E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)11-3524-11

论文引用格式:Li M K, Liu W and Chen W D. 2025. Image denoising via Swin Transformer V2 and feature fusion U-Net. Journal of Image and Graphics, 30(11):3524-3534(利铭康, 柳薇, 陈卫东. 2025. Swin Transformer V2和特征融合的U-Net图像去噪方法. 中国图象图形学报, 30(11):3524-3534)[DOI:10.11834/jig.240659]

Swin Transformer V2和特征融合的U-Net图像去噪方法

利铭康,柳薇*,陈卫东

华南师范大学计算机学院,广州　510631

摘　要:**目的**　Transformer神经网络在图像去噪上效果显著,但要进一步提升去噪质量,需要增加大量的训练和预测资源;另外,原始Swin Transformer对高分辨率图像输入缺少良好的适应性。对此,设计了一种基于Swin Transformer V2的U-Net图像去噪深度学习网络。**方法**　该网络在下采样阶段设计了一种包括Swin Transformer V2和卷积并行提取特征的Transformer块,然后在上采样阶段设计了一种特征融合机制以提升网络的特征学习能力。针对图像去噪任务对Transformer块修改了归一化位置及采用镜像填充机制,提高Swin Transformer V2块的适应性。**结果**　在CBSD68(color Berkeley segmentation dataset)、Kodak24、McMaster和彩色Urban100这4个图像去噪常用测试集上进行去噪实验,选择峰值信噪比(peak signal-to-noise ratio, PSNR)作为去噪效果的评价指标,在噪声等级为50的去噪实验中,得到的平均PSNR值分别为28.59 dB、29.87 dB、30.27 dB和29.88 dB,并与几种流行的基于卷积和基于Transformer的去噪方法进行比较。本文的去噪算法优于基于卷积的去噪方法,而相比于性能接近的基于Transformer方法,本文去噪算法所需浮点运算量仅为26.12%。**结论**　本文方法使用的Swin Transformer V2和特征融合机制均可以有效提升图像去噪效果。与现有方法相比,本文方法在保证或提升图像去噪效果的前提下,大幅度降低了训练和预测所需要的计算资源。

关键词:深度学习;图像去噪;Swin Transformer;U-Net;特征融合

Image denoising via Swin Transformer V2 and feature fusion U-Net

Li Mingkang, Liu Wei*, Chen Weidong

School of Computer Science, South China Normal University, Guangzhou　510631, China

Abstract: **Objective** Image denoising represents a fundamental challenge in the field of image processing, with the primary goal of recovering clear images from their noise-degraded counterparts. Throughout the image acquisition and formation processes, multiple factors such as suboptimal lighting conditions, temperature fluctuations, and imaging system corrections can considerably contribute to the presence of noise in the final images. The impact of image noise extends beyond mere visual perception degradation; specifically, it substantially affects the accuracy of advanced image processing tasks, including image segmentation and object recognition. Traditional denoising approaches, which require manual tuning of numerous parameters, are complex and time consuming. While convolutional neural network (CNN)-based denoising methods have demonstrated promising results, pure Transformer neural networks have shown great effectiveness in image

收稿日期:2024-11-18;**修回日期**:2025-03-07;**预印本日期**:2025-03-15

*通信作者:柳薇　liuwei@m.scnu.edu.cn

基金项目:国家自然科学基金项目(62377015)

Supported by: National Natural Science Foundation of China (62377015)

denoising. Moreover, CNN-based denoising methods are inherently constrained by their convolution kernel sizes, which limits their capability to utilize global image information. Conversely, Transformer-based methods effectively leverage global image information, but they demand exponentially increasing computational resources for enhanced detail restoration. In addition, the original Swin Transformer lacks good adaptability to high-resolution image inputs. In response to these challenges, we have developed a U-Net image denoising method based on Swin Transformer V2, which successfully integrates Transformer features with conventional convolutional features. Thus, it achieves remarkable denoising performance and visual quality across standard image denoising datasets. **Method** We present a novel image denoising network method based on Swin Transformer V2. The network consists of downsampling and upsampling stages. During downsampling, images undergo feature extraction in progressive feature spaces. Each encoder layer contains a different number of DB-Transformer and Transformer blocks. In each DB-Transformer block, parallel Transformer and local convolution branches independently extract Transformer and local convolution feature maps, respectively, and these features interact before being passed to the next block. During upsampling, the network reconstructs images from extracted features. The upsampling decoders contain only Transformer blocks, with each decoder preceded by a feature fusion module that receives features from downsampling and upsampling stages. The feature fusion module incorporates a global average pooling component and a multilayer perceptron, which, through a softmax function, generate dynamic weights that enable the network to adaptively select more informative features from different feature maps. Long-skip connections are employed before the final output, as noisy and clean images share considerable information, and these connections prevent gradient vanishing. To enhance the adaptability and denoising performance of Swin Transformer V2 blocks within our network, we strategically position layer normalization before self-attention computation. This way accelerates network convergence and optimizes it for small-scale model training. Furthermore, instead of utilizing masked shifted window-based multi-head self-attention, we implement mirror padding for incomplete window sections, which enhances the contribution of edge pixels to training. This approach ensures their equal importance in image denoising tasks. We train on a combined dataset of Berkeley segmentation dataset 500(BSD500), diverse 2K resolution high quality images(DIV2K), Flickr2K, and Waterloo exploration database(WaterlooED), with random patch selection from each image. Our experiments utilize the Charbonnier loss function and progressive training mechanism, and they are conducted on a single NVIDIA GeForce RTX 4070Ti Super GPU. **Result** To validate the effectiveness of the model, we conducted comprehensive testing using four widely recognized datasets in the image denoising domain: color Berkeley segmentation dataset (CBSD68), Kodak24, McMaster, and color Urban100. With peak signal-to-noise ratio (PSNR) as the primary evaluation metric, our denoising experiments achieved impressive average PSNR values of 28.59 dB, 29.87 dB, 30.27 dB, and 29.88 dB across these datasets with a noise level of 50. Compared with traditional algorithms, our approach demonstrates significantly enhanced denoising effects and visual perception, with PSNR metrics surpassing those of CNN-based denoising methods. Notably, while achieving comparable performance to Transformer-based methods, our denoising algorithm requires only 26.12% of the floating-point operations. We also conducted extensive ablation studies to verify the effectiveness of our proposed method. In particular, various aspects including the number of convolution blocks, feature fusion modules, and Transformer block improvements were examined. The experimental results convincingly demonstrate that our approach effectively balances training efficiency with image denoising performance. **Conclusion** We have successfully developed and implemented a U-Net deep learning network model based on Swin Transformer V2 for image denoising, which definitively establishes the viability of Swin Transformer V2 in the image denoising domain. Our network architecture effectively combines the strengths of local convolution and Transformer. Thus, it not only efficiently extracts valuable information from both structural feature maps but also achieves superior training efficiency. The experimental results comprehensively demonstrate that our proposed network architecture offers significant advantages in detail restoration and operational efficiency.

Key words: deep learning; image denoising; Swin Transformer; U-Net; feature fusion

0 引 言

图像去噪是图像处理中的一项重要工作，目的是从受到噪声退化的图像中恢复原本的清晰图像。在图像采集和成像过程中，由于传感器的固有噪声、环境光照和温度条件的变化，以及成像系统的校正误差等因素，最终得到的图像可能会含有较多噪声。图像噪声会影响更高级的图像处理任务的精确度，如图像分割（Brempong等，2022）、目标识别（Zuo等，2019）等。因此，图像去噪在各个专业领域都具有研究意义，如图像去噪经常运用于医学影像处理上（Mohd Sagheer和George，2020）。

传统的图像去噪方法通常基于数学模型的算法进行处理。其中具有代表性的基于空间域的方法有非局部空间滤波技术（non-local means，NLM）算法（Buades等，2005）和三维块匹配滤波（block method of 3-dimension，BM3D）算法（Dabov等，2007）。基于频域的方法则以傅里叶变换和小波变换为代表（Kivanc Mihcak等，1999），这类方法通常需要完全分离有用信号和干扰信号才能取得良好的去噪效果。这些传统方法能有效地去除简单的图像噪声，但随着图像噪声强度和图像质量的变化，这些方法的通用性有限，容易导致去噪图像出现失真现象，并且它们往往具有较高的计算成本。

随着硬件的计算能力发展，深度学习广泛运用到图像处理中。有许多研究将深度学习应用到图像去噪任务中以克服传统去噪方法的不足，取得了较好的去噪效果。在早期研究中，这些方法使用卷积神经网络（convolutional neural network，CNN）将噪声图像直接映射到清晰图像，无需额外的图像先验亦能取得更好的图像去噪效果。DnCNN（denoising convolutional neural network）采用多层卷积和残差连接，以输出与噪声的范数为损失函数来训练网络（Zhang等，2017）。RDN（deep residual channel attention network）在网络中通过密集级联多个残差密集块以充分利用分层功能取得更好的去噪效果（Zhang等，2018c），但网络计算复杂度随着残差密集块增加而显著上升，去噪效果提升很有限。FFD-Net（fast and flexible denoising convolutional neural network）在训练过程中将噪声图和退化图像一起输入网络，得到了较好的去噪效果（Zhang等，2018a）。BRDNet（deep CNN with batch renormalization）则结合了两个不同的网络以增加模型的宽度改善去噪性能，并且该方法使用批处理重归一化解决小型迷你批处理问题，还采用扩张卷积节省计算量（Tian等，2020）。一些方法（Cheng等，2021）将注意力机制引入到CNN中，包括通道注意力（Zhang等，2018b）和非局部注意力（Mei等，2021），取得了较好的去噪效果。有研究选择损失函数为切入角度，将感知损失和残差连接结合起来进行图像去噪（吴从中 等，2018）。除此之外，面向真实图像噪声的去噪任务也是备受关注的研究之一（丁岳皓 等，2023）。以上方法都使用卷积神经网络对图像去除噪声，而受限于卷积核的大小，这些方法不能从全局信息中学习特征。

近年来，基于多头自注意力机制的Transformer（Vaswani等，2017）在自然语言处理任务中取得了巨大进展。基于此发展的Vision Transformer在高级计算机视觉领域展现了巨大潜力，包括图像分割（Dosovitskiy等，2021）、目标检测（Lin等，2022）等常见的视觉任务，这是因为Transformer可以在图像中注意长距离的特征依赖程度。目前已有研究将Transformer应用到低级的图像处理任务（Yang等，2020；Yin和Ma，2022；Zhang等，2023）。与局部卷积不同，Transformer网络会同时观察图像中所有位置的像素，得到长距离的像素相关性。IPT（pre-trained image processing Transformer）首次将Transformer运用到图像增强任务中，该方法将像素点图像分割为多个小块嵌入到特征空间中，使用多头自注意力机制取得了优于基于传统卷积神经网络的去噪效果（Chen等，2021）。由于图像增强是基于像素的处理工作，所以使用原始Transformer具有非常高的计算成本。随后，SwinIR（image restoration using Swin Transformer）使用Swin Transformer（Liu等，2021）作为骨干网络，采用可移动划分窗口的多头自注意力机制，极大降低了计算复杂度（Liang等，2021）。Uformer（Wang等，2022）是一个基于U-Net架构的神经网络，该方法通过堆叠LeWin Transformer模块，同样有效降低原始Transformer的高计算复杂度。Swin Transformer V2（Liu等，2022）能显著降低原始Transformer的高计算复杂度，并提高对高分辨率图像输入的适应性，但目前仍然缺少将该方法应用到图像去噪的研究。然而，纯Transformer对于局部特征表示并非最佳的。如果为了进一步增

加局部去噪的精细程度而提高图像嵌入大小，会导致计算复杂度呈指数级增长（Yuan等，2021），延长训练和预测时间，但去噪效果提高则非常有限。考虑到局部特征对于纹理和边缘恢复仍然是非常重要的，本文提出一种去除图像噪声的神经网络模型，该模型使用U-Net神经网络架构（Ronneberger等，2015）融合了Swin Transformer V2和传统卷积特征进行去噪，使网络能并行学习全局特征和局部特征，以此提高去噪效果，同时控制计算复杂度的增长。本文方法在流行的图像去噪数据集上取得了更好的实验评价指标和视觉表现。

本文贡献如下：1）首次将Swin Transformer V2嵌入U-Net结构，用于图像去噪任务，探索了Swin Transformer V2在图像去噪领域的可行性，利用其对高分辨率图像输入更好的适应性，为处理高分辨率图像噪声问题提供了新思路。2）针对Transformer堆叠带来的巨大计算开销，设计了一种融合Swin Transformer V2和局部卷积特征的交互模式。此方法在降低网络浮点运算次数的同时，提升了去噪效果。3）结合图像去噪任务的特性，对Swin Transformer V2进行了针对性的优化，包括归一化层前置，以及在窗口位移后的边缘区域执行镜像填充，增强了模型的实用性。

1 算法模型

1.1 整体网络结构

本文方法使用U-Net网络构建了一个5层神经网络用于图像去噪。虽然池化操作在下采样过程中能够增强高层语义信息的提取，但其同时可能导致低层细节信息的部分丢失。为此，在设计网络结构时将下采样次数控制为两次，以在获取抽象特征的同时最大限度地保留图像的细节特征。

图1展示了本文方法的整体网络结构。含噪图像可表示为$\boldsymbol{I}_{\text{noised}} \in \mathbf{R}^{B\times 3\times H\times W}$，其中，$B$是批次大小，3表示彩色图像的3个颜色通道，$H$和$W$是图像的空间维度。该图像首先经过非重叠的嵌入层$M_{\text{embed}}(\cdot)$得到低层次的特征图$\boldsymbol{R}_0 \in \mathbf{R}^{B\times C\times H\times W}$，其中，$C$表示嵌入后的特征图通道数量。该过程表示为

$$\boldsymbol{F}_0 = M_{\text{embed}}\left(\boldsymbol{I}_{\text{noised}}\right) \tag{1}$$

式中，$M_{\text{embed}}(\cdot)$包含一个卷积核大小为3×3、步长为1的卷积层以及层归一化。通过嵌入层，将图像分割成不重叠的小块并映射到特征空间，在该嵌入层加入归一化可以有效避免梯度消失或梯度爆炸问题。经过嵌入层得到的特征图$\boldsymbol{F}_0$被输入到第1层编码器模块中进行处理，即可得到该层编码器的输出特征图$\boldsymbol{F}_1$。本文设计的5层编码器与解码器之间均采用跳跃连接，可以充分融合各层特征信息，促进浅层细节与深层语义的互补，从而提升重建图像的精度与细节恢复能力。最后一层得到的特征图$\boldsymbol{F}_5$将被输入到反嵌入模块$M_{\text{unembed}}(\cdot)$中重构图像。$M_{\text{unembed}}(\cdot)$包含一个卷积核大小为3×3、步长为1的卷积操作以及层归一化模块，通过反嵌入模块可以将特征图的通道数量恢复到与输入图像一致。考虑到噪声图像和预测结果之间共享很多信息，因此在最后引入残差连接，使网络专注于学习两者之间的差异，进而更精确地恢复原始图像细节。令$\boldsymbol{I}_{\text{estimate}}$表示输出的预测图像，该过程的表达式为

$$\boldsymbol{I}_{\text{estimate}} = M_{\text{unembed}}\left(\boldsymbol{F}_5\right) + \boldsymbol{I}_{\text{noised}} \tag{2}$$

1.2 DB-Transformer模块

在编码器中，本文设计了DB-Transformer模块来取代传统的Transformer模块，这是一种具有两个特征提取分支的模块。该模块的结构如图2所示。具体而言，该模块包含并行的Transformer分支T和卷积分支D。首先，将输入特征经过一个卷积层，用于扩展通道数量，以便为后续的分支处理提供更丰富的特征表达。然后将特征图沿通道维度均分（Chunk）为两部分，分别送入Transformer分支T和卷积分支D进行处理。该过程表示为

$$\left\{\boldsymbol{F}_T^{\text{in}}, \boldsymbol{F}_D^{\text{in}}\right\} = f_{\text{Chunk}}\left(f_{\text{Conv }1\times 1}\left(\boldsymbol{F}_{\text{in}}\right)\right) \tag{3}$$

式中，$\boldsymbol{F}_T^{\text{in}}$表示输入到Transformer分支的特征子集，$\boldsymbol{F}_D^{\text{in}}$表示输入到卷积分支的特征子集，$\boldsymbol{F}_{\text{in}}$表示输入特征，$f_{\text{Conv }1\times 1}(\cdot)$表示卷积核大小为1×1的卷积操作。

对于卷积分支D，先通过一个1×1卷积层进一步增加通道数；然后利用一个3×3可分离深度卷积模块（depthwise separable convolution，DConv）提取局部空间特征；再经过另一个1×1卷积层恢复原有通道数；并采用GELU（Gaussian error linear unit）激活函数引入非线性变换，避免ReLU（rectified linear units）激活函数中可能出现的神经元不再学习的现象。最后通过残差连接将卷积分支的输入与经过上

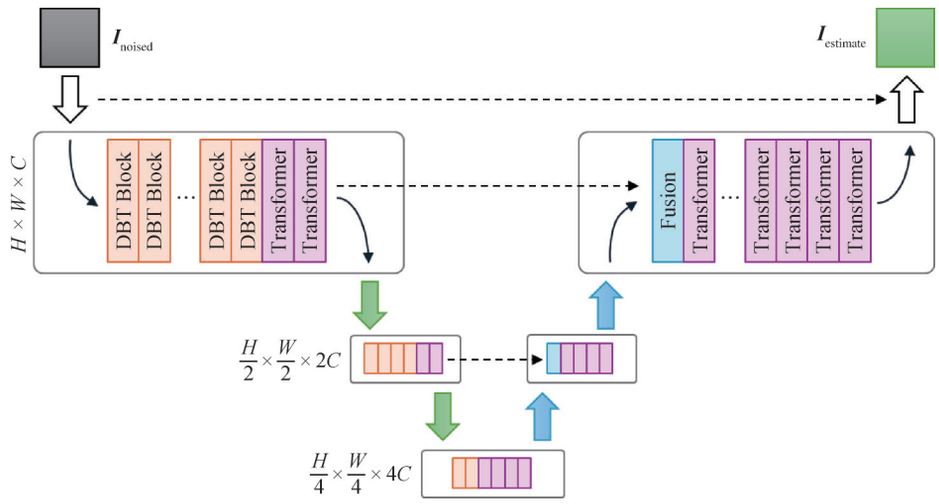

图1　网络结构图

Fig. 1　Network model diagram

述操作后的输出相加，得到卷积分支的最终输出 $\boldsymbol{F}_D$，过程记为

$$\boldsymbol{F}_D = f_{\mathrm{GELU}}\left(f_{\mathrm{Conv}\ 1\times1}\left(f_{\mathrm{DConv}\ 3\times3}\left(f_{\mathrm{Conv}\ 1\times1}\left(\boldsymbol{F}_D^{\mathrm{in}}\right)\right)\right)\right) + \boldsymbol{F}_D^{\mathrm{in}} \quad (4)$$

为了避免层归一化过程中减均值和除以标准差的操作可能会对图像对比度造成影响，因此此处的卷积操作均去除了归一化。

对于 Transformer 分支 $\boldsymbol{T}$，其处理过程则包括多头自注意力机制和前馈神经网络模块，通过一系列变换后输出特征 $\boldsymbol{F}_T$。最后，将 $\boldsymbol{F}_T$ 与 $\boldsymbol{F}_D$ 沿通道维度拼接（concat），再经过一个卷积核大小为 1 × 1 的卷积层进行融合，最终得到该模块的输出特征 $\boldsymbol{F}_{\mathrm{out}}$。该过程的数学表达式为

$$\boldsymbol{F}_{\mathrm{out}} = f_{\mathrm{conv}\ 1\times1}\left(f_{\mathrm{concat}}\left(\boldsymbol{F}_T, \boldsymbol{F}_D\right)\right) \quad (5)$$

这种设计通过并行处理局部细节与全局依赖，有效结合了卷积神经网络与 Transformer 的优势，有助于提高模型在细节恢复和全局信息捕捉方面的性能。

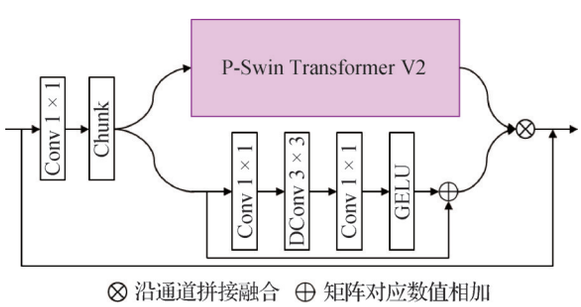

图2　DB-Transformer 模块结构图

Fig. 2　DB-Transformer module diagram

1.3　P-Swin Transformer V2 模块

针对图像去噪任务对 Swin Transformer V2 作出适应性改进，并提出 P-Swin Transformer V2 模块。

首先进一步描述该模块的结构，如图 3 所示。该模块的具体设计遵循传统的“Norm-MHSA-Norm-FFN”结构，主要包括两个子模块：划分窗口的多头自注意力层和多层感知机模块（multilayer perceptron，MLP）。图中，标星号的子模块仅在带位移划分窗口的 P-Swin Tranformer 块中使用。在多头自注意力部分，采用了 Swin Transformer V2 余弦注意力，利用余弦相似度（cos）替代点积，并通过可学习的缩放系数 τ 动态调整注意力分布。注意力具体计算为

$$f_{\mathrm{Attention}}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = f_{\mathrm{softmax}}\left(\frac{\cos\left(\boldsymbol{Q}, \boldsymbol{K}\right)}{\tau} + \boldsymbol{B}\right) \cdot \boldsymbol{V} \quad (6)$$

式中，$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 是对特征图进行线性变换得到的矩阵，并且 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 需进行 L2 正则化，确保每个向量的模长为 1，$\boldsymbol{B}$ 表示采用小型的 MLP 网络生成的位置偏置，softmax 函数的结果与 $\boldsymbol{V}$ 相乘即可得到注意力分数。

在完成自注意力计算后，特征图传入多层感知

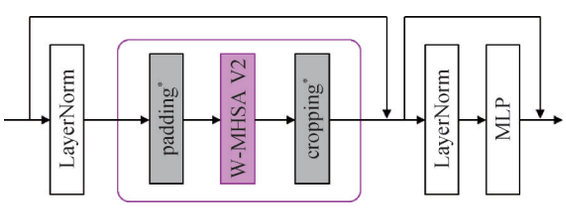

图3　P-Swin Transformer V2 模块结构图

Fig. 3　P-Swin Transformer V2 module diagram

机模块。该模块采用卷积操作来增强局部特征捕捉能力。首先对输入应用第1个卷积操作，然后采用LeakyReLU（leaky rectified linear unit）激活函数对卷积后的结果进行非线性变换，再使用第2个卷积层将特征映射回原始维度。

1.4 层归一化位置

Swin Transformer V2为了能够训练一个规模极大的通用模型来应对图像分割等高级视觉任务，在原始Swin Transformer的基础上将传统的点积注意力替换成了缩放余弦相似注意力，并且在计算自注意力后再进行层归一化（LayerNorm），从而稳定大模型的训练过程。由于缩放余弦相似注意力本身带有对输入的正则化操作，因此，Swin Transformer V2将层归一化置于自注意力模块之后来稳定每一层的输出，在网络很深的情况下可以抑制每层的输出不断变大的趋势，从而降低浅层输出和深层输出的差距。考虑到原始Swin Transformer V2需要训练一个参数量超过30亿的大模型，然而对于提出的模型规模远小于此的大部分中小模型而言，采用后置层归一化在实践中会因正则化方式和更新策略不同而显著减缓收敛速度，这一点从后续消融实验结果中得到了验证。基于这一考虑，为了更快地训练并保持模型的准确度与稳定性，在Swin Transformer V2的基础上重新回归到层归一化位置前置，充分结合了原始Swin Transformer的高效训练过程和Swin Transformer V2对高分辨率输入的良好适应性。

1.5 镜像填充

回顾原始Swin Transformer，当采用位移窗口策略计算窗口自注意力时，特征图边缘区域无法形成完整的规则窗口。原始Swin Transformer为解决此问题，采用循环位移（cyclic shift）对边缘块进行位置重组，通过将多余的特征块循环移动到对侧，拼凑形成完整的窗口结构。但该操作会破坏特征块的真实空间拓扑关系，导致重组窗口内部存在不连续的位置编码。为消除这种伪相关性的干扰，Swin Transformer引入了掩码机制（mask）来屏蔽不需要的注意力权重，这导致边缘区域的特征学习受到人为抑制。在图像恢复任务中，边缘区域往往承载着同样关键的结构信息，例如物体轮廓、纹理边界等。为此，提出采用镜像填充（padding）策略替代传统循环位移方法。即在特征图边界处进行镜像对称扩展，确保每个边缘窗口都能通过真实特征扩展形成完整窗口。该机制保留了边缘区域的真实特征分布，这有助于精确重建边缘细节。在自注意力计算后，特征图需要被裁剪（cropping）以恢复到原始大小。

1.6 特征融合模块

本文方法只在下采样阶段部署DB-Transformer块，而在上采样重建图像阶段中则使用只包含Transformer块的解码器。本文方法在上采样阶段中每个解码器前设计了一个多特征图融合模块。该模块可以通过通道注意力机制动态选择来自不同阶段和分支的特征图。因此，神经网络使用该模块学习各个阶段提取的对图像去噪任务更有用的特征。首先，对先前阶段的特征图做一次卷积核大小1 × 1、步长为1的卷积操作，即

$$\boldsymbol{F}'_{l^*} = f_{\text{Conv }1\times1}\left(\boldsymbol{F}_{l^*}\right) \tag{7}$$

式中，l^*是与第l层融合模块位于同一级的编码器所在层数。本文方法是5层网络，那么第5层特征融合模块应当接收第1层并行特征编码器的特征图。然后，使用平均池化$AP(\cdot)$计算通道级统计量$\boldsymbol{s}$，即

$$\boldsymbol{s} = AP\left(\boldsymbol{F}'_{l^*}, \boldsymbol{F}_{l-1}\right), \quad \boldsymbol{s} \in \mathbf{R}^{1\times1\times C} \tag{8}$$

式中，C与特征图通道数量一致，令$\boldsymbol{s}$经过一个多层感知器，其中包含一个卷积核大小为1 × 1、步长为1的点卷积，一个LeakyReLU激活函数和一个将通道扩大到原来的2倍的1 × 1卷积。紧接着使用softmax函数对多层感知器的输出预测权重矩阵$\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$。该过程的表达式为

$$\{\boldsymbol{a}_1, \boldsymbol{a}_2\} = f_{\text{softmax}}\left(MLP(\boldsymbol{s})\right), \quad \boldsymbol{a}_n \in \mathbf{R}^{1\times1\times C} \tag{9}$$

最后，使用这些权重对特征图计算加权和，以此抑制或增强各个阶段的特征，并与上一层的输出特征图$\boldsymbol{F}_{l-1}$进行残差连接。令$\boldsymbol{F}_{\text{fusion}}$表示特征融合模块的输出特征，那么该过程定义为

$$\boldsymbol{F}_{\text{fusion}} = \boldsymbol{a}_1\boldsymbol{F}'_{l^*} + \boldsymbol{a}_2\boldsymbol{F}_{l-1} + \boldsymbol{F}_{l-1} \tag{10}$$

1.7 训练模式

本文方法使用Charbonnier损失训练网络，具体为

$$L_{\text{Charbonnie}}(x) = \sqrt{x^2 + \varepsilon^2} \tag{11}$$

式中，x表示真实值与预测值之间的差异，ε^2是一个非常小的常数，在实验中ε设置为10^{-3}。图像去噪任务常用的损失函数中，L1损失函数计算真实值与预测值的绝对值之和，对异常值具有鲁棒性；L2损失函数是图像中每个像素真实值与预测值之间插值的平方和，具有更平滑的优点。Charbonnier损失函数

结合了 L1 损失函数和 L2 函数损失的特点。

本文采用渐进训练机制，其核心逻辑是：随着训练迭代次数的增加，逐步使用更大尺寸的图像块对网络进行训练。该机制的优势在于，更大的图像块能让 CNN 捕获更多边缘细节信息，从而有效改进模型的最终结果。从本质上看，这是一种让模型先学习简单任务、再过渡到复杂任务的科学学习模式。

2 实验结果

2.1 实验设置

本文使用由 BSD500（Berkeley segmentation dataset 500）（Arbeláez 等，2011）、DIV2K（diverse 2K resolution high quality images）（Agustsson 和 Timofte，2017）、Flickr2K（Lim 等，2017）和 WaterlooED（Waterloo exploration database）（Liang 等，2021）复合组成的数据集作为训练数据集，包括不同分辨率的图像 8 694 幅。训练和测试均在清晰图像上添加噪声水平为 σ 的高斯噪声来合成噪声图像，然后在配对的清晰和噪声图像上裁剪特定大小的补丁用于训练。为了验证方法的去噪效果，本文在 CBSD68（color Berkeley segmentation dataset）（Martin 等，2001）、Kodak24、McMaster（Zhang 等，2011）和彩色 Urban100 测试集（Huang 等，2015）上测试图像的去噪效果，并选择了多种经典或最新的去噪方法作为对比，包括传统算法的 BM3D（Dabov 等，2007）；基于 CNN 的 DnCNN（Zhang 等，2017）、FFDNet（Zhang 等，2018a）、BRDNet（Tian 等，2020）、DRUNet（Zhang 等，2022）；以及基于 Transformer 的 SwinIR（Liang 等，2021）。以上实验均使用峰值信噪比（peak signal-to-noise ratio，PSNR）作为定量指标。另外，为了衡量各个网络的训练效率，本文还测试了除 BM3D 以外全部算法的浮点运算量（floating point operations，FLOPs）指标，即提供相同的输入到初始化后的网络中，然后计算网络在一次前向传播需要执行的所有浮点运算的总数。较高的 FLOPs 指标通常意味着需要更多的计算资源，需要更长的训练和推理时间。

在具体实现中，分别设置初始通道数、自注意力窗口大小为 48 和 16，并使用 AdamW 作为训练的优化器。初始学习率设置为 0.000 3，使用分阶段逐渐缩小学习率的策略。FLOPs 测试中的输入大小设置为 $\boldsymbol{I} \in \mathbf{R}^{1\times 3\times 128\times 128}$。以上实验均使用单块 NVIDIA GeForce RTX 4070Ti Super 在 Python 3.11 和 PyTorch 2.0 框架上进行。

2.2 去噪效果和对比

表 1 对比了各方法在 CBSD68、Kodak24、McMaster 和彩色 Urban100 测试集中不同噪声强度 σ = 15, 25, 50 的去噪结果。结果显示，在噪声水平 σ = 25 和 σ = 50 的测试中，本文方法均取得了最好的 PSNR 指标，即便在噪声水平 σ = 15 的一些数据集测试中未取得最优，所得结果也是次优水平。表 2 统计了各方法的 FLOPs 以及在实验中预测去噪图像的耗时总和。在不受到其他硬件瓶颈的影响下，耗时指标通常与 FLOPs 呈正相关。该对比实验显示了

表 1 不同方法在各数据集上的去噪结果对比（PSNR）

Table 1 Comparison of denoising results of different methods on various datasets (PSNR)

/dB

方法	CBSD68			Kodak24			McMaster			Urban100		
	σ = 15	σ = 25	σ = 50	σ = 15	σ = 25	σ = 50	σ = 15	σ = 25	σ = 50	σ = 15	σ = 25	σ = 50
含噪图像	24.83	20.53	15.00	24.75	20.44	14.87	25.13	20.90	15.38	24.95	20.69	15.18
BM3D	33.52	30.71	27.38	34.28	32.15	28.46	34.06	31.66	28.51	33.93	31.36	27.93
DnCNN	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
FFDNet	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
BRDNet	34.10	31.43	28.16	34.88	32.41	29.22	35.08	32.75	29.52	34.42	31.99	28.56
DRUNet	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
SwinIR	**34.42**	31.78	28.56	35.34	32.89	29.79	**35.61**	33.20	30.22	**35.13**	32.90	29.82
本文	34.37	**31.83**	**28.61**	**35.41**	**32.95**	**29.89**	35.57	**33.36**	**30.29**	34.98	**32.93**	**29.90**

注：加粗字体表示各列最优结果。

本文方法在保持去噪效果的前提下，运行效率上有显著的优势。

图4和图5展示了两组去噪结果，原图分别来自测试集CBSD68和Urban100。对于每幅图像均放大了特定的部分来对比局部细节。在噪声强度σ = 25中，DnCNN、FFDNet、DRUNet方法的结果均呈现出较明显的边缘模糊。而SwinIR和本文方法得到了更清晰的边缘，与原图更接近。在更高级别噪声强度σ = 50的去噪结果中，SwinIR在人影部分丢失了更多细节。虽然本文方法不能还原云彩细节，但是得到了更清晰的细节和更锐利的边缘。

表2　不同方法的FLOPs和运行实验用时对比

Table 2　Comparison of FLOPs and experiment runtime for different methods

实验设置	实验用时/s	FLOPs/次
BM3D	1 922.45	–
DnCNN	24.09	110.038亿
FFDNet	24.08	20.219亿
BRDNet	33.26	–
DRUNet	39.68	359.022亿
SwinIR	888.35	1 880.340亿
本文	89.71	491.176亿

注：“–”表示无相关实验结果。

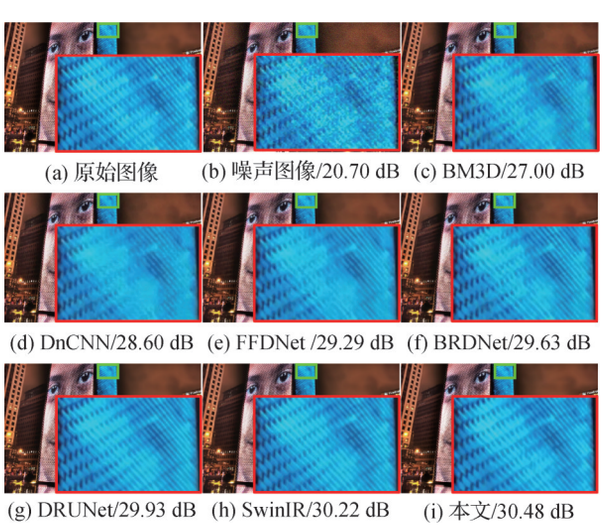

图4　不同方法在噪声等级为25时的去噪结果

Fig. 4　Denoising results of different methods at noise level 25 ((a)original image; (b)noised image 20. 70 dB; (c)BM3D/27. 00 dB; (d)DnCNN/28. 60 dB; (e)FFDNet/29. 29 dB; (f)BRDNet/29. 63 dB; (g)DRUNet/29. 93 dB; (h)SwinIR/30. 22 dB; (i)ours/30. 48 dB)

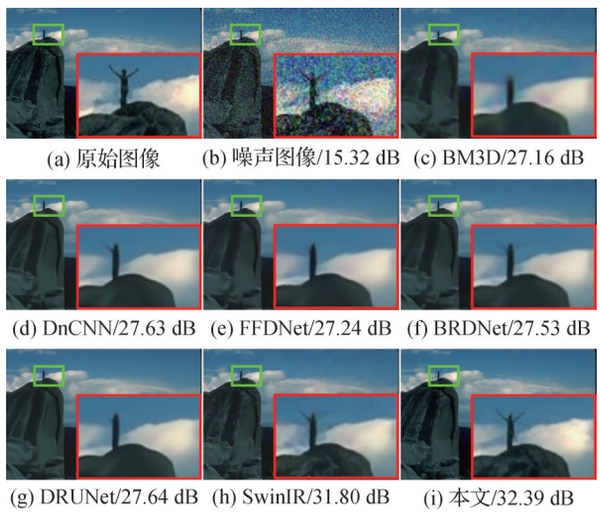

图5　不同方法在噪声等级为50时的去噪结果

Fig. 5　Denoising results of different methods at noise level 50 ((a)original image; (b)noised image 15. 32 dB; (c)BM3D/27. 16 dB; (d)DnCNN/27. 63 dB; (e)FFDNet/27. 24 dB; (f)BRDNet/27. 53 dB; (g)DRUNet/27. 64 dB; (h)SwinIR/31. 80 dB; (i)ours/32. 39 dB)

2.3　消融实验和讨论

对所提方法的机制进行多项消融实验。该部分实验使用相对小规模的网络，并且在噪声强度σ = 50的CBSD68测试集上进行。在对特定参数做消融实验时，其他实验设置均保持不变，仅修改指定参数。

2.3.1　DB-Transformer模块数量消融实验

测试并比较了下采样阶段中编码器选择不同的DB-Transformer模块数量对去噪结果的影响。令第1层到第3层的编码器DB-Transformer模块数量表示为$\{N_1, N_2, N_3\}$。其中，$\{0, 0, 0\}$表示这一项实验相当于去除了所有DB-Transformer模块。实验结果如表3所示。结果表明，在浅层编码器中部署更多的DB-Transformer模块对结果更有利。在卷积块数量达到一定程度后，继续增加DB-Transformer模块不会给去噪效果带来明显的收益。在实验中，$\{16, 12, 8\}$设置比$\{8, 6, 4\}$多出一倍DB-Transformer块，但PSNR和结构相似性(structural similarity index，SSIM)结果几乎持平，FLOPs指标则在进一步增加。因此，为了平衡计算效率和去噪效果，本文使用$\{8, 6, 4\}$为最佳设置。

2.3.2　特征融合消融实验

本文测试并比较了上采样阶段中不同的特征融合策略对结果的影响。该项实验将融合模块依次替

表3　DB-Transformer模块数量消融实验

Table 3　Ablation study on numbers of DB-Transformer blocks

实验设置	PSNR/dB	SSIM	FLOPs/亿次
{16, 12, 8}	28.20	0.753	170.99
{12, 10, 6}	28.21	0.772	148.71
{8, 6, 4}	28.21	0.769	124.85
{4, 6, 8}	28.17	0.748	124.64
{4, 4, 2}	28.12	0.752	118.43
{2, 1, 1}	28.08	0.741	115.19
{0, 0, 0}	28.06	0.730	110.45

换为：拼接卷积、作残差连接共两种策略。实验结果如表4所示。结果显示，本文使用的融合模块在去噪结果上要优于其他设置，这验证了动态学习特征图权重的可行性。融合机制提高模型的特征提取能力，比简单地拼接卷积或直接相加都取得了更好的去噪效果。

表4　特征融合消融实验

Table 4　Ablation study on feature fusion

实验设置	PSNR/dB	SSIM
融合模块	28.21	0.769
拼接卷积	28.16	0.761
残差连接	28.05	0.740

2.3.3　P-Swin Transformer模块消融实验

本文对原始Swin Transformer V2块做了一定改进，为了验证这些改进对图像去噪的提升效果，本文调整归一化位置与镜像填充的设置，做了多项实验。实验结果如表5所示。图6展示了不同层归一化位置的训练过程结果对比。

表5　P-Swin Transformer模块消融实验

Table 5　Ablation study on P-Swin Transformer module

归一化位置	镜像填充	PSNR/dB	SSIM
自注意力前	√	28.21	0.769
自注意力后	√	28.06	0.738
自注意力前	–	28.13	0.757

注："√"表示使用，"–"表示不使用。

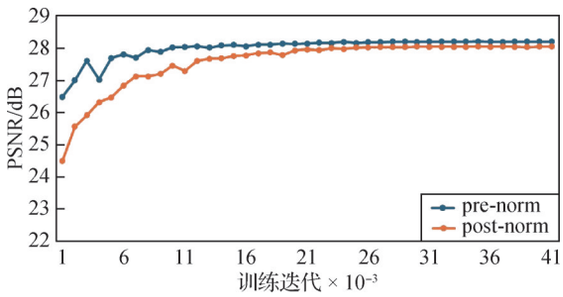

图6　不同层归一化位置的训练过程比较

Fig. 6　Comparison of training processes with different normalization positions

3　结　论

本文提出一个Swin Transformer V2和特征融合的U-Net深度学习网络用于去除图像噪声。由于局部卷积提取的特征对图像去噪任务仍然有重要作用，将卷积和Transformer网络相结合，设计了Transformer特征和局部卷积特征的双分支机制，从两种结构的特征中学习对图像去噪更有用的特征信息。除此之外，还对Swin Transformer V2 block做出了适应性改进，使得模型收敛更为快速、去噪效果更好。

设计的去噪网络相较于其他主流的基于CNN的方法取得了更好的去噪效果。而相比较于优秀的基于Transformer的SwinIR，本文方法在部分去噪测试集上优于或接近SwinIR的去噪效果，并且本文方法所需的浮点运算量远远低于SwinIR所需。该模型生成的去噪图像的视觉效果有了很大的改善，在图像结构内容和边缘细节方面都取得了较好的还原质量。理论分析和实验结果表明，本文提出的图像去噪算法在提升去噪效果的同时计算代价更小。如何更高效充分地利用两者的特征提取能力是今后主要研究方向。

参考文献（References）

Agustsson E and Timofte R. 2017. NTIRE 2017 challenge on single image super-resolution: dataset and study//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE: 1122-1131 [DOI: 10.1109/CVPRW.2017.150]

Arbeláez P, Maire M, Fowlkes C and Malik J. 2011. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 33(5): 898-916 [DOI: 10.1109/TPAMI.2010.161]

Brempong E A, Kornblith S, Chen T, Parmar N, Minderer M and Norouzi M. 2022. Denoising pretraining for semantic segmentation// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, USA: IEEE: 4174-4185 [DOI: 10.1109/CVPRW56347.2022.00462]

Buades A, Coll B and Morel J M. 2005. A non-local algorithm for image denoising//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, USA: IEEE: 60-65 [DOI: 10.1109/CVPR.2005.38]

Chen H T, Wang Y H, Guo T Y, Xu C, Deng Y P, Liu Z H, Ma S W, Xu C J, Xu C and Gao W. 2021. Pre-trained image processing transformer//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 12294-12305 [DOI: 10.1109/CVPR46437.2021.01212]

Cheng S, Wang Y Z, Huang H B, Liu D H, Fan H Q and Liu S C. 2021. NBNet: noise basis learning for image denoising with subspace projection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 4894-4904 [DOI: 10.1109/CVPR46437.2021.00486]

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

Ding Y H, Wu H, Kong F L, Xu D and Yuan G W. 2023. A dual of real image noise-related blind denoising technique. Journal of Image and Graphics, 28(7): 2026-2036 (丁岳皓, 吴昊, 孔凤玲, 徐丹, 袁国武. 2023. 面向真实图像噪声的两阶段盲去噪. 中国图象图形学报, 28(7): 2026-2036) [DOI: 10.11834/jig.211020]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR

Huang J B, Singh A and Ahuja N. 2015. Single image super-resolution from transformed self-exemplars//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 5197-5206 [DOI: 10.1109/CVPR. 2015.7299156]

Kivanc Mihcak M, Kozintsev I, Ramchandran K and Moulin P. 1999. Low-complexity image denoising based on statistical modeling of wavelet coefficients. IEEE Signal Processing Letters, 6(12): 300-303 [DOI: 10.1109/97.803428]

Liang J Y, Cao J Z, Sun G L, Zhang K, Van Gool L and Timofte R. 2021. SwinIR: image restoration using swin transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE: 1833-1844 [DOI: 10.1109/ICCVW54120.2021.00210]

Lim B, Son S, Kim H, Nah S and Lee K M. 2017. Enhanced deep residual networks for single image super-resolution//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE: 1132-1140 [DOI: 10.1109/CVPRW.2017.151]

Lin H, Ma Z H, Ji R R, Wang Y W and Hong X P. 2022. Boosting crowd counting via multifaceted attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 19596-19605 [DOI: 10.1109/CVPR52688.2022.01901]

Liu Z, Hu H, Lin Y T, Yao Z L, Xie Z D, Wei Y X, Ning J, Cao Y, Zhang Z, Dong L, Wei F R and Guo B N. 2022. Swin transformer V2: scaling up capacity and resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 11999-12009 [DOI: 10.1109/CVPR52688.2022.01170]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Martin D, Fowlkes C, Tal D and Malik J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics//Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver, Canada: IEEE: 416-423 [DOI: 10.1109/ICCV. 2001.937655]

Mei Y Q, Fan Y C and Zhou Y Q. 2021. Image super-resolution with non-local sparse attention//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 3516-3525 [DOI: 10.1109/CVPR46437.2021.00352]

Mohd Sagheer S V and George S N. 2020. A review on medical image denoising algorithms. Biomedical Signal Processing and Control, 61: #102036 [DOI: 10.1016/j.bspc.2020.102036]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Tian C W, Xu Y and Zuo W M. 2020. Image denoising using deep CNN with batch renormalization. Neural Networks, 121: 461-473 [DOI: 10.1016/j.neunet.2019.08.022]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-

6010

Wang Z D, Cun X D, Bao J M, Zhou W G, Liu J Z and Li H Q. 2022. Uformer: a general u-shaped transformer for image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 17662-17672 [DOI: 10.1109/CVPR52688.2022.01716]

Wu C Z, Chen X, Ji D and Zhan S. 2018. Image denoising via residual network based on perceptual loss. Journal of Image and Graphics, 23(10): 1483-1491 (吴从中, 陈曦, 季栋, 詹曙. 2018. 结合深度残差学习和感知损失的图像去噪. 中国图象图形学报, 23(10): 1483-1491) [DOI: 10.11834/jig.180069]

Yang F Z, Yang H, Fu J L, Lu H T and Guo B N. 2020. Learning texture transformer network for image super-resolution//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 5790-5799 [DOI: 10.1109/CVPR42600.2020.00583]

Yin H T and Ma S Y. 2022. CSformer: cross-scale features fusion based transformer for image denoising. IEEE Signal Processing Letters, 29: 1809-1813 [DOI: 10.1109/LSP.2022.3199145]

Yuan L, Chen Y P, Wang T, Yu W H, Shi Y J, Jiang Z H, Tay F E H, Feng J S and Yan S C. 2021. Tokens-to-token ViT: training vision transformers from scratch on ImageNet//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 538-547 [DOI: 10.1109/ICCV48922.2021.00060]

Zhang F, Chen G G, Wang H, Li J J and Zhang C M. 2023. Multi-scale video super-resolution transformer with polynomial approximation. IEEE Transactions on Circuits and Systems for Video Technology, 33(9): 4496-4506 [DOI: 10.1109/TCSVT.2023.3278131]

Zhang K, Li Y W, Zuo W M, Zhang L, Van Gool L and Timofte R. 2022. Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 6360-6376 [DOI: 10.1109/TPAMI.2021.3088914]

Zhang K, Zuo W M, Chen Y J, Meng D Y and Zhang L. 2017. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 26(7): 3142-3155 [DOI: 10.1109/TIP.2017.2662206]

Zhang K, Zuo W M and Zhang L. 2018a. FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Transactions on Image Processing, 27(9): 4608-4622 [DOI: 10.1109/TIP.2018.2839891]

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

Zhang Y L, Li K P, Li K, Wang L C, Zhong B N and Fu Y. 2018b. Image super-resolution using very deep residual channel attention networks//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 294-310 [DOI: 10.1007/978-3-030-01234-2_18]

Zhang Y L, Tian Y P, Kong Y, Zhong B N and Fu Y. 2018c. Residual dense network for image super-resolution//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2472-2481 [DOI: 10.1109/CVPR.2018.00262]

Zuo J H, Jia Z H, Yang J and Kasabov N. 2019. Moving target detection based on improved Gaussian mixture background subtraction in video images. IEEE Access, 7: 152612-152623 [DOI: 10.1109/ACCESS.2019.2946230]

作者简介

利铭康，男，硕士研究生，主要研究方向为深度学习图像处理。E-mail: 2022023211@m.scnu.edu.cn

柳薇，通信作者，女，副教授，主要研究方向为多媒体信息处理和机器学习。E-mail: liuwei@m.scnu.edu.cn

陈卫东，男，教授，主要研究方向为图论与复杂网络、组合优化、算法与计算复杂性。E-mail: chenwd@scnu.edu.cn