SwinT-Unet:基于双通道自注意力机制的超声图像分割方法

宋艳涛[1,2],路云里[1]
(1. 山西大学大数据科学与产业研究院,山西太原 030006;2. 山西大学计算机与信息技术学院,山西太原 030006)

摘 要: 超声图像分割在疾病诊断和治疗中扮演着关键的角色,但由于超声图像的低对比度、噪声干扰以及病灶在形状、大小和位置上的差异等特点,导致准确地分割出感兴趣的区域仍然是一个具有挑战性的任务. 为了解决这一问题,本文提出了一种双通道自注意力机制U型网络(SwinT-Unet),该网络利用Swin-Transformer与Unet编码器同时进行特征提取. 为了有效融合Swin-Transformer和Unet编码器提取到的不同层级的特征,本文还提出了一个门控双层特征融合模块(Gated Dual-layer Feature Fusion,GDFF),通过门控机制实现了整体特征与局部特征的有效融合,从而提高分割结果的精确度和鲁棒性. 本文在2个不同的超声图像分割数据集上进行了实验,结果表明,本研究所提出的模型在分割准确性和鲁棒性方面均优于现有的卷积神经网络和基于Transformer的网络模型. 本文为超声图像分割领域提供了一种新的方法,并为临床医学诊断和治疗提供了更准确、可靠的支持.

关键词: 超声图像;Unet;Swin-Transformer;图像分割;医学图像
基金项目: 山西省回国留学人员科研教研资助项目(No.2023-015)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)11-3835-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230904

SwinT-Unet: Ultrasound Image Segmentation Based on Two-Channel Self-Attention Mechanism

SONG Yan-tao[1,2], LU Yun-li[1]
(1. *Institute of Big Data Science & Industry*, *Shanxi University*, *Taiyuan*, *Shanxi* 030006, *China*;
2. *School of Computer & Information Technology*, *Shanxi University*, *Taiyuan*, *Shanxi* 030006, *China*)

Abstract: Ultrasound image segmentation plays a key role in disease diagnosis and treatment, but accurately segmenting the regions of interest is still a challenging task due to the low contrast, noise interference, and variability in shape, size, and location of the lesions in ultrasound images. To address this problem, we propose a dual-channel self-attention mechanism U-shaped network (SwinT-Unet), which utilizes Swin-Transformer and Unet encoder to simultaneously extract features. To effectively fuse the different-level features extracted by Swin-Transformer and Unet encoder, we also propose a gated dual-layer feature fusion module (GDFF), which achieves the effective fusion of global and local features through the gating mechanism, thereby improving the accuracy and robustness of the segmentation results. We conduct experiments on two different ultrasound image datasets, and the results show that our proposed model outperforms the existing convolutional neural network and Transformer-based models in terms of segmentation accuracy and robustness. Our paper provides a new method for ultrasound image segmentation, and offers more accurate and reliable support for clinical medical diagnosis and treatment.

Key words: ultrasound image; Unet; Swin-Transformer; image segmentation; medical image
Foundation Item(s): Shanxi Scholarship Council of China (No.2023-015)

1 引言

超声图像是通过超声设备利用超声声束扫描人体,对反射信号进行接收、处理、获得的器官图像,进而得到人体相应位置的器官信息,具有无创、无辐射、便

收稿日期:2023-09-26;修回日期:2024-07-24;责任编辑:梅志强

携、低成本等优点. 超声影像设备约占医学影像设备总保有量的74.44%[1],是医学图像中最为广泛使用的一种成像方式,在疾病诊断和治疗中扮演着关键的角色,广泛应用于心脏病学、妇产科、肿瘤学、神经科等领域. 然而,由于超声图像的低对比度、噪声干扰以及疾病在形状、大小和位置上的差异,导致准确地分割出超声图像中感兴趣的区域仍然是一个具有挑战性的任务. 目前超声图像的识别与分割工作仍然主要依靠富有经验的医生,因此对超声图像的自动分割算法具有很强的应用前景与现实意义.

为了有效地对医学图像进行分割,研究者们提出了多种方法,其中,深度学习模型因其对复杂数据的卓越学习能力备受国内外学者关注[2]. 2015 年 Ronneberger 等人[3]提出了一种称为 Unet 的基于卷积神经网络(Convolutional Neural Network, CNN)的图像分割方法,该方法采用了一种对称的编码器-解码器结构,通过跳跃连接将编码器的特征图与解码器的特征图进行拼接,以保留更丰富的上下文信息和精确的定位能力. 由于其可以较好地融合底层特征,Unet 被广泛应用于医学图像分割. 随后,基于 Unet 的医学图像分割模型被接连提出,如 R2u-net[4], V-net[5]等. Vaswani 等人[6]在 2016 年提出了注意力机制,由于其具有在全局提取特征的特点而引起了研究者的广泛关注. 随后,Rajamani 等人[7]于 2020 年将注意力增强卷积(Attention Augmented Convolution, AAC)模块嵌入到 Unet 的瓶颈层中,得到基于 Unet 的改进模型 Attention-Augmented U-Net 以实现更精确的空间聚合. AAC 将自注意力特征图与卷积特征图进行拼接,从而捕捉长距离的交互进而可以有效地提高分割任务中复杂任务和低对比度的分割性能.

随着 Transformer 在自然语言处理领域的成功应用,研究者们开始思考将其引入计算机视觉领域. 其中, Dosovitskiy 等人[8]在 2020 年提出了 ViT(Vision Tansformer),ViT 通过将图像打成图像块并且利用自注意力机制来提取图像中的全局特征,采用多层 Transformer 结构进行特征建模. ViT 在大规模图像任务中取得了出色的性能,并在图像分割领域引起了广泛关注. 在此基础上,Chen 等人[9]提出了 Trans-Unet,该网络利用 Transformer 对 Unet 所提取的深层次特征进行进一步编码,并使用 Unet 解码器模块对提取的特征进行解码,在多个医学图像数据集上取得了较好的分割效果. Valanarasu 等人[10]将 Transformer 应用于医学图像领域提出了 MedT(Medical Transformer),该模型利用门控轴向注意力和局部全局训练策略,同时考虑医学图像中的远程依赖关系和细节信息. Liu 等人[11]在 2021 年提出了基于滑动窗口自注意力机制的 Swin-T(Swin-Transformer),该网络引入了分层注意力机制和窗口化的卷积操作,以有效解决 ViT 在图像分割任务中的限制. Swin-T 在精度和效率方面取得了显著的提升,成为图像分割领域的重要模型之一. 随后,Cao 等人[12]在此研究的基础上提出了 Swin-Unet,该网络使用 Swin-T 构建了一个 U 形架构,来处理医学图像中的局部和全局语义信息. 尽管神经网络模型尤其是基于 Transformer 的方法在图像分割方面已经取得了较好的效果,但由于超声图像较低的信噪比和对比度特点,目前基于 Transformer 结构的医学图像分割网络往往缺乏对局部特征的充分利用与融合而导致在超声图像分割上表现不佳.

文献[13]将 CNN 与基于注意力机制的 ViT 网络进行了对比,研究结果显示 ViT 比标准的 CNN 网络更能模仿人类的视觉系统. 这是由于卷积操作的固有特点,使得网络往往从一个卷积核大小的区域中进行特征提取,进而导致卷积操作更加偏向于从细节纹理的角度来理解图像信息. 另一方面,由于自注意力机制的特点,使得 ViT 更加偏向于从整体形状上来理解图像信息. 综上考虑,为了将传统卷积操作与自注意力机制两者的优势相结合,从而提升对于数量较少,纹理模糊,容易受到噪声影响的超声图像的分割精度,本文提出了一种双通道注意力 U 型网络(SwinT-Unet). 该网络整体上为 U 形结构,使用两个通道的编码器,Unet 编码器与 Swin-T 编码器来分别对图像进行特征提取,为了进一步增强模型的特征表示能力,本文提出了一个门控双层特征融合模块(Gated Dual-layer Feature Fusion, GDFF),用于耦合 Swin-T 和 Unet 编码器提取到的不同层级的特征信息,通过门控机制选择性地对特征进行融合,从而有效提高分割结果的精度.

本文的主要贡献可以总结如下:(1)提出一种融合了 Swin-T 和 Unet 的医学超声图像分割方法 SwinT-Unet,有效地结合 CNN 和 Transformer 两种机制,既保留了 CNN 对于图像细节信息的捕捉能力,又利用了 Transformer 对于图像全局信息的感知能力,解决传统的医学图像分割方法对超声图像边缘分割不准确的问题,进一步提高分割准确性和鲁棒性. (2)构造了门控双层特征融合方法,通过带有门控机制的双层特征融合模块,将 Swin-T 编码器与 Unet 编码器所提取到的不同通道、不同层级的语义特征进行融合,实现了整体特征与局部特征的有效融合,进一步提升了特征表示能力. (3)在 TN3K(Thyroid Nodule region segmentation dataset)与 BUSI(Breast UltraSound Image)两个医学超声图像数据集上进行了验证,实验证明本文提出的 SwinT-Unet 网络能够更加精确地对超声图像进行分割,实验结果优于所有对比模型.

2　相关工作

2.1　Unet以及衍生网络

Unet是一种广泛应用于生物医学图像分割的全卷积网络,它有一个对称的U形结构,包括一个压缩路径和一个扩展路径,具体如图1所示.

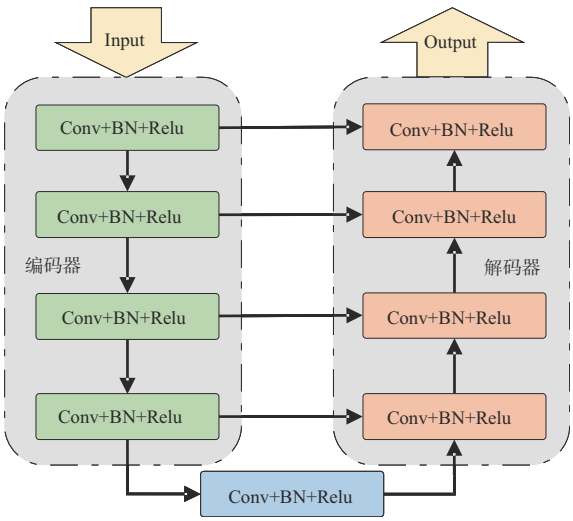

图1　Unet网络结构图

在Unet网络中,压缩路径用于提取图像特征,扩展路径用于恢复图像分辨率.在扩展路径中,每一层都会与压缩路径中对应层的特征进行拼接,以融合位置信息和语义信息.此外,Unet还通过带边界权值的损失函数和数据扩充的方法来提高分割的精度和鲁棒性.Unet的提出为医学图像分割领域带来了重要的突破.它在很大程度上解决了传统方法中对上下文信息利用不足和分割细节损失的问题.随后,基于Unet思想的各种改进方法相继被提出[14~18].其中,2017年Zhang等人[14]提出了Res-UNet,该网络通过在编码器和解码器中加入多个残差连接,有效缓解了梯度消失和低频信息提取不足等问题,进一步提高了网络分割精度和鲁棒性.2018年Zhou等人[15]提出了Unet++,该网络使用了相互联系的解码器子网络,使得不同深度的Unet可以共享同一个编码器,并通过密集连接来实现更多层次的特征融合,从而增强了特征表示能力和细节恢复能力,在监督方法上使用了深度监督机制,使得每个解码器子网络都可以输出一个分割结果,并通过加权求和来得到最终的分割结果,从而增强了梯度传播和优化效果,在多个医学图像分割任务上都表现出了优于Unet的性能.2020年Cai等人[16]提出了Dense-Unet,该网络在Unet的基础上,使用了密集连接的方式,增强了特征的传递和复用,提高了网络的性能和效率.在医学图像分割方面,2021年Baccouche等人[17]提出了Connected-Unets,该网络通过连接多个Unet网络,并且通过修改后的跳跃连接来实现更多层次的特征融合,从而提升医学图像分割性能.Rehman等人[18]提出了BU-Net,该网络使用残差扩展跳跃(Residual Extended Skip,RES)和宽上下文(Wide Context,WC)模块,以及定制的损失函数来改进基准Unet架构,从而挖掘更丰富的特征,并增加有效感受野.尽管基于Unet网络的方法在医学图像分割任务中取得了较大的成功,但是由于Unet采用的常规卷积操作受限于CNN本身的缺陷,其本质上仅限于局部特征的获取,因而不能充分利用空间信息与上下文信息,而这些信息对于超声图像这样的复杂图像分割任务十分重要.

2.2　自注意力机制

自注意力机制可以应用于不同的维度或者层面,如通道、空间、时间等,通过计算每个单元(通道与通道之间、像素点与像素点之间、词语与词语之间)的值,进而对远程空间依赖关系进行建模,提高精确度或者表达能力.自注意力机制的计算公式如下:

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

其中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别为查询矩阵、键矩阵和值矩阵,d_k表示键矩阵的维度.其计算过程如下:首先对于每个输入元素(比如一个向量),分别乘以3个系数$W^{\boldsymbol{Q}}$、$W^{\boldsymbol{K}}$、$W^{\boldsymbol{V}}$得到$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的值,这里的系数就是模型需要学习的参数.其次,利用得到的$\boldsymbol{Q}$和$\boldsymbol{K}$计算每两个输入元素之间的相关性,即计算注意力的值$\boldsymbol{\alpha}$,再对$\boldsymbol{\alpha}$矩阵进行softmax操作得到$\boldsymbol{\alpha}'$,这样可以保证每个元素的权重在0到1之间,并且和为1.最后利用得到的$\boldsymbol{\alpha}'$和$\boldsymbol{V}$计算每个输入元素对应的自注意力层的输出$\boldsymbol{B}$,计算式如下:

$$\boldsymbol{B}=\boldsymbol{V}\cdot\boldsymbol{\alpha}' \tag{2}$$

自注意力机制可以看作是一种网络结构,它可以嵌入到其他模型中,比如Transformer[6]、ViT[8]等.自注意力机制可以更好地捕捉输入数据中不同部分之间的关系和依赖,提高模型的性能和泛化能力,因此,自2016年提出后被广泛应用于计算机视觉等相关领域[19~22].其中,Wang等人[19]首次将自注意力机制引入计算机视觉任务中,提出了Non-Local模块,通过非局部操作捕获长距离依赖关系.在医学图像分割领域,2018年Oktay等人[20]提出了At-Unet(Attention-Unet),这是一种在跳跃连接中通过使用注意力门(Attention-Gate)来重新调整编码器输出特征的方法,该方法在腹部胰腺CT图像分割中取得了较好的分割结果.Schlemper等人[21]提出了注意力门网络,通过注意力门控模块来选择性地聚焦于感兴趣的区域,并过滤掉不相关的区域,在超声波扫描平面检测数据集上取得了很好的效果.2021年Petit等人[22]提出了U-Transformer,该网络利

用基于自注意力与交叉注意力的Transformer结构构建了一个U形架构的网络，并在Unet解码器中进行精细的空间恢复，从而克服Unet无法建模长距离上下文交互和空间依赖的问题.

2.3 Swin-Transformer

近年来，Transformer模型在自然语言处理领域展现出了强大的建模能力，它利用自注意力机制来捕捉序列中的长程依赖关系，同时具有并行化和可解释性强的优势. 因此，将Transformer应用于计算机视觉领域成为了一个研究趋势[10~14,23]. 其中，Xie等人[23]在2021年提出了SegFormer，该网络使用了多个不同层次的Transformer编码器，从而产生多个不同尺度的输出，并且在解码器中使用了MLP(MultiLayer Perceptron)层，使模型最终能够有效结合全局与局部注意力，来提升图像分割性能. 此外，Transformer也被应用于医学图像分割任务. 其中，MedT[10]网络提出了门控轴向注意力模型，通过在自注意力模块中引入一个额外的门控机制来扩展现有的架构. ViT则是将Transformer直接应用于图像分类任务的方法，它将图像切分为固定大小的图像块，并将每个图像块视为一个token，然后输入到Transformer编码器中. 此外，ViT使用了预训练和微调的策略，利用大规模的图像数据集来学习通用的视觉特征表示，然后在小规模的数据集上进行微调. ViT证明了Transformer在图像分类任务上的有效性，但也存在一些问题，如忽略了图像中的局部信息，以及对数据量和计算资源的依赖等. 针对ViT存在的问题，文献[11]提出了一种改进方法Swin-T模型，它使用了分层和移位的窗口来提取多尺度和长程的视觉特征，其模块结构如图2所示. Swin-T将图像切分为不重叠的窗口，并在每个窗口内部进行自注意力计算，然后通过移位操作来交换相邻窗口之间的信息. 此外，Swin-T还采用了金字塔结构，逐渐减少窗口数量和增加窗口大小，从而形成多层次的特征表示，因此Swin-T具有灵活性和线性复杂度，可以作为通用的视觉主干网络. 另一方面，文献[12]通过将图像分块输入基于Swin-T的U形编码器-解码器结构，并利用跳跃连接进行连接，使网络能够对图像的局部与全局特征进行学习. 这些研究证明了Transformer结构在医学图像分割任务上的潜力，但也存在一些问题，如忽略了局部卷积信息，以及对显存消耗较大等.

3 SwinT-Unet网络

如图3所示，SwinT-Unet网络由三部分组成：Unet与Swin-T结合的双通道编码器模块、特征融合模块以及解码器模块. 在图3中，编码器模块由Swin-T与Unet构成，两个模块会分别通过滑动窗口自注意力操作和

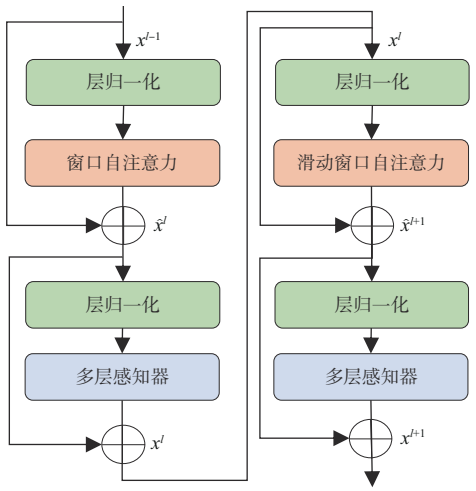

图2 Swin-T模块结构图

卷积操作对图像进行特征提取，再通过下采样操作得到更加深层次的特征图并使通道数翻倍. 在获得两个特征提取模块所提取到的不同层级的特征图后，通过GDFF，将两者所提取到的特征进行融合，再使用与编码器结构相对应的解码器拼接两个编码器所提取的特征，并通过逐层的上采样与卷积来实现对两部分编码器提取特征的解码，减少通道数，逐渐恢复特征图的尺寸，从而得到最终的分割结果.

相比于传统的Transformer结构，Swin-T使用了分层特征表示和移位窗口自注意力机制，使得其具有更高的计算效率和更低的内存消耗，促进了其在视觉任务上的优异表现. 此外，使用基于卷积操作的Unet作为分割头，相比于传统的全连接层或者上采样层，Unet能够更好地保留图像特征信息，并且通过跳跃连接实现了低层特征和高层特征的融合. 因此，本文提出的SwinT-Unet网络将CNN和Transformer两种机制相结合，同时考虑图像的细节和全局特征，从而提高了分割精度和鲁棒性.

3.1 编码器

编码器旨在从输入图像中提取高级语义特征，用于后续的分类或分割任务. 本模型同时使用Transformer与CNN来分别提取全局特征与局部特征，即对于输入图像$\boldsymbol{X}\in\mathbb{R}^{(H\times W\times 3)}$($H$和$W$分别为图像的高和宽)分别从Swin-T和Unet两个通道进行编码提取不同的层次特征.

3.1.1 Swin-T编码器

在Swin-T编码器中，包括四个阶段的处理，每个阶段主要由线性嵌入、堆叠的Swin-T模块以及图像块合并(Patch Merging)模块等操作组成，进而通过窗口以及滑动窗口的自注意力机制来提取图像的信息得到最终的特征. 具体过程如下.

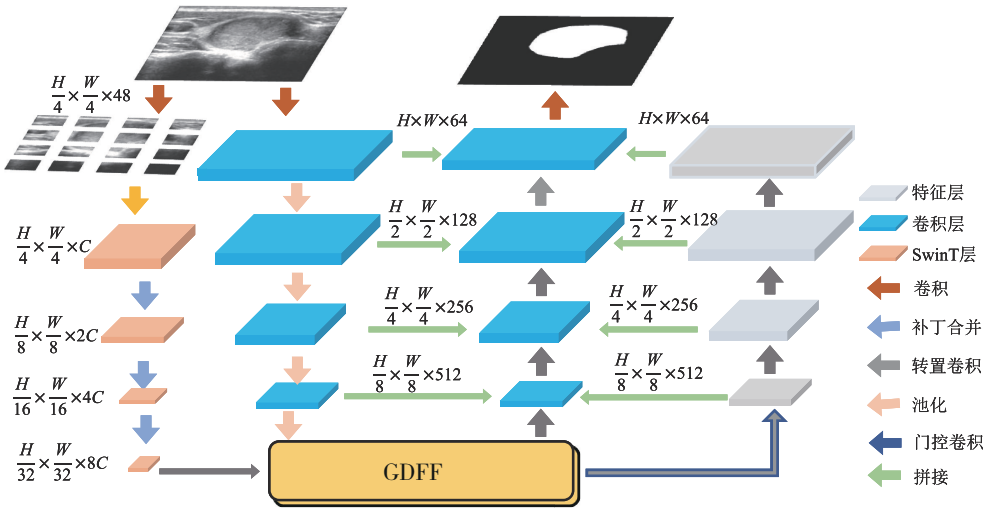

图 3　本文网络结构图

在 Swin-T 编码器的第一阶段，首先进行输入图像序列化，图像块分割（Patch Partition）中的卷积操作将尺寸为 $H\times W\times 3$ 的输入图像 $\boldsymbol{X}$ 分解为大小为 $P\times P$ 的 N 个图像块 $N=\frac{H\times W}{P\times P}$，并将其展平为线性序列 $\boldsymbol{x}=\left\{\boldsymbol{x}_i\in\mathbb{R}^{3\times P^2}|i=1,2,\cdots,N\right\}$. 之后在线性嵌入（Linear Embedding）层通过卷积操作对上述序列 $\boldsymbol{x}$ 做线性变换，并将其映射到维度为 C（C 预设为 96）的空间中，得到新维度为 $\frac{H}{P}\times\frac{W}{P}\times C$ 的序列 $\boldsymbol{x}^l$，l 为编码层. 随后，如图 2 所示，将上述的新序列输入到 Swin-T 中的窗口自注意力和滑动窗口自注意力两个模块中，通过这两个模块对序列 $\boldsymbol{x}^{l-1}$ 进行更新得到 $\boldsymbol{x}^l$ 和 $\boldsymbol{x}^{l+1}$ 两层特征向量，具体计算公式如下所示：

$$\hat{\boldsymbol{x}}^l=\text{W-MSA}(\text{LN}(\boldsymbol{x}^{l-1}))+\boldsymbol{x}^{l-1} \tag{3}$$

$$\boldsymbol{x}^l=\text{MLP}(\text{LN}(\hat{\boldsymbol{x}}^l))+\hat{\boldsymbol{x}}^l \tag{4}$$

$$\hat{\boldsymbol{x}}^{l+1}=\text{SW-MSA}(\text{LN}(\boldsymbol{x}^l))+\boldsymbol{x}^l \tag{5}$$

$$\boldsymbol{x}^{l+1}=\text{MLP}(\text{LN}(\hat{\boldsymbol{x}}^{l+1}))+\hat{\boldsymbol{x}}^{l+1} \tag{6}$$

其中，$\hat{\boldsymbol{x}}^l$ 与 $\boldsymbol{x}^l$ 表示第 l 层（滑动）窗口自注意力模块和多层感知机模块的输出，$\hat{\boldsymbol{x}}^{l+1}$ 与 $\boldsymbol{x}^{l+1}$ 则表示更新后第 $l+1$ 层的输出. 函数 LN（Layer Norm）为层归一化，函数 MSA（Multi-Head Self-Attention）为多头自注意力，函数 W-MSA（Window based Multi-head Self-Attention）为窗口自注意力，函数 SW-MSA（Shifted Window based Multi-head Self-Attention）为滑动窗口自注意力，这样就完成了一个基本的滑动窗口自注意力层.

第二阶段首先通过对图像块进行补丁合并操作，即将相邻的四个图像块按深度拼接成一个新的图像块，从而实现对输入特征的 2 倍下采样，同时将特征维度增加 1 倍. 经过这一步操作后，输入特征的尺寸变为 $\frac{H}{8}\times\frac{W}{8}\times2C$，随后通过对补丁合并后的图像块进行 Swin-T 模块的处理，该模块包含一个窗口自注意力层和一个滑动窗口自注意力层，实现对图像特征的深层次提取.

第三和第四阶段与第二阶段类似，首先都会通过补丁合并操作对特征进行深度拼接，从而实现特征下采样，使特征维度分别调整为 $\frac{H}{16}\times\frac{W}{16}\times4C$ 和 $\frac{H}{32}\times\frac{W}{32}\times8C$，在第三阶段中 Swin-T 模块层数为 6 层，而在第四阶段中 Swin-T 模块层数为 2 层. 在经过整个 Swin-T 编码器模块后，就得到了维度为 $\frac{H}{32}\times\frac{W}{32}\times8C$ 的深层次特征 $\boldsymbol{F}_{\text{Swin-T}}$.

3.1.2　Unet 编码器

在 Unet 编码阶段，编码器通过堆叠的两个 3×3 卷积、批归一化、激活函数以及一个以 2×2 最大池化实现的下采样层来提取图像不同层级的特征. 图像进入编码器后，每经过一个池化层，图像的分辨率就会减半，特征的通道数就会加倍，因此，在 Unet 编码器进行特征提取后，各个阶段分别会得到特征维度为 $\frac{H}{2}\times\frac{W}{2}\times128$，$\frac{H}{4}\times\frac{W}{4}\times256$，$\frac{H}{8}\times\frac{W}{8}\times512$，$\frac{H}{16}\times\frac{W}{16}\times1\,024$ 的特征向量 $\boldsymbol{F}_{\text{Unet}}$.

3.2　GDFF

经过 Swin-T 编码器与 Unet 编码器进行特征提取后，得到了不同尺度的特征 $\boldsymbol{F}_{\text{Swin-T}}$ 和 $\boldsymbol{F}_{\text{Unet}}$，为了将不同层级不同编码器的特征进行深度融合，本文提出了 GDFF 模块. 如图 4 所示，首先使用转置卷积将 Swin-T 编码器所提取到的维度为 $\frac{H}{32}\times\frac{W}{32}\times8C$ 的特征 $\boldsymbol{F}_{\text{Swin-T}}$

进行上采样，从而将其特征维度调整为$\frac{H}{16}\times\frac{W}{16}\times 8C$，之后将Unet编码器所提取到的特征$\boldsymbol{F}_{\text{Unet}}$的特征维度通过卷积调整为$\frac{H}{16}\times\frac{W}{16}\times 1\,024$，将所得到的两组特征向量进行拼接．具体计算如下：

$$\boldsymbol{F}_{\text{conv}}=\text{DC}(\text{TP}(\boldsymbol{F}_{\text{Swin-T}})) \tag{7}$$

$$\boldsymbol{F}_{\text{Concat}}=\text{Concat}(\boldsymbol{F}_{\text{conv}},\boldsymbol{F}_{\text{Unet}}) \tag{8}$$

其中，函数DC由两个卷积层、批量归一化以及激活函数所组成，TP表示将特征进行转置卷积操作，函数Concat表示拼接操作．

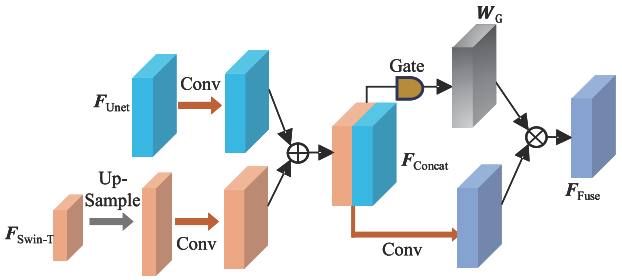

图4 门控双层特征融合模块结构图

拼接后的模块会经过带有门控的卷积模块进行选择性的特征提取．为了进一步对提取的特征进行优化，从而实现对有用特征的加强和无用特征的弃用．本模块提出了一种基于卷积操作的门控机制Gate．具体来说，通过一组可学习参数即带有Sigmoid激活函数的卷积层，来计算输入特征的权重．在同一个局部区域内，特征之间的相关性越高，那么这些特征的贡献就越大，对应的权重也越大，反之亦然．则$\boldsymbol{F}_{\text{Concat}}$经过Gate后得到特征权重$\boldsymbol{W}_{\text{G}}$，其计算式如下：

$$\boldsymbol{W}_{\text{G}}=\text{Sigmoid}(\text{Conv}(\boldsymbol{F}_{\text{Concat}})) \tag{9}$$

其中，Conv表示1×1卷积层，Sigmoid表示Sigmoid激活函数．因此，经过门控卷积模块的处理，模型可以根据学习到的特征权重$\boldsymbol{W}_{\text{G}}$，实现有选择性的特征提取．随后，将这些权重与输入特征相乘，得到最终融合特征$\boldsymbol{F}_{\text{Fuse}}$：

$$\boldsymbol{F}_{\text{Fuse}}=\boldsymbol{W}_{\text{G}}\otimes\text{Conv}(\boldsymbol{F}_{\text{Concat}}) \tag{10}$$

经过门控卷积模块后，得到维度为$\frac{H}{16}\times\frac{W}{16}\times 1\,024$的双通道融合特征，该特征融合了自注意力机制对于全局信息的特征以及卷积机制对于图像细节纹理的特征．

3.3 解码器

解码器模块的作用是将编码器提取的高级语义特征进行逐步恢复和重建，以获得与输入图像相同分辨率的分割结果．本网络中所使用的解码器，与编码器类似，主要由双通道、四个阶段所构成，在每一个阶段都使用了编码器下采样过程中的信息以及特征融合模块所构建的特征向量，从而实现对医学超声图像的精准分割．在解码过程中，首先将门控卷积特征融合模块所提取出的特征与经过上采样的Swin-T编码器所提取到的特征，通过转置卷积进行×2上采样将特征维度转换为$\frac{H}{8}\times\frac{W}{8}\times 512$，与Unet编码器模块对应层次所提取的特征进行拼接与双层卷积、批归一化与激活操作．在得到本层级的输出之后，下一层解码器会再次将上一层的特征与Swin-T编码器进行上采样操作，再次与对应层次的Unet编码器模块进行拼接并卷积．这样在之后的三个阶段中，就会分别得到维度为$\frac{H}{4}\times\frac{W}{4}\times 256$，$\frac{H}{2}\times\frac{W}{2}\times 128$以及$H\times W\times 64$的特征向量，之后在经过两次$3\times 3$卷积，得到最后维度为$H\times W\times 1$的分割预测结果．

4 实验结果

4.1 数据集

本实验采用了TN3K[24]数据集和BUSI[25]数据集作为实验数据．其中TN3K数据集由3 493张包含甲状腺结节的超声图像、3 585张甲状腺超声图像以及由专家手工标注的真值标签组成．BUSI数据集则包含了780张乳腺超声图像，其中133例为正常病例、487例为良性病例和210例恶性病例，以及相应的真值标签．本实验选取了BUSI数据集中的良性病例和恶性病例进行分割．所有数据集按照8∶1∶1的比例随机划分为训练集、验证集与测试集．

4.2 参数设置与评价指标

在模型训练的过程中，将大小不一的图像分辨率统一调整为224×224，并且进行随机水平翻转与垂直翻转来增强网络的泛化性能，初始学习率设置为0.01，训练轮次为200，批量大小为8，使用动量为0.9的随机梯度算法，权重衰退率为10^{-4}，实验中使用了余弦退火法来进行学习率衰退．所有的实验都在单张NVIDIA GTX TITAN GPU上进行训练．

实验使用Dice系数，召回率Recall，交并比（Intersection over Union，IoU），精确度Precision与豪斯多夫距离（Hausdorff Distance，HD）来评估模型预测结果与真值标签的相似程度．具体计算公式如下：

$$\text{Dice}=2\times\frac{\text{precision}\times\text{Recall}}{\text{precision}+\text{Recall}} \tag{11}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{12}$$

$$\text{IoU}=\frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{13}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{14}$$

$$\mathrm{HD}=\max\left(\max_{x\in \boldsymbol{X}}\min_{y\in \boldsymbol{Y}} d(x,y),\ \max_{y\in \boldsymbol{Y}}\min_{x\in \boldsymbol{X}} d(y,x)\right) \quad (15)$$

其中,TP为预测结果中的真正例,FP与FN为假正例与假负例. $\boldsymbol{X}$和$\boldsymbol{Y}$分别表示分割结果边界与分割标签边界,x和y则分别为$\boldsymbol{X}$和$\boldsymbol{Y}$中的像素.

实验中使用Dice损失,其定义如下:

$$L_{\mathrm{Dice}}=1-\frac{2\sum_{i=1}^{N}p_i g_i}{\sum_{i=1}^{N}{p_i}^2+\sum_{i=1}^{N}{g_i}^2} \quad (16)$$

其中,p为预测结果,g为真实标签.

4.3　实验结果与分析

4.3.1　与其他方法的对比

首先,将文本方法与经典网络模型进行对比,对比模型包括以CNN为骨干结构的Unet[3]、Unet++[15]、At-Unet[20],以及以Transformer为骨干结构的Trans-Unet[9]、MedT[10]、Swin-Unet[12]网络. 表1展示了各种方法在TN3K数据集上的分割结果.

从表1中可以看出,在甲状腺结节的超声图像分割任务中,引入注意力机制的分割网络(如At-Unet、Trans-net和MedT)的性能普遍优于仅仅依靠卷积的网络(如Unet和Unet++),在Dice、Recall和IoU这三项指标上均有2个百分点以上的提升. 这是由于引入注意力机制的基于卷积操作的网络(如At-Unet)能够利用注意力机制的全局视野感知能力,加强网络对于高噪声、低信噪比的医学超声图像的理解能力,从而提高分割精度;而基于纯Transformer机制的网络(如Swin-Unet)并不一定能够构建一个有效的U型网络,其分割效果可能弱于基于卷积操作的网络(如At-Unet和Unet++),可能是由于Transformer机制对于医学超声图像的细节信息捕捉能力不足. 与引入注意力机制的分割网络At-Unet相比,本文网络在Dice、Recall和IoU指标上存在0.41、0.05和0.80个百分点的优势. 此外,本文网络结果整体优于同样采用混合网络结构的TransUnet方法,具体来说在Dice、Recall和IoU指标上分别高于TransUnet方法0.65、0.61和1.06个百分点,Precision高出Unet方法2.95个百分点. 这意味着本文所提出的SwinT-Unet能够更好地融合CNN和Transformer的优点,通过Swin-Transformer作为骨干结构对图像的整体信息进行提取,同时使用基于卷积操作的Unet作为分割头对图像的局部信息进行提取,进而获得了最佳的分割效果.

表1　本文网络与对比方法在TN3K数据集上的分割精度

网络		Dice/%	Recall/%	IoU/%	Precision/%	HD
CNN	Unet	77.67±0.46	80.06±0.46	67.01±0.47	81.04±0.34	23.00±0.25
	Unet++	79.43±0.37	82.11±0.42	69.58±0.44	82.01±0.36	19.46±0.45
	At-Unet	80.85±0.26	83.92±0.28	71.37±0.35	83.95±0.24	**13.23±0.41**
Transformer	SwinUnet	78.61±0.35	81.54±0.38	70.96±0.54	81.06±0.37	22.67±0.46
	MedT	79.86±0.31	82.21±0.29	70.02±0.39	81.80±0.24	16.23±0.34
Hybrid	TransUnet	80.61±0.45	83.36±0.43	71.11±0.34	82.71±0.35	14.26±0.52
	SwinT-Unet	**81.26±0.33**	**83.97±0.27**	**72.17±0.41**	**83.99±0.26**	13.63±0.32

对于BUSI数据集,同样使用上述不同骨干网络模型与本文网络进行对比实验,实验结果如表2所示. 与上述实验结果一致,采用CNN与Transformer混合结构的网络(如Trans-Unet和At-Unet)的分割结果整体优于单独的CNN网络结构或者Transformer网络结构. 与基线方法Unet相比,本网络在Dice、Recall以及IoU指标上分别提高了6.52、6.46和7.40个百分点. 本文方法在Dice、Recall、IoU指标上分别优于第2名0.91、0.13和1.15个百分点,HD相比于Unet与第2名的TransUnet分别降低了3.73与1.32. 进一步验证了本文模型在图像分割方面的有效性.

4.3.2　实验结果可视化

图5展示了本文所提出的分割网络在TN3K与BUSI数据集上的部分分割结果,其中(A)、(B)两行为

表2　本文网络与对比方法在BUSI数据集上的分割精度

网络		Dice/%	Recall/%	IoU/%	Precision/%	HD
CNN	Unet	75.34±0.41	76.93±0.40	65.01±0.37	78.06±0.31	23.01±0.47
	Unet++	77.44±0.36	78.26±0.42	68.28±0.32	83.35±0.44	21.49±0.41
	At-Unet	80.74±0.27	82.61±0.35	71.37±0.35	**85.58±0.23**	26.40±0.31
Transformer	SwinUnet	76.23±0.35	77.81±0.54	65.69±0.42	79.35±0.33	25.26±0.42
	MedT	79.46±0.49	82.57±0.26	70.77±0.37	79.14±0.31	22.67±0.27
Hybrid	TransUnet	80.95±0.36	83.26±0.34	71.26±0.37	83.74±0.24	20.60±0.26
	SwinT-Unet	**81.86±0.31**	**83.39**±0.36	**72.41±0.31**	84.39±0.28	**19.28±0.36**

甲状腺结节的超声图像,(C)、(D)两行为乳腺癌变的超声图像. 从左到右分别为原图、Unet、Unet++、At-Unet、Swin-Unet、Trans-Unet、本文网络与Ground Truth.

从图中可以看出,本文所提出的SwinT-Unet模型在各种情况下都能够得到与数据标注结果最为接近的分割效果. 具体来说,图5(A)行中,基于Unet的网络容易将无关区域也划分为甲状腺结节,尤其是Unet误将超声的阴影当成了甲状腺结节,而Unet++和Attention Unet虽然有所改善,但仍然存在这一问题. 而基于Transformer的网络相对于Unet能够更准确地分割出甲状腺结节. 尤其是本文所提出的网络能够更细致地刻画分割边缘,避免过分割的现象. 在图5(B)行中,所有的网络对于图像的大致位置都有了准确的把握,但是在具体细节上基于Unet的网络仍然因为过于注重纹理而丢失了一部分甲状腺结节区域,而基于Transformer的网络能够更好地捕捉到甲状腺结节的形状信息,尤其是本文所提出的网络能够更精确地分割出甲状腺结节的边界. 在图5(C)行中,分割难度较大的区域为乳腺癌变区域的右侧,该区域的上半部分为大片的阴影,下半部分为与背景区分度不高的渐变区域. 对于该图像,基于Unet的网络难以处理阴影和渐变区域,导致分割结果不准确或不完整. 而基于Transformer的网络如Trans-Unet能够更好地分割出图像的边缘细节区域,尤其是本文所提出的网络能够更完整地分割出乳腺癌变区域. 在图5(D)行中,乳腺癌变区域面积过小导致基于Unet的网络普遍无法关注到图像中的目标区域,而基于Transformer的网络能够更敏感地捕捉到图像中的细微变化,得到了较好的分割结果. 综上所述,本文所提出的超声图像分割网络能够有效地利用Swin-T编码器提取深层次特征,并且在解码器中进行多层次特征融合,从而提高了对超声图像细节信息和整体形状信息的理解和表达能力,在信噪比低、目标边缘模糊不清等复杂情况下仍然能够实现高精度、高鲁棒性、高泛化性的超声图像分割.

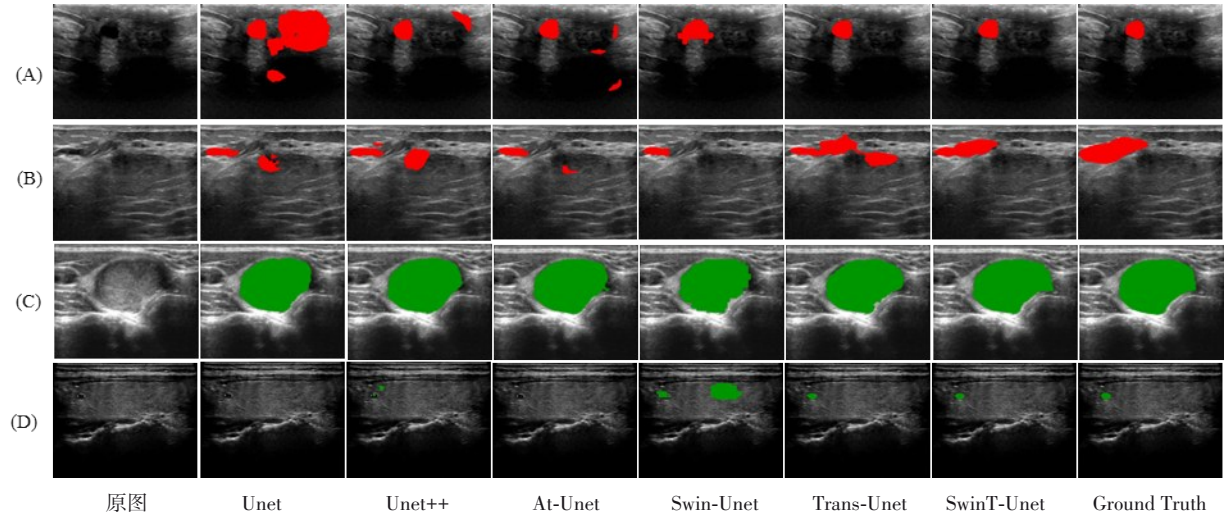

图5 本文网络与对比方法在TN3K与BUSI数据集上的分割可视化结果

4.3.3 Swin-T编码器性能分析

为了验证Swin-T编码器所提取的深层次特征对网络解码器的有效性,本文对不同层次的解码器是否使用Swin-T编码器所提取的深层次特征进行了消融实验. 针对四层解码器中使用的Swin-T特征,设置了4种情况:(1)不融合Swin-T编码器提取的特征,即解码器部分不使用Swin-T编码器所提取的特征(0层);(2)仅在解码器的第一层融合Swin-T编码器提取的特征(1层);(3)仅在解码器的第一、二层融合Swin-T编码器提取的特征(2层);(4)仅在解码器的第一、二、三层融合Swin-T编码器提取的特征(3层). 实验结果如表3所示,最后一行为本文四层解码器均使用Swin-T编码器所提取特征的网络模型.

从表3可以看出,随着解码器对深层次特征的融合层数的减少,在两个数据集中的分割精度都呈现下降趋势,即不使用Swin-T编码器以及减少都会造成分割精度的损失. 当完全不使用Swin-T编码器所提取的特征时,分割精度接近Unet,侧面验证了本文网络中的Swin-T编码器所提取的深层次特征对网络解码器有着重要的指导作用,能够帮助网络更好地理解图像的细

表3 不同层数Swin-T编码器引导解码器消融实验结果 单位:%

使用层数	TN3K		BUSI	
	Dice	IoU	Dice	IoU
0	78.94	69.24	77.41	67.61
1	79.77	70.05	78.41	68.98
2	80.26	70.81	80.24	70.16
3	80.75	71.64	81.01	71.23
SwinT-Unet	81.26	72.17	81.86	72.41

节信息和整体形状信息,从而提升分割精度. 解码器消融实验可视化结果如图6所示,该图展示了逐层消融Swin-T编码器所提取的特征后的网络在BUSI数据集上的部分分割结果,从图中可以看出,当解码器使用的编码器信息逐渐减少时,网络降低了对乳腺癌区域整体形状等长距离语义信息的理解,分割结果的边缘区域逐渐变得不平滑,并且出现了较大的空洞,甚至将一些正常的区域也识别为了癌变区域. 进一步验证了解码器对于Swin-T编码器所提取的特征的使用能够提升网络避免卷积神经网络的缺陷,实现更加精准地分割.

4.3.4　不同特征融合方式性能分析

为了研究特征融合方式对网络性能的影响,本实验选取了三种特征融合方式进行对比实验:(1)门控融合;(2)双层特征融合;(3)本网络融合方式:GDFF. 具体来说,在将两个层级不同通道的特征图进行上采样以及卷积转换为维度与通道数相同的特征图后,门控融合机制会将两个特征直接相加,进而通过门控机制进行特征筛选;而双层特征融合仅仅通过卷积操作来实现特征融合;门控双层特征融合方法使用了带有门控机制的卷积操作来将特征进行有选择性地融合两个层级的特征. 具体实验结果如表4所示.

由表4可知,本文提出的GDFF模块在TN3K与BUSI两个数据集上与门控机制相比,在Dice与IoU指标方面分别提升了2.59个百分点、2.76个百分点与4.48个百分点、4.08个百分点;比双层特征融合机制提升了0.47个百分点、1.16个百分点与0.55个百分点、0.54个百分点. 实验结果表明GDFF模块能够更好地对两个编码器所提取到的深层次特征进行融合,加强了网络对于纹理细节与整体结构的理解,有效地提升了模型的性能.

表4　不同特征融合方式在TN3K与BUSI上的分割精度　单位:%

网络	TN3K		BUSI	
	Dice	IoU	Dice	IoU
门控融合	78.67	69.41	77.32	68.33
双层特征融合	80.79	71.01	81.31	71.87
GDFF	81.26	72.17	81.86	72.41

为了进一步验证本文特征融合算法的有效性,图7展示了一组TN3K数据集中甲状腺结节的超声图像在不同的特征融合方式下的分割结果的比较. 其中,图7(A)行为门控机制特征融合后的分割图像,图7(B)行为双层特征融合后的分割结果,图7(C)行为本文特征融合算法GDFF的分割结果. 同理为了更好地显示分割结果,在原图中选取了一些图像块进行放大对比,其中,黄色曲线为不同融合方式下的分割结果,蓝色曲线为标准分割结果. 通过细节对比图可知,本文所提出的融合算法的分割结果与专家手工分割结果更为接近,其余算法尤其是在边缘区域均产生了较多的误分割. 因此本文提出的特征融合算法能够在获得较好分割结果的同时最大可能地保留图像的细节信息.

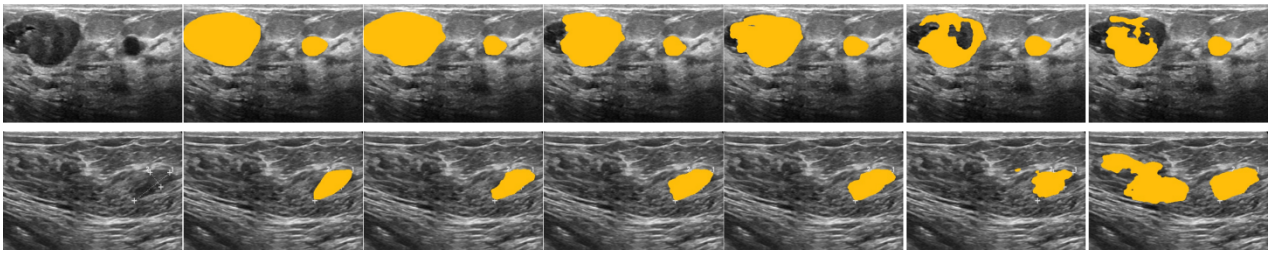

图6　解码器消融实验可视化

4.3.5　模型规模的性能分析

为了研究不同的Swin-T编码器模型大小对于网络性能的影响,本文选取了Swin-T模型的四个版本进行对比,分别是T(Tiny)、S(Small)、B(Base)和L(Large),它们分别对应不同的特征维度、滑动窗口自注意力模块层数以及注意力头数量,具体参数设置如表5所示. 在实验中,由于L版本的参数过多,因此将批量大小统一设置为4. 实验结果如表6所示.

从表6中可以看出,Swin-T模型的大小与其分割性能呈正相关,即模型越大,分割效果越好. 四个版本的模型从T到L,对应的参数量从29 M逐渐增加到197 M. Dice和IoU指数也随着版本规模的扩大而逐步提高. 但相应的,编码器的规模越大,训练时间也越长. 例如,在BUSI数据集上,L的Dice和IoU指标分别达到了82.98%和74.04%,但训练时间为11 h,而T的Dice和IoU分别为81.32%和71.98%,训练时间仅为4 h. Swin-T模型能够随着模型大小的增加而提高分割精度,但同时也会显著增加计算量和内存消耗. 为了兼顾性能与分割精度,在本实验中,选取Tiny网络作为骨干网络.

4.3.6　对复杂样本的处理能力

为了更全面地评估模型的有效性,验证模型对于复杂样本的处理能力. 根据文献[26],本文将Dice分数

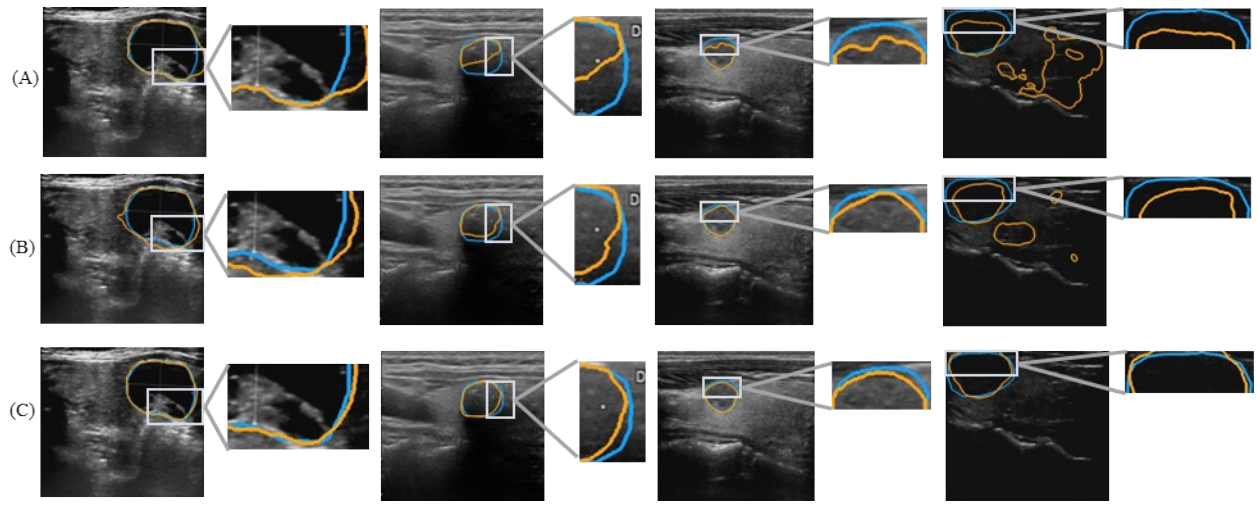

图7 不同特征融合方式下分割结果比较

表5 Swin-T编码器参数设置

网络	维度	层数	注意力头
T	96	{2, 2, 6, 2}	{3, 6, 12, 24}
S	96	{2, 2, 18, 2}	{3, 6, 12, 24}
B	128	{2, 2, 18, 2}	{4, 8, 16, 32}
L	192	{2, 2, 18, 2}	{6, 12, 24, 48}

表6 不同Swin-T编码器规模的分割对比结果

网络	参数/M	TN3K			BUSI		
		Dice /%	IoU /%	时间 /h	Dice /%	IoU /%	时间 /h
T	29	80.74	71.26	11	81.32	71.98	4
S	50	81.32	72.25	17	81.97	72.45	6
B	88	81.98	72.68	23	82.47	73.17	8
L	197	82.24	73.01	40	82.98	74.04	11

低于第2优方法(即表1中,TN3K数据集上At-Unet网络的80.85%和表2中BUSI数据集上TransUnet网络的80.95%)的样本定义为复杂样本.TN3K与BUSI数据集上针对复杂样本的定量比较结果如表7所示,其中黑体表示最优结果,下划线为次优.

在所有比较方法中,本文所提出的网络达到了整体最优性能,在TN3K数据集上Dice、Recall、IoU指标分别为65.97%、69.62%、52.56%.尽管SwinT-Unet的Recall略低,但是本网络的Dice与IoU与次优成绩相比分别提升了1.54个百分点与2.11个百分点.而在BUSI数据集上本网络的Dice、Recall、IoU分别为67.08%、67.66%与55.38%,这些指标分别与次优成绩相比提升了2.31、0.20、1.27个百分点.实验表明本文所提出的网络,在复杂样本上的有效性整体优于所对比的网络,进一步验证了本文模型的有效性.

表7 不同模型针对复杂样本的分割精度比较 单位:%

网络		TN3K			BUSI		
		Dice	Recall	IoU	Dice	Recall	IoU
CNN	Unet	57.18	66.05	44.69	60.86	64.85	49.09
	Unet++	63.87	64.83	48.68	<u>64.77</u>	51.85	53.10
	At-Unet	64.34	66.55	49.70	64.17	52.29	53.83
Trans-former	SwinUnet	59.00	**69.81**	49.88	64.23	58.72	51.41
	MedT	61.64	68.41	47.44	62.75	60.97	52.94
Hybrid	TransUnet	<u>64.43</u>	68.96	<u>50.45</u>	64.49	<u>67.46</u>	<u>54.11</u>
	SwinT-Unet	**65.97**	<u>69.62</u>	**52.56**	**67.08**	**67.66**	**55.38**

5 结论

本文提出了一种基于双通道自注意力机制的超声图像分割方法,将传统的Unet网络与滑动窗口自注意力机制相结合,形成一个具有双分支的编码器结构,进而获得更具鉴别性的图像纹理细节与整体形状特征.为了将两个编码器所提取的特征进行融合,本文提出了门控双层特征融合模块,与两个编码器一起指导Unet解码器进行解码,解决了传统的医学图像分割方法对超声图像边缘分割不准确的问题.在TN3K与BUSI数据集上进行模型性能验证,实验结果表明,本文所提出的SwinT-Unet网络能够更好地平衡超声图像的整体形状与局部细节,取得了优于现有算法的分割性能.进一步的消融实验结果验证了本文提出的门控双层特征融合以及解码器结构的有效性.在复杂样本上进行的实验进一步证明了本模型在处理高噪声、低对比度等复杂医学超声图像上的优异性能.尽管本文使用了小规模Transformer模型,但是依然需要较长训练时间,本文在今后的工作中将对网络模型的轻量化继续进行改进与优化.

参考文献

[1] 黄鑫, 胡艳波. 国产医疗设备临床应用现状与对策探讨[J]. 医疗卫生装备, 2018, 39(9): 75-78.

HUANG X, HU Y B. Problems and countermeasure of clinical application of domestic medical equipment in China[J]. Chinese Medical Equipment Journal, 2018, 39(9): 75-78. (in Chinese)

[2] 张淑军, 彭中, 李辉. SAU-Net:基于U-Net和自注意力机制的医学图像分割方法[J]. 电子学报, 2022, 50(10): 2433-2442.
ZHANG S J, PENG Z, LI H. SAU-Net: Medical image segmentation method based on U-Net and self-attention[J]. Acta Electronica Sinica, 2022, 50(10): 2433-2442. (in Chinese)

[3] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[4] ALOM M Z, HASAN M, YAKOPCIC C, et al. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation[EB/OL]. (2018-05-29) [2023-05-24]. http://arxiv.org/abs/1802.06955v5.

[5] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV). Piscataway: IEEE, 2016: 565-571.

[6] 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. Long Beach: Curran Associates Inc, 2017: 6000-6010.

[7] RAJAMANI K T, RANI P, SIEBERT H, et al. Attention-augmented U-Net (AA-U-Net) for semantic segmentation[J]. Signal, Image and Video Processing, 2023, 17(4): 981-989.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2023-05-24]. http://arxiv.org/abs/2010.11929v2.

[9] CHEN J N, LU Y Y, YU Q H, et al. TransUNet: Transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08) [2023-05-24]. http://arxiv. org/abs/2102.04306.

[10] VALANARASU J M J, OZA P, HACIHALILOGLU I, et al. Medical transformer: Gated axial-attention for medical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 36-46.

[11] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 9992-10002.

[12] CAO H, WANG Y Y, CHEN J, et al. Swin-Unet: Unetlike pure transformer for medical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023: 205-218.

[13] TULI S, DASGUPTA I, GRANT E, et al. Are convolutional neural networks or transformers more like human vision? [EB/OL]. (2021-07-01) [2023-05-24]. http://arxiv.org/abs/2105.07197v2.

[14] ZHANG Z X, LIU Q J, WANG Y H. Road extraction by deep residual U-net[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.

[15] ZHOU Z W, SIDDIQUEE M M R, TAJBAKHSH N, et al. UNet++: A nested u-net architecture for medical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, 11045: 3-11.

[16] CAI S J, TIAN Y X, LUI H, et al. Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network[J]. Quantitative Imaging in Medicine and Surgery, 2020, 10(6): 1275-1285.

[17] BACCOUCHE A, GARCIA-ZAPIRAIN B, CASTILLO OLEA C, et al. Connected-UNets: A deep learning architecture for breast mass segmentation[J]. NPJ Breast Cancer, 2021, 7(1): 151.

[18] REHMAN M U, CHO S, KIM J H, et al. BU-net: Brain tumor segmentation using modified U-net architecture[J]. Electronics, 2020, 9(12): 2203.

[19] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.

[20] OKTAY O, SCHLEMPER J, LE FOLGOC L, et al. Attention U-net: Learning where to look for the pancreas[EB/OL]. (2018-05-20) [2023-05-24]. http://arxiv. org/abs/1804.03999v3.

[21] SCHLEMPER J, OKTAY O, CHEN L, et al. Attention-gated networks for improving ultrasound scan plane detection[EB/OL]. (2018-04-15) [2023-05-24]. http://arxiv.org/abs/1804.05338.

[22] PETIT O, THOME N, RAMBOUR C, et al. U-Net transformer: Self and cross attention for medical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 267-276.

[23] XIE E Z, WANG W H, YU Z D, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[EB/OL]. (2021-10-28) [2023-05-24]. http://arxiv.org/abs/2105.15203v3.
[24] GONG H F, CHEN G Q, WANG R R, et al. Multi-task learning for thyroid nodule segmentation with thyroid region prior[C]//2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). Piscataway: IEEE, 2021: 257-261.
[25] AL-DHABYANI W, GOMAA M, KHALED H, et al. Dataset of breast ultrasound images[J]. Data in Brief, 2020, 28: 104863.
[26] DU J, GUAN K, LIU P, et al. Boundary-sensitive loss function with location constraint for hard region segmentation[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(2): 992-1003.

作者简介

宋艳涛 女,1989年7月出生于山西省临汾市. 现为山西大学大数据科学与产业研究院副教授、硕士生导师. 主要研究方向为医学图像处理、计算机视觉、机器学习.
E-mail: songyantao@sxu.edu.cn

路云里 男,1999年1月出生于山西省长治市. 现为山西大学计算机与信息技术学院研究生. 主要研究方向为计算机视觉.
E-mail: 1454408685@qq.com