软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(4):1501−1515 [doi: 10.13328/j.cnki.jos.006469]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向手术器械语义分割的半监督时空 Transformer 网络*

李耀仟[1], 李才子[1], 刘瑞强[1], 司伟鑫[1], 金玥明[2], 王平安[1,3]

[1](中国科学院 深圳先进技术研究院, 广东 深圳 518055)

[2](Department of Computer Science, University College London, United Kingdom)

[3](香港中文大学 计算机科学与工程学系, 香港 999077)

通信作者: 司伟鑫, E-mail: wx.si@siat.ac.cn

摘 要: 基于内窥镜的微创手术机器人在临床上的应用日益广泛, 为医生提供内窥镜视频中精准的手术器械分割信息, 对提高医生操作的准确度、改善患者预后有重要意义. 现阶段, 深度学习框架训练手术器械分割模型需要大量精准标注的术中视频数据, 然而视频数据标注成本较高, 在一定程度上限制了深度学习在该任务上的应用. 目前的半监督方法通过预测与插帧, 可以改善稀疏标注视频的时序信息与数据多样性, 从而在有限标注数据下提高分割精度, 但是这些方法在插帧质量与对连续帧时序特征方面存在一定缺陷. 针对此问题, 提出了一种带有时空 Transformer 的半监督分割框架, 该方法可以通过高精度插帧与生成伪标签来提高稀疏标注视频数据集的时序一致性与数据多样性, 在分割网络 bottleneck 位置使用 Transformer 模块, 并利用其自我注意力机制, 从时间与空间两个角度分析全局上下文信息, 增强高级语义特征, 改善分割网络对复杂环境的感知能力, 克服手术视频中各类干扰从而提高分割效果. 提出的半监督时空 Transformer 网络在仅使用 30%带标签数据的情况下, 在 MICCAI 2017 手术器械分割挑战赛数据集上取得了平均 DICE 为 82.42%、平均 IoU 为 72.01%的分割结果, 分别超过现有方法 7.68%与 8.19%, 并且优于全监督方法.

关键词: 视频序列; 时空特征; 手术器械分割; Transformer; 半监督学习

中图法分类号: TP391

中文引用格式: 李耀仟, 李才子, 刘瑞强, 司伟鑫, 金玥明, 王平安. 面向手术器械语义分割的半监督时空 Transformer 网络. 软件学报, 2022, 33(4): 1501−1515. http://www.jos.org.cn/1000-9825/6469.htm

英文引用格式: Li YQ, Li CZ, Liu RQ, Si WX, Jin YM, Heng PA. Semi-supervised Spatiotemporal Transformer Networks for Semantic Segmentation of Surgical Instrument. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1501−1515 (in Chinese). http://www.jos.org.cn/1000-9825/6469.htm

Semi-supervised Spatiotemporal Transformer Networks for Semantic Segmentation of Surgical Instrument

LI Yao-Qian[1], LI Cai-Zi[1], LIU Rui-Qiang[1], SI Wei-Xin[1], JIN Yue-Ming[2], HENG Pheng-Ann[1,3]

[1](Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China)

[2](Department of Computer Science, University College London, United Kingdom)

[3](Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China)

Abstract: With the increasingly wide application of surgical robots in clinical practice, it is of great significance to provide doctors with precise semantic segmentation information of surgical instrument in endoscopic video to improve the clinicians' operation accuracy and patients' prognosis. Training surgical instrument segmentation models requires a large amount of accurately labeled video frames, which

* 基金项目: 深圳市基础研究重点项目(JCYJ20200109110208764, JCYJ20200109110420626); 国家自然科学基金(U1813204, 61802385); 广东省自然科学基金(2021A1515012604)

本文由"面向开放场景的鲁棒机器学习"专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.

收稿时间: 2021-05-10; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-26

limits the application of deep learning in the surgical instrument segmentation task due to the high cost of video data labeling. The current semi-supervised methods enhance the temporal information and data diversity of sparsely labeled videos by predicting and interpolating frames, which can improve the segmentation accuracy with limited labeled data. However, these semi-supervised methods suffer from the drawbacks of frame interpolation quality and temporal feature extraction from sequential frames. To tackle this issue, this study proposes a semi-supervised segmentation framework with spatiotemporal Transformer, which can improve the temporal consistency and data diversity of sparsely labeled video datasets by interpolating frames with high accuracy and generating pseudo-labels. Here the Transformer module is integrated at the bottleneck position of the segmentation network to analyze global contextual information from both temporal and spatial perspectives, enhancing advanced semantic features while improving the perception to complex environments of the segmentation network, which can overcome various types of distractions in surgical videos and thus improve the segmentation effect. The proposed semi-supervised segmentation framework with Transformer achieves an average *DICE* of 82.42% and an average *IOU* of 72.01% on the MICCAI 2017 Surgical Instrument Segmentation Challenge dataset using only 30% labeled data, which exceeds the state-of-the-art method by 7.68% and 8.19%, respectively, and outperforms the fully supervised methods.

Key words: video sequences; spatiotemporal feature; surgical instruments segmentation; Transformer; semi-supervised learning

达芬奇手术机器人是一种可以实施微创外科手术的高级机器人平台, 不但具有高灵活与稳定性, 还对患者创伤小, 具有恢复快、切除彻底、并发症少等特点, 目前已广泛应用于成人与儿童的普通外科、妇产科、心脏手术等多个细分领域[1]. 通过图像分割技术对达芬奇手术机器人导航用的内窥镜实时影像进行分析, 分割标记视频中手术器械的位置, 对帮助医生在手术操作过程中定位器械、提高手术操作精准度以及参与器械控制等下游任务都有重要意义[2,3].

近年来, 基于编码器-解码器架构的卷积神经网络(CNN)被应用到包括手术视频分割在内的一系列图像分割任务中, 该架构使用跳跃连接将低级特征与高级特征相结合, 实现像素级别的精确定位, 上采样中大量的特征通道可以将上下文信息传播至更高分辨率的层[4]. 该结构在许多分割竞赛中取得了出色的效果[5], 但由于卷积运算具有局部性, 因此在处理长距离关系依赖时存在局限性. 此外, 精确标注数据的不足也是限制分割算法表现的重要因素之一, 现有的高精度手术器械视频分割方法往往需要依赖大量的标注视频数据集, 但这些标注的数据需要经验丰富的专业人员逐帧逐像素进行标注, 耗费时力[6,7].

为了解决标注数据不足的问题, 相关研究证明了: 采用光流辅助标签向无标注图像迁移, 有助于半监督分割[7−10]. 科研人员基于此提出一种基于光流预测及插帧的半监督方法, 来为带有少量标签的视频数据集自动标注伪标签, 增加训练集样本数量. 同时, 插帧可以平滑帧间的运动, 增强连续帧之间的时序一致性, 并通过增加带标签数据来提高数据集的数据多样性, 能有效提高分割网络的分割效果[9]. 但该方法使用光流进行插帧, 插帧之后, 图片中的手术器械外形出现扭曲、变色等情况, 使连续帧之间相同的对象存在不同特征, 对分割网络学习手术器械外形特征造成负面作用[2].

在手术器械视频分割任务中, 对连续帧时序特征提取是提高分割准确率的一种重要策略, 现阶段, 这一策略主要通过 ConvLSTM 或 ConvGRU 模块实现[11,12]. 具体来说, 就是在分割网络的 bottleneck 处中使用相关模块对连续多帧的高级特征进行时序信息建模, 提取多帧间时序上的依赖关系. 然而, LSTM 的长距离遗忘特性无法有效保持视频中的长距离依赖性, 在计算视频时序特征, 特别是对同一时间段内距离较远两帧之间的长期依赖关系时存在局限性[13], 这导致在使用 ConvLSTM 模块处理手术器械视频时, 分割网络无法完整地学习到手术器械的运动特征, 从而影响最后的分割结果. 因此, 克服距离对时序信息的影响, 从全局角度对连续帧时间信息建模, 是提高时序特征提取的关键.

针对现有半监督方法在手术器械术中视频分割任务上插帧质量不高和 ConvLSTM 无法有效处理时序信息的缺陷, 本文提出了改进的插帧机制, 并引入基于全局注意力的 Transformer 模块的内窥镜视频半监督分割框架, 在插帧方法中加入光流细化与语义特征来提高生成帧的质量, 同时引入时空 Transformer 提取视频中空间与时间上的全局依赖性, 以便更好地学习手术器械运动特征. 该框架包含一个插帧模块与含有时空 Transformer 模块的编码器-解码器架构分割网络. 插帧模块包含一个光流估计器与一个优化器, 相对于之前的工作[9], 本文在插帧时加入语义特征提高生成帧的质量, 使得两帧间过度更平滑, 提升训练集的时序一致性

与数据多样性. 尽管近段时间以来, Transformer 因其可以提取全局上下文特征的优势被广泛使用在各类视觉任务中并取得良好效果[14,15], 但并未被应用到本任务中. 本文分割网络在利用 CNN 卷积层提取图像的高级语义特征后, 利用时空 Transformer 进一步提取时空上下文依赖信息. 具体来讲, 我们首先计算同一时间不同位置上的全局依赖性, 之后再从时间维度计算经空间 Transformer 增强过后的特征来获取时间上即某段时间内连续帧之间的一致性信息, 这样可以将高级语义特征从时间与空间两个方面进行增强, 使分割网络更好地识别全局与局部的信息. 最后, 将生成的高级语义特征在解码器上与先前编码器生成的特征拼接, 补充丢失的细节信息, 强化对边缘的分割能力.

本文的贡献包括以下 3 个部分:

(1) 提出一种基于语义强化的插帧及伪标签生成算法, 通过学习连续帧间的光流, 对连续帧进行插帧并生成对应的伪标签, 提高生成帧的质量, 平滑两帧间的过度, 促进分割网络提取手术器械特征;

(2) 针对在插帧之后增加的时序信息与空间信息, 提出了带有时空 Transformer 模块的内窥镜视频分割网络, 从时间和空间两个维度构建视频的全局上下文依赖关系, 对编码器输出的高级语义特征进行再提取, 强化对数据的解析能力, 从而提高复杂环境下手术器械, 特别是钳口部分的分割能力;

(3) 在此基础上, 我们提出了带有时空 Transformer 半监督手术器械分割框架, 在使用少量标注数据的情况下, 利用插帧和光流得到的无标注数据, 取得了优于基准模型的分割结果, 并超过了其他 state-of-the-art 方法, 并通过消融实验验证了本文方法的实际效果与有效性.

1 相关工作

本节将对本文所涉及的背景知识与相关技术进行简要介绍, 主要包括视频分割、半监督学习以及 Transformer 这 3 个方面.

1.1 视频分割

与语义分割相比, 视频分割增加了对连续帧时间一致性的考虑. 现有的方法主要分类两类. 第 1 类是通过利用以前帧中的特征来进行计算. Shelhamer 等人提出了一种 Clockwork 网络[16], 该网络基于多层 FCN, 并利用之前帧的特征来节省计算量. Zhu 等人提出了深度特征流, 通过在 FlowNet 中学习到的光流, 将高级特征从关键帧传播到当前帧[17]. Gadde 等人提出了一个 NetWarp 模块, 将先前帧的特征与光流和当前帧的特征相结合, 辅助分割[18]. 另一类是通过利用视频连续帧之间的时序特征聚合或时序模块来提高分割的准确性. Fayyazd 等人使用时序 LSTM 模块结合连续帧的 CNN 特征[19]. Jin 等人提出通过预测特征学习模型以无监督的方式从未标记的视频数据中学习[9]. 在机器人视频分割方面, LWANet, BASNet 等工作将轻量模型与注意力模块引入到分割任务中, 提高了分割网络的精确度与计算速度[20,21].

1.2 半监督学习

实际应用中, 无标签数据易于获取, 而有标签的数据往往需要耗费大量人力与时间进行标注. 为了解决这些问题, 半监督学习在深度学习领域的应用逐渐受到重视. 一般来说, 半监督学习需要数据满足三大假设才能建立预测样例与学习目标之间的关系, 即平滑假设、聚类假设与流型假设. 常用的半监督学习方法主要有两种. 第 1 种为一致性正则化, 即通过使用无标签数据强化训练已有的深度学习模型, 来使模型符合聚类假设. Pi-Model 利用正则化通常不会改变模型输出概率分布的特性来提升模型在不同扰动下的一致性[22]. Mean teacher 使用时序组合模型, 对学生模型进行滑动指数平均来消除扰动并稳定当前值, 在图像分类数据集 SVHN 上实现了 4.35%的错误率[23]. 第 2 种方法为代理标签法, 这种方法使用预测模型生成一些代理标签, 为训练提供一些额外的信息. Self-training with Noisy Student 首先训练一个 EfficientNet 来为无标签数据生成伪标签, 再利用有标签数据和伪标签数据训练一个学生模型, 并在训练过程中加入噪声, 该方法在 ImageNet 数据集上取得了 88.4%的正确率[24].

1.3 Transformer

Transformer 是一种基于 multi-head 注意力机制与前馈神经网络的端到端模型, 起初被设计用于解决 seq2seq 的机器翻译问题. 此后, 根据其能构建全局依赖性的特点开发的 BERT, GPT 模型在自然语言处理任务上取得了良好的效果. 近年来, 由于该架构在自然语言处理领域上出色的表现, 该思想被逐渐应用于计算机视觉领域. 在图像分类任务上, ViT[25]在完整大小的图像上使用 Transformer 来实现全局自注意, 在 ImageNet 分类上取得了最优的结果. Chen 等人提出了一种基于 Transformer 和二分图最大匹配损失函数的框架 DERT[26], 用于解决直接集合预测问题, 在 COCO 数据集挑战赛上取得了与优化 Faster RCNN 相当的结果. 在图像分割任务上, VisTR 首次提出一种基于 Transformer 的视频实例分割框架[15], 将视频实例分割问题转化为直接端到端的并行序列解码/预测问题. Bertasius 使用 Transformer 除了视频的时序与空间信息来进行视频分类[27]. Han 等人首次在医学图像分割上将 UNet 与 Transformer 相结合[14], 这样不仅可以将图像特征作为序列获取上下文信息, 也可以通过 UNet 的卷积层获取低级局部特征.

2 方 法

为了解决现有的手术器械视频半监督分割方法无法插帧、精度低以及分割网络上下文感知能力差的问题, 针对目前广泛使用的间隔若干帧带有标签的稀疏标注数据集, 本文研究并实现了一种基于插帧的半监督数据扩增方法以及基于时空 Transformer 的分割网络. 图 1 为本方法的整体框架示意图, 针对稀疏标注视频数据集中无标注的数据, 我们通过学习并预测光流迁移方向的半监督方法, 将有标注帧上的标签通过光流方向迁移至邻近无标注的帧上生成伪标签.

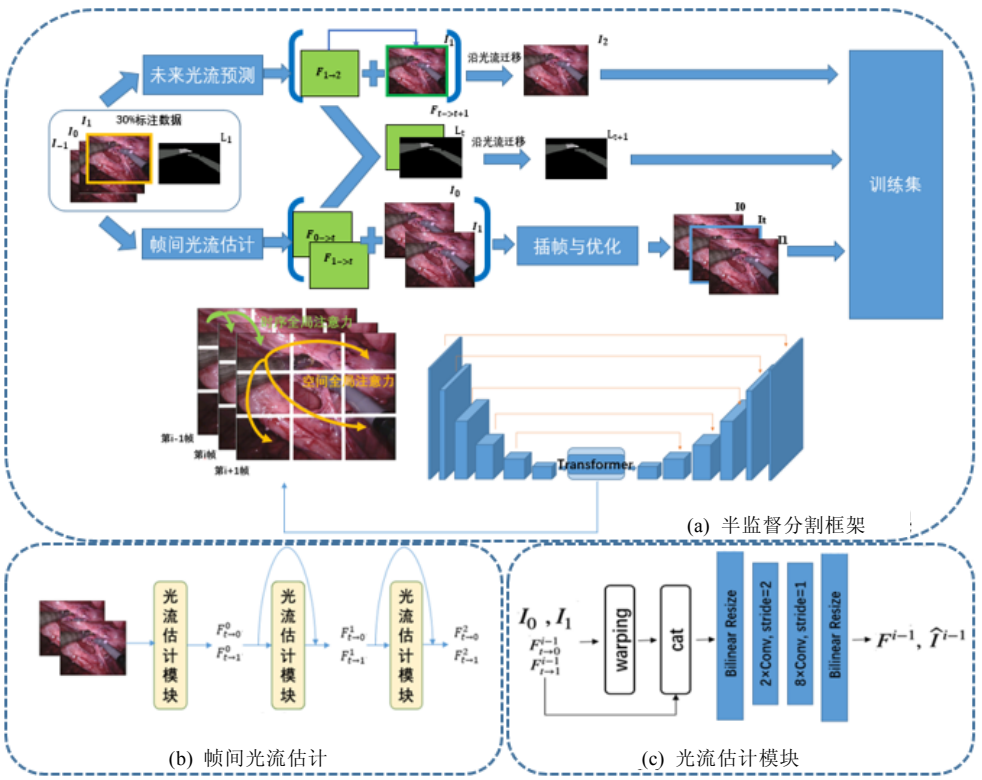


图 1 半监督分割框架结构图

2.1 面向半监督学习的视频插帧方法

现有半监督分割框架[9]为了使分割网络学习到更多手术器械运动的特征, 使用插帧模块对连续帧进行插帧, 以平滑连续帧之间手术器械的快速运动. 但该插帧方法存在光流估计不准确、插帧结果质量不高的问题. 为更好地平滑手术器械的帧间运动, 改善半监督框架生成数据的时序信息, 提高分割网络的分割效果, 本文提出一种新的面向半监督学习视频插帧方法, 在此前半监督分割框架的光流预测的基础上, 从光流估计与图像优化两个方面对其插帧机制进行了改进, 以此提高经半监督方法处理后数据的时序一致性与数据多样性.

本节将介绍基于插帧与未来帧预测的半监督方法. 本文通过提取数据集中未标注帧的光流与相关特征, 预测未来帧与构造中间帧, 再将标注数据的标签通过光流迁移来构造对应的伪标签, 从而完成对稀疏标注数据集无标签数据的标注. 本文首先通过预测未来帧与伪标签的方法[9]来对稀疏标注数据进行标注, 该方法通过学习当前帧及过去若干帧的光流来预测未来的光流, 然后将当前帧与标签沿预测的光流迁移, 生成新帧与伪标签, 步骤如图 1 所示.

2.1.1 未来帧预测与伪标签生成

参考此前的半监督框架中未来帧预测的方法[9], 假设数据集中的稀疏标注数据集连续帧 $\boldsymbol{I}=\{I_0,I_1,\ldots,I_{T-1}\}$, 每间隔若干帧含有一个带有标签的帧, 例如, 带有 33%有标注数据的$\{I_0,I_3,I_6,\ldots\}$间隔为 2. 针对标注帧之间的无标注数据, 我们使用光流预测的方法构造一个编码器-解码器结构的 CNN 网络, 通过学习当前有标注帧 I_i 及之前若干连续帧的光流变化, 预测到未来无标注帧 I_{i+1} 的光流 $\hat{F}_{t\to t+1}$, 即下一帧手术器械的运动方向. 预测网络输出光流 $\hat{F}_{t\to t+1}$之后, 将当前帧及标签(I_i,L_i)沿预测的光流 $\hat{F}_{t\to t+1}$向 I_{i+1} 的位置进行扭曲, 得到该位置的预测帧与伪标签 $(\hat{I}_{t+1},\hat{L}_{t+1})$, 而原始数据 I_{i+1}仅用于光流的预测. 同理, 使用该方法可以获得其他无标注帧对应位置的预测帧与伪标签, 如 I_{i+2} 或 $I_{i\text{-}1}$.

2.1.2 视频插帧

在插帧工作中, 准确的光流估计有助于构建高质量的伪帧和伪标签, 在训练分割网络中起着重要的作用. 特别是在低帧率的视频中, 有助于平滑运动和保持帧序列的一致性. 本文采用光流场向后扭曲的方法进行插帧. 但传统的做法从预训练的光流估计模型计算双向的光流[28], 然后对其进行反向和细化来生成中间的光流. 此外, 由于物体位置的变化, 这些方法往往无法准确估计物体的运动. 因此, 本文采用从粗到细多尺度光流网络, 从低分辨率到高分辨率估计光流. 利用神经网络在低分辨率上有着更大感受野的特点, 在低分辨率的图像上捕捉大范围的运动, 之后再高分辨率的图像上对光流的细节进行细化. 这样的光流估计器可以较好地估计大幅度运动. 以往的中间光流估计往往直接利用反向光流加以计算, 而视频中物体复杂的运动使得反向光流不能保证准确.

在本文的工作中, 我们使用一个编码器-解码器结构的网络来直接估计光流, 如图 1(b)及图 1(c)所示. 给定一对连续的 RGB 帧, 我们的目标是在中间时刻合成一个中间帧. 通过光流估计器输入帧直接估计中间光流, 然后通过向后扭曲输入帧得到两个粗略的结果. 为了处理大范围运动引起的光流估计偏差, 我们先在低分辨率上的对帧间的光流进行估计, 从而更好地捕捉大范围运动, 再逐步提高分辨率优化相关细节. 具体操作如公式所示:

$$F^i = F^{i-1} + g^i(F^{i-1},\hat{I}^{i-1}),$$

其中, F^{i-1} 为当前从第 i–1 个光流估计模块中产生的光流, g^i 代表第 i 个光流估计模块. 将输入图像与 F^{i-1} 扭曲之后生成 $\hat{I}^{i-1}$, 之后将 F^{i-1} 与 $\hat{I}^{i-1}$ 一同输入到 g^i 之中来估计下一个光流 F^i.

仅依靠光流进行插值得到的帧会导致图像存在模糊或不对齐的情况. 为了解决这个问题, 如图2所示, 得到光流之后, 首先将对应帧与光流进行一个后向扭曲, 产生一个粗糙的生成帧. 为了提高插值帧的质量, 除光流外, 本文参考此前的研究[29−32], 引入上下文特征、visibility maps M 与 reconstruction residual Δ来优化插帧. 我们通特征提取器模块来提取前后两个原始帧不同分辨率上的金字塔语义特征:

$$C_0:\{C_0^0,C_0^1,C_0^2,C_0^3\},\ C_1:\{C_1^0,C_1^1,C_1^2,C_1^3\}.$$

再将这些语义特征与之前生成的不同分辨率的光流进行后向扭曲至插帧位置，得到 $C_{t\leftarrow 0}$ 和 $C_{t\leftarrow 1}$，以参与插值帧的构造.

我们构造了一个 UNet 结构的插帧优化器，包含一个编码器与一个解码器，解码器部分包括 4 个上采样模块. 将前后两个原始帧沿估计光流生成的粗糙生成帧 $\hat{I}_{t\leftarrow 0},\hat{I}_{t\leftarrow 0}$ 与光流 $F_{t\to 0}$ 拼接输入到编码器，并在每次下采样前与此前提取的语义特征拼接，输出采用 PReLU[30]函数进行激活，产生可以用于改善图像质量的 visibility map M 与 reconstruction residual Δ，之后对原始插帧进行如下处理，得到优化后的插帧 $\hat{I}_t$：

$$\hat{I}_t = M \odot \hat{I}_{t\leftarrow 0} + (1-M) \odot \hat{I}_{t\leftarrow 1} + \Delta,$$

其中，$\odot$为 element-wise 乘法，$0\leqslant M$，$1\leqslant \Delta$.

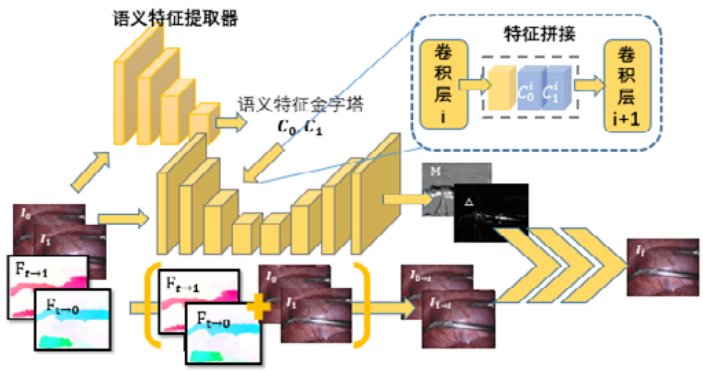


图 2　插帧与优化模块

2.1.3　视频插帧后处理

我们将稀疏标注视频数据集的帧序列定义为 $\boldsymbol{I}=\{I_0,I_1,\ldots,I_{T-1}\}$，每间隔 h 帧含有一个带标签的帧，整个数据集中的 N 组图像与对应标签表示为 $\mathcal{D}_L=\{(I_t,L_t)\}_{t=hn}$，不含标签的 M 张图像表示为 $\mathcal{D}_U=\{I_t\}_{t\neq hn}$，其中，$M=hN$. 基于 Zhao 等人[7]的预测光流并传播图像与标签的方法[7]，得到 M 组预测的帧与伪标签 $\mathcal{D}_R=\{\hat{I}_t,\hat{L}_t\}_{t\neq hn}$，在 $\mathcal{D}_L$ 与 $\mathcal{D}_U$ 的基础上，使用本文提出的插帧方法对 $\mathcal{D}_L$ 标注帧与相邻的 $\mathcal{D}_U$ 无标注帧之间进行插帧，并利用插帧时产生的光流，与 $\mathcal{D}_L$ 中的标签向新生成的帧进行后向扭曲生成匹配的标签，得到 $N+M-1$ 组图像与伪标签：

$$\mathcal{D}_I=\{\tilde{I}_{t_0},\tilde{L}_{t_0}\}_{t=1}^{T-1}.$$

2.2　分割网络

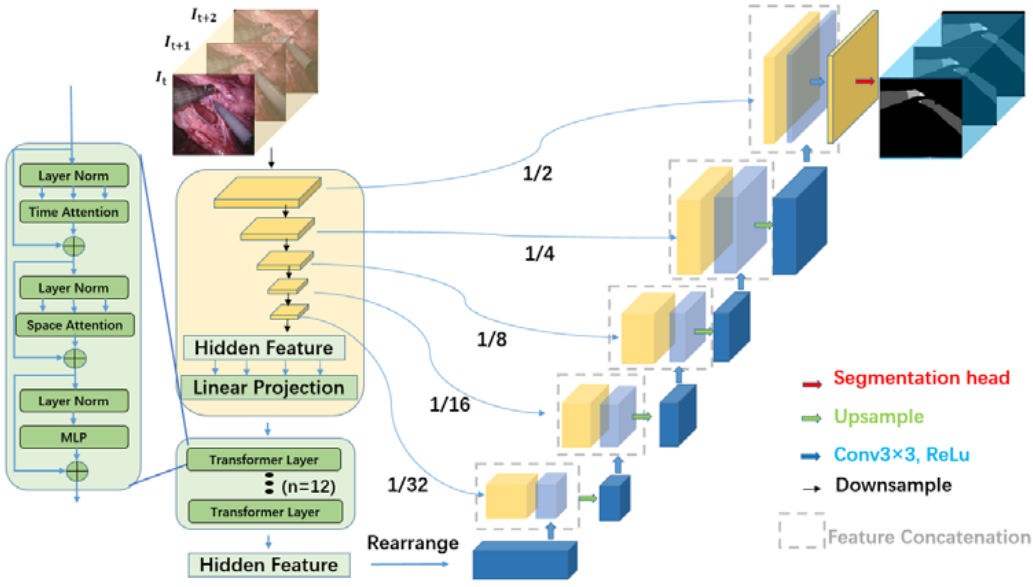
给定一段视频 $\boldsymbol{X}\in\mathbb{R}^{F\times H\times W\times C}$，即 F 张分辨率为 $H\times W$，通道为 C 的连续帧，本文的目标是将这些帧逐像素进行分类.

常用的 CNN 分类模型，例如 U-Net，首先使用编码器将帧序列下采样为高级语义特征，之后再使用解码器将其上采样至原分辨率. 与现有方法不同，为了对编码器输出的高级语义特征构建全局依赖关系，本文将 Transformer 模块引入到分割网络的编码器中. 本文提出的分割网络结构如图 3 所示，我们在下面各小节中详细阐述编码器、Transformer 模块以及解码器的构造与具体设置.

2.2.1　编码器

我们的分割网络整体结构如图 3 所示，分割网络的编码器是 5 个 Conv-ReLU 模块，前两个模块中每个模块包含一个 2D 卷积层和一个 ReLU 激活函数，后 3 个模块中每个模块包含两个 2D 卷积层和 2 个 ReLU 激活函数，呈 Conv-ReLU-Conv-ReLU 结构. 每个模块后使用一个 2×2 的 max pooling 操作来进行下采样，使得每次卷积过后的特征图尺寸减半，同时通道数加倍，第 1 次卷积产生 64 个通道，整个过程总共进行 5 次下采样，直到达到 512 个通道. 这种方法通过逐步采样的方式从编码为低分辨率特征图的原始图像中提取语义信息，同时获得周围区域的空间信息. 在最后一个 Conv-ReLU 模块之后的 bottleneck 层，采用 Transformer 来从高特

征图中进步一步提取空间、时间上下文信息, 从而建立长距离依赖.



(a) Transformer 模块 (b) 分割网络

图 3 分割网络结构图

2.2.2 Transformer

由于卷积计算自身的局限性, 基于 CNN 的分割网络常常无法有效处理全局信息, 因此在处理某些问题时效能不佳, 特别是对于形状大小以及颜色都与肌肉组织有较大区别的手术器械时. 有效处理全局关系、联系远距离关系上手术器械的整体结构, 会对分割工作提供较大帮助. 此外, 视频中的时序信息也有助于提高分割网络对当前帧的理解, 但此前对时序信息进行处理的模块往往会丢失部分信息, 导致无法对时序上的全局信息进行有效处理, 对使用插帧方法增强的时序信息无法有效理解. 因此, 我们在分割网络的 bottleneck 位置引入了空间 Transformer 模块, 该模块由时序 Transformer 与空间 Transformer 组成, 从时序与空间两个角度对视频进行全局建模.

编码器处理后的特征图经 Linear Projection 后输入到 Transformer 中, 每一段视频由 F 张连续帧组成. 按照 ViT 的方法[25], 我们将每一帧分解为 N 个不重叠的 patch, 每个 patch 的大小为 $D=h\times w$. 一帧图像由 N 个 patch $x_i\in\mathbb{R}^{h\times w}$ 组成, 即 $N=HW/hw$, 将这些 patch 进行 linear projection 操作, 转化为一维 tokens $z_i\in\mathbb{R}^d$, 之后进行 Patch Embedding 操作:

$$\boldsymbol{z}=[\boldsymbol{E}x_1,\boldsymbol{E}x_2,\ldots,\boldsymbol{E}x_N]+\boldsymbol{p},$$

其中, $\boldsymbol{E}\in\mathbb{R}^{h\times w}$ 为 patch embedding projection, $\boldsymbol{P}\in\mathbb{R}^{N\times d}$ 为 positional embedding. 之后, 使用 L 层 Transformer 模块计算 token 全局上下文关系:

$$y^{\ell}=MSA(LN(z^{\ell}))+z^{\ell},$$
$$z^{\ell+1}=MLP(LN(y^{\ell}))+y^{\ell}.$$

每一次层包括 Multi-head self-attention 模块(MSA)、layer normalization 模块(LN)及多层感知机 MLP 模块.

基于 ViT[25]的基本操作, 我们从时序与空间两个角度来计算输入特征的全局上下文关系. 我们将每次输入的视频 $\boldsymbol{V}\in\mathbb{R}^{T\times H\times W\times C}$ 转换为 token 序列 $z\in\mathbb{R}^{1\times n_t\cdot n_h\cdot n_w\cdot d}$, 之后根据先时序后空间的顺序, 将输入的 token z 从 $\mathbb{R}^{1\times n_t\cdot n_h\cdot n_w\cdot d}$ reshape 为 $\boldsymbol{Z}_t\in\mathbb{R}^{n_h\cdot n_w\times n_t\cdot d}$, 使用时序 Transformer 依次计算视频中连续帧 $m_h\cdot n_w$ 个位置不同时间上 n_t 个 patch $\{x_1,x_2,\ldots,x_{n_t}\}$ 之间的时序全局上下文关系, 之后 reshape 为 $\boldsymbol{Z}_s\in\mathbb{R}^{n_t\cdot n_h\times n_w\cdot d}$, 使用空间

Transformer 计算依次视频中 n_t 帧不同位置上 n_s 个 patch $\{x_1, x_2, ..., x_{n_h \times n_w}\}$ 之间的空间全局上下文关系, n_s=n_h×n_w, 如图 4 所示. 输出结果前, 使用多层感知机(MLP)进行激活:

$$y_t^\ell = MSA(LN(z_t^\ell)) + z_t^\ell,$$

$$y_s^\ell = MSALN(y_t^\ell) + y_t^\ell,$$

$$Z^{\ell+1} = MLPLN(y_s^\ell) + y_s^\ell.$$

这一先后顺序的设置, 参考了此前的相关研究[27].

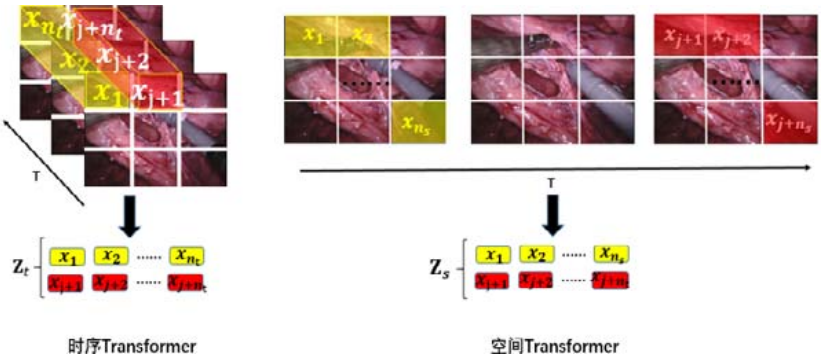


图 4 时空 Transformer

2.2.3 解码器

在 bottleneck 位置使用 Transformer 建立获取时间空间上下文依赖关系之后, 我们通过解码器来将高级特征图降维和恢复到原始输入大小来输出分割标签. 解码器由 5 个上采样模块组成, 每个模块包含一个 ConvReLU、一个 ConvTranspose2d 和一个 ReLU 激活函数组成. 每个上采样模块将特征图与编码器中对应的同样大小的特征图拼接后进行上采样. 之后使用一个 3×3 的 conv2d 模块将其解码为与原始输入一样大小的特征图, 最后使用 log_softmax 函数进行激活, 得到最后的分割结果.

3 实 验

3.1 实验设置

本文实验使用 MICCAI 2017 EndoVis 挑战赛手术器械分割数据集[33], 该数据集包括 10 段由达芬奇手术机器人操作的猪腹部手术视频, 每段 300 帧, 频率为 1Hz, 分辨率为 1280×1024. 其中, 8 段视频的前 225 帧作为训练集, 余下 75 帧和另外两段视频的 300 帧作为测试集. 数据集的标签包括手术器械的手柄、关节、钳口这 3 个部位与背景. 考虑到服务器的配置, 训练时我们将图片大小调整为 960×960, 视频的长度设定为 3 帧. 为了更好地学习手术器械的运动特征, 截取视频时, 我们按照长度为 3、步长为 1 的设置对正序视频和倒序视频进行截取, 并随机进行水平和垂直方向的翻转. 为了直接和公平地比较, 我们遵循文献[9]中相同的评估方式, 使用发布的 8×225 帧的视频进行四折交叉验证. 评估指标为平均交并比 IoU 与平均 DICE 系数[34]. 计算方式如下:

$$DICE = \frac{2\times|X \cap Y|}{|X|+|Y|},$$

$$IOU = \frac{|X \cap Y|}{|X \cup Y|},$$

其中, X 为预测结果, Y 为金标准.

本文方法基于 pytorch1.4.0 实现, 使用两张 NVIDIA GTX 2080 显卡进行训练, 训练时采用 Adam 优化方法. 实验中, batch size 设置为 2, 初始学习率设置为 1e–4, 实验总迭代次数为 50 次, 每 25 次迭代学习率减小

为原来的 1/10. 参考此前相关研究的结果, 结合服务器实际配置, 我们将 Transformer 的 Patch Size 设置为 6, blocks 设置为 12.

为了对比之前的插帧方法, 我们使用 Zhao 等人论文中提到的插帧方法[9]对数据进行扩增, 得到 $N+M-1$ 组图像与伪标签 $\mathcal{D}_C=\{\tilde{I}_{t_0},\tilde{L}_{t_0}\}_{t=1}^{T-1}$. 我们将图像与标签的集合$\mathcal{D}_L\cup\mathcal{D}_R\cup\mathcal{D}_C$ 作为 Zhao 的方法[9]的训练集, 将 $\mathcal{D}_L\cup\mathcal{D}_R\cup\mathcal{D}_I$作为本文提出方法的训练集. 实验中, 我们采用的系数表述视频数据集的标签间 h 为 2, 即使用原数据集约 30%的标签. 同时, 在实验部分也对 Zhao 等人方法[9]中使用的分割网络 UNet11[5]在不同大小数据集上的表现进行了比较.

3.2 半监督时空Transformer网络与其他方法对比实验结果

为了证明本文提出的半监督时空 Transformer 网络 SSTNet(semi-supervised spatiotemporal transformer networks)在手术器械分割方面的优势, 我们将本文提出的的方法与现有的半监督方法及基础分割网络的分割结果进行对比, 分别是本文与 Zhao 等人[9]方法中使用的基础分割网络 UNet11[5]使用全部标注数据及使用 30%标注数据时的分割结果、使用 30%标注数据时 Zhao 等人[9]提出的方法及加入 ConvLSTM [11]模块后的分割结果结果, 如表 1 所示. 可以看出, 我们的方法在 DICE 与 IoU 指标上取得了 82.42%和 72.01%的结果. 该结果不仅优于 Zhao 等人的半监督框架[9]的分割结果, 对比全监督分割方法也有一定的优势. 本文对部分分割结果进行了可视化, 如图 5 所示. 可以看出, 本文提出的方法能够改善手术器械钳口及手柄处的分割准确率.

表 1 本文提出的方法与其他方法的分割结果

方法	标注数据百分比 (%)	DICE (%)	IoU (%)
UNet11[5]	100	76.08	64.87
UNet11[5]	30	72.44	61.75
Zhao 等人[9]	30	74.74	63.82
SSTNet	**30**	**82.42**	**72.01**

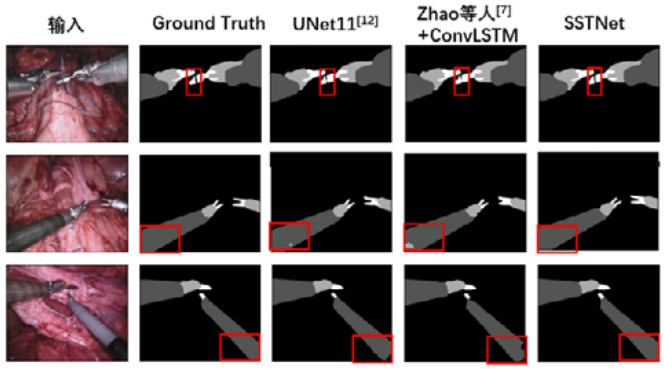


图 5 SSTNet 与其他方法的分割结果

3.3 插帧机制对结果的影响分析

我们将本文提出的插帧方法与 Zhao 等人的插帧方法[9]实验结果相比较, 以此来验证本文提出的插帧方法在提高视频数据集时间一致性及数据多样性方面的能力. 为了直观对比两种方法的效果, 我们采用 Zhao 等人[7]论文中使用的分割网络 UNet11[5]来对两组数据进行分割. 结果如表 2 所示: 在 UNet11[5]网络上, 使用 30%的标注数据, 本文提出的插帧方法在 DICE 和 IoU 两个指标上分别领先 Zhao 等人的插帧方法[9]1.18%和 0.92%. 该结果证明了我们的插帧方法在提高数据多样性进而提高分割训练结果的能力.

表 2 不同插帧方法的分割结果

方法	DICE (%)	IoU (%)
Zhao 等人[9]+插帧[9]	74.74	63.82
本文提出的方法	**75.92**	**64.74**

3.4 时序机制对结果的影响分析

为了验证本文提出的半监督时空 Transformer 分割框架中时序机制, 我们对框架内的插帧方法以及时序 Transformer 进行了实验. 通过与 Zhao 等人[9]加插帧[9]以及目前经常被用于视频时序处理的 ConvLSTM 模块[11]进行对比, 分别证明了本文提出的插帧方法与时序 Transformer 对提高分割结果的贡献, 验证了这两个模块的有效性.

我们将 ConvLSTM 模块[11]加入到分割网络中, 该模块具有多个 ConvLSTM 层[11], 在输入到状态和状态到状态的转换中使用卷积结构, 可用于分析连续的图像序列, 提取其中的关键特征. 在这里, 我们将其用于验证我们的插帧方法在提升数据集时序一致性方面的能力. 实验时, 我们对两种方法生成的数据使用相同的分割网络, 即在 bottleneck 位置嵌入 ConvLSTM 模块[11]的 UNet11[5]网络. 实验结果由表 3 所示: 在 bottleneck 位置加入 ConvLSTM 模块[11]后, Zhao 等人的方法[9]在 DICE 和 IoU 上分别取得 76.33%, 64.89%的结果; 我们的方法在 DICE 和 IoU 上分别取得 76.82%, 66.13%的结果, 均优于 Zhao 等人的方法[9]. 实验结果表明: 我们的插帧方法能够生成精度更高的图像与对应的伪标签, 为 ConvLSTM 模块[11]提供了更多时间上下文信息, 使其更好地挖掘连续帧之间手术器械的相关特征, 提高了分割网络的分割效果, 验证了该方法在增强稀疏视频数据集时序一致性上的有效性.

为了对比时序 Transformer 模块与 ConvLSTM[11]模块处理时序信息的能力, 验证时序 Transformer 效果时, 我们在 UNet11[5]分割网络的 bottleneck 位置嵌入时序 Transformer 模块, 使用本文提出的半监督插帧方法生成的数据, 来对插帧之后的数据集进行分割. 结果见表 3: 时序 Transformer 的分割结果较 ConvLSTM[11]的结果, 在 mDICE 与 mIoU 方面分别有 2.12%和 2.90%的提高.

表 3 不同时序处理模块的分割结果

方法	DICE (%)	IoU (%)
Zhao 等人[9]+插帧[9]+ConvLSTM[11]	76.33	64.89
本文提出的半监督方法+ConvLSTM[11]	76.82	66.13
本文提出的半监督方法+时序 Transformer	**78.94**	**69.03**

图 6 展示了分割网络在分别加入 ConvLSTM[11]和时序 Transformer 之后, 对连续 3 帧的分割结果. 视频中, 两件手术器械正在进行缝合操作. 从分割结果中可以看出: 两种方法对手柄和连接部分的分割效果正确率较高, 但钳口部分的分割效果仍有些欠缺, 特别是使用 ConvLSTM[11]的分割结果, 会将连接部分误认为钳口部分. 第 1 帧黄色方框内, 使用 ConvLSTM[11]的分割网络将反光的缝合线识别为钳口部分, 而使用时序 Transformer 的分割网络则有效处理了这个问题. 第 2 帧第 3 帧红色与绿色方框中, 使用 ConvLSTM[11]的分割网络将部分连接部分识别为钳口部分, 因为钳口与连接部分都是金属材质, 增加了分割网络的识别难度. 但使用时序 Transformer 的分割网络能对这部分进行正确地判断. 由此可见: 时序 Transformer 能更好地处理时序上的上下分关系, 强化分割网络对手术器械相同材质不同部分的识别能力, 以及对其他反光物体的鉴别能力. 而 ConvLSTM[11]因其具有遗忘的特性, 无法构建时序上的完整全局依赖关系, 丢失了部分有效信息, 导致对相同材质的部位以及反光物体的分割效果上存在缺陷.

以上实验验证了本文提出的插帧方法与时序 Transformer 在增强和处理时序信息上的能力, 证明了半监督时空 Transformer 框架中时序机制的有效性.

3.5 消融实验

基于本文中提出的时空 Transformer 模块, 本文做了相关的对比实验, 验证时序 Transformer 与空间 Transformer 两个子模块的性能, 结果见表 4. 可以看到: 本文设计的时序 Transformer 模块与空间 Transformer 模块相比原始分割网络, 分割效果均有不同幅度的提升. 而同时使用时空 Transformer 模块的分割网络取得了最好的分割效果, 相比于原始分割网络, 在 DICE 和 IoU 上提升 75.92%与 64.74%, 单独使用空间 Transformer 可将原始分割网络的结果提升 3.54%与 4.5%, 单独使用时序 Transformer 可将原始分割网络的结果提升 3.02%与 4.49%.

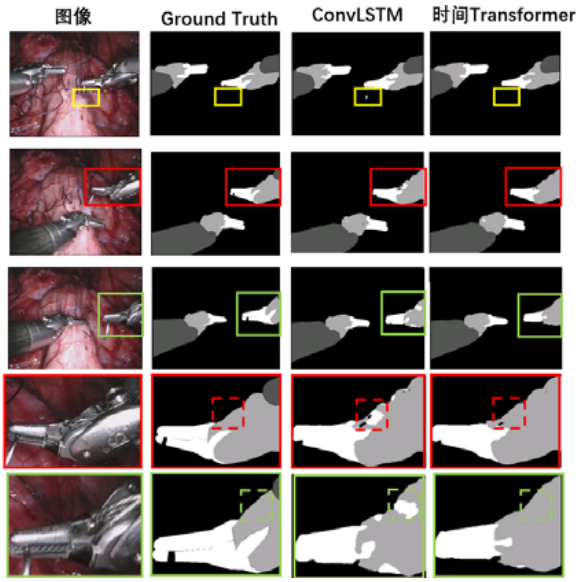


图 6 使用 ConvLSTM[9]与使用时序 Transformer 分割结果可视化

表 4 使用不同模块组合的分割结果

插帧	空间	时序	DICE (%)	IoU (%)
√	–	–	75.92	64.74
√	√	–	79.46	69.24
√	–	√	78.94	69.03
√	√	√	**82.42**	**72.1**

图 7 展示了仅使用时序 Transformer 模块的分割网络与原始分割网络的分割结果. 结果选自连续的 3 帧图像, 视频中手术器械正在牵拉组织, 钳口与组织接触, 部分被组织遮挡. 由分割结果可知: 仅插帧的方法能基本完成对手术器械的手柄与连接部分的分割, 但钳口部分分割效果较差; 而在分割网络加入时序 Transformer 后, 分割网络能够对钳口部分进行有效的分割, 提高了钳口位置的分割精确度. 此外, 第 2 帧中手柄出现反光的情况, 此前研究结果表明, 手术器械表面的反光会影响分割网络的表现. 在第 2 帧中, 仅使用插帧的分割网络对反光区域产生了错误的分割结果; 而加入时序 Transformer 后, 分割网络可以在反光位置进行正确地分割. 以上结果证明了: 时序 Transformer 模块可以帮助分割网络构建对视频连续帧时序全局上下文的理解, 提高分割网络处理复杂环境、识别细小部件的能力.

图 8 展示了仅使用空间 Transformer 模块的分割网络与原始分割网络的分割结果, 3 张图片中, 手术器械均在进行牵拉、缝合组织等复杂的手术操作, 金属钳口也出现反光、遮挡等情况. 图 1 中, 原始分割网络无法有效区分左边器械的钳口与连接部分和识别阴暗条件下右边器械的钳口部分; 图 2 中, 原始分割网络无法区分缝合线与钳口部分, 并将右边钳口附近反光的组织也误识别为钳口; 图 3 中, 下方器械的手柄部分被原始分割网络错误识别为钳口. 而原始分割网络加入空间 Transformer 之后, 对复杂环境中的手术器械分割效果得到提高, 能够很好地解决上述原始分割网络中存在的问题.

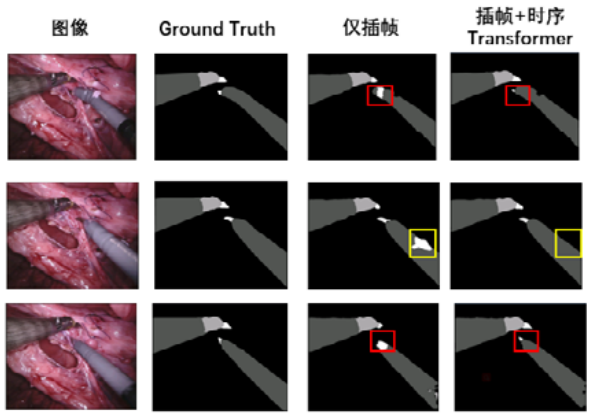


图 7 使用时序 Transformer 的分割结果可视化

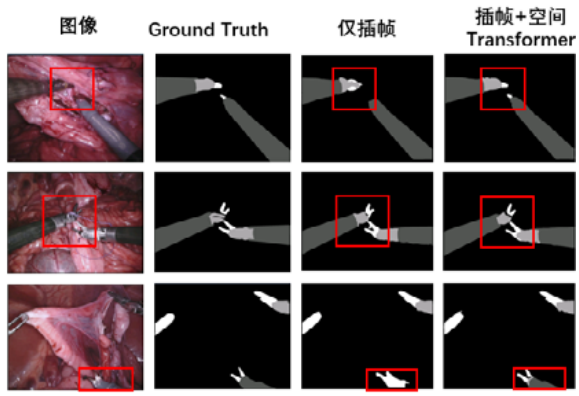


图 8 使用空间 Transformer 的分割结果可视化

从实验结果与效果分析中可以得出: 原始分割网络在加入时序 Transformer 或空间 Transformer 后, 模型性能得到明显提升, 处理复杂条件下的钳口部分分割能力得到增强. 这说明时序 Transformer 或空间 Transformer 能在时间或空间维度建立全局依赖关系, 提取时间或空间上下分信息, 强化了模型对高级语义特征的理解.

3.6 讨 论

本文针对目前研究中对数据不足、复杂环境下分割效果差等问题, 提出了一种新的半监督分割框架, 通过优化插帧的效果来为无标签数据标注, 克服了标注数据少的问题, 增强了数据的时序一致性与多样性. 为充分利用增强后的数据, 学习手术器械内在运动特征, 提出了时序 Transformer 网络, 强化对视频中的时序上下文信息以及空间上下文信息的学习, 克服了 CNN 网络感受野小以及一般时序信息学习模块遗忘丢失信息的缺点, 提高了分割网络的分割效果.

尽管在分割网络中加入时空 Transformer 特征学习之后, 手术器械视频分割的正确率得到提高, 但考虑到内窥镜手术的复杂环境, 视频仍有一些情况是本文无法解决的. 如图 9 所示, 图 9(a)左边的手术器械无法被正确识别与分割, 原因可能是器械运动过快导致画面模糊. 图 9(b)中, 视频下方的手术器械各部分识别效果较差, 原因可能是离光源过近导致相关区域过度曝光丢失了画面细节. 图 9(c)左边出现不明的障碍. 图 9(d)中光滑的组织表面出现了手术器械的倒影被分割网络错误识别. 造成相关错误的原因主要有二: 一是训练集中不含有类似情况, 导致模型没有学习到这类图像的分割方法; 二是模型在全局信息感知上仍存在缺失, 导致泛化能力不强. 考虑到这些缺陷, 下一步我们的研究方向是改进 Transformer 与分割网络的连接方式, 实现对高低级语义特征的更加有效的结合, 提高模型的全局信息感知能力, 从而实现更加精准的分割效果.

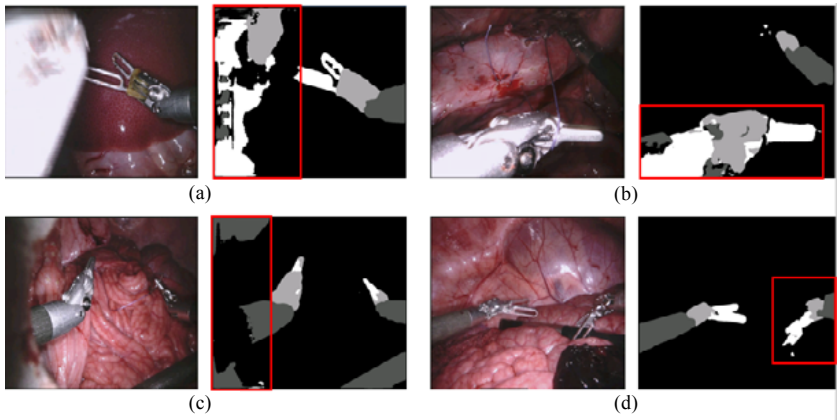


图 9 分割结果中表现较差的情况

4 结 论

本文针对内窥镜视频分割问题, 结合插帧方法与时空 Transformer 模块, 提出了一个基于时空 Transformer 半监督视频分割框架. 我们利用预测光流在连续帧中插帧及生成伪标签来提高稀疏标注数据集的时序一致性与数据多样性, 在分割网络的 bottleneck 位置中加入时空 Transformer 模块, 将编码器生成的高级语义特征输入 Transformer 模块, 依次从时间与空间两个角度对高级语义特征进行增强与构建全局依赖; 同时, 在上采样过程中结合分割网络卷积层提取的局部细节信息, 有效克服了手术中各类因素的影像, 提高了内窥镜视频, 特别是钳口部分的分割效果. 我们通过消融实验, 分别验证了插帧方法、时序 Transformer 与空间 Transformer 模块各自的有效性. 最后, 我们在公开数据集上对比了本框架与其他半监督与全监督模型的分割性能, 实验结果表明: 本框架在仅使用 30%标注数据进行训练的情况下, 达到并超过同类半监督或全监督分割网络的分割准确率.

针对第 3.6 节中提到的本框架无法有效解决的情况, 我们下一步的研究方向致力于通过调节分割网络各种特征之间的连接方式, 改进对整体上下文信息以及局部细节的感知能力, 来提高分割网络与 Transformer 特征计算能力解决模型泛化能力不强的问题.

References:

[1] Tan M, Wang S. Research progress on robotics. Acta automatica sinica, 2013, 39(7): 963−972 (in Chinese with English abstract).

[2] Ross T, Reinke A, Full PM, *et al*. Robust medical instrument segmentation challenge 2019. arXiv: 2003.10299, 2020.

[3] Allan M, Kondo S, Bodenstedt S, *et al*. 2018 robotic scene segmentation challenge. arXiv: 2001.11190, 2020.

[4] Chen J, Chen YS, Li WH, *et al*. Application and prospect of deep learning in video object segmentation. Chinese Journal of Computers, 2021, 44(3): 609−631 (in Chinese with English abstract).

[5] Shvets AA, Rakhlin A, Kalinin AA, *et al*. Automatic instrument segmentation in robot-assisted surgery using deep learning. In: Proc. of the 17th IEEE Int'l Conf. on Machine Learning and Applications (ICMLA). IEEE, 2018. 624−628.

[6] Liu DC, Wei YH, Jiang TT, *et al*. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In: Proc. of the Int'l Conf. on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020. 657−667.

[7] da Costa Rocha C, Padoy N, Rosa B. Self-supervised surgical tool segmentation using kinematic information. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation (ICRA). IEEE, 2019. 8720−8726.

[8] Yan PX, Li GB, Xie Y, *et al*. Semi-supervised video salient object detection using pseudo-labels. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2019. 7284−7293.

[9] Zhao ZX, Jin YM, Gao XJ, *et al*. Learning motion flows for semi-supervised instrument segmentation from robotic surgical video. In: Proc. of the Int'l Conf. on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020. 679−689.

[10] Jin YM, Cheng KY, Dou Q, *et al*. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: Proc. of the Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Springer, 2019. 440−448.

[11] Kim S, Hong S, Joh M, *et al*. Deeprain: Convlstm network for precipitation prediction using multi-channel radar data. arXiv: 1711.02316, 2017.

[12] Tokmakov P, Alahari K, Schmid C. Learning video object segmentation with visual memory. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 4481−4490.

[13] Liu Z, Luo S, Li W, *et al*. Convtransformer: A convolutional transformer network for video frame synthesis. arXiv: 2011.10185, 2020.

[14] Han K, Wang YH, Chen HT, *et al*. A survey on visual transformer. arXiv: 2012.12556, 2020.

[15] Wang YQ, Xu ZL, Wang XL, *et al*. End-to-end video instance segmentation with transformers. arXiv: 2011.14503, 2020.

[16] Shelhamer E, Rakelly K, Hoffman J, *et al*. Clockwork convnets for video semantic segmentation. In: Proc. of the European Conf. on Computer Vision. Springer, 2016. 852−868.

[17] Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 207. 2223−2232.

[18] Gadde R, Jampani V, Gehler PV. Semantic video CNNS through representation warping. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 4453−4462.

[19] Fayyaz M, Saffar MH, Sabokrou M, *et al*. STFCN: Spatiotemporal fcn for semantic video segmentation. arXiv: 1608.05971, 2016.

[20] Ni ZL, Bian GB, Hou ZG, *et al*. Attention-guided lightweight network for real-time segmentation of robotic surgical instruments. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). IEEE, 2020. 9939−9945.

[21] Ni ZL, Bian GB, Xie XL, *et al*. Rasnet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In: Proc. of the 41st Annual Int'l Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019. 5735−5738.

[22] Laine S, Aila T. Temporal ensembling for semi-supervised learning. arXiv: 1610.02242, 2016.

[23] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. 2017. 1195−1204.

[24] Xie QZ, Luong MT, Hovy E, *et al*. Self-training with noisy student improves imagenet classification. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 10687−10698.

[25] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv: 2010.11929, 2020.

[26] Chen HT, Wang YH, Guo TY, *et al*. Pre-trained image processing transformer. arXiv: 2012.00364, 2020.

[27] Bertasius G, Wang H, Torresani L. Is space time attention all you need for video understanding? arXiv: 2102.05095, 2021.

[28] Jiang HZ, Sun DQ, Jampani V, *et al*. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 9000−9008.

[29] Huang ZW, Zhang TY, Heng W, *et al*. Rife: Real-time intermediate flow estimation for video frame interpolation. arXiv: 2011.06294, 2020.

[30] He KM, Zhang XY, Ren SQ, *et al*. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1026−1034.

[31] Niklaus S, Liu F. Softmax splatting for video frame interpolation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 5437−5446.

[32] Bao WB, Lai WS, Ma C, *et al*. Depth-aware video frame interpolation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2019. 3703−3712.

[33] Allan M, Shvets A, Kurmann T, *et al*. 2017 robotic instrument segmentation challenge. arXiv: 1902.06426, 2019.

[34] Song J, Xiao L, Lian ZC, *et al*. Overview and prospect of deep learning for image segmentation in digital pathology. Ruan Jian Xue Bao/Journal of Software, 2021, 32(5): 1427−1460 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6205.htm [doi: 10.13328/j.cnki.jos.006205]

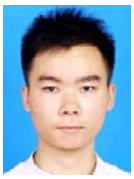
附中文参考文献:

[1] 谭民, 王硕. 机器人技术研究进展. 自动化学报, 2013, 39(7): 963−972.

[4] 陈加, 陈亚松, 李伟浩, 等. 深度学习在视频对象分割中的应用与展望. 计算机学报, 2021, 44(3): 609−631.

[34] 宋杰, 肖亮, 练智超, 等. 基于深度学习的数字病理图像分割综述与展望. 软件学报, 2021, 32(5): 1427−1460. http://www.jos.org.cn/1000-9825/6205.htm [doi: 10.13328/j.cnki.jos.006205]

李耀仟(1997－), 男, 硕士生, 主要研究领域为手术器械分割, 人工智能.



李才子(1993－), 男, 博士生, 主要研究领域为医学影像分析, 人工智能.

刘瑞强(1997－), 男, 硕士生, CCF 学生会员, 主要研究领域为医学图像处理, 人工智能.

司伟鑫(1990－), 男, 博士, 副研究员, 博士生导师, CCF 高级会员, 主要研究领域为医学影像分析, 计算机辅助介入.

金玥明(1994－), 女, 博士, 主要研究领域为机器人视频感知, 人工智能.

王平安(1961－), 男, 博士, 教授, 主要研究领域为 AI 和 VR 的医学应用, 手术仿真, 可视化, 图形学, 人机交互.