2025 年 第 25 卷 第 21 期
2025, 25(21):08973-07

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404319

引用格式:宋政, 孟晓亮, 张立晔, 等. 基于改进 Swin Transformer 的双编码网络手术器械分割方法[J]. 科学技术与工程, 2025, 25(21): 8973-8979.

Song Zheng, Meng Xiaoliang, Zhang Liye, et al. Surgical instrument segmentation method of double encoding network based on improved Swin Transformer[J]. Science Technology and Engineering, 2025, 25(21): 8973-8979.

基于改进 Swin Transformer 的双编码网络手术器械分割方法

宋政, 孟晓亮*, 张立晔, 王晓雨, 韩储屹

(山东理工大学计算机科学与技术学院, 淄博 255000)

摘 要 为实现手术器械的准确分割,提出一种基于改进 Swin Transformer 的双编码网络手术器械分割方法,利用 Swin Transformer 和卷积神经网络(convolutional neural network,CNN)不同的编码优势,有效捕获图像特征的全局语义信息和局部细节信息,提高模型的表达能力。为尽可能弥补下采样过程中的特征细节损失,设计了一个多尺度分辨率特征金字塔池化(multi-resolution feature pyramid pooling,MFPP)模块,通过结合不同维度特征,获取更多尺度的上下文信息,增强局部细节信息表达。最后,在跳跃连接中增加一个坐标注意力模块,将目标位置信息与通道信息相融合,对手术器械目标进行精确感知。实验结果表明,所提方法在手术器械二值分割和部分分割中,均获得了更为精确的分割结果,进一步验证了所提方法的有效性和准确性。

关键词 手术器械; 语义分割; Swin Transformer; 深度学习; 注意力机制

中图法分类号 TP391; **文献标志码** A

Surgical Instrument Segmentation Method of Double Encoding Network Based on Improved Swin Transformer

SONG Zheng, MENG Xiao-liang*, ZHANG Li-ye, WANG Xiao-yu, HAN Chu-qi

(School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China)

[**Abstract**] In order to achieve accurate segmentation of surgical instruments, a dual-encoding network surgical instrument segmentation method was proposed based on improved Swin Transformer. By taking advantage of different coding advantages of Swin Transformer and convolutional neural network(CNN), the global semantic information and local details of image features can be effectively captured to improve the expression ability of the model. To compensate for the loss of feature details during the downsampling process as much as possible, the multi-resolution feature pyramid pooling(MFPP) block was constructed to obtain more scale context information by combining different dimensional features and enhance the expression of local detail information. Finally, a coordinate attention block was added in the skip connection to fuse target position information with feature information for precise perception of the surgical instrument targets. The experimental results show that the proposed method achieves more accurate segmentation results in both binary and parts segmentation of surgical instruments, further verifying the effectiveness and accuracy of the proposed method.

[**Keywords**] surgical instruments; semantic segmentation; Swin Transformer; deep learning; attention mechanism

随着机器人技术的发展,外科辅助手术机器人已成为微创手术领域的重要研究方向[1]。与传统的手术相比,微创手术不仅具有切口小、恢复快等优点,还可以对手术过程中不必要和不安全的操作提供实时告警,进一步提高手术的安全性。目前,微创手术已被广泛应用于各种外科手术中[2-3]。对手术器械的准确分割是外科辅助手术机器人的重要组成部分,在手术器械跟踪、姿态估计和增强现实中也起着关键作用。因此,准确的手术器械分割在微创手术中具有重要的研究意义。

近年来,卷积神经网络(convolutional neural network,CNN)在图像检测、分类、语义分割等领域都取得了优异的成绩[4-6]。在微创手术中内窥镜视角下的器械分割不同于一般场景的语义分割,手术器械

收稿日期: 2024-06-11 **修订日期**: 2025-04-09

基金项目: 国家自然科学基金(62001272)

第一作者: 宋政(1997—),男,汉族,山东菏泽人,硕士研究生。研究方向:计算机视觉。E-mail:22505040006@stumail.sdut.edu.cn。

* **通信作者**: 孟晓亮(1988—),男,汉族,山东潍坊人,博士,讲师。研究方向:视觉检测与图像处理、深度学习。E-mail:xiaoliang@sdut.edu.cn。

反光引起的图像亮度变化、相机镜头雾化造成的视野遮挡以及背景软组织的动态多变性等因素都会影响到手术器械的分割准确度[7]。此外,内窥镜图像还存在类别不平衡的问题,手术器械在图像中一般占比较小,其背景像素数量通常要远远大于手术器械的像素数量。针对手术器械图像的分割问题,Mahmood 等[8]提出一种双流残差密集网络,采用稠密连接网络 DenseNet 和空洞空间金字塔池化强化语义信息提取,但 DenseNet 进行多次拼接,每块 DenseBlock 需要接收前面所有层的信息,导致过多地占用内存和计算资源。Shen 等[9]提出一种分支聚合注意力网络 BAANet,在编码阶段通过多尺度分支特征的聚合来提高特征语义信息的表达,而在解码阶段利用双分支注意力机制感知的语义信息和临界特征图对手术器械进行精准分割。邓健志等[10]提出一种结合拆分注意力跨通道特征融合的分割网络,通过拆分注意力模块使不同拆分组的特征图被赋予不同权重比,关注特征通道间的重要特征,但忽略了对全局特征信息的获取。Zhou 等[11]提出一种含有文本提示的新型混合机制分割方法,以预训练的图像编码器和文本编码器作为主干网络,使用文本提示掩膜解码器获取器械目标的文本预测信息。但是,在下采样过程中上下文信息的丢失会随着下采样层数的增加而变大,导致大量细节信息提取不充分,造成手术器械分割边缘模糊。而在自然语言处理领域获得较为成功的 Transformer 网络[12],因其擅长处理全局间的依赖关系,在计算机视觉领域也得到广泛的应用和发展。Liu 等[13]提出的 Swin Transformer 架构,在非重叠的局部窗口进行自注意力计算,减少计算开销,并使用滑动窗口机制关联多个相邻窗口,实现夸窗口间的特征信息交互。Cao 等[14]提出一种基于 Swin Transformer 搭建的 U 形分割网络,其中补丁合并层和 Swin Transformer 负责下采样和增加维度,用于语义特征的学习,在解码部分使用补丁扩展层和 Swin Transformer 进行上采样。虽然 Transformer 网络在捕捉长距离依赖关系方面表现出良好的性能,但它对特征局部细节信息的捕捉不如 CNN 那么出色,且需要使用大量数据来提高 Transformer 网络的泛化能力。

为实现对手术器械的准确分割,现提出一种改进 Swin Transformer 双编码网络结构的手术器械分割方法。利用 CNN 对局部细节特征进行学习的同时,使用 Swin Transformer 对全局上下文信息进行捕获,有效提高模型对特征的学习能力。针对下采样过程中存在的空间细节信息丢失问题,构造一个多尺度分辨率特征金字塔池化(multi-resolution feature pyramid pooling,MFPP)模块以提取特征的多尺度上下文信息,增强网络模型的鲁棒性。同时,在跳跃连接过程中设置一个坐标注意力(coordinate attention,CA)[15]模块,以有效捕获目标位置信息,精确定位手术器械在空间中的位置。

1 方法概述

1.1 Swin Transformer 编码分支

传统 Transformer 计算图像全局每个图像块(Tokens)间的依赖关系,容易产生较高的冗余和较大的计算量,而 Swin Transformer 将图像分割为若干个不重叠窗口,在各自窗口内进行图像块间的自注意力计算,使得计算量从平方量级变成线性量级,降低计算开销,并使用滑动窗口机制,以增加特征间的信息交换,结构如图 1 所示。

Swin Transformer 由两个连续的 Swin Transformer 块构成,由层归一化(layer normalization,LN)、多层感知机(multi-layer perception,MLP)、窗口多头自注意力(windows multi-head self-attention,W-MSA)和滑动窗口多头自注意力(shifted windows multi-head self-attention,SW-MSA)组成。将 Swin Transformer 初始维度设置为 96,窗口大小设置为 8,W-MSA 和 SW-MSA 中相应的 head 数为 3、6、12、24。

对于 Swin Transformer 模块,第 l 层 W-MSA、SW-MSA 的输出分别为 $\boldsymbol{Z}^l$、$\boldsymbol{Z}^{l+1}$ 的编码向量,具体计算方式如下。

$$\boldsymbol{Z}^l = \text{W-MSA}[\text{LN}(\boldsymbol{Z}^{l-1})] + \boldsymbol{Z}^{l-1} \tag{1}$$

$$\boldsymbol{Z}^l = \text{MLP}[\text{LN}(\hat{\boldsymbol{Z}}^l)] + \hat{\boldsymbol{Z}}^l \tag{2}$$

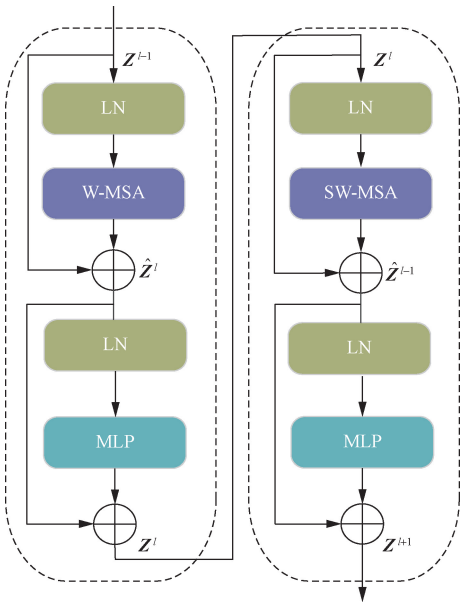


图 1 Swin Transformer 模块结构

Fig. 1 Swin Transformer blocks structure

$$\hat{\boldsymbol{Z}}^{l+1} = \mathrm{SW-MSA}[\mathrm{LN}(\boldsymbol{Z}^{l})] + \boldsymbol{Z}^{l} \quad (3)$$

$$\boldsymbol{Z}^{l+1} = \mathrm{MLP}[\mathrm{LN}(\hat{\boldsymbol{Z}}^{l+1})] + \hat{\boldsymbol{Z}}^{l+1} \quad (4)$$

1.2　CNN 编码分支

在手术器械分割任务中,随着网络深度的增加,镜头雾化、遮挡等问题都会影响网络对特征的学习。而 ResNet 网络能够有效增强特征的表达,避免梯度消失。因此,使用 ResNet34 的编码块作为 CNN 编码分支提取局部细节特征,从上至下编码部分每层残差单元的特征数量分别为 64、128、256 和 512。

1.3　双编码网络

本文算法的整体双编码网络结构如图 2 所示。对于一幅给定图像 $\boldsymbol{X} \in \mathbf{R}^{H\times W\times C}$,在提取其图像浅层语义信息后,输入由 Swin Transformer 和 CNN 构成的双编码模块中进行编码,同时在模型底部构造一个 MFPP 模块,以融合不同尺度大小的特征图,以获取丰富的上下文信息。对 CNN 生成的特征图和 Swin Transformer 产生的向量进行拼接后,传入 CA 注意力模块,将位置信息融入特征通道中,最后通过逐步上采样恢复特征图大小,获得准确的分割结果。

1.4　MFPP 模块

网络模型随着下采样层数的增加,特征信息变得越来越抽象的同时,也会丢失许多局部细节特征,导致网络模型对特征的语义信息学习不充分,影响手术器械的分割精度。为更好地处理不同尺度特征,充分提取特征的多尺度语义信息,加强不同尺度特征信息交流。基于 DeepLabV2[16] 模型中的空洞空间池化模块,构造一个多尺度分辨率特征金字塔池化(MFPP)模块,如图 3 所示。

首先,通过卷积操作将不同尺度的特征进行融合拼接。然后,输入一个 $\mathrm{Conv}_{1\times1}$ 卷积和 4 个不同感受野大小的 $\mathrm{Conv}_{3\times3}$ 卷积中,其中 $\mathrm{Conv}_{3\times3}$ 卷积的空洞率分别为 1、6、12 和 18。通过捕获不同感受域中的上下文语义信息,提高模型的表达能力和泛化能力。面对不同空洞率的特征,为学习不同通道间的通道比重,进一步引入 SE(squeeze-and-excitation)注意力

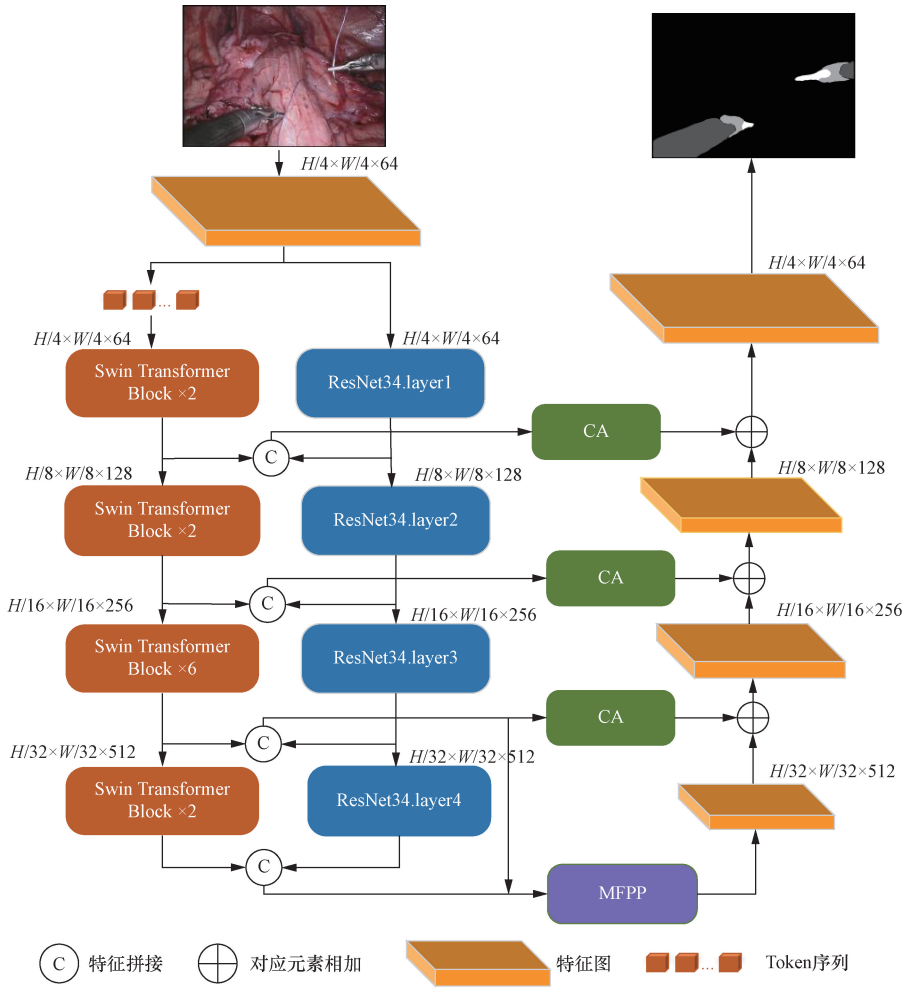
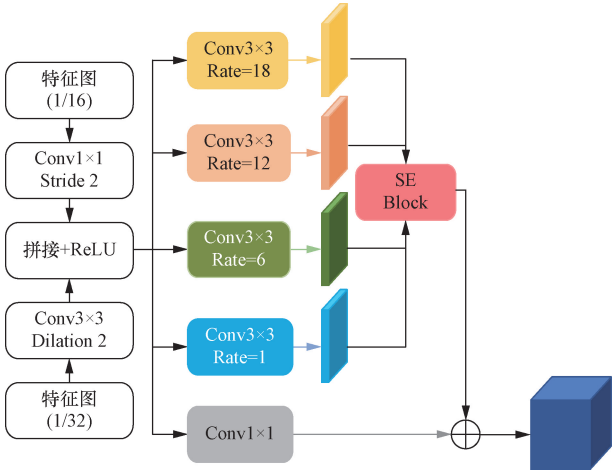


图 2　本文网络模型结构

Fig. 2　Network model structure of this paper



Stride 为步长;Rate 为速率;Dilation(也称为空洞卷积)为在卷积操作中插入“空洞”或零值,以增加卷积核的感受野(receptive field)而不增加参数数量

图 3 多尺度分辨率特征金字塔池化模块

Fig. 3 Multi-resolution feature pyramid pooling block

模块[17],通过动态地调整不同特征通道间的关系,以增加感兴趣通道比重。最后,将经过 SE 注意力模块调整的特征和经过 $\mathrm{Conv}_{1\times1}$ 卷积的特征进行拼接,具体公式如下。

$$\boldsymbol{X}_3 = \mathrm{Concate}[\boldsymbol{X}_1, \mathrm{Conv}_{2\times2}(\boldsymbol{X}_2)] \tag{5}$$

$$\boldsymbol{Y}_1 = f_{\mathrm{se}}[\mathrm{Concate}(\mathrm{Conv}_{\mathrm{dr}}(\boldsymbol{X}_3, i)] \tag{6}$$

$$\boldsymbol{Y}_2 = \boldsymbol{Y}_1 + \mathrm{Conv}_{1\times1}(\boldsymbol{X}_3) \tag{7}$$

式中:$\boldsymbol{X}_1$、$\boldsymbol{X}_2$、$\boldsymbol{X}_3$ 为不同尺度的特征向量;Concate( )为特征拼接;$\mathrm{Conv}_{\mathrm{dr}}$( )为空洞卷积操作;$i$ 为空洞率,分别取 1、6、12 和 18;f_{se}( )为 SE 模块的功能函数;$\boldsymbol{Y}_1$ 为 f_{se}( )功能函数调整通道权重后的特征向量;$\boldsymbol{Y}_2$ 为最后输出的特征向量。

1.5 空间坐标注意力模块

为高效融合来自 ResNet34 和 Swin Transformer 的分支特征,引入坐标注意力(coordinate attention,CA)模块。常见的注意力机制只着重于通道注意力的提升,而忽略位置信息的获取。坐标注意力模块对特征进行两个方向上的全局平均池化,一个在水平空间方向上捕捉长距离依赖关系,另一个在垂直空间方向上保存精确的位置信息,形成一对方向感知和位置敏感的特征图,经过拼接、卷积和标准化等操作,增强对手术器械区域的特征表示。

2 实验与结果分析

2.1 实验数据集

在 EndoVis 2017 公共数据集上测试所提出的模型,该数据集包含 8 个不同视频序列,每个视频序列含有 225 帧,共 1 800 帧。手术器械每个部分又可以分为轴(shaft)、腕(wrist)和末端尖端(end tip),且已标记在每幅图像中。

2.2 实验配置和训练参数

本文模型训练使用的 GPU 型号为 NVIDIA RTX A6000(48 GB 内存),CUDA 版本为 11.6。在进行模型训练之前,通过水平翻转和垂直翻转对数据集进行扩充,并对图像的每个通道进行归一化处理,数据集中每幅图像分辨率为 1 280 × 1 024。Adam 优化器的初始学习率设置为 1×10^{-5},迭代次数和批处理大小分别设置为 160 和 8。

2.3 模型衡量指标

为评估所提方法的分割性能,采用 Dice 和 mIoU 进行定量评估。Dice 是一个集合度量函数,通常用于计算两个样本的相似度,其值范围为[0,1]。mIoU 计算真实值和预测值两集合的交集与并集之比。Dice 和 mIoU 得分越高,分割效果越好,具体计算公式如下。

$$\mathrm{Dice} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{2\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+2\mathrm{TP}} \tag{8}$$

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}} \tag{9}$$

式中:k 为像素点数量;TP、FP、FN 分别为真阳性、假阳性和假阴性的数量。

2.4 分割结果与分析

首先,将所提方法与其他先进的分割方法进行手术器械二值分割实验对比,结果如表 1 所示。

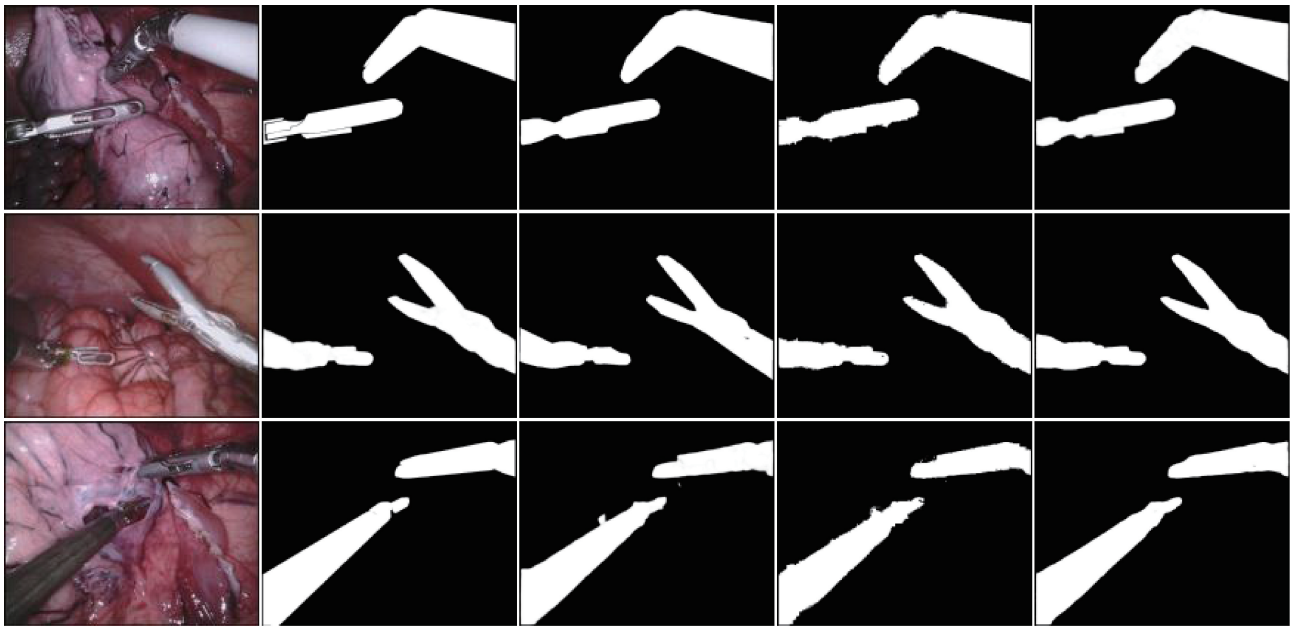
从表 1 可知,本文方法在二值分割中取得了良好的分割效果,Dice 系数为 89.60%,mIoU 为 83.15%,器械的分割精度得到有效提升。与基于 Swin Transformer 构建的 Swin-UNet 相比,Dice 系数和 mIoU 分别提升 2.21%、3.47%;比 U-NetPlus 分割方法分别高出 1.33% 和 1.83%。相较于表 1 中其他分割方法,本文方法的手术器械二值分割准确度得到了有效改善。

为更直观地呈现本文方法在该数据集二值分割效果,选取测试集上的部分图像为样本,给出不同二值分割方法的可视化预测图,如图 4 所示。可以看出,相较于其他二值分割方法出现的像素预测错

表 1 不同方法在 EndoVis 2017 数据集上的二值分割结果

Table 1 Binary segmentation results of different methods on the EndoVis 2017 datasets

二值分割	Dice/%	mIoU/%
Swin-UNet[14]	87.39	79.68
U-Net[18]	84.37	75.44
Unet + +[19]	87.09	78.91
TernauNet[20]	88.07	81.14
U-NetPlus[21]	88.27	81.32
本文方法	89.60	83.15



(a) 原始图像　(b) 真实值　(c) TernauNet　(d) Swin-UNet　(e) 本文方法

图 4　不同方法在 EndoVis 2017 数据集上二值分割可视化结果比较

Fig. 4　Comparison of visualization results of binary segmentation using different methods on the EndoVis 2017 datasets

误和手术器械边缘粗糙问题,本文方法的预测图有着更为平滑的整体手术器械目标,像素预测错误的问题明显减少,手术器械边缘更加精确。

在有限的内窥镜空间下,对手术器械的轴、腕、尖端末端部位的精确分割,可以有效确定器械各部位在图像中的位置,为医师实施缝合、切割等操作提供实时指导,实现手术器械工具的跟踪。其次,手术器械部分分割实验相较于手术器械二值分割实验而言,是通过不同标签数据集分别训练得到。在手术器械部分分割实验时模型的预测类别为 4(轴、腕、尖端末端、背景),通过 softmax 函数取对数得到不同类别的预测结果,softmax 函数计算公式为

$$\mathrm{softmax}(x_i) = \frac{\mathrm{e}^{x_i}}{\sum_{i=1}^{N} \mathrm{e}^{x_i}} \tag{10}$$

式(10)中:N 为类别数;x_i 为第 i 个特征分量。

在该数据集上进行的手术器械部分分割实验与其他先进方法比较的结果如表 2 所示。

由表 2 可知,本文方法在手术器械部分分割中取得了良好的分割效果,Dice 系数和 mIoU 分别为 75.64% 和 65.03%,比 TernauNet 分别高出 1.39% 和 2.80%。与 Swin-UNet 相比,Dice 系数和 mIoU 分别提升 7.58% 和 8.24%。本文方法相较于其他分割方法,在部分分割试验中也取得了较好的分割效果,手术器械部分分割的精度得到进一步的提升。

表 2　不同方法在 EndoVis 2017 数据集上的部分分割结果

Table 2　Parts segmentation results of different methods on the EndoVis 2017 datasets

部分分割	Dice/%	mIoU/%
Swin-UNet[14]	68.06	56.79
U-Net[18]	60.75	48.41
U-Net + +[19]	53.69	42.52
TernauNet[20]	74.25	62.23
PlainNet[22]	73.53	64.73
PaI-Net[23]	73.02	61.45
本文方法	75.64	65.03

此外,将本文方法与 Swin-UNet 和 TernauNet 手术器械分割方法每个部位的 Dice 系数、mIoU 及 Hausdorff 距离进行详细比较,Hausdorff 距离表示两空间子集的距离,衡量预测边界与真实边界的重合程度,值越小,代表相似度越高。结果如表 3 所示。

表 3　手术器械不同部位的分割结果比较

Table 3　Comparison of segmentation results for different parts of surgical instruments

方法	Dice/%			mIoU/%			Hausdorff 距离		
	Shaft	Wrist	End tip	Shaft	Wrist	End tip	Shaft	Wrist	End tip
Swin-UNet[14]	68.91	74.52	60.74	56.45	66.29	47.62	8.55	10.20	10.60
TernauNet[20]	74.20	80.63	67.92	60.14	73.08	53.47	8.13	8.70	10.08
本文方法	74.56	80.88	71.48	63.03	73.46	58.61	7.88	8.54	9.76

由表3可以看出,相较于Swin-UNet和TernauNet方法,在手术器械复杂的末端尖端(end tip)部位,本文方法与两者相比Dice系数分别提高10.74%和3.56%,mIoU分别提高10.99%和5.14%,Hausdorff距离分别降低0.84和0.32,手术器械其他部位的分割精度也得到相应改善。为更直观地呈现部分分割的效果,给出了不同方法得到的部分分割结果,如图5所示。

本文方法的部分分割预测图中,不同手术器械部位间的分界边际更加清晰、明显,边际间像素值分类错误的问题得到有效控制,手术器械的不同部位得到准确定位,轮廓更加完整、平滑,产生更加精确的边缘细节。

2.5 消融实验

为验证Swin Transformer和CNN双编码结构、MFPP模块和CA模块的有效性。以双编码结构为基线模型,设置4组消融实验进行对比论证,结果如表4所示。首先,基线模型的Dice、mIoU和Hausdorff距离得分依次为73.19%、62.36%和8.96,说明双编码结构对手术器械的分割具有明显的提升作用。其次,基线模型+CA和基线模型+MFPP在基线模型的基础上分割性能均得到了进一步的优化。可见,坐标注意力模块在水平和垂直两个方向上能够有效捕获长距离依赖关系和位置信息,帮助模型精确定位和识别手术器械。MFPP模块通过引入更多尺度的特征图进行多尺度特征提取,从而获取更多的语义信息,以优化特征在模型解码过程中细节语义信息丢失的问题。最后,基线模型+MFPP+CA即本文方法,Dice和mIoU分别为75.64%和65.03%,在消融实验中取得最高得分,Hausdorff距离得分为8.73,也为消融实验的最优得分,证明了本文方法的双编码结构、坐标注意力模块和MFPP模块的有效性。

3 结论

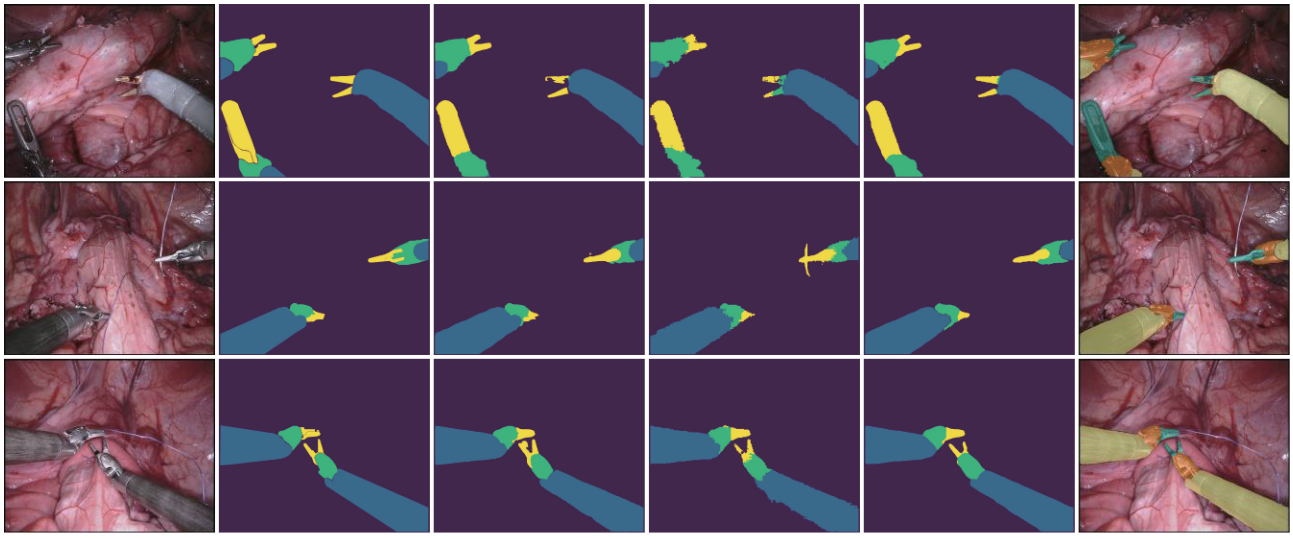
提出一种由Swin Transformer和CNN构成的双编码网络手术器械分割方法。基于Swin Transformer强大的自注意力机制,弥补CNN捕获全局上下文信息不足的问题,通过设计的MFPP模块,有效地融合不同编码层的特征图,提取更多尺度的语义信息。坐标注意力模块解决以往注意力模块只关注通道信息,而忽略相关位置信息的问题,增强对手术

表4 不同模块的有效性分析

Table 4 Effectiveness analysis of different blocks

模型	CA	MFPP	Dice/%	mIoU/%	Hausdorff距离
基线模型	—	—	73.19	62.36	8.96
基线模型+CA	√	—	74.68	63.86	8.84
基线模型+MFPP	—	√	75.36	64.73	8.78
基线模型+MFPP+CA	√	√	75.64	65.03	8.73

注:√表示采用该模块;—表示未采用该模块。



(a)原始图像 (b)真实值 (c)TernauNet (d)Swin-UNet (e)本文方法 (f)本文方法+原始图像

黄色代表手术器械的末端尖端部分;绿色为手术器械的腕部;深蓝色表示为手术器械的轴部

图5 不同方法在EndoVis 2017数据集上的部分分割可视化结果比较

Fig. 5 Comparison of parts segmentation results of different methods on the EndoVis 2017 datasets

器械目标的感知。本文方法在 EndoVis 2017 数据集上进行的二值分割,Dice 系数为 89.60%,mIoU 为 83.15%,部分分割的 Dice 系数为 75.64%,mIoU 为 65.03%,均表现出良好的分割性能,并有效抑制手术器械边缘粗糙的问题,手术器械分割的准确度得到进一步提升。

参 考 文 献

[1] 刘哲, 石钰, 林延带, 等. 智能医学的现状与未来[J]. 科学通报, 2023, 68(10): 1165-1181.
Liu Zhe, Shi Yu, Lin Yandai, et al. The current status and future of intelligent medicine[J]. Chinese Science Bulletin, 2023, 68(10): 1165-1181.

[2] Zhao Z, Jin Y, Heng P A. Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery[C]//2022 International Conference on Robotics and Automation (ICRA). Harrisburg: IEEE, 2022: 11186-11193.

[3] Hashimoto F, Ote K, Onishi Y. PET image reconstruction incorporating deep image prior and a forward projection model[J]. IEEE Transactions on Radiation and Plasma Medical Sciences, 2022, 6(8): 841-846.

[4] 孟晓亮, 赵吉康, 王晓雨, 等. 基于改进 YOLOv5s 的手术器械检测与分割方法[J]. 液晶与显示, 2023, 38(12): 1698-1706.
Meng Xiaoliang, Zhao Jikang, Wang Xiaoyu, et al. Surgical instrument detection and segmentation method based on improved YOLOv5s[J]. Chinese Journal of Liquid Crystals and Displays, 2023, 38(12): 1698-1706.

[5] 董涛涛, 宋宇博. 基于自适应阈值的循环增长地面点云分割算法[J]. 科学技术与工程, 2024, 24(9): 3526-3532.
Dong Taotao, Song Yubo. A recurrent growth ground point cloud segmentation algorithm based on adaptive thresholding[J]. Science and Engineering, 2024, 24(9): 3526-3532.

[6] 刘畅, 党淑雯, 陈丽. 改进位姿估计环节的 ORB-SLAM 稠密建图算法[J]. 科学技术与工程, 2024, 24(7): 2782-2789.
Liu Chang, Dang Shuwen, Chen Li. ORB-SLAM dense mapping algorithm for improving pose estimation[J]. Science Technology and Engineering, 2024, 24(7): 2782-2789.

[7] Baby B, Thapar D, Chasmai M, et al. From forks to forceps: a new framework for instance segmentation of surgical instruments[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Hawaii: IEEE, 2023: 6180-6190.

[8] Mahmood T, Cho S W, Park K R. DSRD-Net: dual-stream residual dense network for semantic segmentation of instruments in robot-assisted surgery[J]. Expert Systems with Applications, 2022, 202: 117420.

[9] Shen W, Wang Y, Liu M, et al. Branch aggregation attention network for robotic surgical instrument segmentation[J]. IEEE Transactions on Medical Imaging, 2023, 42(11): 3408-3419.

[10] 邓健志, 支佩佩, 张峰铭, 等. 结合拆分注意力特征融合的病理图像分割网络[J]. 科学技术与工程, 2023, 23(7): 2922-2931.
Deng Jianzhi, Zhi Peipei, Zhang Fengming, et al. Pathological image segmentation network combining split attention feature fusion[J]. Science Technology and Engineering, 2023, 23(7): 2922-2931.

[11] Zhou Z, Alabi O, Wei M, et al. Text promptable surgical instrument segmentation with vision-language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 28611-28623.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[13] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 9992-10002.

[14] Cao H, Wang Y, Chen J, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation[C]//European Conference on Computer Vision. Cham: Springer, 2022: 205-218.

[15] Gu R, Wang G, Song T, et al. CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation[J]. IEEE Transactions on Medical Imaging, 2020, 40(2): 699-711.

[16] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[17] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.

[18] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference. Germany: Springer, 2015: 234-241.

[19] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. UNet + +: a nested U-Net architecture for medical image segmentation[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Granada: Springer, 2018: 3-11.

[20] Shvets A A, Rakhlin A, Kalinin A A, et al. Automatic instrument segmentation in robot-assisted surgery using deep learning[C]//2018 17th IEEE international conference on machine learning and applications (ICMLA). Orlando: IEEE, 2018: 624-628.

[21] Hasan S M K, Linte C A. U-NetPlus: a modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images[C]//2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Berlin: IEEE, 2019: 7205-7211.

[22] Jin Y, Cheng K, Dou Q, et al. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2019: 22nd International Conference. Shenzhen: Springer, 2019: 440-448.

[23] Wang X, Wang L, Zhong X, et al. PaI-Net: a modified U-Net of reducing semantic gap for surgical instrument segmentation[J]. IET Image Processing, 2021, 15(12): 2959-2969.