计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):39−51 [doi: 10.15888/j.cnki.csa.009555]

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

融合 CNN 和 Transformer 的图像去噪网络①

姜文涛, 卜艺凡

(辽宁工程技术大学 软件学院, 葫芦岛 125105)
通信作者: 卜艺凡, E-mail: byf1098309193@qq.com

摘　要: 目前基于深度学习的图像去噪算法无法综合考虑局部和全局的特征信息, 进而影响细节处的图像去噪效果, 针对该问题, 提出了融合 CNN 和 Transformer 的图像去噪网络 (hybrid CNN and Transformer image denoising network, HCT-Net). 首先, 提出 CNN 和 Transformer 耦合模块 (CNN and Transformer coupling block, CTB), 构造融合卷积和通道自注意力的双分支结构, 缓解单纯依赖 Transformer 造成的高额计算开销, 同时动态分配注意力权重使网络关注重要图像特征. 其次, 设计自注意力增强卷积模块 (self-attention enhanced convolution module, SAConv), 采用递进式组合模块和非线性变换, 减弱噪声信号干扰, 提升在复杂噪声水平下识别局部特征的能力. 在 6 个基准数据集上的实验结果表明, HCT-Net 相比当前一些先进的去噪方法具有更好的特征感知能力, 能够抑制高频的噪声信号从而恢复图像的边缘和细节信息.

关键词: 图像去噪; 深度学习; Transformer; 卷积神经网络; 注意力机制

引用格式: 姜文涛, 卜艺凡.融合 CNN 和 Transformer 的图像去噪网络.计算机系统应用,2024,33(7):39–51. http://www.c-s-a.org.cn/1003-3254/9555.html

Image Denoising Network Fusing with CNN and Transformer

JIANG Wen-Tao, BU Yi-Fan

(Software College, Liaoning Technology University, Huludao 125105, China)

Abstract: The current image denoising algorithms based on deep learning are unable to consider the local and global feature information comprehensively, which in turn affects the image denoising effect at the details. To address this problem, this study proposes a hybrid CNN and Transformer image denoising network (HCT-Net). First, CNN and Transformer coupling block (CTB) is proposed to construct a two-branch structure that integrates convolution and channel self-attention to alleviate the high computational overhead caused by relying solely on the Transformer. At the same time, the attention weights are dynamically allocated so that the network focuses on important feature information. Secondly, the self-attention enhanced convolution module (SAConv) is designed to adopt the progressive combination of modules and nonlinear transformations to attenuate the noise signal interference and identify local features under complex noise levels. Experimental results on six benchmark datasets show that HCT-Net has better feature perception ability than some current advanced denoising methods and can suppress high-frequency noise signals to recover the edge and detail information of images.

Key words: image denoising; deep learning; Transformer; convolutional neural network (CNN); attention mechanism

① 基金项目: 国家自然科学基金 (61172144); 辽宁省自然科学基金 (20170540426); 辽宁省教育厅重点基金 (LJYL049)
收稿时间: 2024-01-08; 修改时间: 2024-02-04; 采用时间: 2024-02-26; csa 在线出版时间: 2024-06-05
CNKI 网络首发时间: 2024-06-07

图像在采集、处理和传输过程中受外部环境以及成像设备技术等因素影响, 产生信息或亮度随机突变的像素点, 恶化图像质量. 图像去噪旨在利用图像序列的上下文信息去除噪声信号、恢复真实图像, 在视频监控、医学影像以及现实工业场景等领域具有重要应用价值.

传统图像去噪算法主要包括基于滤波的方法和基于变分模型的方法. 基于滤波的方法利用低通滤波器将噪声信号从图像中分离出来, 重构干净图像, 如高斯滤波[1]、双边滤波[2]、非局部均值滤波 (non-local means, NLM)[3]、BM3D[4]等. 基于变分模型的方法采用贝叶斯观点将去噪任务转变成最大后验概率 (maximum a posteriori, MAP)[5]问题, 然后通过图像先验信息构造正则化最小优化算法, 根据不同的先验约束可分为全变分模型[6]、稀疏模型[7]、自相似性[8]等. 上述方法需要手动调节参数, 且往往只针对特定的图像结构, 难以解决具有复杂分布的真实噪声问题.

基于深度学习的图像去噪算法主要采用卷积神经网络 (convolutional neural network, CNN), 依靠 CNN 强大的学习能力和良好的网络泛化能力, 克服了传统方法依赖于先验信息和人工特征的局限性, 将去噪问题的研究重点转向了难以参数化的真实噪声. Zhang 等[9]提出一种端到端的卷积神经网络去噪算法 (residual learning of deep CNN for image denoising, DnCNN), 通过残差学习和批量归一化加速去噪网络的训练过程. Zhang 等[10]提出处理空间变化噪声的子图像去噪网络 (toward a fast and flexible solution for CNN-based image denoising, FFDNet), 以噪声水平估计作为网络输入, 采用正交正则化抑制扩张卷积产生的边界伪影. Guo 等[11]提出分阶段盲去噪网络 (toward convolutional blind denoising of real photographs, CBDNet), 通过噪声估计子网络调整噪声水平从而实现交互式去噪, 非盲去噪子网络采用残差学习防止网络过拟合. Anwar 等[12]提出模块化结构的新型单阶段盲去噪网络 (real image denoising with feature attention, RIDNet), 采用特征重标定策略[13]实现通道间的相互依赖关系建模. 但基于 CNN 的图像去噪算法受卷积核范围的限制, 只能处理空间上局部邻域的构造块, 无法捕获长距离特征之间的上下文联系, 使网络缺乏全局信息感知能力, 难以恢复图像的整体形状和结构.

上述方法所提出的去噪模型缺乏对远程特征相关性的建模能力, 而 Transformer[14]能够利用其注意力机制捕获远程像素间的强弱语义关系, 实现上下文信息的全局交互, 弥补了 CNN 在特征映射能力上的不足. Wang 等[15]提出一种基于 Transformer 的 U 型图像恢复架构 (a general U-shaped Transformer for image restoration, Uformer), 在一定程度上缓解了 CNN 的局部限制, 但在复杂情况下仍缺乏对高频噪声信号的抑制能力. Liang 等[16]在 Swin Transformer 的基础上, 提出采用滑动窗口机制的强基线模型 SwinIR (image restoration using Swin Transformer), 通过限制注意力的作用范围来平衡计算效率, 但这种局部自注意力机制与实现上下文信息的全局交互始终相互矛盾. Zamir 等[17]提出针对高分辨率图像的恢复算法 (efficient Transformer for high resolution image restoration, Restormer), 采用通道自注意力和双路门控机制, 改进前馈网络的特征筛选方式, 但完全依赖注意力捕获图像特征的方式造成了高额的计算开销.

目前基于 CNN 和 Transformer 的图像去噪算法, 面对复杂分布的真实噪声时, 主要通过叠加多层池化和降采样操作以获得高级语义信息, 但在前向传播过程中逐渐丢失像素数较少的细节特征, 无法从局部和全局出发综合考虑特征信息, 导致输出的去噪图像仍存在边缘模糊或过度平滑现象, 难以恢复真实图像的细节纹理信息. 另外, 自注意力机制在捕获像素间的远程依赖关系时, 将不可避免的导致 GPU 内存不足和计算效率低下.

针对上述问题, 本文提出融合 CNN 和 Transformer 的图像去噪网络 (hybrid CNN and Transformer image denoising network, HCT-Net). 首先, 提出 CNN 和 Transformer 耦合模块 (CTB), 构造一种卷积和通道自注意力的并行双分支结构, 利用局部连接和共享权重的特性高效提取浅层特征, 捕获通道维度的关键特征、抑制噪声信号. 其次, 设计自注意力增强卷积模块 (SAConv) 提升网络对局部特征的关注度, 在实现多频信号感知融合的同时优化网络前馈信息, 保留更多细微特征从而恢复图像的细节和边缘结构. 此外, HCT-Net 采用编解码器[18]提取低分辨率和多尺度特征, 利用层级连接融合不同粒度的相邻特征, 最大程度保留原始图像的结构信息, 进一步增强前馈网络的特征表达能力和模型泛化能力. 在高斯噪声数据集和真实噪声数据集上的实验结果均验证了本文算法的去噪性能, 并通过算法效率分析证明了本文网络的合理性.

1 相关工作

1.1 Transformer

Transformer 是一种完全基于自注意力机制的编解码网络模型, 每个基本模块由多头自注意力机制 (multi-head self-attention, *MSA*) 和简单的全连接前馈网络 (feed-forward network, FFN) 组成.

自注意力机制通过查询和键值的相似性程度来确定值的权重分布, 计算每一个序列对于其他序列的注意力系数, 从而捕捉数据或特征间的内部相关性. 首先, 使用参数矩阵 W_q、W_k、W_v 对输入矩阵 X 进行线性变换, 得到输入矩阵 X 在查询、键值、值上的投影矩阵 Q、K、V, 其维度分别为 d_q、d_k、d_v. 然后通过 Q 矩阵和 K^{T} 矩阵的点积交互进行相似度计算, 得到自注意力权重矩阵 QK^{T}, 为了防止内积过大, 将 QK^{T} 除以 $\sqrt{d_k}$, 之后对这个权值进行归一化得到 (0, 1) 之间的注意力权重. 通过注意力权重分布判断对应位置信息的重要程度, 与矩阵 V 进行计算得到一个融合注意力的更好的值, 从而提升网络对关键特征的识别能力. 自注意力机制可表示为:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

多头自注意力机制由多个自注意力层并行组成, 首先将输入矩阵 X 分别传递到 h 个不同的自注意力层中, 分别执行式 (1) 的点积交互操作得到注意力权重矩阵 $head_i$, 然后将 h 个输出矩阵 $head_1$ 到 $head_h$ 拼接起来得到一个高维矩阵, 聚合不同维度的特征信息. 再传入线性变换层, 通过点乘权重矩阵 W^O 压缩输出矩阵的高维信息, 得到与输入矩阵 X 维度相同的输出矩阵 X'. 多头自注意力机制可表示为:

$$MSA(Q,K,V)=Concat(head_1,\cdots,head_h)W^O \tag{2}$$

其中, $head_i=Attention(Q_i,K_i,V_i)$, $Q_i=QW_i^Q$, $K_i=KW_i^k$, $V_i=VW_i^v$ 表示不同自注意力层输出的注意力权重矩阵, $i=1,\cdots,h$ 表示多头自注意力的头部数量, 每个并行的自注意力模块捕获不同维度的特征关系.

全连接前馈网络是一种单向多层网络结构, 其中每一层包含若干个神经元, 信号从输入层到输出层单向传播, 整个网络中无反馈信号. 全连接网络具有两层线性层, 第 1 层的激活函数为 ReLU, 第 2 层不使用激活函数, 通过前馈全连接层后输入和输出的维度不变, 得到输出矩阵 X''可表示为:

$$FFN(X')=\max(0,X'W_1+b_1)W_2+b_2 \tag{3}$$

此后, VIT (vision Transformer)[19−21]将 Transformer 迁移到计算机视觉领域, 并在此基础上不断提出新的网络结构, 如 DeiT[22]、Swin Transformer[23]和 LeVIT[24]等.

1.2 深度可分离卷积

如图 1, 深度可分离卷积 (depthwise separable convolution, DSC) 将传统卷积操作拆分成两个独立的步骤, 分别是深度卷积和逐点卷积, 旨在减少神经网络的参数量和计算复杂度, 同时保持模型性能.

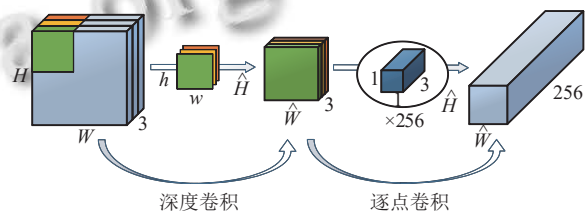


图 1 深度可分离卷积

在传统卷积中, 卷积核是立方体状的, 同时对所有输入通道执行卷积操作, 而在深度卷积中, 每个输入通道都有一个单独的卷积核, 用来在空间上对该通道进行卷积, 使每个通道的卷积核独立地处理输入数据, 产生相应通道的特征图. 由于每个输入通道都有一个独立的卷积核, 大大减少了卷积层的参数数量, 所以深度卷积可以减小模型的复杂度, 降低计算和内存需求. 同时, 深度卷积也可以提高对空间特征的捕捉能力, 因为它在每个通道上分别处理数据, 这样更有助于模型理解输入的局部结构.

逐点卷积采用 1×1 的卷积核, 将不同通道的特征图汇聚在一起, 通过线性组合融合各通道的特征信息, 从而生成最终的输出特征图. 由于逐点卷积允许在不同通道之间进行线性组合和特征交互, 所以其具有更丰富的网络表达能力. 此外, 逐点卷积还可以用来调整输出通道的数量, 进一步降低计算复杂度.

深度可分离卷积将卷积分为两个步骤, 降低了网络的复杂程度, 并减少参数之间的相关性, 降低过拟合的风险, 并使网络拥有更好的泛化能力.

2 融合 CNN 和 Transformer 的图像去噪网络

2.1 网络模型

本文提出融合 CNN 和 Transformer 的图像去噪网

络, 如图 2 所示. 首先, 将噪声图像$I \in R^{H\times W\times 3}$划分为若干可重叠的图像块, 缓解了 Transformer 处理图像序列时会在块周围引入边界伪影的问题. 然后, 通过 3×3 卷积层提取图像的浅层特征$F_0 \in R^{H\times W\times C}$, 其中, H、W 为特征图的空间维度, C 为通道数. 将浅层特征 F_0 输入编解码器中进一步还原图像的真实特征, 采用像素重组作为下采样操作, 每层由 CNN 和 Transformer 耦合模块 (CTB)、自注意力增强卷积模块 (SAConv) 串联组成, 经过多层编码器生成具有局部以及全局依赖关系的特征图. 将低分辨率的潜在特征 F_1 作为解码器输入, 经过 3 次上采样操作转化为具有高级语义的深度特征$F_d \in R^{H\times W\times C}$.

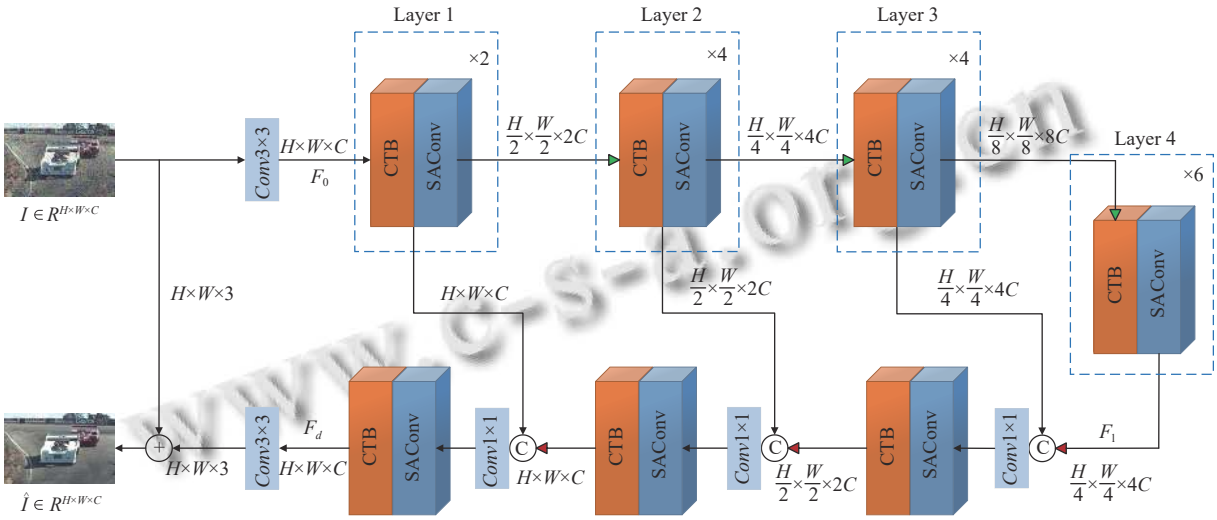


图 2 HCT-Net 网络结构

为了重构下采样阶段造成的信息损失, 将不同层级的特征图进行特征融合, 以确保各层之间的信息流最大化, 避免丢失重要的细节特征. 然后通过 1×1 卷积实现跨通道的信息交互和线性组合, 降低特征通道维度. 最后, 将 3×3 卷积层生成的残差图像$R \in R^{H\times W\times 3}$与噪声图像 I 残差连接, 保留更多原始图像的结构信息, 恢复潜在的干净图像$\hat{I} = I + R$.

2.2 CNN 和 Transformer 耦合模块

CNN 和 Transformer 耦合模块 (CNN and Transformer coupling block, CTB) 结构如图 3 所示, 是一种融合卷积和通道自注意力的并行双分支结构, 并提取二者的相同部分作为公共模块, 从而实现利用效率最大化. 在并行分支中, 一方面利用深度可分离卷积实现高效的浅层特征提取, 另一方面, 采用通道自注意力模块提取特征维度的显著纹理信息, 并缓解全局注意力机制所造成的高额计算开销. 最后, 采用非平均混合策略赋予重要特征更高的权值, 抑制无用特征, 使网络在动态关注重要特征信息的同时减少复杂噪声信号的干扰.

CTB 模块主要分为两个阶段. 第 1 阶段基于卷积和自注意力机制之间的内在相关性, 两种方法共同执行 3 个 1×1 卷积操作, 强化特征提取, 使每组卷积得到 N 个中间特征 F_i:

$$F_i = Conv1\times 1(LayerNorm(F_0))$$

其中, $i = q,k,v$, $Conv1\times 1(\cdot)$代表 1×1 卷积操作. 在这一阶段, CNN 实现了将卷积核大小为 K 的卷积操作分解为 K^2 个 1×1 卷积, 使输入特征沿不同位置的核权重实现线性投影. 而自注意力机制则通过 1×1 卷积跨通道聚合像素级上下文信息, 输出 3 组特征图分别对应自注意力机制查询、键和值的投影, 将输入特征图投影到更深的空间维度.

第 2 阶段卷积和自注意力模块分别采用不同的聚合操作, 重用上一阶段获得的 3×N 个中间特征. 在第 1 分支中, 采用通道维度的全连接操作将通道数从 3×N 扩张到 K^2×N, 对应 CNN 上一阶段的特征维度, 然后将所有输入特征和卷积核分别连接起来, 通过位移操作进一步结合不同方向的特征, 得到移位特征:

$$Shift(F_i) = Concat(F_q, F_k, F_v) * Concat(K_c)$$

其中, $*$代表单组卷积, K_c代表核权重. 单组卷积用移位核作为初始化位移操作, 将投影后的特征图根据核位置进行位移、求和操作, 释放卷积核的可学习权值, 在提高模型容量的同时保持了原始移位操作的能力. 最

后, 使用深度可分离卷积来匹配自注意力路径的输出通道维度, 得到卷积模块的特征输出

$$F_{\text{conv}}=\prod_{h=1}^{N}W_d(Shift(F_i))$$

其中, $\prod$ 代表 N 个注意力输出的连接, $W_d(\cdot)$ 代表 3×3 深度可分离卷积. 在第 2 分支中, 通道自注意力模块将中间特征图分为 N 组, N 即为“头部”数量, 每组包括 3 个 3×3 深度可分离卷积, 组间采用无偏置的深度可分离卷积编码通道级空间上下文, 并行学习单独的注意力特征图. 通过对查询和关键投影进行重塑, 计算通道间的交叉协方差, 基于特征维度的注意力如下:

$$CA\left(F_q,F_k,F_v\right)=Softmax\left(\frac{W_d^q(F_q)\cdot W_d^k(F_k)^{\mathrm{T}}}{\sqrt{d_k}}\right)W_d^v(F_v) \tag{4}$$

该模块生成具有线性复杂度的转置注意力特征图, 然后将注意力权重特征图与原始图像残差连接, 避免网络出现过拟合的现象, 得到注意力模块的特征输出:

$$F_{\text{att}}=\prod_{h=1}^{N}W_p(CA\left(F_q,F_k,F_v\right))\oplus F_0$$

其中, $W_p(\cdot)$代表 1×1 逐点卷积, ⊕代表残差连接. 另外, 通道自注意力的头部数量随着下采样过程中通道数的扩充也保持逐层递增趋势, 从而在高维空间中捕获更多特征信息, 提升网络的图像重建质量.

在训练过程中发现, 不同网络深度的卷积和自注意力模块捕获特征的能力存在差异性. 因此, 设置学习参数 α、β, 将两条路径的输出非平均混合, 通过自主学习实现网络对重要特征的动态关注, 并抑制无用特征, 该模块的输出结果为

$$F_{\text{out}}=\alpha F_{\text{conv}}+\beta F_{\text{att}} \tag{5}$$

其中, $F_{\text{conv}}\in R^{H\times W\times C}$, $F_{\text{att}}\in R^{H\times W\times C}$, α 和 β 反映了模型对不同模块捕获特征的关注度, 使网络自适应地偏向较优方法捕获的特征信息. 在浅层阶段, 卷积实现了高效的特征提取, 此时网络更关注局部特征; 在中间阶段自注意力模块的权重逐渐增加, 模型倾向于为两种路径的输出赋予相似权重; 而在深层阶段, 自注意力的特征提取能力明显优于卷积, 主要采用自注意力机制代替原来的 3×3 卷积操作, 提升了网络的全局信息感知能力.

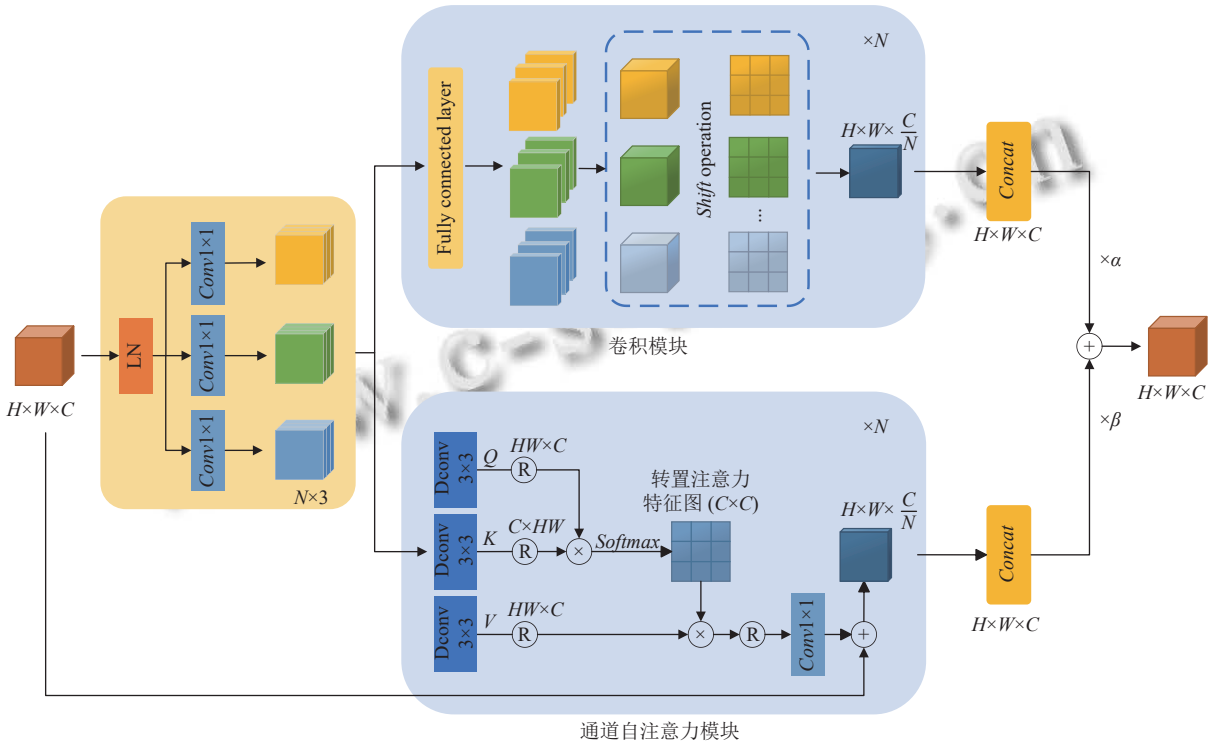


图 3 CNN 和 Transformer 耦合模块结构

从计算效率来看, CTB 模块采用并行化分支代替传统单分支结构, 在融合 CNN 时通过整合公共模块, 使网络具有更小的计算开销. 在第 1 阶段, 两种方法通过执行相同操作完成了模块的主要计算量, 并利用卷

积的共享权重特性, 来减少模型的内存占用. 第 2 阶段, 卷积模块采用轻量级的聚合操作几乎不造成额外的参数负担, 自注意力模块利用跨特征维度的深度可分离卷积避免了计算效率受图像分辨率 H、W 的影响, 计算成本和训练参数呈通道数的线性相关. 另外, 采用非平均混合策略整合输出特征图, 使网络在关注重要特征的同时降低通道维度, 进一步提升去噪模型的推理速度.

2.3 自注意力增强卷积模块

自注意力机制能够提供全局信息使网络实现远程依赖建模, 但也在一定程度上恶化高频信号, 导致基于 Transformer 的图像去噪算法不能明确的辨别噪声特征, 在复杂背景情况下, 将一些图像细节信息误判为噪声信号, 丢失对局部纹理信息的关注, 导致输出的图像过度平滑甚至模糊了重要细节.

对此, 本文设计了一种即插即用的自注意力增强卷积模块 (self-attention enhanced convolution module, SAConv), 利用轻量级的多头自注意力捕获低频信号, 增强后续卷积模块对局部特征的提取. SAConv 通过构建邻域像素的相似性程度、放大高频信号, 识别出边缘和轮廓等细微特征. 然后, 在捕获图像多频信号的基础上, 进行局部特征和全局特征融合, 提升网络的特征表达能力, 代替单一的前馈网络, 在关注局部关键特征的基础上筛选显著的特征纹理信息, 从而恢复更多图像细节和边缘信息.

如图 4 所示, SAConv 主要由高效的多头自注意力模块 (efficient multi-head self attention module, *EMSA*)、分组卷积模块 (multi-group convolutional block, *MGCB*) 和多层感知器 (multilayer perceptron, *MLP*) 串联组成.

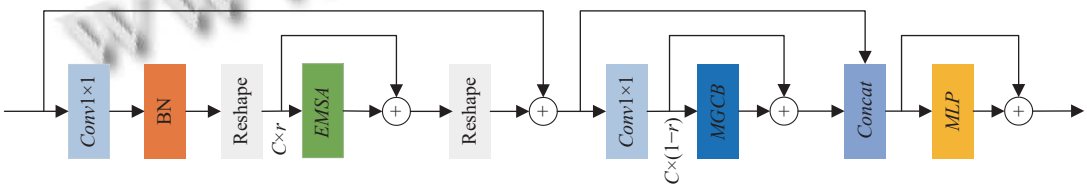


图 4 SAConv 模块结构

首先, 采用 1×1 逐点卷积降低输入特征通道数, 通过批量归一化 (BN) 聚合空间维度的冗余信息.

$$X = BatchNorm(Conv1\times1(F_{\text{out}}))$$

将经过上述预处理的图像特征 X 作为模型输入, 以提升后续 *EMSA* 模块的计算效率. *EMSA* 模块通过引入平均池化层聚合特征映射的空间信息, 压缩输入特征图的空间维数, 并逐元素加权合并, 以产生轻量级的通道注意力图. *EMSA* 可表示为:

$$SA(X) = Softmax\left(\frac{(X\cdot W^q)\otimes Avgpool(X\cdot W^k)^{\mathrm{T}}}{\sqrt{d_k}}\right)Avgpool(X\cdot W^v) \tag{6}$$

$$EMSA(z) = W_p(Concat(SA_1(z_1),SA_2(z_2),\cdots,SA_h(z_h))) \tag{7}$$

其中, $z=[z_1,z_2,\cdots,z_h]$代表多头注意力机制, $Avgpool(\cdot)$ 代表平均池化操作, $\otimes$代表点积交互运算. *EMSA* 模块利用平均池化操作实现轻量级的多头注意力机制, 捕获低频信号, 然后采用长距离连接, 将捕获的全局特征作为 *MGCB* 模块的直接输入, 使 *MGCB* 模块专注于提取高频信号, 捕获具有重要语义信息的局部特征. 通过级联操作聚合 *EMSA* 和 *MGCB* 模块的特征输出, 混合高低频信息从而实现局部和非局部的像素交互, 保留恢复图像中的精细结构和纹理细节. 最后, 利用多层感知器的全连接层模拟复杂的非线性函数, 提取更基本和明显的图像特征, 优化前向传播的特征信息, 进一步提升网络的特征表达能力. SAConv 模块步骤如下.

输入: 经过预处理的集合 X.

输出: 局部和全局特征融合的完整特征图$\widehat{X}$.

Step 1: $\dot{X} = EMSA(X)\oplus X$

Step 2: $\ddot{X} = Conv1\times1(\dot{X}\oplus F_{\text{out}})$

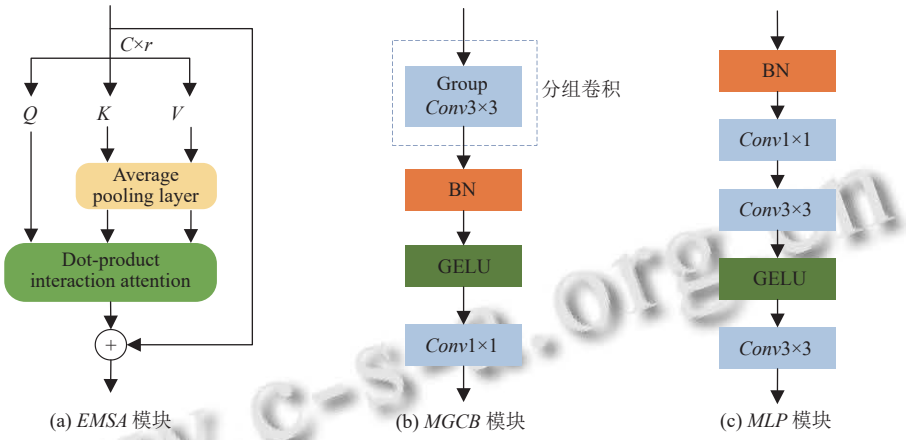
Step 3: $\bar{X} = MGCB(\ddot{X})\oplus\ddot{X}$

Step 4: $\tilde{X} = Concat(\dot{X}\oplus F_{\text{out}},\bar{X})$

Step 5: $\widehat{X} = MLP(\tilde{X})\oplus\tilde{X}$

此外, 通过引入随机收缩因子 r 降低 *EMSA* 和 *MGCB* 模块的输入通道数, 使用 BN 和 GELU 作为规范化层和非线性激活函数, 利用平均池化、深度可分离卷积等进一步提高模块的训练速率和稳定性.

(1) 轻量级的多头自注意力模块 (*EMSA*): *EMSA* 是一种轻量级的高效自注意力模块, 如图 5(a) 所示. 将输入特征图按特征维度划分为多头自注意力, 跨通道执行查询关键特征交互, 并行地学习不同维度的注意力特征图. 在点积交互计算前采用平均池化层对空间维度进行下采样, 降低多头自注意力机制带来的高额计算消耗, 并通过残差连接弥补下采样造成的特征损失.



(a) *EMSA* 模块 (b) *MGCB* 模块 (c) *MLP* 模块

图 5 SAConv 子模块结构

(2) 分组卷积模块 (*MGCB*): *MGCB* 主要包括分组卷积和逐点卷积, 如图 5(b) 所示. 首先, 将输入特征图按通道维度划分为多组卷积的形式, 从多组并行子空间中捕获高频信息. 通过关注不同子空间的不同位置的信息表示, 将空间联系密切的邻域像素高效聚合, 实现有效的局部特征学习, 然后采用逐点卷积促进多组卷积之间的信息交互. 另外, 在训练过程中将所有 *MGCB* 模块的通道维度统一设置为 32, 以实现快速推理.

SAConv 在传统前馈网络的基础上增加聚焦局部特征的功能模块, 采用 *EMSA* 和 *MGCB* 模块组合的方式增强模型对输入序列中不同特征的理解和提取能力, 降低噪声信号对图像重建的干扰, 利用 CNN 的平移不变性捕获图像局部特征, 使网络关注局部和全局特征之间的交互, 识别边缘细节信息从而更好地还原真实图像.

2.4 损失函数

本文算法采用均方误差损失 (mean square error, *MSE*) 进行训练, 损失函数如下:

$$\min_{\Theta}\frac{1}{M}\sum_{i=1}^{M}\|X_i-f_{\Theta}(Y_i)\|_F^2 \tag{8}$$

其中, M 是训练样本的数量, $f_{\Theta}(Y_i)$代表 HCT-Net 输出的恢复图像, Θ表示所有可学习参数, Y_i 为第 i 个噪声图像, X_i 为对应的干净参考图像. 训练过程中使用 ReduceLROnPlateau 动态调整学习率, 设置初始学习率为 1E–4, 最小学习率为 1E–6, 并采用 AdamW 优化器.

3 实验分析

3.1 评价指标

本文采用峰值信噪比 (peak signal to noise ratio, *PSNR*) 和结构相似性 (structural similarity, *SSIM*) 作为算法去噪性能的客观评价指标, 将去噪效果对比图作为主观评价依据. 在计算评价指标时, 高斯噪声数据集以干净图像作为参照图像, 真实噪声数据集以原始图像作为参照图像, 分别计算 *PSNR*、*SSIM* 指标.

PSNR 通过计算参照图像与去噪图像$\hat{I}$之间的峰值信噪比, 衡量二者的相似性程度. *PSNR* 公式如下:

$$PSNR=10\cdot\log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{9}$$

其中, *MAX* 代表参照图像的可能最大像素值, *MSE* 代表大小为 $H\times W$ 的参照图像与去噪图像之间的均方误差, 以均方误差判断图像的失真程度. *MSE* 公式如下:

$$MSE=\frac{1}{H\times W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}[I(i,j),k(i,j)]^2 \tag{10}$$

SSIM 仿造人类视觉系统, 从亮度、对比度以及结构量化图像的属性, 感知图像在去噪过程中发生的局部结构性改变. *SSIM* 公式如下:

$$SSIM(x,y)=\frac{\left(2\mu_x\mu_y+c_1\right)\left(2\sigma_{xy}+c_2\right)}{\left(\mu_x^2+\mu_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)} \tag{11}$$

其中, x、y 分别代表参照图像和去噪图像, μ_x和μ_y是图像的平均值, σ_x^2和σ_y^2是图像方差, σ_{xy}代表图像间的协

方差, c_1、c_2是用于维持稳定的两个常数项, 且 *SSIM* 取值范围应在 0–1 之间.

3.2 实验设置及数据集

本文采用 Windows 操作系统, 在 GeForce GTX 1080 Ti 处理器的 Python 环境下, 搭建 PyTorch 深度学习框架进行 HCT-Net 的训练和测试.

在 6 个基准数据集下验证本文算法的有效性. 灰度图像合成噪声实验采用 BSD400 为训练集, Set12 和 BSD68 为测试集. BSD400 数据集包括 400 张灰度图像, Set12 数据集包括 12 张不同尺寸的灰度图像, BSD68 数据集包括 68 张不同尺寸的灰度图像. 彩色图像合成噪声实验采用 CBSD400 为训练集, Kodak24 和 CBSD68 为测试集. CBSD400 数据集包括 400 张彩色图像, Kodak24 数据集包括 24 张 500×500 的彩色图像, CBSD68 数据集包括 68 张不同尺寸的彩色图像. 真实图像去噪实验采用 SIDD 为训练集, SIDD、DND 为测试集. SIDD、DND 数据集分别由 320 对、50 对噪声图像和相应的真实图像组成.

3.3 实验结果与分析

为验证融合 CNN 和 Transformer 的图像去噪算法去噪性能的有效性和运行效率的可行性, 本文分别选取近年来各类主流图像去噪方法进行对比, 包括基于图像先验的传统去噪算法、基于卷积神经网络的图像去噪算法和基于 Transformer 的图像去噪算法, 并通过设置不同噪声类型和噪声等级验证本文算法的普适性.

实验测试包括高斯合成噪声的图像去噪实验和真实噪声的图像去噪实验. 其中, 高斯合成噪声分为灰度图像合成噪声和彩色图像合成噪声. 在合成噪声图像实验时, 通过对原始图像添加不同等级的高斯噪声 (σ = {5, 25, 50, 75}) 得到合成的噪声图像, 并将原始图像作为参考图像.

3.3.1 灰度图像合成噪声的实验结果

表 1 列出了在高斯灰度图像测试集上不同算法的平均 *PSNR* 和 *SSIM* 指标. 实验结果表明, 在 Set12 数据集上, 除噪声等级为 75 时本文算法的平均 *PSNR* 指标略低于 FFDNet, 在其他噪声等级下均取得了最好的 *PSNR* 和 *SSIM* 指标. 在 BSD68 数据集上, 在噪声等级为 75 时, 本文算法的 *SSIM* 指标略低于 FFDNet 算法, 但 *PSNR* 指标的结果是最好的. 此外, 在其他噪声等级下均取得了最好的 *PSNR* 和 *SSIM* 指标值.

表 1 在 Set12 和 BSD68 数据集上不同方法的平均 *PSNR* 和 *SSIM* 结果

数据集	方法	σ = 5		σ = 25		σ = 50		σ = 75	
		PSNR (dB)	*SSIM*	*PSNR* (dB)	*SSIM*	*PSNR* (dB)	*SSIM*	*PSNR* (dB)	*SSIM*
Set12	BM3D	38.06	0.959	29.97	0.854	26.72	0.778	24.91	0.717
	DnCNN	38.15	0.960	30.34	0.857	27.18	0.785	25.20	0.735
	FFDNet	38.07	0.958	30.45	0.859	27.31	0.789	**25.47**	0.734
	IRCNN	38.10	0.959	30.43	0.824	27.12	0.745	25.09	0.714
	BRDNet	38.25	0.960	30.61	0.846	27.45	0.761	25.34	0.732
	ADNet	38.21	0.959	30.58	0.838	27.37	0.759	25.29	0.729
	Restormer	38.38	0.961	31.02	0.855	27.85	0.780	25.41	0.734
	Ours	**38.72**	**0.961**	**31.63**	**0.863**	**27.96**	**0.791**	25.45	**0.737**
BSD68	BM3D	37.59	0.964	28.57	0.807	25.62	0.695	24.21	0.632
	DnCNN	37.68	0.956	29.16	0.819	26.23	0.707	24.64	0.656
	FFDNet	37.75	0.966	29.21	0.829	26.27	0.724	24.74	**0.755**
	IRCNN	37.63	0.959	29.15	0.769	26.19	0.702	24.55	0.627
	BRDNet	37.74	0.965	29.28	0.816	26.34	0.719	24.69	0.733
	ADNet	37.70	0.962	29.25	0.813	26.29	0.717	24.63	0.731
	Restormer	37.94	0.965	29.51	0.829	26.62	0.721	24.83	0.746
	Ours	**38.12**	**0.968**	**30.39**	**0.832**	**27.33**	**0.733**	**24.87**	0.751

图 6 给出了在 Set12 数据集中部分灰度图像的去噪结果 (σ = 25), 通过对比可以发现, 本文算法在第 1 组图像中保留了更多房屋外部的纹理细节信息, 在第 2 组图像中对蝴蝶花纹的边缘部分恢复得更加清晰.

图 7 给出了在 BSD68 数据集中部分灰度图像的去噪结果 (σ = 50), 可以观察到在较高的噪声水平下, BM3D 和 FFDNet 的去噪结果中仍存在较为明显的噪点, IRCNN 和 Restormer 存在一定的过度平滑的现象, 导致处理后的图像模糊了边界等重要信息. 而本文方法的去噪结果显示噪声残留更少、细节表达更加清晰, 输出图像更接近于真实图像, 使主观上具有更好的去噪视觉效果.

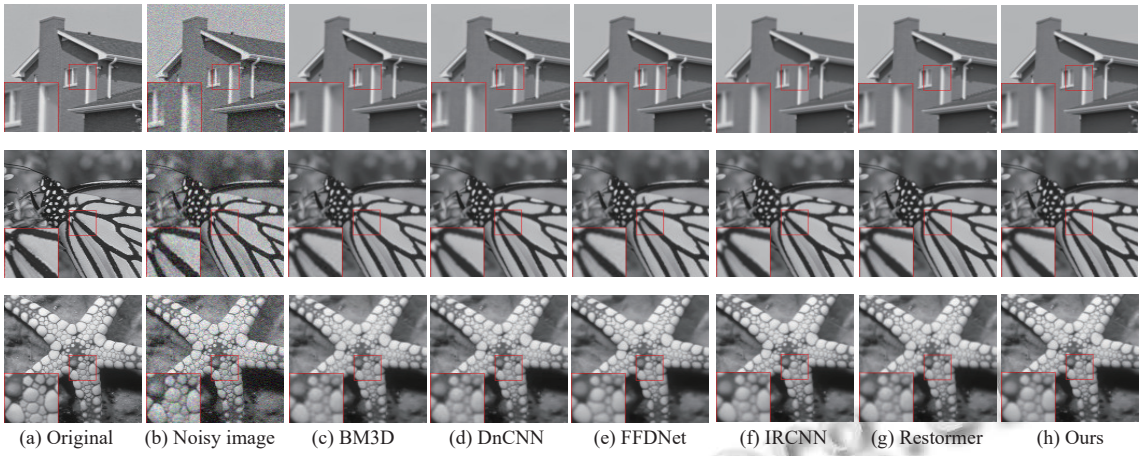


图 6 Set12 中的部分图像去噪结果 (σ=25)

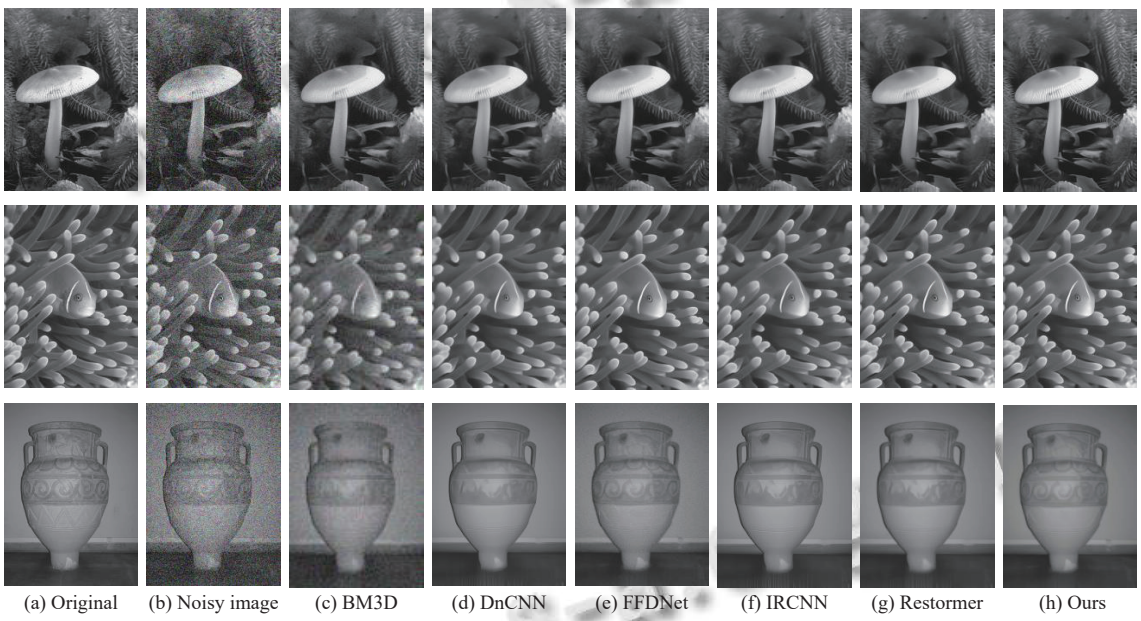


图 7 BSD68 中的部分图像去噪结果 (σ=50)

3.3.2 彩色图像合成噪声的实验结果

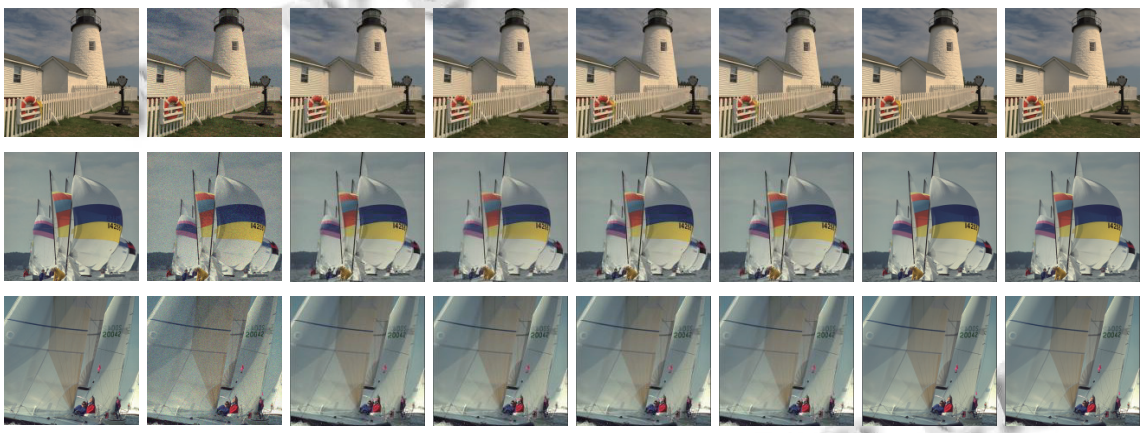
表 2 列出了各种算法在高斯彩色图像测试集上的平均 *PSNR* 和 *SSIM* 指标. 通过表 2 可以看出, 本文算法在 Kodak24 数据集的不同噪声等级下, 均取得了较好的去噪性能. 在 CBSD68 数据集上, 除了噪声水平为 50 时, 本文算法的 *PSNR* 指标略低于 CDnCNN, 其他情况下的各项性能指标均表现为最优, 验证了本文算法对于彩色高斯噪声的有效性.

图 8 给出了在 Kodak24 数据集上的部分图像去噪结果 (σ=25), 与灰度合成噪声实验结果相似, 本文算法在彩色图像去噪任务中仍获得了较好的视觉效果. 在图 8 的第 1 组图像中, CBM3D 的去噪结果中仍存在噪点, CDnCNN 和 FFDNet 在云朵区域丢失了一些细节信息. 在第 2 组和第 3 组图像中, 对比其他方法的细节部分处理, 本文算法恢复出更多的帆面纹理信息, 使图像更具真实性.

图 9 给出了 CBSD68 数据集上的部分图像去噪结果 (σ=50), 通过对比可以发现, 在第 1 组图像的背景山脉部分、第 2 组图像的大象皮肤部分以及第 3 组图像的花瓣区域, IRCNN 和 Restormer 等方法均存在局部失真和背景过度平滑现象, 而本文算法通过抑制高频噪声保留了更丰富的图像细节信息, 使得纹路层次更加清晰准确. 另外, 本文算法得到的恢复图像不存在色调偏移和饱和度失衡现象, 在图像色彩和对比度等方面更接近于干净图像.

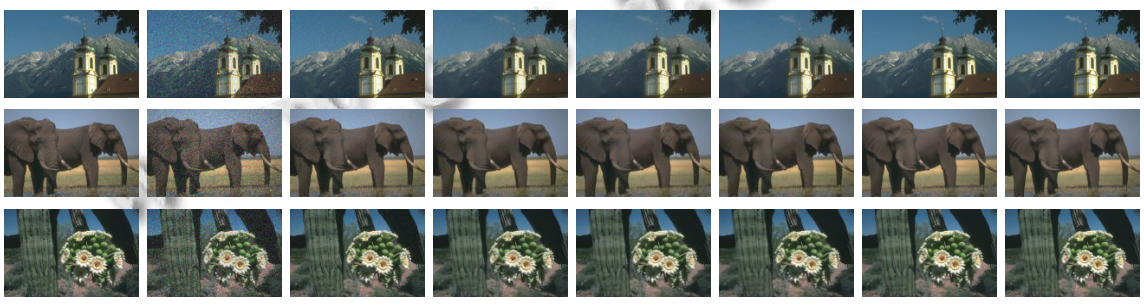
表 2 在 Kodak24 和 CBSD68 数据集上不同方法的平均 *PSNR* 和 *SSIM* 结果

Datasets	Methods	$\sigma=5$		$\sigma=25$		$\sigma=50$		$\sigma=75$	
		PSNR (dB)	*SSIM*	*PSNR* (dB)	*SSIM*	*PSNR* (dB)	*SSIM*	*PSNR* (dB)	*SSIM*
Kodak24	CBM3D	40.26	0.971	31.67	0.869	28.45	0.781	26.81	0.724
	CDnCNN	39.86	0.928	32.14	0.845	28.95	0.768	25.84	0.496
	FFDNet	40.21	0.971	32.13	0.879	28.98	0.798	27.25	0.738
	IRCNN	40.32	0.969	32.18	0.873	28.93	0.782	26.46	0.735
	BRDNet	40.27	0.967	32.41	0.881	29.22	0.796	27.49	0.739
	ADNet	40.25	0.963	32.37	0.879	29.17	0.793	27.46	0.735
	Restormer	40.43	0.971	33.02	0.882	30.00	0.801	27.39	0.741
	Ours	**40.82**	**0.972**	**33.69**	**0.885**	**30.06**	**0.826**	**27.65**	**0.762**
CBSD68	CBM3D	40.24	0.979	30.71	0.871	27.38	0.769	25.73	0.704
	CDnCNN	39.81	0.978	31.23	0.888	27.92	0.778	25.06	0.546
	FFDNet	40.16	0.979	31.21	0.883	27.86	**0.791**	26.23	0.723
	IRCNN	40.76	0.976	31.16	0.873	27.86	0.784	26.09	0.711
	BRDNet	40.64	0.981	31.43	0.885	28.16	0.781	26.34	0.728
	ADNet	40.59	0.980	31.39	0.880	28.09	0.779	26.30	0.724
	Restormer	40.45	0.982	31.78	0.890	28.59	0.786	26.29	0.727
	Ours	**40.92**	**0.985**	**31.86**	**0.891**	**28.97**	0.789	**26.36**	**0.731**



(a) Original (b) Noisy image (c) CBM3D (d) CDnCNN (e) FFDNet (f) IRCNN (g) Restormer (h) Ours

图 8 Kodak24 中的部分图像去噪结果 ($\sigma=25$)



(a) Original (b) Noisy image (c) CBM3D (d) CDnCNN (e) FFDNet (f) IRCNN (g) Restormer (h) Ours

图 9 CBSD68 中的部分图像去噪结果 ($\sigma=50$)

3.3.3 真实噪声图像的实验结果

为了评估本文算法在真实噪声情况下的去噪性能, 本文测试了 SIDD 和 DND 两个真实噪声数据集, 所有算法仅使用 SIDD 数据集上 160 张噪声图像作为训练集, 然后将剩余图像和 DND 数据集作为测试集.

表 3 给出了在 SIDD 和 DND 数据集上不同方法的平均 *PSNR* 和 *SSIM* 指标, 结果表明, 在 SIDD 数据集上本文算法的平均 *PSNR* 和 *SSIM* 指标均处于最高, 相比之前去噪性能最优的卷积神经网络去噪算法 APD-Nets-C224 和基于 Transformer 的图像去噪算法

Restormer, *PSNR* 指标分别提升了 2.1 dB 和 1.8 dB. 另外, 在 DND 数据集上的结果显示, 本文算法依旧取得了最优的去噪性能.

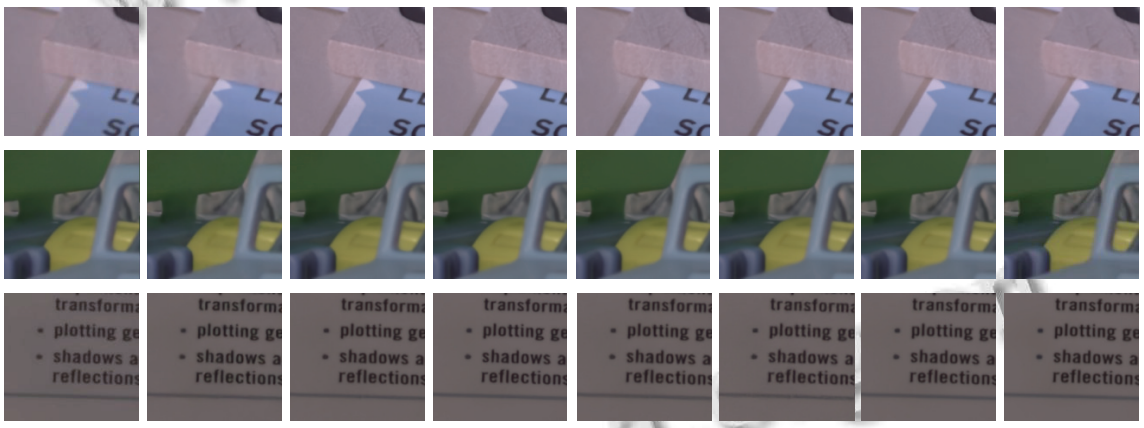
图 10 和图 11 分别给出了在 SIDD 和 DND 数据集上部分真实图像的去噪效果. 通过对比图 10 中第 1 组和第 3 组图像的字母边界部分、图 11 的 3 组建筑图像的图案花纹部分, 观察到本文算法相比于 Uformer 和 Restormer 等算法, 更关注图像的局部特征, 善于捕捉图像细节和边缘等关键信息, 从而在不影响精细纹理的情况下生成更为清晰的图像, 避免出现高频噪声恶化细节纹理信息和模糊边界的情况.

3.4 算法效率分析

图像去噪作为图像预处理中的重要步骤, 在现实场景中具有广泛的应用价值, 而网络的复杂程度和运行速率作为能否实施部署的关键影响因素, 决定了图像去噪网络的真实性能.

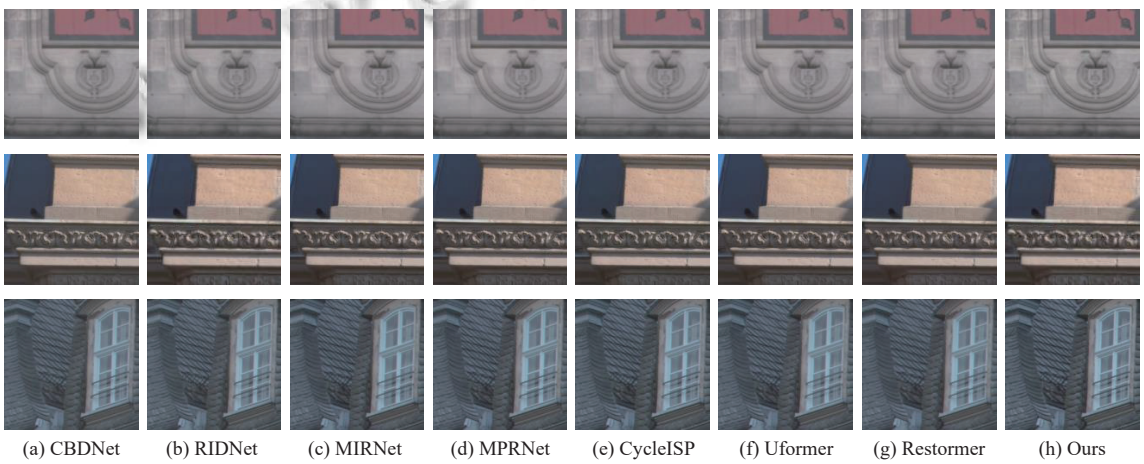
表 3 在 SIDD 和 DND 数据集上不同方法的平均 *PSNR* 和 *SSIM* 结果

方法	SIDD数据集		DND数据集	
	PSNR (dB)	*SSIM*	*PSNR* (dB)	*SSIM*
BM3D	25.65	0.685	34.51	0.851
WNNM	25.78	0.809	34.67	0.865
K-SVD	26.88	0.842	36.49	0.899
DnCNN	23.66	0.583	32.43	0.790
CBDNet	30.78	0.801	38.06	0.942
RIDNet	38.71	0.951	39.26	0.953
VDN	39.26	0.955	39.38	0.952
MIRNet	39.72	0.959	39.88	0.956
MPRNet	39.71	0.958	39.80	0.954
CycleISP	39.52	0.957	39.56	0.956
Uformer	39.77	0.959	39.96	0.956
Densformer	39.68	0.958	39.87	0.955
PromptIR	39.64	0.961	39.71	0.956
DANet	39.43	0.956	39.58	0.955
COLA-E	39.04	0.951	39.64	0.954
APD-Nets-C224	39.75	0.961	39.92	0.954
Restormer	40.02	0.960	40.03	0.956
Ours	**41.83**	**0.962**	**41.91**	**0.956**



(a) CBDNet (b) RIDNet (c) MIRNet (d) MPRNet (e) CycleISP (f) Uformer (g) Restormer (h) Ours

图 10 SIDD 中部分真实图像去噪结果



(a) CBDNet (b) RIDNet (c) MIRNet (d) MPRNet (e) CycleISP (f) Uformer (g) Restormer (h) Ours

图 11 DND 中部分真实图像去噪结果

表 4 给出了在 SIDD 真实数据集上部分网络框架的运行效率对比, 分别比较不同去噪方法的参数量 (param) 和每秒 10 亿次的浮点运算数 (GFLOPs). 实验结果表明, 本文提出的去噪算法所需的参数量相较于目前基于 Transformer 的图像去噪算法最少, 运行速率仅次于 CBDNet, 而 CycleISP、MIRNet 和 APD-Nets-C224 等算法依赖于强大的硬件能力, 在处理高分辨率图像时将造成高额的计算开销. 综合分析, 本文算法在复杂的现实场景中依然能保证良好的计算效率, 实现了图像去噪的最优性能.

表 4 不同算法的计算效率对比

Methods	Param (MB)	GFLOPs	*PSNR* (dB)
CBDNet	4.37	40.28	30.78
RIDNet	1.50	97.95	38.71
CycleISP	2.84	335.01	39.52
MIRNet	31.79	816.75	39.72
Uformer	20.63	80.32	39.77
Restormer	27.14	92.09	40.02
APD-Nets-C224	18.61	574.98	39.75
Ours	18.02	75.25	41.83

3.5 消融实验

HCT-Net 是由 CTB 模块和 SAConv 模块串联组成的网络结构, 为验证各个模块的有效性, 本文在 SIDD 数据集上, 保持编解码网络深度不变, 进行消融实验, 模块间的组合方式如图 12 所示.

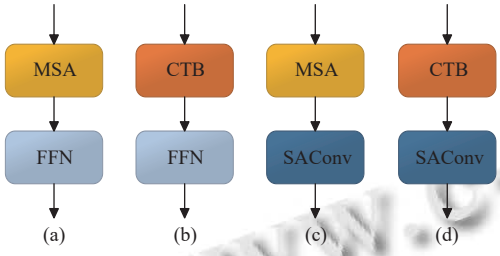


图 12 模块的组合情况

实验结果如表 5 所示, 表中黑色加粗字体为最优值, 验证了 HCT-Net 中各个模块对图像去噪算法的有效性. 对比不同模块对网络去噪性能的影响可以看出, CTB 模块能够抑制噪声信号, 增强网络对整体结构和高级语义信息的理解, 提升重建图像的质量和相似度, SAConv 模块则更擅于关注局部特征, 减少边界伪影和细节误判, 恢复更多纹理细节信息. 本文网络在 CTB 模块和 SAConv 模块串联组合时图像去噪效果最优.

此外, CTB 模块构造了融合卷积和通道自注意力的双分支结构, 并采用非平均混合策略进行特征融合. 为验证并行分支和混合策略对图像去噪性能的影响不同, 对 CTB 模块进行不同分支间的消融实验.

结果如表 6 所示, 对比第 1 组、第 2 组、第 3 组, 在单分支情况下, 两种特征提取方式的去噪指标均明显下降, 而双分支结构能够抑制噪声信号、捕捉关键特征, 从而恢复更多图像特征. 对比第 3 组、第 4 组, 并行融合机制通过非平均混合策略不断优化深层迭代, 提升了重建图像质量, 验证了动态注意力分配机制的有效性.

表 5 各模块之间的消融实验

Group	CTB	SAConv	*PSNR* (dB)	*SSIM*
a	—	—	38.87	0.943
b	√	—	40.94	0.945
c	—	√	40.31	0.960
d	√	√	**41.83**	**0.962**

表 6 CTB 模块不同分支的消融实验

Group	卷积模块	通道注意力模块	非平均混合策略	*PSNR* (dB)	*SSIM*
1	√	—	—	40.02	0.949
2	—	√	—	40.93	0.945
3	√	√	—	41.25	0.958
4	√	√	√	**41.83**	**0.962**

4 结论与展望

本文提出了融合 CNN 和 Transformer 的图像去噪网络 (HCT-Net), 采用 CNN 和 Transformer 耦合模块 (CTB) 抑制噪声信号、捕捉关键特征, 并缓解全局注意力机制导致计算效率低下的问题, 达到准确性和计算效率的平衡. 通过自注意力增强卷积模块 (SAConv) 提升网络对局部特征的关注度, 捕获多频信号从而实现局部和全局信息融合, 使网络既能关注局部特征又能捕获远程依赖关系. HCT-Net 与目前先进的深度学习去噪算法相比, 具有较好的去噪效果和一定的性能优势. 在未来的工作中, 将进一步减少网络参数, 优化算法的网络结构.

参考文献

1 Pronina V, Kokkinos F, Dylov DV, *et al*. Microscopy image restoration with deep Wiener-Kolmogorov filters. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 185–201.

2 Gavaskar RG, Chaudhury KN. Fast adaptive bilateral filtering. IEEE Transactions on Image Processing, 2019, 28(2): 779–790. [doi: 10.1109/TIP.2018.2871597]

3 Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 60–65.
4 Dabov K, Foi A, Katkovnik V, *et al*. Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on Image Processing, 2007, 16(8): 2080–2095. [doi: 10.1109/TIP.2007.901238]
5 Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, 1994, 2(2): 291–298. [doi: 10.1109/89.279278]
6 Chan TF, Shen JH, Zhou HM. Total variation wavelet inpainting. Journal of Mathematical Imaging and Vision, 2006, 25(1): 107–125. [doi: 10.1007/s10851-006-5257-3]
7 Elad M, Aharon M. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing, 2006, 15(12): 3736–3745. [doi: 10.1109/TIP.2006.881969]
8 Hou YK, Xu J, Liu MX, *et al*. NLH: A blind pixel-level non-local method for real-world image denoising. IEEE Transactions on Image Processing, 2020, 29: 5121–5135. [doi: 10.1109/TIP.2020.2980116]
9 Zhang K, Zuo WM, Chen YJ, *et al*. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 2017, 26(7): 3142–3155. [doi: 10.1109/TIP.2017.2662206]
10 Zhang K, Zuo WM, Zhang L. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. IEEE Transactions on Image Processing, 2018, 27(9): 4608–4622. [doi: 10.1109/TIP.2018.2839891]
11 Guo S, Yan ZF, Zhang K, *et al*. Toward convolutional blind denoising of real photographs. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1712–1722.
12 Anwar S, Barnes N. Real image denoising with feature attention. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3155–3164.
13 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
14 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
15 Wang ZD, Cun XD, Bao JM, *et al*. Uformer: A general U-shaped Transformer for image restoration. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17662–17672.
16 Liang JY, Cao JZ, Sun GL, *et al*. SwinIR: Image restoration using Swin Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 1833–1844.
17 Zamir SW, Arora A, Khan S, *et al*. Restormer: Efficient Transformer for high-resolution image restoration. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5718–5729.
18 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.
19 Pan XR, Ge CJ, Lu R, *et al*. On the integration of self-attention and convolution. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 805–815.
20 Park N, Kim S. How do vision Transformers work? Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.
21 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
22 Touvron H, Cord M, Douze M, *et al*. Training data-efficient image Transformers & distillation through attention. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 10347–10357.
23 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.
24 Graham B, El-Nouby A, Touvron H, *et al*. LeViT: A vision Transformer in ConvNet's clothing for faster inference. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 12239–12249.

(校对责编: 张重毅)