文章编号:1671-4598(2026)10-0282-06 DOI:10.16526/j.cnki.11-4762/tp.2026.10.036 中图分类号:R615;TP183 文献标识码:A

基于时空注意力模型的腹腔镜手术缝合技能评估方法

陈士涵[1]，夏建福[2]，彭橼平[2]，刘学锋[1]，陆小锋[1]
(1. 上海大学 通信与信息工程学院，上海 200444；
2. 温州医科大学附属温州中心医院 普外科，浙江 温州 325000)

摘要：对腹腔镜缝合技能的客观一致评估是外科培训的关键，但现有自动化评估方法在捕捉长时序依赖与操作细节方面存在不足，限制了评估的全面性与准确性；为解决该问题，提出了一种基于时空注意力机制的智能评估框架，采用TimeSformer作为骨干网络，提取更具判别性的时空特征以克服长时序动作依赖捕捉难点，并设计AdaPool模块，实现视频级特征的自适应汇聚，满足评估对关键操作片段的敏感度需求；基于LapSuture数据集的实验结果表明，该方法在技能分级任务中取得了83.0%的准确率和81.6%的F_1分数，优于传统方法；研究结果凸显了注意力驱动架构在实现稳健、可扩展且客观的手术技能评估方面的潜力，为提升外科教育与能力评价提供了新的思路。

关键词：腹腔镜手术；手术技能评估；深度学习；图像处理；Transformer

Assessment Method for Laparoscopic Surgical Suturing Skills Based on a Spatiotemporal Attention Model

CHEN Shihan[1], XIA Jianfu[2], PENG Xiaoping[2], LIU Xuefeng[1], LU Xiaofeng[1]
(1. School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China;
2. Department of General Surgery, The Wenzhou Central Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China)

Abstract: It is crucial for an objective and consistent assessment of laparoscopic suturing skills to surgical training. However, existing automated assessment methods have the shortages in capturing long-term temporal dependencies and operational details, limiting the comprehensiveness and accuracy of assessments. To address this issue, an intelligent assessment framework based on spatiotemporal attention mechanism is proposed. Which adopts the TimeSformer as the backbone network to extract more discriminative spatiotemporal features, overcoming the difficulty in capturing long-term temporal action dependencies. Additionally, the AdaPool module is designed to adaptively aggregate video-level features, satisfying the sensitivity assessment requirements for key operational segments. Experimental results on the LapSuture dataset demonstrate that this method achieves an accuracy of 83.0% and an F_1-score of 81.6% in the skill grading task, outperforming traditional methods. The research results highlight the potential of attention-driven architectures in achieving robust, scalable, and objective surgical skill assessments, providing new insights for enhancing surgical education and competency evaluation.

Keywords: laparoscopic surgery; surgical skill assessment; deep learning; image processing; Transformer

0 引言

随着医疗技术的进步，腹腔镜手术因其创伤小、恢复快等优势在临床中得到广泛应用。然而，该类手术操作复杂，对外科医生的操作技能提出了极高要求[1]。研究表明，术者的熟练程度与并发症发生率密切相关[2]。因此，构建客观、精准的技能评估体系在外科培训和手术安全中具有重要意义。传统的外科规培通常包括两个

收稿日期:2025-10-28； **修回日期**:2025-12-08。
作者简介:陈士涵(2001-),男,硕士研究生。
通讯作者:夏建福(1983-),男,硕士,教授。
引用格式:陈士涵,夏建福,彭橼平,等. 基于时空注意力模型的腹腔镜手术缝合技能评估方法[J]. 计算机测量与控制,2026,34(10):282-287,295.

环节：一是规培医生或医学生在手术室协助主治医生参与临床操作；二是在微创技能训练系统上进行模拟训练[3]。训练过程结束后，主治医生或导师会根据其表现提供反馈，并结合标准化评估量表打分。虽然这一模式在实践中被广泛采用，但其评估结果往往依赖导师的主观判断，耗时、成本高且难以实现大规模应用。

为此，研究者们提出了多种自动化技能评估方法，主要分为两大类[4]：一类是基于运动学数据的方法，通过传感器或眼动仪采集操作过程中的运动轨迹和速度等参数，通过这些设备，可以精确捕捉到医生的操作细节，包括手术工具的运动轨迹、速度、力度等参数。例如，Karl-Friedrich 等人[5]使用传感器记录了不同经验水平的医生在进行腹腔镜结扎任务时的运动数据，结果表明，新手、中级和专家外科医生在时间、路径长度、平均速度、角度路径、运动次数以及每秒运动次数等多个方面存在显著差异。这些差异反映了不同技能水平医生在操作中的效率和精确度。另一项研究中 Kuo 等人[6]采用形状变换算法处理眼动追踪数据的时间序列，提取特征并将其作为分类器的输入，结合集成学习与卷积神经网络算法来检测医生的技能水平。

另一类是基于视频数据的方法，利用深度学习从手术视频中自动提取语义和时序信息。前者在区分不同水平的外科医生方面具有较好效果，但对昂贵设备的依赖限制了其临床应用；相比之下，视频方法因成本低、可扩展性强而更具潜力[7]。视频数据的使用通常依赖于两种主要策略：一种是通过追踪视频中的手术工具，分析其运动数据来评估手术技能。具体来说，许多研究采用了视觉工具跟踪算法来捕捉手术中工具的运动。例如，Jin 等人[8]采用基于区域的卷积神经网络进行工具的检测和定位，并通过分析工具的使用模式、运动范围以及运动的经济性来评估手术技能。另外，Adam 等人[9]对 YOLO 网络进行了修改，增加了多任务学习能力，使其不仅可以进行工具定位，还能检测手－工具的交互。他们进一步计算了工具的使用持续时间、路径长度、移动次数以及保持角度等指标来评估手术技能。另一种策略是直接从视频数据中学习特征表示。Hassan 等人[10]在 2018 年通过分析 JIGSAWS 数据集[11]中的手术动作，如缝合和穿针等，使用卷积神经网络来提取与手术技能相关的空间特征。这些特征包括边缘、纹理和形状等，与手术操作密切相关，可以有效反映医生的技能水平。此外，Funke[12]等人将时间分段网络（TSN，temporal segment networks）[13]与 3D 卷积神经网络（3DCNN，convolutional neural network）相结合，仅使用手术视频来将使用机器人器械的外科医生分为 3 个技能级别。TSN 框架通过对连续几帧视频的堆叠进行分类，再通过聚合片段级别的预测来得到整体视频的分类结果。Erim[14]等人则利用自监督对比学习模型从手术视频中提取时空特征，并结合一维残差神经网络进行手术结果的分类与性能评分预测。

然而，CNN 等传统结构在建模长时序依赖方面仍存在不足：卷积神经网络受限于局部感受野，难以捕捉全局上下文。腹腔镜手术操作具有强时序性，且包含多个阶段，每个阶段对医生的技术要求不同。CNN 结构由于其局部感受野的限制，无法有效建模整个操作过程中的长期依赖关系，尤其是在视频中跨越多个时间段的关键操作。鉴于传统方法的局限性，本文创新性地引入时空注意力模型，其在处理长时序视频数据及关键细节捕捉方面具有独特优势。具体而言，Transformer 架构能够通过全局自注意力机制捕捉视频中任意两个时间点和空间位置之间的依赖关系，非常适合于建模手术操作过程中复杂的长距离时序关联，从而提取更具判别性的技能特征。基于此，本研究提出一种基于时序建模的缝合技能智能评估框架。在 TSN 分段采样的基础上，引入 TimeSformer 作为主干网络，以更好地捕捉跨片段的长时程依赖与复杂视觉语义，从而提升技能判别特征的提取质量；同时，在视频级特征聚合阶段引入自适应指数池化（AdaPool，adaptive exponential pooling），通过在全局一致性与局部显著性之间动态平衡，突出关键操作片段，进而提高评估结果的准确性与可解释性。该设计不仅提升了视频表征的判别性与鲁棒性，也为实现端到端的智能化技能评估提供了新的思路。

1　腹腔镜手术缝合视频数据集构建

1.1　参与者与任务设定

本研究所使用的数据集 Lapsuture，是在与温州某医院外科合作下，于其微创技能训练中心专门采集和构建的。该数据集旨在为腹腔镜手术缝合技能的自动化评估提供一个真实、多样且具有专业评级的基准。

本研究共招募了 21 名不同技能水平的医生参与视频录制，并根据其临床经验划分为 3 个层级：15 名处于规范化培训阶段的规培医生；3 名已具备独立手术经验的住院医生；3 名经验丰富的主治医生。

所有参与者均被要求完成标准的腹腔镜缝合与打结任务。为保证任务的标准化和动作的完整性，每个完整的缝合过程被明确定义为 6 个连续的步骤：针头定位、针头处理、进针出针、打结 1（打线）、打结 2（打结）、剪线。所有参与者在执行任务前均已签署知情同意书，且本研究已获得温州某医院伦理委员会的批准。

1.2　采集设备与数据量

视频采集是在一套由上海惟肖医疗科技有限公司研发的多学科微创技能训练集成装置系统上进行的。该系统集成了多种专业硬件，确保了采集环境的标准化和数

据质量的一致性。主要设备包括：高清训练内窥镜成像系统、模拟体腔箱、专业训练器械套装。

共采集了 158 段高质量的腹腔镜缝合视频，图 1 为一次腹腔镜手术缝合视频的其中一帧。腹腔镜缝合视频与由经验丰富的主任医生根据缝合和打结错误的客观评价量表所完成的结构化评分数据共同构成了本研究的数据集。该评分数据记录在 Excel 电子表格中，评价量表对缝合过程中的各项任务进行多维度评分，涵盖器械使用、组织处理等方面。缝合任务共 4 项，包括：正确握持针头进行穿刺、通过组织两边进行针穿刺，确保咬合位置和组织处理得当、打结的第一次双圈以及打结操作。评分表中的每一项都与操作的具体细节密切相关，评分标准根据 1 到 4 的等级进行评定，其中 1 分表示最佳表现，4 分表示最差表现。每个学生和视频都有自己的标识符。对于每个视频，相应的学生证与由 3 名独立评分者评分的技能分数一起记录在 Excel 电子表格中。为建立数据的初步基准，分析重点聚焦于总分（TS，total score）。通过计算每个视频中各评分者的平均分，将样本划分为 3 个等级——新手、中级和熟练。具体划分标准为：新手（$24<TS<32$）、中级（$16<TS<24$）及熟练（$8<TS<16$）。基于此标准，数据集包含：初级 43 段，中级 47 段，熟练 68 段。

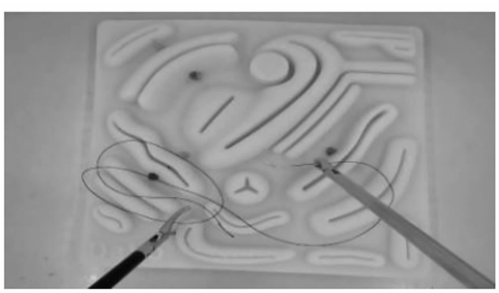


图 1 腹腔镜手术缝合视频帧

2 结合时空注意力模型的腹腔镜手术缝合技能评估模型

2.1 整体架构

在腹腔镜缝合等长时序、强阶段性的外科操作视频中，经典 TSN 通常将整段视频划分为 T 个时间片段，并对每个片段提取特征，分别生成类别或质量评分，最终采用“简单平均”作为视频级汇聚函数。平均聚合隐含假设“各片段对最终评估同等重要”。然而，缝合动作包含关键子阶段（如持针定位、穿刺入点、拉线收紧、打结定型）与非关键过渡阶段（如器械移动准备等），二者在信息密度与判别性上存在显著差异。平均聚合会使关键片段的判别性信号被冗余或噪声片段稀释，从而导致对错位或失败操作的敏感性下降；且难以在长视频中突出“短而关键”的操作失误。在评分或分级任务中，平均聚合尤其容易导致边界样本评估偏差，从而限制了 TSN 在外科技能评估中的性能上限。为克服上述缺陷，本研究采用 AdaPool[16] 替代 TSN 的简单平均汇聚策略，引入 TimeSformer[17] 作为主干网络，对腹腔镜缝合技能进行更具表现力的特征表示。

在以上两方面改进的基础上，本文将其整合为一个统一的框架，并命名为 Ada-TSN，如图 2 所示。其整体工作流程仍遵循 TSN 的分段采样思路：在训练过程中，输入视频被平均划分为若干个片段（Segment），每个片段中随机选取连续的帧（Snippet）作为样本，片段长度由预设参数决定。通过合理设置片段数与长度，使模型在训练时能够覆盖视频的大约一半内容。随后，来自同一视频的所有片段会被组合成一个批次，并依次送入主干网络 TimeSformer 进行特征提取和片段级别的预测。输入数据可表示为张量形式 $(B\times S)\times C\times T\times H\times W$，其中 B 为批量大小，S 为片段数，$C=3$ 为 RGB 通道数，T 为片段长度，H 与 W 分别为帧的高和宽。

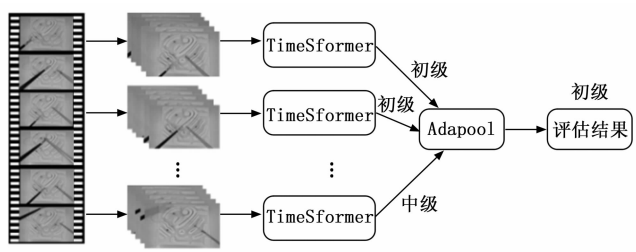


图 2 Ada-TSN 网络流程图

在视频级预测阶段，AdaPool 能够自适应地为不同片段分配权重，从而在聚合时突出关键操作片段、削弱冗余或无关片段的影响。最终，AdaPool 输出的视频级特征用于计算交叉熵损失，并通过网络进行反向传播。借助这种机制，模型不仅学习如何对片段进行准确分类，还能够在全局层面更精准地完成视频级技能评估。在推理阶段，模型会处理完整视频的所有片段，以实现更加稳定和可靠的技能分级。

2.2 主干网络

为了提升视频特征提取的时空建模能力，本文在传统 TSN 框架中引入 TimeSformer 作为主干网络，对腹腔镜缝合技能进行更具表现力的特征表示。

TSN 网络的特征提取阶段通常采用预训练的卷积神经网络从视频片段中提取时空特征。传统方法依赖双流 CNN 架构[18-19]，该架构包含两个并行分支：空间流网络处理 RGB 图像以捕获静态外观信息，时间流网络处理光流图像以建模动态运动模式。这种设计能够有效分离和融合视频中的空间与时间信息，在复杂视觉场景理解中展现出良好性能。然而，双流架构存在若干固有局限：首先，光流计算需要对原始视频进行密集的预处

理操作，通常采用 TVL1 或 Farneback 等算法逐帧估计像素级运动向量，显著增加了数据预处理的计算开销和存储需求；此外，两个流之间的信息交互往往局限于最后的融合层，缺乏深层次的时空特征协同建模能力。为克服上述限制，采用 TimeSformer 网络结构替代传统双流 CNN 作为特征提取骨干。

2.2.1　视频序列化与时空嵌入

TimeSformer 是一种完全基于自注意力机制[20]的视频理解模型，如图 3 所示，其核心思想是将视频表示为帧级 patch 序列，每个 patch 经过线性映射后形成 token，再输入 Transformer 编码器进行时空联合建模。与卷积操作依赖局部感受野不同，TimeSformer 能够直接计算任意时空位置间的全局依赖，从而更好地捕捉到外科操作中长距离的动作联系。其提出的“分离时空注意力（Divided Space-Time Attention）”机制在每个 Transformerblock 内分阶段执行时间注意力与空间注意力：首先在时间维度上对同一空间位置的不同帧建模以获取动作演化特征，然后在空间维度上对单帧的不同位置建模以提取操作细节。这种分而治之的注意力策略不仅在计算复杂度上优于联合时空注意力，还在多个视频分类基准上取得了最优表现。

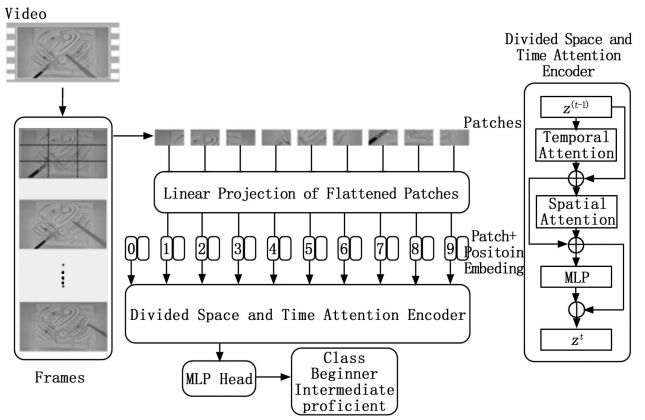


图 3　TimeSformer 网络结构图

TimeSformer 接收的输入为一段固定长度的视频片段 $X\in\mathbb{R}^{T\times H\times W\times C}$，其中 T 为帧数，H、W 为帧的空间分辨率，C 为通道数（本研究中 T 为 TSN 每个时间片段中所包含的连续视频帧数，输入分辨率调整为 224×224）。首先，将每一帧图像切分为 N 个不重叠的图块（Patch），每个图块的大小为 $P\times P$，则每帧的图块数量为 $N=(H\times W)/P^2$。随后，将这些图块展平并映射为一维向量，加上可学习的位置嵌入以保留时空位置信息。最终，输入序列被表示为 $Z^{(0)}\in\mathbb{R}^{T\cdot N\times D}$，其中 D 为隐藏层特征维度。

2.2.2　分离时空注意力机制

为了在计算效率与建模能力之间取得平衡，TimeSformer 采用“分离时空注意力”机制。该机制在每个 Transformer 编码器块中交替执行时间注意力（Temporal Attention）和空间注意力（Spatial Attention）。

假设第 l 层的输入特征为 $Z^{(l-1)}$，多头自注意力（MSA）的计算公式如下：

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

其中：$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 分别为查询、键和值矩阵，d_k 为注意力头的维度。

在 TimeSformer 的编码块中，计算过程分解为两步：时间注意力：仅在同一空间位置的不同时间帧之间计算注意力。此步骤用于捕捉如持针动作、打结轨迹等时间维度的动态演变。LN 为层归一化。

$$\hat{Z}^{(l)}_{\mathrm{time}}=\mathrm{MSA}_{\mathrm{time}}[\mathrm{LN}(Z^{(l-1)})]+Z^{l-1} \tag{2}$$

空间注意力：仅在同一帧内的不同空间图块之间计算注意力。此步骤用于聚合手术器械与组织的空间上下文信息。

$$Z^{(l)}=\mathrm{MSA}_{\mathrm{space}}[\mathrm{LN}(\hat{Z}^{(l)}_{\mathrm{time}})]+\hat{Z}^{(l)}_{\mathrm{time}} \tag{3}$$

本研究采用 TimeSformer-Base 架构配置，具体参数设置如下：Transformer 层数 $L=12$，隐藏层维度 $D=768$，注意力头数 $A=12$，图块大小 $P=16$。该配置在保证特征提取深度的同时，有效控制了计算开销。

2.2.3　变长视频输入的适配策略

Lapsuture 腹腔镜手术视频的时长差异较大，直接输入固定长度的模型难以适用。为此，本框架采用了稀疏采样适配策略（见 2.1 节）：遵循 TSN 的采样逻辑，将任意长度的原始视频平均划分为 S 个片段。从每个片段中随机抽取 T 帧组成一个输入张量。无论原始视频多长，输入到 TimeSformer 模型的始终是规范化的 $(B\times S)\times C\times T\times H\times W$。对于输入帧数 T 与预训练模型设定不一致的情况，采用双线性插值法对时间位置编码进行调整，以适配不同时间分辨率的输入需求。通过上述机制，模型既能利用 Transformer 强大的全局建模能力，又能灵活处理不同时长的手术视频，实现对全过程技能的覆盖评估。

2.3　自适应聚合策略

本研究采用 AdaPool 替代 TSN 的简单平均汇聚策略。相较于简单平均将所有片段一视同仁，AdaPool 能够动态地识别并突出视频中那些对最终技能评估具有决定性意义的‘关键操作’片段，同时降低冗余或非关键片段的权重，从而实现更精准的视频级汇聚。AdaPool 的核心思想在于学习两组池化核的区域特定融合：基于指数化 Dice-Sørensen 系数的池化核与基于指数最大值的池化核，并通过可学习的权重掩码实现自适应加权。具体而言，在接收 TSN 骨干网络为 T 个视频片段生成的特征 x_1，x_2，x_3，…，x_i 后，AdaPool 将其作为输

入，并计算两种池化结果，以捕获视频流中不同性质的时序信息。

1）指数化 Dice-S∅rensen 系数加权池化：基于片段特征与区域均值的相似性进行加权，其公式为：

$$\mathrm{eDSCW}=\sum_{i\in R}\left\{x_i\cdot\frac{\exp[DSC(\boldsymbol{x}_i,\bar{\boldsymbol{x}})]}{\sum_j\exp[DSC(\boldsymbol{x}_j,\bar{\boldsymbol{x}})]}\right\}\tag{4}$$

其中：$\boldsymbol{x}_i$ 为第 i 个片段的特征向量，$DSC(\boldsymbol{x}_i,\bar{\boldsymbol{x}})$ 表示第 i 个片段特征与区域均值的 Dice-S∅rensen 相似性系数，可根据特征相似性分配权重，保持区域一致性。

2）指数最大值池化（eM）：此策略专注于突出视频中“短而关键”的高响应片段。它类似一种可微分的最大池化，能通过指数函数放大显著激活的片段，有效捕捉那些对最终技能评估具有决定性作用的时刻。

$$\mathrm{eM}=\sum_{i\in R}\left[x_i\cdot\frac{\exp(x_i)}{\sum_j\exp(x_j)}\right]\tag{5}$$

该方法基于 Softmax 机制，确保高响应片段的权重远高于低响应片段。

3）自适应融合：

AdaPool 通过可学习的权重掩码 β，在两种池化方式间自适应插值，公式为：

$$\mathrm{AdaPool}_c=\beta\cdot\mathrm{eDSCW}_c+(1-\beta)\cdot\mathrm{eM}_c\tag{6}$$

其中：$\beta\in[0,1]$ 是与空间位置相关的可学习参数，可根据不同区域特性自动调整两种池化的贡献比例。当 β 接近 1 时，模型倾向于使用 Dice-S∅rensen 池化，保持时序连贯性；当 β 接近 0 时，则更侧重于突出高响应片段，捕捉关键瞬间。

可学习权重 β 是实现 AdaPool 自适应融合机制的核心，其优化过程直接影响模型的收敛速度和评估的稳健性。

1）参数初始化和更新：

由于 β 必须严格约束在 $[0,1]$ 范围内以保证权重分配的物理意义，采用 Sigmoid 函数对一个无约束的原始可学习参数 w 进行映射：

$$\beta=\mathrm{Sigmoid}(w)=\frac{1}{1+e^{-w}}\tag{7}$$

在初始化阶段，原始参数 w 被初始化为 0。根据 Sigmoid 函数的特性，这意味着可学习权重 β 被初始化为 0.5。这种初始化策略保证了在训练开始时，模型对 eDSCW_c 和 eM 两种池化方式采取平均分配的策略，避免了先验偏置，让模型从数据中自主学习最优的融合比例。可学习参数 w 作为整个 Ada-TSN 框架的一部分，直接参与反向传播过程。它的更新策略与其他网络参数（如 TimeSformer 的权重）一致。

2）过拟合与权重失衡避免：

为避免 β 在训练初期就过度拟合于训练数据，导致其值迅速趋向 0 或 1，本研究在优化过程中对原始参数 w 施加了 L2 正则化。

$$L_{\mathrm{total}}=L_{CE}+\lambda\cdot\|w\|_2^2\tag{8}$$

其中：L_{CE} 为交叉熵损失，λ 为正则化系数。

综上所述，AdaPool 的自适应融合机制使其能够超越传统平均聚合的局限性，它赋予了 TSN 汇聚层“动态关注”的能力，在缝合视频中根据内容重要性进行加权，最终显著提升了模型对关键操作的捕捉能力，从而大幅改善腹腔镜缝合手术评估的准确性与鲁棒性。

3 实验结果与评估

3.1 实验环境

本研究中，实验操作系统为 Windows 11，GPU 为 NVIDIA GeForce GTX 3090，采用的开发环境为 Python 3.9、Pytorch 1.8.0、CUDA 11.1。

3.2 参数设置及评估指标

实验在 LapSuture（158 个视频）数据集上训练，轮次共 1 200 轮，学习率 0.001，批次 64。TSN 中的共识函数选择 Adapool 共识策略，如式（9）所示：

$$x_{\mathrm{Ada}}=\beta\cdot\sum_{i\in R}\frac{\exp[\mathrm{DSC}(x_i,\bar{x})]}{\sum_{j\in R}\exp[\mathrm{DSC}(x_j,\bar{x})]}\cdot x_i+(1-\beta)\cdot\sum_{i\in R}\frac{\exp(x_i)}{\sum_{j\in R}\exp(x_j)}\cdot x_i\tag{9}$$

其中：x_i 表示第 i 个时间片段的特征或预测结果。

损失函数为多分类交叉熵损失函数，其中 y_i 和 $\hat{Y}_i$ 分别是样本真实类别和预测概率，C 为类别总数：

$$L(y,\hat{y})=-\sum_{I=1}^{c}y_i\log(\hat{y}_i)\tag{10}$$

在后续性能评估中主要使用 F_1 分数和 $Accuracy$ 准确率作为算法的评估指标，公式如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}\tag{11}$$

F_1 分数是精确率（Precision）和召回率（Recall）的调和平均值，它提供了一个综合评价模型性能的方法，特别是在类别分布不均衡的情况下更为适用。$Precision$ 和 $Recall$ 如（12）和（13）所示：

$$Precision=\frac{TP}{TP+FP}\tag{12}$$

$$Recall=\frac{TP}{TP+FN}\tag{13}$$

所以 F_1 的计算公式为：

$$F_1=2\times\frac{Precision\times Recall}{Precision+Recall}\tag{14}$$

3.3 实验结果与分析

为全面评估所提出 Ada-TSN 框架的有效性，在 LapSuture 数据集上进行了一系列实验。本节首先将该方法与基线模型进行整体性能对比，随后通过组件分析以剥离式地验证每个创新模块的贡献，最后探讨了关键

超参数的影响。

3.3.1　整体性能对比

为了证明模型改进的有效性，首先对 Ada-TSN 网络与传统的 TSN 网络在 Lapsuture 手术缝合数据集上进行性能对比，实验结果如表 1 所示。

表 1　传统 TSN 与改进 TSN 性能对比

Model	Snippet	Segment	Acc/%	F_1/%
Ada-TSN	32	10	83	81.6
TSN[12]	32	10	75	74.3

由表 1 可知，该方法在准确率和 F_1 上均显著优于传统基线模型。具体而言，准确率提升了 8.0%，F_1 分数提升了 7.3%。F_1 的显著提升尤为关键，它表明本模型在数据集类别分布可能不平衡的情况下，能更好地在精确率与召回率之间取得平衡。这一结果强有力地证明了，将 TimeSformer 的全局时空建模能力与 AdaPool 的自适应聚焦机制相结合，能够更有效地从腹腔镜缝合视频中提取与技能水平高度相关的判别性特征。除此之外，为了验证每个模块的性能，还针对改进后的模块进行实验。

3.3.2　关键模块贡献分析

为进一步厘清两个核心模块——TimeSformer 骨干网络与 AdaPool 聚合模块——各自的贡献，设计了对比实验。

首先，评估了 TimeSformer 骨干网络相对于传统卷积网络的优越性。在均使用简单平均聚合的相同条件下，对比 TimeSformer 与 3D-CNN 的表现。结果如表 2 所示。

表 2　主干网络性能对比

Model	Snippet	Segment	Acc/%	F_1/%
TSN(TimeSformer)	32	10	81.3	78.9
TSN(3DCNN)	32	10	75.0	74.3

表 2 的结果展示了 TimeSformer 的强大表征能力。在相同的聚合条件下，仅将主干网络从 3D-CNN 替换为 TimeSformer，即可带来准确率 6.3%和 F_1 分数 4.6%的显著提升。这充分证明了时空注意力架构在建模腹腔镜缝合这类长时序、强上下文依赖任务中的巨大优势，其捕获的全局时空特征远比 3D-CNN 的局部特征更具判别力。

其次，评估了 AdaPool 模块的有效性。在保持 TimeSformer 作为骨干网络不变的情况下，将视频级汇聚策略从 AdaPool 替换为 TSN 默认的简单平均法（Averaging），结果如表 3 所示。

表 3 的结果表明，仅通过引入 AdaPool 模块，模型性能便获得了稳定的提升：准确率提高了 1.7%，F_1 分数提高了 2.7%。F_1 分数的更大增幅说明 AdaPool 通过动态加权突出关键片段、抑制冗余信息，不仅提升了整体分类精度，更显著增强了模型对关键技能动作的识别能力与分类的稳健性。

表 3　自适应聚合模块性能对比

Model	Snippet	Segment	Acc/%	F_1/%
Ada-TSN	32	10	83.0	81.6
TSN(TimeSformer)	32	10	81.3	78.9

3.3.3　超参数影响分析

如表 4 所示，尽管 Snippets 和 Segments 是网络训练过程中可调节的参数，但经过多次实验对比，结果表明改变片段长度和片段数量对模型性能有微妙的影响。模型的最佳准确率波动在 82.6%～83.5%之间，最大波动幅度小于 0.9%。这一现象说明：1）Segment 数量的增加通常有助于模型捕捉更丰富的时序上下文，从而带来性能的轻微提升；2）Snippet 长度的增加可能在带来信息冗余的同时也增加了优化难度，导致性能略有回调。

表 4　不同 Snippets 值和 Segments 值的网络性能对比

Model	Snippet	Segment	Acc/%	F_1/%
Ada-TSN	32	10	83.0	81.6
Ada-TSN	64	10	82.6	81.0
Ada-TSN	32	12	83.5	81.8
Ada-TSN	64	12	83.2	81.3

尽管超参数的变化会对模型性能产生可观察到的影响，但在一个合理的取值区间内，本文提出的 Ada-TSN 框架始终保持了稳定的高性能水平，性能波动幅度有限。

4　结束语

本文提出的 Ada-TSN 框架通过引入 TimeSformer 和 AdaPool 模块，显著提升了腹腔镜手术技能的自动化评估能力。该方法有效解决了传统视频评估方法在长时序建模中的不足。尽管本研究已取得一定成果，未来仍可在以下两个方面进一步改进：一是扩大数据集范围，提升模型的泛化能力；二是优化模型的计算效率，减少训练时间。最终，本文的工作为智能化手术评估提供了新的思路，推动了医疗领域中更高效、客观的技能评估体系的建设，对外科教育与医疗质量的提升具有重要意义。

参考文献：

[1] YAMAZAKI Y，KANAJI S，KUDO T，et al. Quantitative comparison of surgical device usage in laparoscopic gastrectomy between surgeons′skill levels：an automated analysis using a neural network [J]. Journal of Gastrointestinal Surgery，2022，26 (5)：1006－1014.

（下转第 295 页）