CT理论与应用研究
Computerized Tomography Theory and Applications

基于Transformer增强型U-net的CT图像稀疏重建与伪影抑制

樊雪林，文昱齐，乔志伟

Sparse Reconstruction of CT Images with Transformer Enhanced U-net

FAN Xuelin, WEN Yuqi, and QIAO Zhiwei

在线阅读 View online: https://doi.org/10.15953/j.ctta.2023.183

樊雪林, 文昱齐, 乔志伟. 基于 Transformer 增强型 U-net 的 CT 图像稀疏重建与伪影抑制[J]. CT 理论与应用研究（中英文）, 2024, 33(1): 1-12. DOI:10.15953/j.ctta.2023.183.
FAN X L, WEN Y Q, QIAO Z W. Sparse Reconstruction of Computed Tomography Images with Transformer Enhanced U-net[J]. CT Theory and Applications, 2024, 33(1): 1-12. DOI:10.15953/j.ctta.2023.183. (in Chinese).

基于 Transformer 增强型 U-net 的 CT 图像稀疏重建与伪影抑制

樊雪林[1], 文昱齐[2], 乔志伟[1]✉

1. 山西大学计算机与信息技术学院, 太原 030006
2. 北京理工大学材料学院, 北京 102401

摘要：实现低剂量计算机断层成像（CT）的一个有效办法是减少投影角度，但投影角度较少会产生严重的条状伪影，降低图像的临床使用价值。针对该问题，提出一种耦合卷积神经网络（CNN）和多种注意力机制的 U 型网络（TE-unet）。首先采用 U 型架构提取多尺度特征信息；其次提出一个包含 CNN 和多种注意力的模块提取图像特征；最后在跳跃连接处加入 Transformer 块过滤信息，抑制不相关特征，突出重要特征。所提网络结合 CNN 的局部特征提取能力和 Transformer 的全局信息捕获能力，辅以多种注意力机制，实现了良好的去条状伪影能力。在 60 个投影角度下，与经典的 Uformer 网络相比，峰值信噪比（PSNR）高出 0.3178dB，结构相似度（SSIM）高出 0.002，均方根误差（RMSE）降低 0.0005。实验结果表明，所提 TE-unet 重建的图像精度更高，图像细节保留的更好，可以更好地压制条状伪影。

关键词：稀疏重建；计算机断层成像；Transformer；多注意力机制；条状伪影

DOI:10.15953/j.ctta.2023.183 **中图分类号：**O242；TP391 **文献标识码：**A

计算机断层成像（computed tomography，CT）[1]是当前应用最为广泛的医学成像模态。然而，过高的 X 射线剂量会对人体造成损害。因此，为了在减少对人体损害的同时满足临床诊断的需求，低剂量 CT 成为了研究的一个重点。当前，低剂量 CT 有两种实现方法，一种是降低每个投影角度下的辐射剂量，另一种是在保持每个投影角度下辐射剂量不变的前提下，减少投影角度个数。在稀疏角度下，减少了图像数据的采集和传输量，从而降低了辐射剂量。然而由于投影角度的不足，使用传统解析法稀疏重建的图像会产生严重的条状伪影，降低了图像的可用性。因此，稀疏重建对临床医学诊断有着非常重要的意义。

当前 CT 图像稀疏重建算法主要有两种，一种是以压缩感知（compressed sensing，CS）[2]为基础的迭代重建算法。迭代法重建精度高，但存在迭代时间长、速度慢和计算成本高等不足。自 2006 年以来，Sidky 等[3-4]提出了扇束和锥束 CT 总变差（total variation，TV）最小化算法，实现了高精度的 CT 稀疏重建。随后，学者们在此基础上提出了自适应加权 TV（adaptive-weighted total variation，AwTV）[5]、保边 TV（edge-preserving TV，EPTV）[6]和高阶 TV（high order TV，HOTV）[7]等算法，极大地推动了迭代法的发展。

另一种方法是深度学习法。近年来，基于深度学习的方法在图像恢复任务上取得了杰出的效果。卷积神经网络（convolutional neural network，CNN）由于局部感知、权重共享等优点占据主导地位多年。2017 年 Chen 等[8]提出的（residual encoder-decoder convolutional neural network）网络，将残差连接运用于编码器和解码器之间，在稀疏重建方面取得了不错的效果。

Zhang 等[9]提出的 DnCNN（denoising convolutional neural network）网络强调了残差学习和

收稿日期：2023-09-26。
基金项目：国家自然科学基金面上项目（模型与数据耦合驱动的快速四维 EPRI 肿瘤氧成像（62071281））；中央引导地方科技发展资金项目（新型 TV 和学习先验联合约束的快速四维 EPRI 成像方法（YDZJSX2021A003））；山西省回国留学人员科研资助项目（基于新型四维 TV 正则机理的快速 EPRI 肿瘤氧成像方法研究（2020-008））。

Batch Normalization 在图像复原中的作用，在较深网络条件下，依然可以较快的收敛并取得良好的性能。Jin 等[10]提出的FBPConvNet 将传统的滤波反投影（filtered back Projection，FBP）算法与残差 U-net 结合起来，可以很好地压制条状伪影。Wolterink 等[11]将生成对抗网络（generative adversarial network，GAN）应用于低剂量 CT 图像重建任务，取得了良好的效果。2018 年，Oktay 等[12]提出的Attention U-Net 在跳跃连接处加入 Attention 对信息进行过滤，使得效果得到了进一步提升。2022 年，Chen 等[13]提出的 NAFnet（nonlinear activ-ation free network）为图像恢复任务提出了一个由 CNN 和通道注意力组成的基线（Baseline），同时通过 SimpleGate 替换激活函数，获得了性能的提升。

虽然 CNN 在图像恢复领域已经取得了令人瞩目的成果，但是 CNN 对长程依赖建模的效果并不是很理想。Transformer 通过自注意力的方式捕获全局信息可以很好的解决以上问题。

2017 年，Google 团队首先在自然语言处理（natural language pro-cessing，NLP）中提出了 Transformer[14]，通过自注意力机制，缩短了训练时间，大幅提升了机器翻译的性能。2020 年，Gulat 等[15]提出的Conformer 将 Transformer 结构中的前向反馈层替换为两个半步的前向反馈层，以提高网络的性能。同年 Google 团队提出 Vision Transformer（ViT）[16]，该网络模型首次将 Transformer 应用于计算机视觉领域中的图像分类任务，通过将图片划分为更小的图像块，然后将小图像块的线性序列作为输入进行训练，取得了很好地效果。ViT 的开创性工作表明，纯粹的基于 Transformer 的架构也可以取得很好地结果。2021 年 Liu 等[17]提出Swin Transformer 网络架构，通过对图片进行划分窗口，将注意力的计算限制在一个窗口中，然后利用滑动窗口的操作实现与窗口外像素注意力的计算，在减少了计算量的同时实现了很好地效果。Liang 等[18]提出的 SwinIR 将 Transformer 应用于图像恢复任务，在浅层阶段采用卷积块，随后在深度特征提取方面使用 Transformer，取得了很好地效果。图像恢复任务通常依赖每个阶段的特征来获得更好的结果，因此在保持较低的计算成本的同时，有效的实现大接受域是非常重要的。2021 年 Wang 等[19]提出的 Uformer 网络结构包括具有局部增强能力的 Transformer 模块，很好地提取了局部信息，同时使用跳跃连接机制将编码器的信息传递到解码器，取得了良好的图像恢复效果。

但是在计算机视觉中，Transformer 的计算复杂度与图像分辨率息息相关。所以在医学图像任务中 Transformer 面临计算复杂度高的难题，同时 Transformer 有着提取细粒度局部特征能力较弱的缺点。针对此问题，本文在引入 Transformer 的同时，通过窗注意力的方式减少计算复杂度，通过耦合 CNN 和深度卷积弥补其在局部特征提取能力的不足。

综上所述，本文的主要工作如下：

（1）提出一个包含 CNN、Transformer 和多种注意力[20]的U 型网络（transformer enhanced U-net，TE-unet），结合 CNN 的局部特征提取能力和 Transformer 的全局特征提取能力，并加入残差连接、特征融合和多种注意力机制，弥补传统 CNN 和 Transformer 在处理条状伪影时的一些不足，取得了良好的去条状伪影效果。

（2）设计一个耦合 CNN、深度卷积、通道注意力和空间注意力的 CCA 块（CNN Coupled Attention Block）。

（3）将编码器传递的信息与完成上采样的信息拼接后，引入 Transformer 块处理拼接后的信息。利用 Transformer 可以很好地建模长程依赖的优点，对信息进行过滤，抑制无关信息的同时突出重要特征。

1 本文方法

去条状伪影的本质是要从低质量图片中分离出条状伪影，保留有用信息。本文假定 $\boldsymbol{I}_{\mathrm{r}}$ 表示恢复的高质量图片，$\boldsymbol{I}$ 表示含条状伪影的图片，$\boldsymbol{A}$ 表示条状伪影。那么 $\boldsymbol{I}_{\mathrm{r}}$ 和 $\boldsymbol{I}$ 的关系可以表示为：

$$\boldsymbol{I}_{\mathrm{r}}=\boldsymbol{I}-\boldsymbol{A}。\tag{1}$$

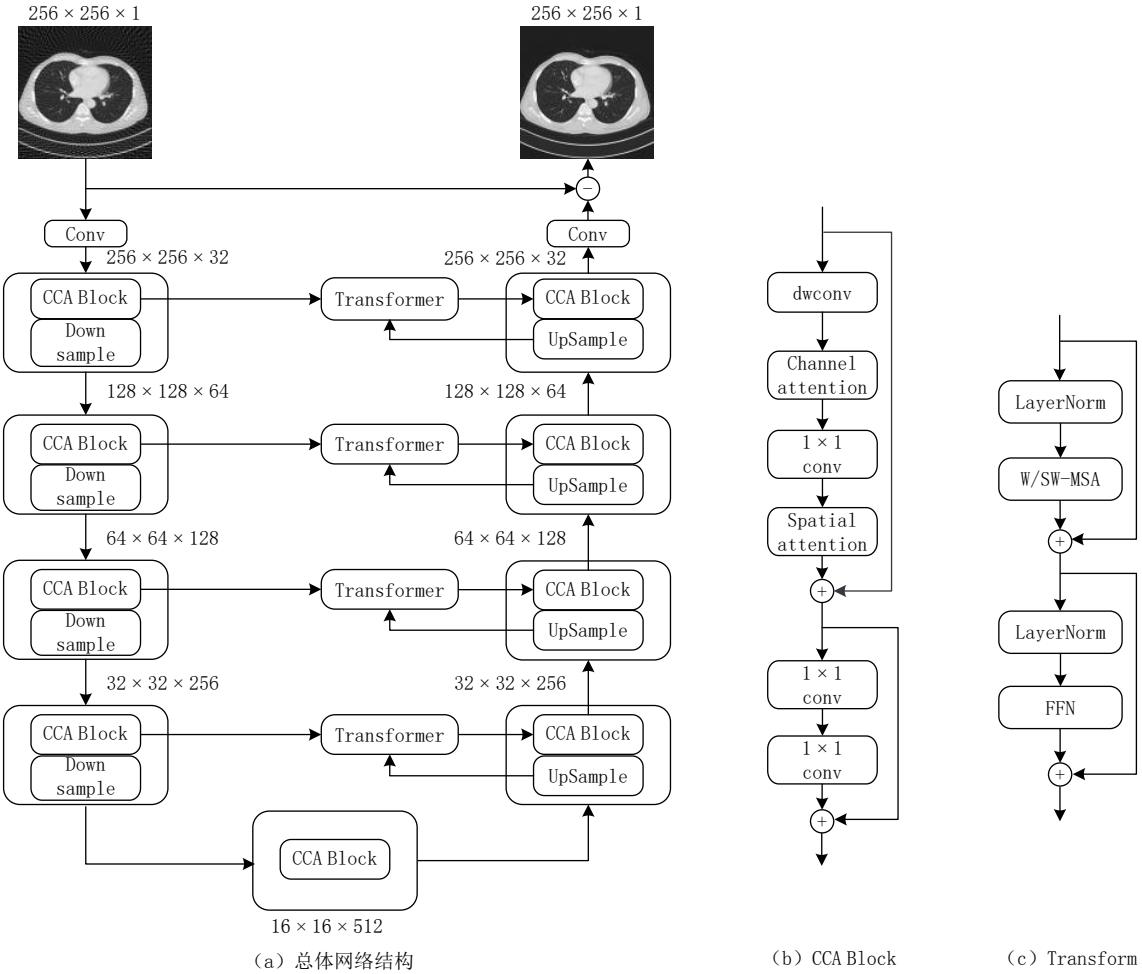
当前大多数深度学习网络通过设计不同的网络结构直接学习 $\boldsymbol{I}$ 到 $\boldsymbol{I}_{\mathrm{r}}$ 之间的复杂映射：$\boldsymbol{I}_{\mathrm{r}}=f_1(\boldsymbol{I})$。然而，研究表明，通过学习含条状伪影图像 $\boldsymbol{I}$ 与条状伪影 $\boldsymbol{A}$ 之间的映射关系，可以得到更好的效果[21]。$\boldsymbol{I}_{\mathrm{r}}$ 和 $\boldsymbol{A}$ 之间的关系可以表示为：

$$\boldsymbol{A}=f_2(\boldsymbol{I}),\tag{2}$$

$$\boldsymbol{I}_{\mathrm{r}}=\boldsymbol{I}-\boldsymbol{A}=\boldsymbol{I}-f_2(\boldsymbol{I})。\tag{3}$$

1.1　网络整体结构

本文提出的 TE-unet 总体结构如图 1（a）所示。网络分为特征提取的编码器阶段和特征融合的解码器阶段。



（a）总体网络结构　　（b）CCA Block　　（c）Transform

图 1　TE-unet 的网络结构

Fig. 1　Network structure of TE-unet

具体过程为：给定含条状伪影图像 $\boldsymbol{I}\in R^{H\times W\times 1}$，经过3 × 3 卷积提取浅层特征为 $\boldsymbol{F}_{\mathrm{s}}\in R^{H\times W\times C}$，其中 $H\times W$ 是图像的大小，C 是通道数。随后 $\boldsymbol{F}_{\mathrm{s}}$ 将通过第一层编码器，编码器包含了若干个 CCA 块和下采样。CCA 块首先利用深度卷积增强局部信息，再利用通道自注意力和空间注意力捕获全局信息。最后使用步长为 2 的 4 × 4 卷积核进行下采样，得到一级编码器的输出特征 $\boldsymbol{F}_1\in R^{H/2\times W/2\times 2C}$。经过 4 层编码器提取特征后得到深层特征 $\boldsymbol{F}_{\mathrm{d}}\in R^{H/16\times W/16\times 16C}$，随后经过中间转换层处理后，进入到 4 层解码器中，每一层解码器包含上采样层和多个 CCA 块。使用步长为 2 的 2 × 2 转置卷积进行上采样，与对

应编码器传递的信息在通道维度拼接后得到特征 $\boldsymbol{F}_1' \in \boldsymbol{R}^{H/8\times W/8\times 16C}$，随后将 $\boldsymbol{F}_1'$ 输入到 Transformer 块中进行处理，包括层归一化（LayerNorm）、窗注意力（W/SW-MSA）、和前向反馈网络（FFN）。突出对图像恢复有帮助的特征，抑制不相关的特征。然后再通过若干 CCA 块恢复图像特征。在通过 4 级 Transformer 块和解码器的处理后，再利用一个 3×3 卷积得到条状伪影 $\boldsymbol{A} \in \boldsymbol{R}^{H\times W\times 1}$。最终得到干净图像 $\boldsymbol{I}_r = \boldsymbol{I} - \boldsymbol{A}$。

1.2 CCA 块

CCA 块结构如图 1（b）所示。对于退化图像来说，受损像素点的邻域像素可以被用来恢复图像，所以局部上下文信息在图像恢复中占据着非常重要的地位。因此，本文在 CCA 块中首先加入深度卷积提取局部特征。深度可分离卷积分为两部分，如图 2 所示。

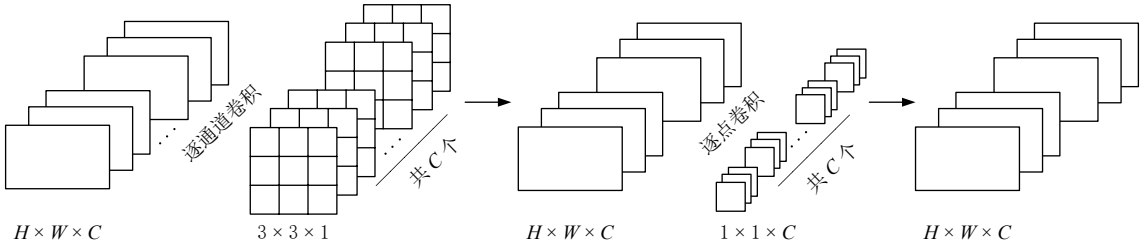


图 2 深度可分离卷积结构

Fig.2 Depthwise separable convolution structure

首先是逐通道进行 3×3 卷积，对输出进行拼接后使用逐点卷积（pointwise convolution）得到特征图 $\boldsymbol{x}'$。通过拆分空间维度和通道维度的相关性，减少卷积计算所需要的参数个数，提高卷积核参数的使用效率。然后通过一个简化的通道注意力强调不同通道间的重要性。简化的通道注意力主要分为压缩与恢复两部分。压缩部分使用全局平均池化对 $H\times W\times C$ 的特征图进行压缩得到 $1\times 1\times C$ 的特征图。恢复部分对压缩后的特征图进行 1×1 的卷积，得到通道权重信息，用于对特征图进行加权。随后通过空间注意力得到特征图 $\boldsymbol{x}''$，空间注意力通过关注不同的空间位置来捕获重要区域的特征，使得网络关注更加重要的区域。最后，通过两个 1×1 卷积得到 CCA 块的输出为 $\boldsymbol{x}'''$。两个 1×1 卷积通过升维和降维，实现跨通道的信息交互。CCA 块通过几种机制的综合运用，提高了网络学习能力和泛化能力。

CCA 块计算公式如下：

$$\boldsymbol{x}' = \mathrm{dwconv}(\boldsymbol{x}), \tag{4}$$

$$\boldsymbol{x}'' = \mathrm{SA}\left(\mathrm{Conv}\left(\mathrm{CA}(\boldsymbol{x}')\right)\right) + x, \tag{5}$$

$$\boldsymbol{x}''' = \mathrm{Conv}(\boldsymbol{x}'') + \boldsymbol{x}'', \tag{6}$$

其中，dwconv 表示深度卷积，SA 表示空间注意力，CA 表示通道注意力，Conv 表示卷积操作。

1.3 Transformer Block

考虑到编码器传递的信息与上采样之后的信息中存在一些与图像恢复任务不相关的信息，本文将两者进行拼接后，送入 Transformer 块进行处理。利用其中的多头机制将特征表示映射到不同的特征子空间，增强模型的表达能力。通过自注意力机制捕获全局信息，对需要重点关注的区域投入更多资源的同时抑制其他区域的信息，以获取更重要的信息。

Transformer Block 的结构如图 1（c）所示，包括层归一化、多头自注意力和前馈层。由于 Transformer 中的全局自注意力机制与图像的分辨率呈二次方关系，导致在高分辨率图片中计算自注意力会带来巨大的计算复杂度。因此，本文将输入划分为 $M\times M$ 大小的不重叠的局部窗口，窗口

总数为 HW/M^2，则输入转化为 $(HW/M^2)\times M^2\times C$。在窗口内进行自注意力 Attention 的计算（W-MSA），对于窗口特征 $\boldsymbol{X}\in R^{M^2\times C}$ 对应的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵计算公式为：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{Q}}\\ \boldsymbol{K}=\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{K}}\\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}}\end{cases},\tag{7}$$

其中，$\boldsymbol{P}_{\boldsymbol{Q}}$、$\boldsymbol{P}_{\boldsymbol{K}}$、$\boldsymbol{P}_{\boldsymbol{V}}$ 是跨不同窗口的投影矩阵，且 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}\in R^{M^2\times d}$，其中 d 为 $\boldsymbol{Q}/\boldsymbol{K}$ 的维度。则自注意力矩阵在对应窗口内的计算公式为：

$$\text{Attention}=\text{soft max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}+\boldsymbol{B}\right)\boldsymbol{V},\tag{8}$$

其中 $\boldsymbol{B}$ 是可学习的相对位置编码。但是划分窗口会限制窗内外的信息交互，因此本文加入滑动窗口的操作实现窗内外元素的信息交互。

传统前向反馈层为两层全连接层，并在其中加入激活函数。但是，这样利用局部上下文信息的能力有限。因此，本文将前向反馈层改为由卷积实现：即两个卷积块加深度卷积。第一个 1×1 卷积用于将通道扩大 4 倍，随后通过卷积核为 3×3 大小的逐通道卷积和 GELU（gaussian error linear unit）激活函数，最后用一个 1×1 的卷积恢复通道数，融合通道信息。通过卷积提取图像的局部特征，可以更好地利用局部上下文信息，使得恢复图像细节信息更好。

Transformer 块通过自注意力机制对传递的信息进行处理和提取，自适应的调整不同特征的权重，抑制不相关的特征，突出重要特征。

Transformer 块的计算公式如下：

$$t'=\text{W/SW}-\text{MSA}\left(\text{LayerNorm}(\boldsymbol{t})\right)+\boldsymbol{t},\tag{9}$$

$$\boldsymbol{t}''=\text{FFN}\left(\text{LayerNorm}(t')\right)+t',\tag{10}$$

其中，$\text{W/SW}-\text{MSA}$ 表示窗多头自注意力，LayerNorm 表示层归一化，FFN 表示前馈层。

2　实验结果分析

2.1　数据集创建

实验所用数据来自 TCIA 数据集（https://www.cancerima-gingarchive.net）。本文从中选取了 5600 张 256 × 256 图像，包括头部、胸部和腹部的完备投影角度下的高精度图像。对高精度图像进行 Radon 变换得到其稀疏投影数据，再用 FBP 稀疏重建得到对应的含条状伪影图像。

本文从中选取 5000 对图像作为训练集，300 对图像作为验证集，300 对图像作为测试集。

2.2　网络训练超参数设定和实验平台

实验配置 CPU 是 Inter(R) Xeon(R) CPU E5-2620v4 @ 2.10 GHz，GPU 是 NVIDIA Geforce GTX 3090，使用 Pytorch 库，在 Python 上进行训练。初始学习率 $\text{lr}=4.5\times10^{-4}$，batch size = 8，epoch = 100。

2.3　评价指标

本文采用峰值信噪比（peak signal to noise ratio，PSNR）、结构相似性（structural similarity，SSIM）、均方根误差（root mean square error，RMSE）、参数量和训练时长这 5 个指标来评价网络性能。其中，PSNR 是衡量重建图像质量的一个重要指标；SSIM 是一种衡量两幅图像相似度的指标；RMSE 则衡量两幅图像之间的偏差；参数量可以在一定程度上衡量网络大小；训练时长可以衡量网络训练速度。公式如下：

$$\mathrm{PNSR}(\boldsymbol{x},\boldsymbol{y})=10\times\lg\left\{\frac{\mathrm{MAX}^2}{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_{i,j}-y_{i,j}\right)^2}\right\},\tag{11}$$

$$\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y})=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)},\tag{12}$$

$$\mathrm{RMSE}(\boldsymbol{x},\boldsymbol{y})=\sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_{i,j}-y_{i,j}\right)^2},\tag{13}$$

其中，MAX 表示图像的最大灰度值，$\boldsymbol{x}$ 表示经过网络训练得到的图像，$\boldsymbol{y}$ 表示高质量图像；μ_x 表示 $\boldsymbol{x}$ 的均值，μ_y 表示 $\boldsymbol{y}$ 的均值，σ_x^2 表示 x 的方差，σ_y^2 表示 y 的方差，σ_{xy} 表示 $\boldsymbol{xy}$ 的协方差，$\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$ 是常数。

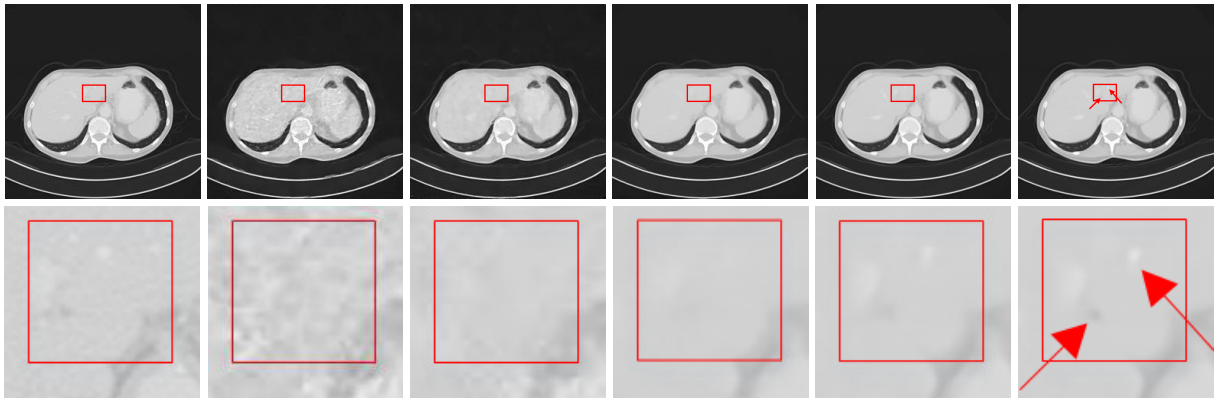
2.4 实验结果分析

在该实验中，输入是 5000 对［0,π］范围内等间隔采集 60 个角度下的投影数据进行稀疏重建的含条状伪影 CT 图像和对应的完备投影角度下的清晰图像。

2.4.1 不同重建算法对比分析实验

本文选取 DnCNN[9]、RED-CNN[8]、SwinIR[18]、Uformer[19]4 个经典网络进行对比实验，同时使用 PSNR、SSIM 和 RMSE 评估算法的去条状伪影能力和重建的稀疏 CT 图像的质量。

从测试集中随机的挑选一张腹部图片以展示不同算法的实验结果。由图 3 可见，DnCNN 重建图像存在的条状伪影比较明显；RED-CNN 重建图像仍存在肉眼可见条状伪影；SwinIR 重建图像通过肉眼已经很难观察到条状伪影的存在，但某些图像细节恢复的不够好；Uformer 重建图像已可以恢复部分局部组织的结构，细节信息保留的更多；TE-unet 重建出图像局部组织的保留最多，细节恢复的更多，在上述网络中达到最好的效果。图 4 为图 3 的伪彩色显示，从图 4 中也可以看出 TE-unet 重建出的图像效果更好。



（a）高清图 （b）DnCNN （c）RED-CNN （d）SwinIR （e）Uformer （f）TE-unet

图 3 不同算法的腹部图像实验结果图及局部放大图

Fig.3 Experimental results and local enlarged images of abdominal images using different algorithms

由表 1 可知，本文所提 TE-unet 在 PSNR、SSIM 和 RMSE 等多个指标上都优于其他模型。其中，PSNR 比 Uformer 要高出 0.3178 dB，SSIM 高出 0.002，RMSE 降低 0.0005。结果表明该模型可以在保留更多图像细节的同时有效去除条状伪影。

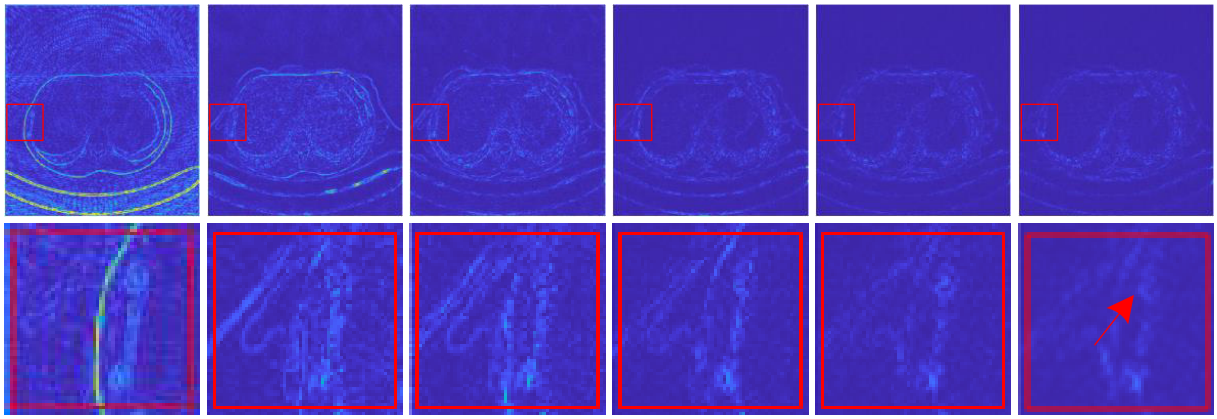


图 4　不同算法的腹部图像实验结果残差图及局部放大图（伪彩色显示，窗口为 [0,128]）

Fig.4　Residual and local enlarged images of abdominal image experimental results using different algorithms (Pseudo color display with a window of [0,128])

表 1　不同算法下测试集的实验结果

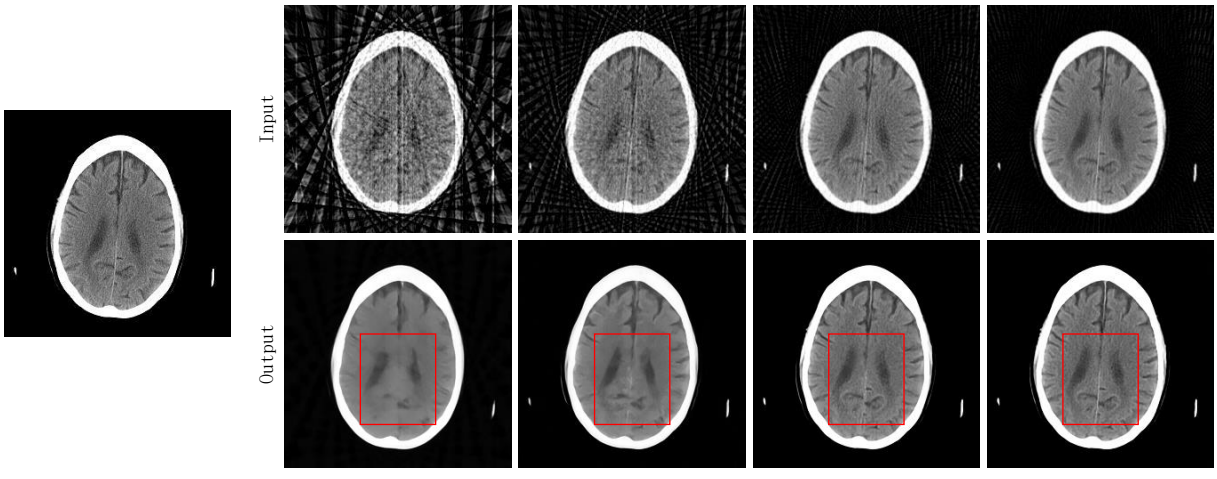
Table 1　Experimental results of different algorithms

网络	PSNR/dB	SSIM	RMSE	参数量/M	训练时长/h
DnCNN	29.0412	0.8161	0.0359	6.73	2.07
RED-CNN	31.9144	0.8813	0.0255	10.41	8.05
SwinIR	35.1524	0.9427	0.0176	27.28	11.28
Uformer	38.2551	0.9563	0.0124	20.75	13.12
TE-unet	38.5729	0.9583	0.0119	53.55	14.13

2.4.2　不同稀疏角度下的实验结果

为了探索 TE-unet 在不同稀疏角度下恢复图像的能力，本文分别以 15、30、60 和 90 个稀疏角度下的含条状伪影图作为输入进行训练，最后对实验结果进行分析。

在测试集中随机挑选一张头部图片以展示网络在不同稀疏角度下的重建性能。如图 5 可见，在 15 个稀疏角度下恢复的图像质量很差，仍然可见明显的条状伪影。当稀疏角度为 30 时，可以恢复部分细节信息。当稀疏角度为 60 时，恢复的图像质量显著提高，细节信息保留的更多。当稀疏角度为 90 时，图像的结构与纹理信息大部分被保留下来，得到的图像质量最高。



（a）高精度　（b）15 角度　（c）30 角度　（d）60 角度　（e）90 角度

图 5　不同稀疏角度下头部图像实验结果可视化

Fig.5　Experimental results under different sparse angles

表 2 中指标均为测试集中所得结果的均值。通过对比这些数值可知，当稀疏角度为 15 时，各个指标均为最低。与 30 个角度下的重建图相比，PSNR 低 3.5545 dB，SSIM 低 0.0493，RMSE 高 0.009；与 60 个角度下的重建图相比，PSNR 低 6.9832 dB，SSIM 低 0.0769，RMSE 高 0.0148；与 90 个角度下相比 PSNR 低 7.8405 dB，SSIM 低 0.0827，RMSE 高 0.0159。实验结果表明，当稀疏角度增大时，网络恢复的图像效果在逐渐变好。

表 2 不同稀疏角度下测试集的实验结果
Table 2 Experimental results of test sets under different sparse angles

稀疏角度	PSNR/dB	SSIM	RMSE
15	31.8786	0.8954	0.0258
30	35.5670	0.9341	0.0169
60	38.5729	0.9583	0.0119
90	39.6324	0.9650	0.0106

2.5 网络内部规律探索

本小节讨论 TE-unet 内部机制对去噪能力的影响。在其他参数一致的情况下，使用 PSNR、SSIM 和 RMSE 3 个指标来评估去噪效果。

2.5.1 Transformer 块的不同个数

如图 6（a），将上采样后的信息直接输入编码器，而不经过 Transformer 块处理，也就是 Transformer 在跳跃连接处只处理编码器传递过来的信息的网络记为 Only-Skip。如图 6（b），在编码器与解码器的每一个阶段都加入 Transformer 块，并将网络标记为 TE-unet^{+}。

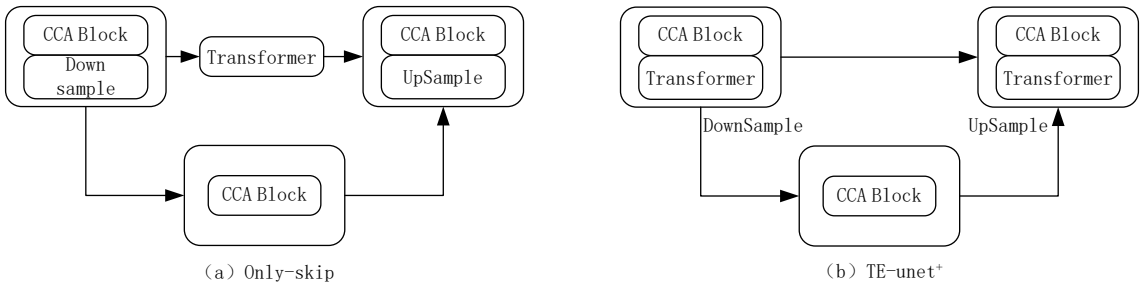


图 6 不同 Transformer 块的网络结构
Fig.6 Network structure of different Transformer blocks

从测试集中随机抽取一张腹部图片以展示 Transformer 块的个数对网络重建结果的影响。如图 7 所示，Only-Skip 可以恢复部分图形细节，但是效果不明显。TE-unet^{+}与前者相比较而言，恢复的图像细节有所增加，纹理结构也更加明显，但仍有一些信息没有恢复。TE-unet 则可以恢复更多的图像特征和纹理结构，在一些图像的细节信息上也恢复的更好。

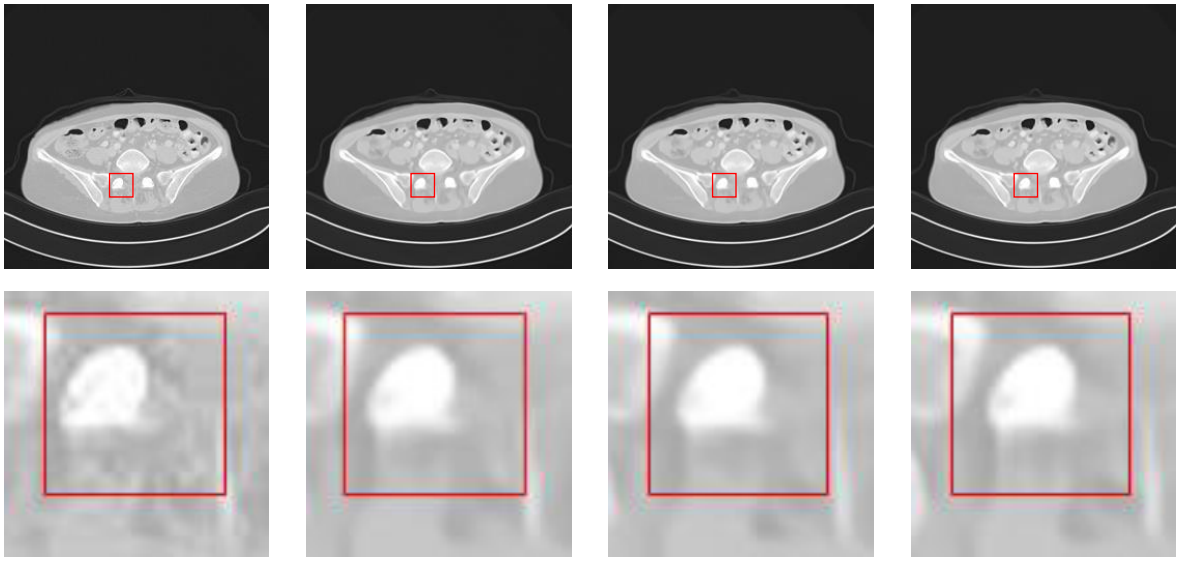
表 3 为不同结构的网络在不同指标下的数值。定量比较表 3 中结果，可见 TE-unet 在 PSNR 和 SSIM 两个指标上达到了最高，RMSE 达到了最小，同时网络参数量和训练时长也达到了一个较好的平衡状态。结果表明，本文所提网络可以更好地学习到图像特征，恢复的图像效果更好，细节更加明显。

2.5.2 不同前馈层

前馈层对网络性能有着很大的影响，因此本小节讨论以不同方式构建的前馈层对网络性能的影响。

图 8 是网络使用不同前馈层重建出的测试集中一张腹部图像。可见，以多层感知机（multiLayer perception，MLP）作为前馈层重建出的图片在细节上仍旧有一些模糊；以 Uformer 中局部增强的前馈网络（locally-enhanced feed-forward，Leff）作为前馈层重建出的图片在细节上有所提高；同

时可见使用 Conv 改进的前馈层，恢复的图像细节信息更多，质量最好。



（a）高清图　（b）Only-Skip　（c）TE-unet^{+}　（d）TE-unet

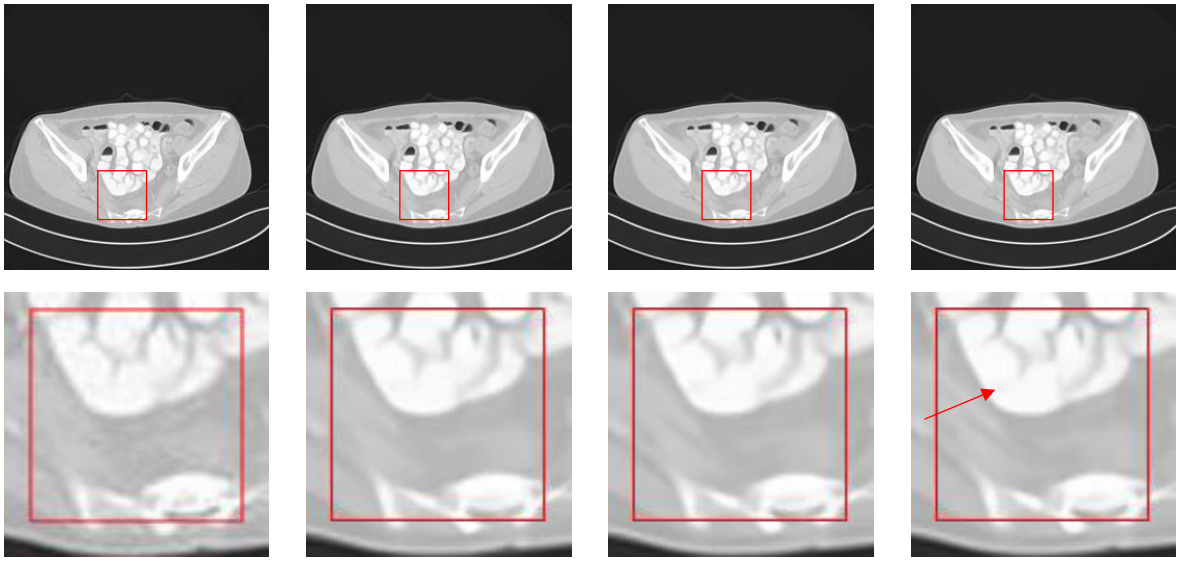
图 7　不同数量 Transformer 块的实验结果及其放大图

Fig.7　Experimental results and enlarged images of different number of Transformer blocks

表 3　不同数量 Transformer 块在测试集中的实验结果

Table 3　Experimental results of different number of Transformer blocks

不同连接方式	PSNR/dB	SSIM	RMSE	参数量/M	训练时长/h
Only-Skip	38.2681	0.9571	0.0124	29.25	11.65
TE-unet^{+}	38.4930	0.9578	0.0121	63.94	15.98
TE-unet	38.5729	0.9583	0.0119	53.55	14.13



（a）高清图　（b）MLP　（c）Leff　（d）Conv

图 8　不同前馈层的实验结果及其放大图

Fig.8　Experimental results and enlarged images of different feedforward layers

从表 4 可见，使用卷积和深度卷积组成的前馈层在 PSNR、SSIM 和 RMSE 3 个指标上都达到了最优。实验结果表明，采用本文所使用的前馈层可以得到更好地重建结果。

表 4 不同前馈层在测试集中的实验结果
Table 4 Experimental results of different feedforward layers

不同前馈层	PSNR/dB	SSIM	RMSE	参数量/M	训练时长/h
MLP	38.1269	0.9569	0.0126	53.29	14.17
Leff	38.5563	0.9569	0.0119	53.55	14.38
Conv	38.5729	0.9583	0.0119	53.55	14.13

2.6 消融实验

为了进一步探索 TE-unet 构成部件对 CT 图像稀疏重建的影响，取消 Transformer 块并标记为 No-Trans。在保持其他参数不变的情况下，使用 PSNR、SSIM、RMSE、参数量和训练时长作为评估指标，可以做出定量比较。

由图 9 可见，不在跳跃连接处加入 Transformer 块得到的实验结果图中，一些细节信息并未被很好地恢复，纹理结构有缺失，而本文方法则可以更好地恢复图像细节信息。

（a）高精度 （b）No-Trans （c）TE-unet

图 9 缺少 Transformer 块的实验结果图可视化
Fig.9 Experimental result diagram of Transformer block is missing

表 5 显示本文贡献所产生的性能改进。No-Trans 网络的 PSNR 值比本文方法低了 0.5158 dB，SSIM 低了 0.0023，RMSE 高了 0.0008。实验结果表明，在跳跃连接处加入 Transformer 块会极大地提升性能，可以更好地重建图像。

表 5 缺少 Transformer 块在测试集中的实验结果
Table 5 Experimental results of missing Transformer block

消融实验	PSNR/dB	SSIM	RMSE	参数量/M	训练时长/h
No-Trans	38.0571	0.9560	0.0127	20.85	9.5
TE-unet	38.5729	0.9583	0.0119	53.55	14.13

3 结语

本文提出的 TE-unet，耦合了 CNN 的局部建模能力、Transformer 的全局建模能力和多种注意力机制，使得网络可以获得很好的去条状伪影能力。首先，通过 CNN、深度可分离卷积、通道注意力和空间注意力构建了一个 CCA 块，其中深度可分离卷积在减少了参数量的同时可以更好地提取局部特征，通道注意力计算出各个通道间的权重，提高特征表示能力，空间注意力使网络关注更感兴趣区域和更重要特征。随后在跳跃连接处加入 Transformer 块融合信息，通过 Transformer 块的处理，去除了一些冗余信息，保留了更加重要的信息。通过整个网络的训练，最终得到更清晰的 CT 重建图像。与现有 4 个经典网络相比，本文提出的 TE-unet，可以在保留更多图像细节的同时去除更多的条状伪影。未来将基于 TE-unet，引入更合适的机制，进一步探索更优的网络结构。

参考文献

[1] BRENNER D J, HALL E J. Computed tomography: An increasing source of radiation exposure[J]. New England Journal of Medicine, 2007, 357(22): 2277-2284. DOI:10.1056/NEJMra072149.

[2] DONOHO D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306. DOI:10.1109/TIT.2006.871582.

[3] SIDKY E Y, KAO C M, PAN X. Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT[J]. Journal of X-ray Science and Technology, 2006, 14: 119-139.

[4] SIDKY E Y, PAN X. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization[J]. Physics in Medicine & Biology, 2008, 53(17): 4777-4807.

[5] LIU Y, MA J, FAN Y, et al. Adaptive-weighted total variation minimization for sparse data toward low-dose X-ray computed tomography image reconstruction[J]. Physics in Medicine & Biology, 2012, 57(23): 7923-7956.

[6] DAVID S, TONY C. Edge-preserving and scale-dependent properties of total variation regularization[J]. Inverse Problems, 2003, 19(6): 165-187. DOI:10.1088/0266-5611/19/6/059.

[7] ZHANG Y, ZHANG W H, CHEN H, et al. Few-view image reconstruction combining total variation and a high-order norm[J]. International Journal of Imaging Systems and Technology, 2013, 23(3): 249-255. DOI:10.1002/ima.22058.

[8] CHEN H, ZHANG Y, KALRA M K, et al. Low-dose CT with a residual encoder-decoder convolutional neural network[J]. IEEE Transactions on Medical Imaging, 2017, 36(12): 2524-2535. DOI:10.1109/TMI.2017.2715284.

[9] ZHANG K, ZUO W, CHEN Y, 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. DOI:10.1109/TIP.2017.2662206.

[10] JIN K H, MCCANN M T, FROUSTEY E, et al. Deep convolutional neural network for inverse problems in imaging[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4509-4522. DOI:10.1109/TIP.2017.2713099.

[11] WOLTERINK J M, LEINER T, VIERGEVER M A, et al. Generative adversarial networks for noise reduction in low-dose CT[J]. IEEE Transactions on Medical Imaging, 2017, 36(12): 2536-2545. DOI:10.1109/TMI.2017.2708987.

[12] OKTAY O, SCHLEMPER J, FOLGOC L L, et al. Attention U-Net: Learning Where to Look for the Pancreas[EB/OL]. (2018-04-11)[2023-02-28]. https://arxiv.org/pdf/1804.03999.

[13] CHEN L, CHU X, ZHANG X, et al. Simple baselines for image restoration[C]//Proceedings of the 2022 European Conference on Computer Vision. Cham: Springer, 2022: 17-33.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information process systems. 2017, 30: 5998-6008

[15] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[EB/OL]. (2020-05-16)[2022-11-22]. https://arxiv.org/pdf/2005.08100.

[16] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations. Austria: OpenReview.net, 2021.

[17] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Piscataway: IEEE, 2021: 9992-10002.

[18] LIANG J, CAO J, SUN G, et al. SwinIR: Image restoration using swin transformer[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE, 2021: 1833-1844.

[19] WANG Z, CUN X, BAO J, et al. Uformer: A general u-shaped transformer for image restoration[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 17683-17693.

[20] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the 2018 European conference on computer vision. Cham: Springer, 2018: 3-19.

[21] HAN Y, YOO J, YE J C. Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis[EB/OL]. (2016-11-19)[2022-10-18]. https://arxiv.org/pdf/1611.06391.

Sparse Reconstruction of Computed Tomography Images with Transformer Enhanced U-net

FAN Xuelin[1], WEN Yuqi[2], QIAO Zhiwei[1]✉

1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
2. School of Materials Science & Engineering, Beijing Institute of Technology, Beijing 102401, China

Abstract: An effective way to achieve low-dose computed tomography (CT) is to reduce the projection angle while maintaining the same radiation dose at each angle. However, a fewer projection angle can result in severe strip artifacts, reducing the practicality and clinical value of the image. To address this issue, a U-shaped network (Transformer Enhanced U-net, TE-unet) coupled with convolutional neural network (CNN) and multiple attention mechanisms was proposed. Firstly, a U-shaped architecture was adopted to fuse multi-scale feature information; Secondly, a module that includes CNN and multiple types of attention was proposed to extract image features; Finally, transformer blocks were added at skip connections to filter information, suppress irrelevant features, and highlight important features. This network combines the local feature extraction ability of CNN and the global information capture ability of Transformer, supplemented by various attention mechanisms, to achieve good ability to remove stripe artifacts. At 60 projection angles, compared to the classic uformer network, peak signal to noise ratio (PSNR) is 0.3178 dB higher, Structural Similarity (SSIM) is 0.002 higher, and Root Mean Square Error (RMSE) is 0.0005 lower. The experimental results show that the proposed TE-unet network reconstructs images with higher accuracy, preserves better image details, and can better suppress strip artifacts.

Keywords: sparse reconstruction; computed tomography; Transformer; multiple attention mechanism; strip artifact

作者简介：樊雪林，男，山西大学计算机与科学技术专业硕士研究生，主要从事医学图像重建、图像处理，E-mail：172953677@qq.com；乔志伟✉，男，博士，山西大学计算机与信息技术学院教授、博士生导师，主要从事医学图像重建、信号处理、大规模最优化等方面的研究，E-mail：zqiao@sxu.edu.cn。