基于超像素引导的 Transformer 低光图像去噪方法

宋泉臻，陈作钧，秦品乐，曾建潮
（中北大学计算机科学与技术学院，山西 太原 030051）

摘　要： 现有的低光图像去噪方法主要使用 Transformer 和卷积神经网络（CNN）的特征提取和去噪机制，会面临两个问题：基于局部窗口的自注意力机制未能充分捕捉图像中的非局部自相似性；通道维度上的自注意力计算未充分利用图像的空间关联性。针对上述问题，在基于窗口划分的视觉 Transformer 方法上提出一种超像素引导的策略，其可以自适应地选择相关窗口进行全局交互。首先，设计基于窗口交互的 Top-N 交叉注意力机制（TNCA），动态选择与目标图像窗口最相似的前 N 个窗口，并在通道维度上聚合图像窗口的信息，充分考虑图像非局部自相似性；其次，通过超像素分割引导的方式，显著提升窗口内局部特征的表达力，同时在通道维度上增强空间特征的关联性；最后，构建一个层次化的自适应交互超像素引导的 Transformer 去噪网络（AISGFormer）。实验结果表明，AISGFormer 在 SIDD 和 DND 真实图像数据集上的峰值信噪比（PSNR）分别为 39.98 dB 和 40.06 dB，与其他先进网络相比分别提升了 0.02 dB～14.33 dB 和 0.02 dB～7.63 dB，AISGFormer 更能交互局部与全局的信息和细节，自适应地利用自相似性来抑制区域相似噪声。

关键词： 低光图像去噪；Transformer；交叉注意力；非局部自相似性；真实图像噪声；超像素
中图分类号： TP391.41　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0070426

Superpixel Guide for Transformer Low-Light Image Denoising Method

SONG Quanzhen, CHEN Zuojun, QIN Pinle, ZENG Jianchao
(School of Computer Science and Technology, North University of China, Taiyuan 030051, Shanxi, China)

【Abstract】 Existing low-light image denoising methods mainly use the feature extraction and denoising mechanisms of Transformer and Convolutional Neural Networks (CNN). They face two problems: the self attention mechanism based on local windows fails to fully capture the nonlocal self-similarity in images, and the calculation of self-attention in the channel dimension does not fully utilize the spatial correlation of images. To address these issues, this study proposes a superpixel guided strategy for a window partition-based visual Transformer method; the strategy can adaptively select relevant windows for global interactions. First, a Top-N Cross Attention mechanism (TNCA) is designed based on window interactions, the top N windows that are most similar to the target image window are selected dynamically, and the information related to the image windows in the channel dimension are aggregated, fully considering the nonlocal self-similarity of the image. Second, through superpixel segmentation guidance, the expressive power of local features within the window is significantly improved while enhancing the correlation of spatial features in the channel dimension. Finally, a hierarchical Adaptive Interaction Superpixel Guide Transformer (AISGFormer) is constructed. Experimental results show that AISGFormer achieves a Peak Signal-to-Noise Ratio (PSNR) of 39.98 dB and 40.06 dB on the SIDD and DND real image datasets, respectively. Compared with other advanced networks, the PSNR improves by 0.02 dB–14.33 dB and 0.02 dB–7.63 dB, respectively. AISGFormer interacts with local and global information and details more effectively, and it adaptively utilizes self-similarity to suppress region similarity noise.

【Key words】 low-light image denoising; Transformer; cross-attention; non-local self-similarity; real image noise; superpixel

0　引言

低光图像噪声是指在光照不足的情况下相机捕获的图像中出现的随机变化或异常像素点，这种噪声具有随机性和非均匀性，会降低图像的对比度并掩盖细节，从而影响目标检测和追踪等下游任务的准确性[1-2]。同时，在低光环境下的夜视成像等技术逐渐成为安全监控、生物医学成像、夜间军事侦察等领域不可或缺的重要组成部分。因此，如何在光照条件受限的情况下有效抑制噪声并恢复图像细节，

基金项目： 山西省科技重大专项计划“揭榜挂帅”项目（202101010101018）；山西省长治市“揭榜挂帅”项目。
作者简介： 宋泉臻（CCF 学生会员），男，硕士，主研方向为底层计算机视觉、低光图像去噪；陈作钧，博士；秦品乐（通信作者），曾建潮，教授、博士。
收稿日期： 2024-09-30　　**修回日期：** 2024-11-27　**E-mail：** qpl@nuc.edu.cn

成为夜视成像过程和计算机视觉领域亟需解决的问题[3]。

针对低光图像的噪声问题，许多解决方法逐渐被提出。传统的图像去噪方法主要是基于结构自相似性和使用图像先验的方式，但当面临复杂多变且高度动态的噪声场景，尤其是叠加于低光条件下的环境时，现有方法普遍显现出细节丢失和适应性差的局限性[4]。近年来，深度学习在计算机视觉领域取得了重大的突破，卷积神经网络(CNN)因其卓越的局部特征提取能力，在图像去噪领域一度占据核心地位[5]。然而，CNN 主要依赖于卷积操作来提取图像的局部特征，对于图像中的长距离依赖关系，其捕捉能力有限。在低光条件下，图像中的噪声往往呈现出复杂的空间分布和纹理特征，需要全局的上下文信息来进行准确去噪，CNN 在处理这类图像时往往难以达到理想的去噪效果[6]。

Transformer 模型的出现弥补了 CNN 在全局特征捕捉上的不足。Transformer 模型通过自注意力机制，能够捕捉图像的全局依赖关系，这对于处理低光条件下的图像去噪问题具有重要意义。然而，目前标准的 Transformer 模型仍然存在许多问题：

1)首先，Transformer 模型依赖于空间维度上的全局自注意力机制，其计算复杂度随 token 数量的增加而呈二次方增长，导致处理高分辨率图像时计算成本显著提高。尽管基于局部窗口的自注意力策略[7-8]尝试通过将注意力计算限制在局部窗口内来降低计算复杂度，但未能充分利用图像的非局部自相似性和图像先验，限制了其对低光大噪声图像的去噪性能。同时，虽然有一些方法通过影响注意力得分来推理自相似性，但仍未考虑图像本身的非局部自相似性[9]。

2)其次，传统 Transformer 方法在通道维度上的自注意力计算时未充分考虑图像的空间关联性，忽视了图像内部空间结构信息对去噪性能的潜在贡献[10-12]。

本文提出一种基于超像素引导和自适应交互的 Transformer 模型(AISGFormer)，旨在通过精细化的全局交互策略，解决现有 Transformer 模型在低光图像去噪中图像自相似性利用不足的问题。具体来说，本文提出一种基于窗口交互的 Top-N 交叉注意力机制(TNCA)，通过动态选择与目标区域最相关的图像窗口，在保持计算效率的基础上，对图像进行高效且精确的特征提取，并有效抑制噪声。最终构建的 AISGFormer 通过分层的编码器-解码器网络架构，在不同尺度上有效计算窗口内部和窗口间的自注意力，实现对低光图像噪声的精确去除及细节恢复。本文的主要贡献如下：

1)提出了一种新的低光图像去噪模型，将超像素聚类和 Transformer 结构相结合，构建一个具有更强表达性和高效的框架 AISGFormer。

2)通过基于图像先验的超像素技术对图像中的局部相似像素进行聚类，并引导后续注意力计算在基于目标窗口的前 N 个窗口上完成，显著提升了窗口间信息交互的表达力，在通道维度上增强了其空间关联性。

3)设计了 TNCA，通过动态选择与目标窗口最相似的前 N 个窗口，实现在通道维度上的交叉注意力聚合，从而在降低计算复杂度的同时，有效利用图像的非局部自相似性。

1 相关工作

1.1 基于图像先验和深度学习的低光图像去噪算法

早期基于图像先验的传统去噪方法主要涉及非局部自相似性。文献[13]提出的 NLM 算法率先使用了非局部自相似性，通过在相似区域加权平均来去除噪声。文献[14]通过融合 NLM 的相似块搜索和三维小波变换的频域处理，有效保留了图像的纹理细节。但这些方法往往过于依赖噪声强度的估计，对于未知噪声尤其是低光下噪声的泛化性能较差。

近年来，学者们对基于深度学习的低光图像去噪方法进行了广泛研究。在 CNN 中，DnCNN[15]利用深度学习有效学习噪声分布，首次证明了深度学习在图像去噪领域的有效性。文献[16]引入噪声估计子网络，通过非对称损失增强对真实噪声的适应性，并支持高效的交互式去噪。文献[17]借助编解码结构提取图像多尺度特征。

相比于 CNN 的方法，基于 Transformer 的低光图像去噪方法具有更优的性能。ViT[18]作为首个将 Transformer 应用于计算机视觉的模型，展示了自注意力机制在图像全局特征提取上的潜力。随后，为了适应特定的图像去噪任务，IPT[19]和 Uformer[20]等变体模型通过引入局部窗口和 U 型网络结构，提升了 Transformer 在低光图像去噪任务中的性能。LRT[11]通过多视角信息融合和多尺度特征提取，学习低光图像噪声分布，并估计照度进行亮度调整，实现图像的高质量恢复。Xformer[21]综合使用局部窗口注意力和通道注意力，同时关注了图像的局部和全局信息。尽管这些模型在提取图像全局依赖性方面表现出色，但它们在处理低光噪

声图像时仍有特定噪声模式适应性不足的问题。本文将图像先验和 Transformer 相结合，帮助模型在优化全局特征提取的同时保留低光图像中阴影等处的细节纹理。

1.2 注意力机制

注意力机制使模型能够识别并优先处理最重要的信息[19]。为了解决空间自注意力的计算复杂性和权重问题，通道注意力根据通道之间的相关程度，动态调整不同通道的特征权重，模型可以根据图像内容自适应分配不同通道的重要性，并将注意力集中在最相关的通道上[21]。同时，为了关注不同输入窗口间的交叉关系，交叉注意力通过计算查询向量与键值向量的相关性，来生成将注意力权重分配到值序列上的权重。这可以使模型在处理一个模态的输入时，同时考虑到其他模态的信息，从而更好地捕捉输入窗口中不同部分之间的关联，提高多模态模型的表现[22]。

对于真实世界中低光噪声图像的处理，通常需要多尺度和不同语义细节的关联方法。本文巧妙融合 2 种注意力机制，实现了对全局和局部特征的充分提取，使得精细去噪任务在特征表达上更为精确和全面。

1.3 超像素分割

超像素分割是 REN 和 MALIK 在 2003 年首次提出的。基于聚类的算法使用传统的聚类技术，如 k 均值聚类，计算锚像素与其邻近像素之间的连接性，常见的算法有 SLIC[23]、Mani-fold-SLIC[24] 和 SNIC[25]。作为弱标签或图像先验的一种形式，预先计算的超像素分割有助于众多下游任务的实现。通过将超像素集成到深度学习中作为指导，可以更好地保留一些重要的图像属性，例如像素相似信息[26]。

文献[27]开始探索将超像素与注意力机制相结合的方法，以提高图像去噪的效果。利用超像素的结构信息来引导注意力模型的聚焦区域，从而更有效地处理低光条件下的关键图像信息。通过分析超像素结构，注意力机制能更准确地定位到与目标像素相关的区域，从而优先处理这些区域以提高去噪性能。这种策略不仅显著降低了计算的复杂性，而且巧妙地保持了对图像全局信息的全面利用[28]。本文将基于图像先验的超像素与注意力机制相结合，以一种局部与全局窗口交互的策略，使得模型能够精确捕捉并利用图像的相似结构性信息，实现更为出色的噪声抑制和细节保留效果。

2 基于超像素引导和自适应交互的低光图像去噪模型

2.1 AISGFormer 网络结构设计

AISGFormer 建立在基于分层的 U 型网络架构上，充分利用多尺度特征提取的方式，使得全局和局部特征的交互更加有效。此外，受图像先验和窗口划分方法的启发，这种多尺度的网络结构不仅能够增强模型对于图像自相似性的特征利用能力，还提高了通道维度上注意力计算的空间关联性，使得模型在处理去噪任务时更加精准和高效。

本文提出的 AISGFormer 网络结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。整体网络结构由 3×3 卷积层、多尺度输入编码器、超像素分割模块、多尺度输出解码器和特征增强模块组成。其中编码器采用 4 个 Top-N 交叉 Transformer 模块(TNCTB)，桥接模块(Bottleneck Block)采用 1 个 TNCTB，解码器采用 4 个 TNCTB 和 4 个 1×1 卷积层，特征增强模块由 3×3 普通卷积层、3×3 可分离卷积层和线性层组成。

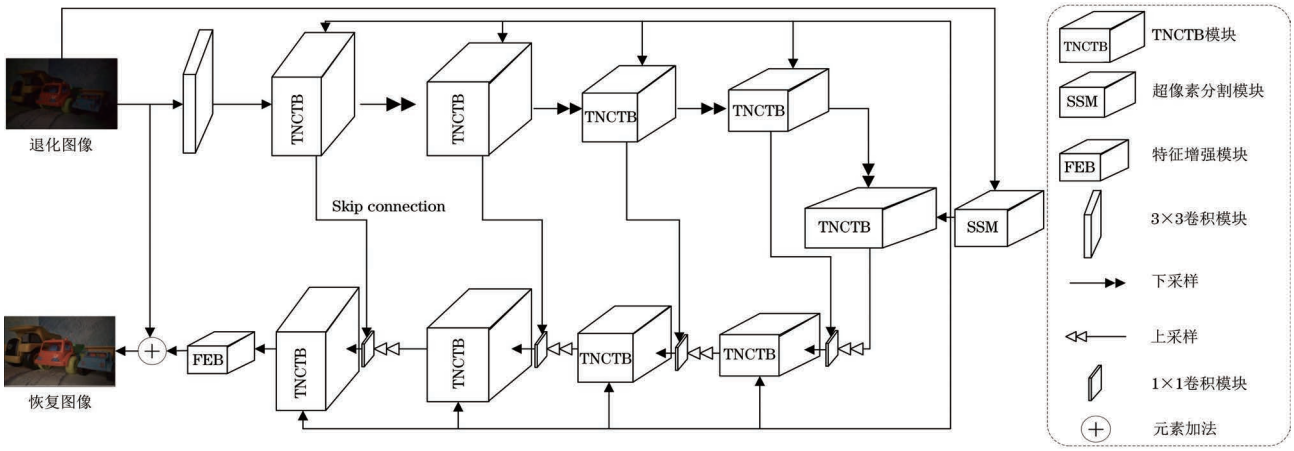

图 1 AISGFormer 网络架构

Fig. 1 Architecture of AISGFormer

首先，输入退化图像 $I_{noise}\in R^{H\times W\times 3}$ 通过超像素分割模块，得到具有超像素聚合块的特征信息，作为图像先验知识输入编解码器层，为 TNCTB 实现全局交互作准备。同时，整体网络的第一层采用标准的 3×3 卷积，提取图像的浅层信息，得到浅层特征映射 $x_0\in R^{H\times W\times C}$，其中，$H$ 为图像高度，W 为图像宽度，C 为通道数。浅层特征 x_0 的提取过程如式(1)所示：

$$x_0=f_{embed}(I_{noise}) \tag{1}$$

式中：f_{embed} 为 3×3 卷积层；I_{noise} 为退化图像。

浅层特征映射 x_0 传入编码器层，其中包含 4 个 TNCTB 模块和 4 个下采样。输入的特征图 x_0 通过 4 个下采样单元进行压缩，每个下采样单元后的 Shuffle 操作实现特征通道数增加一倍和空间尺度减半，以高效提取多尺度特征。其中，编码器部分逐步降低特征图的空间分辨率，输出一系列不同层级的特征图。之后在网络中加入一个桥接层(Bottleneck stage)，用于捕捉在较大空间范围内的特征依赖性。

解码器层同样包含 4 个 TNCTB 模块和 4 个上采样，特征图通过 4 个解码器逐步重建特征信息。在每个上采样的过程中，特征图通过 Unshuffle 操作使其空间尺度放大为原先的 2 倍，通道数变为原来的一半。上采样后，1×1 卷积用于在通道维度上合并特征图，包括通过跳跃连接引入的特征图和当前层级的特征图。在高质量特征重建阶段，特征增强模块用于进一步细致地优化特征图的细节，得到残差图像。最终，将噪声图像添加到残差图像中以获得干净图像。

2.2 超像素分割模块

Transformer 模块的局部建模能力较弱，其中的自注意力主要关注输入序列内部的关系，忽略了像素之间的非局部自相似性，导致局部细节失真。针对此问题，本文设计了能够引导自注意力机制的超像素分割模块。超像素分割结果使其能预先识别与每个窗口最相关的前 N 个窗口，这为 TNCA 的自注意力计算提供了重要的前提条件。超像素分割部分对图像的处理过程如图 2 所示。

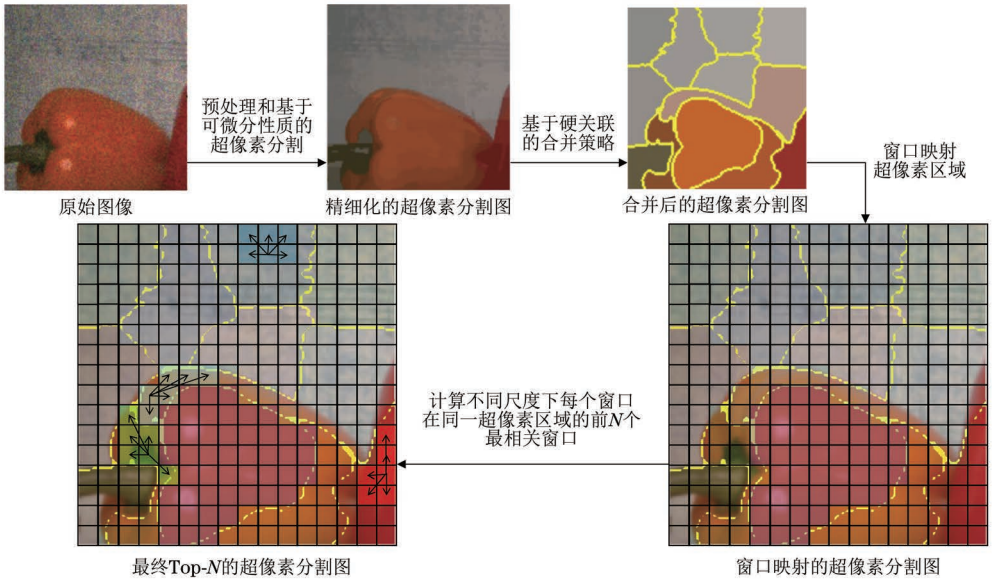

图 2 超像素分割模块

Fig. 2 Superpixel segmentation module

首先，使用模糊去噪技术对原始噪声图像进行预处理，之后使用可微分的 SLIC 算法对去噪后的图像进行超像素分割，得到精细化的超像素分割特征图。该算法的核心是引入了可微分的相似度度量 S 和迭代的关联参数 Q，使得算法能够在迭代过程中根据具体的图像内容和噪声条件调整参数。软性关联 Q 和相似度 S 的计算公式分别如式(2)和式(3)所示：

$$Q(p,i)=e^{-\frac{\|f_p-f_i\|^2}{2\sigma^2}} \tag{2}$$

$$S(p,i)=\frac{1}{Z_i}\sum_{p=1}^{n}Q(p,i)\boldsymbol{f}_p \tag{3}$$

式中：$\boldsymbol{f}_p$ 和 $\boldsymbol{f}_i$ 分别是像素点 p 和超像素中心 i 的特征向量；σ 是超参数，用于调节相似度的敏感度；$Z_i=\sum_p Q(p,i)$ 表示归一化项。这种方法使得分割过程可以在训练中自适应地优化，提高了分割的准确性和鲁棒性。之后完成基于硬关联的合并，旨在构建更大、更具代表性的超像素区域，得到合并后的超像素分割图。合并算法的具体执行过程如式(4)所示：

$$\text{if } \|c_i - c_j\| < \theta \text{ and } \text{dist}(s_i, s_j) < \varphi$$
$$\text{then averaging} \tag{4}$$

式中：c_i 和 c_j 分别是超像素区域 i 和 j 的颜色中心；$\text{dist}(s_i, s_j)$ 是空间欧氏距离；θ 和 φ 分别是颜色和空间合并阈值。

合并后的超像素区域将利用可微分 SLIC 算法的标签关联信息，进行精确的窗口映射，使每个窗口都能精准地反映出其所属的超像素区域。如图 2 所示，在最终 Top-N 的超像素分割图中，目标窗口指向其他窗口，其关键在于通过空间索引策略精准识别与每个窗口最为匹配的前 N 个相关窗口，这为后续在空间特征上增强关联性提供了支撑。

2.3 基于超像素引导的深度特征提取模块

低光照条件下噪声图像的非局部自相似性表现为图像中不同区域间存在颜色和阴影等重复相似的纹理和结构特征[13]，这种特性为图像去噪提供了重要的图像先验信息，因为去噪算法可以利用这些特征来识别和消除噪声，同时保留图像的细节[14]。此外，低光照图像中的噪声分布往往具有高度的不均匀性，这使得去噪过程变得更加复杂和具有挑战性。然而，传统的 Transformer 结构在处理图像时，往往忽略了这种非局部自相似性和图像先验，导致去噪效果未能达到预期。受上述工作启发，本文提出 TNCA 模块，其结构如图 3 所示。引入超像素分割作为图像先验知识，采用一种创新的方法，既克服全局自注意力的高计算成本，又优化基于通道注意力的方法在捕获空间局部细节方面的不足。同时，这种超像素引导的方式充分利用图像的非局部自相似性，能够有效提取低光图像中颜色和阴影等具有相似性纹理的特征。

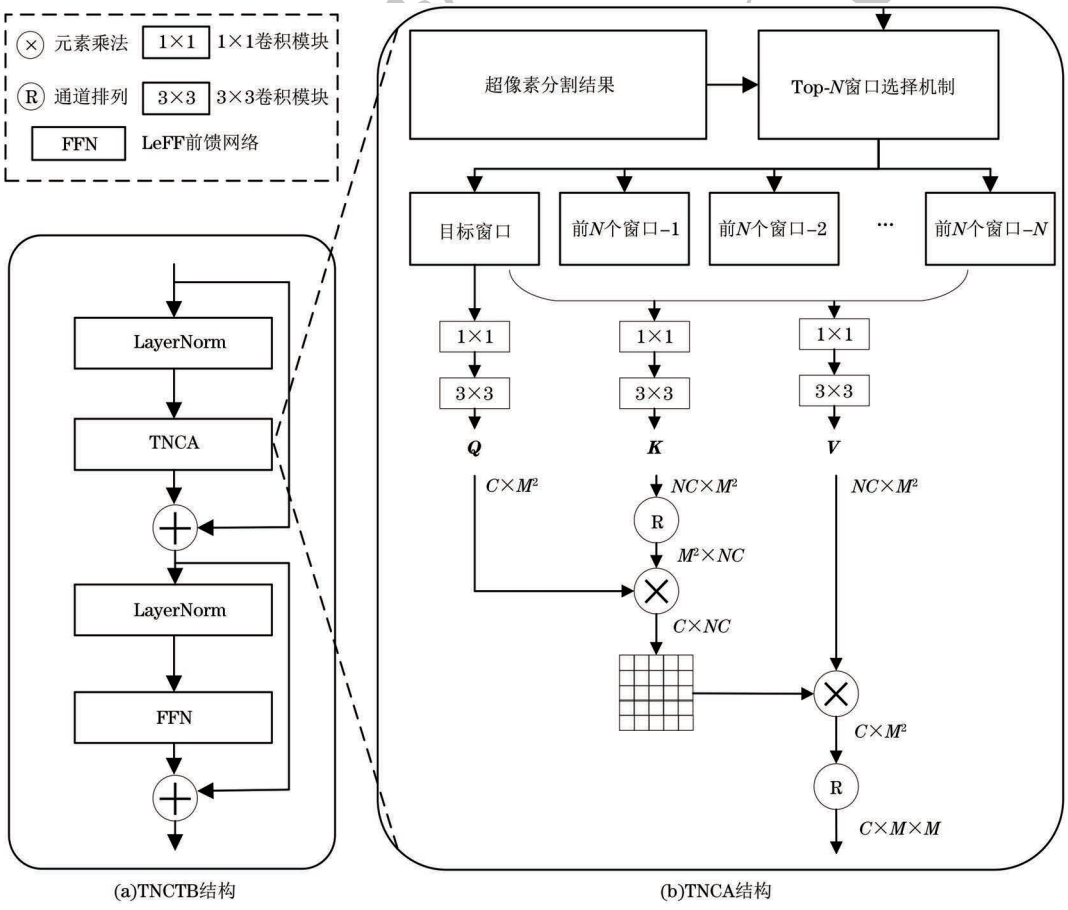

图 3 2 种模块结构

Fig. 3 Two types of module structures

TNCTB 由一个汇聚前 N 个窗口信息到目标窗口的交叉注意力 TNCA、一个前馈网络和两个层归一化串联而成。TNCA 是深度特征提取模块的核心组成部分，它利用超像素分割的结果来交互性地引导注意力机制，在目标窗口及其最相关的 N 个窗口之间进行基于通道维度的注意力计算。与使用窗口划分的计算方法[20-21]相比，通过这种全局交互性引导的计算方式，不仅能够有效捕捉低光图像中阴影和噪声等非局部相似性特征，还在达到稀疏性的同时显著降低了计算复杂度[14]。

如图 3(b)所示，对于输入的退化图像，首先通过超像素分割模块获取目标窗口及其最相关的

N 个窗口的信息，然后在每个目标窗口上应用交叉注意力计算，其中查询向量($\boldsymbol{Q}$)由目标窗口生成，而键($\boldsymbol{K}$)和值($\boldsymbol{V}$)则由前 N 个与目标窗口最相关的窗口在通道维度聚合而成的特征图生成。这种方法在维持计算效率的同时，有效利用了局部信息和全局上下文信息，精确恢复图像细节。

在计算过程中，目标窗口的特征图为 $T\in R^{M\times M\times C}$，而基于目标窗口的前 N 个窗口的特征图为 $T_n\in R^{M\times M\times NC}$。首先，将其输入到 1×1 卷积层和 3×3 深度卷积层，进行线性投影转换的同时增强模型的局部特征提取能力，并生成查询投影 $\boldsymbol{Q}_t=W_d^qW_p^qT^{C\times M^2}$、键投影 $\boldsymbol{K}_t=W_d^kW_p^kT_n^{NC\times M^2}$ 和值投影 $\boldsymbol{V}_t=W_d^vW_p^vT_n^{NC\times M^2}$，其中，$W_p$ 为 1×1 卷积层，W_d 为 3×3 深度卷积层。通过这种方式，将 Top-N 窗口的上下文信息有效地传递到目标窗口中，实现特征增强和丰富。以编码阶段的第一个 TNCA 为例，输入是一个经超像素分割模块处理后的五维张量，形状为[1 024，7，8，8，32]。其中：1 024 代表原始 256×256 尺寸的输入特征图被分割成的窗口数量，即 $M=8$ 的窗口数量；7 表示与目标窗口最相关的 7 个窗口，并包括目标窗口本身，即 $N=7$；[8，8，32]分别对应每个窗口的空间尺寸和通道数，即 M 和 C。

接着，对生成的查询投影、键投影和值投影进行重塑，通过查询投影和键投影的点积运算生成大小为 $C\times NC$ 的注意力分数矩阵，相比于大小为 $HW\times HW$ 的常规空间注意力矩阵，这样做可以有效缩减参数并提高模型效率。注意力分数反映目标窗口与前 N 个窗口中每个窗口之间的相似度，通道维度上计算得到的注意力分数也指导了信息从 Top-N 窗口到目标窗口的传递，解决了通道注意力机制对局部上下文细节敏感性不足的问题，提高了注意力计算决策过程中的透明度，由此在通道维度上增强了其空间关联性和表达性。此过程的表示如式(5)所示：

$$A_{\text{Attention Scores}}=\text{Softmax}\left(\frac{\boldsymbol{Q}_t\boldsymbol{K}_t^{\text{T}}}{\sqrt{D}}\right) \tag{5}$$

最后，注意力分数与值向量 $\boldsymbol{V}$ 相乘，获取目标窗口更新后的特征表示。最终的输出计算过程如式(6)所示：

$$A_{\text{Attention Output}}=\text{Softmax}\left(\frac{\boldsymbol{Q}_t\boldsymbol{K}_t^{\text{T}}}{\sqrt{D}}\right)\cdot\boldsymbol{V}_t \tag{6}$$

式中：$\boldsymbol{Q}$ 为查询；$\boldsymbol{K}$ 为键；$\boldsymbol{V}$ 为值；D 为缩放因子。通过这样的交叉注意力机制，TNCA 能够有效地将局部相关窗口的信息集成到目标窗口，这不仅提升了局部特征的表达力，还保持了计算效率。在 TNCTB 中，紧随自注意力模块的是前馈网络。本文采用了 Uformer[28] 中的 LeFF 前馈网络，它旨在通过引入深度卷积和门控机制 GELU，进一步提升信息的流动效率和特征的表达能力。

2.4 特征增强模块

在低光噪声图像中，Transformer 的深度特征提取在全局范围内计算，由于大量有用信息被大噪声干扰，导致去除噪声后的图像细节丢失严重，同时可能会导致高频信息丢失，从而造成特征表达不充分。为此，本文设计了基于 CNN 的特征增强模块(FEB)，其结构如图 4 所示。FEB 采用串行结构，利用线性层将可分离卷积的输出特征映射到不同的维度，并进行特征转换，从而提取更有用的特征表示，以此增强图像的局部特征表达，更精确地捕捉图像细节，完成高质量特征重建。

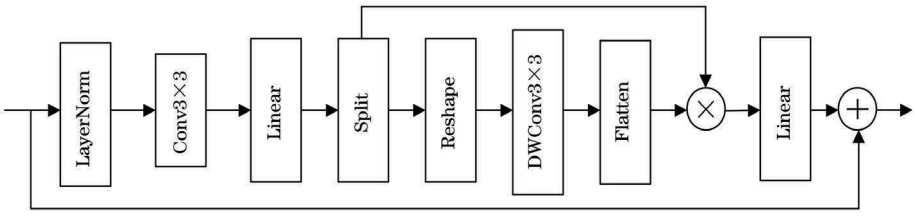

图 4 特征增强模块

Fig. 4 Feature enhancement block

3 实验与结果分析

3.1 评价指标

参考之前的图像去噪任务，本文选用峰值信噪比(PSNR，P_{PSNR})和结构相似性(SSIM)作为去噪后图像质量的评价指标，选用 Params 和 GFLOPs 作为衡量模型复杂度的指标。峰值信噪比、结构相似性定义分别为：

$$M_{\text{MSE}}=\frac{1}{hw}\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[I(i,j)-K(i,j)]^2 \tag{7}$$

$$P_{\text{PSNR}}=10\times\lg\left[\frac{(2^n-1)^2}{M_{\text{MSE}}}\right] \tag{8}$$

$$\text{SSIM}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{9}$$

式中：M_{MSE} 表示大小为 $h\times w$ 的原始图像 I 与去噪后的图像 K 之间的均方误差；μ_x 和 μ_y 是图像的均值；σ_x^2 和 σ_y^2 是图像的方差；σ_{xy} 是图像的协方差；c_1 和 c_2 是维持稳定的常数。

均方误差的值越低，峰值信噪比的值就越高，表明更高的峰值信噪比值有更好的视觉质量与去噪效果；结构相似性值越大，表明去噪后的图像与原始无噪声图像之间的失真越小，即图像的相似性越高。Params 表示模型需要学习的参数量。GFLOPs 表示模型的计算量。

3.2 实验设置及数据集

本文选用 SIDD[29] 作为训练集，该数据集专门用于图像去噪研究，同时这些图像在不同光照条件下进行拍摄，分为低光、正常光、高曝光 3 种光照条件，且真实世界中拍摄照片的噪声大多是由低光导致的。对 SIDD 训练集中的原始图像进行随机裁剪，将每一个图像随机裁剪为 300 个大小为 256×256 像素的图像块，并随机增加训练数据的多样性。本文使用 SIDD 的 Medium 版本，在测试算法性能时使用 SIDD 的测试集以及 DND[30] 数据集。

本文训练采用 AdamW 优化器，动量项为(0.9，0.999)。采用余弦退火策略将学习率从初始的 2×10^{-4} 衰减到 1×10^{-6}。使用 PyTorch 深度学习框架，并在 NVIDIA V100 显卡上运行，在训练 200 个 Epoch 后达到收敛。

3.3 对比实验

表 1 给出了本文方法与其他先进的图像去噪方法的比较结果(最优结果加粗标注，次优结果加下划线标注)，对比方法包括基于先验的 BM3D[31]，基于 CNN 的 DnCNN[15]、CBDNet[16]、SERLNet[32]、CycleISP[33]、MIRNet[34]、DeamNet[35]，以及基于 Transformer 的 Uformer[20]、MAXIM[10]、Xformer[21]、LRT[11]。

从表 1 可以观察到，基于 Transformer 的方法指标更好，并且一些 Transformer 方法通过限制注意力在局部窗口或通道内来减少计算量，但这可能忽略了图像的非局部自相似性。相比之下，在兼顾局部与全局信息特征提取的情况下，本文所提方法在含有低光照明环境的 SIDD 训练集中更具优越性。本文方法与 Xformer 有同样好的指标效果；与 BM3D、DnCNN 等经典方法相比，本文方法的去噪指标均有显著提升；与 CycleISP、MIRNet 等近年来的先进方法相比，本文方法的去噪性能同样呈现出不同程度的提高。在 SIDD 数据集上，与其他先进方法相比，本文方法的 PSNR 提升了 0.02 dB～14.33 dB，本文方法的 PSNR 和 SSIM 比经典算法 BM3D 分别高出 14.33 dB 和 0.276，比最经典的深度学习方法 DnCNN 分别高出 16.32 dB 和 0.378，比交互性先驱模型 CBDNet 分别高出 9.20 dB 和 0.160。在 DND 数据集上，相比于其他先进方法，本文方法的 PSNR 提升了 0.02 dB～7.63 dB，其中 PSNR 和 SSIM 相比于 BM3D 分别提高了 5.55 dB 和 0.107，相比于 DnCNN 分别提升了 5.55 dB 和 0.107。本文方法通过超像素引导的方式，对具有高度关联性的窗口进行聚合，实现了更精细的局部信息处理，同时保持了对全局内容的敏感性。

表 1 不同方法在 SIDD 和 DND 数据集上的定量结果

Table 1 Quantitative results of different methods on SIDD and DND datasets

方法	SIDD		DND	
	PSNR/dB	SSIM	PSNR/dB	SSIM
BM3D	25.65	0.685	34.51	0.851
DnCNN	23.66	0.583	34.51	0.851
CBDNet	30.78	0.801	38.06	0.942
SERLNet	34.61	0.883	38.95	0.947
DeamNet	39.47	0.957	39.63	0.953
CycleISP	39.52	0.957	39.56	0.956
MIRNet	39.72	0.959	39.88	0.956
LRT	39.82	0.959	39.91	0.956
Uformer	39.89	<u>0.960</u>	39.96	0.956
MAXIM	<u>39.96</u>	<u>0.960</u>	40.04	0.956
Xformer	**39.98**	<u>0.960</u>	**40.19**	<u>0.957</u>
AISGFormer	**39.98**	**0.961**	<u>40.06</u>	**0.958**

图 5 展示了几种方法在 SIDD 和 DND 数据集上的去噪效果对比。结果表明，当图像中超像素区域明显、图像有较强的非局部自相似特性时，所提 AISGFormer 可以有效地恢复被噪声污染的区域。相比之下，将注意力限制在窗口内的方法，无法利用与目标窗口相似的其他窗口的信息，因而不能达到很好的去噪效果。例如，SIDD 数据集上的结果表明，本文方法的去噪效果优于其他方法，进一步说明了非局部自相似性的利用对于低光图像去噪有着重要的作用。

在表 2 中，相较于其他先进图像去噪模型，本文模型的计算量处于较低水平，并且相较于 MAXIM，本文模型在计算量低 47.2%的情况下，PSNR 指标提升 0.02 dB。相比于 CBDNet、Uformer、Xformer，本文模型需要执行更大的参数量和计算量，但由于采用了先进的全局交互模块和网络结构设计，本文网络具有更优的特征表示能力和相似性

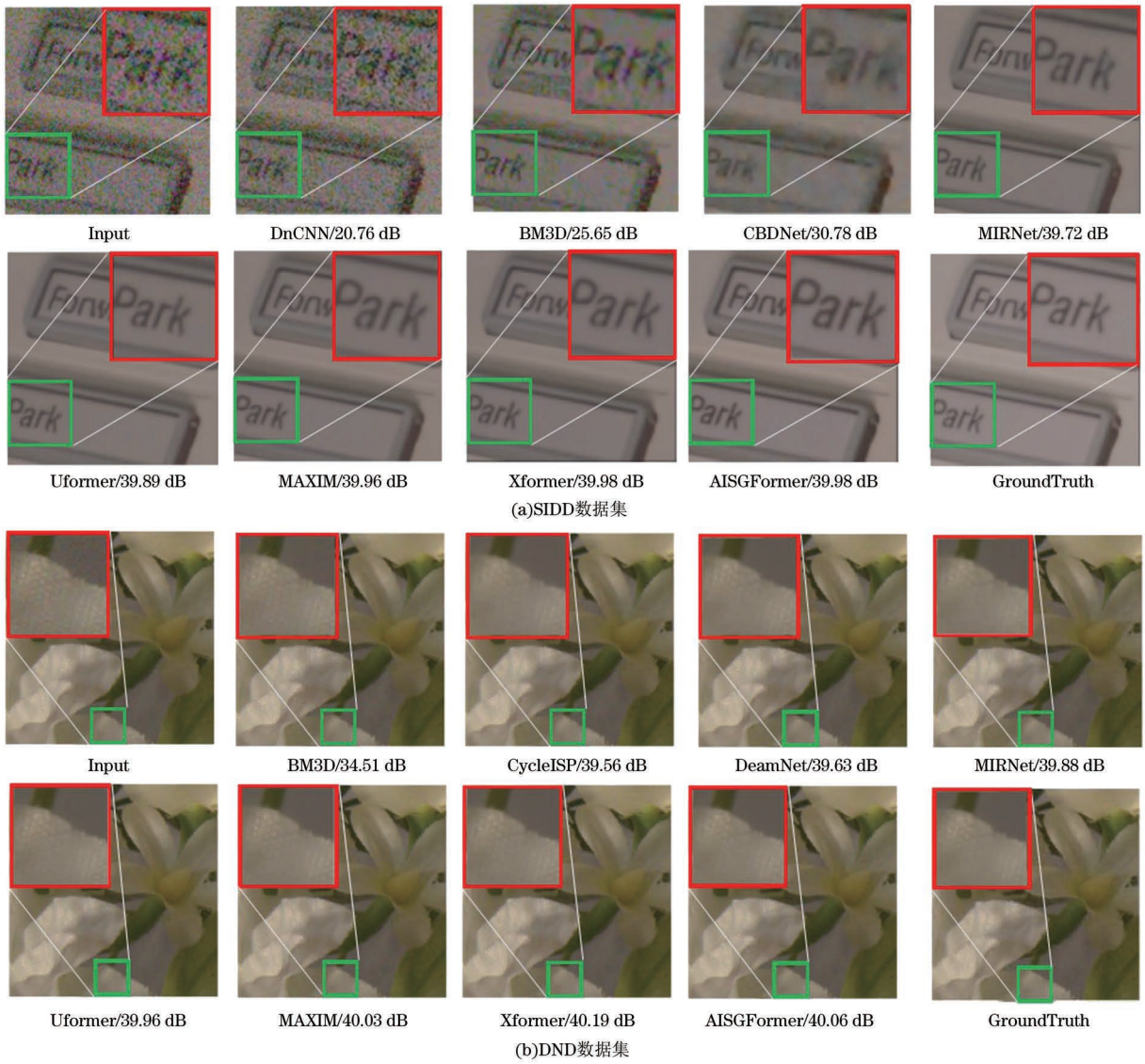

图 5 不同方法的去噪效果

Fig. 5 The denoising effect of different methods

提取能力。因此，本文模型在强大的硬件支持下得到的结果会更好。综上，本文模型去噪指标最佳，能够处理复杂的低光图像，在处理多样化的真实噪声任务中效果出色。

表 2 不同模型的参数量和计算量对比

Table 2 Comparison of parameter and computational complexity among different models

对比项	CBDNet	CycleISP	MIRNet	MPRNet	SwinIR	Uformer	MAXIM	Xformer	AISGFormer
Params/10^6	4.37	2.84	20.10	15.74	11.50	50.88	22.20	25.23	34.35
GFLOPs	40.28	335.01	196.76	1 393.83	201.20	89.46	169.50	42.20	89.42

SIDD 数据集同时提供了 PNG(sRGB)中 3 类光种的数据，在表 3 中，展现了使用这 3 种输入数据进行训练的结果。除输入数据外，其余训练参数均保持一致。实验结果表明，在低光条件下，性能数据虽不及正常光照但优势依然明显，这是因为低光条件下噪声更大且图像整体更暗，具有更多的、重复相似的纹理和结构特征，使用 sRGB 会使得集中在较低像素值处的信息丢失。但本文方法利用低光图像的非局部自相似性这一特性，聚焦于低光噪声的相似噪声区域。因此，使用本文方法更有利于网络学习从低光噪声图像到干净图像的映射。

表 3 不同输入光种对网络性能的影响

Table 3 The impact of different input light types on network performance

格式	低光		正常光		高曝光	
	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
sRGB	39.83	0.959	39.96	0.960	40.01	0.961

3.4 消融实验

影响 AISGFormer 性能的主要参数包括网络深度、窗口尺寸和各模块的使用。本文采用 SIDD 作为测试数据集，进行网络性能影响因素的消融实验。

3.4.1 网络深度消融实验

本文编解码器网络采用分级结构设计，其不同层级的深度在图像恢复过程中有不同的重建程度。网络层级过浅会导致特征表达不足，而层级过深则会增加模型过拟合的风险，进而影响其泛化能力。合适的网络层数有助于网络更深入地学习图像的特征和上下文信息，从而有效执行更为复杂的去噪任务。图 6 展示了不同网络深度（编解码器部分的级数 n 分别为 3、4、5、6）对去噪性能的影响。由图 6 可知，当编解码器架构级数取 4 时，网络性能最优。

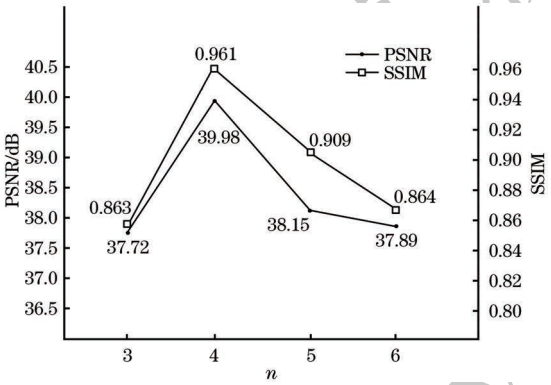

图 6 编解码器层数对网络性能的影响

Fig. 6 The impact of encoder decoder layers on network performance

3.4.2 窗口尺寸消融实验

在网络模型训练过程中，较大的窗口尺寸意味着每次处理较多的图像数据，这可能导致较高的计算复杂度，需要较多的计算资源。较小的窗口尺寸可以保留更多的细节信息，但也可能导致模型难以捕捉全局结构。因此，要选择合适的窗口尺寸大小。

从图 7 可以得知，当窗口尺寸为 8 时，模型在 PSNR 和 SSIM 上达到最佳性能，比窗口尺寸为 4 和 16 时分别高出了 3.36 和 4.45 dB。这说明中等大小的窗口在保持较低计算复杂度的同时，能够有效捕获上下文信息，从而优化图像恢复质量。较小的窗口尺寸（如 4）虽然减少了计算负担，但限制了模型捕获上下文信息的能力，从而影响模型性能；较大的窗口尺寸（如 16）虽然能够覆盖更广泛的上下文，但是增加了计算负担，并且可能引入过多信息，影响了模型的去噪效果。

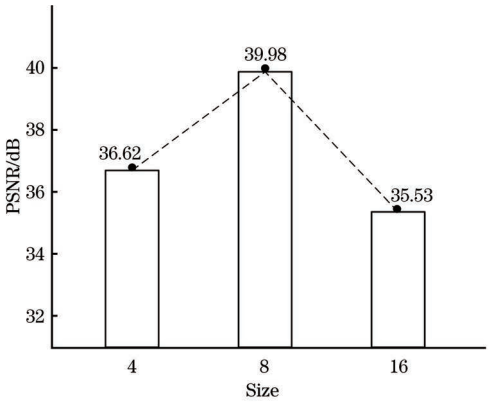

图 7 不同窗口大小的分析结果

Fig. 7 Analysis results of different window sizes

3.4.3 各模块的消融实验

为验证 AISGFormer 中各个部分的有效性，本节在 SIDD 数据集上对 SSM、TNCA 和 FEB 进行消融实验，具体来说，将 TNCA 替换为基于滑动窗口的多头自注意力机制，将 FEB 改为普通 3×3 卷积层，将 SSM 删掉。实验结果如表 4 所示。

表 4 各模块的消融实验结果

Table 4 The ablation experiment results of each module

序号	SSM	TNCA	FEB	Params/10^6	FLOPs/10^9	PSNR/dB	SSIM
1	—	—	√	39.46	88.65	39.53	0.954
2	—	√	√	33.29	81.21	39.72	0.956
3	√	—	√	40.52	96.86	39.68	0.956
4	√	√	—	33.23	87.52	39.93	0.960
5	√	√	√	34.35	89.42	**39.98**	**0.961**

由表 4 可知，缺少任何模块都会使得网络性能下降。第 1 组实验表明，基于窗口交互的交叉注意力机制是 AISGFormer 的核心组件，它将同一超像素区域内窗口的信息聚合到目标窗口，从而自适应地利用自相似性抑制区域相似噪声。同时第 1 组实验表明，引入了增强计算过程透明度和空间关联性的注意力机制，能够更准确地评估模型的行为，优化模型性能。第 2 组实验表明，通过 SSM 对注意力机制进行引导与交互，充分利用图像的自相似性，能够对图像局部细节及特征进行更精确的表征。第 3 组实验表明，使用 SSM 引导交叉注意力机制而非普通的基于窗口的注意力机制，更能交互局部与全局的

信息和细节。第 4 组实验表明，特征增强模块能够增强图像的局部特征表达，完成高质量特征重建。

4 结束语

本文提出一种基于超像素引导和自适应交互的 AISGFormer 模型，该模型融合 TNCA，解决了低光图像去噪过程中面临的图像自相似性利用不足的问题。AISGFormer 通过精细化的全局交互策略，有效捕获低光噪声图像非局部自相似性和细节信息。实验结果表明，相较于其他先进的去噪方法，AISGFormer 在真实噪声数据集上取得了更优的效果，以独特的交互形式，在兼顾全局信息和局部细节的同时，有效地去除了复杂的噪声。但 AISGFormer 在模型效率方面存在局限性，未来将在参数性能和损失函数方面进行优化，继续探索超像素分割对注意力机制的指导作用。

参考文献

[1] 陈钱. 先进夜视成像技术发展探讨[J]. 红外与激光工程, 2022, 51(2): 9-16.
CHEN Q. Discussions on the development of advanced night vision imaging technology [J]. Infrared and Laser Engineering, 2022, 51(2): 9-16. (in Chinese)

[2] GUO P Y, ASIF M S, MA Z. Low-light color imaging via cross-camera synthesis[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(4): 828-842.

[3] FENG H S, WANG L Z, WANG Y Z, et al. Learnability enhancement for low-light raw image denoising: a data perspective[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 370-387.

[4] YINGKUN H, JUN X, MINGXIA L, et al. NLH: a blind pixel-level non-local method for real-world image denoising[J]. IEEE Transactions on Image Processing, 2020, 29: 5121-5135.

[5] YANG W, WANG W, HUANG H, et al. Sparse gradient regularized deep retinex network for robust low-light image enhancement[J]. IEEE Transactions on Image Processing, 2021, 30: 2072-2086.

[6] HUANG H F, YANG W H, HU Y Y, et al. Towards low light enhancement with RAW images[J]. IEEE Transactions on Image Processing, 2022, 31: 1391-1405.

[7] CUI Y, KNOLL A. PSNet: towards efficient image restoration with self-attention [J]. IEEE Robotics and Automation Letters, 2023, 8(9): 5735-5742.

[8] HASSANI A, WALTON S, LI J C, et al. Neighborhood attention transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2023: 6185-6194.

[9] XIAO Y, YUAN Q Q, JIANG K, et al. TTST: a top-k token selective transformer for remote sensing image super-resolution[J]. IEEE Transactions on Image Processing, 2024, 33: 738-752.

[10] TU Z Z, TALEBI H, ZHANG H, et al. MAXIM: multi-axis MLP for image processing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2022: 5759-5770.

[11] ZHANG S S, MENG N, LAM E Y. LRT: an efficient low-light restoration transformer for dark light field images[J]. IEEE Transactions on Image Processing, 2023, 32: 4314-4326.

[12] 朱凯, 李理, 张彤, 等. 基于 Transformer 的多阶段运动模糊图像修复网络[J]. 计算机工程, 2024, 50(9): 276-285.
ZHU K, LI L, ZHANG T, et al. Multi-stage motion blur image restoration network based on Transformer [J]. Computer Engineering, 2024, 50(9): 276-285. (in Chinese)

[13] ANTONI B, BARTOMEU C, JEAN M M. A non-local algorithm for image denoising[J]. Computer Vision and Pattern Recognition, 2005, 2: 60-65.

[14] DABOV K, FOI A, KATKOVNIK V, et al. Image denoising by sparse 3-D transform-domain collaborative filtering[J]. IEEE Transactions on Image Processing, 2007, 16(8): 2080-2095.

[15] ZHANG K, ZUO W M, CHEN Y J, et al. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155.

[16] GUO S, YAN Z F, ZHANG K, et al. Toward convolutional blind denoising of real photographs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2019: 1712-1722.

[17] 高煜宝, 文志诚. 基于注意力机制的双路解码器图像去噪方法[J]. 计算机工程, 2024, 50(9): 324-332.
GAO Y B, WEN Z C. Dual decoder image denoising method based on attention mechanism[J]. Computer Engineering, 2024, 50(9): 324-332. (in Chinese)

[18] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2010.11929.

[19] CHEN H T, WANG Y H, GUO T Y, et al. Pre-trained image processing transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 12294-12305.

[20] WANG Z, CUN X, BAO J, et al. Uformer: a general U-shaped transformer for image restoration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022: 17662-17672.

[21] JIALE Z, YULUN Z, JINJIN G, et al. Xformer: hybrid X-shaped transformer for image denoising[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2303.06440.

[22] 刘凯, 任洪逸, 李莹, 等. 基于交叉模态注意力特征增强的医学视觉问答[J]. 计算机工程, 2025, 51(6): 49-56.
LIU K, REN H Y, LI Y, et al. Medical visual question answering based on cross-modal attention feature enhancement[J]. Computer Engineering, 2025, 51(6): 49-56. (in Chinese)

[23] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.

[24] LIU Y J, YU C C, YU M J, et al. Manifold SLIC: a fast method to compute content-sensitive superpixels [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2016: 651-659.

[25] ACHANTA R, SUSSTRUNK S. Superpixels and polygons using simple non-iterative clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2017: 4895-4904.

[26] ZHANG A, REN W, LIU Y, et al. Lightweight image

super-resolution with superpixel token interaction[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2023: 12682-12691.
[27] PAN Y J, WEN C, ZHAO X L, et al. Irregular tensor representation for superpixel- guided hyperspectral image denoising[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 12-23.
[28] ZHOU M, XU Z, TONG R K. Superpixel-guided class-level denoising for unsupervised domain adaptive fundus image segmentation without source data[J]. Computers in Biology and Medicine, 2023, 162: 107061.
[29] ABDELHAMED A, LIN S, BROWN M S. A high-quality denoising dataset for smartphone cameras[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 1692-1700.
[30] PLOTZ T, ROTH S. Benchmarking denoising algorithms with real photographs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2017: 2750-2759.
[31] DABOV K, FOI A, KATKOVNIK V, et al. Image denoising by sparse 3-D transform-domain collaborative filtering[J]. IEEE Transactions on Image Processing, 2007, 16(8): 2080-2095.
[32] PARAS M, ZHU L, NING X, et al. Improving extreme low-light image denoising via residual learning[C]// Proceedings of the IEEE International Conference on Multimedia and Exposition. Washington D. C., USA: IEEE Press, 2019: 916-921.
[33] ZAMIR S W, ARORA A, KHAN S, et al. CycleISP: real image restoration via improved data synthesis[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2020: 2693-2702.
[34] ZAMIR S W, ARORA A, KHAN S, et al. Learning enriched features for real image restoration and enhancement [EB/OL]. [2024-07-05]. https://arxiv.org/abs/2003.06792.
[35] REN C, HE X H, WANG C C, et al. Adaptive consistency prior based deep network for image denoising[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2021: 8592-8602.

文字编辑 吴云芳
栏目编辑 宋 圆