周梦雨, 孙丽萍, 刘坤,等. 基于多头注意力机制的手术器械图像分割方法[J]. 智能计算机与应用,2024,14(7):145-150.
DOI:10.20169/j.issn.2095-2163.240722

基于多头注意力机制的手术器械图像分割方法

周梦雨[1], 孙丽萍[2], 刘 坤[1], 徐乃岳[1], 雷雪怡[1]
(1 上海理工大学 健康科学与工程学院, 上海 200093; 2 上海健康医学院 医疗器械学院, 上海 201318)

摘 要: 对手术器械的自动分割是微创手术机器人稳定运行的保障,目前的手术器械分割方法都由串联连接的,容易造成细节丢失。因此,本文提出了一种基于多头注意力机制的手术器械分割方法(ST-HRNet),采用 HRNet 结构构建并行子网络直接输出高分辨率特征图,防止细节丢失;还融合滑动窗口多头注意力机制来获取全局信息进一步提高分割精度。在 Endovis2017 手术器械数据集和私有数据集上与 Unet、TransUNet、GCnet、HRnet 方法进行了对比实验,实验表明 ST-HRNet 方法效果最佳。

关键词: 手术器械分割; 多头注意力机制; 高分辨率特征图
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)07-0145-06

ST-HRNet: Multi-head attention mechanism and high-resolution network for surgical instrument image segmentation

ZHOU Mengyu[1], SUN Liping[2], LIU Kun[1], XU Naiyue[1], LEI Xueyi[1]
(1 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 School of Medical Instrumentation, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China)

Abstract: The automatic segmentation of surgical instruments is the guarantee for the stable operation of minimally invasive surgical robots. The current surgical instrument segmentation methods are composed of high-to-low-resolution sub-networks connected in series, which is prone to loss of details. Therefore, this article proposes a surgical instrument segmentation method based on multi-attention mechanism and high resolution (ST-HRNet). The model uses the HRNet structure and constructs parallel sub-networks to directly output high-resolution feature maps, preventing the loss of details. In addition, the model also integrates the sliding window multi-attention mechanism to obtain global information, which further improves the accuracy of model segmentation. On the Endovis2017 surgical instrument data set and the privatedata set, the average intersection and union ratios were 95.86% and 97.17%, respectively, and the remaining indicators also exceeded the existing methods. Experiments show that ST-HRNet has better results than other segmentation methods.

Key words: surgical instrument segmentation; multi-head attention; high-resolution feature maps

0 引 言

近年来,内窥镜微创手术机器人被广泛应用在临床中,手术器械分割技术在机器人辅助手术中起着至关重要的作用。利用手术器械分割技术,临床医生可以更精确地控制手术器械,从而缩短手术时间,减少术中出血量,提高手术效率和准确性[1]。然而,由于手术器械结构精密,工作姿态不断变化,人体组织结构复杂等问题,手术器械的语义分割相比医学中其他分割任务更加具有挑战性。

深度学习广泛应用于医学影像的辅助诊断中[2]。对于手术器械的分割,相关领域的研究人员提出了许多方法,如:Zhen 等[3]提出采用双重注意力模块和金字塔上采样模块分别捕获联合语义信息和全局上下文进行建模的金字塔注意力聚集网络。为了突出手术器械区域,Yang 等[4]提出一种 U-net

基金项目: 国家重点研发计划资助项目(2018YFB1307700)。
作者简介: 周梦雨(1999-),女,硕士研究生.主要研究方向:医学图像分割; 刘 坤(1998-),男,硕士研究生,主要研究方向:医疗数据分析; 徐乃岳(1999-),男,硕士研究生,主要研究方向:图像处理; 雷雪怡(1999-),女,硕士研究生,主要研究方向:医学图像分割。
通讯作者: 孙丽萍(1971-),女,硕士,教授,主要研究方向:智能医疗机器人,医疗健康大数据。Email:zhou_meng_yu_66@163.com
收稿日期: 2023-05-25

变体网络,引入非局部注意块和双关注模块;Yu 等[5]提出一种采用密集上采样卷积代替反卷积进行采样,并且在每个侧输出层上设置侧损失函数的 U 型变体网络。但目前应用于手术器械分割的方法通常由串联连接的高到低分辨率子网络组成,均有一个低到高恢复高分辨率的过程,这个过程往往借助插值法或者转置卷积实现,导致输出图片的精度较低,丢失大量信息,无法保证有效的分割手术器械尖端部位。

针对目前算法存在的一些局限性,本文基于 HRnet[6]网络框架,结合 Swin transformer[7]滑动窗口自注意力模块,提出了一种基于多头注意力机制的手术器械图像分割方法,命名为 ST-HRNet。通过对输入图像进行并行且密集的预测,直接输出高分辨率图像,减少信息损失,对手术器械尖端的分割更为准确。在公开数据集 Endovis2017 和私有数据集上进行了对比实验,得出的各项指标均为最高,证明本文提出的方法的综合性能优于其他对比方法。

1 方法理论

本文基于 HRnet 并行网络结构构建编码器提取特征,进行重复的多尺度融合,输出高分辨率图像,能够对细节特征进行提取;用 Swin-transformer 模块替换原始 HRnet 网络中大量使用的卷积层,利用分层设计和移动窗口减少了计算复杂度,增强模型的全局感受野,突出手术器械边缘特征。

1.1 改进的 HRnet 网络模型结构

手术机器人中的手术器械尖端是非常重要的部分,通常很微小且形状和尺寸变化剧烈,需要兼顾细节的处理和全局特征的把握。传统的对称性编解码器网络往往先经过卷积池化得到低分辨率抽象特征,再逐步上采样增加特征图的分辨率,还原图片尺寸。但是简单的上采样会丢失较多的全局信息,跳跃连接逐级融合低层特征时又容易引入背景噪声[8]。HRnet 是一种高分辨率网络,从一个高分辨率的子网络开始作为第一阶段,逐步增加高分辨率到低分辨率的子网络,形成新的阶段,在每个阶段中间对并行的多分辨率子网络进行尺度融合。这种高到低分辨率子网络并行的网络结构,不需要从低恢复到高分辨率的过程,因此不会丢失太多细节特征[9]。此外,HRnet 模型执行重复的多尺度融合,利用相同深度的中分辨率和低分辨率特征图以提高高分辨率的表示,使高分辨率也有丰富的语义信息。该网络的有效性在 COCO 关键点检测数据集和 MPII 人体姿势数据集上已经得到证明,但目前还未用于内窥镜图像手术器械的分割研究中。

为了将 HRnet 网络模型更好地应用在内窥镜下手术器械分割任务中,且考虑到医学图像分割的数据集较少的现实,本文选取该网络结构体系的基线网络 HRNet-W32 进行结构改进,本文提出的 ST-HRNet 网络模型的结构,如图 1 所示。ST-HRNet 与 HRnet 具有相同的网络模型架构,将降采样拓展分支之间的卷积过程称为一个阶段,共有 4 个阶段。第一阶段由瓶颈模块组成,从第二阶段开始,除了第一分支用原始的基本卷积模块,其余分支都替换成 Swin transformer 模块,增加模型的全局感受野,通过不同注意力信息来提高模型表达能力。网络中一共有 4 个并行分支,各分支的分辨率分别为输入图像的 1/2、1/4 、1/8 、1/16,为了避免分辨率下降造成特征图的丢失,每个新增的低分辨率分支通道数增加一倍。Swin transformer 模块中包含了滑动窗口多头注意力机制,随着通道数的增加,注意力头数也成倍增加,在第四阶段的最后将所有分辨率的特征图上采样到相同尺寸进行相加,再经过一次上采样得到高质量分割图。ST-HRNet 网络模型具体结构配置见表 1,其中[1×1,64]表示输出通道为 64 的 1×1 的卷积;SW-SMA 表示 Swin transformer 模块中的滑动窗口多头自注意力机制,窗口大小默认为 7;(C_1,C_2,C_3,C_4)、(H_1,H_2,H_3)、(R_1,R_2,R_3) 表示不同分辨率相关联的变换器块中的通道数、注意力头数、多层感知机的扩展比,分别为(30,60,120,240)、(3,6,12)、(4,4,4);(M_1,M_2,M_3,M_4) 和(B_1,B_2,B_3,B_4) 表示每个阶段的模块数和每个阶段重复的次数,分别为(4,4,4,4)和(1,1,4,2)。

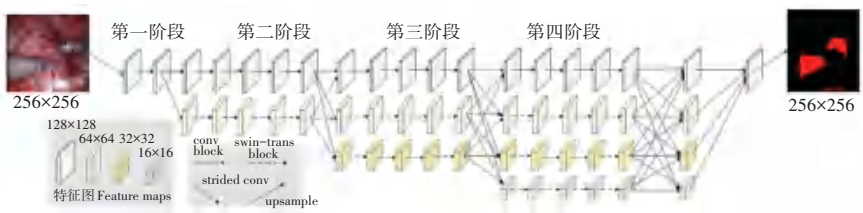

图 1 ST-HRNet 模型结构

Fig. 1 Structure of the ST-HRNet model

表 1　ST-HRNet 模型的具体结构配置

Table 1　Architecture configuration of ST-HRNet

模型结构	第一阶段	第二阶段	第三阶段	第四阶段
一分支	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times M_1\times B_1$	$\begin{bmatrix}3\times3,C_1\\3\times3,C_1\end{bmatrix}\times M_2\times B_2$	$\begin{bmatrix}3\times3,C_1\\3\times3,C_1\end{bmatrix}\times M_3\times B_3$	$\begin{bmatrix}3\times3,C_1\\3\times3,C_1\end{bmatrix}\times M_4\times B_4$
二分支		$\begin{bmatrix}\text{SW-MSA},H_1,C_2\\\text{MLP},R_1\end{bmatrix}\times M_2\times B_2$	$\begin{bmatrix}\text{SW-MSA},H_1,C_2\\\text{MLP},R_1\end{bmatrix}\times M_3\times B_3$	$\begin{bmatrix}\text{SW-MSA},H_1,C_2\\\text{MLP},R_1\end{bmatrix}\times M_4\times B_4$
三分支			$\begin{bmatrix}\text{SW-MSA},H_2,C_3\\\text{MLP},R_2\end{bmatrix}\times M_3\times B_3$	$\begin{bmatrix}\text{SW-MSA},H_2,C_3\\\text{MLP},R_2\end{bmatrix}\times M_4\times B_4$
四分支				$\begin{bmatrix}\text{SW-MSA},H_3,C_4\\\text{MLP},R_3\end{bmatrix}\times M_4\times B_4$

1.2　融合 Swin block 提高模型精度

随着 Transformer 模型在自然语言处理任务中大获成功，各种融合 Transformer 模型的方法出现在计算机视觉领域中[10]。2020 年 Dosovitskiy 等[11]提出 ViT(Vision Transformer)模型，首次使用一种完全基于自注意力机制的 Transformer 模型用于图像分类，之后越来越多的研究者提出许多 Transformer 的变种模型用于图像视觉领域。Swin transformerm 模型通过基于移位窗口的自注意力具有线性计算复杂度，并在图像识别，密集预测任务，如对象检测和语义分割中表现良好[7]。

与大多数基于 Transformer 的模型不同，Swin transformer 是一种分层架构。为了高效建模，Swin transformer 模型提出了基于窗口的多头自注意力机制(W-MSA)和基于滑动窗口的多头注意力机制(SW-MSA)。Swin Block 是 Swin transformer 模型中的核心部分，其结构如图 2 所示，Swin Block 由 W-MSA、SW-MSA 和多层感知机(MLP)组成，W-MSA 和 SW-MSA 交替执行，并且在每个注意力模块和 MLP 中间应用 LayerNorm(LN)层。在 W-MSA 中，输入要素将被划分为不重叠的窗口，每个窗口包含 $C\times C$ 个像素点(默认设置为 7)，W-MSA 仅在本地窗口内进行自注意力操作。SW-MSA 解决了 W-MSA 中窗口之间缺乏有效信息交互的问题，采用循环位移的方式来将特征映射到多个相邻的子窗口中，同时保持窗口的数目和规则划分相同，能在相邻区域间建立空间依赖关系，提高模型表达能力。

图像的上下文信息对于提高语义分割准确性是至关重要的，长过程的语义信息可以作为鉴别辅助，从而允许模型不仅仅依赖于图像的局部信息。由于本文的数据集不是很大，且手术器械受环境影响容易泛白、形状和尺寸变化剧烈，故提出使用 Swin Block 中滑动窗口多头自注意力模块代替 HRNet 中原来的卷积层模块。

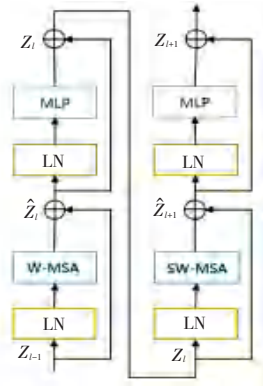

图 2　Swin Block 的结构

Fig. 2　Structure of Swin Block

2　实验

2.1　数据集

为了验证本文所提出的基于多头注意力机制的手术器械分割方法的有效性和鲁棒性，在公共数据集 Endovis2017 和私有数据集上进行对比分割实验。

Endovis2017 数据集：该数据集是来源于 2017 年内窥镜视觉挑战赛的机器人仪器分割子挑战部分，是手术器械分割领域最知名的数据集之一[12]。Endovis2017 数据集的图片包含 7 种大小不同的手术器械，由 1 800 张图像分辨率为 1 280×1 204 的手术图像组成，其中 1 080 张图片用于训练，360 张用于验证，360 张用于测试。

私有数据集：是在消化内镜手术机器人平台收集的，该平台基于国家重点研发计划资助项目(2018YFB1307700)开发。在做离体实验过程中，共选取差异性比较大的 2 344 帧，以便于更好的代表

整个手术过程。每一帧都包含电刀和电夹两种手术器械,图片分辨率为 1 312×1 020,将采集的图片利用标注工具(labelme 软件)进行注释。

本文基于 Ubuntu22.04 系统,使用深度学习框架 Pytorch1.12.0、Python3.8 b 版本以及高效便捷的框架 Ptorch-Lighting1.7.7 进行实验。每个模型在不同的数据上训练 100 个 epoch,每 2 个 epoch 验证一次,保存下 5 次最好的权重参数,再用这 5 次权重参数分别来预测手术器械图片,得到各种指标取平均数。在训练过程中,改变学习率来防止过拟合。本实验的线程参数为 16,初始学习速率为 0.000 1。损失函数使用 FocalLoss,可以减少容易分类的样本的权重,增加难以分类的样本的权重[13]。损失函数如公式(1):

$$FL_{(p_t)} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

其中,$p \in [0,1]$ 是模型对标记类的估计概率;γ 是可调聚焦参数;α 是平衡参数。

本文将 γ 设为 2,α 设为 0.25。

2.2 评价指标

为了从多个角度检测本文所提出的基于多头注意力机制的手术器械分割方法的各方面的性能,本文使用语义分割中常用的评价指标即精确率(*Precision*)、召回率(*Recall*)、*F*1 分数(*F*1 - Score)、平均交并比(*MIOU*)。

精确率表示模型识别为正的样本中,真正为正类的样本所占的比例, 如公式(2)。精确率越高说明模型的效果越好。

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

其中,*TP* 表示真正类的类别数目,*FP* 表示假正性类别数目。

召回率表示模型正确识别出为正类的样本的数量占总的正类样本数量的比值,如公式(3)。召回率越高,说明有更多的正类样本被模型预测正确,模型的效果越好。

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

其中,*FN* 表示假负类类别数目。

*F*1 分数被定义为精确率和召回率的调和平均数,在尽可能的提高精确率和召回率的同时,降低两者之间的差异,*F*1 - Score 的计算公式(4):

$$F1 - \text{Score} = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

平均交并比为数据集中的每一个类的真实掩码和预测掩码的交集和并集之比的平均值,如公式(5):

$$MIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN + FP + TP} \tag{5}$$

其中,k 表示分割结果的类别数。

2.3 实验结果与分析

为了验证本文所提出的基于多头注意力机制的手术器械分割方法的有效性,在公共数据集 Endovis2017 和私有数据集上进行实验与分析,包括定量和定性分析,对比实验的方法有 Unet 方法[14]、TransUNet 方法[15]、GCNet 方法[16]、HRnet 方法。定量分析提供了客观的数据支持和结论验证,而定性分析则能够帮助我们深入理解现象背后的意义和机制,从而更好地解释和丰富定量分析的结果。

2.3.1 定量分析

在公共数据集 Endovis2017 上的实验结果见表 2。由表 2 可以看出,本文所提出的方法在综合性能上表现最好。与基线模型 HRNet 相比,平均交并比提高了 0.73%,平均 *F*1 分数提高了 0.35%,平均精确率提高了 0.17%,平均召回率下降了0.24%,对局部特征图的有效处理,总体性能得到提升。与其他方法相比,ST-HRNet 方法在各项指标的数据均为最高,分割性能最好。

表 2 Endovis2017 数据集上的性能比较

Table 2 Performance comparison on Endovis2017 data set

方法	平均交并比/%	平均 *F*1 分数/%	平均精确率/%	平均召回率/%
Unet	95.11	97.45	97.34	97.57
TransUNet	94.83	97.30	97.34	97.26
GCNet	94.25	96.98	96.19	97.81
HRnet	95.13	97.51	97.45	**98.34**
ST-HRNet	**95.86**	**97.86**	**97.62**	98.10

为了验证在不同组织环境和手术器械的场景中 ST-HRNet 方法是否具有较高的普适性,本文在私有数据集上也做了对比实验,见表 3。由表 3 可知,在私有数据集上 ST-HRNet 方法各项性能均为最优,与经典医学分割方法 Unet 相比,平均交并比提高了 0.72%,平均 *F*1 分数提高了 0.37%,平均精确率和召回率分别提高了 0.51%和 0.24%;与基线方法 HRNet 相比,平均交并比提高了 0.67%,调和平均值提高了 0.34%,平均精确率和召回率分别提高

了 0.59%和 0.25%，说明本文所提出的方法具有鲁棒性，能够在复杂的环境中有效的分割手术器械。

表 3　私有数据集上的性能比较

Table 3　Performance comparison on private data set　%

方法	平均交并比	平均 *F*1 分数	平均精确率	平均召回率
Unet	96.45	98.18	98.11	98.25
TransUNet	93.72	96.70	97.10	96.35
GCnet	95.23	97.52	97.89	97.17
HRnet	96.50	98.19	98.03	98.24
ST-HRNet	**97.17**	**98.55**	**98.62**	**98.49**

2.3.2　定性分析

本文在 Endovis2017 数据集和私有手术器械数据集的测试集中分别提取 3 张图片进行可视化结果展示，可视化结果如图 3 和图 4 所示。

从可视化结果展示中可以看出，由于手术器械移动造成了边缘界限模糊，本文所提方法能够完整细致地分割出手术器械，而其余模型则对手术器械分割较粗糙，或者对尖端分割不细致。此外，在不同照明和形状尺寸不一致的图片中可以看出，本文所提方法对手术器械尖端的分割更为精确细致，且没有噪音点。与 GCNNet 方法、HRNet 方法、TrandUnet 方法、Unet 方法相比，本文所提出方法分割结果与真实情况最为一致，符合临床手术场景的要求。

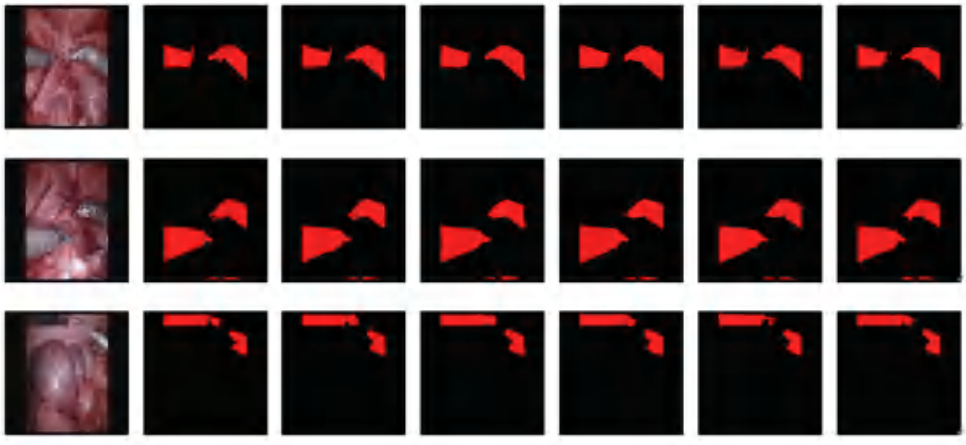

图 3　在公共 Endovis2017 数据集上做对比实验的可视化结果

Fig. 3　Visualization results of comparison experiment on the public Endovis2017 data set

图 4　在私有数据集上做对比实验的可视化结果

Fig. 4　Visualization results of comparison experiment on the private data set

通过定量和定性的分析基于多头注意力机制的手术器械分割方法的有效性。

3　结束语

针对手术器械分割领域中尖端部位分割不细致等问题，本文基于 HRNet 网络结构进行改进，提出了一种提高精度分割的有效方案。基于 HRNet 结构构建 4 条并行且分辨率不同的分支网络，将不同分辨率特征图之间进行多尺度融合，使其高分辨率特征图也具有丰富的语义信息，保证细节不被丢失；

对第二、三、四阶段的后 3 个分支的卷积层替换成 Swin Block,采用滑动窗口多头注意力机制扩大感受野,考虑全局信息,能够把形状和尺寸变化剧烈的手术器械进行完整地分割。在公共数据集 Endovis2017 和私有数据集上都做了验证,无论定量和定性分析,皆验证了本文所提出方法的有效性,为相关手术器械分割领域的研究奠定良好的基础。

参考文献

[1] 李耀仟,李才子,刘瑞强,等.面向手术器械语义分割的半监督时空 Transformer 网络[J].软件学报,2022,33(4):1501-1515.

[2] ESTEVA A, CHOU K, YEUNG S, et al. Deep learning-enabled medical computer vision[J]. NPJ Digital Medicine, 2021, 4(1): 5.

[3] ZHEN L N, BIAN G B, WANG G A, et al. Pyramid attention aggregation network for semantic segmentation of surgical instruments[C]// Proceedings of Assoc Advancement Artificial Intelligence. AAAI Conference on Artificial Intelligence. CA, USA: AAAI, 2020:11782-11790.

[4] YANG L, GU Y, BIAN G, et al. An attention-guided network for surgical instrument segmentation from endoscopic images[J]. Computers in Biology and Medicine, 2022, 151: 106216.

[5] YU L, WANG P, YU X, et al. A holistically-nested U-Net: Surgical instrument segmentation based on convolutional neural network[J]. Journal of Digital Imaging, 2020, 33: 341-347.

[6] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation [C]// Proceedings of Institute of Electrical and Electronics Engineers. Seoul, Republic of Korea: IEEE, 2019: 5693-5703.

[7] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 10012-10022.

[8] DAI Z, LIU H, LE Q V, et al. Coatnet: Marrying convolution and attention for all data sizes[J]. Advances in Neural Information Processing Systems, 2021, 34: 3965-3977.

[9] XIAO T, SINGH M, MINTUN E, et al. Early convolutions help transformers see better[J]. Advances in Neural Information Processing Systems, 2021, 34: 30392-30400.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J].arXiv preprint arXiv:2010.11929, 2020.

[12] ALLAN M, SHVETS A, KURMANN T, et al. 2017 robotic instrument segmentation challenge[J].arXiv preprint arXiv:1902.06426, 2019.

[13] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of Institute of Electrical and Electronics Engineers. Proceedings of the IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 2980-2988.

[14] RONNEBERGE O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[J]. Medical Image Computing and Computer Assisted Intervention, 2015, 9351(65): 234-241.

[15] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[16] CAO Y, XU J, LIN S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. New York, USA: IEEE, 2019: 1-3.