第 50 卷 第 15 期/2023 年 8 月/中国激光 研究论文

中国激光

基于金字塔长程Transformer的OCT图像超分辨率重建

芦焱琦，陈明惠*，秦楷博，吴玉全，尹志杰，杨政奇

上海理工大学健康科学与工程学院，上海介入医疗器械工程技术研究中心，教育部医学光学工程中心，上海 200093

摘要 光学相干层析成像(OCT)在眼科方面的应用通常受到散斑噪声和低分辨率的影响。目前主流的OCT图像超分辨率重建方法多基于卷积神经网络，往往存在成像质量低、图像过度平滑和边缘细节缺失等情况。本文提出了基于Transformer的OCT视网膜图像超分辨率网络——TESR。TESR加入了边缘增强模块，以加强边缘信息对模型的影响，提高视网膜各层边缘的清晰度；新提出的金字塔长程Transformer模块融合了局部特征和全局表示，对图像的内部信息进行长程建模，能更有效地学习更丰富的图像特征。实验结果表示：本文所提TESR模型在峰值信噪比和结构相似度这两个指标上比其他经典模型均有一定程度的提高，在学习感知图像块相似度这一指标上表现优秀，同时在主观视觉质量上也有明显提高，泛化能力较强。

关键词 医用光学；光学相干断层成像；超分辨率；Transformer；自注意力；深度学习

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/CJL230624

1 引 言

光学相干层析成像(OCT)是利用光的弱相干特性获得入射光在目标组织不同深度位置的散射和反射信号从而进行成像的技术。OCT具有非接触、无创、高灵敏度等优点，在眼科成像和诊断中得到了广泛应用[1-2]。然而，两个主要问题阻碍了眼科OCT诊断的发展。首先，OCT成像的干涉性质会导致散斑噪声，不仅降低了对比度，而且会使视网膜的精细结构特征变得模糊。其次，OCT扫描仪的空间采样率往往较低，导致图像的分辨率降低，不利于临床诊断。因此，现实中获得的大多数OCT图像在信噪比和分辨率方面都不是最优的[3]。

研究人员提出了基于软件和基于硬件的方法，以期将低质量的OCT图像恢复成高信噪比、高分辨率的图像。基于硬件的方法包括极化分集[4-5]、空间复合[6]和频率复合[7-8]，这些方法通过改进成像系统中的光源、探测器和其他硬件的结构来提高OCT图像的质量，但无法完全消除图像中的散斑噪声或系统中的白噪声[9]。基于软件的方法包括基于重建[10-11]、滤波[12]、变换[13]、图像稀疏性[14]的方法，这些方法易受其他条件的影响，易产生伪影，而且往往需要复杂的正则化设计，难以应用于临床实践。

近年来，随着深度学习的不断发展，深度学习模型显示出了由低质量输入生成高质量图像的卓越能力[15-16]。其中，卷积神经网络(CNN)已经成为OCT图像超分辨率重建的主力[17-21]。Huang等[17]首次提出了一种基于深度学习的监督方法(SDSR-OCT)，该方法在去除OCT图像中斑点噪声的同时实现了图像的超分辨率重建。Ma等[18]提出了一种边缘敏感的条件生成对抗网络(cGAN)，并用其对OCT图像进行去噪。此外，Qiu等[19]提出了一种具有感知敏感损失函数的深度神经网络(N2NSR)，以减少OCT图像中的噪声，实现OCT图像超分辨率重建。然而，基于卷积神经网络的模型存在两个源于基本卷积层的基本问题，往往会导致重建后的OCT图像出现平滑、模糊和细节缺失，无法可靠地重建病理性结构[22]。作为卷积神经网络的替代方案，Transformer[23]中的自注意机制可以有效捕捉上下文之间的全局交互，弥补了卷积神经网络在长距离依赖上的不足，在一些视觉问题上表现出了良好的性能[24]。

为了解决重建图像边缘细节缺失的问题，弥补基于卷积神经网络的OCT超分辨率重建网络存在的不足，笔者提出了一个基于Transformer的OCT图像超分辨率模型——TESR。首先，加入边缘增强模块以满足OCT图像重视每层结构边缘的要求；其次，提出了一种新的Transformer模块用于深层特征提取，将局部特征提取模块与金字塔池化自注意力模块结合，捕

收稿日期：2023-03-16；**修回日期**：2023-04-18；**录用日期**：2023-04-23；**网络首发日期**：2023-05-05

基金项目：上海市科委产学研医项目(15DZ1940400)

通信作者：*cmhui.43@163.com

获图像的局部和整体特征信息，为后续图像重建提供丰富的信息。将该模型在真实的OCT图像上进行了评估，结果显示：该模型经过训练后，可以在OCT成像中实现超分辨率应用。

2 实验方法与原理

提出的基于Transformer的边缘增强OCT图像超分辨率重建网络TESR如图1所示。图1给出了TESR模型框架的整体结构，包括浅层特征提取模块、深层特征提取模块和图像重建模块(IR)共三部分。

该模型首先通过边缘增强模块将输入图像与提取出来的边缘细节融合，然后经由一个基本的3×3卷积块进行浅层特征提取，即进行图像颜色、纹理、边缘和棱角等低频信息的提取。深层特征提取模块由特征融合模块(FIB)和卷积块组成，用于提取更抽象的语义信息。其中，FIB模块由新提出的金字塔长程Transformer层(PLT模块)和卷积块组成。PLT模块将局部信息获取和全局信息获取两个机制融合。PLT模块中的移位卷积提取模块(Shift-Conv)用于扩大感受野，有效提取图像的局部特征；金字塔池化自注意力模块(P-MHSA)用于加强图像不同部分之间的注意力关系，捕获长距离的特征依赖关系。两个模块相结合为后续图像重建提供细节和整体信息。最后通过卷积模块、ReLU函数以及像素混洗模块完成图像重建。

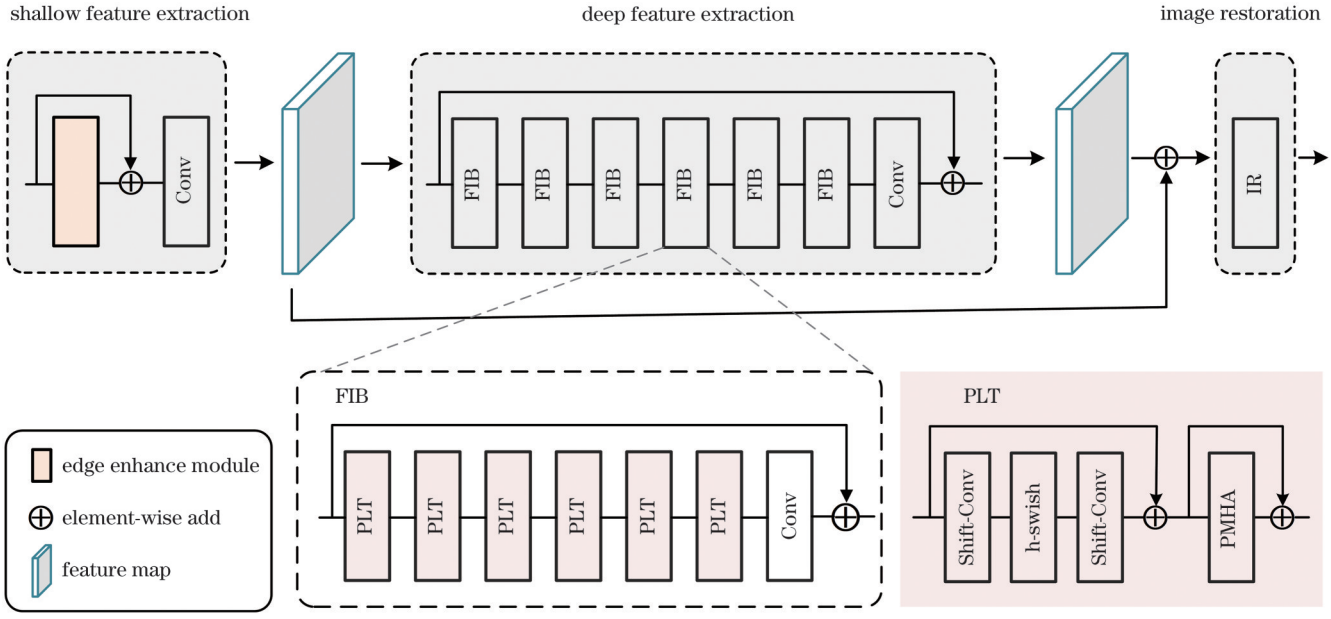

图1 TESR整体框架

Fig. 1 Whole frame of TESR

2.1 边缘增强模块

受Liang等[25]使用EDCNN对低剂量CT图像去噪的启发，笔者在浅层特征提取模块中加入边缘增强功能，并在边缘增强模块中选用Sobel-Feldman算子(也称为Sobel滤波器)。如图2(a)所示，与传统的固定值Sobel算子不同，所提边缘增强模块在Sobel算子中加入了可学习参数α，该参数也被称为“Sobel因子”。该参数的值可以在训练时的优化过程中进行自适应调整，从而提取出不同强度的边缘信息。同时，所提边缘增强模块在传统Sobel算子拥有水平、垂直两个卷积核的基础上，增加了两个对角方向的算子，进一步提高了Sobel算子提取边缘的能力[26]。

所提模块将4个Sobel算子定义为一组，可使用多组可训练的Sobel算子。如图2(b)所示，在边缘增强模块处理过程中，首先在输入图像上使用一定数量(4的倍数)的可训练Sobel算子执行卷积运算，以获得一定数量保存有边缘信息的特征图；然后，该模块将这些特征图通过残差连接与输入的低分辨率带噪声的OCT图像在通道维度上叠加在一起，以获得边缘增强的图像。该模块可以在数据源级别丰富模型的输入信息，为后续浅层特征的提取增强边缘信息，以满足OCT视网膜图像对视网膜层级结构分层的高精度要求。

2.2 金字塔长程Transformer层(PLT模块)

在提取图像浅层特征之后，笔者提出了一个新的PLT模块，用于深层特征提取。该模块可以有效地学习更丰富的图像特征，获取图像的局部信息，建立多尺度的长期依赖关系，从而辅助重建。将多个PLT模块与卷积层连接，可以完善图像细节和整体特征。PLT模块是由局部特征提取模块和金字塔池化自注意力模块组成的，如图3所示。与其他自注意力模块相比，PLT模块更加轻量且高效，它不仅可以减少多头自注意力的计算负载，而且可以通过金字塔池化捕获丰富的上下文信息；同时，移位卷积的使用增加了感受野，

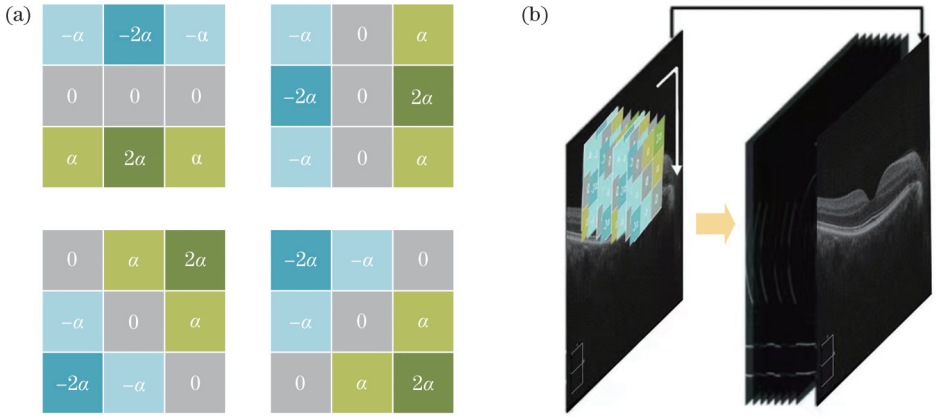

图 2　边缘增强模块。(a)4 种可训练的 Sobel 算子;(b)处理流程

Fig. 2　Edge enhancement module. (a) Four kinds of trainable Sobel operators; (b) process of our module

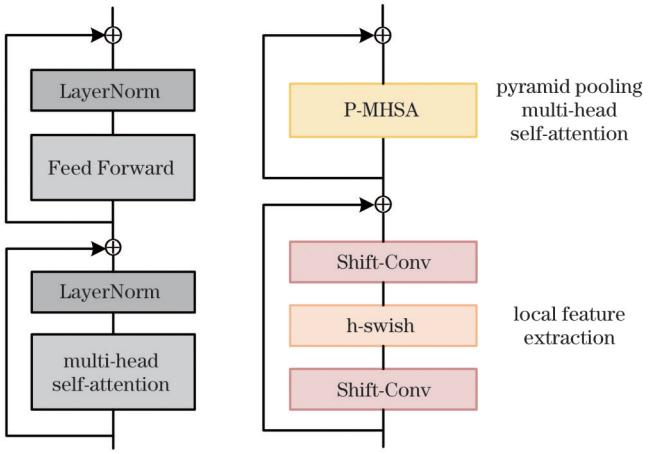

图 3　传统 Transformer 模块(左)和 PLT 模块(右)的对比图

Fig. 3　Comparison of traditional Transformer module (left) and PLT module (right)

使得 PLT 模块能够有效地对局部特征进行提取,弥补了自注意力机制侧重于全局信息获取而对于局部信息获取不足的缺点。PLT 模块极大地增强了局部特征的全局感知能力和全局表示的局部细节。此外,相比于传统 Transformer 模块,PLT 模块删去了部分对超分辨率任务不具有成本效益的组件[27],例如层归一化等,达到了简洁高效的目的。

输入图像首先通过两个移位卷积和一个 h-swish 函数(即 SHS 模块)有效地提取局部特征,然后通过金字塔池化自注意力模块捕获丰富的全局特征信息,最后得到输出。其中 SHS 和金字塔池化自注意力各自使用残差连接,将来自不同尺度特征图的低级细节与高级语义结合起来。

2.2.1　高效的局部特征提取模块(SHS 模块)

以往的研究大多通过多层感知机或两个级联的 1×1 卷积来提取局部特征,然而这些卷积往往只有 1×1 的感受野。本文利用两个移位卷积,并在它们之间进行 h-swish 激活,可以在扩大感受野的同时更有效地提取局部特征。

如图 4 所示,移位卷积[28]由一组移位操作和一个 1×1 卷积组成,其中移位操作在空间上收集数据,1×1 卷积在通道上混合信息。移位操作在逻辑上可表示为

$$\tilde{\boldsymbol{G}}_{k,l,m}=\sum_{i,j}\tilde{\boldsymbol{K}}_{i,j,m}\boldsymbol{F}_{k+\tilde{i},l+\tilde{j},m},\tag{1}$$

式中:k、l 和 i、j 是沿空间维度的索引;m 是通道维度的索引;$\tilde{i}$、$\tilde{j}$ 是重新定义的空间索引;$\tilde{\boldsymbol{G}}_{k,l,m}$ 为移位操作的

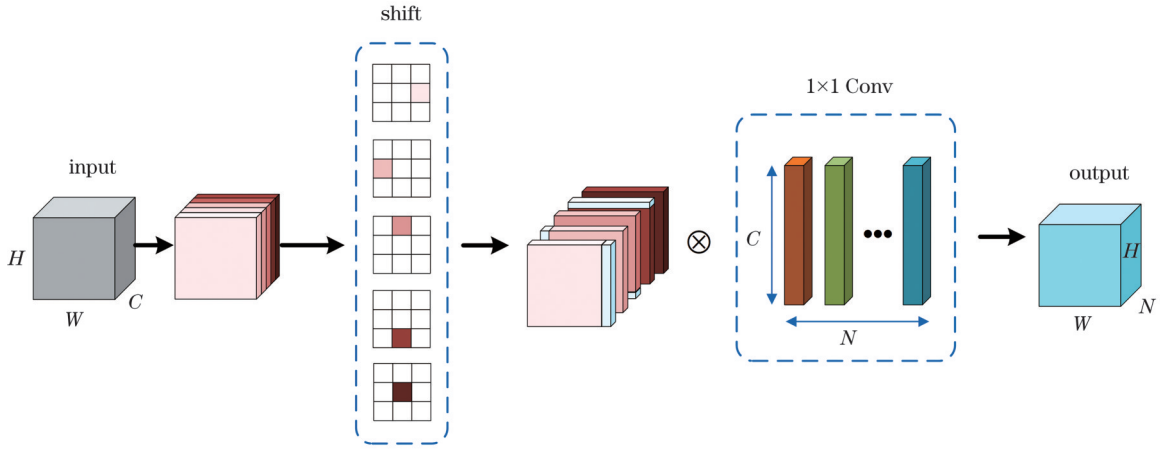

图 4　移位卷积

Fig. 4　Shift-convolution

输出；$F_{k+\tilde{i},l+\tilde{j},m}$为移位操作的输入；$\tilde{K}_{i,j,m}$为移位操作的核，可表示为

$$\tilde{K}_{i,j,m}=\begin{cases}1, & \text{if } i=i_m \text{ and } j=j_m\\ 0, & \text{otherwise}\end{cases},\tag{2}$$

在这里，i_m和j_m是与通道相关的索引，这些索引将$\tilde{K}_{:,:,m}$中的一个值分配为1，将其余值分配为0。$\tilde{K}_{:,:,m}$也被称为移位矩阵。

具体流程如下：先将输入特征按通道平均分成5组，并沿不同的空间维度移动前4组特征，即左、右、上、下4个方位，最后一组通道保持不变。移位运算后再利用1×1卷积获取来自相邻像素的信息。与普通1×1卷积相比，移位卷积本身不需要参数或浮点运算，在不引入额外的可学习参数和大量计算的情况下提供了更大的感受野，提取了更丰富的局部信息，同时保持了与1×1卷积几乎相同的算术复杂度。

此外，在两个移位卷积之间，使用h-swish激活函数[29]引入非线性，提高学习强度。h-swish激活函数的数学表达式为

$$\text{h-swish}(x)=x\frac{\text{ReLU6}(x+3)}{6}。\tag{3}$$

h-swish函数作为swish激活函数的改进，不仅保留了swish函数提高神经网络准确性的优点，并且改善了其计算成本高、速度慢等的缺点[29]。在实践中，h-swish激活函数可以实现分段功能，以减少内存访问次数，从而大大缩短了等待时间。

2.2.2　金字塔池化自注意力

将金字塔池化用于传统自注意力模块中[30]，以捕获图像的整体特征，如图5(a)所示。金字塔池化以其强大的抽象上下文能力在各类视觉任务上的表现都十分出色，而且其空间不变性的自然属性适合用于解决结构信息的丢失问题[31-32]。

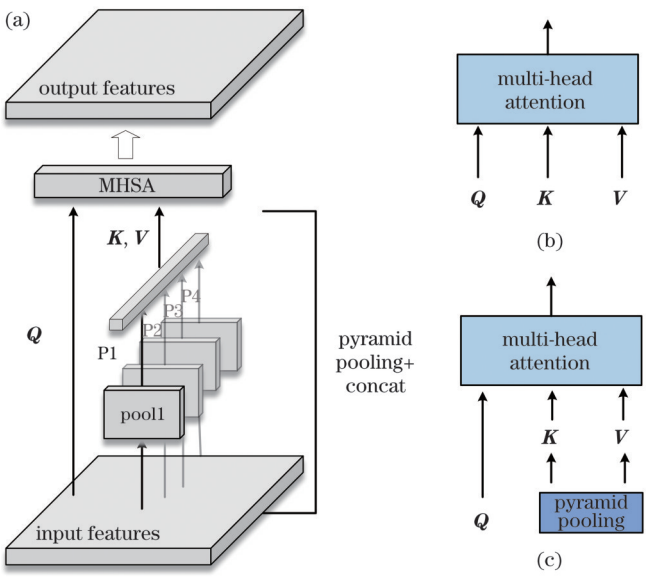

图5　金字塔池化自注意力图解。(a)自注意力机制；(b)多头注意力(MHA)；(c)金字塔池化自注意力

Fig. 5　Diagrams of P-MHSA. (a) Self-attention mechanism; (b) multi-head attention (MHA); (c) P-MHSA

金字塔池化自注意力机制中的关键步骤是将具有不同比率的多个平均池化层应用到输入图像$\boldsymbol{X}$上，以生成金字塔特征图，即

$$\begin{cases}\boldsymbol{P}_1=\text{AvgPool}_1(\boldsymbol{X})\\ \boldsymbol{P}_2=\text{AvgPool}_2(\boldsymbol{X})\\ \cdots\\ \boldsymbol{P}_n=\text{AvgPool}_n(\boldsymbol{X})\end{cases},\tag{4}$$

式中：$\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n$表示生成的金字塔特征图；$n$是池化层的数量。按照经验，本文设置金字塔池化自注意力中并行池化层的数量为4。第一阶段的池化比例设置为$\{32,48,64,80\}$，之后阶段的池化比例分别是第一阶段的1/2，同时将最后阶段的池化比例设置为$\{1,2,3,4\}$。

之后使用深度卷积对生成的金字塔特征图进行相对位置编码，得到$\boldsymbol{P}_i^{\text{encode}}$，$i\in n$。由于$\boldsymbol{P}_i$是池化后的特征图，因此进行位置编码时只需要很小的计算成本。之后，4个金字塔特征图通过展平连接起来。这样得到的$\boldsymbol{P}$序列长度虽然短于$\boldsymbol{X}$，但是包含$\boldsymbol{X}$的上下文抽象信息，在计算自注意力时可以作为$\boldsymbol{X}$的有力替代。即在计算多头自注意力机制中的$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$时，本文使用

$$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=(\boldsymbol{X}\boldsymbol{W}^q,\boldsymbol{P}\boldsymbol{W}^k,\boldsymbol{P}\boldsymbol{W}^v),\tag{5}$$

式中：$\boldsymbol{W}^q$、$\boldsymbol{W}^k$和$\boldsymbol{W}^v$分别表示用于生成$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的线性变换的权重矩阵。相比传统算法，式(5)所示算法进一步减少了计算量。

最后，将得到的$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$用于计算自注意力机制(A)，数学表达式为

$$A=\text{Softmax}\left(\frac{\boldsymbol{Q}\times\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\times\boldsymbol{V},\tag{6}$$

式中：d_K是$\boldsymbol{K}$的通道尺寸，$\sqrt{d_K}$用作近似归一化。Softmax函数沿式(5)所示矩阵的行进行计算。

金字塔池化自注意力和传统多头注意力的区别如图5(b)、(c)所示。由于金字塔池化自注意力中用于计算$\boldsymbol{K}$和$\boldsymbol{V}$的$\boldsymbol{P}$序列长度小于$\boldsymbol{X}$，因此金字塔池化自注意力的计算量显著减少。此外，经由金字塔池化得到的$\boldsymbol{K}$和$\boldsymbol{V}$包含高度抽象的多尺度信息，这使得金字塔池化自注意力在全局上下文依赖性建模中具有更强的能力，有助于深层特征的提取。利用Transformer进行金字塔池化，可为超分辨率重建任务提供强大的特征学习能力。

2.3　图像重建模块

在重建超分辨率图像(SR图像)之前，本文将浅层特征和深层特征聚合，用公式表示为

$$I_{\text{SR}}=H_{\text{REC}}(F_{\text{SF}}+F_{\text{DF}}),\tag{7}$$

式中：$H_{\text{REC}}(\cdot)$是重建模块的函数；F_{SF}是浅层特征模

块的输出；F_{DF} 是深层特征模块的输出。浅层特征主要包含低频信息，而深层特征侧重于恢复丢失的高频信息。通过长跳连接，TESR可以将低频信息直接传输到重建模块，帮助深度特征提取模块专注于高频信息并稳定训练。

本文使用两个卷积块、一个ReLU函数和一个亚像素卷积层构成图像重建模块。亚像素卷积层是一种应用于超分辨率重建任务的具有上采样功能的卷积层，如图6(a)所示，可以将输入特征图像素重组输出高分辨率特征图。

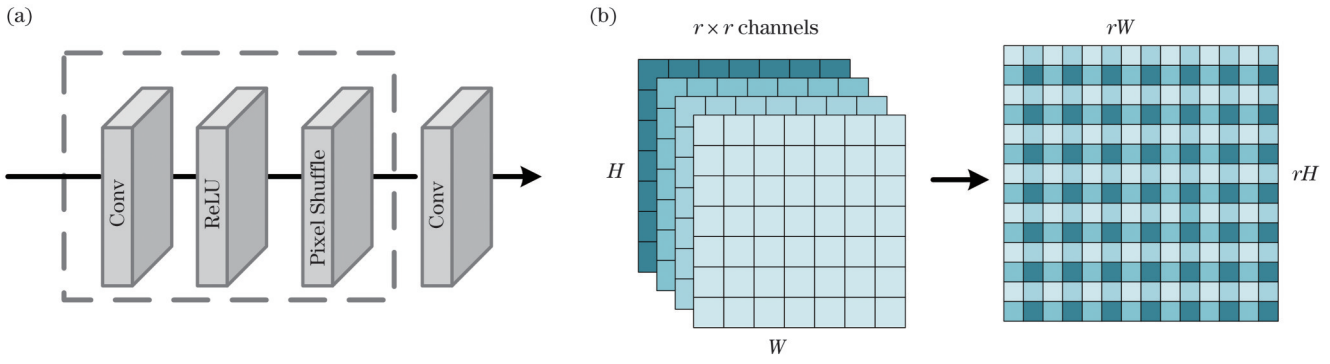

图6　图像重建。(a)整体图；(b)亚像素卷积层示意图

Fig. 6　Image reconstruction. (a) Overall module; (b) schematic of sub-pixel convolution layer

2.4　损失函数

用L1损失函数、感知损失函数和生成对抗损失函数进行组合训练[33]。实验结果证明损失函数的对抗损失权重取0.1时本文模型性能最佳，因此损失函数的计算公式为

$$\mathcal{L}=\mathcal{L}_1+\mathcal{L}_{\text{feat}}^{\phi,j}+10^{-2}\mathcal{L}_{\text{G}},\tag{8}$$

式中：$\mathcal{L}$ 为本算法的损失函数；$\mathcal{L}_1$ 为L1损失函数；$\mathcal{L}_{\text{feat}}^{\phi,j}$ 为感知损失函数；j 表示网络的第 j 层；ϕ 表示第 j 层的特征图像；$\mathcal{L}_{\text{G}}$ 为GAN损失函数。各部分损失函数的具体计算公式为

$$\mathcal{L}_1=\frac{1}{n}\sum_{1}^{n}\left|\left(I_{\text{RHQ}}-I_{\text{HQ}}\right)\right|,\tag{9}$$

$$\mathcal{L}_{\text{feat}}^{\phi,j}\left(I_{\text{RHQ}},I_{\text{HQ}}\right)=\frac{1}{C_jH_jW_j}\left\|\phi_j\left(I_{\text{RHQ}}\right)-\phi_j\left(I_{\text{HQ}}\right)\right\|_2^2,\tag{10}$$

$$\mathcal{L}_{\text{G}}=\sum_{n=1}^{N}\left[-\lg D_{\theta_D}\left(G_{\theta_D}I_{\text{LQ}}\right)\right],\tag{11}$$

式中：I_{RHQ} 表示模型输出图像；I_{HQ} 表示真实的高分辨率图像；I_{LQ} 表示低分辨率图像；C_j 是第 j 层特征图的通道数；H_j、W_j 是第 j 层特征图的维度；D 为判别操作。本文将U-Net网络作为判别器，其主要结构包括解码器、编码器、瓶颈层三部分，是深度学习的经典网络之一。

3　实验结果与分析

3.1　实验设置

1) 实验平台。网络训练所用编程框架为Pytorch1.8，GPU为NVIDIA TESLA V100，32 GB显存。

2) 数据集。本实验中用到了两个真实数据集，分别是由绵阳市第三人民医院和福建医科大学附属协和医院提供的OCT视网膜图像。两个数据集中包含了成对的含噪和干净OCT图像，图像均由临床使用的德国ZEISS视网膜光学相干断层扫描仪获取，图像原始尺寸为1055 pixel×703 pixel。将原始图像裁去白边和外围部分的模糊结构，裁剪后的尺寸为960 pixel×640 pixel。对原始含噪图像进行2倍和4倍下采样来模拟临床实践中的低采样率，生成2倍和4倍的含噪低分辨率图像(LR图像)。下采样过程如图7所示。实验中采用两个数据集中的1000对图像作为训练集，100对图像作为验证集，100对图像作为测试集。

TESR模型的规模较大，因此本文采用Yoo等[34]提出的适用于真实图像超分辨率的混合数据增强方法MoA来提高模型的泛化性能，防止模型出现过拟合。由于OCT图像为灰度图像，本文删去了用于RGB图像的RGB permute方法，设置其余数据增强方法的使用比例为 $P_{\text{CutBlur}}=0.5$，$P_{\text{Cutout}}=P_{\text{CutMix}}=P_{\text{Mixup}}=P_{\text{CutMixup}}=P_{\text{Blend}}=0.1$。其中CutBlur方法将LR图像补丁剪切并粘贴到其相应的高分辨率图像(HR图像)补丁中，通过正则化模型来减小不切实际的失真，对超分辨率模型的增益效果最佳。CutBlur操作过程及可视化结果如图8所示。其余数据增强方法对模型也有一定的增益效果，按经验设置为相同的比例。

3) 训练参数设置。实验中采用Adam(Adaptive Moment Estimation，自适应矩估计)作为优化器，其具体参数设置为：$\beta_1=0.9$，$\beta_2=0.999$，$\varepsilon=10^{-8}$。训练过程中，图像批处理大小设置为16，训练的最大周期为400，初始学习率为 1×10^{-4}，每80个周期学习率衰减一半。

4) 评价指标。采用峰值信噪比(PSNR)、结构相似性(SSIM)和学习感知图像块相似度(LPIPS)来客观地描述模型的性能。

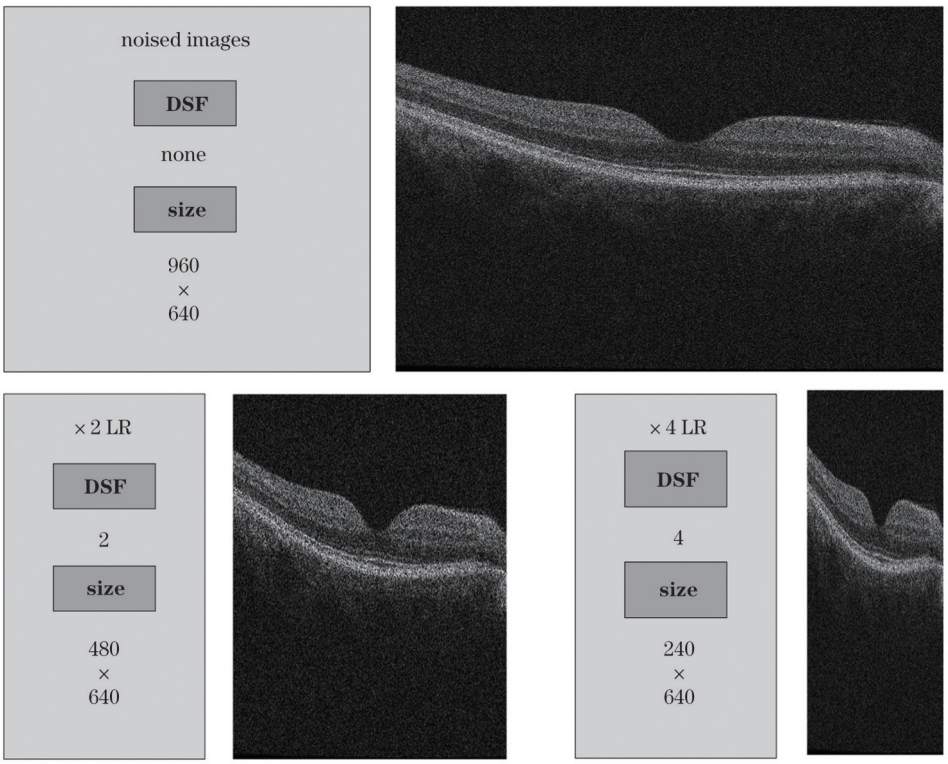

图 7　图像下采样处理(DSF:下采样因子)

Fig. 7　Image down-sampling (DSF: down-sampling factor)

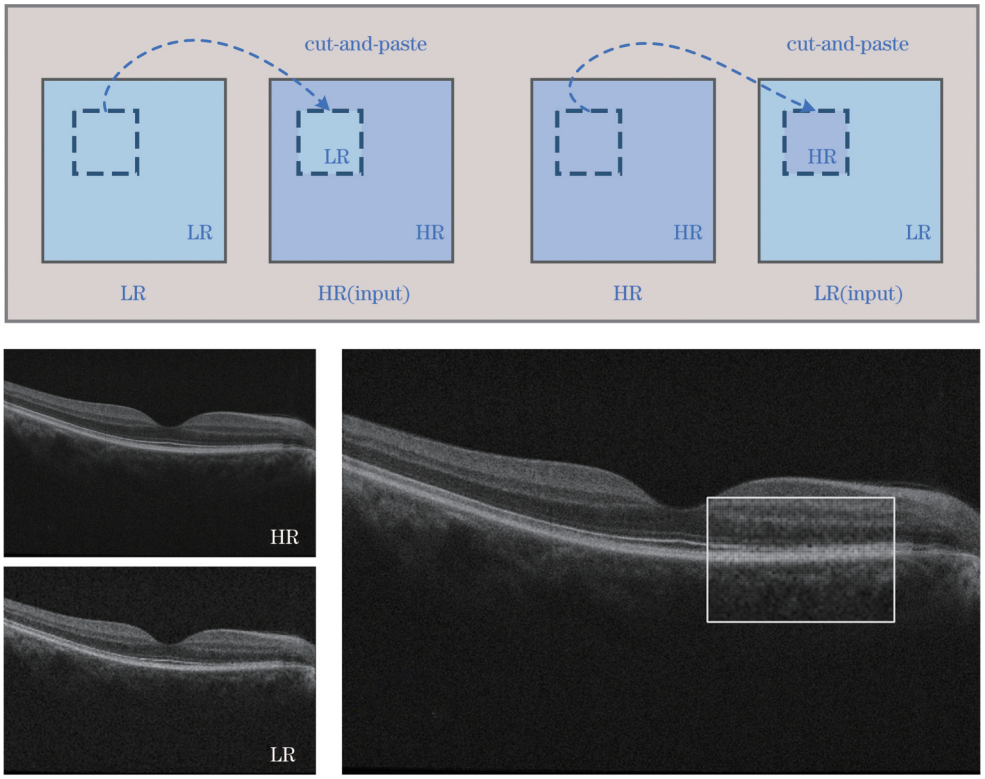

图 8　CutBlur操作过程及可视化

Fig. 8　Process and visualization of CutBlur

5) 对比算法。为验证本文所提模型 TESR 在 OCT 图像超分辨率重建上的优越性,将其与其他经典模型进行对比。对比模型包括 SRGAN[35]、RCAN[36]、IPT[37]、SwinIR[38]。

3.2　定量评价

在两个数据集上分别进行了三次训练和测试,表 1 列出了上述 5 种模型分别在数据集 1 和数据集 2 上进行 2 倍和 4 倍超分辨率处理后重建图像的平均 PSNR

和 SSIM 对比结果。加粗字体为相同条件下的最优值。由测试数据集上的评价结果可以看出，本文所提 TESR 模型在数据集 1 和数据集 2 上均实现了更好的 PSNR 和 SSIM 结果，说明其性能优越。

表 1 各种超分辨率模型重建图像的平均 PSNR 和 SSIM 值

Table 1 Average PSNR and SSIM values of various super-resolution models reconstructed images

Scale	Model	Dataset 1		Dataset 2	
		PSNR / dB	SSIM	PSNR /dB	SSIM
2×	SRGAN	33.05	0.8950	32.48	0.9090
	RCAN	32.41	0.8920	32.42	0.8900
	IPT	33.76	0.9005	34.35	0.8998
	SwinIR	34.83	**0.9130**	34.28	0.9096
	TESR(ours)	**35.53**	0.9124	**35.12**	**0.9140**
4×	SRGAN	30.96	0.7998	30.37	0.7814
	RCAN	30.92	0.7914	31.35	0.7928
	IPT	31.83	0.8112	31.76	0.8068
	SwinIR	32.29	0.8279	32.13	0.8114
	TESR(ours)	**32.91**	**0.8452**	**32.77**	**0.8309**

以数据集 1 中的 4 倍(4×)超分辨率结果为例，本文模型与基于卷积神经网络的 RCAN 相比，在 PSNR 上高出 1.99 dB，在 SSIM 上高出 0.0538；与使用生成对抗网络的 SRGAN 相比，在 PSNR 上高出 1.95 dB，在 SSIM 上高出 0.0454；与同样使用 Transformer 的 IPT 和 SwinIR 相比，在 PSNR 上分别高出 1.08 dB 和 0.62 dB，在 SSIM 上分别高出 0.034 和 0.0173。在数据集 2 上的结果类似。

除了使用 PSNR 和 SSIM 评估重建图像的质量以外，还使用 LPIPS 对重建图像的感知质量进行了评价。表 2 列出了 5 种方法分别在数据集 1 和数据集 2 上进行 4 倍超分辨率处理后重建图像的平均 LPIPS 的对比结果，LPIPS 值越小，说明图像的感知质量越好。可以看出，SwinIR 和使用生成对抗网络的 SRGAN 模型在 LPIPS 指标上比较突出，明显优于 RCAN 模型。IPT 模型重建图像的 LPIPS 指标则介于 SRGAN 和 SwinIR 之间。整体来看，虽然在数据集 2 中 TESR 模型得到的 LPIPS 指标值略高于 SwinIR 模型，但仍然可以说明 TESR 模型在 LPIPS 指标上是比较优秀的，符合人眼的感知情况。

表 2 各种超分辨率模型 4 倍重建图像的平均 LPIPS 值

Table 2 Average LPIPS value of reconstructed images by various super-resolution models after 4× reconstruction

Scale	Model	LPIPS value	
		Dataset 1	Dataset 2
4×	SRGAN	0.214	0.298
	RCAN	0.265	0.306
	IPT	0.203	0.215
	SwinIR	0.170	**0.144**
	TESR(ours)	**0.156**	0.147

TESR 模型的损失函数曲线如图 9 所示，可以看出经过 400 次迭代后，损失函数曲线收敛，模型的稳定性较好。

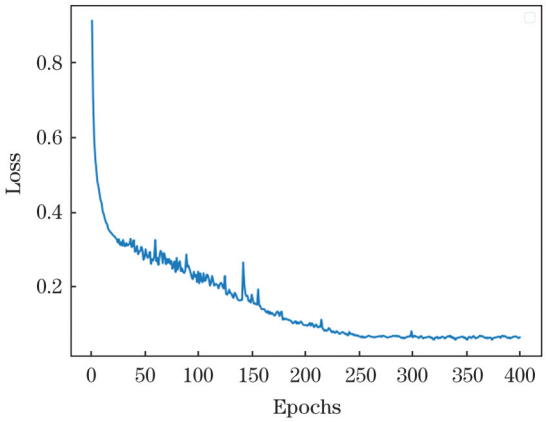

图 9 损失函数曲线

Fig. 9 Loss function curve

3.3 定性评价

为了验证 TESR 的有效性，本节将对重建后的图像进行定性分析，将本文所提模型 TESR 与 RCAN、IPT、SRGAN、SwinIR 模型重建图像进行视觉效果上的直观对比。分别从测试集 1 和测试集 2 中各选取了一幅图像，对 4 倍超分辨率重建结果进行比较。从局部重建图像的对比可以看出，RCAN、IPT、SRGAN、SwinIR 模型重建的超分辨率图像存在明显的模糊、失真、细节不完整等问题。

图 10 展示了不同模型对测试集 1 中正常视网膜 OCT 图像的超分辨率重建结果。与原始的真实图像相比，经 SRGAN 模型重建出的图像最为模糊，而且可以明显看出层次边缘处损失严重；RCAN 模型在层次边缘处有所补偿，但整体的清晰度仍然不够高；IPT 模型和 SwinIR 模型重建出的图像虽然在清晰度上得到了提高，没有明显的噪声问题，而且在空间域中展示了较为清晰的纹理，但是存在伪影，同时有图像过度平滑、细节缺失等缺点。相比之下，本文提出的 TESR 模型利用边缘增强模块和图像特征提取更好地还原出了视网膜的层次信息，边缘锐利，纹理细节清晰，同时没有明显的噪声和伪影问题，整体画面更加干净明了，与 HR 参考图像最为接近。

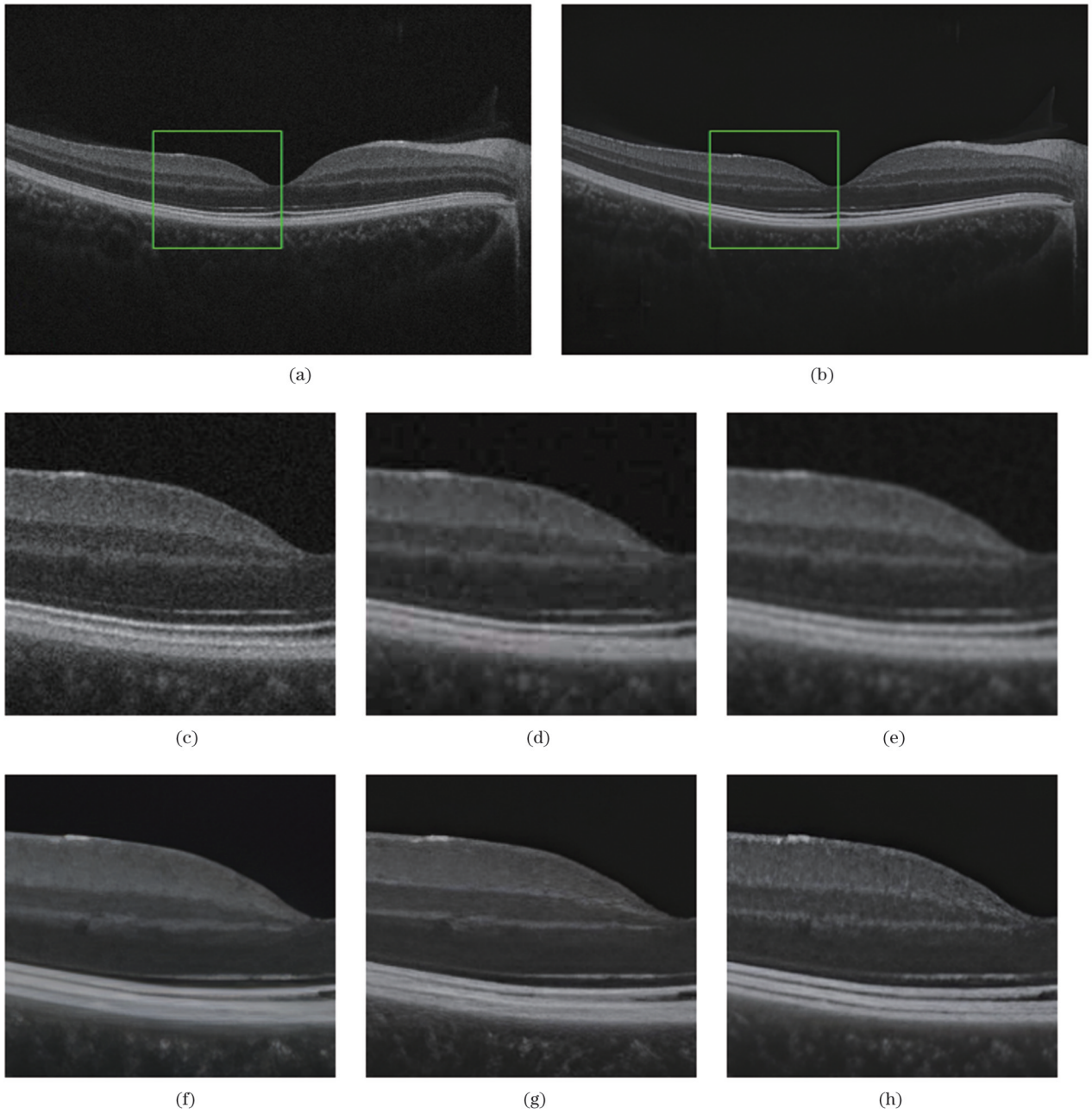

图 10　不同模型对正常视网膜 OCT 图像的超分辨率重建结果。(a)HR 图像;(b)TESR 的重建结果;(c)HR 图像的细节图;(d)～(h)SRGAN、RCAN、IPT、SwinIR、TESR 的局部重建效果

Fig. 10　Super-resolution reconstruction images of a normal retinal OCT image by different models. (a) HR image; (b) TESR reconstructed image; (c) detail of HR image; (d) - (h) local reconstruction effect of SRGAN, RCAN, IPT, SwinIR and TESR

图 11 展示了不同模型对测试集 2 中病变视网膜 OCT 图像超分辨率重建的结果,重建结果与图 10 所示结果相似。放大重点重建部位可以看出:TESR 模型重建图像的视觉质量优于其他经典模型重建图像,清晰度更好,病变部分的位置和大小准确直观,边缘锐利程度更高,没有噪声和伪影,同时基本还原了 HR 图像中的细节信息,为后续病理分析提供了更好的选择。

3.4　分析与讨论

将 TESR 模型与 RCAN、IPT、SRGAN、SwinIR 等 4 种模型的 2 倍和 4 倍超分辨率重建结果进行了对比,SRGAN 的训练时长最短,SwinIR 次之,RCAN 和 TESR 居中,IPT 训练时长最长。TESR 能更好地适应重建难度更大的 4 倍放大倍数的图像,同时更善于重建存在细节特征的图片。TESR 在 PSNR、SSIM 和 LPIPS 这三个图像质量评价指标上均表现优异,这得益于 PLT 模块对图像细节信息的重构能力以及该模块对图像整体特征的整合能力,改善了重建超分辨率图像的质量。主观上,TESR 重建图像的质量相对其他模型更符合人眼视觉的满意程度,对层次结构的恢复尤为明显,这得益于其添加了基于 Sobel 算子的边缘增强模块,加强了

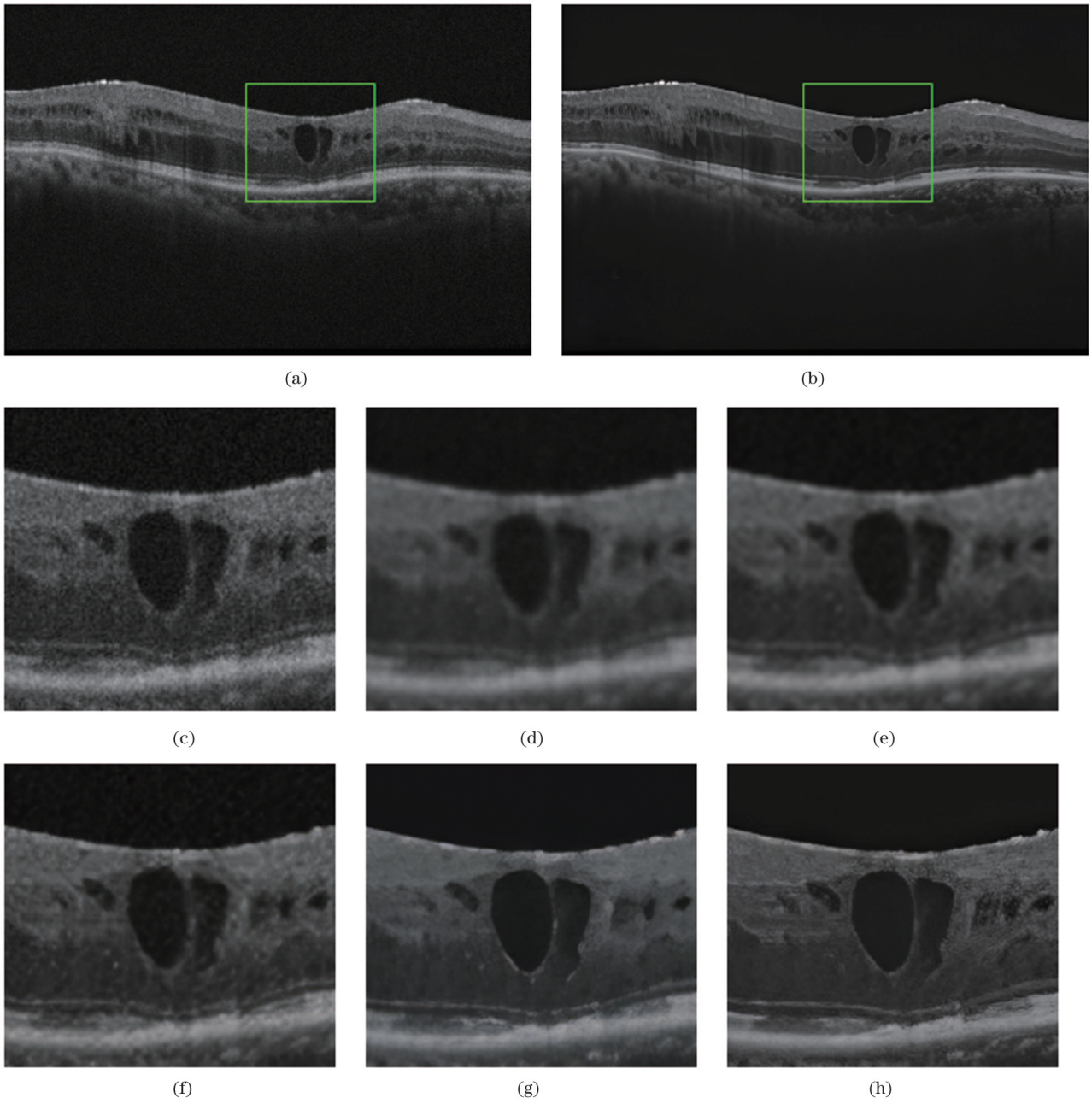

图 11　不同模型对病变视网膜 OCT 图像的超分辨率重建结果。(a)HR 图像;(b)TESR 的重建结果;(c)HR 图像的细节图;(d)～(h)SRGAN、RCAN、IPT、SwinIR、TESR 的局部重建效果

Fig. 11　Super-resolution reconstruction results of a pathological retina OCT images by different models. (a) HR image; (b) TESR reconstructed image; (c) detail of HR image; (d)–(h) local reconstruction effect of SRGAN, RCAN, IPT, SwinIR and TESR

模型对视网膜层次的提取能力。

4　结　　论

针对 OCT 图像超分辨率重建算法过于关注局部特征、忽视整体图像的内部信息、缺乏对视网膜边缘细节的提取等问题,本文提出了一种基于 Transformer 的边缘增强 OCT 图像超分辨率重建模型——TESR。TESR 通过新增的边缘增强模块高质量地还原 OCT 图像的边缘细节信息,同时抑制图像的噪声问题;深层特征提取部分使用的 PLT 模块进一步融合了图像的局部信息和全局信息,对图像的内部整体信息进行长程建模,消除了以往模型易出现的伪影问题,提高了重建图像的真实性。实验时对数据集原有图像进行了数据增强,提高了模型的泛化能力。实验结果表明,本文所提模型可以有效地提高 OCT 视网膜图像的分辨率,以 4 倍重建为例,所提模型与 RCAN、IPT、SRGAN、SwinIR 等经典模型相比在 PSNR 和 SSIM 这两个指标上均有一定程度的提升,并且在 LPIPS 指标上表现优秀,同时在主观视觉上也有明显改善。尽管 TESR 取得了不错的重建结果,但与基于卷积神经网络的轻量级模型相比,该模型中的金字塔池化自注意力仍然是计算和内存密集型模块。未来拟将进一步探索更有效

的自注意力模块，以降低Transformer的计算复杂度，使超分辨率重建技术更好地应用于临床实践。

参 考 文 献

[1] Rebolleda G, Diez-Alvarez L, Casado A, et al. OCT: new perspectives in neuro-ophthalmology[J]. Saudi Journal of Ophthalmology, 2015, 29(1): 9-25.

[2] 王成, 周岐, 陈奕君, 等. 基于低相干干涉测量的眼轴长度测量系统研究[J]. 中国激光, 2022, 49(5): 0507106.
Wang C, Zhou Q, Chen Y J, et al. Axial eye length measurement system based on low coherence interferometry[J]. Chinese Journal of Lasers, 2022, 49(5): 0507106.

[3] Das V, Dandapat S, Bora P K. Unsupervised super-resolution of OCT images using generative adversarial network for improved age-related macular degeneration diagnosis[J]. IEEE Sensors Journal, 2020, 20(15): 8746-8756.

[4] Kobayashi M, Hanafusa H, Takada K, et al. Polarization-independent interferometric optical-time-domain reflectometer[J]. Journal of Lightwave Technology, 1991, 9(5): 623-628.

[5] de Boer J F, Tearney G J, Bouma B E. Apparatus and method for ranging and noise reduction of low coherence interferometry LCI and optical coherence tomography OCT signals by parallel detection of spectral bands: US7643152[P]. 2010-01-05.

[6] Klein T, André R, Wieser W, et al. Joint aperture detection for speckle reduction and increased collection efficiency in ophthalmic MHz OCT[J]. Biomedical Optics Express, 2013, 4(4): 619-634.

[7] Pircher M, Götzinger E, Leitgeb R A, et al. Speckle reduction in optical coherence tomography by frequency compounding[J]. Journal of Biomedical Optics, 2003, 8(3): 565-569.

[8] Bajraszewski T, Wojtkowski M, Szkulmowski M, et al. Improved spectral optical coherence tomography using optical frequency comb[J]. Optics Express, 2008, 16(6): 4163-4176.

[9] Schmitt J M, Xiang S H, Yung K M. Speckle in optical coherence tomography[J]. Journal of biomedical optics, 1999, 4(1): 95-105.

[10] Ozcan A, Bilenca A, Desjardins A E, et al. Speckle reduction in optical coherence tomography images using digital filtering[J]. Journal of the Optical Society of America A, 2007, 24(7): 1901-1910.

[11] Zhao W D, Lu H C. Medical image fusion and denoising with alternating sequential filter and adaptive fractional order total variation[J]. IEEE Transactions on Instrumentation and Measurement, 2017, 66(9): 2283-2294.

[12] Adabi S, Rashedi E, Clayton A, et al. Learnable despeckling framework for optical coherence tomography images[J]. Journal of Biomedical Optics, 2018, 23(1): 016013.

[13] Zhang A Q, Xi J F, Sun J T, et al. Pixel-based speckle adjustment for noise reduction in Fourier-domain OCT images[J]. Biomedical Optics Express, 2017, 8(3): 1721-1730.

[14] Abbasi A, Monadjemi A, Fang L Y, et al. Optical coherence tomography retinal image reconstruction via nonlocal weighted sparse representation[J]. Journal of Biomedical Optics, 2018, 23(3): 036011.

[15] Shah Z H, Müller M, Wang T C, et al. Deep-learning based denoising and reconstruction of super-resolution structured illumination microscopy images[J]. Photonics Research, 2021, 9(5): B168-B181.

[16] 柯舒婷, 陈明惠, 郑泽希, 等. 生成对抗网络对OCT视网膜图像的超分辨率重建[J]. 中国激光, 2022, 49(15): 1507203.
Ke S T, Chen M H, Zheng Z X, et al. Super-resolution reconstruction of optical coherence tomography retinal images by generating adversarial network[J]. Chinese Journal of Lasers, 2022, 49(15): 1507203.

[17] Huang Y Q, Lu Z X, Shao Z M, et al. Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network[J]. Optics Express, 2019, 27(9): 12289-12307.

[18] Ma C, Rao Y M, Cheng Y A, et al. Structure-preserving super resolution with gradient guidance[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 7766-7775.

[19] Qiu B, You Y F, Huang Z Y, et al. N2NSR-OCT: simultaneous denoising and super-resolution in optical coherence tomography images using semisupervised deep learning[J]. Journal of Biophotonics, 2021, 14(1): e202000282.

[20] Das V, Dandapat S, Bora P K. A diagnostic information based framework for super-resolution and quality assessment of retinal OCT images[J]. Computerized Medical Imaging and Graphics, 2021, 94: 101997.

[21] Zhang W W, Yang D W, Cheung C Y, et al. Frequency-aware inverse-consistent deep learning for OCT-angiogram super-resolution[M]//Wang L, Dou Q, Fletcher P T, et al. Medical image comp uting and computer-assisted intervention-MICCAI 2022. Lecture notes in computer science. Cham: Springer, 2022, 13432: 645-655.

[22] 黄永强. 基于深度学习的OCT图像恢复研究[D]. 成都: 四川大学, 2021.
Huang Y Q. Research on OCT image restoration based on deep learning[D]. Chengdu: Sichuan University, 2021.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[24] Wang Z D, Cun X D, Bao J M, et al. Uformer: a general U-shaped transformer for image restoration[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 17662-17672.

[25] Liang T F, Jin Y, Li Y D, et al. EDCNN: edge enhancement-based densely connected network with compound loss for low-dose CT denoising[C]//2020 15th IEEE International Conference on Signal Processing (ICSP), December 6-9, 2020, Beijing, China. New York: IEEE Press, 2021: 193-198.

[26] Luthra A, Sulakhe H, Mittal T, et al. Eformer: edge enhancement based transformer for medical image denoising[EB/OL]. (2021-09-16)[2022-12-05]. https://arxiv.org/abs/2109.08044.

[27] Zhang X D, Zeng H, Guo S, et al. Efficient long-range attention network for image super-resolution[M]//Avidan S, Brostow G, Cissé M, et al. Computer vision-ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13677: 649-667.

[28] Wu B C, Wan A, Yue X Y, et al. Shift: a zero FLOP, zero parameter alternative to spatial convolutions[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 9127-9135.

[29] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3 [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 1314-1324.

[30] Wu Y H, Liu Y, Zhan X, et al. P2T: pyramid pooling transformer for scene understanding[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence: 1-12[2022-12-05]. https://ieeexplore.ieee.org/document/9870559.

[31] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[32] Grauman K, Darrell T. The pyramid match kernel: discriminative classification with sets of image features[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, October 17-21, 2005, Beijing, China. New York: IEEE Press,

2005: 1458-1465.
[33] Wang X T, Xie L B, Dong C, et al. Real-ESRGAN: training real-world blind super-resolution with pure synthetic data[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), October 11-17, 2021, Montreal, BC, Canada. New York: IEEE Press, 2021: 1905-1914.
[34] Yoo J, Ahn N, Sohn K A. Rethinking data augmentation for image super-resolution: a comprehensive analysis and a new strategy[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 8372-8381.
[35] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 105-114.
[36] Zhang Y L, Li K P, Li K, et al. Image super-resolution using very deep residual channel attention networks[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 294-310.
[37] Chen H T, Wang Y H, Guo T Y, et al. Pre-trained image processing transformer[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021. Nashville, TN, USA. New York: IEEE Press, 2021.
[38] Liang J Y, Cao J Z, Sun G L, et al. SwinIR: image restoration using swin transformer[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), October 11-17, 2021. Montreal, BC, Canada. New York: IEEE Press, 2021.

Super-Resolution Reconstruction of OCT Image Based on Pyramid Long-Range Transformer

Lu Yanqi, Chen Minghui*, Qin Kaibo, Wu Yuquan, Yang Zhengqi

Shanghai Engineering Research Center of Interventional Medical Device, the Ministry of Education of Medical Optical Engineering Center, School of Health Sciences and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, *China*

Abstract

Objective Optical coherence tomography (OCT) is widely employed for ophthalmic imaging and diagnosis because of its low latency, noncontact nature, noninvasiveness, high resolution, and high sensitivity. However, two major issues have hindered the development of OCT diagnostics for ophthalmology. First, OCT images are inevitably corrupted by scattering noise owing to the low-coherence interferometric imaging modality, which severely degrades the quality of OCT images. Second, low sampling rates are often used to accelerate the acquisition process and reduce the impact of unconscious motion in clinical practice. This practice leads to a reduction in the resolution of OCT images. With the development of deep learning, the use of neural networks to achieve super-resolution reconstruction of OCT images has compensated for the shortcomings of traditional methods and has gradually become mainstream. Most current mainstream super-resolution OCT image reconstruction networks adopt convolutional neural networks, which mainly use local feature extraction to recover low-resolution OCT images. However, traditional models based on convolutional neural networks typically encounter two fundamental problems that originate from the underlying convolutional layers. First, the interaction between the image and convolutional kernel is content-independent, and second, using the same convolutional kernel to recover different image regions may not be the best choice. This often leads to problems, such as excessive image smoothing, missing edge structures, and failure to reliably reconstruct pathological structures. In addition, acquiring real OCT images affects the training effectiveness of previous models. First, deep learning models usually require a large amount of training data to avoid overfitting; however, it is difficult to obtain a large number of real OCT images. Second, even if the results are excellent, it is meaningless to train the model without using images acquired from OCT devices commonly used in today's clinics. To address the above problems, this study proposes a new OCT image super-resolution model that has the advantages of a convolutional neural network and incorporates a transformer to compensate for its disadvantages, while simultaneously solving the data aspect problem considering recent real clinical images and data enhancement methods during training to increase the generalizability of the model.

Methods In this study, a transformer-based TESR for OCT image super-resolution network was constructed. It constituting three parts: a shallow feature extraction module, a deep feature extraction module, and an image reconstruction module. First, the input image is fused with the extracted edge details using the edge enhancement module, and then shallow feature extraction is performed using a basic 3×3 convolution block. The deep feature extraction module comprises six feature fusion modules, FIB, and a convolution block to extract more abstract semantic information. The FIB module comprises six newly proposed pyramidal long-range transformer layers, PLT, and a convolutional block. The PLT module fuses two mechanisms of local and global information acquisition, where the shifted convolutional extraction module is used to expand the perceptual field and effectively extract local features of the image, and the pyramidal pooling self-attention module is used to strengthen the attentional relationships between different parts of the image and capture feature dependencies over long distances. Finally, image reconstruction was completed using a pixel-blending module.

Results and Discussions We compare our model with four classical super-resolution reconstruction models for 2× and 4× reconstruction, namely, SRGAN, RCAN, IPT, and SwinIR. Quantitative evaluation metrics include the peak signal-to-noise ratio

(PSNR), structural similarity (SSIM), and learning perceptual image patch similarity (LPIPS). For qualitative evaluation, we provide 4× reconstructed images sampled from both datasets for comparison. The experimental results show that TESR outperformed the other methods on both datasets. Objectively, the PSNR results of TESR improved by 7.1%, 6.5%, 3.2%, and 1.9%, the SSIM results improved by 5.9%, 5.3%, 3.5%, and 2.2% (Table 1), and the LPIPS results decreased by 0.1, 0.13, 0.06, and 0.01 (Table 2) for the 4× image reconstruction. Similar results are obtained for 2× image reconstruction. Zooming in on the key reconstructed areas, it is clear that the TESR-reconstructed images can better restore the hierarchical information of the retina using the edge enhancement module and image feature extraction (Fig. 9). The retinal edge structure is sharp, the texture details are clear, and there are no obvious noise or artifact problems (Fig. 10). The overall image is clean with high realism and is close to the HR reference image. The experiment verifies the effectiveness and superiority of TESR for super-resolution reconstruction of OCT images.

Conclusions To address the problems that OCT image super-resolution reconstruction algorithms focus too much on local features and ignore the internal knowledge of the overall image, while lacking the extraction of retinal edge details, we proposes a transformer-based edge enhancement OCT image super-resolution network TESR. TESR restores the edge detail information of OCT images with high quality through the new edge enhancement module, while suppressing the noise problem of the images. The PLT module used in deep feature extraction further fuses the local and global information of the image to model the overall internal information of the image over a long range. This approach eliminates the artifact problem that tended to occur in previous algorithms and improves the realism of the reconstructed images. The experiment shows that the TESR model proposed in this study is better than other classical methods in terms of PSNR and SSIM, respectively. It is excellent in terms of LPIPS, and has a significant improvement in subjective visual quality. Additionally, the model has a strong generalization ability. In the future, more effective self-attentive implementations will be explored to reduce the computational complexity of the transformer and improve the convenience of the super-resolution reconstruction technique for clinical practice.

Key words medical optics; optical coherence tomography; super-resolution; Transformer; self-attention; deep learning