted}} \end{cases} \tag{7}$$

其中, Y_{shifted} 表示移位后的特征图, $Rearrange(\cdot)$ 表示特征在空间维度上的移位操作. 然而, 大卷积核在捕获长程依赖的同时其通道维度可能存在冗余与噪声残留. 为了增强模型的特征选择能力与鲁棒性, ERFB 在 BN 层后引入了轻量级通道注意力模块. 该模块通过动态抑制背景响应, 完善了基于方向性分解的细粒度优化, 形成完整去噪机制: 大核负责提取粗粒度上下文, 小核补充局部细节, 通道注意力则聚焦有效信息. 该设计实现了全局上下文与局部纹理的精细平衡, 克服了大卷积核对细微特征建模的不足, 从而在保持大感受野优势的同时显著提升了对图像结构与纹理信息的感知能力. 具体实现如公式 (8) 所示, 首先对输入特征进行全局平均池化操作, 随后通过两层全连接网络 (中间使用 ReLU 激活函数) 生成通道缩放因子, 最终经 Sigmoid 函数归一化后与原始特征进行逐通道乘法操作, 完成通道维度的自适应重校准.

$$\begin{cases} scale = \sigma(FC(GAP(Y))) \\ O_{\text{ERFB}} = scale \cdot Y \end{cases} \tag{8}$$

其中, $GAP(\cdot)$ 表示全局平均池化, $FC(\cdot)$ 包含两层 MLP, ReLU 在两层 MLP 之间, σ 表示 Sigmoid 函数, $scale$ 对特征图各通道进行重新加权. 与 RepLKNet[29]和 SLaK[30]等仅依赖大卷积核建模长距离依赖的方法相比, ERFB 通过通道注意力机制有效缓解了大核可能引起的信息冗余问题. 该通道注意力模块结构紧凑, 具有极低的参数量与计算开销, 不会增加额外的运行时间或内存负担. 因此, ERFB 在保持大感受野和高计算效率的基础上进一步强化了特征选择能力与去噪鲁棒性.

在推理阶段, 为提升计算效率, 小卷积核层与 BN 层进行参数融合, 生成等效大卷积核避免重复计算. 融合后的权重和偏置计算如公式 (9) 所示:

$$\begin{cases} W_{\text{small\_fused}} = \gamma_{\text{small}} \cdot \dfrac{W_{\text{small}}}{\sqrt{\sigma^2_{\text{small}} + \varepsilon}} \\ b_{\text{small\_fused}} = \beta_{\text{small}} - \dfrac{\gamma_{\text{small}}\mu_{\text{small}}}{\sqrt{\sigma^2_{\text{small}} + \varepsilon}} \end{cases} \tag{9}$$

其中, $W_{\text{small\_fused}}$ 和 $b_{\text{small\_fused}}$ 分别表示融合后的权重和偏置参数, γ_{small}、W_{small} 和 β_{small} 分别是小卷积核层的可学习缩放因子、权重以及偏置参数, μ_{small} 和 σ^2_{small} 是其在训练阶段统计得到的均值和方差. 随后, 将融合后的小卷积核权重置于大卷积核中心位置, 构建等效大卷积核 K_{eq}:

$$K_{\text{eq}} = K_{\text{large}} + Padding(K_{\text{small\_fused}}, pad) \tag{10}$$

其中, $Padding(\cdot, pad)$ 表示将小卷积核置于大卷积核的中心并进行零填充的操作, 使用融合后的权重和偏置参数

初始化等效大卷积核. 最后, 移除训练阶段的小卷积核分支, 通过减少卷积运算次数, 显著降低内存占用并提升推理速度.

3.3 双阶段融合模块

针对局部特征与全局上下文建模的失衡问题, 双阶段融合模块 DSFM 协同整合 U-Net[48]的层次化特征提取和 Transformer[6]的全局依赖建模优势. 编码器-解码器架构在图像去噪中面临信息衰减与语义鸿沟的固有挑战: 编码器通过下采样获得深层语义表征时, 受信息瓶颈原理约束而丢弃部分空间细节; 而解码器进行上采样重建时却亟需这些细节以实现精确恢复. 传统跳跃连接通过直接传递编码器特征以弥补信息损失, 但其粗粒度的补偿机制未考虑编码器特征本身可能含噪, 且与解码器特征存在语义层级错位, 直接融合易导致特征冲突. 尽管跳跃注意力机制[49]通过窗口自注意力提升融合效果, 但仍面临计算复杂度随窗口大小呈二次增长, 以及在局部特征捕捉方面存在不足, 进而影响精细纹理恢复的问题. 为此, 本文设计了跨特征融合模块 (cross feature fusion block, CFFB) 对跨层级信息进行重加权与再校准以实现精细化融合. CFFB 的核心是卷积注意力模块 (convolutional attention block, CAB). 标准 Transformer 的自注意力机制虽能建模全局依赖, 但其计算密集型特性不可避免地带来了较高的计算开销, 且可能导致细节过度平滑. 为此, 本文通过空间-通道联合调制, 以线性复杂度实现高效上下文建模, 从而融合多尺度边缘与语义特征, 提升去噪的细节保真度. 给定输入特征图 $x \in \mathbb{R}^{h\times w\times c}$, 首先通过层归一化 (LayerNorm, *LN*) 层根据通道进行归一化, 得到特征图, 如公式 (11) 所示:

$$\hat{x} = LN(x) \tag{11}$$

其中, $LN(\cdot)$ 表示层归一化操作, 用以提高训练过程的稳定性. 随后, $\hat{x}$ 通过空间-通道混合调制块, 包括相似度分支以及值分支, 联合建模空间上下文和通道交互. 相似度分支由一个 1×1 卷积、GELU 激活函数和一个使用环形填充策略的 11×11 分组卷积层组成, 用于建模空间上下文关系, 而值分支仅包括一个 1×1 卷积, 对特征进行线性投影. 两个分支的输出通过哈达玛积进行融合, 再经 1×1 卷积投影. 最后, 注意力分支乘上一个可学习的缩放因子 α, 并通过残差连接与输入相加, 得到注意力块输出 x_1. 上述过程可以通过公式 (12) 表示:

$$\begin{cases} SCModulation(\hat{x}) = Conv_{1\times1}(Similarity(\hat{x})\odot Value(\hat{x})) \\ \qquad\qquad = Conv_{1\times1}(Conv_{11\times11}(GELU(Conv_{1\times1}(\hat{x})))\odot Conv_{1\times1}(\hat{x})) \\ x_1 = \hat{x} + DropPath(\alpha\cdot SCModulation(\hat{x})) \end{cases} \tag{12}$$

其中, $SCModulation(\cdot)$ 是空间-通道混合调制块, $Similarity(\cdot)$ 和 $Value(\cdot)$ 分别代表相似度分支和值分支, $\odot$ 代表哈达玛积操作, $GELU(\cdot)$ 是 GELU 激活函数, $Conv_{1\times1}$ 和 $Conv_{11\times11}$ 分别为 1×1 和 11×11 卷积. 此外, $DropPath(\cdot)$ 是 DropPath 正则化操作, 被用于增强模型的泛化能力和鲁棒性. 接下来, x_1 通过 MLP 块, 同样乘上一个可学习缩放因子 β, 并通过残差连接与输入相加, 得到输出特征图.

$$O_{\mathrm{CAB}} = x_1 + DropPath(\beta\cdot MLP(x_1)) \tag{13}$$

其中, $MLP(\cdot)$ 表示多层感知机, 用于提取特征的非线性关系. 与使用标准自注意力机制或其变体的方法相比, 卷积注意力模块通过结合卷积的局部性与哈达玛积的交互性在保持较高计算效率的同时有效地捕获全局依赖, 避免了传统自注意力机制带来的高昂计算成本, 构成了本方法在效率上的关键优势.

基于 CAB, CFFB 以编码器特征图 $e \in \mathbb{R}^{h_2\times w_2\times c_2}$ 和解码器特征图 $d \in \mathbb{R}^{h_2\times w_2\times c_2}$ 为输入. 首先通过投影操作将两个输入映射至同一特征空间以对齐语义, 并统一特征通道数, 随后进行特征相加获得初步融合结果. 该融合特征经由 CAB 处理得到最终融合后的特征. 不同于 Restormer[21]等方法的简单特征相加或拼接, CFFB 在投影与初步融合后, 进一步通过卷积注意力块对特征进行深度处理. 该设计促进了不同层级特征的有效交互, 显著增强了细节和纹理的恢复能力, 同时避免了简单融合可能导致的信息冗余. 该过程可以通过公式 (14) 表示:

$$\begin{cases} e' = proj_e(e),\ d' = proj_d(d) \\ fused = e' + d' \\ fused' = fused + DropPath(\alpha\cdot SCModulation(fused)) \\ O_{\mathrm{CFFB}} = fused + LN(fused' + DropPath(\beta\cdot MLP(fused'))) \end{cases} \tag{14}$$

其中, e' 和 d' 分别为 e 和 d 经过投影 $proj(\cdot)$ 后的特征, $fused(\cdot)$ 为初步融合结果, O_{CFFB} 代表模块的最终输出.

DSFM 整体采用对称的解码器-编码器结构. 来自高效感受野块的输出特征图 O_{ERFB} 首先被输入到编码器阶段, 该阶段由 CAB 和下采样层组成, 使用 4×4 卷积 (步长为 2) 进行下采样, 在减少空间维度的同时扩大感受野. 解码器各阶段由上采样层和 CAB 组成, 使用 2×2 转置卷积 (步长为 2) 进行上采样以减少可能出现的棋盘状伪影[50]且恢复图像细节. 各层级间通过 CFFB 连接, 实现跨层级的信息交互. 此设计通过 CAB 的线性复杂度注意力与 CFFB 的精细化融合机制的协同工作, 在保持计算效率的同时有效提升了模型对复杂噪声的鲁棒性及细节恢复能力.

4 实验结果及分析

4.1 数据集

训练数据集: 对于合成灰度与彩色噪声图像的训练, 本文方法遵循 SwinIR[7], 并使用以下数据集: 包含 432 张自然图像的伯克利分割数据集 (BSD)[51]、包含 4744 张自然图像的滑铁卢探索数据集 (WED)[52]、包含 800 张自然图像的 DIV2K 数据集[53]以及包含 2650 张自然图像的 Flickr2K 数据集[54]. 在真实图像去噪方面, 本文选择一个包含 100 张自然图像的真实噪声数据集[55]. 此外, 为了避免过拟合, 在训练过程中以上 5 个数据集中的每张图像均被裁剪为大小为 128×128 的图像块. 每个图像块都将随机应用 8 种数据增强技术中的一种进行处理以增强所提出的 FSCformer 的鲁棒性, 与 DnCNN[2]的实现方式一致.

测试数据集: 相对地, 本文使用公开数据集 BSD68[56]、Set12[56]、CBSD68[56]、Kodak24[57]、McMaster[58]和 Urban100[59]来测试合成灰度与彩色图像去噪模型的有效性, 同时使用 CC[60]数据集来测试真实图像去噪模型的性能.

4.2 实验设置

为了评估本文所提出的 FSCformer 的性能, 本文使用 PyTorch 2.4.0 和 Python 3.9.19 进行了一系列实验. 实验运行在 NVIDIA CUDA 12.1 和 CentOS Linux 7 环境下, 硬件配置为 40 核 Intel(R) Xeon(R) Gold 6248 CPU (主频 2.50 GHz)、384 GB 内存以及一块 NVIDIA Tesla v100 显卡. 图像批量大小和训练轮数分别设置为 16 和 40. 此外, 学习率初始化为 2×10^{-4}, 并在第 15、22、24、26、28 和 30 轮时减半.

4.3 方法分析

本节对所提出的 FSCformer 进行全面的性能分析, 以验证其各个组件的有效性和优越性.

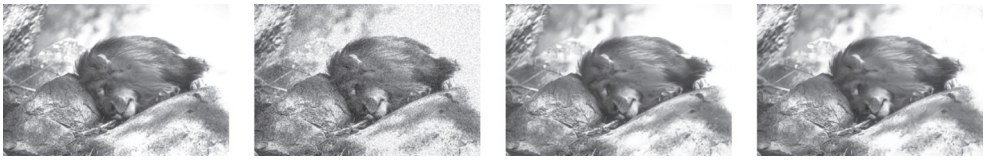
高效感受野模块: 扩大感受野有助于模型捕获更广泛的上下文信息, 但直接使用大卷积核会导致计算量显著增加, 尤其是当网络深度增加时, 计算成本会呈指数级增长. 为此, 本文使用高效感受野模块, 通过将较大 11×11 卷积核分解为水平和垂直方向, 并额外使用 3×3 小卷积核增强局部特征提取. 该分解式设计结合基于组的空间移位操作, 有效捕捉特征空间不同位置的上下文信息, 在提升空间感知能力的同时降低计算开销, 从而改善去噪性能. 如表 1 所示, 包含高效感受野模块的 FSCformer 的表现优于变体 1, 峰值信噪比 PSNR (peak signal-to-noise ratio) 提升了 0.28 dB, 验证了该模块的有效性.

表 1 不同变体在噪声水平为 25 时在 BSD 数据集上的平均 PSNR 值和复杂度

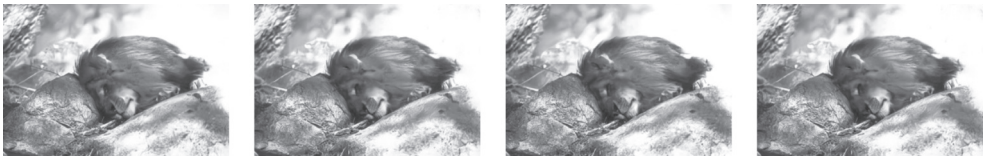
简称	方法	PSNR (dB)	FLOPs (G)	参数量 (M)
Baseline	FSCformer	**29.50**	**11.497**	**28.622**
变体1	FSCformer中缺少高效感受野模块	29.22	11.456	28.618
变体2	FSCformer中用11×11大卷积核替换高效感受野模块	29.33	14.025	28.897
变体3	FSCformer中用3×3小卷积核替换高效感受野模块	29.09	10.167	28.126
变体4	高效感受野模块中缺少额外的3×3小卷积核	29.44	10.917	28.528
变体5	高效感受野模块中缺少通道注意力	29.46	11.497	28.622
变体6	FSCformer中将高效感受野模块放置在网络末端	29.03	11.497	28.622

分解策略在保持表征能力的同时优化了计算效率, 支撑这一结论的证据见表 1 中 FSCformer 和变体 2 的对比, FSCformer 的 FLOPs 和参数量分别下降了 2.528G 和 0.275M, 但基本保持相当的去噪性能. 在图像去噪任务中, 大卷积核通过扩大感受野能够编码更广泛的上下文信息, 增强像素间空间关系的建模能力, 从而提升网络对整体结构的表征能力, 而小卷积核则专注于捕捉高频局部细节. 这一现象可以通过表 1 中变体 1、变体 2 和变体 3 的对比证实. 特别地, 对比 FSCformer 和变体 4 的实验结果, PSNR 值提升了 0.06 dB, 进一步验证了小卷积核对细节恢复的重要贡献. 因此, 将大卷积核的全局感知能力与小卷积核的局部细节提取能力结合, 能够有效增强特征多样性, 提升去噪性能. 此外, 集成于高效感受野模块的通道注意力机制通过增强对图像结构和纹理恢复关键通道的响应, 进一步优化了去噪性能. 消融实验表明, 移除该机制导致 PSNR 下降 0.04 dB, 验证了其在抑制噪声与保留细节方面的有效作用.

模块的位置设计对模型性能具有关键影响. 网络浅层主要负责提取基础特征, 如边缘和纹理等, 而在第 2 层部署高效感受野模块, 可使模型在特征提取早期建立全局感知, 为深层网络理解图像结构提供重要支持. 随着网络层次加深, 特征表征趋于抽象, 此时大感受野对高层特征的增强效果有限, 甚至可能引入不必要的计算复杂度. 如表 1 所示, FSCformer 在 PSNR 值上明显优于变体 6, 说明其位置设计合理且关键. 视觉对比结果 (图 2) 进一步表明, 完整模型在边缘保持和纹理恢复方面表现最优. 当移除 ERFB 时, 图像细节出现明显模糊; 使用 11×11 大卷积核直接替换 ERFB 虽可获得近似去噪效果, 但计算开销增加 22%, 验证了分解策略的效率优势; 而仅使用 3×3 小卷积核则因感受野有限导致结构信息丢失, 在复杂纹理区域产生过度平滑现象. 此外, 缺少辅助 3×3 小卷积核会削弱局部细节捕获能力, 边缘锐化效果下降; 将 ERFB 置于网络末端则因缺乏早期全局建模, 导致整体结构恢复不完整. 这些对比结果共同验证了模块设计和位置安排的有效性.



(a) 给定图像 (b) 含噪图像 (c) FSCformer (d) 缺少ERFB (变体1)



(e) 11×11卷积替换ERFB (变体2) (f) 3×3卷积替换ERFB (变体3) (g) ERFB缺少额外的3×3卷 (变体4) (h) ERFB在网络末端 (变体6)

图 2　不同变体的去噪效果对比

双阶段融合模块: 双阶段融合模块基于 U-Net 架构构建, 充分利用其层次化特征提取与融合能力, 有效平衡局部细节修复与全局结构建模. 然而, 直接堆叠 Transformer 模块会显著增加计算复杂度, 并影响多尺度特征交互效率与细节保留能力. 为此, 本文在 U-Net 框架的特定层级嵌入经过优化的 Transformer 模块, 既保留了卷积操作对局部特征的提取优势, 又通过全局注意力机制提升了去噪性能. 如表 2 所示, 通过 FSCformer 与变体 7 的 PSNR 值对比结果验证了 U-Net 架构与 Transformer 结合的重要性. 当使用标准 Transformer 模块替换卷积注意力块时, FSCformer 中用标准 Transformer (即变体 8) 相较于 FSCformer 的 PSNR 下降了 0.25 dB, 参数量提高了 1.307M, 计算开销也明显上升. 这主要是由于标准 Transformer 缺乏有效的局部特征建模能力, 难以捕捉高频噪声细节, 导致边缘模糊和噪声残留. 同时, 全局注意力的平方复杂度使得模型计算量呈指数增长, 导致计算效率低下. 相比之下, 卷积注意力块通过卷积约束与全局注意力形成互补, 在保持局部细节感知的同时实现高效的全局依赖建模. 此外, FSCformer 采用带有可学习参数的卷积层进行下采样, 相较于传统池化操作能更好地保留关键特征. 实验数据显示, 尽管变体 9 的参数量下降了, 但 FSCformer 在 PSNR 上提升了 0.74 dB, 验证了该设计在减少细节丢失和像素错位方面的有效性.

表 2 不同变体在噪声水平为 15 时在 Urban100 数据集上的平均 PSNR 值和复杂度

简称	方法	PSNR (dB)	FLOPs (G)	参数量 (M)
Baseline	FSCformer	**33.61**	11.497	28.622
变体7	FSCformer中用堆叠的卷积注意力块替代双阶段融合模块	33.23	10.244	28.548
变体8	FSCformer中用标准Transformer替代卷积注意力块	32.36	13.436	29.929
变体9	FSCformer中用池化层替代卷积下采样层	32.87	11.158	25.488
变体10	FSCformer中用通道拼接替代跨特征融合机制	33.21	**4.442**	**12.707**
变体11	FSCformer中使用原始跳跃注意力块替代跨特征融合机制	33.49	11.866	29.184

在特征融合方面, 传统跳跃连接常采用通道拼接方式融合编码器和解码器特征以补偿上采样过程中的细节损失. 然而, 浅层局部细节与深层全局语义的直接拼接容易导致特征冗余与语义不匹配, 影响融合效率与模型收敛. 原始跳跃注意力块[54]通过动态加权机制改进了这一过程, 在边缘区域强化编码器的细节特征, 而在平滑区域侧重解码器全局上下文, 并通过注意力权重抑制噪声区域. 但该方法的计算复杂度与窗口尺寸平方成正比, 全局注意力机制难以针对性处理局部噪声. 本文提出的跨特征融合机制对跳跃注意力机制进行改进, 采用卷积注意力块替代原有结构, 结合大卷积核扩展感受野, 在更高效地整合局部与全局特征的同时降低计算复杂度. 通过环形填充技术有效避免了边界伪影问题, 提升了去噪一致性. 如表 2 所示, 相对于变体 10, 变体 11 的 PSNR 值提升了 0.28 dB, 而本文的跨特征融合机制进一步将性能提高了 0.12 dB. 同时, FSCformer 比变体 11 的参数量减少了 0.562 M, 计算开销也有所降低, 验证了该设计在效率与性能方面的双重优势.

4.4 模型泛化性分析

为评估 FSCformer 对复杂噪声的零样本泛化能力, 本节在 3 个典型图像退化场景进行测试: 暗光噪声、雨纹干扰和雾霾退化. 使用在噪声水平从 0 变化到 55 的条件下训练的盲去噪模型 FSCformer-B 在 LOL[61]、Rain100L[62]、NH-Haze[63]数据集上直接测试, 未进行任何任务特定的微调. 如表 3 所示, 在暗光图像增强任务中, FSCformer 的 PSNR 值达到 19.26 dB, 相较专门设计的 RetinexNet[61]提升 2.49 dB, 这表明 ERFB 的大感受野设计能够有效捕捉暗光条件下的全局结构信息, 而 DSFM 的多尺度特征融合机制有助于在保持细节的同时抑制暗部噪声. 在去雨任务中, FSCformer 在 Rain100L 数据集上取得 29.98 dB 的 PSNR 值, 超越 DerainNet[64] 2.95 dB, 这主要得益于分解大卷积核的水平/垂直分支形成方向感知特性, 能自然适配线性雨纹结构, 而 CFFB 通过多尺度特征加权有效弱化雨线响应. 在去雾任务中, 模型在 NH-Haze 数据集上获得 19.73 dB 的 PSNR 值, 相比 DehazeNet[65]提升了 3.11 dB, 表明 FSCformer 在处理大气散射和颜色偏移等复杂退化模式的鲁棒性. 值得注意的是, FSCformer 虽为图像去噪任务设计, 但在不同类型图像恢复任务中均表现优秀, 体现良好的架构归纳偏置. ERFB 通过分解大卷积核获得的多尺度感受野使模型能够自适应地处理点状噪声、线性雨纹和全局雾气遮挡等不同退化模式; DSFM 的编码器-解码器结构结合混合空间-通道调制机制提供了强大的特征表示能力; 推理阶段的卷积融合策略则保证计算效率的同时增强了模型稳定性. 这些跨场景实验结果验证了 FSCformer 在复杂实际环境中的广泛应用潜力.

表 3 不同应用场景下, FSCformer-B 在不同数据集上的 PSNR 值 (dB)

应用场景	测试数据集	FSCformer-B的PSNR	基线方法	基线方法的PSNR
暗光增强	LOL	19.26	RetinexNet	16.77
去雨	Rain100L	29.98	DerainNet	27.03
去雾	NH-Haze	19.73	DehazeNet	6.62

4.5 与现有方法对比

为了全面评估 FSCformer 的去噪效果, 本节通过 3 个关键指标对其进行定量和定性实验: 定量评估去噪性能(通过 PSNR 评估)、模型效率和定性评估视觉效果. 具体而言, 本节将 FSCformer 与多个现有图像去噪模型进行比较, 包括 TNRD[66]、BM3D[67]、WNNM[68]、HDCNN[69]、FFDNet[70]、DnCNN[2]、ADNet[71]、LKSVD[72]、FOCNet[73]、N^{3}Net[74]、IRCNN[75]、EWT[76]、CDNet[77]、CTNet[78]、DEFformer[79]、NLRN[80]、IPT[15]、NHNet[81]和 DSNet[82]等方法. 此外, 所有表格中的加粗标注表示最佳去噪效果.

为评估 FSCformer 在灰度图像去噪任务中的性能, 本文在基准数据集 BSD68 和 Set12 以及数据集 Urban100 上进行了系统实验. 通过向清晰图像添加特定水平的高斯噪声生成测试样本, 随即计算去噪结果与真实图像之间的 PSNR 值进行定量评估. 表 4 和表 5 的实验结果表明, FSCformer 在噪声水平 (σ) 分别为 15、25 和 50 条件下均表现出优秀的去噪性能. 如表 4 所示, 在 BSD68 数据集上, 当噪声水平为 25 时, FSCformer 相较基线模型 DnCNN[2]提升 0.27 dB, 在噪声水平为 50 的高噪声条件下, 较次优模型 SwinIR[7]提升了 0.04 dB, 这主要得益于 ERFB 的通道注意力机制对背景噪声的有效抑制. 在 Urban100 数据集上, 当噪声水平为 25 时, FSCformer 也超越了几种被广泛认可的去噪算法, 包括 ADNet[71]、NLRN[80]和 LKSVD[72]. 当噪声水平为 15 和 25 时, FSCformer 在 Urban100 数据集上分别以 33.75 dB 和 31.32 dB 的 PSNR 值超越 SwinIR 的 33.70 dB 和 31.30 dB, 体现了 DSFM 的空间-通道联合调制对建筑物规则边缘的强化重建能力. 虽然当噪声水平为 50 时, FSCformer 的性能略低于 SwinIR, 主要原因是在极端噪声破坏结构连续性的情况下, ERFB 分解式大核 (等效 11×11) 的有效感受野 (约 23×23) 的全局建模能力弱于 SwinIR 的窗口自注意力机制. 表 5 显示, 在 Set12 数据集上, 当噪声水平为 15 时, FSCformer 不仅超过 DnCNN 0.39 dB, 还相较于次优去噪模型 FocNet[73]获得了 0.18 dB 的提升. 需要强调的是, 尽管在个别高噪声场景下存在微小性能波动, 但 FSCformer 的计算开销仅为 SwinIR 的约 11% (如后文模型效率比较部分所示, FLOPs 45.99G vs. 420.12G), 在大多数测试条件下达到优于或持平的去噪效果. 这种精度与效率的平衡使得 FSCformer 在资源受限的实际应用场景具有重要实用价值. 因此, FSCformer 在灰度图像去噪任务中被证明是既有效且具有稳定而强大的泛化能力.

表 4　不同方法在噪声水平 15、25 和 50 时在 BSD68 和 Urban100 数据集上的平均 PSNR 值 (dB)

方法	BSD68			Urban100		
	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$
TNRD	31.42	28.92	25.97	—	—	—
BM3D	31.07	28.57	25.62	32.35	29.7	25.95
WNNM	31.37	28.83	25.84	32.97	30.39	26.83
FFDNet	31.63	29.19	26.29	32.43	29.92	26.52
DnCNN	31.72	29.23	26.23	32.64	29.95	26.26
ADNet	31.74	29.25	26.29	32.87	30.24	26.64
LKSVD	31.54	29.07	26.13	—	—	—
FOCNet	31.83	29.38	26.50	33.15	30.64	27.40
N^3Net	31.78	29.30	26.39	33.08	30.19	26.82
IRCNN	31.63	29.15	26.19	32.46	29.80	26.22
SwinIR	**31.97**	**29.50**	26.49	33.70	31.30	**27.98**
EWT	31.90	29.43	26.55	33.57	31.10	27.72
CDNet	31.87	29.41	26.52	33.36	30.68	27.02
CTNet	31.94	29.46	26.49	33.72	31.28	27.80
DEFormer	31.80	29.33	26.40	—	—	—
NLRN	31.88	29.41	26.47	33.45	30.94	27.49
FSCformer	31.94	**29.50**	**26.53**	**33.75**	**31.32**	27.82
FSCformer-B	31.85	29.40	26.42	33.61	30.81	27.75

在彩色高斯噪声图像去噪任务中, 为全面评估 FSCformer 的性能, 本文在 CBSD68、Kodak24 和 McMaster 数据集上进行了扩展测试, 噪声水平覆盖 15、25、35、50 和 75 等多个强度. 如表 6–表 8 所示, 表明 FSCformer 依然在所有测试条件下均展现出强大的竞争力. 具体而言, 在 CBSD68 数据集上, 当噪声水平为 50 时, FSCformer 的 PSNR 值较基于 Transformer 的 IPT[15]模型提升 0.05 dB, 但计算开销仅为其 25% (具体见后文效率分析内容), 大幅降低了推理成本; 在 Kodak24 数据集上, 当噪声水平为 15 时, FSCformer 在噪声水平 15 时超越基线模型 DnCNN[2] 0.10 dB; 在 McMaster 数据集的高噪声条件 (噪声水平为 75) 下, 相较次优模型 ADNet[71]提升了 0.10 dB. 综合数据集的定量结果, FSCformer 在从低到高的所有噪声水平下均保持较优性能, 证明了其在彩色图像高斯去噪任务中的强大泛化能力和实用价值.

表 5 不同方法在噪声水平 15、25 和 50 时在 Set12 数据集上的 PSNR 值 (dB)

噪声水平	方法	图片												平均值
		摄影师	房子	胡椒	海星	蝴蝶	飞机	鹦鹉	贝利	芭芭拉	船	男人	夫妻	
$\sigma = 15$	TNRD	32.19	34.53	33.04	31.75	32.56	31.46	31.63	34.24	32.13	32.14	32.23	32.11	32.50
	BM3D	31.91	34.93	32.69	31.14	31.85	31.07	31.37	34.26	33.10	32.13	31.92	32.10	32.37
	WNNM	32.17	35.13	32.99	31.82	32.71	31.39	31.62	34.27	33.60	32.27	32.11	32.17	32.70
	FFDNet	32.43	35.07	33.25	31.99	32.66	31.57	31.81	34.62	32.54	32.38	32.41	32.46	32.77
	DnCNN	32.61	34.97	33.30	32.20	33.09	31.70	31.83	34.62	32.64	32.42	32.46	32.47	32.86
	HDCNN	32.51	35.17	33.22	32.23	33.20	31.69	31.86	34.57	32.60	32.39	32.36	32.46	32.86
	LKSVD	32.16	34.59	32.92	31.78	32.78	31.54	31.66	34.24	32.22	32.18	32.22	32.11	32.53
	FOCNet	32.71	35.44	33.41	32.40	33.29	31.82	31.98	34.85	33.09	32.62	32.56	32.64	33.07
	DEFormer	32.68	35.57	33.34	32.40	33.12	31.85	31.96	34.85	32.81	32.54	32.48	32.64	33.02
	FSCformer	**32.93**	**35.71**	**33.64**	**32.64**	**33.60**	**32.08**	**32.16**	**34.97**	**33.58**	**32.76**	**32.62**	**32.78**	**33.29**
	FSCformer-B	32.56	34.64	33.32	32.25	33.12	31.75	31.85	34.80	32.97	32.60	32.45	32.58	33.05
$\sigma = 25$	TNRD	29.72	32.53	30.57	29.02	29.85	28.88	29.18	32.00	29.41	29.91	29.87	29.71	30.06
	BM3D	29.45	32.85	30.16	28.56	29.25	28.42	28.93	32.07	30.71	29.90	29.61	29.71	29.97
	WNNM	29.64	33.22	30.42	29.03	29.84	28.69	29.15	32.24	31.24	30.03	29.76	29.82	30.26
	FFDNet	30.10	33.28	30.93	29.32	30.08	29.04	29.44	32.57	30.01	30.25	30.11	30.20	30.44
	DnCNN	30.18	33.06	30.87	29.41	30.28	29.13	29.43	32.44	30.00	30.21	30.10	30.12	30.43
	HDCNN	30.03	33.28	30.75	29.42	30.37	29.11	29.43	32.53	30.03	30.23	30.01	30.14	30.44
	LKSVD	29.70	32.53	30.35	28.99	30.15	28.92	29.13	31.99	29.36	29.95	29.85	29.71	30.05
	FOCNet	30.35	33.63	31.00	29.75	30.49	**29.26**	29.58	32.83	30.74	30.46	**30.22**	30.40	30.73
	N^3Net	30.08	33.25	30.90	29.55	30.45	29.02	29.45	32.59	30.22	30.26	30.12	30.12	30.55
	DEFormer	30.07	33.71	30.90	29.74	30.44	29.28	29.44	32.82	30.45	30.39	30.15	30.38	30.65
	FSCformer	**30.39**	**33.66**	**31.08**	**29.79**	**30.61**	29.25	**29.63**	**32.88**	**30.78**	**30.60**	30.21	**30.41**	**30.79**
	FSCformer-B	30.21	33.52	30.76	29.63	30.54	29.13	29.46	32.63	30.69	30.41	30.12	30.26	30.56
$\sigma = 50$	TNRD	26.62	29.48	27.10	25.42	26.31	25.59	26.16	28.93	25.70	26.94	26.98	26.50	26.81
	BM3D	26.13	29.69	26.68	25.04	25.82	25.10	25.90	29.05	27.22	26.78	26.81	26.46	26.72
	WNNM	26.45	30.33	26.95	25.44	26.32	25.42	26.14	29.25	27.79	26.97	26.94	26.64	27.05
	FFDNet	27.05	30.37	27.54	25.75	26.81	25.89	26.57	29.66	26.45	27.33	27.29	27.08	27.32
	DnCNN	27.03	30.00	27.32	25.70	26.78	25.87	26.48	29.39	26.22	27.20	27.24	26.90	27.18
	HDCNN	27.20	30.04	27.47	25.73	26.89	25.82	26.29	29.50	26.14	27.16	27.23	26.93	27.20
	LKSVD	26.68	29.37	26.96	25.38	26.54	25.62	25.99	28.85	25.73	26.99	26.95	26.55	26.80
	FOCNet	27.36	30.91	27.57	26.19	27.10	26.06	26.75	29.98	27.60	27.53	27.42	27.39	27.68
	N^3Net	27.14	30.50	27.58	26.00	27.03	25.75	26.50	29.76	27.01	27.32	27.33	27.04	27.43
	DEFormer	27.02	31.14	27.58	26.16	27.12	26.10	26.60	29.96	26.86	27.50	27.35	27.39	27.57
	FSCformer	**27.54**	**30.99**	**27.78**	**26.37**	**27.13**	**26.28**	**26.78**	**30.10**	**27.97**	**27.65**	**27.44**	**27.56**	**27.80**
	FSCformer-B	27.15	30.89	27.26	26.13	27.01	25.62	26.50	29.41	27.31	27.35	27.32	27.27	27.49

在真实图像去噪任务中, 本文将 FSCformer 与 TID[83]、CBM3D[67]、ADNet[71]、DnCNN[2]和 DeCapsGAN[84]进行对比实验, 即, 从真实含噪图像数据集 CC 中选择给定的参考图像, 并在不同 ISO 设置下计算去噪图像的 PSNR 值, 以测试 FSCformer 的去噪性能. 如表 9 所示, FSCformer 在大多数测试条件下均优于对比方法, 相较于次优去噪模型 DeCapsGAN[84]提升了 0.36 dB, 在 Nikon D800 ISO 3200 场景下较基线模型 DnCNN[2]提升 3.90 dB. 实验结果表明, FSCformer 对不同 ISO 条件下的真实噪声均展现出优异的适应性与鲁棒性.

在模型效率方面, FSCformer 展现出较大的计算与内存优势. 如表 10 所示, 尽管 FSCformer 的参数量略高于 SwinIR、EDT-B 和 Uformer, 但远低于 IPT 的大规模模型. 随着现代硬件存储成本的逐渐降低, 适度增加的参数量对实际部署的限制影响有限, 因此在参数量上的差异不应成为决定性劣势. 在 FLOPs 指标方面, FSCformer 表现出显著优势, 其 45.99G 的计算量远低于 SwinIR 的 420.12G 和 IPT 的 528.16G, 也显著低于 EDT-B 的 416.06G 和 Uformer 的 156.70G, 表明其在理论计算复杂度上具备较佳的轻量性. 这主要得益于 FSCformer 的高效感受野模块通过分解式大卷积核与分组空间移位操作在保持大感受野的同时控制计算复杂度; 卷积注意力模块采用哈达玛积

实现全局上下文建模, 避免了传统自注意力的大规模矩阵运算. 这些设计不仅降低了计算负担也优化了内存使用效率. 传统 Transformer 的内存复杂度通常为 $O((H\times W)^2)$, 其中 $H\times W$ 是图像像素数量. 相比之下, FSCformer 的核心模块将内存复杂度控制在 $O(H\times W)$ 的线性范围内, 显著降低了处理高分辨率图像时的峰值内存需求. 本文进一步测试了不同模型在 192×192 图片上的实际推理速度. 实验结果表明, FSCformer 的平均推理时间为 0.689 s, 虽略高于 SwinIR 的 0.589 s, 但明显优于 IPT 的 1.055 s 和 EDT-B 的 1.091 s, 虽然与轻量级 Uformer 存在一定差距, 但体现出 FSCformer 在实际速度上的良好平衡. 这种在参数量、计算复杂度和推理速度之间的平衡设计, 使 FSCformer 适合在移动设备、嵌入式系统等资源受限环境中部署, 为实时图像去噪应用提供了实用解决方案.

表 6　不同方法在噪声水平 15、25、35、50 和 75 时在 CBSD68 数据集上的平均 PSNR 值 (dB)

方法	$\sigma=15$	$\sigma=25$	$\sigma=35$	$\sigma=50$	$\sigma=75$
CBM3D	33.51	30.69	28.89	27.36	25.74
FFDNet	33.87	31.21	29.57	27.96	26.24
DnCNN	33.90	31.24	29.65	27.95	—
ADNet	33.99	31.31	29.66	28.04	26.33
IPT	—	—	—	28.39	—
DSNet	33.91	31.28	—	28.05	—
NHNet	34.20	31.54	—	28.31	—
IRCNN	33.86	31.16	29.5	27.86	—
IRPDNN	—	31.24	29.64	28.26	26.39
DEFormer	34.28	31.65	—	**28.47**	—
FSCformer	**34.29**	**31.70**	**30.06**	28.44	**26.62**
FSCformer-B	34.21	31.52	29.93	28.30	26.41

表 7　不同方法在噪声水平 15、25、35、50 和 75 时在 Kodak24 数据集上的平均 PSNR 值 (dB)

方法	$\sigma=15$	$\sigma=25$	$\sigma=35$	$\sigma=50$	$\sigma=75$
CBM3D	34.26	31.67	30.56	28.46	26.82
FFDNet	34.63	32.13	30.56	28.98	27.25
DnCNN	34.60	32.14	30.64	28.95	—
ADNet	34.76	32.26	30.44	29.10	27.40
IPT	—	—	—	29.64	—
DSNet	34.63	32.16	—	29.05	—
NHNet	35.02	32.54	—	29.41	—
IRCNN	34.69	32.18	28.81	29.93	—
IRPDNN	—	32.34	30.81	29.25	—
FSCformer	**34.70**	**32.41**	**31.19**	**29.29**	**27.86**
FSCformer-B	34.61	32.33	30.72	29.12	27.38

表 8　不同方法在噪声水平 15、25、35、50 和 75 时在 McMaster 数据集上的平均 PSNR 值 (dB)

方法	$\sigma=15$	$\sigma=25$	$\sigma=35$	$\sigma=50$	$\sigma=75$
CBM3D	34.03	31.63	29.92	28.48	26.79
FFDNet	34.66	32.35	30.76	29.18	27.29
DnCNN	33.45	31.52	30.91	28.62	—
ADNet	34.93	32.56	31.00	29.36	27.53
IPT	—	—	—	29.98	—
DSNet	34.67	32.40	—	29.28	—
IRCNN	33.86	31.16	29.5	29.98	—
IRPDNN	—	32.33	30.90	29.33	27.59
FSCformer	**34.99**	**32.69**	**31.21**	**30.11**	**28.03**
FSCformer-B	34.75	32.41	31.00	29.89	27.42

表 9 不同方法在真实图片数据集 CC 上的平均 PSNR 值 (dB)

相机型号	TID	CBM3D	ADNet	DnCNN	DeCapsGAN	FSCformer
Canon 5D ISO=3200	37.22	39.76	35.96	37.26	35.74	**37.80**
	34.54	36.40	36.11	34.13	37.02	**37.19**
	34.25	36.37	34.49	34.09	**36.47**	36.31
Nikon D600 ISO=3200	32.99	34.18	33.94	33.62	**35.71**	35.35
	34.20	35.07	34.33	34.48	35.83	**36.46**
	35.58	37.13	38.87	35.41	36.93	**40.76**
Nikon D800 ISO=1600	34.49	36.81	37.61	37.95	38.41	**39.24**
	35.19	37.76	38.24	36.08	39.14	**40.54**
	35.26	37.51	36.89	35.48	37.19	**39.25**
Nikon D800 ISO=3200	33.70	35.05	37.20	34.08	37.68	**39.06**
	31.04	34.07	35.67	33.70	36.85	**37.63**
	33.07	34.42	38.09	33.31	36.85	**39.88**
Nikon D800 ISO=6400	29.40	31.13	32.24	29.83	**33.32**	33.29
	29.86	31.22	32.59	30.55	31.81	**33.93**
	29.21	30.97	33.14	30.09	**33.67**	33.66
平均值	33.63	35.19	35.69	33.86	36.18	**37.36**

表 10 输入图像大小为 192×192 时, 不同方法在图像上的参数量、FLOPs 和推理时间

方法	FLOPs (G)	参数量 (M)	推理时间 (s)
SwinIR	420.12	11.50	0.589
EDT-B	416.06	**11.33**	1.091
Uformer	156.70	11.74	**0.094**
IPT	528.16	115.31	1.055
FSCformer	**45.99**	28.62	0.689

在定性评估方面, 本文将 FSCformer 与 BM3D[67]、FFDNet[70]、DnCNN[2]、IRCNN[75]、ADNet[71]等去噪方法在 BSD68、Set12、Kodak24 和 McMaster 数据集上进行视觉效果对比以验证所提出方法的去噪效果. 为突出去噪效果, 随机选取去噪图像的局部区域进行放大展示, 以纹理细节和轮廓清晰度作为视觉评估标准. 图 3 和图 4 的灰度图像去噪结果显示, FSCformer 在纹理保持方面优于 BM3D[67]、FFDNet[70]、DnCNN[2]、IRCNN[75]和 ADNet[71]. 图 5 和图 6 的彩色图像去噪结果进一步表明, FSCformer 能够恢复出更为清晰的边缘和轮廓信息. 这些视觉对比充分证明了 FSCformer 在细节保留方面的优势.

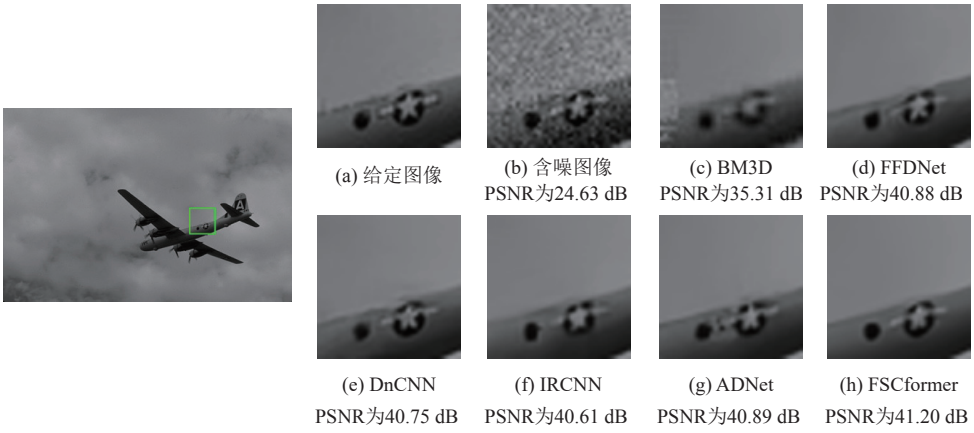


图 3 当噪声水平为 15 时, 不同方法在 BSD68 数据集上的灰度图像去噪结果

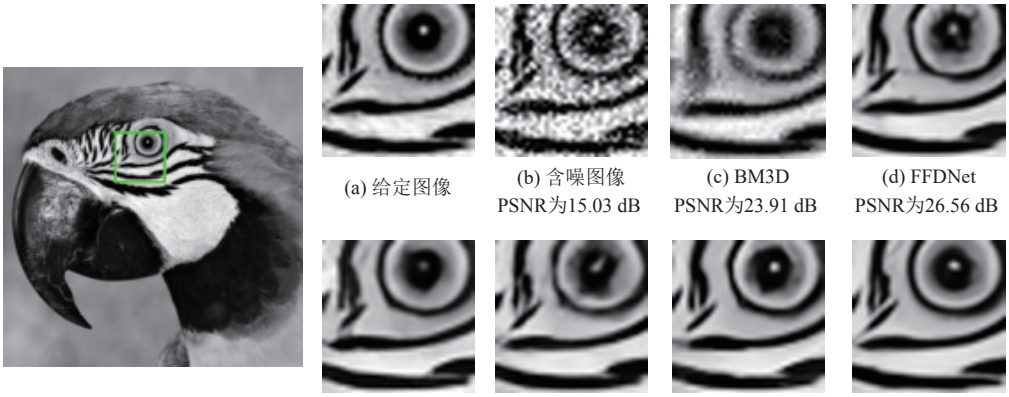


图 4　当噪声水平为 50 时, 不同方法在 Set12 数据集上的灰度图像去噪结果

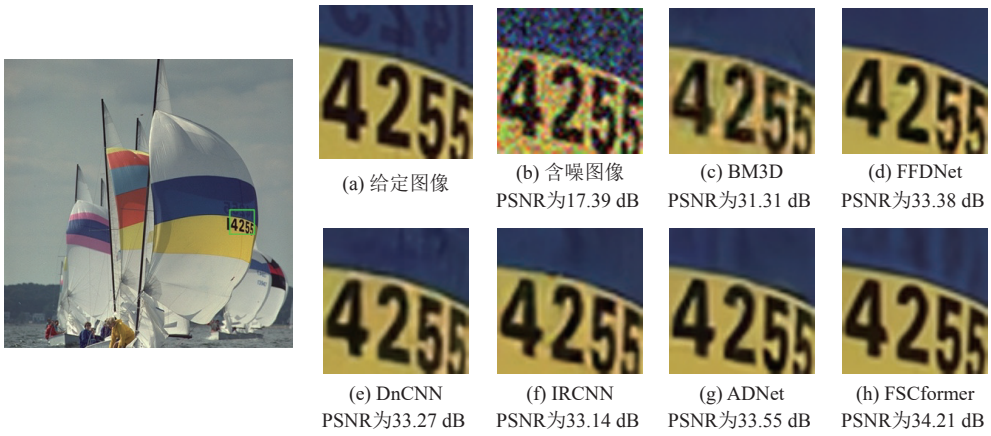


图 5　当噪声水平为 35 时, 不同方法在 Kodak24 数据集上的彩色图像去噪结果

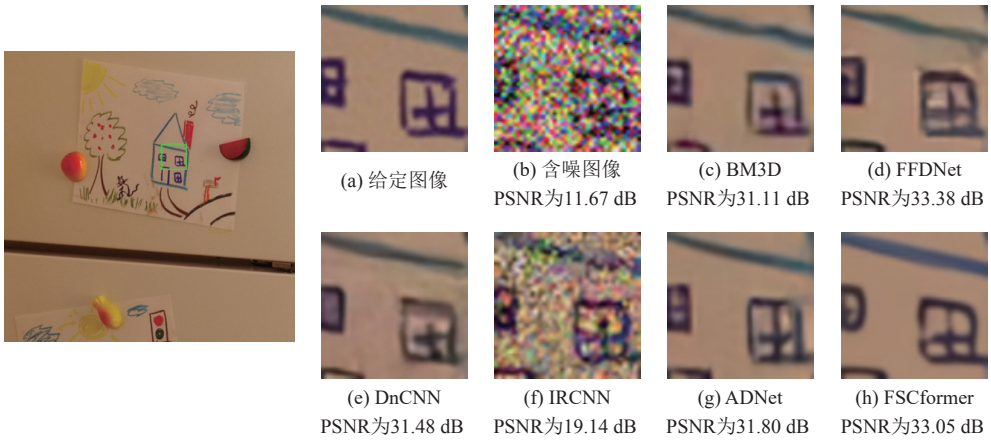


图 6　当噪声水平为 75 时, 不同方法在 McMaster 数据集上的彩色图像去噪结果

综合主客观评估结果, FSCformer 在定量指标、视觉质量和计算效率这 3 个方面均表现出色, 展现了优秀的综合性能. 其在保持竞争性去噪精度的同时实现显著的计算效率提升, 使其在实际应用场景中具有重要价值.

5 结 论

针对图像去噪任务中 CNN 的局部性约束与 Transformer 全局建模的高计算复杂度问题, 本文提出了一种基于特征空间上下文 Transformer 端到端去噪框架, FSCformer. 通过分析现有方法的局限性, 我们从特征表达、上下文建模与多尺度信息融合这 3 个核心维度展开: 高效感受野模块通过分解式大卷积核、辅助小卷积核和基于组的空间移位操作, 并结合通道注意力实现了感受野的动态调整, 有效解决了传统卷积神经网络固定权重带来的适应性不足问题; 卷积注意力模块通过卷积操作和哈达玛积的结合在保持局部特征提取的同时显著降低了自注意力机制的计算冗余, 增强了模型对复杂噪声模式的鲁棒性; 跨特征融合机制通过编码器-解码器间多层级特征的精细化交互, 在抑制噪声传递的同时提升了细节恢复能力. 大量实验结果表明, FSCformer 在合成噪声与真实噪声场景下均展现出优秀性能, 尤其在纹理复杂区域的表现优于现有方法. 未来研究将着重探索动态感受野调整机制与轻量化注意力模块的结合. 以进一步提升模型在移动端设备上的适用性, 并拓展其在低光照增强、医学影像重建等领域的应用潜力.

References

[1] Han Q, Fan ZJ, Dai Q, Sun L, Cheng MM, Liu JY, Wang JD. On the connection between local attention and dynamic depth-wise convolution. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[2] Zhang K, Zuo WM, Chen YJ, Meng DY, Zhang L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. on Image Processing, 2017, 26(7): 3142–3155. [doi: 10.1109/TIP.2017.2662206]

[3] Yu S, Park B, Jeong J. Deep iterative down-up CNN for image denoising. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019. 2095–2103. [doi: 10.1109/CVPRW.2019.00262]

[4] Hu YX, Tian CW, Zhang CY, Zhang SC. Efficient feature redundancy reduction for image denoising. World Wide Web, 2024, 27(2): 20. [doi: 10.1007/s11280-024-01258-3]

[5] Zhao HY, Gou YB, Li BY, Peng DZ, Lv JC, Peng X. Comprehensive and delicate: An efficient Transformer for image restoration. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14122–14132. [doi: 10.1109/CVPR52729.2023.01357]

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[7] Liang JY, Cao JZ, Sun GL, Zhang K, Van Gool L, Timofte R. SwinIR: Image restoration using swin Transformer. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision Workshops. Montreal: IEEE, 2021. 1833–1844. [doi: 10.1109/ICCVW54120.2021.00210]

[8] Zhao M, Cao G, Huang XL, Yang LF. Hybrid Transformer-CNN for real image denoising. IEEE Signal Processing Letters, 2022, 29: 1252–1256. [doi: 10.1109/LSP.2022.3176486]

[9] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[10] Wang YQ, Xu ZL, Wang XL, Shen CH, Cheng BS, Shen H, Xia HX. End-to-end video instance segmentation with Transformers. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8737–8746. [doi: 10.1109/CVPR46437.2021.00863]

[11] Lin K, Wang LJ, Liu ZC. End-to-end human pose and mesh reconstruction with Transformers. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1954–1963. [doi: 10.1109/CVPR46437.2021.00199]

[12] Zhu XZ, Su WJ, Lu LW, Li B, Wang XG, Dai JF. Deformable DETR: Deformable Transformers for end-to-end object detection. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[13] Xie EZ, Wang WH, Yu ZD, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with Transformers. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 924.

[14] Wang ZD, Cun XD, Bao JM, Zhou WG, Liu JZ, Li HQ. Uformer: A general U-shaped Transformer for image restoration. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17662–17672. [doi: 10.1109/CVPR52688.2022.01716]

[15] Chen HT, Wang YH, Guo TY, Xu C, Deng YP, Liu ZH, Ma SW, Xu CJ, Xu C, Gao W. Pre-trained image processing Transformer. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12294–12305. [doi: 10.1109/CVPR46437.2021.01212]

[16] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[17] Chen LY, Chu XJ, Zhang XY, Sun J. Simple baselines for image restoration. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 17–33. [doi: 10.1007/978-3-031-20071-7_2]

[18] Liang YW, Ge CJ, Tong Z, Song YB, Wang J, Xie PT. Not all patches are what you need: Expediting vision Transformers via token reorganizations. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022. 1–21.

[19] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH, Shao L. Multi-stage progressive image restoration. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14816–14826. [doi: 10.1109/CVPR46437.2021.01458]

[20] Hu YX, Zhang SC. Feature representation learning for image denoising. Neurocomputing, 2026, 659: 131787. [doi: 10.1016/j.neucom.2025.131787]

[21] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH. Restormer: Efficient Transformer for high-resolution image restoration. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5718–5729. [doi: 10.1109/CVPR52688.2022.00564]

[22] Chen Z, Zhang YL, Gu JJ, Zhang YB, Kong LH, Yuan X. Cross aggregation Transformer for image restoration. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1847.

[23] Yu DB, Li QW, Wang XL, Zhang ZL, Qian YX, Xu C. DSTrans: Dual-stream Transformer for hyperspectral image restoration. In: Proc. of the 2023 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2023. 3728–3738. [doi: 10.1109/WACV56688.2023.00373]

[24] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

[25] Li YY, Yuan G, Wen Y, Hu J, Evangelidis G, Tulyakov S, Wang YZ, Ren J. EfficientFormer: Vision Transformers at MobileNet speed. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 940.

[26] Liu Z, Mao HZ, Wu CY, Feichtenhofer C, Darrell T, Xie SN. A ConvNet for the 2020s. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976. [doi: 10.1109/CVPR52688.2022.01167]

[27] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[28] Guo MH, Lu CZ, Liu ZN, Cheng MM, Hu SM. Visual attention network. Computational Visual Media, 2023, 9(4): 733–752. [doi: 10.1007/s41095-023-0364-2]

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

[30] Liu S, Chen TL, Chen XH, Chen XX, Xiao Q, Wu BQ, Kärkkäinen T, Pechenizkiy M, Mocanu D, Wang ZY. More ConvNets in the 2020s: Scaling up kernels beyond 51×51 using sparsity. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[31] Chen HH, Chu XX, Ren YJ, Zhao X, Huang KQ. PeLK: Parameter-efficient large kernel convnets with peripheral convolution. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 5557–5567. [doi: 10.1109/CVPR52733.2024.00531]

[32] Hu YX, Tian CW, Zhang J, Zhang SC. Efficient image denoising with heterogeneous kernel-based cnn. Neurocomputing, 2024, 592: 127799. [doi: 10.1016/j.neucom.2024.127799]

[33] Rao YM, Zhao WL, Tang YS, Zhou J, Lim SN, Lu JW. HorNet: Efficient high-order spatial interactions with recursive gated convolutions. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 752.

[34] Hou QB, Lu CZ, Cheng MM, Feng JS. Conv2Former: A simple Transformer-style ConvNet for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8274–8283. [doi: 10.1109/TPAMI.2024.3401450]

[35] Li G, Chen RY, Zhang J, Liu KL, Geng C, Lyu L. Fusing enhanced Transformer and large kernel CNN for malignant thyroid nodule segmentation. Biomedical Signal Processing and Control, 2023, 83: 104636. [doi: 10.1016/j.bspc.2023.104636]

[36] Wang WH, Dai JF, Chen Z, Huang ZH, Li ZQ, Zhu XZ, Hu XW, Lu T, Lu LW, Li HS, Wang XG, Qiao Y. InternImage: Exploring large-

scale vision foundation models with deformable convolutions. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 14408–14419. [doi: 10.1109/CVPR52729.2023.01385]

[37] Lee D, Yun S, Ro Y. Emulating self-attention with convolution for efficient image super-resolution. In: Proc. of the 2025 Int'l Conf. on Computer Vision. Honolulu: IEEE, 2025. 24467–24477.

[38] Lin TY, Dollár P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944. [doi: 10.1109/CVPR.2017.106]

[39] Su YZ, Cheng J, Zhong CQ, Jiang CZ, Ye J, He JJ. Accurate polyp segmentation through enhancing feature fusion and boosting boundary performance. Neurocomputing, 2023, 545: 126233. [doi: 10.1016/j.neucom.2023.126233]

[40] Zhao HS, Shi JP, Qi XJ, Wang XG, Jia JY. Pyramid scene parsing network. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239. [doi: 10.1109/CVPR.2017.660]

[41] Fu J, Liu J, Tian HJ, Li Y, Bao YJ, Fang ZW, Lu HQ. Dual attention network for scene segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149. [doi: 10.1109/CVPR.2019.00326]

[42] Yin ZZ, Li BH, Wang M, Huang RL, Wu WL, Wang HF. Partial multimodal hashing based on fine-grained feature fusion. Ruan Jian Xue Bao/Journal of Software, 2023, 35(3): 1074–1089 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7076.htm [doi: 10.13328/j.cnki.jos.007076]

[43] Kazakos E, Nikou C, Kakadiaris IA. On the fusion of RGB and depth information for hand pose estimation. In: Proc. of the 25th IEEE Int'l Conf. on Image Processing. Athens: IEEE, 2018. 868–872. [doi: 10.1109/ICIP.2018.8451022]

[44] Sun HF, Mu ZY, Qi Q, Wang JY, Liu C, Liao JX. Fast and accurate depth completion method based on dynamic gated fusion strategy. Ruan Jian Xue Bao/Journal of Software, 2022, 34(4): 1765–1778 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6399.htm [doi: 10.13328/j.cnki.jos.006399]

[45] Li Y, Liu JY, Tang ZY, Lei BY. Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification. IEEE Trans. on Medical Imaging, 2020, 39(9): 2818–2830. [doi: 10.1109/TMI.2020.2976825]

[46] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: OpenReview.net, 2015.

[47] Li DC, Li L, Chen ZZ, Li JQ. ShiftwiseConv: Small convolutional kernel with large kernel effect. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 25281–25291. [doi: 10.1109/CVPR52734.2025.02354]

[48] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[49] Agarwal A, Arora C. Attention attention everywhere: Monocular depth prediction with skip attention. In: Proc. of the 2023 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2023. 5850–5859. [doi: 10.1109/WACV56688.2023.00581]

[50] Odena A, Dumoulin V, Olah C. Deconvolution and checkerboard artifacts. Distill, 2016, 1(10): e3. [doi: 10.23915/distill.00003]

[51] Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. of the 8th IEEE Int'l Conf. on Computer Vision. Vancouver: IEEE, 2001. 416–423. [doi: 10.1109/ICCV.2001.937655]

[52] Ma KD, Duanmu ZF, Wu QB, Wang Z, Yong HW, Li HL, Zhang L. Waterloo exploration database: New challenges for image quality assessment models. IEEE Trans. on Image Processing, 2017, 26(2): 1004–1016. [doi: 10.1109/TIP.2016.2631888]

[53] Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1122–1131. [doi: 10.1109/CVPRW.2017.150]

[54] Timofte R, Agustsson E, van Gool L, *et al.* NTIRE 2017 challenge on single image super-resolution: Methods and results. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1110–1121. [doi: 10.1109/CVPRW.2017.149]

[55] Xu J, Li H, Liang ZT, Zhang D, Zhang L. Real-world noisy image denoising: A new benchmark. arXiv:1804.02603, 2018.

[56] Roth S, Black MJ. Fields of experts: A framework for learning image priors. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 860–867. [doi: 10.1109/CVPR.2005.160]

[57] Franzen R. Kodak lossless true color image suite. 1999. https://r0k.us/graphics/kodak/

[58] Zhang L, Wu XL, Buades A, Li X. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. Journal of Electronic Imaging, 2011, 20(2): 023016. [doi: 10.1117/1.3600632]

[59] Huang JB, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5197–5206. [doi: 10.1109/CVPR.2015.7299156]

[60] Nam S, Hwang Y, Matsushita Y, Kim SJ. A holistic approach to cross-channel image noise modeling and its application to image denoising. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1683–1691. [doi: 10.1109/CVPR.2016.186]

[61] Wei C, Wang WJ, Yang WH, Liu JY. Deep retinex decomposition for low-light enhancement. In: Proc. of the 2018 British Machine Vision Conf. Newcastle: BMVA, 2018. 155.

[62] Yang WH, Tan RT, Feng JS, Liu JY, Guo ZM, Yan SC. Deep joint rain detection and removal from a single image. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1685–1694. [doi: 10.1109/CVPR.2017.183]

[63] Ancuti CO, Ancuti C, Timofte R. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020. 1798–1805. [doi: 10.1109/CVPRW50498.2020.00230]

[64] Fu XY, Huang JB, Ding XH, Liao YH, Paisley J. Clearing the skies: A deep network architecture for single-image rain removal. IEEE Trans. on Image Processing, 2017, 26(6): 2944–2956. [doi: 10.1109/TIP.2017.2691802]

[65] Cai BL, Xu XM, Jia K, Qing CM, Tao DC. DehazeNet: An end-to-end system for single image haze removal. IEEE Trans. on Image Processing, 2016, 25(11): 5187–5198. [doi: 10.1109/TIP.2016.2598681]

[66] Chen YJ, Pock T. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1256–1272. [doi: 10.1109/TPAMI.2016.2596743]

[67] Dabov K, Foi A, Katkovnik V, Egiazarian K. Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. on Image Processing, 2007, 16(8): 2080–2095. [doi: 10.1109/TIP.2007.901238]

[68] Gu SH, Zhang L, Zuo WM, Feng XC. Weighted nuclear norm minimization with application to image denoising. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 2862–2869. [doi: 10.1109/CVPR.2014.366]

[69] Zheng MH, Zhi KY, Zeng JW, Tian CW, You L. A hybrid CNN for image denoising. Journal of Artificial Intelligence and Technology, 2022, 2(3): 93–99. [doi: 10.37965/jait.2022.0101]

[70] Zhang K, Zuo WM, Zhang L. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. on Image Processing, 2018, 27(9): 4608–4622. [doi: 10.1109/TIP.2018.2839891]

[71] Tian CW, Xu Y, Li ZY, Zuo WM, Fei LK, Liu H. Attention-guided CNN for image denoising. Neural Networks, 2020, 124: 117–129. [doi: 10.1016/j.neunet.2019.12.024]

[72] Scetbon M, Elad M, Milanfar P. Deep K-SVD denoising. IEEE Trans. on Image Processing, 2021, 30: 5944–5955. [doi: 10.1109/TIP.2021.3090531]

[73] Jia XX, Liu SY, Feng XC, Zhang L. FOCNet: A fractional optimal control network for image denoising. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6047–6056. [doi: 10.1109/CVPR.2019.00621]

[74] Plötz T, Roth S. Neural nearest neighbors networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates, Inc., 2018. 1095–1106.

[75] Zhang K, Zuo WM, Gu SH, Zhang L. Learning deep CNN denoiser prior for image restoration. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2808–2817. [doi: 10.1109/CVPR.2017.300]

[76] Li JC, Cheng BD, Chen Y, Gao GW, Shi J, Zeng TY. EWT: Efficient wavelet-Transformer for single image denoising. Neural Networks, 2024, 177: 106378. [doi: 10.1016/j.neunet.2024.106378]

[77] Ko K, Koh YJ, Kim CS. Blind and compact denoising network based on noise order learning. IEEE Trans. on Image Processing, 2022, 31: 1657–1670. [doi: 10.1109/TIP.2022.3145160]

[78] Tian CW, Zheng MH, Zuo WM, Zhang SC, Zhang YN, Lin CW. A cross Transformer for image denoising. Information Fusion, 2024, 102: 102043. [doi: 10.1016/j.inffus.2023.102043]

[79] Gao YB, Wen ZC, Duan XS, Ma P. Image denoising method based on multi-scale dual-branch Transformer. Journal of Yunnan University (Natural Sciences Edition), 2025, 47(3): 465–474 (in Chinese with English abstract). [doi: 10.7540/j.ynu.20240236]

[80] Liu D, Wen BH, Fan YC, Loy CC, Huang TS. Non-local recurrent network for image restoration. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates, Inc. , 2018. 1680–1689.

[81] Zhang JH, Cao LH, Wang T, Fu WL, Shen WH. NHNet: A non-local hierarchical network for image denoising. IET Image Processing, 2022,16(9): 2446–2456. [doi: 10.1049/ipr2.12499]

[82] Peng YL, Zhang L, Liu SG, Wu XJ, Zhang Y, Wang XL. Dilated residual networks with symmetric skip connection for image denoising. Neurocomputing, 2019, 345: 67–76. [doi: 10.1016/j.neucom.2018.12.075]

[83] Luo EM, Chan SH, Nguyen TQ. Adaptive image denoising by targeted databases. IEEE Trans. on Image Processing, 2015, 24(7): 2167–2181. [doi: 10.1109/TIP.2015.2414873]

[84] Lyu QS, Guo M, Ma M, Mankin R. DeCapsGAN: Generative adversarial capsule network for image denoising. Journal of Electronic Imaging, 2021, 30(3): 033016. [doi: 10.1117/1.JEI.30.3.033016]

附中文参考文献

[42] 殷崭祚, 李博涵, 王萌, 黄瑞龙, 吴文隆, 王昊奋. 基于细粒度特征融合的部分多模态哈希. 软件学报, 2024, 35(3): 1074–1089. http://www.jos.org.cn/1000-9825/7076.htm [doi: 10.13328/j.cnki.jos.007076]

[44] 孙海峰, 穆正阳, 戚琦, 王敬宇, 刘聪, 廖建新. 基于动态门控特征融合的轻量深度补全算法. 软件学报, 2023, 34(4): 1765–1778. http://www.jos.org.cn/1000-9825/6399.htm [doi: 10.13328/j.cnki.jos.006399]

[79] 高煜宝, 文志诚, 段旭升, 马跑. 基于多尺度双分支 Transformer 的图像去噪方法. 云南大学学报 (自然科学版), 2025, 47(3): 465–474. [doi: 10.7540/j.ynu.20240236]

作者简介

胡雨轩, 博士, 主要研究领域为深度学习, 图像复原与增强.

张师超, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为数据挖掘, 知识发现.