中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Hu Zhonghua, Li Xiaoning, Song Jiangling, Zhang Rui. Cross-scale temporal interaction and boundary probability refinement for continuous action segmentation and prediction[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260266.(胡中华, 李晓宁, 宋江玲, 张瑞. 跨尺度时序交互与边界概率细化的连续动作分割与预测[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260266.)[DOI:10.11834/jig.260266]

跨尺度时序交互与边界概率细化的连续动作分割与预测

胡中华,李晓宁,宋江玲,张瑞

西北大学大数据与认知计算研究中心,陕西西安　710127

摘　要:**目的**　针对连续视频动作分割与预测任务中局部动作动态与长时语义上下文之间交互建模不足、连续动作过渡阶段边界表征不足以及未来序列预测稳定性不足等问题,提出一种融合跨尺度时序交互与边界概率细化的连续动作分割与预测方法。**方法**　面向动作分割与预测两类任务,构建由局部时空编码、跨尺度时序交互和边界概率细化组成的统一框架。对于分割任务,采用膨胀三维卷积网络(inflated 3D convolutional network,I3D)对RGB视频序列进行时空特征编码;对于预测任务,采用X3D网络对RGB序列与光流序列同时进行时空特征提取,并通过双向交叉注意力机制实现多模态信息融合。在此基础上,设计跨尺度时序交互模块,对不同层级时空特征进行编码—解码式建模,刻画不同时间尺度下的上下文依赖关系,增强模型对局部动态变化与全局时序结构的协同表征能力;进一步引入边界概率细化模块,对动作边界进行显式建模,从而提升动作分割与预测结果的稳定性及边界定位精度。**结果**　在50Salads、GTEA、Breakfast等生活场景连续动作分割数据集以及JIGSAWS手术手势数据集上进行了实验验证。所提方法在50Salads、GTEA、Breakfast和JIGSAWS分割任务上的准确率(accuracy,Acc)分别达到92.3%、88.0%、83.2%和90.5%,Edit分别达到89.0%、95.8%、84.7%和95.2%;在JIGSAWS手术动作预测任务中,预测准确率达到90.3%。实验结果表明,所提方法在多数指标上优于现有代表性方法。**结论**　所提方法能够有效兼顾局部动态特征建模、长程时序依赖建模与边界结构优化,在连续动作分割与预测任务中表现出良好的泛化能力与稳定性。

关键词:连续动作分割与预测;跨尺度时序交互;边界概率细化;时空特征编码;多模态融合

Cross-scale temporal interaction and boundary probability refinement for continuous action segmentation and prediction

Hu Zhonghua, Li Xiaoning, Song Jiangling, Zhang Rui

Research Center for Big Data and Cognitive Computing, Northwest University, Xi'an, Shaanxi 710127, China

Abstract: **Objective** Continuous action segmentation and prediction are important for human-computer interaction, intelligent perception, and surgical activity understanding. Continuous action videos usually contain rapid local motion changes, long-range temporal dependencies, ambiguous transitions, and large variations in action duration. Although existing meth-

收稿日期:2026-05-13;**修回日期**:2026-09-03

基金项目:陕西基础科学(数学、物理学)研究院基础研究计划项目(项目编号:22JSZ008,25JSY036);陕西省自然科学基金项目(项目编号:2025JCYBMS060);陕西省教育厅重点科学研究计划项目(项目编号:24JR156)。

Supported by: Shaanxi Fundamental Science Research Institute for Mathematics and Physics Basic Research Program Project (Grant Nos. 22JSZ008 and 25JSY036); Natural Science Foundation of Shaanxi Province (Grant No. 2025JCYBMS060); Key Scientific Research Program of the Shaanxi Provincial Department of Education (Grant No. 24JR156).

ods based on three-dimensional convolutional networks, recurrent networks, temporal convolutional networks, or Transformer structures have improved spatiotemporal representation learning, they still have difficulty in jointly modeling short-term dynamic details and long-term semantic context. Weak boundary cues in transition regions may also cause temporal jitter, over-segmentation, and inaccurate boundary localization. For action prediction, future categories must be inferred from partially observed sequences, which further requires stable temporal modeling and effective representation of motion trends. To address these issues, a video action segmentation and prediction method is proposed based on cross-scale temporal interaction and boundary probability refinement, aiming to improve discriminative representation, temporal consistency, boundary localization, and future prediction stability. **Method** The proposed method establishes a unified framework consisting of local spatiotemporal encoding, cross-scale temporal interaction, and boundary probability refinement. For action segmentation, RGB video sequences are encoded by an I3D network to obtain frame-level spatiotemporal features. For action prediction, RGB and optical flow sequences are used as dual-modal inputs, and an X3D network is used to extract features from both modalities. A bidirectional cross-attention mechanism enhances multimodal interaction, in which RGB features aggregate motion information from optical flow features and optical flow features obtain complementary appearance information from RGB features. A gated fusion strategy integrates the enhanced RGB and flow representations. Based on the extracted feature sequence, the cross-scale temporal interaction module adopts an encoder-decoder structure to model temporal dependencies at different hierarchical levels. The encoder aggregates global contextual information, whereas the decoder fuses high-level context with lower-level temporal details, thereby preserving both long-range dependencies and fine-grained local dynamics. The boundary probability refinement module explicitly models action transition boundaries through refinement stages based on dilated convolution and residual connections. The refined features are mapped to frame-level boundary probabilities. A weighted binary cross-entropy loss is used for boundary supervision, and a boundary-gated temporal consistency constraint suppresses unnecessary prediction fluctuations within action segments while allowing category changes near true boundaries. The whole framework is optimized by classification loss, boundary prediction loss, and boundary-gated temporal consistency loss. **Result** Experiments were conducted on four public continuous action segmentation datasets, including 50Salads, GTEA, Breakfast, and JIGSAWS. Frame-wise accuracy, edit score, and F1 scores under different intersection-over-union thresholds were used as evaluation metrics. On 50Salads, the proposed method achieved 93.6, 92.2, and 88.8 in F1@10, F1@25, and F1@50, respectively, with an edit score of 89.0 and an accuracy of 92.3. On GTEA, the method obtained the best results in F1@10, F1@25, edit score, and accuracy, indicating its effectiveness for fine-grained first-person action segmentation. On Breakfast, the method achieved 86.4 in F1@10, 82.5 in F1@25, 84.7 in edit score, and 83.2 in accuracy, demonstrating robustness under complex long-sequence conditions. On JIGSAWS, it achieved 97.1, 95.7, and 90.6 in F1@10, F1@25, and F1@50, respectively, with an edit score of 95.2 and an accuracy of 90.5, suggesting that it can effectively model fine-grained surgical action sequences. For action prediction, the method achieved an accuracy of 90.3 on JIGSAWS, outperforming representative prediction methods. Ablation experiments confirmed the contribution of each key component. Removing the cross-scale temporal interaction module caused performance degradation on all datasets, especially in F1@50 and edit score, indicating that cross-scale temporal modeling improves boundary recovery and sequence-level consistency. Removing the boundary probability refinement module also reduced performance, particularly in high-threshold F1 and edit score, verifying that explicit boundary modeling can reduce over-segmentation and improve transition localization. RGB and optical flow also showed complementary effects in action prediction, and their fusion achieved higher accuracy than either modality alone. **Conclusion** The proposed method integrates local spatiotemporal encoding, cross-scale temporal interaction, and boundary probability refinement into a unified framework for continuous video action segmentation and prediction. By combining short-term dynamic information with long-range temporal context, the model enhances the representation of action evolution across different temporal scales. By explicitly learning boundary probabilities and introducing boundary-gated temporal consistency constraints, it improves frame-level classification stability and action boundary localization. For action prediction, the fusion of RGB appearance information and optical flow motion information strengthens the perception of future action trends. Experimental results on multiple public datasets demonstrate that the proposed method achieves competitive or superior performance compared with existing methods, especially in temporal consistency, boundary local-

ization，and future action prediction. The method provides an effective solution for continuous action understanding in complex video scenarios.

Key words: continuous action segmentation and prediction; cross-scale temporal interaction; boundary probability refinement; spatiotemporal feature encoding; multimodal fusion

论文引用格式:[DOI:10. 11834/jig. 260266]

0 引　言

连续动作分割与预测是视频理解领域的重要研究任务，旨在对连续视频中的动作片段进行帧级或片段级类别标注，并根据已观测序列推断未来动作状态(Abu Farha 等，2019)，在人机交互、智能感知、生活行为理解和手术辅助分析等场景中具有重要应用价值。

早期连续动作分割研究通常将输入视频划分为单帧图像或短时间的视频片段，并利用三维卷积神经网络(3D convolutional neural network，3D CNN)对片段进行建模，以提取局部时空联合特征，从而实现对动作的识别。常用的3D CNN模型包括C3D(Tran 等，2015)、I3D(Carreira 等，2017)以及X3D(Feichtenhofer，2020)等。这类方法在刻画短时运动模式方面具有较强能力。然而，动作本质上是一种连续动态行为，动作之间往往存在明显的上下文依赖、过渡关系以及较长时间范围内的时序变化。因此，仅依赖单帧视频图像(即RGB视频图像)或短时视频片段特征来完成孤立动作分割的方法，难以充分刻画连续视频中相邻动作之间的时间关联与较长时间范围内的时序变化。

基于此，越来越多的研究开始关注对连续视频中帧级特征或片段级特征的时序关系建模，即引入专门的序列建模模块，以进一步捕获跨时间范围的上下文依赖信息。例如，基于长短时记忆网络(long short-term memory，LSTM)及其变体的连续动作分割方法(Hochreiter 等，1997；Donahue 等，2015)，该类方法通过递归结构对视频特征进行时序累积，能够在一定程度上建模动作演化过程中的长期依赖；基于3D CNN与LSTM相结合的相关方法(Molchanov 等，2016；Núñez 等，2018)，通过融合局部时空表征与全局时序建模能力，已成为连续动作分割中的经典框架。此外，近年来，基于Transformer及多尺度上下文建模的相关研究逐渐增多。这类方法通常借助注意力机制、扩张卷积或多尺度特征融合等策略，加强对长距离时序依赖及多尺度动作变化的建模能力，在连续动作分割和相关时空动作理解任务中也取得了一定进展(Men 等，2026；程勇 等，2025；任小龙 等，2025)。然而，这类方法在强化长距离时序依赖建模的同时，往往更侧重全局上下文信息的聚合，而对短时间范围内的局部动态变化及动作过渡阶段关注不足，难以兼顾不同时间尺度下的动作演化特征，导致模型在捕捉细粒度局部动态变化、整体时序演化关系以及相邻动作边界变化时仍存在不足，从而影响动作分割的连贯性与边界定位精度。

与此同时，预测任务作为分割与识别任务的延伸，需要在仅获得部分历史信息的情况下推断未来动作类别，因此对时序建模能力、未来动作变化趋势的刻画能力以及长程上下文依赖的捕捉提出了更高要求。现有方法一方面聚焦于仅利用视频模态信息，通过增强时序建模能力来提高未来动作预测性能，另一方面也逐渐从单一RGB视频信息建模拓展到融合视觉与运动学等多源信息的多模态建模(Qin 等，2020；Shi 等，2022；Weerasinghe 等，2024；Chen 等，2025)。然而，在实际应用场景中，额外模态信息并不总是易于获取；同时，仅依赖RGB图像进行建模，虽然能够表征场景外观与动作语义信息，却仍难以充分捕捉连续帧之间的细粒度动态变化趋势。因此，如何在仅利用视频信息的条件下进一步增强对未来动作变化趋势的刻画能力，尤其是在动作流程复杂、阶段衔接紧密的手术场景中，仍是动作预测研究中需要解决的关键问题。

综上所述，尽管现有方法在连续动作分割与预测任务中取得了较好的效果，但仍面临两方面挑战。一方面，现有方法往往侧重于增强长时依赖建模能力，尽管能够在较大时间范围内捕获动作演化关系，但对短时间范围内的局部动态变化关注不足，难以兼顾不同时间尺度下的动作变化特征，从而影响模型对细粒度动作的判别能力。另一方面，连续动作分割与预测中相邻动作切换频繁、边界区域模糊，现有方法在边界附近仍易出现预测抖动、类别过渡不

清晰和过分割等问题，进而影响整体结果的连贯性与边界定位精度。进一步，在手术动作预测任务中，仅依赖RGB视频图像往往难以直接反映相邻帧之间的运动方向和速度变化，从而限制了模型对未来动作趋势的建模能力。

为解决上述问题，本文提出一种融合跨尺度时序交互与边界概率细化的连续动作分割与预测方法。在分割任务中，采用时空特征提取网络获取视频序列中的局部动态表示，在此基础上引入跨尺度时序交互策略，对不同时间尺度下的动作演化信息进行联合建模，以增强模型对短时变化信息和长时依赖信息的融合能力。同时，针对连续动作过渡阶段的时序边界表征不足的问题，进一步设计边界概率细化策略，通过显式学习动作起止边界信息，对帧级分类结果进行约束与修正，从而提升动作边界的定位精度与时序一致性，最终提高动作分割的准确性与稳定性。对于预测任务，在共享上述跨尺度时序交互与边界概率细化框架的基础上，融合视频光流信息，以增强模型对未来动作动态变化趋势的刻画能力，从而提升手术动作预测结果的准确性与稳定性。

本文的主要创新点和工作总结如下：

1）提出一种跨尺度时序交互方法，用于解决连续动作分割与预测中长短时间依赖难以兼顾的问题。该方法通过联合建模局部短时动态变化与全局长程依赖关系，增强了模型对不同时间尺度动作演化特征的表达能力。

2）针对连续动作分割与预测中普遍存在的边界抖动与过分割问题，提出一种边界概率细化方法。通过显式建模动作起止边界信息，对帧级分类结果进行约束与修正，从而提高动作分割结果的时序一致性与边界定位精度。

3）设计了一种融合RGB图像与光流信息的动作预测方法，通过引入显式运动信息增强模型对未来动作变化趋势的感知能力，从而提高未来动作预测的准确性与稳定性。

4）在多场景数据集上对所提方法的有效性进行了验证，实验结果表明所提方法既能够在连续动作分割任务中取得更稳定的分割结果与更准确的边界划分，也能在手术动作预测任务中取得更优的预测性能。

1 本文方法

1.1 总体网络框架

整体框架如图1所示，主要由局部时空编码模块、跨尺度时序交互模块和边界概率细化模块组成。首先，在视频特征提取阶段，对输入视频序列进行局部时空模式编码，并通过跨尺度时序交互机制学习动作在不同时间尺度下的动态变化及其上下文依赖关系，从而获得具有判别性的动作时序特征表示，并生成初始的分类结果。随后，在动作标签序列优化阶段，引入边界概率建模方法，对动作起止边界进行显式刻画与约束，对初始分类结果进行逐阶段细化与修正，从而提升动作分段的稳定性和边界定位精度。

1.2 局部时空模式编码模块

连续视频中的动作片段通常具有持续时间短、局部动态变化快等特点。为提取短时间范围内的时空变化信息，在分割任务和预测任务中分别采用I3D和X3D网络作为局部时空模式编码模块，对输入视频序列进行初步特征提取。其中，分割任务以RGB视频序列为输入，采用I3D网络提取局部时空特征；在预测任务中，为了增强对未来动作变化趋势的刻画能力，采用X3D网络分别提取RGB视频序列和光流序列的局部时空特征。

对于动作分割任务，首先将给定视频通过固定间隔抽帧方法转化为RGB视频序列 $\boldsymbol{X}^{\mathrm{r}}$，进而采用I3D网络提取其对应的帧级（局部）时空特征序列：$\boldsymbol{X}=\left\{\boldsymbol{x}_1^{\mathrm{r}},\boldsymbol{x}_2^{\mathrm{r}},\cdots,\boldsymbol{x}_T^{\mathrm{r}}\right\},\boldsymbol{x}_t^{\mathrm{r}}\in\mathbb{R}^D$。式中，$\boldsymbol{X}$ 表示I3D网络提取的RGB时空特征序列，T 表示视频序列长度，$\boldsymbol{x}_t^{\mathrm{r}}$ 表示第 t 帧对应的RGB时空特征向量，D 为特征维度，$\mathbb{R}$ 表示实数集。

对于动作预测任务，在上述基础上，采用TV-L1（total variation-L1）光流算法计算每个像素的水平与垂直位移，用于反映相邻帧之间的运动方向和速度变化，记为 $\boldsymbol{X}^{\mathrm{f}}=(\boldsymbol{x}_1^{\mathrm{f}},\cdots,\boldsymbol{x}_T^{\mathrm{f}})$，进而采用X3D网络分别提取 $\boldsymbol{X}^{\mathrm{r}}$ 和 $\boldsymbol{X}^{\mathrm{f}}$ 的帧级（局部）时空特征 $\boldsymbol{F}^{\mathrm{r}}$ 和 $\boldsymbol{F}^{\mathrm{f}}$：

$$\boldsymbol{F}^{\mathrm{r}}=e\left(\boldsymbol{X}^{\mathrm{r}}\right),\boldsymbol{F}^{\mathrm{f}}=e\left(\boldsymbol{X}^{\mathrm{f}}\right)\#(1)$$

式中，$e(\cdot)$ 表示X3D特征提取函数，r和f分别表示RGB模态和光流模态。

对于提取的 $\boldsymbol{F}^{\mathrm{r}}$ 和 $\boldsymbol{F}^{\mathrm{f}}$，引入双向交叉注意力融合

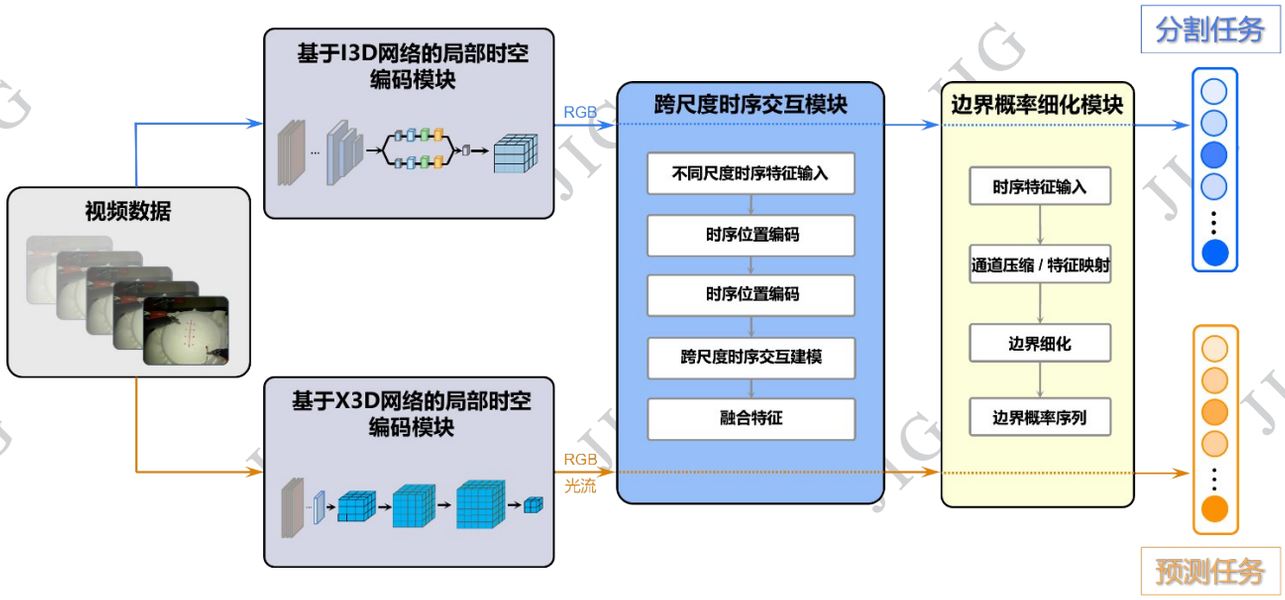

图1　总体网络框架

Fig. 1　Overall network framework

模块对RGB特征和光流特征进行互补建模，如图2所示。

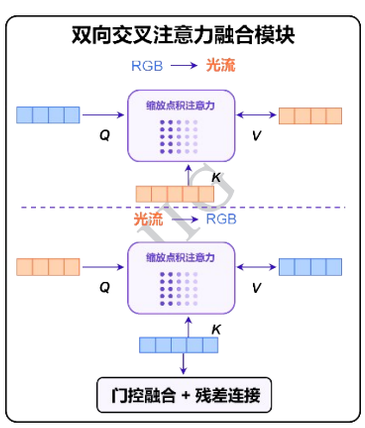

图2　双向交叉注意力融合模块

Fig. 2　Bidirectional cross-attention fusion module

具体而言，给定输入特征 $\boldsymbol{F}\in\{\boldsymbol{F}^{\mathrm{r}},\boldsymbol{F}^{\mathrm{f}}\}$，首先通过一维卷积进行线性映射，得到查询、键和值：

$$\boldsymbol{Q}=\boldsymbol{W}_{\mathrm{q}}\boldsymbol{F},\boldsymbol{K}=\boldsymbol{W}_{\mathrm{k}}\boldsymbol{F},\boldsymbol{V}=\boldsymbol{W}_{\mathrm{v}}\boldsymbol{F}\#(2)$$

式中，$\boldsymbol{W}_{\mathrm{q}},\boldsymbol{W}_{\mathrm{k}},\boldsymbol{W}_{\mathrm{v}}$ 分别为查询、键和值的线性映射权重矩阵；$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别表示查询、键和值矩阵。

本文采用基于缩放点积的单头注意力机制，而非标准Transformer多头注意力，其计算形式为：

$$a\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{d}}\right)\boldsymbol{V}\#(3)$$

式中，d 表示键向量的特征维度。

为增强两种模态之间的信息交互，分别从两个方向构建交叉注意力。一方面，以RGB特征为主导，引入光流特征中的运动变化信息。具体地，以RGB特征 $\tilde{\boldsymbol{F}}^{\mathrm{r}}$ 作为主查询，利用交叉注意力函数 $a(\cdot)$ 从光流特征 $\tilde{\boldsymbol{F}}^{\mathrm{f}}$ 中自适应聚合与其相关的运动信息，生成增强后的RGB表示 $\hat{\boldsymbol{X}}^{\mathrm{r}}$，即：

$$\hat{\boldsymbol{X}}^{\mathrm{r}}=a\left(\tilde{\boldsymbol{F}}^{\mathrm{r}},\tilde{\boldsymbol{F}}^{\mathrm{f}}\right)\#(4)$$

另一方面，光流特征 $\tilde{\boldsymbol{F}}^{\mathrm{f}}$ 作为主查询，通过交叉注意力机制从 $\tilde{\boldsymbol{F}}^{\mathrm{r}}$ 中提取互补外观信息，得到增强表示 $\hat{\boldsymbol{X}}^{\mathrm{f}}$：

$$\hat{\boldsymbol{X}}^{\mathrm{f}}=a\left(\tilde{\boldsymbol{F}}^{\mathrm{f}},\tilde{\boldsymbol{F}}^{\mathrm{r}}\right)\#(5)$$

通过上述双向交互，RGB分支能够关注光流特征中的关键运动变化，光流分支也能够结合RGB特征中的空间外观语义，从而增强两种模态之间的信息互补。

在获得双向增强特征后，进一步采用门控融合策略对RGB与光流特征进行自适应整合。即基于两种模态的联合响应生成门控权重 $\boldsymbol{G}$：

$$\boldsymbol{G}=\sigma\left(\boldsymbol{W}_{\mathrm{g}}\left[\hat{\boldsymbol{X}}^{\mathrm{r}};\hat{\boldsymbol{X}}^{\mathrm{f}}\right]+\boldsymbol{b}_{\mathrm{g}}\right)\#(6)$$

式中，$\boldsymbol{W}_{\mathrm{g}}$ 表示门控映射的权重矩阵，$\boldsymbol{b}_{\mathrm{g}}$ 为偏置向量，$\sigma(\cdot)$ 表示Sigmoid激活函数。再利用该权重对 $\hat{\boldsymbol{X}}^{\mathrm{r}}$ 与 $\hat{\boldsymbol{X}}^{\mathrm{f}}$ 进行自适应加权融合，并结合残差连接与层归一化（layer normalization，LN）操作，得到稳定的融合特

征表示 $\boldsymbol{X}^{\mathrm{p}}$：

$$\boldsymbol{X}^{\mathrm{p}}=n\left(\boldsymbol{G}\odot\hat{\boldsymbol{X}}^{\mathrm{r}}+(1-\boldsymbol{G})\odot\hat{\boldsymbol{X}}^{\mathrm{f}}+\hat{\boldsymbol{X}}^{\mathrm{r}}+\hat{\boldsymbol{X}}^{\mathrm{f}}\right)\#(7)$$

式中，$n(\cdot)$ 表示层归一化操作，1 表示与门控权重矩阵 $\boldsymbol{G}$ 维度相同的全1矩阵，$\odot$ 表示逐元素乘法。

1.3 跨尺度时序交互模块

在获得局部时空模式编码得到的帧级特征序列后，为进一步建模连续动作序列中的长程时序依赖关系，引入跨尺度时序交互模块对序列特征进行高层时序建模，如图3所示。

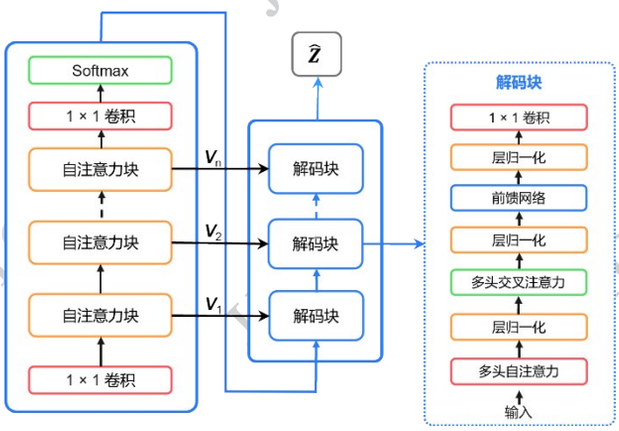

图3 跨尺度时序交互模块

Fig. 3 Cross-scale temporal interaction module

该模块采用编码器—解码器式结构：编码器用于聚合全局上下文信息，解码器用于融合不同层级的时序特征并逐步恢复局部细节，从而增强模型对整体动作结构和细粒度动态变化的联合表征能力。

经过局部时空模式编码后得到的帧级特征序列表示为：

$$\boldsymbol{X}=\{\boldsymbol{x}_t\}_{t=1}^{T},X\in\mathbb{R}^{T\times D}\qquad(8)$$

式中，T 表示视频序列长度，D 表示特征维度，$\boldsymbol{x}_t$ 表示第 t 帧对应的局部时空特征。对于分割任务，$\boldsymbol{X}$ 表示I3D提取的RGB时空特征；对于预测任务，$\boldsymbol{X}$ 表示经双向交叉注意力融合后得到的双模态特征。

在编码阶段，跨尺度时序交互模块由 L 个时序建模层堆叠而成。第 l 层编码过程可表示为：

$$\boldsymbol{X}^{l}=\mathcal{E}_l\left(\boldsymbol{X}^{l-1}\right),\boldsymbol{X}^{0}=\boldsymbol{X},l=1,2,\cdots,L\#(9)$$

式中，$\mathcal{E}_l(\cdot)$ 表示第 l 层编码单元，$\boldsymbol{X}^{l}$ 表示第 l 层编码器输出的时序特征表示。编码单元通过自注意力机制建模不同时间步之间的相关性，并结合前馈映射、残差连接和归一化操作完成特征更新。随着编码层数增加，模型能够逐步聚合更大时间范围内的上下文信息，从而增强对长程时序依赖关系的表达能力。

由于高层编码特征在聚合全局上下文的同时可能削弱局部时间细节，进一步引入解码器结构，对不同层级的时序特征进行逐层融合与细化。设第 l 层解码器输出为 $\boldsymbol{Z}^{l}$，其更新过程可表示为：

$$\boldsymbol{Z}^{l}=\mathcal{D}_l\left(\boldsymbol{Z}^{l+1},\boldsymbol{X}^{l}\right),l=L-1,L-2,\cdots 0\#(10)$$

式中，$\mathcal{D}_l(\cdot)$ 表示第 l 层解码单元，$\boldsymbol{Z}^{l+1}$ 表示上一层解码特征，$\boldsymbol{X}^{l}$ 表示对应层级的编码特征。解码单元通过融合高层上下文特征与对应层级的编码特征，在保留长程时序依赖信息的同时补充局部动态细节，从而提高模型对动作切换区域和细粒度动作变化的表征能力。

最终，解码器输出的多尺度时序特征表示为：

$$\hat{\boldsymbol{Z}}=\phi\left(\boldsymbol{Z}^{0},\boldsymbol{Z}^{1},\cdots,\boldsymbol{Z}^{L}\right)\#(11)$$

式中，$\phi(\cdot)$ 表示通道拼接与1×1卷积映射操作。该特征同时融合了长程上下文依赖信息与局部动态细节，有助于提升后续分类预测和边界优化过程的稳定性。

1.4 边界概率细化模块

在跨尺度时序交互模块输出多尺度时序特征的基础上，为进一步增强动作边界的准确性与序列稳定性，引入逐阶段边界概率细化模块对分割与预测结果进行显式结构约束与边界优化，如图4所示。

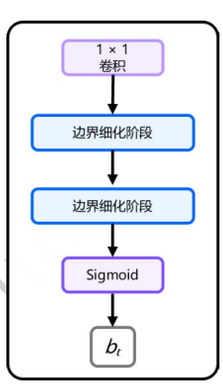

图4 边界概率细化模块

Fig. 4 Boundary probability refinement module

对于多尺度时序特征序列 $\hat{\boldsymbol{Z}}$，边界细化模块由 K 个边界细化阶段单元串联构成。其递推更新过程为：

$$\widehat{\boldsymbol{Z}}_t^{(k+1)}=\widehat{\boldsymbol{Z}}_t^{(k)}+f^{(k)}\left(\widehat{\boldsymbol{Z}}_t^{(k)}\right),k=0,1,\ldots,K-1\#(12)$$

式中，$\widehat{\boldsymbol{Z}}_t^{(k)}$ 表示第 k 个边界细化阶段中第 t 帧的时序特征，$\hat{\boldsymbol{Z}}^{(0)}=\hat{\boldsymbol{Z}}$，$f^{(k)}(\cdot)$ 表示第 k 个边界细化阶段单元的局部细化映射。

每个边界细化阶段单元由膨胀卷积与残差结构组成，用于在局部时间窗口内增强边界敏感特征：

$$f^{(k)}(\hat{\boldsymbol{Z}})=u\left(\rho\left(h\left(\hat{\boldsymbol{Z}}\right)\right)\right)\#(13)$$

式中，$h(\cdot)$ 表示膨胀卷积操作，$\rho(\cdot)$ 表示 ReLU 激活函数，$u(\cdot)$ 表示 1×1 卷积映射。膨胀卷积的引入能够在不显著增加参数量的情况下扩大局部时间感受野，从而使模块在边界附近获得更强的局部上下文建模能力；残差结构则有助于稳定逐阶段优化过程。

完成 K 次逐阶段细化后，模块输出细化后的时序特征序列：

$$\bar{\boldsymbol{Z}}=\hat{\boldsymbol{Z}}^{(K)}\#(14)$$

通过线性映射与 Sigmoid 函数得到每一帧为动作边界的概率：

$$b_t=\sigma\left(q\left(\overline{\boldsymbol{Z}}_t\right)\right),b_t\in[0,1]\#(15)$$

式中，$q(\cdot)$ 表示边界预测的线性映射函数，b_t 表示第 t 帧为动作边界的预测概率。该边界概率序列能够为后续分类结果细化提供显式的结构先验，使模型在动作切换位置附近获得更强的边界响应能力。

为了使预测边界概率与真实动作边界对齐，采用加权二元交叉熵对边界分支进行监督，其损失函数 L_{b} 定义为：

$$L_{\mathrm{b}}=-\sum_{t=2}^{T}\left(w_1 g_t log b_t+w_0\left(1-g_t\right)log\left(1-b_t\right)\right)\#(16)$$

式中，g_t 表示第 $t-1$ 帧与第 t 帧之间是否存在动作边界，w_1 与 w_0 为类别不平衡权重系数，以缓解边界样本与非边界样本数量不均衡带来的训练偏置。通过显式监督边界概率序列，模型能够更准确地感知动作起止位置，从而为后续边界优化提供可靠依据。

为抑制预测序列在非真实边界处的高频波动，引入基于边界门控的时间一致性约束 L_{g}：

$$L_{\mathrm{g}}=\sum_{t=2}^{T}\left(1-b_t\right)^{\gamma}\left\|\boldsymbol{P}_t-\boldsymbol{P}_{t-1}\right\|_2^2\#(17)$$

式中，$\boldsymbol{P}_t$ 为分类概率输出，$\gamma>0$ 为调节系数，$\|\cdot\|_2^2$ 表示二范数平方，用于度量相邻帧分类概率分布之间的差异。该约束的作用在于：当 b_t 较小时，表示当前位置更可能处于非边界区域，此时相邻帧分类概率分布差异将受到更强约束，以增强动作段内部分割结果的一致性；当 b_t 较大时，表示当前位置更可能接近真实边界，此时则允许分类结果发生变化，从而提高边界位置的响应能力与边界对齐精度。

1.5 动作分割与预测方法

动作分割方法：基于跨尺度时序交互与边界概率细化模块，得到逐帧动作类别预测 $\boldsymbol{P}$ 和边界概率输出 $\boldsymbol{B}$：

$$(\boldsymbol{P},\boldsymbol{B})=\Psi(\hat{\boldsymbol{Z}})\#(18)$$

$$\boldsymbol{P}=\{\boldsymbol{P}_t\}_{t=1}^{T},\boldsymbol{P}_t=[p_{t,1},p_{t,2},\cdots,p_{t,C}]\in\boldsymbol{\Delta}^{C-1}(19)\boldsymbol{B}=\{b_t\}_{t=1}^{T},b_t\in[0,1]\qquad(20)$$

式中，$\Psi(\cdot)$ 表示由分类头和边界预测头共同构成的输出映射函数，C 表示动作类别数，$\boldsymbol{P}_t$ 表示第 t 帧属于各类别的概率分布，$p_{t,c}$ 表示第 t 帧属于第 c 类动作的预测概率，$\boldsymbol{\Delta}^{C-1}$ 表示 C 类概率分布构成的概率单纯形，b_t 表示该帧为动作边界的置信度。根据最大后验准则，第 t 帧的动作类别判别结果可表示为：

$$\hat{y}_t=\arg\max_c p_{t,c},c\in\{1,\cdots,C\},t=1,\cdots,T\ (21)$$

对于视频动作分割任务，采用联合损失函数 L 对模型进行训练：

$$L=L_{\mathrm{c}}+\lambda_{\mathrm{b}}L_{\mathrm{b}}+\lambda_{\mathrm{g}}L_{\mathrm{g}}\#(22)$$

式中，L_{c} 表示逐帧分类损失，λ_{b} 与 λ_{g} 为相应损失项的权重系数。设 y_t 表示第 t 帧的真实动作类别标签，则 L_{c} 可写为交叉熵形式：

$$L_{\mathrm{c}}=-\sum_{t=1}^{T}\log p_{t,y_t}\#(23)$$

动作预测方法：对于视频动作预测任务，模型需要根据已观测视频片段推断未来时间段内的动作类别。与分割任务不同，预测任务不仅需要对已观测片段进行时序建模，还需要在观测信息基础上对未来动作序列进行推断。设视频序列由已观测区间和未来预测区间组成，则有：

$$T=T_{\mathrm{o}}+T_{\mathrm{f}}\#(24)$$

式中，T 表示视频序列总长度，T_{o} 表示模型可访问的已观测序列长度，T_{f} 表示需要预测的未来序列长度。已观测 RGB 序列 $\boldsymbol{V}^{\mathrm{r}}$ 和已观测光流序列 $\boldsymbol{V}^{\mathrm{f}}$ 分别表示为：

$$\boldsymbol{V}^{\mathrm{r}}=\left\{\boldsymbol{v}_1^{\mathrm{r}},\boldsymbol{v}_2^{\mathrm{r}},\cdots\boldsymbol{v}_{T_{\mathrm{o}}}^{\mathrm{r}}\right\},\boldsymbol{V}^{\mathrm{f}}=\left\{\boldsymbol{v}_1^{\mathrm{f}},\boldsymbol{v}_2^{\mathrm{f}},\cdots,\boldsymbol{v}_{T_{\mathrm{o}}}^{\mathrm{f}}\right\}\#(25)$$

式中，$\boldsymbol{v}_i^{\mathrm{r}}$ 表示第 i 个时间步对应的 RGB 帧，$\boldsymbol{v}_i^{\mathrm{f}}$ 表示第 i 个时间步对应的光流输入。

同理，基于跨尺度时序交互与边界概率细化模

块，得到逐帧预测概率序列和边界概率序列：

$$\left(\boldsymbol{P}^{\mathrm{p}},\boldsymbol{B}^{\mathrm{p}}\right)=\Psi\left(\hat{\boldsymbol{Z}}^{\mathrm{p}}\right)\#(26)$$

式中，$\hat{\boldsymbol{Z}}^{\mathrm{p}}$ 表示预测任务的时序特征，$\boldsymbol{P}^{\mathrm{p}}=\{\boldsymbol{p}_t^{\mathrm{p}}\}_{t=1}^{T}$ 表示逐帧动作预测概率序列，其中 $\boldsymbol{p}_t^{\mathrm{p}}$ 表示预测任务中第 t 帧的类别概率分布，$\boldsymbol{B}^{\mathrm{p}}=\{b_t^{\mathrm{p}}\}_{t=T_{\mathrm{o}}+1}^{T_{\mathrm{o}}+T_{\mathrm{f}}}$ 表示边界概率序列，用于刻画预测序列中的动作切换位置，并为时间一致性约束提供门控信号。

预测任务关注未来区间 $[T_{\mathrm{o}}+1,T_{\mathrm{o}}+T_{\mathrm{f}}]$ 的动作类别序列，因此最终输出定义为：

$$\hat{y}_t^{\mathrm{p}}=\arg\max_c\ p_{t,c}^{\mathrm{p}},t=T_{\mathrm{o}}+1,...,T_{\mathrm{o}}+T_{\mathrm{f}}\#(27)$$

式中，$p_{t,c}^{\mathrm{p}}$ 表示预测任务中第 t 帧属于第 c 类动作的预测概率。

对于视频动作预测任务，同样采用分类损失、边界预测损失和基于边界门控的时间一致性约束进行联合优化，其损失函数 L_{p} 定义为：

$$L_{\mathrm{p}}=L_{\mathrm{c}}^{\mathrm{p}}+\lambda_{\mathrm{b}}L_{\mathrm{b}}^{\mathrm{p}}+\lambda_{\mathrm{g}}L_{\mathrm{g}}^{\mathrm{p}}\#(28)$$

式中，$L_{\mathrm{c}}^{\mathrm{p}}$、$L_{\mathrm{b}}^{\mathrm{p}}$ 和 $L_{\mathrm{g}}^{\mathrm{p}}$ 分别表示预测任务的分类损失、边界预测损失和基于边界门控的时间一致性约束项。

分类监督与边界监督主要作用于未来预测区间。为统一表达，引入时间权重函数 w_t：

$$w_t=\begin{cases}1,\ t\in\{T_{\mathrm{o}}+1,...,T\}\\ \alpha,\ t\in\{1,...,T_{\mathrm{o}}\}\end{cases},0\leqslant\alpha\leqslant 1\#(29)$$

式中，α 表示观测区间的监督权重系数。则预测任务的逐帧分类损失可写为：

$$L_{\mathrm{c}}^{\mathrm{p}}=-\sum_{t=1}^{T}w_t\log p_{t,y_t}^{\mathrm{p}}\#(30)$$

2 实验结果

2.1 数据集

本文使用的数据集覆盖两类连续视频序列建模场景：一类为生活场景下的连续动作分割数据集，包括50Salads、GTEA和Breakfast；另一类为手术场景下的细粒度手术动作分割与预测数据集，即JIGSAWS。前者主要用于验证模型在通用连续动作分割任务中的时序建模与边界定位能力，后者用于验证模型在精细手术操作序列中的动作分割与预测能力。

50Salads（Stein 等，2013）数据集采集于厨房食物准备场景，以固定俯视视角记录参与者完成沙拉制作的全过程，共包含50段视频，并提供细粒度动作片段的起止位置标注。该数据集序列较长，动作持续时间差异较大，适合用于评估模型对长时依赖关系和动作边界的建模能力。

GTEA（Fathi 等，2011）数据集为第一视角活动数据集，使用佩戴式摄像机记录参与者完成日常操作过程。该数据集包含多类细粒度操作行为，动作切换较快、局部运动差异细微，适合用于评估模型对短时局部动态变化的建模能力及精细边界定位能力。

Breakfast（Kuehne 等，2014）数据集采集于真实家庭厨房环境，覆盖多名参与者、多个场景和多种早餐制作活动，具有视频时长长、类别数量多、场景差异大等特点。该数据集能够用于检验模型在复杂长序列和跨场景条件下的泛化能力。

JIGSAWS（Gao 等，2014）数据集来源于机器人辅助手术操作过程，包含缝合、穿针和打结等任务，并提供细粒度手术动作区间标注。由于该数据集中的动作具有较强的规范性和重复性，因此适合用于分析模型在精细操作序列中的时序一致性和边界刻画能力。

2.2 评价指标

为全面评估方法在连续视频动作分割与预测任务中的性能，采用帧级准确率（Acc）、编辑距离得分（Edit）以及基于交并比阈值的F1分数作为评价指标。该指标体系能够分别从逐帧分类准确性、序列整体一致性以及动作边界定位能力三个方面对模型进行评价。

2.3 实验细节

为保证实验结果的可复现性和公平性，本文在各数据集上均采用预定义的数据划分文件进行训练和测试。具体而言，50Salads数据集采用5个split进行交叉验证；GTEA数据集采用4个split；Breakfast数据集采用4个split；对于JIGSAWS数据集，本文使用Suturing子任务，并采用1个固定split进行实验。

为保持与已有动作分割基准实验协议的一致性，本文未额外从训练数据中划分独立验证集，而是按照预定义的train/test split进行训练与测试。模型仅使用训练集进行参数学习，并在对应测试集上计算帧级准确率Acc、编辑距离Edit以及F1@10、F1@25和F1@50等指标。对于50Salads、GTEA和Breakfast等包含多个split的数据集，本文分别在每个split

上训练和测试模型，并将各split的测试结果进行平均，作为该数据集上的最终性能。JIGSAWS数据集在当前实验设置中仅包含1个固定split，因此直接报告该split上的测试结果。本文方法及消融实验均采用相同的数据划分设置，以保证结果比较的一致性。

实验在单机单卡环境下完成，训练框架采用PyTorch，优化器选用Adam。实验数据主要采用预处理后的RGB图像序列与光流序列。其中，动作分割任务以RGB序列作为主要输入，在50Salads、GTEA、Breakfast和JIGSAWS四个数据集上进行验证；动作预测任务采用RGB+光流双模态特征，主要在JIGSAWS数据集上开展实验。

模型训练批大小设置为1，训练轮数设置为120，默认初始学习率为0.0005。考虑到Breakfast数据集样本规模较大且序列较长，其学习率调整为0.0001，其余数据集保持默认设置。边界概率细化模块中，细化阶段数设置为 $K=2$，每个阶段采用一维膨胀卷积实现，其膨胀率分别为1和2。在损失函数设置方面，本文将主分类交叉熵损失的权重设为1，$\lambda_b=0.15$，$\lambda_g=0.001$，$\gamma=128$。为保证实验结果的可比性，不同数据集上的主要训练配置保持一致。

2.4 对比实验

2.4.1 分割任务对比实验

为验证方法在视频动作分割中的有效性，本文在50Salads、GTEA、Breakfast和JIGSAWS 4个公开数据集上与代表性方法进行了对比实验。对比方法包括MS-TCN、MS-TCN++、ASFormer、UVAST、FACT、BaFormer、EAST以及Bi-LSTM、ED-TCN、RL、SD-Net、Reformer等。评价指标采用2.2节定义的Acc、Edit以及F1@10、F1@25和F1@50。实验结果分别如表1—表4所示。

表1给出了本文方法在50Salads数据集上的实验结果。可以看出，本文方法在5项评价指标上均取得最优结果，其中Edit、Acc和高阈值F1指标提升更为明显。这说明在长序列、动作持续时间差异较大的场景中，所提方法能够同时增强长程上下文建模能力和动作边界定位能力。与现有性能较优的方法相比，本文方法在保持较高逐帧分类准确率的同时，进一步改善了动作切换区域的预测稳定性，表明跨尺度时序交互与边界概率细化在长序列场景下具有较好的协同效果。

表1 50Salads数据集实验结果

Table 1 Experimental results on the 50Salads dataset

/%

50Salads	F1@10	F1@25	F1@50	Edit	Acc
MS-TCN(Abu Farha等,2019)	76.3	74.0	64.5	67.9	80.7
MS-TCN++(Li等,2023)	80.7	78.5	70.1	74.3	83.7
ASFormer(Yi等,2021)	85.1	83.4	76.0	79.6	85.6
UVAST(Behrmann等,2022)	89.1	87.6	81.7	83.9	87.4
FACT(Lu等,2024)	81.4	76.5	66.2	79.7	76.2
BaFormer(Wang等,2024)	89.3	88.4	83.9	84.2	89.5
EAST(Wang等,2025)	93.1	91.9	88.6	88.4	91.9
本文	**93.6**	**92.2**	**88.8**	**89.0**	**92.3**

注：加粗字体表示各列最优结果。

表2 GTEA数据集实验结果

Table 2 Experimental results on the GTEA dataset

/%

GTEA	F1@10	F1@25	F1@50	Edit	Acc
MS-TCN	87.5	85.4	74.6	81.4	79.2
MS-TCN++	88.8	85.7	76.0	83.5	80.1
ASFormer	90.1	88.8	79.2	84.5	79.7
UVAST	92.7	91.3	81.0	92.1	80.2
FACT	93.5	92.1	84.1	91.4	86.1
BaFormer	92.0	91.3	83.5	88.7	83.0
EAST	95.8	95.4	**91.7**	95.4	87.1
本文	**96.5**	**95.7**	90.3	**95.8**	**88.0**

注：加粗字体表示各列最优结果。

表2展示了各方法在GTEA数据集上的实验结果。总体来看，本文方法在F1@10、F1@25、Edit和Acc等4项指标上取得最优结果，说明所提方法能够较好地提升第一视角细粒度动作序列的识别性能。需要指出的是，在F1@50指标上，本文方法略低于最佳对比方法，表明在部分动作切换较快、局部运动差异较小的片段上，高阈值条件下的精确边界切分仍存在一定提升空间。总体而言，该结果说明本文方法在细粒度第一视角场景下兼顾了较强的逐帧判别能力和较好的序列整体一致性。

表3展示了本文方法在Breakfast数据集上的实验结果。与其他数据集相比，Breakfast具有视频序列更长、场景变化更复杂、类别数量更多等特点，因

表3 Breakfast数据集实验结果

Table 3 Experimental results on the Breakfast dataset

/%

Breakfast	F1@10	F1@25	F1@50	Edit	Acc
MS-TCN	52.6	48.1	37.9	61.7	66.3
MS-TCN++	64.1	58.6	45.9	65.6	67.6
ASFormer	76.0	70.6	57.4	75.0	73.5
UVAST	75.9	70.0	57.2	76.5	66.0
FACT	81.4	76.5	66.2	79.7	76.2
BaFormer	79.2	74.9	63.2	77.3	76.6
EAST	86.2	82.2	**71.8**	84.5	82.8
本文	**86.4**	**82.5**	71.6	**84.7**	**83.2**

注:加粗字体表示各列最优结果。

表4 JIGSAWS数据集实验结果

Table 4 Experimental results on the JIGSAWS dataset

/%

JIGSAWS	F1@10	F1@25	F1@50	Edit	Acc
Bi-LSTM(Singh 等,2016)	77.8	77.4	66.8	66.8	77.4
ED-TCN(Lea 等,2017)	89.2	84.7	80.8	84.7	80.8
RL(Liu 等,2018)	92.0	90.5	82.2	88.0	81.4
SD-Net(Zhang 等,2021)	92.5	92.0	88.2	89.9	90.1
Reformer(Men 等,2026)	96.4	**96.4**	90.2	94.0	85.7
本文	**97.1**	95.7	**90.6**	**95.2**	**90.5**

注:加粗字体表示各列最优结果。

此分割难度更高。从结果可以看出,本文方法在F1@10、F1@25、Edit和Acc指标上均取得最优结果,分别达到86.4、82.5、84.7和83.2;在F1@50指标上,本文方法为71.6,与EAST的71.8接近。上述结果表明,本文方法在复杂长序列和跨场景条件下具有较好的鲁棒性和泛化能力,尤其在保持整体序列一致性和逐帧分类准确性方面表现较优。

表4给出了JIGSAWS数据集上的实验结果。可以看出,本文方法在F1@10、F1@50、Edit和Acc等4项指标上取得最优结果,分别达到97.1、90.6、95.2和90.5;在F1@25指标上取得95.7,略低于Reformer的96.4。该结果说明,本文方法在手术动作分割任务中能够较好地兼顾局部细粒度动作刻画与整体序列建模能力,尤其在完整序列结构恢复和逐帧识别精度方面具有明显优势。

2.4.2 预测任务对比实验

为验证本文方法在视频动作预测任务中的有效性,本文在JIGSAWS数据集上与近年来具有代表性的预测方法进行比较。实验结果如表5所示。

表5 预测实验结果

Table 5 Experimental results on action prediction

JIGSAWS	预测设置	Acc/%
daVinciNet(Qin 等,2020)	动作预测	84.3
Transformer (MTMs)(Shi 等,2022)	动作预测	84.6
Multimodal Transformer(Weerasinghe 等,2024)	动作预测	89.5
Transformer-based real-time prediction(Chen 等,2025)	动作预测	84.8
本文	动作预测	**90.3**

注:加粗字体表示各列最优结果。

本文方法在JIGSAWS数据集上的动作预测准确率达到90.3%,在对比方法中取得最优结果。相比daVinciNet、Transformer(MTMs)、Multimodal Transformer以及Transformer-based real-time prediction等方法,本文方法均表现出更高的预测精度,说明所提出模型能够更有效地建模动作序列中的时序变化规律,并提升对后续动作状态的判断能力。

与Multimodal Transformer的89.5%相比,本文方法进一步提升至90.3%,表明在动作预测任务中,引入RGB与光流信息能够增强模型对外观特征和运动变化的联合表征能力。其中,RGB序列提供了手术场景和器械状态的空间信息,光流序列则能够刻画手术操作过程中的动态变化。二者结合后,有助于模型更准确地捕捉连续手术动作之间的转移关系,从而提升未来动作预测性能。总体来看,本文方法在JIGSAWS数据集上取得了较好的预测效果,验

证了所提出时序建模与双模态信息融合策略的有效性。

2.5 消融实验

为进一步验证本文所提出各关键模块的有效性，本文在保持训练策略、损失函数及其余网络设置一致的条件下，分别对跨尺度时序交互模块和边界概率细化模块进行消融实验。针对预测任务，比较了不同模态输入方式的预测结果。

2.5.1 跨尺度时序交互模块有效性分析

为验证本文所提出跨尺度时序交互模块的有效性，本文设计了2组对比设置：未采用跨尺度时序交互模块的I3D + 边界概率细化，以及完整模型。为保证消融实验对比的公平性，两种设置均采用相同的I3D骨干网络、数据划分及训练策略。实验结果如表6所示。

表6 跨尺度时序交互模块有效性验证结果

Table 6 Ablation study of the cross-scale temporal interaction module

/%

数据集	模型设置	F1@10	F1@25	F1@50	Edit	Acc
50Salads	I3D + 边界概率细化	91.9	90.1	84.8	85.7	91.1
	完整模型	**93.6**	**92.2**	**88.8**	**89.0**	**92.3**
GTEA	I3D + 边界概率细化	95.4	94.3	87.9	94.7	87.4
	完整模型	**96.5**	**95.7**	**90.3**	**95.8**	**88.0**
Breakfast	I3D + 边界概率细化	83.1	78.0	65.3	81.8	80.9
	完整模型	**86.4**	**82.5**	**71.6**	**84.7**	**83.2**
JIGSAWS	I3D + 边界概率细化	95.8	94.7	88.2	94.8	89.7
	完整模型	**97.1**	**95.7**	**90.6**	**95.2**	**90.5**

注：加粗字体表示各列最优结果。

由表6可知，完整模型在4个数据集上均取得了最优性能，说明本文提出的跨尺度时序交互模块能够在不同类型的连续视频序列分割任务中稳定提升模型性能。相比I3D + 边界概率细化，完整模型不仅保留了局部短时动态建模能力，还进一步引入了跨尺度时序信息交互，从而增强了模型对长程上下文关系和动作阶段结构的建模能力。

具体来看，在50Salads数据集上，完整模型的F1@10、F1@25、F1@50、Edit和Acc分别达到93.6、92.2、88.8、89.0和92.3，相较I3D + 边界概率细化分别提升1.7、2.1、4.0、3.3和1.2。其中，F1@50和Edit指标提升更为明显，说明跨尺度时序交互能够有效改善动作片段边界和整体序列结构的一致性。在Breakfast数据集上，完整模型相较I3D + 边界概率细化在F1@50上提升6.3，在F1@25上提升4.5，表明在长序列、复杂场景和阶段结构明显的任务中，跨尺度时序交互对性能提升具有更加突出的作用。

此外，本文在4个数据集上分别给出了分割结果可视化，如图5—图8所示，不同颜色表示不同动作类别。可视化结果进一步表明，在去除跨尺度时序交互模块后，模型在长动作片段内部或相邻动作切换区域仍会出现局部类别抖动、误分割或片段合并等现象；相比之下，完整模型在保持整体时序平滑性的同时，能够更准确地恢复动作边界位置，使预测结果更接近真实标签。

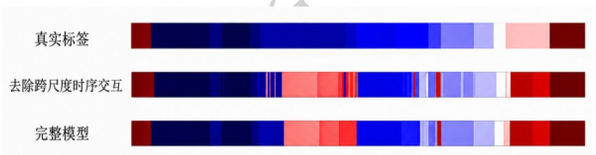

图5 50Salads数据集消融结果可视化

Fig. 5 Visualization of ablation results on the 50Salads dataset

图6 GTEA数据集消融结果可视化

Fig. 6 Visualization of ablation results on the GTEA dataset

图7 Breakfast数据集消融结果可视化

Fig. 7 Visualization of ablation results on the Breakfast dataset

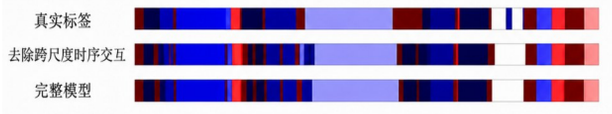

图8 JIGSAWS消融结果可视化

Fig. 8 Visualization of ablation results on the JIGSAWS dataset

2.5.2　边界概率细化模块有效性分析

为验证本文所提出边界概率细化模块的有效性，在保持跨尺度时序交互结构及其余训练设置不变的前提下，本文进一步比较了去除边界概率细化模块和完整模型两种设置，实验结果如表7所示。

表7　边界概率细化模块有效性验证结果

Table 7　Ablation study of the boundary probability refinement module

/%

数据集	模型设置	F1@10	F1@25	F1@50	Edit	Acc
50Salads	I3D + 跨尺度时序交互	93.1	91.4	87.1	87.4	91.8
	完整模型	**93.6**	**92.2**	**88.8**	**89.0**	**92.3**
GTEA	I3D + 跨尺度时序交互	96.1	95.0	88.9	94.9	87.7
	完整模型	**96.5**	**95.7**	**90.3**	**95.8**	**88.0**
Breakfast	I3D + 跨尺度时序交互	85.9	81.4	68.9	82.9	82.6
	完整模型	**86.4**	**82.5**	**71.6**	**84.7**	**83.2**
JIGSAWS	I3D + 跨尺度时序交互	96.2	94.7	89.2	95.0	90.1
	完整模型	**97.1**	**95.7**	**90.6**	**95.2**	**90.5**

注：加粗字体表示各列最优结果。

从表7可以看出，在4个数据集上引入边界概率细化模块后，模型性能均得到不同程度提升。总体来看，该模块对Edit和F1@50两项指标的提升更为明显，而对F1@10和Acc的影响相对较小。这表明边界概率细化模块的主要作用并不在于改变动作类别的整体判别结果，而在于改善动作切换位置的定位精度，减少边界附近的过分割现象。

具体而言，在50Salads数据集上，完整模型相较于去除边界概率细化模块的模型，F1@50由87.1提升至88.8，Edit由87.4提升至89.0；在GTEA数据集上，F1@50和Edit分别由88.9和94.9提升至90.3和95.8；在Breakfast数据集上，F1@50和Edit分别由68.9和82.9提升至71.6和84.7，提升幅度最为明显；在JIGSAWS数据集上，Edit和Acc也分别由95.0和90.1提升至95.2和90.5。上述结果说明，边界概率细化模块能够在不同类型的数据集上稳定改善动作分段的完整性与边界一致性，尤其在长序列、动作切换频繁或阶段结构较复杂的场景中，其作用更加突出。

为进一步展示边界概率细化模块的实际效果，本文以50Salads数据集为例给出可视化结果，如图9所示。

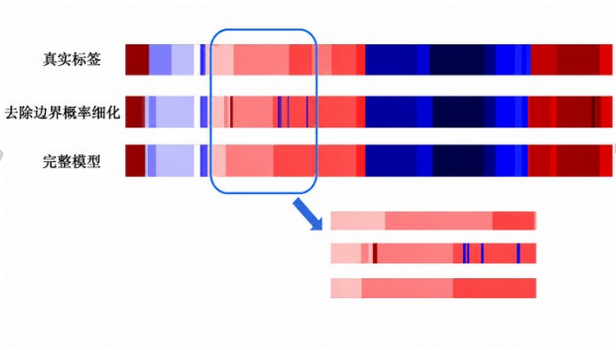

图9　边界概率细化模块消融结果可视化

Fig. 9　Visualization of ablation results for the boundary probability refinement module

由图9可知，在去除边界概率细化模块后，模型在动作切换区域容易出现预测类别的局部跳变，导致真实边界附近产生误分割；引入边界概率细化模块后，边界附近的预测结果更加平滑，也更加接近真实标签。结合定量结果与可视化分析可知，边界概率细化模块能够通过显式建模动作起止边界并对分类结果施加结构化约束，有效提升连续视频动作分割中的边界定位精度与序列稳定性。

2.5.3　多模态输入对动作预测性能的影响

为验证不同输入模态对动作预测性能的影响，本文在JIGSAWS数据集上进一步比较了仅使用RGB、仅使用光流以及RGB与光流融合三种输入方式的预测结果，如表8所示。

由表8可知，仅使用RGB时，模型取得了85.8%的预测准确率，仅使用光流时，模型准确率提升至88.9%，表明相较于静态外观信息，运动变化趋势对

表8　多模态输入下的动作预测结果

Table 8　Action prediction results under different multi-modal inputs

输入模态	Acc/%
仅RGB	85.8
仅光流	88.9
RGB+光流	**90.3**

注：加粗字体表示各列最优结果。

未来短时动作预测效果更好。融合RGB与光流两种模态后，模型准确率达到90.3%，取得最优结果。这表明RGB提供的外观信息与光流表征的变化信息具有互补性，二者结合后能够更全面地刻画未来动作演化过程，从而提高预测性能。

图10给出了不同输入模态下的局部预测结果可视化。

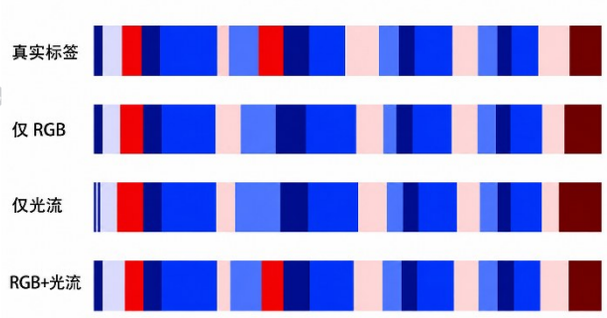

图10　多模态输入下的动作预测可视化对比

Fig. 10　Visualization of action prediction results under different multimodal inputs

由图10可知，仅使用RGB时，预测结果在动作切换位置上存在一定滞后，更倾向于延续当前动作状态；仅使用光流时，模型对动作变化趋势更加敏感，能够更早感知未来动作转移，但存在不稳定现象；相比之下，融合RGB与光流后，预测序列与真实标签更加接近。

3　结　论

本文面向连续动作分割与预测中长短时依赖协同建模不足、动作边界模糊以及未来状态推断不稳定等问题，提出了融合跨尺度时序交互与边界概率细化的方法。该方法通过局部时空编码提取短时动态信息，并利用跨尺度时序交互结构联合建模局部细节与长程上下文；同时，通过边界概率分支和边界门控时间一致性约束增强动作切换区域的结构约束，缓解过分割和类别抖动问题。在预测任务中，进一步引入RGB与光流双模态交互，使模型能够同时利用外观语义和运动趋势信息。多数据集实验和消融结果表明，所提方法不仅适用于生活场景下的连续动作分割任务，也能够迁移到手术场景下的细粒度手术动作分割与预测任务，体现了较好的跨场景泛化能力和序列建模稳定性。

尽管如此，本文方法在快速动作切换、视觉相似动作区分以及光流质量受限场景下仍存在提升空间。后续研究将进一步探索端到端轻量化特征提取、更加稳健的边界感知机制以及低延迟多模态融合策略，以提升模型在实际人机交互和手术辅助场景中的实时应用能力。

参考文献(References)

Abu Farha Y and Gall J. 2019. MS-TCN: multi-stage temporal convolutional network for action segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3575-3584 [DOI: 10.1109/CVPR.2019.00369]

Behrmann N, Golestaneh S A, Kolter Z, Gall J and Noroozi M. 2022. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation//Computer Vision-ECCV 2022. Tel Aviv, Israel: Springer: 52-68 [DOI: 10.1007/978-3-031-19833-5_4]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen K, Bandara D S V and Arata J. 2025. A real-time approach for surgical activity recognition and prediction based on transformer models in robot-assisted surgery. International Journal of Computer Assisted Radiology and Surgery, 20(4): 743-752 [DOI: 10.1007/s11548-024-03306-9]

Cheng Y, Gao Y Y, Wang J, Yang L, Xu X L, Cheng Y, et al. 2025. Combining dilated convolution and multiscale fusion temporal action detection. Journal of Image and Graphics, 30(2): 406-420 (程勇, 高园元, 王军, 杨玲, 许小龙, 程遥, 等. 2025. 结合扩张卷积与多尺度融合的实时时空动作检测. 中国图象图形学报, 30(2): 406-420) [DOI: 10.11834/jig.240098]

Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, et al. 2015. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 2625-2634 [DOI: 10.1109/CVPR.2015.7298878]

Fathi A, Farhadi A and Rehg J M. 2011. Understanding egocentric activities//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE: 407-414 [DOI: 10.1109/ICCV.2011.6126269]

Feichtenhofer C. 2020. X3D: expanding architectures for efficient video recognition//Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition. Seattle, USA: IEEE: 203-213 [DOI:10.1109/CVPR42600.2020.00028]

Gao Y, Vedula S S, Reiley C E, Ahmidi N, Varadarajan B, Lin H C, et al. 2014. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): a surgical activity dataset for human motion modeling//Proceedings of the MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions. Boston, USA: M2CAI: 1-10

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco. 1997.9.8. 1735]

Kuehne H, Arslan A and Serre T. 2014. The language of actions: recovering the syntax and semantics of goal-directed human activities// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 780-787 [DOI: 10. 1109/CVPR.2014.105]

Lea C, Flynn M D, Vidal R, Reiter A and Hager G D. 2017. Temporal convolutional networks for action segmentation and detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 156-165 [DOI: 10.1109/ CVPR.2017.113]

Li S, Abu Farha Y, Liu Y, Cheng M M and Gall J. 2023. MS-TCN++: multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6): 6647-6658 [DOI:10.1109/TPAMI.2020.3021756]

Liu C W, Hou J Y, Wu X X and Jia Y D. 2018. A discriminative structural model for joint segmentation and recognition of human actions. Multimedia Tools and Applications, 77(24): 31627-31645 [DOI: 10.1007/s11042-018-6189-9]

Lu Z and Elhamifar E. 2024. FACT: frame-action cross-attention temporal modeling for efficient action segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18175-18185

Men Y T, Zhang C, Wei L P, Luo J, Zhang G, Wu H, et al. 2026. Research on temporal segmentation of surgical gestures for damage control surgery based on the Reformer method. Biomedical Signal Processing and Control, 114: 109246 [DOI:10.1016/j.bspc.2025. 109246]

Molchanov P, Yang X D, Gupta S, Kim K, Tyree S and Kautz J. 2016. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4207-4215 [DOI:10.1109/CVPR.2016.456]

Núñez J C, Cabido R, Pantrigo J J, Montemayor A S and Vélez J F. 2018. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition, 76: 80-94 [DOI: 10.1016/j. patcog. 2017. 10.033]

Qin Y, Feyzabadi S, Allan M, Burdick J W and Azizian M. 2020. daVinciNet: joint prediction of motion and surgical state in robot-assisted surgery//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, USA: IEEE: 2921-2928

Ren X L, Zhang F F, Zhou W T and Zhou L. 2025. Complementary multi-type prompts for weakly-supervised temporal action location. Journal of Image and Graphics, 30(3): 842-854 (任小龙, 张飞飞, 周琬婷, 周玲. 2025. 多类型提示互补的弱监督时序动作定位. 中国图象图形学报, 30(3): 842-854) [DOI:10.11834/jig. 240354]

Shi C, Zheng Y and Fey A M. 2022. Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan: IEEE: 8017-8023

Singh B, Marks T K, Jones M, Tuzel O and Shao M. 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1961-1970 [DOI:10.1109/CVPR.2016.216]

Stein S and McKenna S J. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities//Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Zurich, Switzerland: ACM: 729-738 [DOI:10.1145/2493432.2493482]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV. 2015.510]

Wang P Y, Lin Y W, Blasch E, Wei J and Ling H B. 2024. Efficient temporal action segmentation via boundary-aware query voting// Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation: 37765-37790 [DOI:10.52202/079017-1192]

Wang T Q and Todorovic S. 2025. End-to-end action segmentation transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Honolulu, USA: IEEE: 2997-3006 [DOI:10.1109/ICCVW69036.2025.00313]

Weerasinghe K, Roodabeh S H R, Hutchinson K and Alemzadeh H. 2024. Multimodal transformers for real-time surgical activity prediction//Proceedings of the IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE: 13323-13330 [DOI: 10.1109/ICRA57147.2024.10611048]

Yi F, Wen H and Jiang T. 2021. ASFormer: transformer for action segmentation//Proceedings of the British Machine Vision Conference (BMVC). Virtual, UK: BMVA Press: 1-15

Zhang J L, Nie Y Y, Lyu Y, Yang X S, Chang J and Zhang J J. 2021. SD-Net: joint surgical gesture recognition and skill assessment.

International Journal of Computer Assisted Radiology and Surgery，16(10)：1675-1682［DOI:10.1007/s11548-021-02495-x］

作者简介

胡中华，第一作者，男，硕士研究生，主要研究方向为面向视频的连续动作分割与预测。E-mail：202533145@stumail.nwu.edu.cn

宋江玲，通讯作者，女，副教授，主要研究方向为深度学习、神经计算建模、临床辅助诊断方法等。E-mail：sjl@nwu.edu.cn

李晓宁，女，硕士研究生，主要研究方向为图像处理与智能应用。E-mail：3151484682@qq.com

张瑞，女，教授，主要研究方向为机器学习理论及算法、神经计算建模、生理信号模式识别等。E-mail：rzhang@nwu.edu.cn