学校代码：10246
学　　号：19110860016

博 士 学 位 论 文

（学术学位）

细粒度医疗行为识别与技能评估技术研究

Research of Fine-grained Medical Action Recognition and Skill Assessment Technologies

院　　系：　工程与应用技术研究院
专　　业：　计算机应用技术
姓　　名：　王顺利
指 导 教 师：　张立华 教授
完 成 日 期：　2024 年 5 月 20 日

指导小组成员名单

张立华　　教授

康晓洋　　副研究员

董志岩　　青年副研究员

目　录

插图目录

表格目录

摘　要

目前我国面临着医疗资源短缺、医疗服务系统承压过重、地域医疗资源分布不均匀等问题。鉴于我国庞大的人口基数与老龄化趋势，未来我国的医务人员数量缺口将会持续扩大。为应对以上问题，一种有效的举措是在保证医疗服务质量的同时尽快填补医务人员巨大的数量缺口。然而，医务人员的培训具有周期长、人力成本高昂的特点。因此，**如何在保证医疗培训质量的前提下提高效率、降低成本**成为有效缓解医疗系统压力的关键。在医疗教学与培训的诸多环节中，临床技能培训有着举足轻重的地位，高质量的技能培训能够有效提升医护人员临床水平、降低医疗事故发生率。

在传统的医疗技能培训与考核模式中，由经验丰富的医师对学员的操作进行全程观察与指导。这种模式固然能够保证高质量的教学效果，但由于始终需要医师参与，存在着人力成本高、医师工作负担重等缺点。高昂的人力成本极大地限制了医疗技能培训的效率与规模。基于人工智能技术的医疗技能评估系统的构建为解决以上问题提供了重要思路：首先通过视觉等传感器对医学生的操作过程进行记录，之后使用人工智能算法实现细粒度的医疗行为识别和精准的自动化技能评估，最终反馈评估结果。智能化的技能评估系统能够有效提升医疗技能的培训与考核效率，从而显著减轻一线医生工作负担、大幅节约教考环节的人力成本。

本文紧密围绕智能医疗技能评估系统构建这一研究主题，在任务范式、数据集构建、算法创新和技能评估系统集成应用四个方面开展了系统性、创新性研究。本文的研究目标为：**构建一套能够同时具备医疗操作质量评估、细粒度医疗行为错误识别、多模态交互评估和时序医疗行为分析功能的细粒度医疗行为识别与技能评估系统**。通过对四个子系统的深入探究，本研究有效改善了现有医疗技能评估研究所面临的细粒度行为识别数据集匮乏、行为划分粒度粗、算法性能受限和人机交互能力弱等问题。

具体而言，本文的主要研究内容如下：

（1）针对现有医疗技能评估模型直接采用传统的视频主干网络，并未进行医疗场景适配的问题，本文将单目标跟踪器引入到行为质量评估任务中，提出了一种基于时空管道机制的稀疏高效特征交互方法，并将其命名为管道自注意力 TSA 模块。单目标跟踪器生成的目标框序列能够为特征增强模块提供位置先验信息，从而实现高效的视频特征增强。本文基于 TSA 模块构建了行为质量评估框架 TSA-Net，并在手术机器人操作技能评估基准和体育行为质量评估基准中开展了性能与计算量对比实验。实验结果证实，相较于传统特征增强方法，TSA-Net 框架能以更少的计算开销取得更精准的行为质量评估结果。

（2）针对医疗技能评估研究所面临的行为划分粒度粗、缺乏对错误行为的探究等问题，本文提出了复合错误行为识别的任务范式，并对心肺复苏术 CPR 中的胸外按压行为进行了深入探究，在专业医师的指导下构建了包含 13 种错误行为和 74 种复合错误行为的标签空间。之后，本文搭建了多视角行为采集平台，构建了包含 RGB、光流和骨骼关键点三种模态信息的数据集 CPR-Coach，该数据集能够同时支持单类错误识别和复合错误识别任务。在算法方面，本研究针对真实技能评估场景所面临的“单类训练，多类测试”监督信息受限设定，提出了基于特征组合训练机制的 ImagineNet 框架。该框架能够有效缓解由“训练集—测试集”间数据分布差异过大所引起的识别性能低下问题。实验结果证实，ImagineNet 框架能够显著提升传统模型在复合错误识别任务中的性能。

（3）针对现有医疗技能评估系统的人机交互能力弱、无法满足实际应用需求等问题，本文在 ImagineNet 框架的基础上引入了多模态预训练方法和提示词工程，并提出了多模态对比预训练框架 CPR-CLIP。该框架首先从错误数量、错误种类、改正建议三个方面构建了错误行为语言描述信息，其次通过最小化对比预训练损失使模型具备了更高的识别精度。在推理阶段中，CPR-CLIP 框架支持通过自然语言对操作视频库进行智能化检索与批量评估。基于以上探究，本研究进一步招募医生开展了随机对照试验，充分的准确率和耗时结果证实了 CPR-CLIP 框架对技能评估的辅助有效性。

（4）针对现有时序医疗行为分析研究中的流程划分标准不统一、时序错误操作研究匮乏等问题，本文首先依据临床技能教材构建了时序医疗行为知识图谱。并以胸腔穿刺术为研究对象，构建了具有高细粒度行为标签的时序医疗行为分析数据集 ThoSet。此数据集能够同时支持时序行为分割、遗漏与错误行为识别等任务。受人类大脑逻辑分区结构的启发，本文提出了基于时序聚类注意力机制的特征增强模块 *k*M-Att，并将其应用在扩散时序行为分割模型中，并进一步依据时序行为分割结果提出了行为合规性评估算法。多个公开数据集和 ThoSet 数据集上的实验结果充分证实了算法的有效性。

综上所述，本研究通过任务范式创新、数据集构建、算法设计和集成应用四个层面的探究，构建了一套细粒度医疗行为识别与技能评估系统。该系统填补了医疗技能评估系统研究领域中的部分空白，为技能评估算法的设计提供了新的思路，为后续技能评估系统的落地应用奠定了基础。本研究有望为医疗技能培训效率的提升、医务人员数量缺口的快速补充和医疗服务系统压力的缓解做出贡献。

关键词：医疗技能评估系统；细粒度行为识别；行为质量评估；视频理解
中图分类号：TP391.4

Abstract

Currently, China is facing problems such as a shortage of medical resources, overpressure on the medical service system, and uneven distribution of regional medical resources. Considering the vast population base of China and the ageing trend, the gap in the number of medical personnel in China will continue to expand. In order to deal with the above problems, there is an effective way to fill the vast number gap as soon as possible while ensuring the quality of medical services. However, the training of medical personnel has the characteristics of a long cycle and high cost. Therefore, improving efficiency and reducing the cost of medical training have become critical issues in alleviating the pressure on the medical system. In many medical teaching and training phases, medical skill training occupies a pivotal position. High-quality skill training can effectively improve the clinical level of medical staff and reduce the incidence of medical accidents.

Under the traditional medical skill training and assessment mode, doctors must observe and guide students' operations. Although this method can ensure high-quality teaching results, it always requires the participation of doctors, which has problems such as high human costs and heavy workloads. The high human cost under the traditional teaching and assessment mode limits the efficiency and scale of medical skill training. The intelligent medical skill assessment system based on artificial intelligence provides a vital idea to solve the above problems: firstly, record the operation process of medical students through vision and other sensors, then use artificial intelligence algorithms to achieve fine-grained action recognition and accurate automatic skill assessment, and finally feedback the skill assessment results. The intelligence medical skill assessment system can effectively improve the efficiency of training and testing, thereby significantly reducing the workload of doctors and human costs in medical skill training.

This thesis takes the construction of an intelligent medical skill assessment system as the research topic. This thesis has carried out a series of systematic and innovative research in four aspects: medical skill assessment task paradigm, dataset construction, algorithm innovation, and the implementation of the skill assessment system. The research goal of this thesis is to build a fine-grained medical action recognition and skill assessment system that can simultaneously support the evaluation of medical operation quality, fine-grained medical action error recognition, multimodal human-computer

interaction and assessment, and temporal medical action analysis functions. Through exploring the four subsystems, this thesis effectively solved the problems faced by the medical skill assessment research, such as lack of fine-grained action analysis research, coarse granularity of action division, poor algorithm adaptation performance, and weak human-computer interaction ability.

Specifically, the main research contents of this thesis are as follows:

(1) The existing medical skill assessment model usually directly adopts the existing video backbones without adaptation, which will cause performance constraints. In this thesis, we introduce the visual object tracker into the skill assessment task and propose an efficient sparse feature aggregation method based on the spatio-temporal tube mechanism named the TSA module. The target bounding box sequences generated through the tracker provide the location prior information for the TSA module to achieve efficient and effective video feature aggregation. Based on the TSA module, this thesis constructs an action quality assessment framework named TSA-Net and carries out performance and computational complexity comparison experiments on medical skill assessment on da Vinci surgical robots and sports skill assessment datasets. Experimental results demonstrate that the proposed TSA-Net framework can achieve better assessment performance with less computational complexity compared with the classical feature aggregation methods.

(2) The existing research on medical skill assessment is faced with problems such as coarse granularity of action division and lack of exploration of error actions. To solve these problems, this thesis first proposes the task of composite error action recognition and makes an in-depth exploration of the external chest compression action in Cardio Pulmonary Resuscitation (CPR). Under the guidance of professional doctors, we construct an action label space containing 13 classes of wrong actions and 74 classes of composite error actions. After that, we built a multi-view action video acquisition platform and a dataset named CPR-Coach. This dataset contains multimodal information: RGB frames, optical flow frames, and 2D key points. This dataset is able to support both single-class error action recognition and composite error recognition tasks. In terms of algorithms, this thesis proposes an ImagineNet framework based on a feature combination training strategy, which aims to solve the setting of *Single-class Training and Multi-class Testing* constraint supervision faced by real skill assessment scenarios. The ImagineNet framework can effectively alleviate the problem of poor recognition performance caused by the large data distribution gap between the training

set and the testing set. The experimental results show that the ImagineNet framework can significantly improve the performance of traditional action recognition models in composite error recognition tasks.

(3) The existing medical skill assessment system has poor human-computer interaction ability and cannot meet the actual assessment application. To tackle this problem, this thesis introduces the multi-modal pre-training framework and prompt engineering to the ImagineNet framework and proposes the multi-modal contrastive pre-training framework named CPR-CLIP. The framework first constructs the error action description statements through the information on error number, error type, and correction suggestions. After that, this framework is equipped with higher recognition accuracy and multimodal ability by minimizing the contrastive pre-training loss. During the inference stage, the CPR-CLIP framework supports intelligent retrieval and batch assessment on large video databases through natural language. In experiments, we recruited doctors and carried out a randomized controlled experiment, which successfully verified the effectiveness of the CPR-CLIP framework for auxiliary assessment.

(4) The existing research on temporal medical action analysis faces problems such as inconsistent action division standards and a lack of study on temporal error detection. To solve these problems, this thesis first constructed the first temporal medical action knowledge graph based on clinical skills textbooks and built the first temporal medical action analysis dataset named ThoSet. This dataset focuses on thoracocentesis and is equipped with fine-grained action labels. The ThoSet dataset can simultaneously support multiple tasks, such as temporal action segmentation, error action recognition, and lost action recognition. In terms of algorithm, inspired by the logical partition structure of human brains, this thesis proposes a feature aggregation module named *k*M-Att based on the temporal clustering attention mechanism and embeds it into the temporal diffusion segmentation model. In addition, this thesis proposes an action compliance assessment algorithm based on temporal action segmentation results and the Dynamic Time Warping (DTW) algorithms. Experimental results on public datasets and the ThoSet dataset fully demonstrate the effectiveness of the proposed algorithms.

To sum up, through the exploration of task paradigm, dataset construction, algorithm design, and system application, this thesis has successfully constructed a fine-grained medical action recognition and skill assessment system. This system fills in some gaps in the research field of medical skill assessment systems, provides some

new ideas for the design of skill assessment algorithms, and lays a foundation for the implementation and application of subsequent skill assessment systems. This study is expected to contribute to the improvement of the efficiency of medical skill training, the rapid supplement of the gap in the number of medical personnel, and the relief of the pressure on the medical service system.

Keywords: Medical Skill Assessment System; Fine-grained Action Recognition; Action Quality Assessment; Video Understanding
CLC number: TP391.4

第1章 绪论

1.1 研究背景与意义

我国正面临着医疗资源短缺、医疗服务系统承压过重、地域医疗资源分布不均匀等问题，填补医务人员数量缺口的举措势在必行。由国家卫生健康委于 2022 年印发的《医疗机构设置规划指导原则（2021—2025 年）》（简称《原则》）在医护比改善方面提出了规划：在 2020 年至 2025 年期间，每千人口执业（助理）医师数由 2.9 提升至 3.2；每千人口注册护士数由 3.34 提升至 3.8；医护比由 1:1.15 提升至 1:1.20；卫生人员比由 1:1.48 提升至 1:1.62。根据第七次全国人口普查数据公布的 14.43 亿人口测算，我国五年间执业医师缺口约为 43.29 万人，护士缺口约为 66.37 万人。考虑到我国老龄化趋势的不断加重，未来国家医疗服务系统中医务人员数量缺口将会持续扩大。《原则》还指出:“要强化信息化的支撑作用，切实落实医院、基层医疗卫生机构信息化建设标准与规范，推动人工智能、大数据、云计算、5G、物联网等新兴信息技术与医疗服务深度融合，构建优质均衡高效的医疗服务体系”。由于医务人员的培养是一个周期长、人力成本高昂的过程，如何充分地利用人工智能和计算机技术，在保证医疗服务质量的前提下提高医务人员的培训效率、降低培训成本成为了缓解医疗服务系统压力的关键问题。

图 1-1　央视报道国家卫健委发布《手术质量安全提升行动方案》

一名合格的医务人员不仅需要掌握深厚的医学理论知识，还需要具备扎实的临床技能功底。与理论知识教学不同，医疗操作技能需要初学者进行反复训练、通过层层考核后才能基本掌握。我国对手术质量安全问题高度重视，如图 1-1 所示，由国家卫健委于 2023 年发布的《手术质量安全提升行动方案》（简称《方案》）从“加强术前风险管理、严格术中风险管理、强化术后风险管理、实现系统持续改进”四方面对医疗机构的手术管理能力提出了更规范的要求，以进一步保障手术质量安全。《方案》对术中操作的风险管理工作进行了反复强调，例如：“强化

手术人员及环节核查，严防手术部位错误、手术用物遗漏、植入物位置不当、手术步骤遗漏等问题；严格执行手术室无菌技术、各项操作流程及技术规范，规范使用抗菌药物、止血药物和耗材”。医务人员对临床技能的熟练程度与真实手术的完成质量息息相关，医疗技能培训质量的提升能够有效减少医疗事故的发生率。

我国现行的住院医师规范化培训（简称规培）制度要求：高等院校医学类专业本科及以上学生，在 5 年医学院校毕业后要以住院医师身份接受 3 年的系统化、规范化培训。培训地点须在由省级以上认定的、具备良好临床医疗和教育培训条件的医院。具体的培训内容是：在经验丰富的上级医师指导下从事临床诊疗，同时接受理论与实践紧密结合的系统化教育培训。完成培训并通过考核的医学生即可获得全国统一的住院规培合格证书。在现行的医疗技能培训与考核模式下，医师往往由医院各科室的一线医生担任。这种以资深医师为主的医疗技能教考方式固然能够保证高质量的教学效果，但是却存在着以下多个方面的问题：

（1）一线医生在诊疗工作之余还需完成医学生的培训与考核任务，导致工作压力过大：考虑到我国目前较低的医护比现状，在保证质量的前提下减轻教学与考核工作压力，从而有效节约医生的宝贵精力至关重要。使用人工智能等技术替代医师完成部分标准化、流程化的教考工作能够有效减轻医师的工作负担。

（2）传统的培训与考核方式难以进行大规模规范化培养，培训效率有待提高：现有的医疗技能培训与考核模式延用传统以资深医师为主的方式，即需要医师对学员的操作进行全程观察与评估。这种方式具有效率低、成本高的缺点，因此难以实现大规模、高效率的医疗技能培养。

（3）我国的医疗教育资源存在地域分布不均衡问题，从而导致医疗技能教育质量不统一：目前我国大城市中顶尖三甲医院所拥有的丰富医疗教育资源难以覆盖至中小城镇，偏远地区面临着医疗教育资源匮乏等问题。传统以资深医师为中心的教考模式会导致地域间技能培训标准不统一、培训质量参差不齐等问题。

本文所探究的**智能医疗技能评估系统**（Intelligent Medical Skill Assessment System）为以上问题提供了重要解决思路。构建智能评估系统的最终目标是替代传统培训与考核方式中的资深医师，其具体流程为：首先通过视频或语音等传感器记录学员的完整操作过程，其次通过人工智能算法对操作过程进行细粒度行为识别和自动化精准评估，最终向医学生或医师反馈评估结果。智能化的医疗技能评估系统能够有效地提升医疗技能培训与考核效率，从而显著减轻一线医生在教考工作中的负担，大幅节约医疗机构在技能培训与考核环节中的人力成本。宏观而言，智能评估系统的应用能够有效提升医院的数字化与信息化建设水平，在一定程度上缩小地域间医疗资源的分布差异，从而有助于提升医疗系统的整体服务质量与效率。

随着机器学习与人工智能技术的繁荣发展，深度学习技术（Deep Learning Technology）在图像识别[1,2]、视频理解[3,4]和内容生成[5,6]等领域中取得了显著成果。在计算机视觉相关研究中，与本文所关注的医疗技能评估系统构建最相关的任务包括：人体行为识别（Human Action Recognition, HAR）、行为质量评估（Action Quality Assessment, AQA）和时序行为分割（Temporal Action Segmentation, TAS）等视频理解任务。图 1-2 对以上三类任务的基本形式进行了展示。其中人体行为识别任务的形式是对剪切后的视频进行分类；行为质量评估任务的形式是对视频中的人的行为进行质量评分或熟练度评级；时序行为分割任务的形式是对未剪辑的长视频进行逐帧行为标签预测。考虑到生活场景中视频数据的易获取性，现有研究对以上任务的探索更多停留在日常行为场景，例如：现有的人体行为识别数据集[7,8]收集了运动、读书、烹饪等日常活动视频；现有的行为质量评估数据集[9,10]关注于跳水、体操等具有录像与评委评分的奥运会场景；现有的时序行为分割数据集[11,12]对做饭等具有一定流程性的日常活动进行了探究。这些日常生活场景中的行为往往具有较小的类内差异（Intra-class Differences）和较大的类间差异（Inter-class Differences），而医疗场景中的行为之间往往具有更小的类间差异和更大的类内差异，由此造成的辨识困难对医疗技能评估系统提出了更高的要求。

图 1-2　人体行为识别、人体行为质量评估和时序行为分割任务形式

目前计算机领域与医学领域中的研究者已开展了初步合作，对智能医疗技能评估系统开展了系列开创性的探索。在数据集构建方面，这些研究基本上延续了以上三类任务（HAR、AQA、TAS）的形式，分别构建了多个医疗技能评估领域中的数据集：医疗行为识别数据集[13,14]、医疗行为质量评估数据集[15–19]和手术流程识别数据集[20–22]；在算法设计方面，这些工作充分参考了视频理解研究领域中的现有算法，并进行了面向医疗场景的迁移和适配。尽管这些研究取得了一定进展，但提出的数据集与算法往往停留在理论研究的层次，现有的医疗技能评估技术距离真正的落地应用仍有较大差距。

针对现有研究的不足，本文围绕智能医疗技能评估系统构建的研究主题，从任务范式创新、数据集构建、算法设计和系统集成应用四个方面开展了创新性探索，通过对**操作质量评估子系统**、**复合错误行为识别子系统**、**多模态复合错误识别子系统**和**时序医疗行为分析子系统**的研究，本文拟构建出具备更广阔应用场景、更精细医疗行为划分粒度、更精准技能评估能力的**细粒度医疗行为识别与技能评估系统**。本研究依托于科技创新 2030—“新一代人工智能”重大项目中的医疗行为感知系统构建任务。本研究对于提升医疗技能培训与考核效率、大幅节约人力成本、缩小地域医疗资源分布差异、提升医院数字化与信息化建设水平等方面均有重要意义。

1.2 国内外研究现状

目前学界中与医疗技能评估系统构建相关的研究主要集中在人体行为识别（HAR）、行为质量评估（AQA）和时序行为分割（TAS）等任务。其中 HAR 和 AQA 均属于静态任务，模型只需对单个视频进行类别预测或质量评估即可；而 TAS 属于动态任务，模型需要对具有较长持续时间的视频进行时序分析。本节对这三类任务进行了充分调研，并对现有的数据集与算法进行了梳理与分类。

1.2.1 人体行为识别技术研究现状

人体行为识别任务旨在通过传感器技术与智能算法实现计算机对人类行为的理解，并对采集到的信息序列进行行为类别预测。现有的行为识别研究根据数据模态可分为两类[23]：视觉模态（Visual Modality）与非视觉模态（Non-visual Modality）。其中前者通过视觉传感器对人体行为进行采集，视觉模态信息包括：RGB 图像与光流[8,24,25]、人体骨架[26,27]、深度图[28]、红外信息[26]、点云[29]和事件流信息[30]（Event Stream），这些不同的传感器往往对应着不同的使用场景。后者通过音频、加速度传感器、雷达和 WiFi 等信号方式对行为进行采集。图 1-3 对这些不同模态下的信息进行了展示。在以上研究中，与本文内容最相关的是基于 RGB 图像、光流图像、人体骨架等视觉信息的人体行为识别算法研究。

由于光流信息可以通过 RGB 图像序列帧获取，因此光流信息一般作为 RGB 图像的辅助信息共同参与行为的预测。基于 RGB 与光流图像的人体行为识别框架按照结构可划分为四类：基于 Two-Stream 双流网络模型的框架[4,8,31]、基于 RNN（Recurrent Neural Network）循环神经网络的框架[18,32,33]、基于 3D 卷积网络的框架[7,34,35]和基于 Transformer 模型的框架[36–38]。研究者在实现了基本的行为识别功能后，分别在主干网络表征能力、视觉特征增强模块、时序信息关联模块和模型训练与推理加速等方面对现有行为识别模型进行了优化。

相较于稠密的 RGB 图像信息，人体骨骼关键点信息（或称 2D/3D 关键点信息）具有高稀疏性、强鲁棒性等优良特性，非常适合作为人体行为的表征模态。现有的基于人体骨骼关键点的行为识别算法按照结构可划分为四个种类：基于 RNN 循环神经网的框架[39–41]、基于卷积神经网络的框架[40,42,43]、基于 GCN（Graph Convolutional Network）图卷积网络的框架[44]和基于 Transformer 模型的框架[45,46]。这些研究探究的主题通常是：如何从可能含有噪声的人体姿势信息中获得稳定的运动表征信息，从而更好地实现人体行为识别任务。

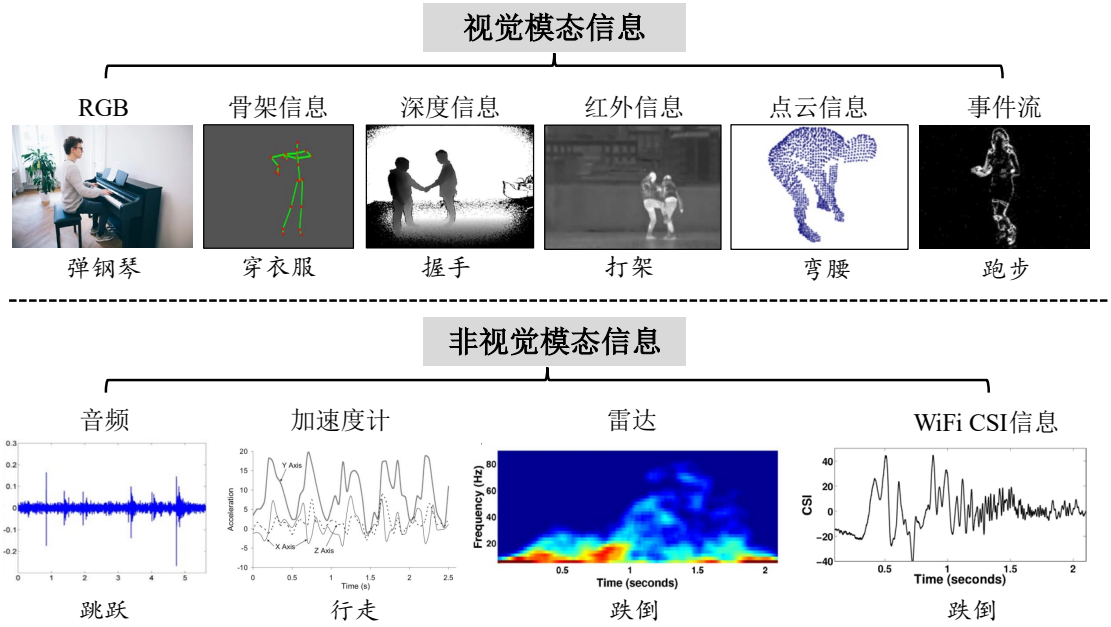


图 1-3　现有行为识别数据集的不同模态数据展示

基于视觉信息的人体行为识别研究虽然再数据集和算法构建方面取得了一定进展，但仍然面临着多方面的挑战，例如：行为标签划分粒度较粗、数据集复杂度低导致性能饱和、算法难以实际应用等问题。针对以上问题，一些研究者开始对**细粒度人体行为识别任务**开展了探究：Shao 等人[47]于 2020 年构建了 FineGym 数据集，使用三个行为层次等级（Events-Sets-Elements）对体操运动中的子行为进行了划分，构建了具备高细粒度、高丰富度的细粒度体操行为识别数据集；Xu 等人[48]于 2022 年构建了 FineDiving 数据集，对完整的跳水过程进行了细致的阶段分割，并提出了时序分割注意力模块用于跳水技能评估任务。

受启发于 FineGym[47]和 FineDiving[48]这类创新性工作，本文对医疗场景下的细粒度行为识别任务进行了探究，拟构建出具有更高行为划分细粒度的医疗行为识别与技能评估数据集，并提出高精度的识别算法。

1.2.2 行为质量评估技术研究现状

相较于行为识别任务，行为质量评估任务（AQA）要对特定行为的完成质量进行优劣性判别，对模型的表征能力有更高的要求。行为质量评估任务旨在通过智能算法自动地完成视频中行为的精准评判，其任务形式通常为：将给定的行为视频映射为指定形式的技能评估结果，例如百分制评分信息、技能等级信息等。行为质量评估系统在医疗、体育和工业等场景中的技能培训环节能够发挥巨大作用。行为质量评估任务与上节所介绍的行为识别任务均属于视频理解任务。

目前学界对人体行为质量评估任务的研究主要关注于体育场景与医疗场景。由于奥林匹克运动会等大型体育赛事通常提供完整的比赛录像回放和专业评委评分结果，研究者们首先探究了体育场景重的行为质量评估数据集构建，并提出了系列算法。这些数据集关注于跳水[9]、滑雪[49]、体操[49]和花样滑冰[50]等丰富的体育运动项目。图 1-4 对以上数据集中的部分案例进行了展示。在算法设计方面，研究者们充分参考了行为识别领域中的模型构建方法，通过常用的视频主干网络对运动视频进行特征提取，最终通过神经网络完成行为的质量预测。

图 1-4　体育场景中的行为质量评估数据集展示

随着行为质量评估模型在体育场景中的有效性被证实，一些研究者开始构建医疗场景中的行为质量评估数据集与算法。这些早期的研究具有非常显著的“机构—主题”关联性。例如约翰斯-霍普金斯大学（Johns Hopkins University, JHU）的 Haro[51]、Zappella[52]和 Malpani[53]等人先后对达芬奇手术机器人（*da Vinci* Robot）的操作技能评估任务进行了探究。之后约翰斯-霍普金斯大学与直觉医疗公司（Intuitive Surgical Inc.）联合发起了 JHU 外科手术语言项目（JHU Language of Surgery Project），旨在推动智能化医疗技能评估技术的发展。在此合作项目中，Gao 等人[15]提出了 JHU-ISI 手术技能识别与评估数据集（JHU-ISI Gesture and Skill

Assessment Working, JIGSAWS），此数据集成为了后续医疗技能评估领域中的重要基准。佐治亚理工学院（Georgia Institute of Technology, GIT）的 Sharma[19,54]和 Zia[16,17,55,56]等人在 OSATS 技能评估框架[57]（Objective Structured Assessment of Technical Skill）下对外科手术基本操作的技能评估任务进行了持续探索，构建了新的基准并提出了评估模型。亚利桑那州立大学（Arizona State University, ASU）的 Islam[58,59]，Chen[60,61]和 Zhang[61–63]等人对腹腔镜手术模拟平台中的技能评估任务进行了探究，并提出了一系列基于概率图模型和多模态融合方法的医疗技能评估算法。在以上三个系列研究之外，Vakanski 等人[64]构建了医疗康复训练中的技能评估数据集 UI-PRMD（Physical Rehabilitation Movement Dataset）。图 1-5 对这些数据集所关注的场景分别进行了展示。整体而言，医疗领域中的技能评估数据集无论是数量还是体量均小于体育领域中的数据集，这主要是因为医疗技能评估数据集的构建需要医生的参与和指导，具有更强的专业性。在体育和医疗场景之外，还有一些研究对日常生活中的行为进行了质量评估探究，例如关注于面团制作、绘画和筷子使用的 Epic Skills 数据集[65]、关注于日常生活技能评估的 BEST 数据集[66]和关注于新生儿抓取能力评估的 Infinite Grasp 数据集[67]。

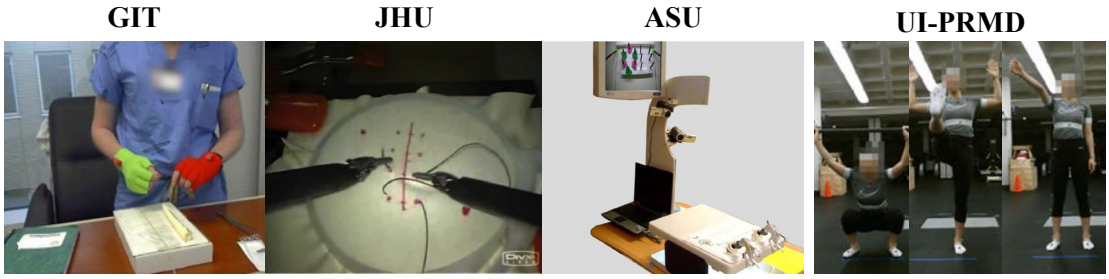


图 1-5　医疗技能评估研究场景

在现有的行为质量评估算法中，研究者通常会直接使用视频理解领域中的主干网络作为特征提取器。这种直接迁移的方法并未考虑行为识别任务与行为质量评估任务之间的差别。传统视频主干网络需要一定的适配才能应用于医疗场景中的行为质量评估任务。

1.2.3 时序行为分割技术研究现状

作为时序行为分析技术的重要前置技术，时序行为分割技术（TAS）在视频理解任务中占有重要地位，其任务形式是：对未剪辑的长视频进行逐帧行为种类预测。现有的时序行为分割任务通常关注于日常生活和外科手术场景。计算机视觉领域的研究者分别构建了 GTEA[12]、50Salads[11]和 Breakfast[68]数据集。这些数据集关注于日常生活中做早餐、制作沙拉等日常活动，为早期的时序行为分割算

法研究提供了测试基准。在医疗场景中，时序行为分割技术是让计算机理解长时手术流程的关键，通常又被称为手术流程识别任务（Surgical Workflow Recognition）。目前医学领域中的研究者对此问题开展了系列探究：国际医学图像处理顶级会议 MICCAI 每年都会发布手术视频识别相关的挑战与竞赛，目前竞赛已经公布了多个数据集：探究胆囊切除术 M2CAI-phase[69]和 Cholec80[20]，同时关注于剖宫产手术、直肠切除手术和乙状结肠切除手术的 HeiCo[70]，探究模拟血管缝合手术的 MISAW[14]，探究腹腔镜胆囊切除术的 CholecT50[71]，探究内窥镜手术的 PETRAW[13]，探究前列腺切除术的 SARAS-MESAD[72]；还有一些研究以白内障手术为研究对象构建了手术流程识别数据集：Cataract-101[21]、CATARACTS[73]；一些工作探究了 *da Vinci* 手术机器人应用场景下的手术流程识别：Nephrec9[22]、ATLAS[74]、RARP45[75]、JIGSAWS[15]，以及其他手术机器人上的流程识别 DESK[76]。图 1-6 对以上数据集的部分案例进行了展示。

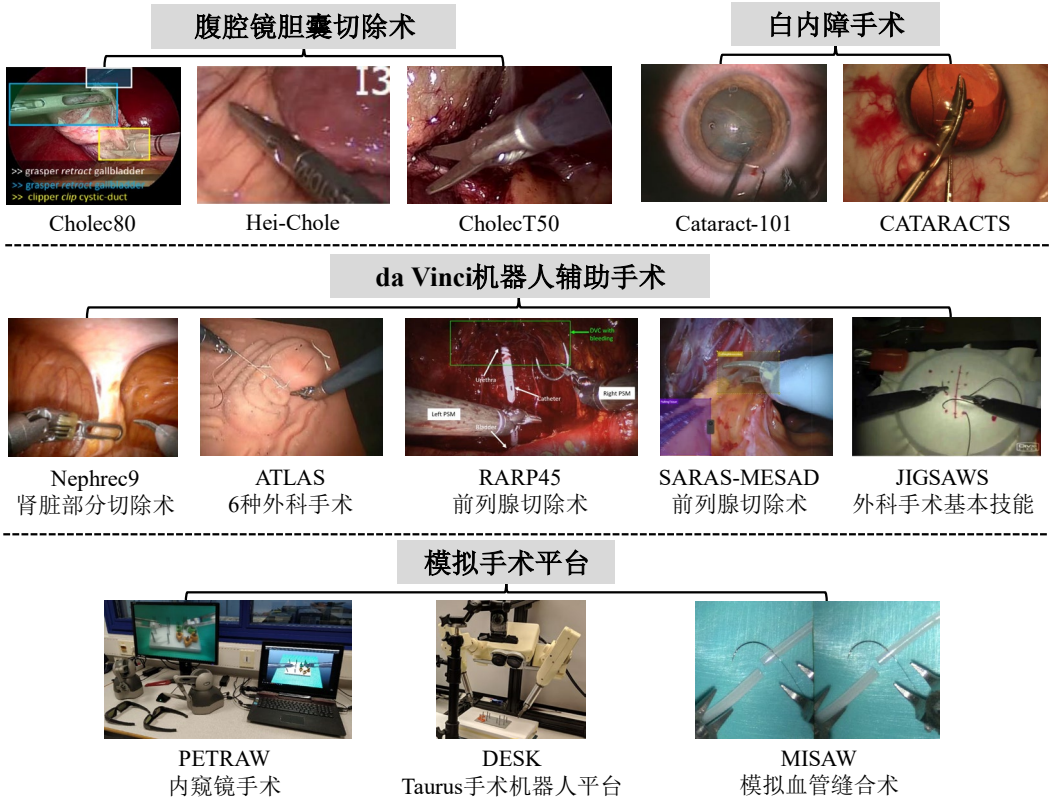


图 1-6　医疗场景中的时序行为分割数据集案例展示

时序行为分割任务与自然语言处理（Natural Language Processing, NLP）任务具有类似的形式：均是对序列信息进行处理和识别。因此早期的时序行为分割算法参考了 NLP 领域中的模型设计。现有的时序行为分割算法按照框架可划分为四类：基于循环神经网络的算法[77]、基于时序卷积的算法[78–84]、基于图卷积神经网络的算法[85,86]和基于 Transformer 模型的算法[84,87–89]。近期计算机视觉领域中

有研究者将多模态预训练[90]、多模态表征学习[91]和扩散模型[92]等新方法引入到了时序行为分割算法中。虽然现有时序行为分割研究能够满足基本的视频理解任务需求，但这些研究只停留在分割（Segment）的层面，并未对行为的正确性与合规性进行判别，即还没有到达分析（Analysis）的层面。构建具有更高时序行为划分粒度、支持更多时序行为分析任务的数据集是未来发展的主要方向。

1.2.4 任务界定与术语辨析

学界中已有一些研究对智能医疗技能评估系统开展了初步探索。整体而言，这些研究存在主题分散、任务界定不清晰、名词术语不明确等问题，核心原因是缺乏统一的任务描述框架。为了更清晰地组织本文研究内容，本文首先对与医疗技能评估系统相关的任务进行了界定和术语辨析。

- **医疗技能评估任务**（Medical Skill Assessment）：本文的研究主题，泛指通过计算机技术实现医疗操作技能的自动评估，又被称为医疗行为分析任务（Medical Action Analysis），包括但不限于以下所有任务。
- **医疗行为质量评估任务**（Medical Action Quality Assessment）：学界中现有的行为质量评估任务根据输出类型可分为两类：医疗技能评分任务（Medical Skill Scoring）和技能等级评定任务（Medical Skill Rating）。行为质量评估模型需要对操作视频进行分数回归或技能等级预测，任务形式较为简单直观。
- **医疗行为识别任务**（Medical Action Recognition）：与视频理解领域中的人体行为识别任务（HAR）形式类似，对采集到的医疗行为视频进行识别与分类。此类任务中每个视频通常只有一个类别标签，无法支持错误行为识别任务。
- **细粒度医疗行为识别任务**（Fine-grained Medical Action Recognition）：任务形式与 HAR 任务相同，主要区别在于：此类任务具有更高的行为标签划分粒度，模型不仅要识别行为的种类，还需要对错误行为进行识别。由于医疗行为之间的微弱差异，此任务相较于 HAR 任务具有更大的难度。
- **复合错误行为识别任务**（Composite Error Action Recognition）：本文首次提出了此任务范式，并构建了心肺复苏术中的复合错误行为识别数据集 CPR-Coach。此任务假定训练集中只含有单类错误样本，而测试集中含有多类复合错误样本。模型需要在监督信息严重受限的情况下实现细粒度的错误辨识。此类任务的探究对医疗技能评估算法的实际应用有重要意义。
- **时序行为分割任务**（Temporal Action Segmentation, TAS）：医疗场景下的时序行为分割技术研究更多关注于外科手术场景，因此又被称为手术流程识别任务（Surgical Workflow Recognition），任务形式是对持续时间较长的手术视频进行逐帧的操作类别预测。

- **时序医疗行为分析任务**（Temporal Medical Action Analysis）：首先对未剪辑的长视频进行时序行为分割，其次在预测的标签序列基础上完成遗漏行为和错误行为的检测，最终实现时序行为的合规性评估。时序行为分割技术（TAS）为此任务的前置技术。目前学界中尚无数据集能够支持此任务。

本研究的总体目标是实现“智能医疗技能评估系统”的构建，由于“细粒度医疗行为识别”任务在研究过程中占据重要地位，因此本文将拟构建的总系统命名为：**细粒度医疗行为识别与技能评估系统**。此系统共由四个子系统构成，分别为：操作质量评估子系统、复合错误行为识别子系统、多模态复合错误识别子系统和时序医疗行为分析子系统。总系统、各子系统与隶属任务之间的对应关系如表 1-1 所示。在各个子系统的构建过程中，本文针对医疗技能评估不同应用场景的难点分别开展了数据集创新与算法创新探究。通过实现四个子系统的构建，本文拟提出一套具备更广阔应用场景、更精细医疗行为划分粒度、更精准技能评估能力的医疗技能评估系统。

表 1-1　本文研究对象、各子系统与隶属任务间的对应关系

研究对象/总系统名称	子系统名称	隶属任务
细粒度医疗行为识别与技能评估系统	操作质量评估子系统	医疗行为质量评估任务 医疗技能评分任务
	复合错误行为识别子系统	复合错误行为识别任务 细粒度医疗行为识别任务
	多模态复合错误识别子系统	复合错误行为识别任务 细粒度医疗行为识别任务
	时序医疗行为分析子系统	时序行为分割任务 时序医疗行为分析任务 细粒度医疗行为识别任务

1.3 研究难点与挑战

上节对医疗技能评估系统的国内外研究现状进行了综述，并对相关任务进行了界定与辨析。整体而言，目前学界中对医疗技能评估技术的研究尚处初期，现有的医疗技能评估数据集面临着行为划分粒度粗、错误行为识别研究匮乏等问题；现有的技能评估算法面临着性能受限、人机交互能力弱、无法满足真实应用需求等问题。为解决以上问题，本文拟构建一套细粒度医疗行为识别与技能评估系统。本文将研究的难点与挑战总结为以下四个方面。图 1-7 展示了研究难点、研究方法与各子系统之间的关联。

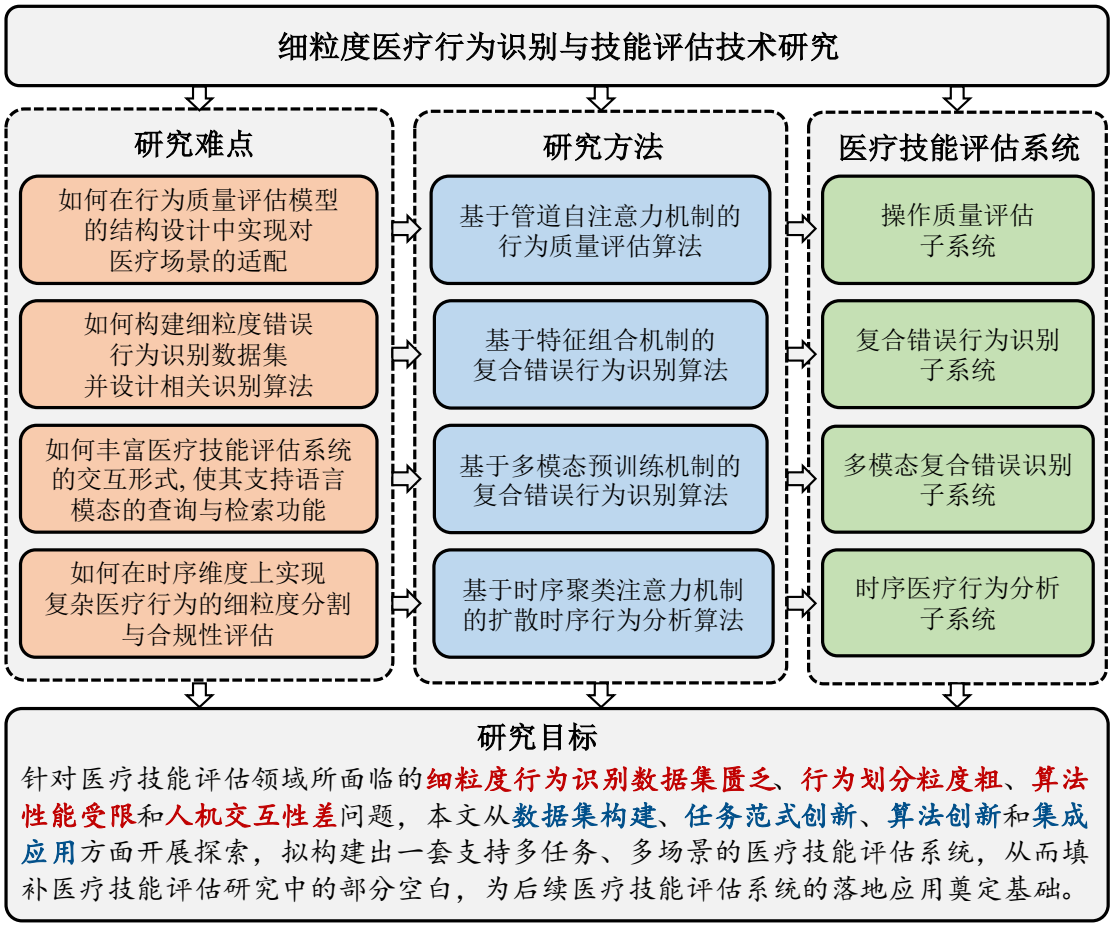


图 1-7　本文的研究难点、研究方法、系统构成与研究目标

研究难点 1：如何在技能评估模型设计环节实现对医疗场景的适配。在现有的医疗技能和行为质量评估模型中，大部分算法直接将计算机视觉领域中的视频主干网络用作为特征提取器。这种直接迁移策略在一方面能够使模型受益于视频主干网络的预训练过程，但在另一方面却会限制模型在医疗场景下的性能，因为适用于普通场景的行为视频框架并不一定适用于医疗场景。通过对比生活场景（图 1-3）和医疗场景（图 1-5）即可发现，医疗行为视频往往具有更低的视觉分辨特性。特征丰富度的缺失导致各医疗行为之间只有细微的差异，从而引起行为质量评估的困难。因此，进一步提升现有模型性能的技术挑战在于：通过视频主干网络的结构改良，实现技能评估模型向医疗场景的适配。

研究难点 2：如何构建细粒度错误行为识别数据集并设计复合错误识别算法。现有医疗行为识别数据集与算法面临着两方面的问题：一方面，现有数据集对医疗行为的划分粒度较粗，基于这些数据集构建的模型无法满足实际应用需求，因为数据集中的行为标签划分细粒度直接决定了模型识别能力的上限；另一方面，现有模型只支持多个行为的分类任务，目前尚无研究对医疗操作过程中发生的错误进行探究。细粒度错误行为识别与组合错误识别研究在学界中实属空白，而这

些任务却恰恰是医疗技能评估系统的重要构成部分。因此，细粒度错误行为识别研究的技术挑战在于：在选定一个医疗行为作为研究对象后，构建高细粒度行为标签空间，完成细粒度错误识别数据集构建，并基于此提出错误行为辨识算法。

研究难点 3：如何丰富医疗技能评估系统的交互形式，使其支持语言模态信息查询和检索等功能。目前学界对医疗技能评估模型的研究仍停留在理论层次，这些研究往往止步于模型的性能测试阶段，因此产出的系统与落地应用之间还有较大鸿沟。目前尚无研究对医疗技能评估系统的易用性进行针对性改善。对于医生而言，一款优秀的辅助评估系统应当具备基本的人机交互功能：即支持通过自然语言对大规模医学生操作案例库进行智能检索与批量评估。多模态学习方法为此功能的实现提供了重要思路。通过不同模态信息之间的交互，多模态模型能够实现异源特征的关联与对齐，从而在提升模型性能的同时拓宽模型的表征能力。因此，有效改善技能评估系统交互能力的技术挑战在于：充分利用多模态学习方法，将多模态学习技术与技能评估模型相耦合，从而显著改善系统的人机交互能力、提高系统易用性。

研究难点 4：如何在时序维度上实现复杂医疗行为的细粒度分割与合规性评估。在现有的时序行为分割研究中，无论是计算机视觉还是医疗领域中的数据集均存在两方面问题：一方面，现有数据集的流程划分标准各异，尚无统一的时序行为划分方法与体系。各异的标准导致现有时序行为分割数据集的复杂度无法进一步提升；另一方面，现有数据集只提供了正确的时序操作，并未提供错误流程案例供算法进行学习与辨识。以上数据集只支持时序行为分割任务，无法支持更复杂的时序行为分析任务。对于本文拟构建的医疗技能评估系统而言，在时间的维度上进行错误操作识别与分析功能不可或缺。因此，时序医疗行为分析系统构建的技术挑战在于：选定一个具有较高复杂度的医疗行为，在时序维度上进行操作的细化拆分，并构建出一套同时支持时序行为分割、错误和遗漏操作检测等功能的数据集，并提出时序行为分析算法。

1.4 研究内容与贡献

本文围绕细粒度医疗行为识别与技能评估系统构建的研究主题，分别对医疗技能质量评估算法、复合错误识别数据集构建与算法、融合多模态技术的医疗技能评估系统、时序医疗行为分析数据集构建与算法四项内容开展探究，并将提出的算法应用在对应的四个任务场景中，构建了相应的操作质量评估子系统、复合错误行为识别子系统、多模态复合错误识别子系统、时序医疗行为分析子系统。最终搭建了一套具备更广阔应用场景、更精细行为划分粒度、更精准评估能力的医疗技能评估系统。图 1-8 对总系统构建的技术路线进行了展示。

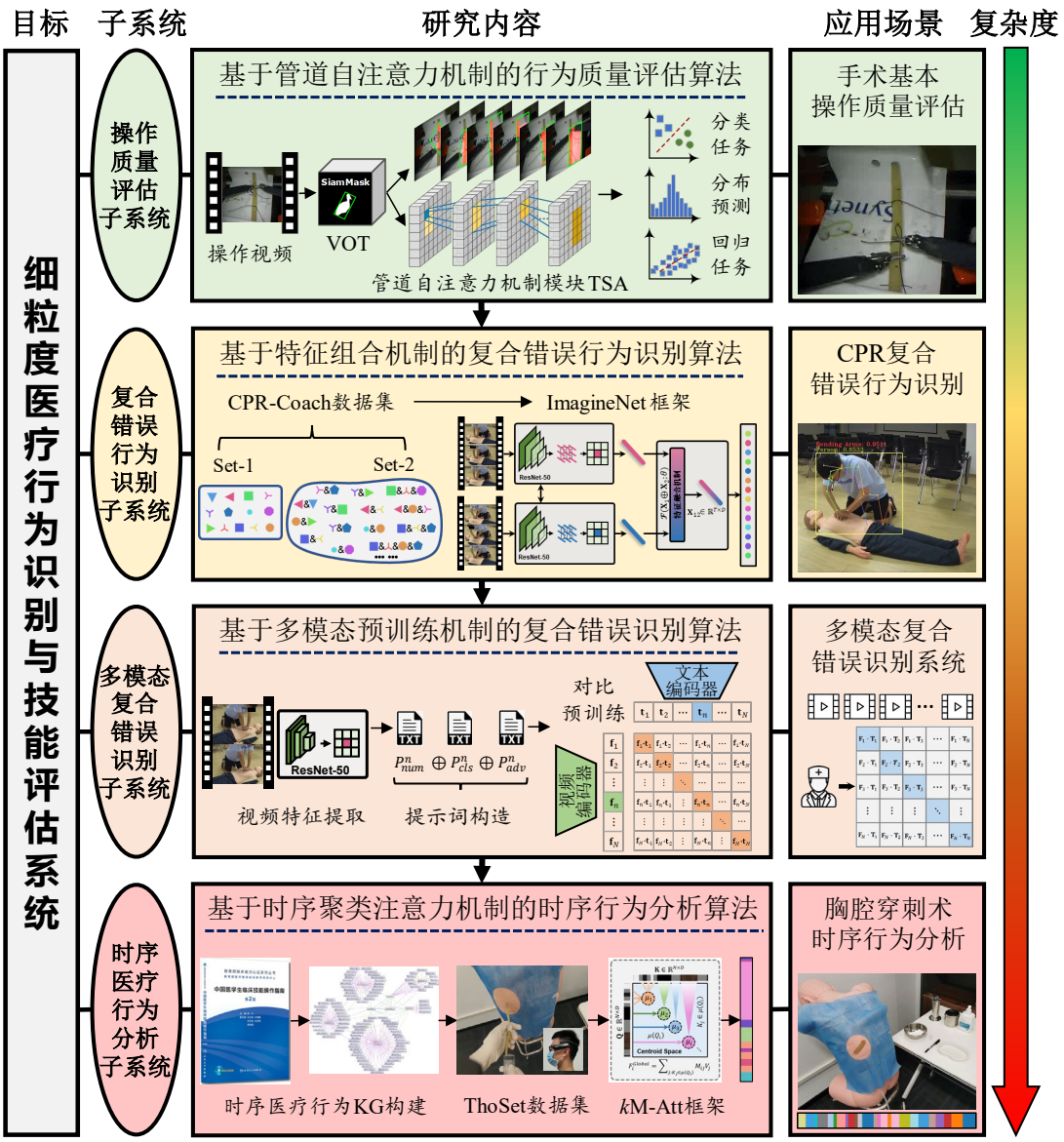


图 1-8 本文的技术路线图

本文的四项核心**研究内容**总结如下：

（1）基于管道自注意力机制的行为质量评估算法

现有医疗技能评估模型通常直接采用视频理解领域中的主干网络，这种做法忽略了行为识别任务与技能评估任务之间的区别，技能评估任务对网络的表征能力有着更高的要求。针对此问题，本文将单目标跟踪技术引入到技能评估模型构建中，提出了一种基于时空管道（Spatial-temporal Tube, ST-Tube）的稀疏高效的特征交互方法，并将其命名为管道自注意力模块（Tube Self-attention Module, TSA）。由于单目标跟踪框能够为评估模型提供位置先验知识，TSA 模块能够高效完成特征增强。本文进一步构建了行为质量评估框架 TSA-Net。在手术机器人操作技能评估和体育运动行为质量评估任务中的实验结果显示，TSA-Net 框架相较于经典特征增强方法能以更少的计算开销取得更优越的性能。

（2）基于特征组合机制的复合错误行为识别算法

针对现有医疗技能评估研究领域中存在的行为种类划分粒度粗、错误操作识别研究匮乏等问题，本文提出了复合错误行为识别（Composite Error Action Recognition）这一任务范式，并将心肺复苏术（CPR）中的胸外按压行为设定为研究对象，构建了同时支持单类错误识别任务和复合错误识别任务的细粒度错误行为辨识数据集 CPR-Coach。此数据集共含有 13 类单错误行为和 74 类复合错误行为。考虑到真实世界医疗技能评估场景中普遍存在的“单类训练，多类测试”这种监督信息受限现象，本文受启发于人类的想象力机制提出了基于特征组合训练机制的 ImagineNet 框架。此框架能够有效缓解由训练集与测试集数据分布差异过大所引起的识别性能低下问题。在 CPR-Coach 数据集上开展的实验充分证实，ImagineNet 框架能够显著提升传统行为识别模型在复合错误行为识别任务中的性能。

（3）基于多模态预训练机制的复合错误行为识别算法

现阶段的医疗技能评估系统研究依然停留在理论层次，距离真正落地应用仍然存在较大差距，其原因在于现有工作仅关注于数据集的体量扩充与算法性能的提升。本文针对现有医疗技能评估系统交互性差、无法满足实际评估应用需求等缺陷，在第二部分研究内容的基础上对技能评估模型的性能和易用性进行了同时改进。本文将多模态对比学习模型和提示词工程引入到了复合错误行为识别任务中，提出了多模态对比预训练框架 CPR-CLIP。具体而言，CPR-CLIP 框架首先通过错误数量、错误种类、改正建议三个不同的角度进行提示语句构建，其次再通过最小化对比预训练损失的方式完成语言与视觉模态中的特征对齐，最终使模型具备更高的复合错误识别精度。此外，CPR-CLIP 框架还能够支持以自然语言的方式对操作视频库进行智能化检索与批量评估。在系统的实际辅助能力探究方面，招募医生开展了随机对照试验，实验结果证实了 CPR-CLIP 框架在实际医疗技能评估应用中的有效性。

（4）基于时序聚类注意力机制的扩散时序行为分析算法

本章针对现有时序医疗行为分析研究中时序行为划分标准不统一、时序错误操作研究匮乏等问题开展探究。本文首先依据《中国医学生临床技能操作指南》教材设计并构建了时序医疗行为知识图谱，此知识图谱能够同时对医疗行为在知识层面和流程层面进行准确表述，为后续医疗时序行为分析研究奠定基础。其次，本文以胸腔穿刺术为研究对象，创建了具有高细粒度行为标签的时序医疗行为分析数据集 ThoSet。此数据集能够同时支持时序行为分割、错误操作识别与遗漏行为检测任务。之后，受启发于人类大脑的逻辑分区结构，本文提出了基于时序聚类注意力机制（*k*-means Clustering Attention Mechanism）的特征增强模块 *k*M-Att，

并将其应用于时序扩散分割模型的构建。在实验部分中，本文将提出的时序分割框架应用于多个计算机视觉中的时序行为分割公开数据集和 ThoSet 数据集，实验结果证实了算法的有效性。

本文各章研究内容与医疗技能评估领域中子任务关联情况如图 1-9 所示。

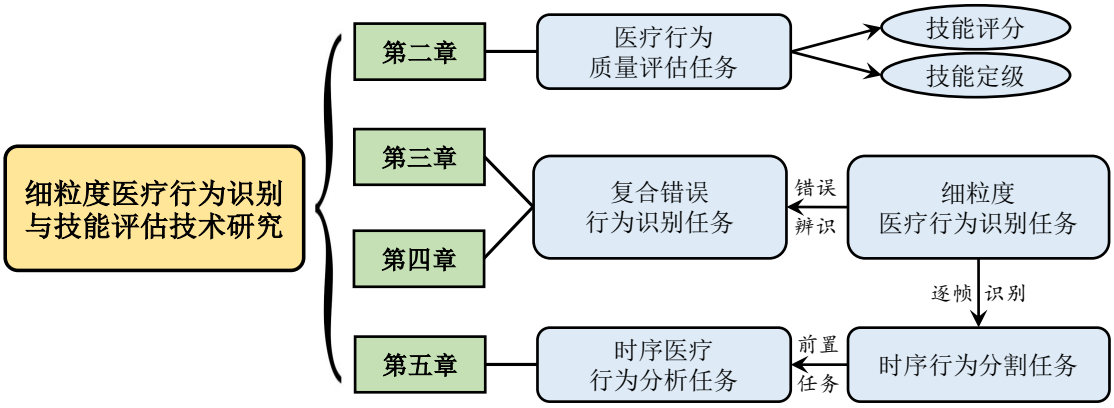


图 1-9 医疗技能评估领域子任务与本文章节关联

本文的**研究目标**为：针对医疗技能评估研究领域所面临的细粒度行为识别数据集匮乏、行为划分粒度粗、算法性能受限、人机交互性差等问题，本文在四项研究内容中从数据集构建、任务范式创新、算法创新和集成应用四个方面开展探索，拟构建出一套支持多任务、多场景的医疗技能评估系统，从而填补医疗技能评估研究领域中的部分空白，为后续技能评估系统的落地应用奠定基础。

本文的**研究贡献**归纳为以下四个方面：

- 创建了 **2 个医疗技能评估数据集**：心肺复苏场景中复合错误行为识别数据集 CPR-Coach、胸腔穿刺术中细粒度时序医疗行为分析数据集 ThoSet，改善了目前医疗技能评估领域所面临的行为划分粒度粗的研究现状。
- 提出了 **4 个医疗技能评估算法**：基于管道自注意力机制的行为质量评估算法 TSA-Net、基于特征组合机制的复合错误行为识别算法 ImagineNet、基于多模态预训练的复合错误识别算法 CPR-CLIP、基于时序聚类注意力机制的扩散时序行为分析算法，这些算法有效地解决了不同医疗技能评估场景中的难点。
- 构建了 **1 个时序医疗行为知识图谱**：为实现 ThoSet 数据集中细粒度时序行为标签体系构建，本文首先依据培训教材构建了具备时序表述能力的医疗行为知识图谱，为后续复杂时序医疗行为的探究提供了思路。
- 实现了 **1 次医疗技能评估系统的应用测试**：为探究多模态框架 CPR-CLIP 在实际应用中的辅助评估能力，本文在第四章中招募医生并设计了随机对照试验，为后续医疗技能评估系统的落地应用研究奠定了基础。

本文的**创新点**总结为以下四个方面：

- 本文将单目标跟踪器引入到行为质量评估任务中，提出了基于时空管道自注意力机制的 TSA 模块和 TSA-Net 框架。TSA 模块能够依据跟踪框结果对视频特征进行选择性增强。在医疗与体育场景中的多个技能评估数据集上的实验结果证实，TSA-Net 以更少的计算开销达到了更优良的行为质量评估性能。
- 本文提出了复合错误行为识别任务范式，并将 CPR 胸外按压行为设定为研究对象，构建了支持细粒度错误辨识任务的数据集 CPR-Coach。针对“单类训练，多类测试”的监督信息受限条件，本文提出了基于特征组合训练机制的 ImagineNet 框架，并通过充分的实验证实了此框架的有效性。
- 本文将多模态预训练框架和提示词工程方法引入到复合错误行为识别任务中，并提出了支持语言检索与批量评估功能的多模态医疗技能评估框架 CPR-CLIP，充分的模型性能对比和随机对照试验结果证实了该框架的有效性。
- 本文提出了一套完整的时序医疗行为分析方法，在设计并构建时序医疗行为知识图谱的基础上，本文以胸腔穿刺术为研究对象构建了同时支持时序行为分割与分析任务、具有高细粒度行为标签的 ThoSet 数据集。在算法创新方面，本文提出了基于时序聚类注意力机制的 *k*M-Att 模块并将其应用在扩散时序行为分割模型中。基于高质量的时序行为分割结果，本文基于 DTW 算法设计了医疗行为合规性评估算法。

1.5 全文组织结构

本文研究内容针对医疗技能评估研究领域所面临的数据集匮乏、算法性能受限、行为标签粒度粗、人机交互性差等问题，在数据集构建、任务范式创新、算法创新和集成应用四个层面进行了探索，填补了目前医疗技能评估研究领域中的部分空白，为医疗技能评估系统的进一步发展和落地应用打下了基础。本文共由六个章节构成，全文组织结构如图 1-10 所示。各个章节的主要工作总结如下：

第一章，绪论。此章介绍了本文的研究背景与意义，对医疗技能评估技术相关的国内外研究现状进行了总结，对现有的技能评估领域中众多任务与术语进行了界定与辨析，并对目前技能评估任务所面临的难点和挑战进行了分析，最终对本文的研究内容、贡献和创新点进行了概述。

第二章，基于管道自注意力机制的行为质量评估算法。针对现有医疗技能评估模型存在的视频主干网络表征能力差的问题，本章引入单目标跟踪技术实现视频运动区域的自动框选，并提出了一种稀疏高效的特征交互机制和管道自注意力模块（TSA）。基于此模块构建的行为质量评估框架 TSA-Net 在医疗技能评估和体育运动行为质量评估任务中取得了优异的性能。

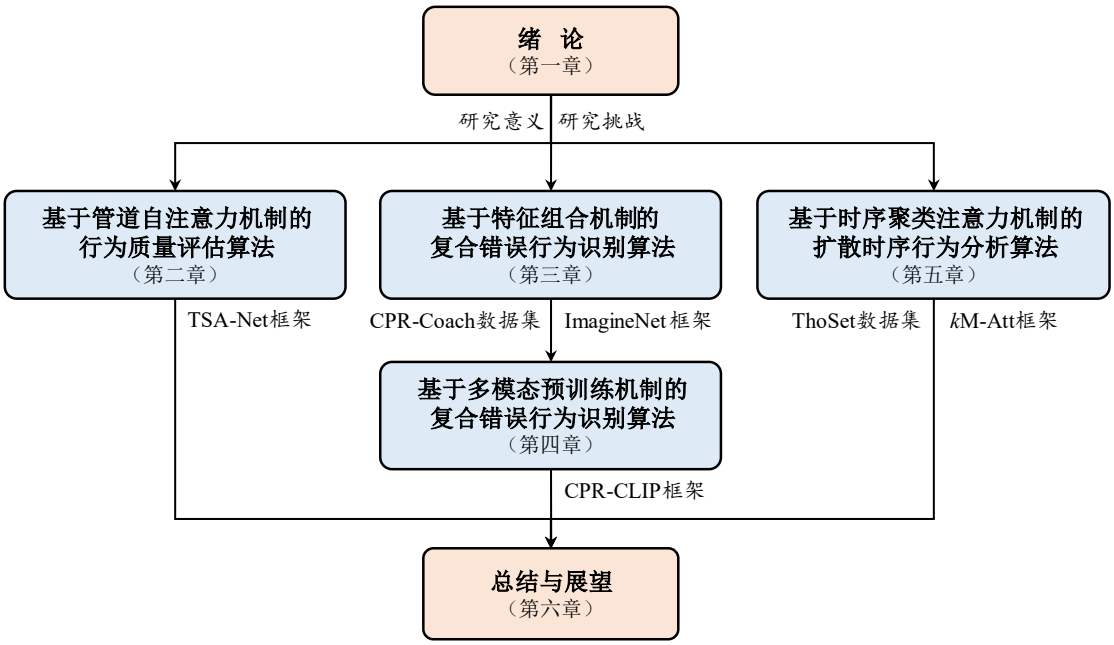


图 1-10　本文组织结构图

第三章，基于特征组合机制的 CPR 复合错误行为识别算法。针对目前医疗技能评估数据集行为划分粒度粗、对错误行为识别研究匮乏的现状，本章提出了复合错误行为识别任务范式，并构建了支持细粒度错误行为识别数据集 CPR-Coach。针对实际应用中普遍存在的“单类训练，多类测试”监督信息限制，本章提出了基于特征组合训练机制的 ImagineNet 框架，并在 CPR-Coach 数据集上开展实验验证了算法的有效性。

第四章，基于多模态预训练机制的复合错误行为识别算法。针对现有医疗技能评估系统的人机交互能力弱、无法满足实际应用需求等问题，本章将多模态对比学习方法与提示词工程引入到了复合错误行为识别任务中，提出了多模态对比预训练框架 CPR-CLIP。此框架的推理模式支持通过自然语言对操作视频库进行智能化检索和批量评估。在 CPR-Coach 数据集上的性能对比和随机对照试验证实了 CPR-CLIP 框架的有效性。

第五章，基于时序聚类注意力机制的扩散时序行为分析算法。针对现有时序医疗行为分析研究所面临的行为划分标准不统一、时序错误操作研究匮乏等问题，本章依据临床技能指南教材构建了时序医疗行为知识图谱，为后续复杂数据集的构建奠定了基础。其次，本章以胸腔穿刺术为研究对象，构建了能够同时支持时序行为分割、错误与遗漏操作识别任务的时序医疗行为分析数据集 ThoSet。之后，受启发于人类大脑结构，本章提出了基于时序聚类注意力机制的时序行为分割框架 *k*M-Att 和行为合规性评估算法。充分的实验证实了算法的有效性。

第六章，总结与展望。此章对全文工作进行了回顾与总结，并且从数据集构建、算法设计、落地应用等多个方面对医疗技能评估技术发展趋势进行了展望。

第2章 基于管道自注意力机制的行为质量评估算法

2.1 引言

医疗行为识别任务（Medical Action Recognition）的目标是对视频中出现的医疗操作进行分类。目前简单的行为识别任务已被充分探究，而根据视频对行为进行质量评估逐渐引起了越来越多研究者的关注。相较于传统的行为识别任务（HAR），人体行为质量评估任务（Action Quality Assessment, AQA）面临着更大的技术挑战：前者只需识别出视频中的特定行为，而后者需要在理解行为的基础上对行为的完成度和质量进行评估。行为质量评估模型在现实世界中有着广泛的应用场景，例如体育领域中的运动技能分析[49,50,93]、医疗领域中的手术操作技能评估[16,53,56]、工厂技能培训场景中的智能考核系统等。稳定且准确的行为质量评估系统能够为各行各业的技能培训环节节省巨额人力成本。

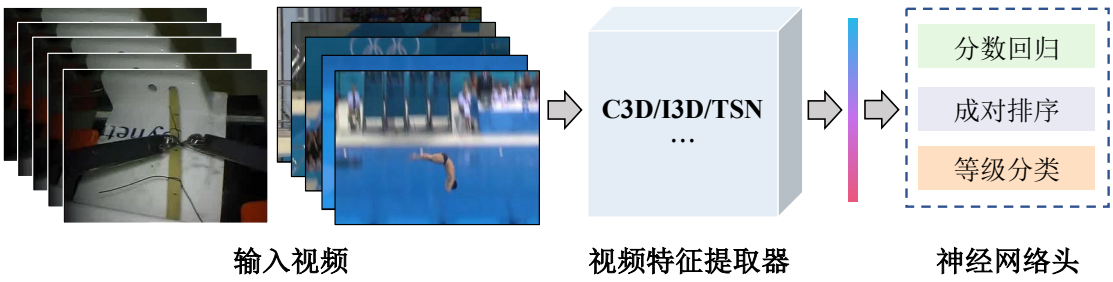


图 2-1　现有行为质量评估模型的通用结构

学界现已对行为质量评估任务开展了初步探究，这些研究主要关注于体育运动场景与医疗技能评估场景，还有少部分研究关注于日常行为。得益于数据的丰富性和易获取性，体育领域中的 AQA 数据集较医疗领域具有更大的体量和丰富度。在模型设计方面，图 2-1 展示了现有行为质量评估模型的通用结构。大部分行为质量评估模型[49,50,93]在特征提取环节直接借用了行为识别模型中的视频主干网络，例如 TSN[4]、I3D[7]、C3D[34]等。虽然这些算法在多个公开数据集中取得了一定的性能，但是仍面临着多方面的问题：在任务差异方面，设计模型时应充分考虑 HAR 和 AQA 任务之间的差异：HAR 模型需要区分不同行为之间的差异，而 AQA 模型需要对同类视频中的优劣性进行判别。后者对视频主干网络的表征能力有更高的要求；在时空上下文信息建模方面，现有模型所使用的视频主干网络通常由卷积层构成，其感受野的尺度严重依赖于卷积核的大小，因此模型无法捕获到远距离的特征关联信息。因此，直接将视频主干网络应用于 AQA 任务会

造成模型性能受限的问题。

针对以上问题，本文提出了时空管道自注意力机制模块（Tube Self-Attention Module, TSA），并基于此模块构建了行为质量评估框架 TSA-Net。TSA 模块的本质是一种基于时空管道机制与自注意机制的高效特征增强方法，其具体实现方式为：首先通过单目标跟踪器[94]（Visual Object Tracking Model, VOT）生成物体的追踪框序列，其次通过检测框与特征图对齐操作获得时空管道（Spatial-temporal Tube, ST-Tube），最终通过自注意力机制完成时空管道内部的特征元素交互。TSA 模块的稀疏特征交互思想来源于对行为质量评估视频的观察：例如，在评估 *da Vinci* 手术机器人的操作熟练程度时，行为质量评估模型关注点应聚焦于两个机械臂和人体组织的接触区域；在跳水技能评估中，行为质量评估模型应重点关注运动员的姿势变化，并忽视背景中的干扰因素。为模拟这种选择性注意力机制，TSA 模块通过引入单目标跟踪器实现了视频中感兴趣区域（Region of Interest, RoI）的自动化框选。此外，TSA 模块由于具备输入输出特征尺寸不变的特性，因此可以被便捷地嵌入在现有视频主干网络中。本文通过 I3D[7]主干网络与 TSA 模块进行组合实现了 TSA-Net 的构建。

本文所提出的 TSA 模块同时具备三个方面的优点：（1）高效率：时空管道机制使模型关注特征图中的元素子集而非全集，从而实现了计算复杂度的大幅降低；（2）有效性：TSA 模块采用自注意机制对时空管道内的特征元素进行特征增强，有效保留了时间和空间维度中的上下文信息，削弱了特征图中冗余信息的影响；（3）灵活性：与 Non-local[35]模块类似，TSA 模块能够以即插即用（Plug-and-Play）的形式嵌入在具有任意视频特征尺寸的神经网络中。在实验部分，本文在医疗技能评估数据集 JIGSAWS[15]和体育行为质量评估数据集 AQA-7[9]、MTL-AQA[49]中，对 TSA-Net 的性能与计算复杂度进行了探究。

本章的主要贡献概括如下：

1、本章在基于视频的行为质量评估任务中提出了一种稀疏高效的特征交互策略，称为管道自注意力 TSA 模块。此模块能够根据单目标跟踪器生成的目标框序列对视频特征进行选择性增强，最终生成具有丰富时空上下文信息的特征。

2、本章基于 TSA 模块和 I3D 视频主干网络构建了行为质量评估框架 TSA-Net。与传统的 Non-local 特征增强框架相比，该框架能够以更低的计算量实现更高的行为质量评估精度。

3、在实验部分，本章分别在医疗行为质量评估和体育行为质量评估任务中的公开数据集对 TSA-Net 模型的性能进行了测试和对比，实验结果充分证实了 TSA-Net 框架的通用性、有效性和高效性。

2.2 行为质量评估任务与自注意力机制

2.2.1 行为质量评估数据集与算法

行为质量评估任务（AQA）的目标为：依据摄像头和运动学传感器记录的信息对操作者的行为进行质量评估。AQA 技术在体育、医疗和工厂技能培训等场景中有着广阔的应用前景。学界中现有的 AQA 研究依据主题可划分为两类：体育场景中的运动员行为质量评估和医疗场景中的手术操作技能评估。得益于国际大型体育赛事提供高清录像与详细评分记录，体育场景中的 AQA 数据集具有较低的构建难度和成本。例如对跳水、滑雪、体操和花样滑冰等运动进行探究的 AQA-7[9]、MTL-AQA[49]、FisV-5[50]、FR-FS[95]数据集。体育场景中的 AQA 算法按照输入模态信息可划分为两类：基于姿势的评估模型[93,96,97]（Pose-based Models）和无姿势评估模型[49,50,98]（Non-pose Models）。前者首先对视频中的运动员进行姿势估计，再根据关键点信息完成行为质量评估任务；而后者省略了姿势估计环节，直接使用神经网络模型对视频进行建模。由于体育运动场景中的人体姿势基本处于极端运动状态，姿势估计结果中会含有大量噪声，最终导致模型的评估性能低下。因此，大部分体育行为质量评估模型会采用第二种方案。

医疗领域中 AQA 数据集的数量和体量均要少于体育领域，这主要是因为医疗数据集的构建需要医生的指导与参与，具有较高的专业性和构建成本。学界中现有的医疗行为质量评估研究具有显著的机构相关特性：约翰斯-霍普金斯大学（JHU）的系列研究[15,17,51,99–102]主要关注于机器人最小介入手术场景（Robotic Minimally Invasive Surgery, RMIS）中的医疗行为质量评估；佐治亚理工学院（GIT）的系列研究[16,17,19,54–56]关注于 OSATS[57]技能评估体系（Objective Structured Assessment of Technical Skill）下的模拟手术技能评估；亚利桑那州立大学（ASU）的系列研究[58–63,103]关注于腹腔镜模拟平台上的手术操作技能评估。此外，Vakanski 等人[64]构建了 UI-PRMD 数据集对医疗康复训练中的人体行为进行评估，此数据集提供了视频、动作捕捉和姿势三种多模态信息。绪论章节中的图 1-5 对这些研究所关注的场景进行了总结。图 2-1 对现有行为质量评估框架的结构进行了归纳总结，这些框架通常直接采用 HAR 模型中的主干网络对视频进行特征提取。这种直接迁移的方法并未充分考虑视频分类任务和质量评估任务之间的差异，从而导致模型性能无法进一步提高。

2.2.2 自注意力机制与上下文信息建模

自注意力机制（Self-attention Mechanism）最初起源于自然语言处理领域：Vaswani 等人[104]于 2017 年提出了 Transformer 模型，此模型在机器翻译任务中取得了显著的性能提升。之后研究者们基于自注意力机制与 Transformer 框架提

出系列语言模型，例如 BERT[105]、GPT[106]、T5[107]等，这些语言模型对机器翻译、对话系统构建、文本生成等任务产生了深远的影响。近期学界中涌现的诸多大语言模型例如 ChatGPT[108]、GPT-4[109]、ChatGLM[110]等均以自注意力机制作为最基本的构成元素。自注意力机制的本质是：在嵌入特征空间中，在序列中的每个位置上计算所有位置的加权平均响应。在计算实现的层面，自注意力机制的计算过程可等效为一系列矩阵乘法，因此相较于传统的循环神经网络（RNNs），自注意力机制具备更加优良的并行特性与训练稳定性。

随着自注意力机制在自然语言处理领域中的广泛应用，研究者们逐渐将其引入到了计算机视觉领域中的诸多任务中，例如图像分类[111]、目标检测[112]和语义分割[113]。Wang 等人[35]于 2018 年提出的 Non-local 网络是视觉任务中引入自注意力机制的开创性工作。Non-local 网络充分借鉴了图像去噪任务中的非局部平均算法[114]（Non-local Means）的思想，并且将此机制与自注意力机制进行了融合。其计算思路为：在视频特征或图像特征图中，在特定位置通过对所有位置上的元素进行加权求和完成响应值计算。这种跨越时间维度和空间维度的注意力操作赋予了模型对超长时空信息连续性建模的能力，从而有效弥补了传统卷积神经网络的感受野受限问题。Non-local 网络在视频分类、目标检测、目标分割和姿势估计任务中均取得了优异的性能。

受启发于 Non-local 模块，Huang 等人[115]提出了十字交叉注意力模块（Criss-Cross Attention Module, CCAM）和语义分割框架 CCNet。通过特征稀疏建模，CCNet 规避了全局注意力机制中的密集运算，高效地实现了视觉特征中的上下文信息建模。具体而言，CCAM 首先从水平与竖直两个方向上对特定位置的特征元素进行注意力机制计算，之后通过双层循环堆叠的方式实现全局信息建模。这种稀疏交互的方式赋予了 CCAM 模块优越的计算速度和内存节省特性。

还有多种模型在上下文信息建模过程中引入了注意力机制的变种方法，例如 Chen 等人[116]通过多通路注意力掩码机制实现多分支特征图的融合操作；PSPNet[117]通过金字塔空间池化操作对图像中的上下文信息进行增强；Zhao 等人[113]提出一种基于空间注意力引导的网络 PSANet 实现时空上下文信息建模。本章所提出的 TSA-Net 框架与以上算法思路一致，均是通过选择性注意力机制在规避密集运算的同时提高特征增强效果。

2.3 基于管道自注意力机制的行为质量评估算法

2.3.1 TSA-Net 网络框架

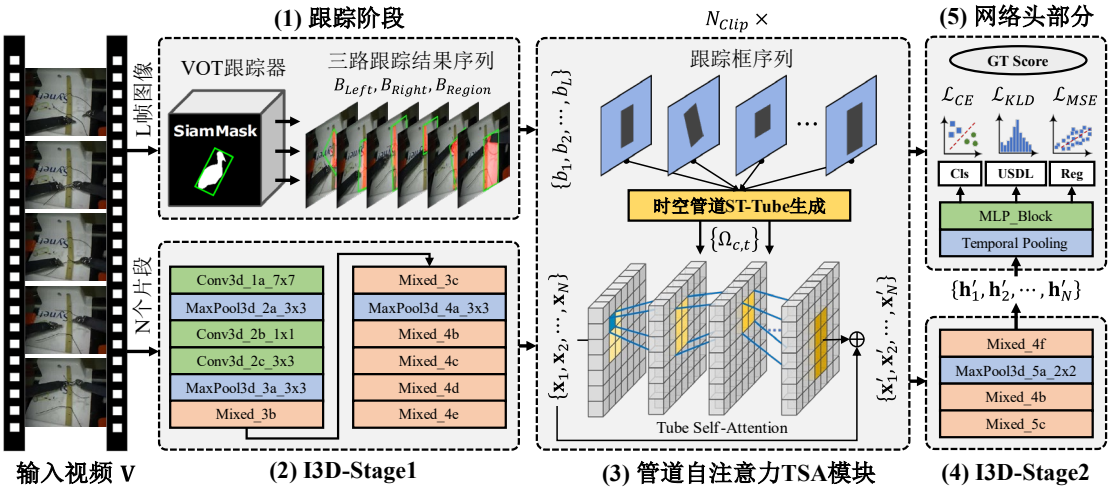


图 2-2 TSA-Net 网络结构图

本文所提出的 TSA-Net 网络结构如图 2-2 所示。给定输入视频$\mathbf{V}=\{F_l\}_{l=1}^{L}$，其中$L$为视频总帧数，$F_l\in\mathbb{R}^{H\times W\times 3}$为视频的第$l$帧图像。首先使用 SiamMask[94]单目标跟踪器对视频中的目标物体进行跟踪，生成跟踪框（Bounding Boxes）集合$B=\{b_l\}_{l=1}^{L}$，其中$b_l=\{(x_p^l,y_p^l)\}_{p=1}^{4}$表示第$l$帧图像中的物体跟踪框，$p$表示跟踪框的四个顶点的编号。图 2-6 以 JIGSAWS 数据集中的缝合视频为案例进行了跟踪结果展示。在 JIGSAWS 数据集中，本文分别对左机械手、右机械手和操作区域共三部分进行了跟踪，分别生成三组跟踪结果序列：$\{B_{Left},B_{Right},B_{Region}\}$。

在特征提取阶段中，将视频$\mathbf{V}$划分为N个视频片段（Clips），其中每个片段含有M张连续的图像。N个视频片段被送入 I3D[7]视频网络的第一阶段（I3D-Stage1）中完成初步的视频特征提取。最终生成视频特征$\mathbf{X}=\{\mathbf{x}_n\}_{n=1}^{N}$，$\mathbf{x}_n\in\mathbb{R}^{T\times H\times W\times C}$。由于视频片段特征$\mathbf{x}_n$的时序维度大小为$T$，$\mathbf{x}_n$也可表示为$\mathbf{x}_n=\{\mathbf{x}_{n,t}\}_{t=1}^{T},\mathbf{x}_{n,t}\in\mathbb{R}^{H\times W\times C}$。

在特征增强阶段，本文所提出的时序管道自注意力模块（TSA）依据跟踪框集合B和视频特征$\mathbf{X}$完成特征增强，最终生成带有丰富时空上下文信息的视频特征$\mathbf{X}'=\{\mathbf{x}'_n\}_{n=1}^{N}$。由于 TSA 模块的嵌入不会影响特征的维度，因此$\mathbf{x}_n$与$\mathbf{x}'_n$具有相同的尺寸，即：$\{\mathbf{x}_n,\mathbf{x}'_n\}\in\mathbb{R}^{T\times H\times W\times C}$。此特性使得 TSA 模块具备即插即用的特性，同时能够使 TSA 模块进行支持多层网络堆叠，从而获取到更优质的视频特征。增强后的视频特征$\mathbf{X}'$会被送入 I3D 视频网络的第二阶段（I3D-Stage2）中完成进一步的特征提取与融合，最终生成全视频特征$\mathbf{H}=\{\mathbf{h}_n\}_{n=1}^{N}$。

在分值预测阶段（神经网络头部分），TSA-Net 使用时序平均池化（Temporal

Average Pooling）操作对$\mathbf{H}$中的N个视频片段特征进行融合，具体的融合过程表示为：

$$\bar{\mathbf{h}}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{h}_n,\ \bar{\mathbf{h}}\in\mathbb{R}^{T\times H\times W\times C} \tag{2.1}$$

融合特征$\bar{\mathbf{h}}$被送入多层感知机（Multilayer Perceptron）网络 MLP_Block 中，最终根据不同的预测任务需求映射为对应形式的输出，例如分类信息、分数回归和得分分布预测。三种预测任务使用相应的损失函数完成预测结果与真实值之间的差异度量。

2.3.2 管道自注意力机制与 TSA 模块

本文所提出的管道自注意力机制与 Non-local[35]机制的核心区别在于：TSA 机制能够依据 SiamMask 单目标跟踪算法生成的跟踪框对视频特征图中的部分元素进行选择性增强；而 Non-local 机制对特征图中所有的元素进行关联建模。TSA 机制所引入的跟踪信息能够有效指导特征增强过程，从而实现抑制视频中背景噪声的功能，促使神经网络关注于视频中的关键行为部分，最终生成更优质的视频特征与更稳定的行为质量评估结果。换言之，可以将提出的管道自注意力机制理解为“Local Non-local”机制，其中第一个 Local 指通过目标跟踪信息生成的时空管道 ST-Tube，第二个 Local 指在 ST-Tube 管道内部对所有的特征元素进行自注意力机制计算。得益于此机制，TSA 模块能够在提升行为质量评估性能的同时大幅降低计算复杂度。

TSA 模块的计算流程可划分为两个阶段：时空管道生成阶段（Spatio-temporal Tube Generation）和管道自注意力机制计算阶段（Tube Self-attention Mechanism）。

时空管道生成阶段

理论而言，在获取到跟踪框信息B和视频特征图$\mathbf{X}$之后即可直接根据跟踪框信息在特征图中进行元素框选。然而由于 I3D-Stage1 中的神经网络含有两个时序池化层，跟踪框序列和特征图之间的对应关系并不是“一对一”而是“多对一”。除此之外，SiamMask 模型生成的跟踪框是倾斜的，对应到特征图上会发生偏移（Misalignment）问题。

为解决以上问题，本文提出了一种跟踪框序列与视频特征图的对齐策略，具体过程如图 2-3 所示。I3D-Stage1 共含有两个时序池化层，因此跟踪框与特征图之间的数量对应关系为 4:1，例如四个跟踪框$\{b_l,b_{l+1},b_{l+2},b_{l+3}\}$与$\mathbf{x}_{c,t}$特征图相对应。在转化过程中，首先将四个坐标形式表示的跟踪框离散化为 0-1 掩码矩阵（0-1 Mask Matrix）。以第l帧的跟踪框b_l为例，将b_l对应到特征图尺寸上生成掩码$M_{c,t}^{l}\in\{0,1\}^{H\times W}$。在此矩阵中，$H\times W$个元素的计算方法为：

$$M_{c,t}^{l}(i,j)=\begin{cases}1, & S(b_l,(i,j))\geq\tau\\ 0, & S(b_l,(i,j))<\tau\end{cases} \tag{2.2}$$

其中$S(\cdot,\cdot)$函数能够对特征网格被跟踪框覆盖的比例进行计算。若覆盖比例高于τ则(i,j)位置的元素被选中；若覆盖比例低于τ则(i,j)位置的元素被排除。本文对所有的 TSA-Net 模型均采取$\tau=0.5$的设定。

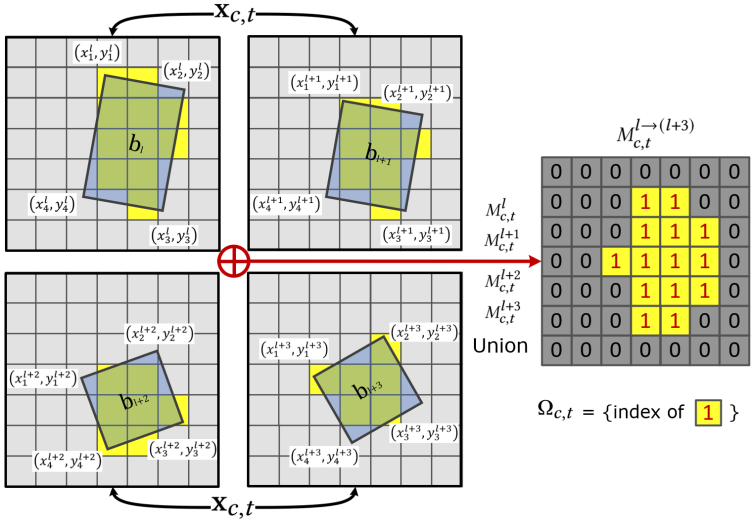


图 2-3 时空管道 ST-Tube 生成示意图

通过对四个跟踪框$\{b_l,b_{l+1},b_{l+2},b_{l+3}\}$分别进行掩码生成，可以获取到四组掩码矩阵$\{M_{c,t}^{l},M_{c,t}^{l+1},M_{c,t}^{l+2},M_{c,t}^{l+3}\}$。四个矩阵通过逐元素或操作（Element-wise OR Operation）合并为一个掩码矩阵$M_{c,t}^{l\to(l+3)}\in\{0,1\}^{H\times W}$：

$$M_{c,t}^{l\to(l+3)}=\text{Union}(M_{c,t}^{l},M_{c,t}^{l+1},M_{c,t}^{l+2},M_{c,t}^{l+3}) \tag{2.3}$$

在体育行为质量评估基准中，每个视频只含有单个运动员，因此只需使用上述公式进行掩码聚合即可。而在 JIGSAWS 医疗行为质量评估基准中，本文依据三个通路的追踪框信息$\{B_{Left},B_{Right},B_{Region}\}$进行时空管道生成，具体实现细节如图 2-8 所示。三个通路生成的掩码矩阵还需要进行一次额外的聚合操作，具体计算方法为：

$$M_{c,t}^{l\to(l+3)}=\text{Union}(M_{c,t}^{Left},M_{c,t}^{Right},M_{c,t}^{Region}) \tag{2.4}$$

此掩码指示矩阵表示了时空管道，即参与自注意力机制计算的所有特征元素位置信息。为后续描述清晰，本文将$M_{c,t}^{l\to(l+3)}$转化为位置信息集合$\Omega_{c,t}$的形式：

$$\Omega_{c,t}=\{(i,j)|M_{c,t}^{l\to(l+3)}(i,j)=1\} \tag{2.5}$$

其中$\Omega_{c,t}$表示时空管道中所有参与自注意力机制的元素位置集合，$|\Omega_{c,t}|$表示特征图$\mathbf{x}_{c,t}$中被选中的元素数量。

管道自注意力机制计算阶段

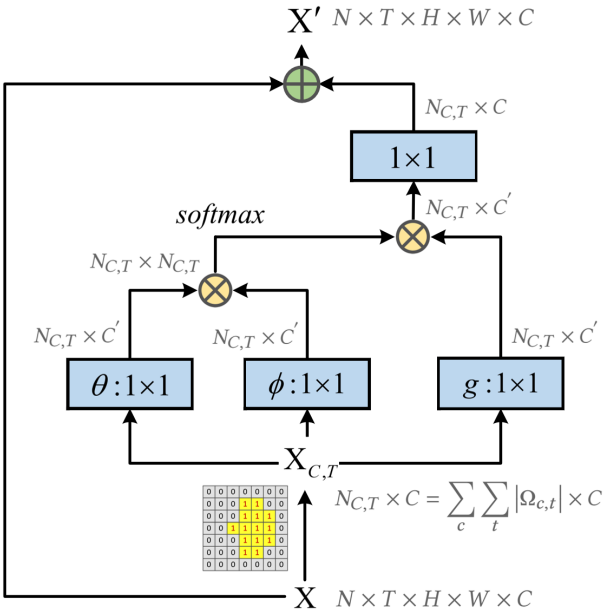


图 2-4　管道自注意力机制计算过程

获取到视频特征$\mathbf{X}$与特征筛选信息$\Omega_{c,t}$后，本文通过自注意力机制对时空管道内的特征元素进行自注意力机制计算，具体计算过程为：

$$\mathbf{y}_p = \frac{1}{C(\mathbf{x})} \sum_{\forall c} \sum_{\forall t} \sum_{\forall (i,j) \in \Omega_{c,t}} f(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)) g(\mathbf{x}_{c,t}(i,j)) \tag{2.6}$$

其中p表示输出特征图中的位置，通过下标元组(c,t,i,j)实现时空管道内所有特征元素的枚举。输出特征$\mathbf{y}$与输入特征$\mathbf{x}$具有相同的特征尺寸。与 Non-local 机制的计算过程类似，$f(\cdot,\cdot)$函数为自注意力机制中的相似度度量函数（Pairwise Function），$g(\cdot)$函数表示单映射函数（Unary Function）。神经网络输出的响应值通过$C(\mathbf{x})$进行归一化，$C(\mathbf{x})$的计算方式为：

$$C(\mathbf{x}) = \sum_{\forall c} \sum_{\forall t} |\Omega_{c,t}| \tag{2.7}$$

为降低计算复杂度，本文在相似度计算函数$f(\cdot,\cdot)$中添加了特征图通道变换操作，具体实现过程为：

$$f(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)) = \theta(\mathbf{x}_p)^T \phi(\mathbf{x}_{c,t}(i,j)) \tag{2.8}$$

其中$\theta(\cdot)$和$\phi(\cdot)$函数均为通道数量减少的1×1卷积层，满足$C' < C$，从而实现网络模型中的特征维度瓶颈设计。最终通过残差连接实现特征映射：

$$\mathbf{x}'_p = \mathbf{W}_z \mathbf{y}_p + \mathbf{x}_p \tag{2.9}$$

其中$\mathbf{W}_z\mathbf{y}_p$表示特征$\mathbf{y}_p$的嵌入映射特征，$\mathbf{x}'_p$与$\mathbf{x}_p$的特征尺寸保持一致。因此 TSA 模块能够被嵌入在任何基于深度神经网络的框架中。为实现计算复杂度和性能之间的平衡，本文在 I3D 网络中的 Mixed_4e 模块后插入 TSA 模块，即$T=4$，$H=W=14$。

医疗行为质量评估中的跟踪机制

在体育场景下的行为质量评估数据集中，视频中的目标较为明显。如图 2-5 所示，通过给定首帧中的目标位置，SiamMask 跟踪器能够为 AQA-7 和 MTL-AQA 数据集中的视频自动生成系列跟踪框。而医疗场景中的情况较为复杂，以 JIGSAWS 数据集为例，所有的视频均是通过 *da Vinci* 手术机器人内置的摄像头录制得到，并没有被试者的肢体运动信息。为妥善处理两个任务之间的差异，本文针对 JIGSAWS 数据集进行了单目标跟踪框生成机制的改进：由于 *da Vinci* 手术机器人的执行部分由两个机械手组成，因此本文在跟踪框生成过程中同时提供了三个初始位置，分别为：左机械手、右机械手与操作区域。SiamMask 跟踪器在 JIGSAWS 三个子任务上的跟踪结果如图 2-6、图 2-7 和图 2-8 所示。在获取到三个独立的跟踪框序列$\{B_{Left}, B_{Right}, B_{Region}\}$之后，本文使用 Union 操作对这些区域进行合并，生成最终的时空管道。详细的合并流程如图 2-9 所示。

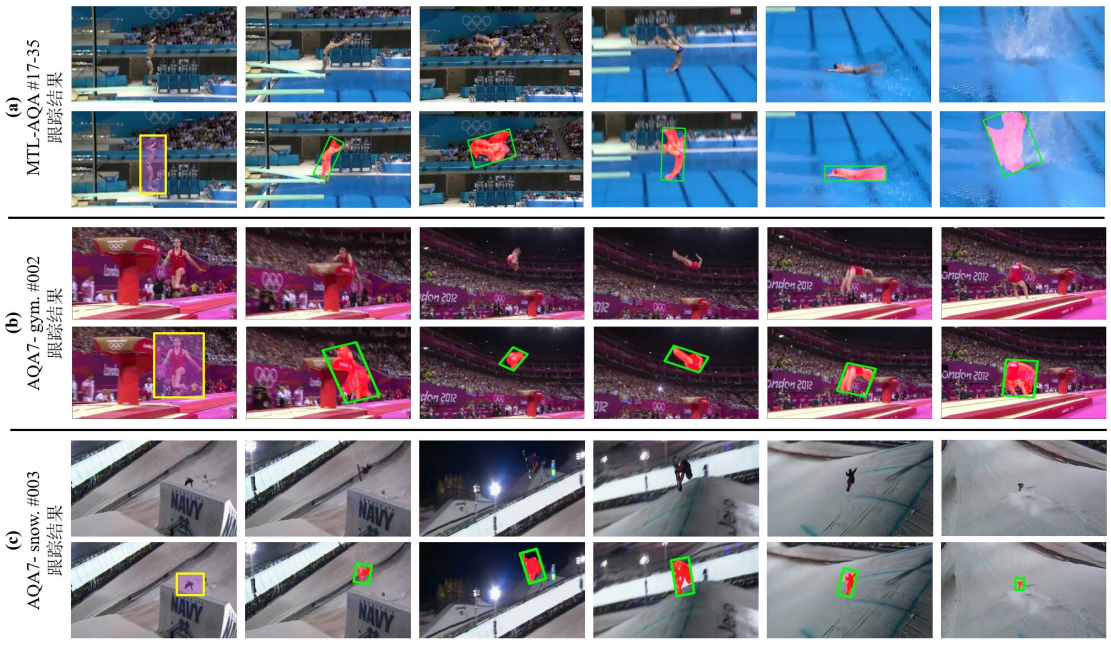


图 2-5 体育场景下的单目标跟踪结果

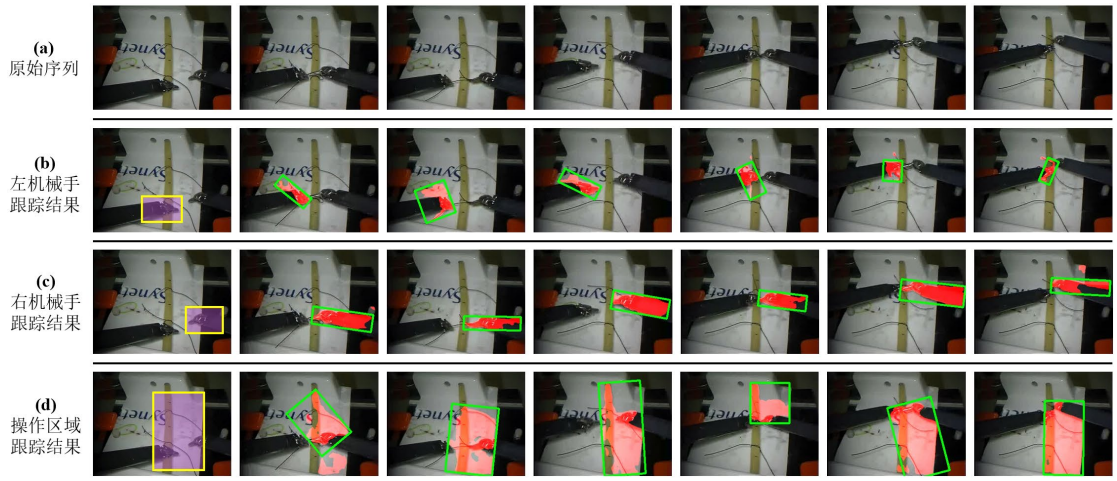


图 2-6　JIGSAWS 数据集中 Knot Tying 案例多重目标跟踪结果

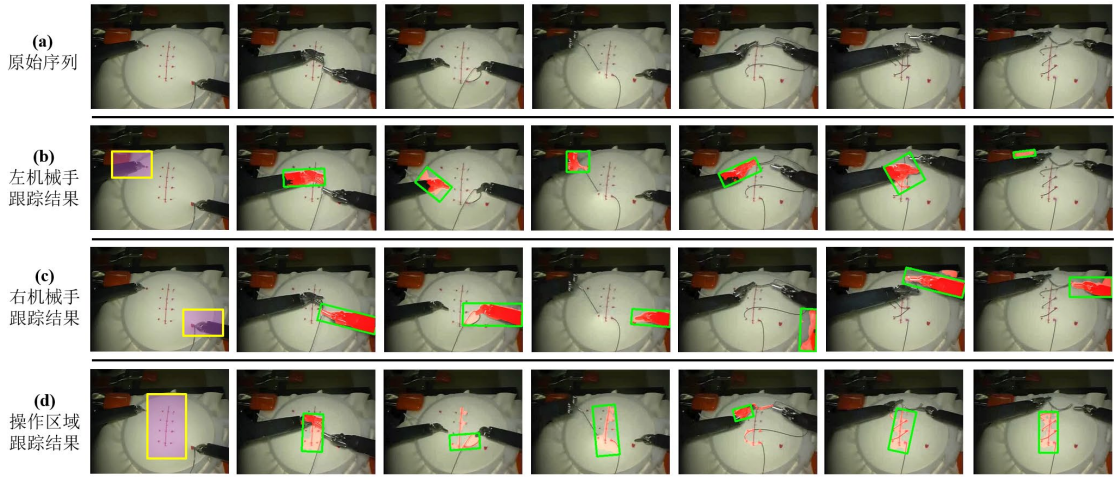


图 2-7　JIGSAWS 数据集中 Needle Passing 案例多重目标跟踪结果

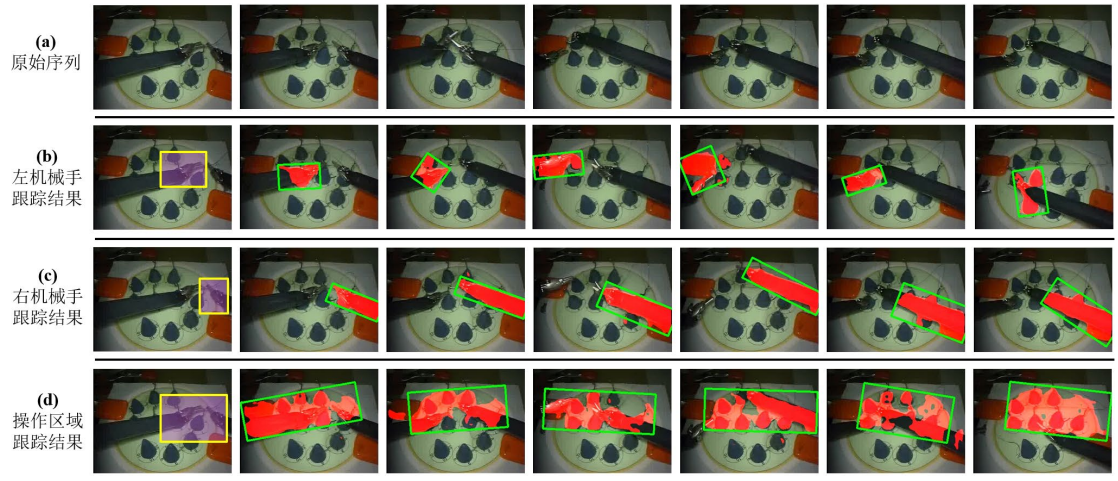


图 2-8　JIGSAWS 数据集中 Suturing 案例多重目标跟踪结果

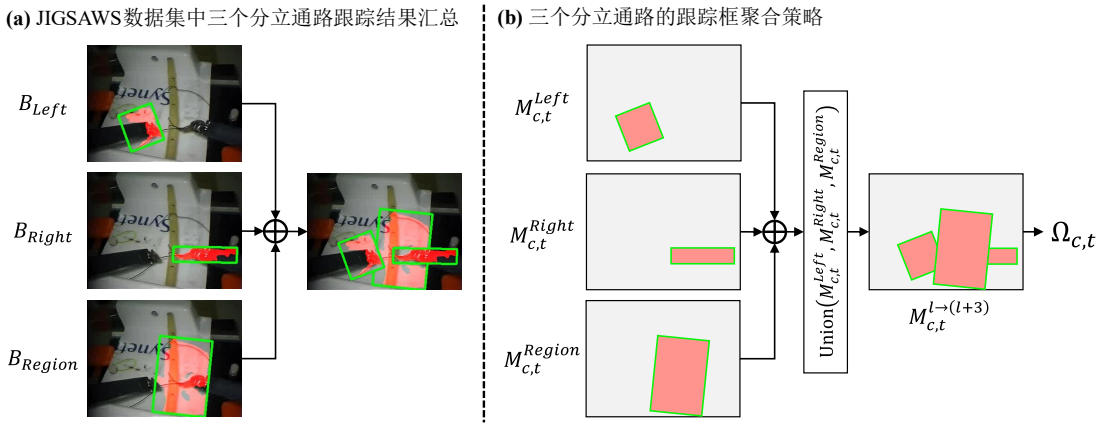


图 2-9 JIGSAWS 数据集中的三区域跟踪框融合策略

2.3.3 网络头设计与损失函数

为充分验证管道自注意力机制的有效性，本文将 TSA-Net 模型拓展为支持多种行为评估任务的框架，包括分类任务、回归任务和分布预测任务。不同的评估模型可以通过修改 MLP_Block 的输出特征图尺寸和损失函数进行切换。三种任务的具体形式如下：

分类任务：行为质量评估任务中的技能评级本质上是对视频进行直接分类，在网络设计方面，只需要将 MLP_Block 的输出特征维度进行修改即可。本文使用交叉熵损失函数对最终类别进行预测。设网络对 M 个类别的预测得分为 $S \in \mathbb{R}^M$，真实类别的独热编码为 $Y \in \{0,1\}^M$，交叉熵损失函数的计算方法为：

$$\mathcal{L}_{CE} = -\sum_{i=1}^{M} Y_i \cdot \log S_i \tag{2.10}$$

回归任务：行为质量评估任务中一些数据集直接对行为进行分数预测，例如体育场景的数据集和遵循 OSATS[57]评估体系的医疗技能评估数据集。在网络设计方面，只需要将 MLP_Block 的输出特征维度设定为 1 即可。在一个含有 N 个样本的批次内，设网络对视频的最终预测分数为 $S \in \mathbb{R}$，真实分值为 $Y \in \mathbb{R}$，均方误差损失（Mean Square Error, MSE）的计算方法为：

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{n=1}^{N}(Y_n - S_n)^2 \tag{2.11}$$

分布预测任务：Tang 等人[118]提出的 USDL 框架支持多个通路同时进行分数的分布预测（即 MUSDL 模型）。虽然这种多通路的方法相较于 USDL 框架[118]取得了优异的性能，但是也引入了成倍的计算量。本文所提出的 TSA 模块能在节省计算量的同时实现特征图中时空信息的特征增强。为验证管道自注意力机制的有效性，本文将 TSA 模块嵌入到 USDL 模型中并进行了性能测试。USDL 框架

中模型输出结果是对每个分数预测的概率，通过 KL 散度（Kullback–Leibler Divergence）对两个概率之间的差异进行度量：

$$KL[P_c \| S_{pre}] = \sum_{i=1}^{M} P(c_i) \log \frac{P(c_i)}{S_{pre}(c_i)} \tag{2.12}$$

其中S_{pre}是 MLP_Block 网络生成的分数分布预测结果，P_c是根据标准标签生成的分数分布。在带有难度系数（Difficulty Degree, DD）的 MTL-AQA 数据集中，预测分数结果需要通过如下过程进行转化：

$$S = S_{DD} \cdot S_{pre} \tag{2.13}$$

2.3.4 计算复杂度分析

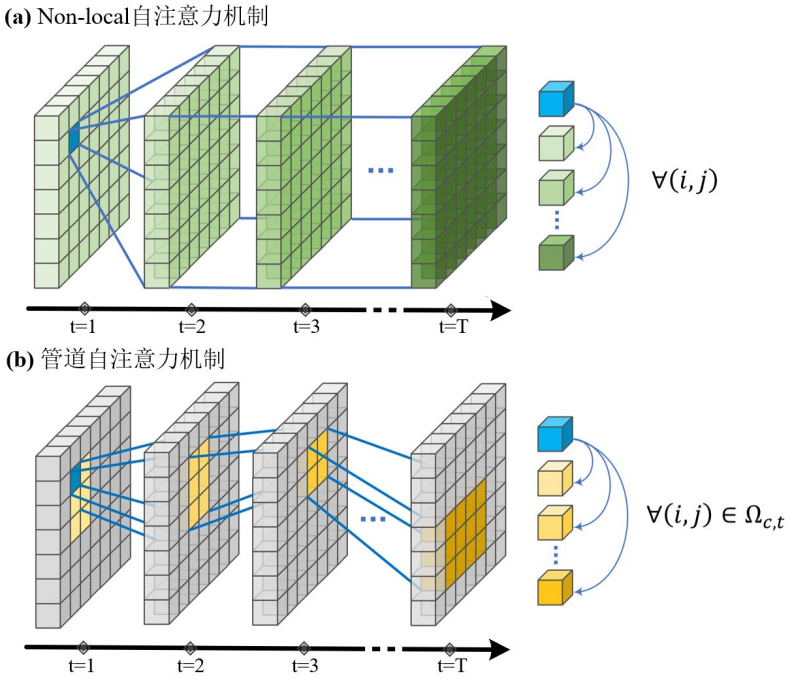


图 2-10　Non-local 与 TSA 机制计算复杂度对比示意图

图 2-10 对 Non-local 机制与本文所提出的管道自注意力机制进行了对比。给定特征图$\mathbf{x}_n \in \mathbb{R}^{T\times H\times W\times C}$，Non-local 机制将每个位置的特征元素与所有特征元素进行交互计算，计算过程表示为：

$$\mathbf{y}_p = \frac{1}{C(\mathbf{x})} \sum_{\forall c} \sum_{\forall t} \sum_{\forall (i,j)} f(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)) g(\mathbf{x}_{c,t}(i,j)) \tag{2.14}$$

而 TSA 模块在追踪框序列的引导下只对时空管道内部的特征元素进行计算，计算过程为公式(2.6)。因此可以从理论层面对两种特征增强策略进行对比。Non-local 自注意力机制的计算复杂度为：

$$O((N\times T\times H\times W)\times(N\times T\times H\times W)) \tag{2.15}$$

而本文所提出的管道自注意力 TSA 模块的计算复杂度为：

$$O\left(\sum_{c}\sum_{t}|\Omega_{c,t}|\times\sum_{c}\sum_{t}|\Omega_{c,t}|\right) \tag{2.16}$$

TSA 机制依赖于 SiamMask 模型生成的跟踪框序列，因此在生成追踪框后即可完成管道自注意力机制的计算复杂度评估。TSA 模块的计算复杂度与时空管道内包含的特征元素数量高度相关：在时空管道体量较大的样本中，计算复杂度较高；而在时空管道体量较小的样本中，计算复杂度较小。此观察在实验部分中的计算复杂度对比中得到了验证。

2.4 实验分析

2.4.1 数据集与评估指标

本文在 3 个开源数据集上对 TSA-Net 的性能开展了测试：1 个医疗技能评估数据集、2 个体育技能评估数据集。

JIGSAWS[15]数据集：全称为 JHU-ISI 手术技能识别与评估数据集（JHU-ISI Gesture and Skill Assessment Working），是由约翰斯·霍普金斯大学与直觉外科公司（Intuitive Surgical Inc.）共同提出的基于 *da Vinci* 手术机器人的外科操作行为质量评估数据集。该数据集共包含三类外科手术培训中的基础行为：缝合（Suturing）、传针（Needle Passing）和打结（Knot Tying），共含有 206 条操作记录。在数据模态方面，JIGSAWS 数据集提供了操作过程中的动力学传感器数据和双目视觉数据两种模态信息。在数据标注方面，JIGSAWS 数据集创建了 15 种手势标签并对视频进行时序标注，并参照 OSATS[57]评估体系从 6 个维度（组织接触、针线抓握、耗时与运动、操作流畅度、整体表现、结局质量）对操作者的技能水平进行评估，其中每个评估的维度按照操作质量划分为三个等级，分别对应 1 分、3 分和 5 分。最终的操作评分通过对各维度得分求和得到。虽然原始数据集提供了双目视觉信息，考虑到两个视角之间信息差异性较小，本文在测试过程中只使用左摄像头拍摄的视频。在训练集与测试集划分方面，本文采用与 USDL[118]模型一致的设置，进行 4 折交叉验证（Four-fold Cross Validation）。

AQA-7[9]数据集：此数据集包含 7 种奥运会运动项目，分别为：跳水（Diving）、体操（Gym）、滑雪（Skiing）、单板滑雪（Snowboard）、同步跳水 3m（Sync. 3m）、同步跳水 10m（Sync. 10m）和蹦床（Trampoline）。AQA-7 数据集共含有 1,189 条视频，在通用的训练配置中，803 个样本用于训练，303 个样本用于测试。考虑

到蹦床项目具有过长的视频时长，本文在模型性能对比中删除了此子项目，只对其余六项运动进行性能平均与汇总。参与性能对比的 USDL 等模型均剔除了蹦床项目，从而保证了性能结果的可比较性。

MTL-AQA[49]数据集：此数据集共含有 16 种不同的跳水项目视频记录，总视频数量为 1,412。MTL-AQA 数据集为每个视频都提供了详细的评委评分、难度系数和文本形式的直播评论。在通用训练配置中，1,059 条视频用于训练，353 条视频用于测试。在现有的体育行为质量评估数据集中，MTL-AQA 拥有最大的数据集体量。

本文使用斯皮尔曼相关性系数（Spearman's Rank Correlation）对模型生成的结果进行评估，其计算方法为：

$$\rho = \frac{\sum (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum (p_i - \bar{p})^2 \sum (q_i - \bar{q})^2}} \tag{2.17}$$

其中 p 和 q 分别表示标准序列和预测分数的排序标号序列。对于含有多个行为的数据集，通过 Fisher's Z-Value[119]对多个行为进行取平均操作。

数据集中的视频分数均通过如下方法进行归一化：

$$S_{norm} = \frac{S - \min(S)}{\max(S) - \min(S)} \times 100 \tag{2.18}$$

其中 $\max(S)$ 和 $\min(S)$ 分别表示数据集中的最大值和最小值。

2.4.2 模型与优化器设定

本章所有的实验均在 Intel Xeon E5-2698 V4 @ 2.20GHz CPU 与单张 NVIDIA Tesla V100 GPU 计算平台上开展，所有的模型均使用 Pytorch 深度学习框架进行构建与训练。

由于管道自注意力 TSA 模块在计算过程中需要使用目标跟踪信息 B，本文使用现有成熟的 SiamMask[94]单目标跟踪器完成跟踪信息生成。SiamMask 跟踪器需要预先给定首帧中的物体框，本文充分结合测评数据集的特点设计了两类策略实现首帧物体框的获取：对于 AQA-7 与 MTL-AQA 数据集，本文使用在 MS-COCO[120]数据集上预训练得到的 Faster-RCNN[121]目标检测模型完成视频首帧的目标检测任务；对于只录制了 *da Vinci* 手术机器人操作过程的 JIGSAWS 数据集，视频中的目标无法通过普适目标检测器进行识别，因此本文使用手动标注的方法分别对左机械臂、右机械臂和操作区域进行标注，只需给定首帧的物体检测框，SiamMask 跟踪器即可为后续所有图像帧生成跟踪信息。

在单目标跟踪器选型方面，通过观察图 2-5 和图 2-6 可知，SiamMask 跟踪

器已经能够实现稳定的目标检测与跟踪功能，生成的追踪结果满足 TSA-Net 框架的需求。此外，SiamMask 跟踪器具备低延时的优点，其运行速度满足实时要求（25 FPS）。因此本文选取 SiamMask 作为目标跟踪器，并未对其他单目标跟踪器进行对比测试。本文所使用的 SiamMask 跟踪器在 DIVAS[122]视频目标分割基准上进行预训练。

本文使用在 Kinetics-400[24]数据集上经过预训练后的 I3D 视频主干网络作为特征提取器。在训练阶段，使用 Adam[123]优化器对网络参数进行优化，初始学习率设定为 1e-4，动量值设定为 0.9，权重衰减值设定为 1e-5，训练轮次设定为 100。考虑到视频时长与模型计算复杂度的区别，本文对不同数据集采取不同的配置：在 JIGSAWS 数据集中，每个视频被划分为 15 个片段（Clips）；而在 AQA-7 和 MTL-AQA 数据集中，每个视频被划分为 10 个片段。

2.4.3 性能对比实验

为探究管道自注意力机制与非局部注意力机制之间的性能和计算复杂度差异，本文在三个数据集上分别测试了 TSA 模块与 Non-local[35]模块，对应的两种模型分别表示为：TSA-Net 与 NL-Net。两个模型除特征增强模块不同外，其他所有设定均保持一致。

表 2-1 列举了 TSA-Net 模型在 JIGSAWS 数据集上的性能，并与现有最优方法进行了对比。实验结果显示，本文所提出的 TSA-Net 模型在 Suturing 和 Needle Passing 两个子项目中均取得了最优性能，并且在平均指标中取得了最高结果。TSA-Net 模型在三个子项目中的性能均超越了 NL-Net，这充分说明了选择性特征增强机制的有效性。在所有对比的方法中，本文并没有将 MUSDL[118]模型列举在内，这是因为 MUSDL 分别使用 6 个独立的 USDL 模型对 JIGSAWS 中的 6 个评估方面进行了单独预测。MUSDL 本质上属于组合模型，而表 2-1 中列举的模型属于单个模型。

表 2-1　JIGSAWS 数据集中 TSA-Net 性能对比结果

模型	Suturing	Needle Pass.	Knot Tying	Avg. Corr.
ST-GCN[124]	0.31	0.39	0.58	0.43
TSN[4]	0.34	0.23	<u>0.72</u>	0.46
JRG[119]	0.36	0.54	**0.75**	0.57
USDL[118]	0.64	0.63	0.61	0.63
NL-Net	<u>0.65</u>	<u>0.64</u>	0.67	<u>0.65</u>
TSA-Net	**0.68**	**0.65**	0.71	**0.67**

表 2-2 列举了 TSA-Net 模型在 AQA-7 数据集各项目中的性能与平均性能。结果显示 TSA-Net 模型取得了 0.8476 的 Avg. Corr.性能，高出当前最优模型 USDL 约 3 个百分点。在所有参与测试的 6 个项目中，只有在 Snowboard 项目中 TSA-

Net 模型的性能要略差于 USDL 模型。这主要是因为滑雪相关项目的视频中目标过小。图 2-13 对 AQA-7 数据集中的 Snowboard #056 序列进行了展示。滑雪项目中远距离的视角导致了运动员在图像中占据的尺寸过小，从而进一步导致 TSA 模块所构建的时空管道体量被压缩，最终影响行为质量评估结果。表 2-2 同时列举了 Non-local 网络的性能。通过对比 TSA-Net 与 NL-Net 模型可发现，TSA-Net 网络的整体性能要优于 NL-Net：在平均指标方面，TSA-Net 较 NL-Net 提高了 0.0016 的 Avg. Corr.性能。但是在 Skiing 项目中 NL-Net 模型优于 TSA-Net 模型，这也是由目标尺寸过小所引起的。以上实验结果表明，TSA-Net 在目标尺寸适中的情况下能够发挥出最优性能，过小的目标尺寸会导致时空管道机制的失灵，进而影响最终的评估结果。需要注意，本文所提出的管道自注意力模块是 Non-local 模块的改进，所以本文将 NL-Net 与 TSA-Net 的性能对比也属于消融实验，因此本章不再单独开辟消融实验小节。

表 2-2　AQA-7 数据集中 TSA-Net 性能对比结果

模型	Diving	Gym	Skiing	Snowboard	Sync. 3m	Sync. 10m	Avg. Corr.
Pose+DCT[96]	0.5300	-	-	-	-	-	-
ST-GCN[124]	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM[119]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR[119]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG[119]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL[118]	0.8099	0.7570	0.6538	**0.7109**	0.9166	0.8878	0.8102
NL-Net	0.8296	0.7938	**0.6698**	0.6856	0.9459	0.9294	0.8418
TSA-Net	**0.8379**	**0.8004**	0.6657	0.6962	**0.9493**	**0.9334**	**0.8476**

表 2-3 列举了 TSA-Net 模型与其他对比方法在 MTL-AQA 数据集中的平均性能。结果显示 TSA-Net 和 NL-Net 均能取得最优性能。在所有对比方法中，MUSDL 模型虽然采用了汇总多个 USDL 模型预测结果的策略，但是由于这些模型的结构同质性较强，最终这种组合策略并不能带来太大的性能提升，最终性能仍然差于 TSA-Net 和 NL-Net。这充分表明了 TSA 特征增强模块的有效性。表 2-3 结果显示，NL-Net 的性能要略高于 TSA-Net。这主要由两部分原因造成：一方面，MTL-AQA 数据集中的视频相较于 AQA-7 有更高的视频分辨率：640×360 和 320×240。MTL-AQA 数据集中的视频具有广阔的视野，从而导致时空管道的体量被压缩，出现了类似 AQA-7 数据集中 Skiing 项目的现象。虽然 TSA-Net 的性能略差于 NL-Net，但在特征增强环节中管道自注意力机制节约了近一半的计算量。综合模型的预测性能和计算复杂度即可发现，本文所提出的管道自注意力机制同时具备高效性和有效性的优点。

表 2-3 MTL-AQA 数据集中 TSA-Net 性能对比结果

模型	Avg. Corr.
Pose+DCT[96]	0.2682
C3D-SVR[119]	0.7716
C3D-LSTM[119]	0.8489
C3D-AVG-STL[49]	0.8960
C3D-AVG-MTL[49]	0.9044
MUSDL[118]	0.9273
NL-Net	**0.9422**
TSA-Net	0.9393

受启发于 Transformer[104]模型中编码器和解码器中的多层堆叠机制，本文进行了多个 TSA 模块堆叠的性能测试。由于 JIGSAWS 数据集的体量较小，网络参数的增加会导致过拟合现象的发生，因此本文只在 AQA-7 和 MTL-AQA 数据集中进行了堆叠性能测试。实验结果汇总在表 2-4 和表 2-5 中。在 AQA-7 数据集中，管道自注意力机制的堆叠数量为 2 时可以取得最优性能，而过多的堆叠数量会导致过拟合现象的发生，从而降低平均指标。可得出结论：管道自注意力模块的数量存在边际效益递减现象。表 2-5 分别列举了 NL-Net、TSA-Net 和多层堆叠网络在 MTL-AQA 数据集中的斯皮尔曼相关性系数（Sp. Corr.）、均方误差（MSE）和计算复杂度对比结果。结果显示，TSA-Net 模型在单个 TSA 模块的设定下能够实现最优的均方误差预测结果；而在相关性系数指标中 TSA-Net 则要略差于 NL-Net。综合对比三个方面的指标即可发现，对 TSA 模块进行双层堆叠可获得整体最优性能，此结论与 AQA-7 数据集中的实验结论保持一致。

表 2-4 AQA-7 数据集中 TSA 模块堆叠实验结果

模型	Diving	Gym	Skiing	Snowboard	Sync. 3m	Sync. 10m	Avg. Corr.
TSA-Net	0.8379	0.8004	0.6657	0.6962	**0.9493**	0.9334	0.8476
TSAx2-Net	0.8380	0.7815	**0.6849**	**0.7254**	0.9483	**0.9423**	**0.8526**
TSAx3-Net	**0.8520**	**0.8014**	0.6437	0.6619	0.9331	0.9249	0.8352

表 2-5 MTL-AQA 数据集中 TSA 模块堆叠实验结果

模型	Sp. Corr.↑	MSE↓	FLOPs↓
NL-Net	**0.9422**	47.83	2.2 G
TSA-Net	0.9393	**37.90**	**1.012** G
TSAx2-Net	0.9412	46.51	2.025 G
TSAx3-Net	0.9403	47.77	3.037 G

2.4.4 计算复杂度对比

表 2-6 汇总了 AQA-7 数据集中各子项目上的 NL-Net 和 TSA-Net 计算量与性能对比结果。最右侧栏中的计算量节省比例与性能增益均以 NL-Net 为基准。在平均性能指标中，管道自注意力机制能够节省 68.7%的计算量，同时带来 0.0058 的相关性系数提升。在所有参与测试的 6 个子项目中，TSA-Net 模型在 Skiing 和

Snowboard 子项目中分别带来了 87.13%和 87.97%的计算量大幅节省。然而时空管道 ST-Tube 尺寸的过度压缩也带来了性能的波动：TSA-Net 在 Snowboard 项目中优于 NL-Net，但是在 Skiing 项目中略差于 NL-Net。表 2-6 只在数据集子集的层次对平均性能进行了探究，为充分观测 TSA 机制在每个数据集子集中带来的计算量节省，本文在图 2-11 中对所有子项目的训练集和测试集案例计算量进行了统计。在所有视频中，由于特征图在网络运行过程中尺寸是固定的，且 Non-local 模块和 TSA 模块所插入的位置也是相同的，所以 NL-Net 模型中特征增强部分的计算复杂度为固定的 2.2 GFLOPs。对比结果显示，管道自注意力能够在所有视频中带来计算量的大幅节省。

表 2-6　TSA-Net 的计算复杂度与性能对比（AQA-7 数据集）

对比项目	NL-Net	TSA-Net	计算量节省	性能提升
Diving	2.2 GFLOPs	0.864 GFLOPs	-60.72%	↑0.0083
Gym	2.2 GFLOPs	0.849 GFLOPs	-61.43%	↑0.0066
Skiing	2.2 GFLOPs	0.283 GFLOPs	-87.13%	↓0.0041
Snowboard	2.2 GFLOPs	0.265 GFLOPs	-87.97%	↑0.0106
Sync. 3m	2.2 GFLOPs	0.952 GFLOPs	-56.74%	↑0.0034
Sync. 10m	2.2 GFLOPs	0.919 GFLOPs	-58.24%	↑0.0040
Average	2.2 GFLOPs	0.689 GFLOPs	-68.70%	↑0.0058

表 2-5 汇总了 NL-Net 和 TSA-Net 模型在 MTL-AQA 数据集上的性能与计算量测试结果。相较于 NL-Net 模型，使用单个管道自注意力机制的 TSA-Net 模型能够降低 46%的计算复杂度以及 9.93 的均方误差。表 2-5 的结果表明，由于过拟合现象的存在，过多的特征增强模块堆叠并不会直接带来相关性系数指标与均方误差性能的提升。以上对比实验结果充分验证了时空管道自注意力机制的高效性与有效性。

表 2-7 汇总了 JIGSAWS 数据集中各子项目上的 NL-Net 和 TSA-Net 网络中特征增强模块计算量与整体性能对比结果。最右侧栏中的计算量节省比例与性能增益计算均以 NL-Net 为基准。结果显示，TSA-Net 能够在节省近 50%计算量的前提下达到和 Non-local 模块相近的性能。由于引入了基于三通路的时空管道构建机制，TSA 模块在 JIGSAWS 数据集中的计算节省比例要少于 MTL-AQA 和 AQA-7 数据集。

表 2-7　TSA-Net 的计算复杂度与性能对比（JIGSAWS 数据集）

对比项目	NL-Net	TSA-Net	计算量节省	性能提升
Suturing	3.52 GFLOPs	1.87 GFLOPs	-46.89%	↑0.03
Needle Pass.	3.52 GFLOPs	1.77 GFLOPs	-49.78%	↑0.01
Knot Tying	3.52 GFLOPs	1.99 GFLOPs	-43.33%	↑0.04
Average	3.52 GFLOPs	1.86 GFLOPs	-46.67%	↑0.02

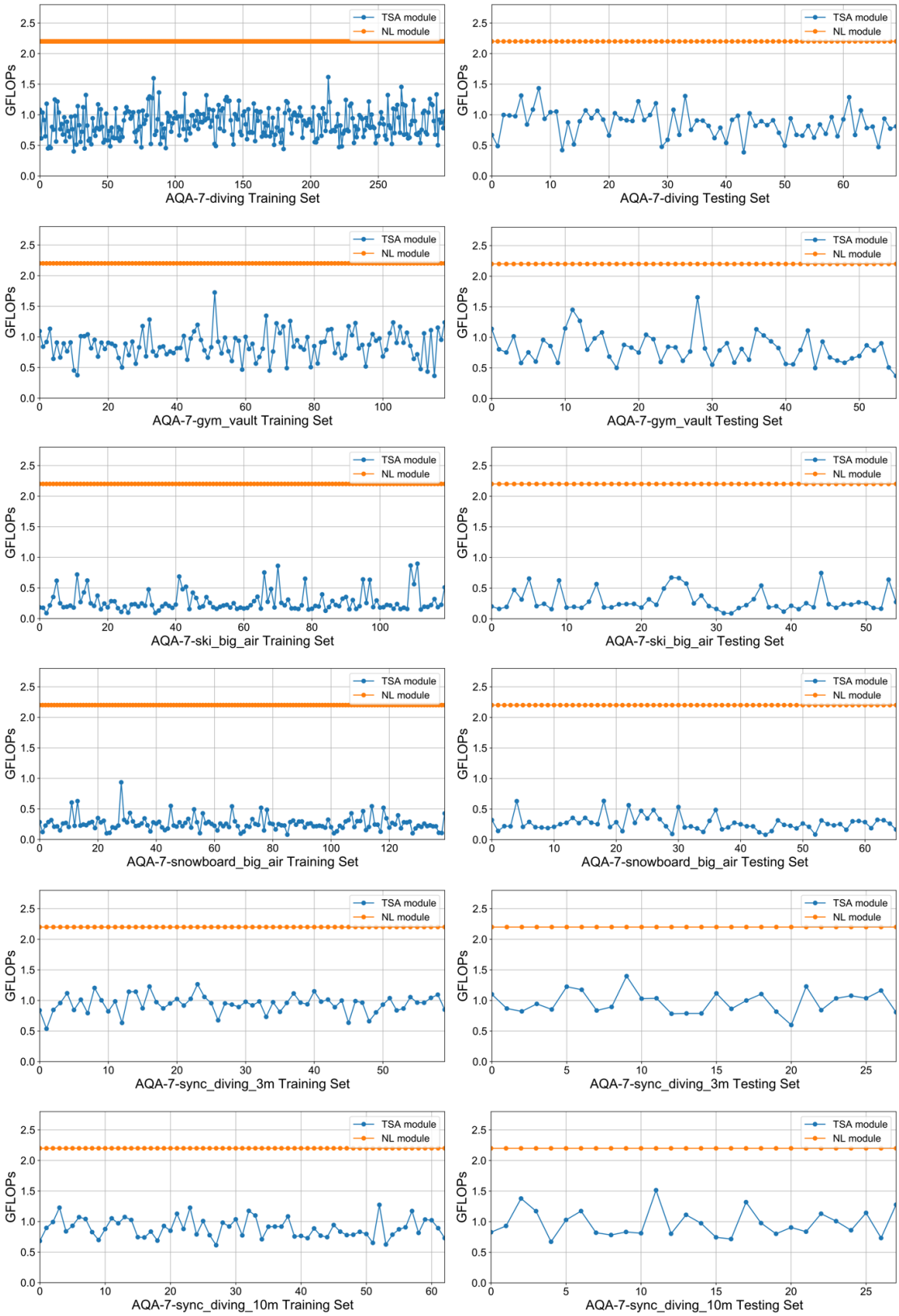


图 2-11 AQA-7 数据集 Non-local 与 TSA 逐样本计算量对比

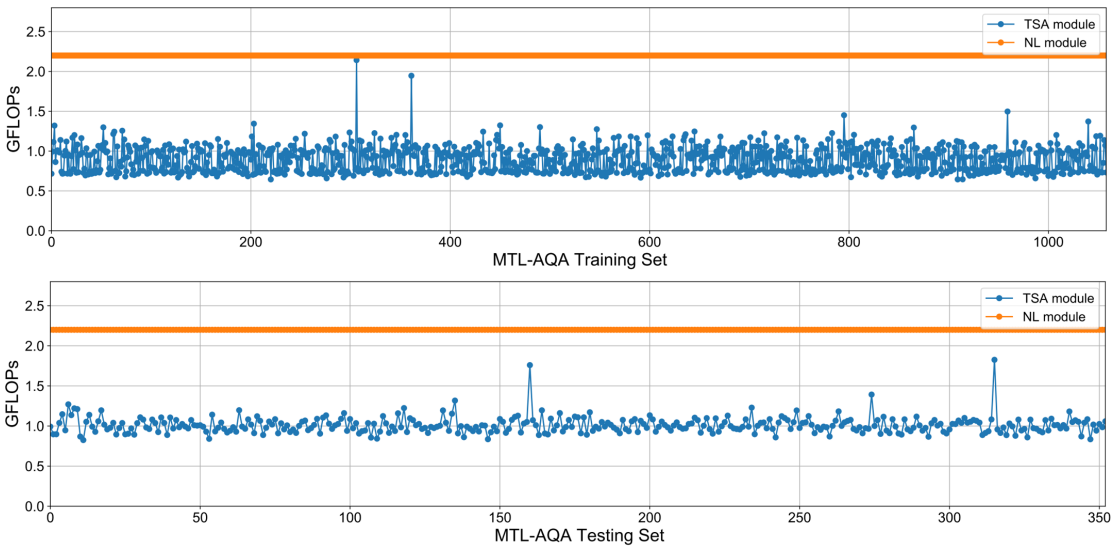


图 2-12　MTL-AQA 数据集 Non-local 与 TSA 逐样本计算量对比

2.4.5 定性对比与可视化结果

图 2-13 展示了 TSA-Net 对 MTL-AQA 和 AQA-7 数据集中部分案例的预测与跟踪结果。观察可发现 SiamMask 跟踪器能够适用于各种体育运动场景，例如单人和双人跳水、体操和滑雪等运动。TSA-Net 能够依据跟踪器生成的追踪框序列对视频特征进行选择性增强，最终实现较高精度的行为质量评估结果。图 2-14 对 SiamMask 跟踪器在 *da Vinci* 手术机器人操作视频中生成的跟踪和预测分数结果进行了展示。这些结果充分说明，本文所提出的管道自注意力机制适用于行为质量评估任务，能够在降低计算复杂度的同时有效提升行为质量评估模型的性能。

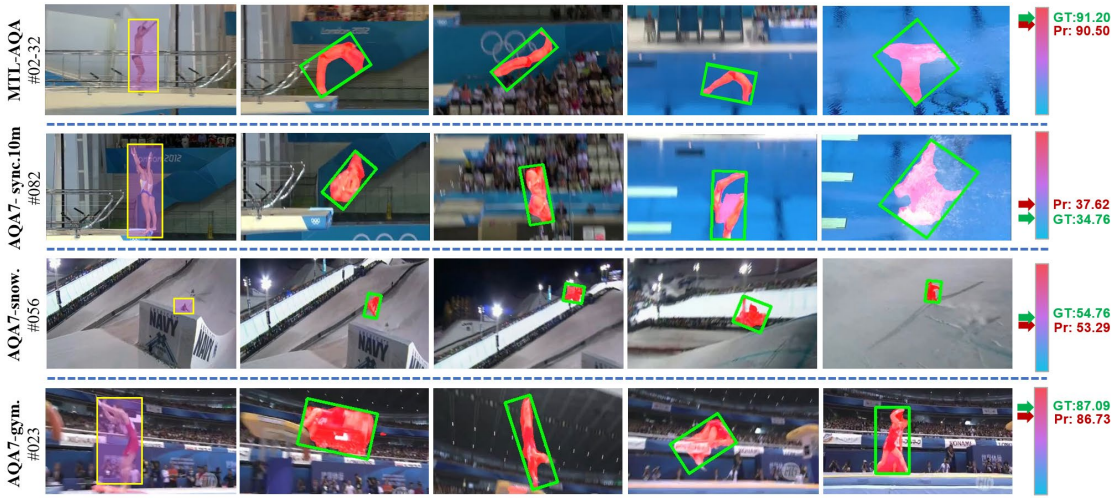


图 2-13　MTL-AQA 与 AQA-7 数据集跟踪与预测结果展示

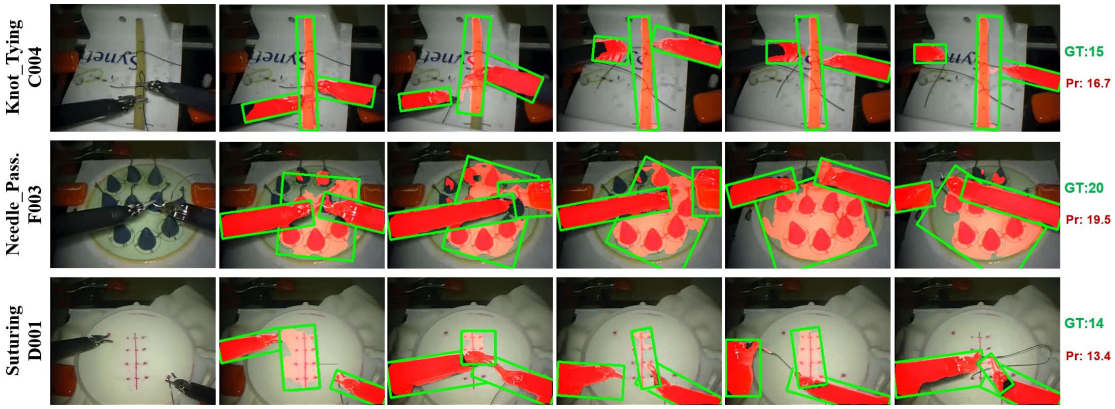


图 2-14 JIGSAWS 数据集跟踪与预测结果展示

2.5 本章小结

本章针对现有医疗技能评估模型视频表征能力弱的问题，提出了一种高效的稀疏特征增强策略：管道自注意力模块 TSA。此模块首先通过单目标跟踪器 SiamMask 生成视频中的目标跟踪框序列，并将序列与特征图进行对齐构建时空管道 ST-Tube。之后，TSA 模块通过对时空管道内部的特征进行自注意力操作完成视频特征的增强。本章将 TSA 模块与 I3D 视频主干网络相结合提出了 TSA-Net。在实验部分，本章在医疗技能评估领域中的 JIGSAWS 数据集和体育技能评估领域中的 AQA-7、MTL-AQA 数据集上进行了算法的性能和计算量对比。充分的实验结果证实了 TSA-Net 的通用性、有效性和高效性。

第3章 基于特征组合机制的复合错误行为识别算法

3.1 引言

第二章对医疗和体育场景中的行为质量评估算法进行了探究。行为质量评估任务的形式较为直接：对视频进行特征提取和分数预测。此类任务形式只适用于简单的医疗技能评估场景，而当技能评估的需求变得复杂时，例如错误医疗行为识别、多种错误相互组合下的识别，以往算法往往无法直接使用。针对此问题，本章将心肺复苏（Cardiopulmonary Resuscitation, CPR）中的胸外按压行为设定为研究对象，首次提出了复合错误行为识别任务（Composite Errors Recognition）：在训练集中只有单类错误行为，而测试集中含有复合错误行为。之后，本章构建了能够同时支持细粒度错误行为识别和复合错误识别的 CPR-Coach 数据集。为解决复合错误行为识别问题，本章提出了基于特征组合训练机制的 ImagineNet 框架本章研究内容隶属于细粒度医疗行为识别研究领域，为后续各种医疗场景下的复合错误行为识别研究奠定了重要基础。

CPR 是一项重要且基本的急救技能，其目的是在患者发生心搏骤停（Cardiac Arrest, CA）时，通过及时的胸外按压与人工呼吸操作在短期内维持患者的血液循环系统与呼吸系统功能，从而为专业抢救争取宝贵时间。不正确或不恰当的 CPR 行为不但会导致无效心肺复苏，甚至会对患者造成严重的二次伤害，例如：按压位置错误、按压频率过慢等错误行为会削弱心肺复苏的预期效果，按压过重、跳跃按压等错误行为会导致患者肋骨骨折。传统的心肺复苏技能考核通常采用“资深医师+假人”的组合方式，其中医师对施救者的肢体动作进行评分，假人胸膛内部安装的力传感器对施救者的按压频率与深度进行精准测量。这种组合式的考核方案人力物力成本较高，很难推广应用到大规模的 CPR 技能训练与考核中。一种理想的评估方案是：无需医师参与，通过智能化的视觉系统替代医师对施救者的行为进行纠错与评估，从而大幅度提升医疗技能的培训考核效率。为实现此方案，本章设计并搭建了一套多视角心肺复苏行为采集系统，并进行了复合错误行为识别数据集的构建。

基于视频的人体行为识别是计算机视觉与视频理解中的代表性任务。得益于样本数据的易获取性，现有行为识别数据集通常关注于日常生活与体育场景，例如 ActivityNet[25]，Kinetics-400[24]，Sports-1M[8]，YouTube-8M[125]，NTU_RGB+D[29]等。基于以上数据集所构建的算法能够为每个视频指派行为标签，例如 TSN[4]，TSM[126]，I3D[7]，C3D[34]，SlowFast[127]等模型；医学领域中的行为识别相关研究

主要聚焦于手术流程识别，即对未剪辑的全程手术视频进行时序分割，例如Cholec80[71]， CholecT50[20]， Hei-Chole[128]， MISAW[14]， Cataract-101[21]，CATARACTS[73]等。手术流程识别算法的功能是为视频中的每一帧进行手术阶段标签预测，例如 MS-TCN++[82]，ASFormer[87]，DiffAct[92]等。

无论是视频级别（Video-level）还是帧级别（Frame-level）的行为种类预测，这些任务均隶属于一般的多分类任务，既没有对行为的正确性进行判断，也没有考虑错误行为之间错综复杂的组合情况。医疗场景中的细粒度错误行为识别任务目前面临着两个方面的困难：一方面，错误行为和复合错误行为标签空间的设计难度较大，需要专业医生的参与和指导；另一方面，错误行为案例的获取难度较大，需要进行专门的数据采集。在实际的医疗技能评估系统构建过程中，考官医师的很大一部分精力放在操作纠错过程中，因此细粒度错误行为识别能力在医疗技能评估系统中具有重要作用。

为填补细粒度医疗错误行为辨识的研究空白，本章对心肺复苏术中的胸外按压行为进行了深入探究，在专业医师的指导下构建了错误行为标签空间，即 13 类单错误行为和 74 类复合错误行为。之后，本章设计并搭建了一套多视角胸外按压行为视觉采集系统，并构建了同时支持单类错误行为识别和复合错误识别的 CPR-Coach 数据集。为缓解真实应用场景中的“单类训练，多类测试”问题，本章提出了基于特征组合训练机制的 ImagineNet 框架，并通过三种不同的网络结构提出了框架的实例化方案。在 CPR-Coach 数据集上开展的实验充分证实，ImagineNet 框架能够有效缓解训练集和测试集之间存在的数据分布差异过大问题，从而帮助现有行为识别框架提升复合错误行为识别精度。本章研究有望为细粒度医疗技能评估领域带来新的启发。

本章的主要贡献概括如下：

1、本章首次提出了复合错误行为识别这一任务范式，并对心肺复苏术 CPR 中的胸外按压行为进行了深入探究，构建了能够同时支持单类错误行为识别和复合错误识别的 CPR-Coach 数据集。

2、本章针对现实医疗技能评估任务所遇到的“单类训练，多类测试”问题，提出了一种基于特征组合训练机制的 ImagineNet 框架。该框架能够有效提升传统行为识别模型在复合错误识别任务中的性能。

3、在实验部分，本章充分探究了基于各种模态的主流视频网络在 CPR-Coach 数据集中的性能。实验结果证实了 ImagineNet 框架的有效性。

3.2 人体行为识别任务与多标签分类问题

3.2.1 人体行为识别任务

人体行为识别技术（Human Action Recognition, HAR）旨在让计算机理解人体的动作和行为。在 1.2.1 小节中，本文对各种模态下的人体行为识别研究进行了汇总与介绍。本章重点关注基于视觉模态信息的人体行为识别任务。

目前学界中的行为识别数据集通常关注于人类日常生活场景，例如关注于运动场景的 Sports-1M[8]、Olympic[129]；关注于日常行为的 NTU_RGB+D[29]、UCF-101[130]、HMDB51[131]；通过大型互联网视频平台收集得到的 ActivityNet[25]、Kinetics-400[24]、YouTube-8M[125]等。尽管这些数据集中最大体量已达数百万条视频，但是行为标签划分粒度较粗，只能够满足理论层面的算法研究，无法满足真正的实际需求。针对行为划分粒度粗的问题，Shao 等人[47]和 Xu 等人[48]分别构建了 FineGym 和 FineDiving 数据集，分别探究体操运动和跳水运动中的高细粒度行为识别任务。然而，目前在医疗领域中尚未有研究对行为进行高粒度划分。

随着深度神经网络的快速发展，学界中已经涌现出了大批人体行为识别算法。这些算法按照网络结构可划分为四种类别：基于双流网络的 Two-Stream 模型[4,8,31]、基于循环神经网络 RNN 的模型[18,32,33]、基于 3D 卷积网络的模型[7,34,35]和基于 Transformer 框架的模型[36–38]。早期研究中，Simonyan 等人[31]首次提出了同时使用光流和 RGB 信息用于行为识别的方案：光流信息能够捕获帧间的人体运动信息，而 RGB 信息则能够提供人体运动的纹理信息，这种方法被形象地命名为 Two-Stream 网络。后续对此类双流网络的研究关注于如何高效地获取光流信息和如何有效地实现两路信息融合等问题；基于循环神经网络的模型[18,32,33]充分参考了自然语言处理领域中的模型结构，将 RNN、LSTM 等网络与行为识别任务进行了结合。由于循环神经网络训练过程具有效率低、不稳定等缺点，此类方法并称为行为识别算法中的主流方向；基于 3D 卷积网络的模型[7,34,35]将视频帧的特征集合视为一个整体，通过 C3D[34]等主干网络对视频进行密集特征提取，从而实现视频特征的获取。由于密集特征提取过程的存在，此类方法具有计算量庞大的缺点；随着 Transformer[104]模型在自然语言处理领域中的各任务取得优异的性能，越来越多的研究者将 Transformer 模型引入到了行为识别领域中，充分利用自注意力机制的运算高效、训练稳定等优点提出了系列行为识别算法[36–38]。虽然以上算法的有效性在公开数据集中得到了验证，但尚未有研究探究过算法对细粒度医疗行为的识别性能。为填补此研究空白，本章对心肺复苏中的胸外按压行为进行了探究，构建了细粒度医疗行为识别数据集 CPR-Coach，并且提出了复合错误行为识别算法。

3.2.2 多标签分类问题与算法

多分类任务（Multi-class Task）假设每个样本只属于多个类别中的一项，而真实世界中更多的问题属于多标签分类任务（Multi-label Task），例如一张高分辨率图像中含有多个类别的物体、一段视频中发生多个行为和事件。目前学界已对经典的多标签分类算法进行了探究[132,133]，并且应用在了计算机视觉[134]、自然语言处理[135]等任务中。多标签分类任务所面临的最大挑战在于标注环节：构建多标签数据集需要对每个样本的所有标签进行详细标注，整个过程耗时耗力且极易发生标签丢失问题。不完整的标注信息会导致模型在预测过程中产生较多的错误负例（False Negatives）。

目前学界中已有一些研究对多标签分类问题的特殊设定进行了探究，这些研究通常关注于减少标注信息或充分使用单样本数据集。Cole 等人[136]针对多标签分类任务中数据集构建成本过高的问题，对“单标签标注—多标签预测”（Single Positive Multi-label Learning）任务进行了探究。在此设定下，训练集中的每个多标签样本仅有单个阳性标签标注信息。若模型能够在这类严峻的监督信息遗漏情况中取得稳定的多标签分类性能，则能大幅降低多标签分类数据集的构建成本。Cole 等人[136]对现有的多标签分类损失进行了改进，实现了损失对单阳性标签标注的适配，并且在四个公开的多标签分类数据集上验证了有效性。虽然此任务的形式与本文所探究的复合错误行为识别任务相类似，但是仍然存在较大的差异：即“训练集—测试集”之间的样本分布差异，Cole 等人[136]所探究的问题中训练集是由多标签样本构成的，本质上是标注信息丢失的设定。而本文所探究的复合错误行为识别任务中训练集—测试集之间的样本分布完全不同，训练集只含有单类样本，而测试集含有多类复合样本。

在生物医学图像分割领域中，Dmitriev 等人[137]针对多类别标注信息匮乏和医学图像数据集标注成本过高的问题，首次提出了一种统一的多类别组合分割框架。该框架能够对现有的多个私有单类别分割数据集进行组合，从而让模型适配多类别分割任务。就问题形式而言，此任务与本章所探究的复合错误行为识别任务有类似之处。具体而言，Dmitriev 等人[137]提出的框架将对抗神经网络构建中（Generative Adversarial Nets, GANs）常用的条件输入方法（Conditions）引入到多类别分割任务中，通过向卷积神经网络的不同阶段引入标签嵌入作为条件信息的模型设计，最终实现多种分割任务的合并与同时学习。在实验部分，此组合分割框架应用在肝脏、胰腺和脾脏的组合分割任务中。

本文结合医疗行为分析场景中的实际问题，对“单类训练—多类测试”问题进行了形式的定义并提出了复合错误行为识别任务，构建了心肺复苏场景中的复合错误行为识别数据集。

3.3 CPR-Coach 数据集构建

现有大部分医疗技能评估数据集只对操作的评分与定级任务进行了探究，并未考虑到被试者犯操作错误的情况。而在真实的医疗操作技能评估中，错误行为的辨识问题往往占据更重要的地位。在医疗场景下，错误行为识别任务面临着两方面问题：一方面，这些错误操作往往具有较高的隐蔽性和较低的辨识度；另一方面，当多种错误相互纠缠组合时，辨识任务就变得更加困难。现有医疗行为识别数据集均无法支持错误行为的辨识任务。

为填补医疗场景下错误行为识别研究的空白，本节构建了首个心肺复苏错误行为识别数据集 CPR-Coach。本节内容的组织情况如下：3.3.1 小节针对目前尚无胸外按压错误行为的明确定义，在收集大量资料的基础上和专业医生的指导下完成了心肺复苏错误标签空间的定义；3.3.2 小节详细描述了多视角采集系统的构建过程；3.3.3 小节描述了 CPR-Coach 数据集的采集与构建过程和适用任务；3.3.4 小节分别对单类错误行为识别和复合错误行为识别的评估指标进行介绍。

3.3.1 CPR 错误行为种类定义

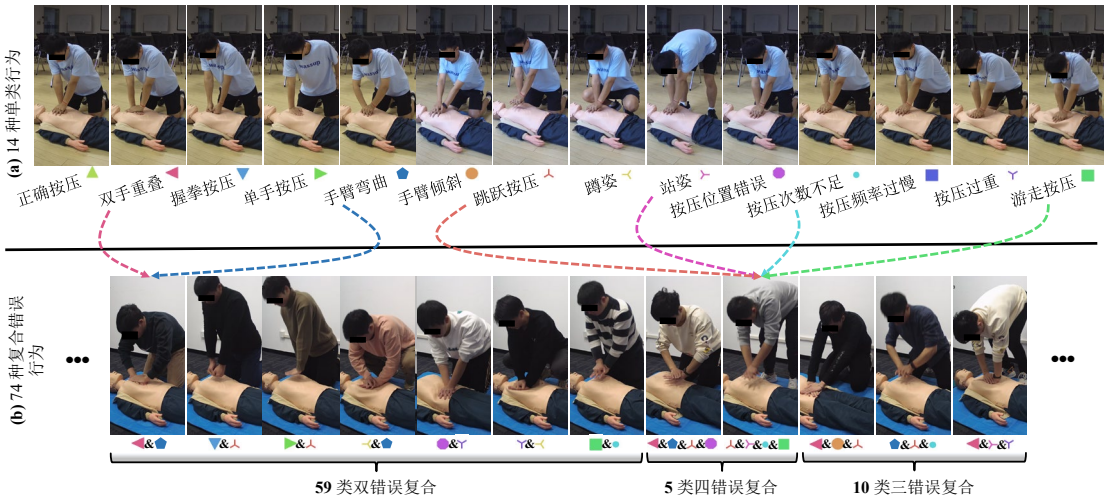


图 3-1　胸外按压单错误行为与复合错误行为展示

一套完整的心肺复苏急救流程包括胸外按压、人工呼吸和 AED 使用，本章重点探究了胸外按压过程中的错误行为识别任务。目前学界还没有研究对胸外按压动作的错误种类进行明确定义，本章在充分参考国内外关于心肺复苏技能教学的资源之后，在中山医院医疗技能教学与培训中心医生的指导下总结出了 13 类胸外按压错误行为，并根据日常培训中多种错误组合出现的频次总结了 59 类双错误复合（Paired-composite Errors）、10 类三错误复合（Triple-composite Errors）和 5 类四错误复合（Quadruple -composite Errors）。图 3-1(a)对胸外按压动作的 1 种正确行为和 13 种单类错误进行了展示。不同错误类型之间的差别较为微小，

需要仔细甄别才能进行区分：例如手臂弯曲、手臂倾斜、按压位置错误等。为提高单类错误之间的辨识度，本文为每种单类错误指派了不同颜色的标记。图 3-1(b)对 10 类三错误复合和 5 类四错误复合进行了列举与展示。复合错误的标记由单类错误标记组合而成。因此本章所构建的数据集共含有59＋10＋5＝74 种复合错误行为。

组合类型	组合错误行为	标记
10类三错误复合 Triple-Comp. Errors	双手重叠 & 手臂弯曲 & 跳跃按压	
	手臂弯曲 & 按压位置错误 & 双手重叠	
	手臂弯曲 & 双手重叠 & 按压次数不足	
	手臂倾斜 & 跳跃按压 & 双手重叠	
	按压位置错误 & 双手重叠 & 手臂倾斜	
	双手重叠 & 手臂倾斜 & 按压次数不足	
	手臂弯曲 & 跳跃按压 & 按压次数不足	
	蹲姿 & 手臂倾斜 & 按压位置错误	
	站姿 & 按压过重 & 双手重叠	
	站姿 & 双手重叠 & 按压次数不足	
5类四错误复合 Quadruple-Comp. Errors	双手重叠 & 手臂弯曲 & 跳跃按压 & 按压位置错误	
	站姿 & 游走按压 & 跳跃按压 & 按压次数不足	
	双手重叠 & 手臂倾斜 & 按压位置错误 & 按压次数不足	
	手臂弯曲 & 跳跃按压 & 按压位置错误 & 按压次数不足	
	手臂倾斜 & 手臂弯曲 & 双手重叠 & 游走按压	

图 3-2　三错误与四错误组合行为列表与对应标记

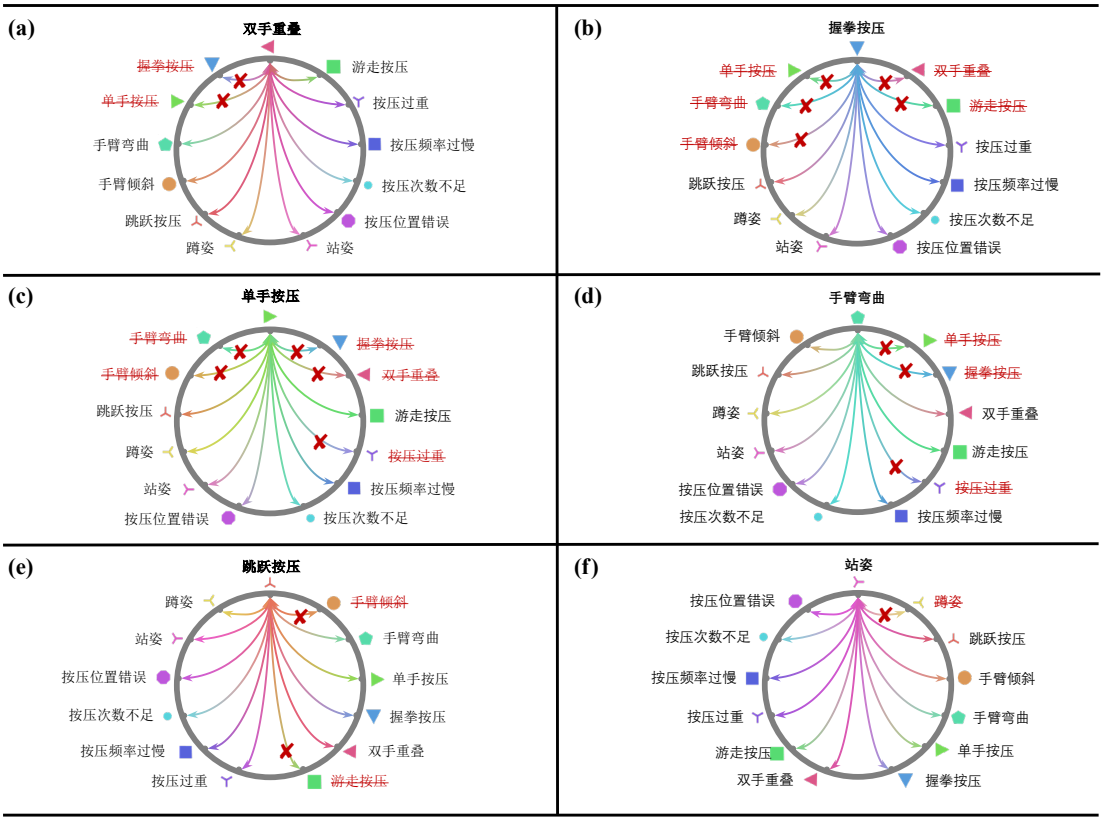
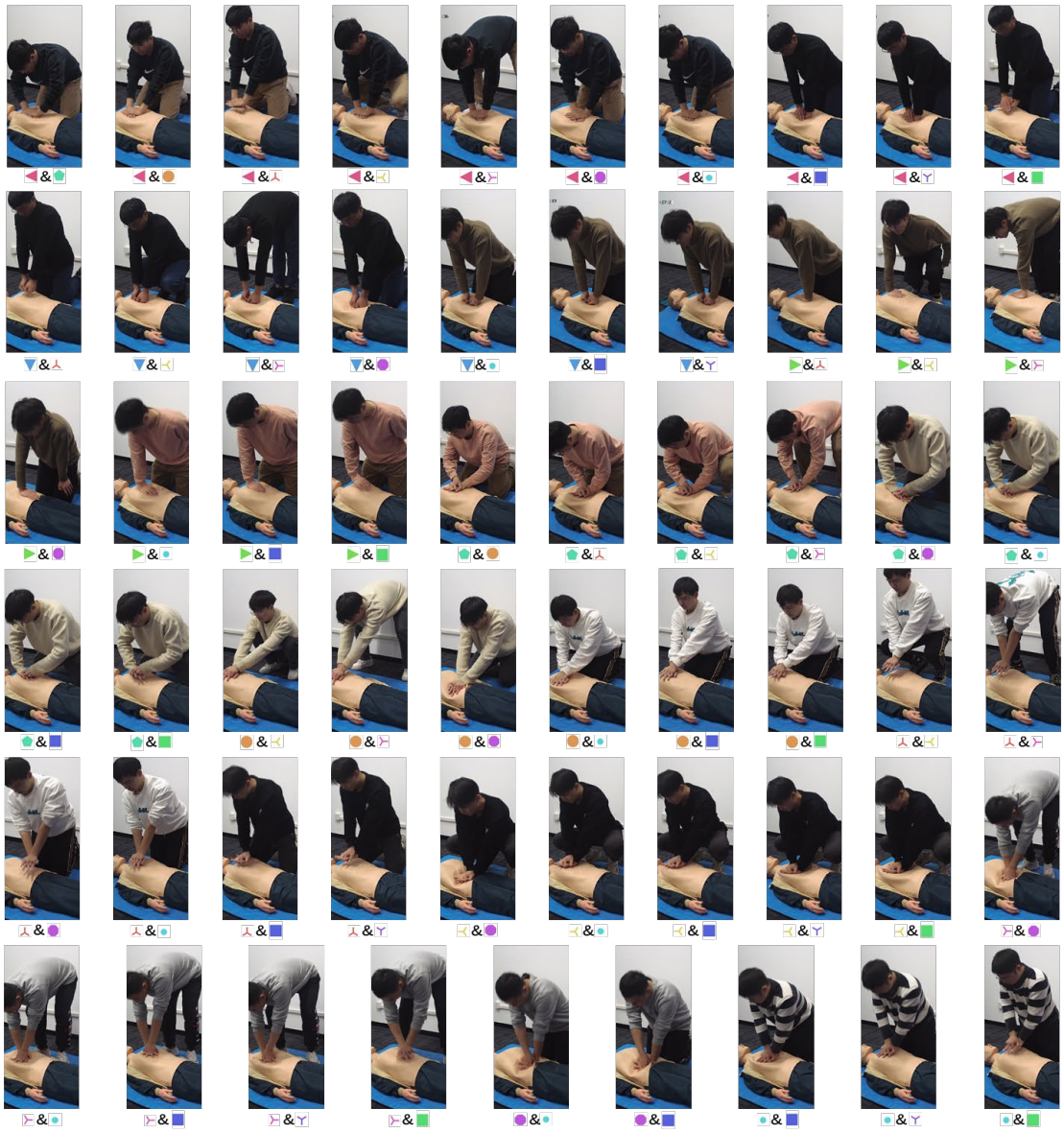
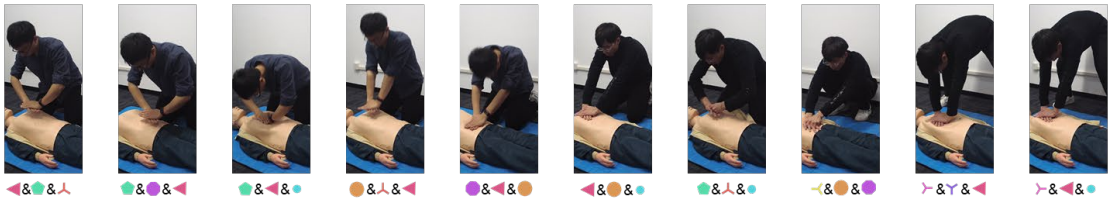


图 3-3　双错误复合种类筛选机制

(a) 59类双错误复合行为



(b) 10类三错误复合行为



(c) 5类四错误复合行为

图 3-4　74 类复合错误案例展示图

根据排列组合计算，13 种单类错误共可产生$C_{13}^{2}=78$种双错误复合行为，然而有一些错误复合情况出现在真实施救过程中的频率较低，例如握拳按压&单手按压、手臂倾斜&按压过重、手臂弯曲&按压过重等组合。本章对 78 种双错误复合行为进行了筛选，最终确定了 59 种双错误复合种类用于 CPR-Coach 数据集构建。图 3-3 选取了六类单错误行为以组合图的方式对筛选机制进行了展示。与双错误复合种类的“列举－筛选”确定形式不同，10 类三错误复合和 5 类四错误复合种类是在临床技能培训医生的指导下，根据错误行为之间的关联进行确定。

本节对胸外按压行为的错误种类进行了明确，即完成了 CPR-Coach 数据集的错误行为标签空间定义。后续的数据采集环节将会依照此标签空间的结构完成 CPR-Coach 数据集构建。如图 3-5 所示，CPR-Coach 数据集由两部分构成：包含单错误行为的 Set-1 数据集和包含多种类复合错误的 Set-2 数据集。在真实的评估任务中，Set-1 内的种类是有限的，而 Set-2 中的错误组合种类是巨量的，因此 Set-1 使用方形框绘制，而 Set-2 使用不规则框绘制。

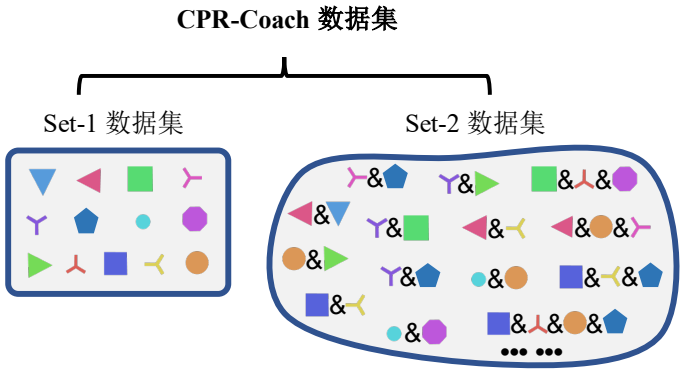


图 3-5　CPR-Coach 数据集结构

3.3.2 视频采集系统搭建

为采集心肺复苏的单类错误行为和复合错误行为视频样本，本章构建了一套多视角视频采集系统。如图 3-6 所示，四个摄像头分别放置在施救区域的正前方、左前方、右前方和侧方。采用多视角采集方案的目的共有两个：在丰富数据集体量方面，多视角的采集方案能够在单次按压中同时获得四条视频样例；在算法探究方面，多视角数据能够支持后续的视角优劣性探究与多视角组合识别实验。摄像头使用海康威视（Hikvision）监控摄像头，具体型号为 DS-2CD3T86FWDV2-I3S（国内标配），焦距为 4mm，录制视频时使用 4K 分辨率与 25 FPS 配置。

在完成胸外按压视频数据集构建之后，本章使用 TV-L1[138]光流提取算法和 Alphapose[139]姿势估计算法分别对 RGB 视频进行光流和人体姿势信息提取。因此在 CPR-Coach 数据集中，每个视角下的胸外按压视频共有三种模态信息：RGB 视频、2D 姿势与光流信息，如图 3-7 所示。丰富的模态信息为后续复合错误行

为识别算法的设计与多模态信息测试提供了基础。

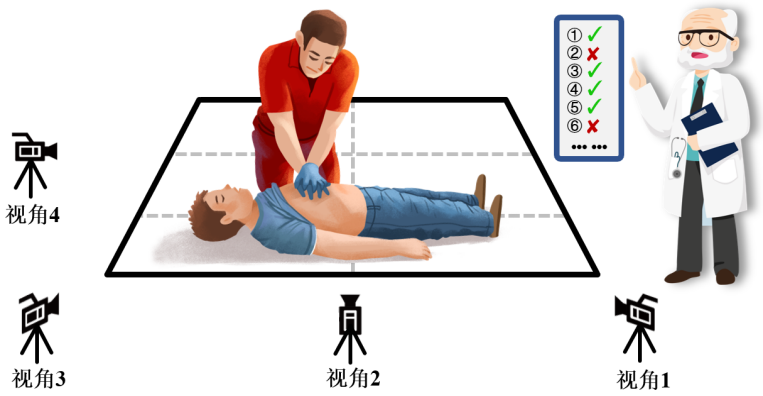


图 3-6 心肺复苏行为视频采集系统示意图

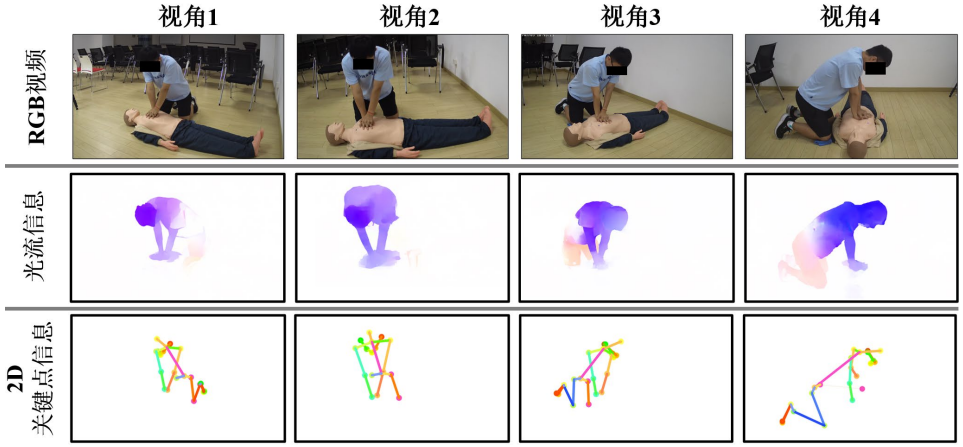


图 3-7 CPR-Coach 数据集提供的三种模态信息

3.3.3 数据采集与数据集结构

本研究共招募 12 名被试者参与 CPR-Coach 数据集构建过程，不同体型、穿着的人员为数据集的视觉特征丰富性提供了保障。3 名被试者参与单类错误行为数据集（Set-1）构建，9 名被试者参与复合错误行为数据集（Set-2）构建。Set-1 中每种单类错误行为的采集次数为 42；Set-2 中双错误行为采集 12 次，三错误和四错误均采集 8 次。在被试者执行心肺复苏操作时，四个摄像头同时进行采集。所有的心肺复苏操作都在医生的指导下进行，以确保各种行为的高质量完成。数据集含有的视频总数为 5,664，计算方法如下。表 3-1 从不同维度对 CPR-Coach 数据集的统计信息进行了汇总。

$$(14\times 42+59\times 12+15\times 8)\times 4\ \text{Views}=5,664\ \text{Videos} \tag{3.1}$$

表 3-2 将 CPR-Coach 数据集与学界中现有的医疗技能评估数据集进行了多维度的对比。对比结果显示，本文所构建的 CPR-Coach 数据集在行为种类数量、

视频数量和模态丰富度均优于现有数据集。基于 CPR-Coach 数据集，本文重点对以下两个任务进行了探究：单类错误行为识别任务与复合错误行为识别任务。图 3-8 展示了两个任务与 CPR-Coach 数据集的关联。单类错误行为识别任务在 Set-1 数据集内部进行训练集与测试集的划分，用于探究现有行为识别模型在心肺复苏场景下对多种错误的辨识能力；复合错误行为识别任务采取 Set-1 为训练集、Set-2 为测试集的设定进行“单类训练，多类测试”模式下的复合错误行为识别能力探究。

表 3-1　CPR-Coach 数据集统计信息

数据集统计项	数据
视角	4
帧率/FPS	25
分辨率	4096×2160 (4K)
参与人数	12
单类行为数量	1+13=14
复合错误行为数量	59+10+5=74
帧数（RGB）	2,217,756
帧数（RGB+光流）	6,644,596
视频数量	5,664
平均视频时长	19.52s
存储空间	450GB

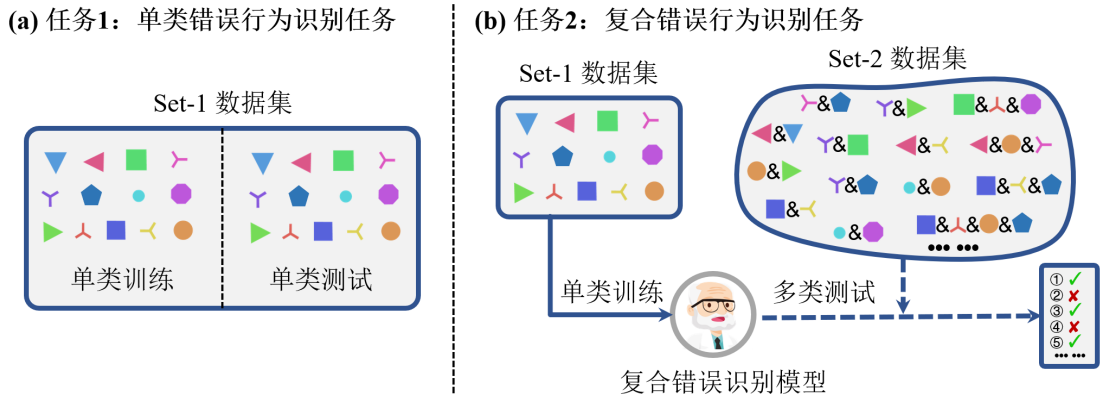


图 3-8　单类错误与复合错误行为识别任务示意图

3.3.4 模型评估指标

对于单类错误行为识别任务，本文使用视频分类任务通常使用的 Top-1 Acc 与 Top-3 Acc 分类准确度作为性能评估指标。Top-1 Acc 指所有测试视频中最高预测分数的类别是类别标签的比例；计算 Top-3 Acc 需要首先对每个视频的预测结果进行排序，若正确标签在前三个种类中则预测正确，最终计算预测正确的案例在所有样本中的比例。

表 3-2 CPR-Coach 数据集与现有医疗技能评估与识别数据集对比

研究主题	数据集	#行为	数据模态	#视频	#视角	评估类型
腹腔镜手术技能评估	FLS-ASU[103]	1	RGB	28	2	技能排序
	Zhang *et al.*[62]	1	RGB	546	1	技能排序
	Chen *et al.*[61]	3	RGB	720	2	技能排序
基本手术技能评估	Sharma *et al.*[19]	2	RGB	33	1	OSATA 分数
	Bettadapura *et al.*[140]	3	RGB	64	2	技能排序
	Zia *et al.*[16]	2	RGB	104	1	技能排序
da Vinci 手术机器人系统操作评估	MISTIC-SL[102]	4	RGB+Kinematics	49	1	技能排序
	JIGSAWS[15]	3	RGB+Kinematics	103	1	技能排序
康复训练评估	UI-PRMD[64]	10	RGB+Kinematics	1,000	4	技能排序
外科手术流程识别	Cataract-101[21]	10	RGB	101	1	手术流程识别
	Hei-Chole[128]	7	RGB	33	1	手术流程识别
	HeiCo[70]	20	RGB	30	1	手术流程识别
	RARP45[75]	8	RGB	45	1	手术流程识别
	Cholec80[71]	7	RGB	80	1	手术流程识别
	GastricBypass[141]	10	RGB	337	1	手术流程识别
	Gastrectomy[142]	8	RGB	461	1	手术流程识别
	Nephrec9[22]	10	RGB	1,262	1	手术流程识别
	CATARACTS[73]	21	RGB	50	1	手术器具识别
	CholecT50[20]	10	RGB	50	1	三元组识别
	Laparo425[143]	9	RGB	425	1	手术阶段预测
	PETRAW[13]	6	RGB+Kinematics	90	1	手术流程识别
	DESK[76]	7	RGB+Kinematics	2,897	1	手术流程识别
心肺复苏	CPR-Coach	14+74	RGB+Flow+Pose	5,664	4	复合错误识别

对于复合错误行为识别任务，本文采用 Moments-in-Time[144]数据集评估体系中的宏平均均值精度（Macro mAP）与微平均均值精度（Micro mAP）对模型的识别性能进行度量。其中 Macro mAP 在类别层次对结果进行度量，其定义方式与多标签分类中的平均均值精度（mean Average Precision, mAP）定义一致：

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{3.2}$$

其中 AP_i 表示模型在第 i 个类别中的识别精度，即在所有预测为正样本中实际正样本的比率，C 表示种类数量。AP_i 的计算方法为：

$$AP_i = \frac{TP_i}{TP_i + FP_i} \tag{3.3}$$

Micro mAP 表示所有视频的识别精度平均值，本文使用 mmit mAP 表示此指标，其计算方式为：

$$mmit\ \ mAP = \frac{\sum_{j=1}^{K} AP_j}{K} \tag{3.4}$$

其中 AP_j 表示模型对第 j 个样本的识别精度，K 表示测试集合中的视频数量。

3.4 基于特征组合机制的复合错误行为识别算法

在真实的医疗操作评估场景中，被试者很有可能在同一个动作中犯多个错误。以本文所构建的 CPR-Coach 数据集为例，13 种单错误行为的自由组合可以产生高达 8191 种复合错误行为：$\sum_{n=1}^{13} C_{13}^{n} = 2^{13} - 1 = 8191$。若仿照传统的多分类数据集构建方法，需要对每种错误组合类别进行样本采集，这种方案是完全不可行的。因此本文提出了“单类训练，多类测试”（Single-class Training & Multi-class Testing）任务，如图 3-8(b)所示：模型在只含有单类错误样本的训练集上完成训练，最终在含有多类复合错误样本的测试集中进行预测。在这种模式下，训练集与测试集之间的巨大差异会影响模型性能。一个良好的识别模型应具备优良的特征迁移能力。本章提出了一种基于特征组合机制的训练框架 ImagineNet 来处理这种训练集和测试集之间存在较大差异的问题。

本节内容组织如下：3.4.1 小节对单分类模型朴素迁移方法进行介绍；3.4.2 小节介绍本文提出的特征组合训练机制，即 ImagineNet 网络框架，并提出了三种网络结构对 ImagineNet 框架进行实例化；3.4.3 小节描述了基于随机线性加权的特征融合机制；3.4.4 小节对 ImagineNet 框架的训练和推理流程进行了阐述。

3.4.1 单分类模型朴素迁移方法

传统的行为识别数据集通常只为每个视频分配单个类别标签，因此行为识别任务是一个多分类问题（Multi-class Task）。人体行为识别模型的功能是对每一个视频进行种类的预测。给定视频序列$\mathbf{V} = \{I_i\}_{i=1}^{N}$，其中$N$表示视频帧数，$I_i$表示视频序列中的第$i$张图像。行为识别模型需要对视频序列进行特征提取，最终将视频特征映射为标签$C \in \{C_1, C_2, \cdots, C_M\}$，其中$M$表示类别数量。随着深度学习技术的不断发展，学界中已涌现出各种各样的行为识别框架。这些框架的主要的区别在于网络主干（Backbone）的结构，越复杂的主干结构通常代表着越强大的特征表示能力与越优异的行为识别性能。

针对本文所提出的“单类训练，多类测试”问题，一种朴素的策略是首先在只含有单类错误样本的数据集（即 CPR-Coach 中的 Set-1 数据集）上完成行为识别模型的训练，再将模型直接迁移到复合错误数据集上（即 CPR-Coach 中的 Set-2 数据集）。如图 3-8 所示，将(a)方式训练得到的模型直接应用于(b)任务。由于训练集与测试集之间存在较大的数据分布差异，这种朴素迁移方法的性能会受到域偏移现象（Domain Shift）的影响。在后续实验中，本文将此朴素迁移策略设定为基线模型，以探究本章所提出的特征组合训练机制能否带来复合错误识别性能的提升。

为充分探究损失函数对行为识别网络的单分类性能和复合错误行为识别性

能的影响，本文采用了三种损失函数设定对行为识别模型进行训练：交叉熵损失（Cross Entropy Loss, CE Loss）、二进制交叉熵损失（Binary Cross Entropy Loss, BCE Loss）和多间隔损失（Multi-Margin Loss）。其中交叉熵损失是多分类任务中最常用的损失，此损失假设各个类别之间具有互斥关系。设共有C个类别，给定模型对各个类别的预测结果$S_i \in [0,1]$与各个类别的指示标签$GT_i \in \{0,1\}$，交叉熵损失的计算过程为：

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} GT_i \cdot \log S_i \tag{3.5}$$

二进制交叉熵损失将各个类别的关系设置为独立，取消了类别之间的约束关系，其计算过程为：

$$\mathcal{L}_{BCE} = -\frac{1}{C}\sum_{i=1}^{C}\left(GT_i \cdot \log S_i + (1-GT_i)\cdot \log(1-S_i)\right) \tag{3.6}$$

多间隔损失函数通常用于多标签分类任务中，与之前两种损失不同，多间隔损失考虑真实类别与其他类别之间的误差，其计算过程为：

$$\mathcal{L}_{MM} = \frac{1}{C}\sum_{i\neq y}^{C} \max(0, margin - S_y + S_i)^{p} \tag{3.7}$$

其中指数p默认设定为 1，$margin$默认设定为 1。

3.4.2 基于 Imagine 机制的特征组合训练策略

在着手处理单类训练—多类测试任务之前，本文首先对真实医疗技能评估中教练医师（Coach）的判断过程进行了思考：一个真正的教练医师不可能见过所有的复合错误种类，但是却可以根据有限的单类错误案例快速地对复合错误案例进行判断。这是由于人具有极强的知识组合与推理能力，从而在给定少量单类参考案例的设定下完成特征组合，实现稳定的多标签识别能力。因此本文提出了一种由人类认知启发的 ImagineNet 框架来解决医疗行为分析任务中的单类训练—多类测试问题。图 3-9(a)展示了 ImagineNet 框架的主体思路，其本质是一种视觉特征组合训练机制，能够充分地利用单类错误数据集中的样本进行特征组合训练，最终有效提升模型对复合错误样本的识别能力。

图 3-9(b)以 TSN 模型为视频主干网络对 ImagineNet 框架的细节进行了展示。ImagineNet 框架可划分为三个部分：视觉特征提取阶段（Visual Feature Extraction Stage）、特征融合阶段（Feature Fusion Stage）与损失计算阶段（Loss Computing Stage）。在视觉特征提取阶段中，首先从 CPR-Coach 的 Set-1 数据集中随机采样两个视频$(\mathbf{V}_1, C_1)$与$(\mathbf{V}_2, C_2)$，这两个视频来自不同错误类型，即采样过程满足

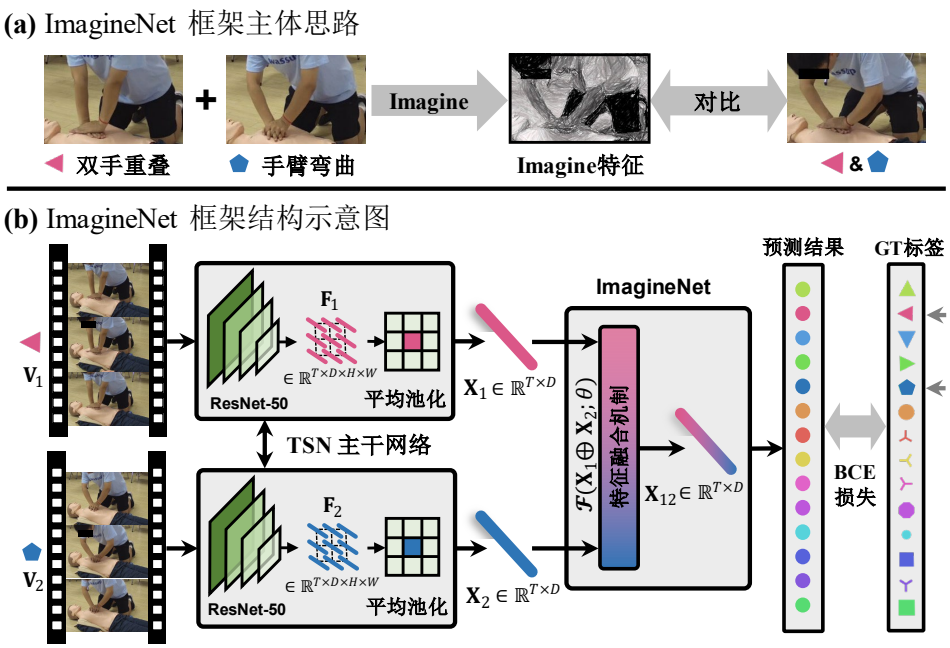


图 3-9　ImagineNet 框架的主体思路与结构示意图

$C_1 \neq C_2$。其中$\mathbf{V}_1 = \{I_i\}_{i=1}^{N_1}$，$\mathbf{V}_2 = \{I_i\}_{i=1}^{N_2}$，$N_1$与$N_2$分别表示两个视频含有的帧数。视频主干网络 TSN 从两个原始视频中分别采样T个视频片段后进行特征提取，最终生成视频特征$\mathbf{X}_1 \in \mathbb{R}^{T\times D}$与$\mathbf{X}_2 \in \mathbb{R}^{T\times D}$，其中$D$表示视频特征维度。在特征融合阶段中，两个视频特征通过特征融合机制映射为$\mathbf{X}_{12}$，特征融合过程表示为$\mathbf{X}_{12} = \mathcal{F}(\mathbf{X}_1 \oplus \mathbf{X}_2, \boldsymbol{\theta})$，其中$\mathcal{F}(\cdot)$表示不同的特征融合网络，$\boldsymbol{\theta}$表示网络参数。本文共提出了三种不同的特征融合机制：基于全连接层的融合网络（Fully Connected Layer based fusion, FC）、基于自注意力机制的融合网络（Self-Attention based fusion, SA）和基于交叉注意力机制的融合网络（Cross-Attention based fusion, CA）。在损失计算阶段中，首先对两个样本的独热编码标签进行取并集操作，再使用二进制交叉熵损失（BCE Loss）对网络预测结果与聚合标签之间的差异进行度量。

3.4.3 特征融合机制设计

本文共设计了三种特征融合网络实现$\mathbf{X}_{12} = \mathcal{F}(\mathbf{X}_1 \oplus \mathbf{X}_2, \boldsymbol{\theta})$映射过程。如图 3-10 所示，分别为 ImagineNet-FC、ImagineNet-SA 和 ImagineNet-CA。清晰起见，此图只展示了双错误组合的训练过程，而 ImagineNet 支持多类错误复合作为输入，只需改变输入种类的数量即可。

基于全连接层的融合网络 ImagineNet-FC

图 3-10(a)展示了基于全连接层的特征融合网络。视频特征$\mathbf{X}_1$和$\mathbf{X}_2$首先通过特征聚合操作进行融合生成$\mathbf{X}_1 \oplus \mathbf{X}_2$，之后使用两个全连接层完成错误类别预测，

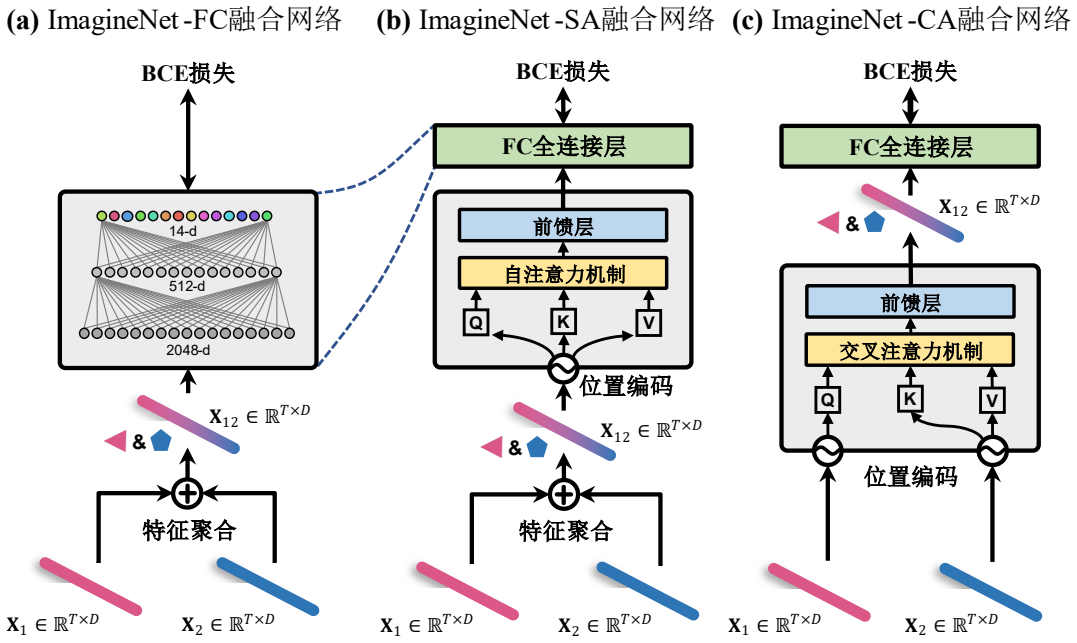


图 3-10 ImagineNet 框架的三种特征融合网络

此映射过程可表示为：

$$S_{FC} = \mathcal{F}_{FC}(\mathbf{X}_1 \oplus \mathbf{X}_2, \boldsymbol{\theta}_{FC}) \tag{3.8}$$

其中 $\mathcal{F}_{FC}(\cdot)$ 表示由全连接层构成的神经网络，$\boldsymbol{\theta}_{FC}$ 表示网络的参数。通过使用 BCE 损失函数对预测分数与标签之间的差异进行度量。网络优化的目标是寻找最优的网络参数使得 BCE 损失函数取得最小：

$$\boldsymbol{\theta}^*_{FC} = \arg\min_{\boldsymbol{\theta}_{FC}} \mathcal{L}_{BCE}(S_{FC},\ GT) \tag{3.9}$$

其中 GT 表示单类错误独热编码的并集：$GT = \text{Onehot}(C_1) \cup \text{Onehot}(C_2)$。清晰起见本文在后续表述中对网络参数 $\boldsymbol{\theta}$ 均进行了省略。

基于自注意力机制的融合网络 ImagineNet-SA

图 3-10(b)展示了基于自注意力机制的特征融合网络。引入自注意力机制的目的是：在全连接层网络的基础上强化模型对时序特征的表示能力，从而获得更优质的融合效果。ImagineNet-SA 的特征映射过程表示为：

$$S_{SA} = \mathcal{F}_{FC}(\mathcal{F}_{SA}(\mathbf{X}_1 \oplus \mathbf{X}_2)) \tag{3.10}$$

其中 $\mathcal{F}_{SA}(\cdot)$ 表示自注意力机制与前馈连接层，$\mathcal{F}_{FC}(\cdot)$ 表示全连接层网络。自注意力机制的计算过程为：

$$\mathbf{X}'_{SA} = LN\left[\mathbf{X}_{12} + softmax\left(\frac{\mathbf{X}_{12}\mathbf{X}_{12}^T}{\sqrt{D}}\right)\mathbf{X}_{12}\right] \tag{3.11}$$

其中D表示视频特征的维度，在 TSN 视频主干网络的设定中，特征维度$D=2048$。$\mathrm{LN}[\cdot]$表示层归一化操作（Layer Normalization）。

前馈连接层的计算过程为：

$$\mathbf{X}_{FFN}=\mathrm{LN}[\mathbf{X}'_{SA}+\mathcal{F}_{FFN}(\mathbf{X}'_{SA})] \tag{3.12}$$

为保证训练的稳定性，本文在层归一化操作之前均添加了残差链接。此部分并未在图 3-10 中展示。最终由全连接网络$\mathcal{F}_{FC}(\cdot)$将前馈连接层特征$\mathbf{X}_{FFN}$映射为各个类别的分数：

$$S_{SA}=\mathcal{F}_{FC}(\mathbf{X}_{FFN}) \tag{3.13}$$

基于交叉注意力机制的融合网络 ImagineNet-CA

图 3-10(c)展示了基于交叉注意力机制的特征融合网络。ImagineNet-CA 与 ImagineNet-SA 最大的差别在于特征融合机制。ImagineNet-SA 网络通过$\mathbf{X}_1\oplus\mathbf{X}_2$操作实现特征聚合，而 ImagineNet-CA 网络通过交叉注意力机制替代特征聚合过程。ImagineNet-CA 网络的映射过程表示为：

$$S_{CA}=\mathcal{F}_{FC}(\mathcal{F}_{CA}(\mathbf{X}_1,\mathbf{X}_2)) \tag{3.14}$$

其中$\mathcal{F}_{CA}(\cdot,\cdot)$网络由一个交叉注意力层与一个前馈连接层级联构成。交叉注意力机制的实现过程表示为：

$$\mathbf{X}'_{CA}=\mathrm{LN}\left[\mathbf{X}_1+softmax\left(\frac{\mathbf{X}_1\mathbf{X}_2^T}{\sqrt{D}}\right)\mathbf{X}_2\right] \tag{3.15}$$

前馈连接层的计算过程为：

$$\mathbf{X}_{FFN}=\mathrm{LN}[\mathbf{X}'_{CA}+\mathcal{F}_{FFN}(\mathbf{X}'_{CA})] \tag{3.16}$$

与 ImagineNet-CA 模型的最后环节保持一致，前馈连接层生成的特征$\mathbf{X}_{FFN}$通过全连接网络$\mathcal{F}_{FC}(\cdot)$映射为各个类别的分数：

$$S_{SA}=\mathcal{F}_{FC}(\mathbf{X}_{FFN}) \tag{3.17}$$

基于随机线性组合机制的特征聚合策略

随机线性加权机制最初起源于 MixUp[145]与 Manifold MixUp[146]等研究。在这些工作中，此机制作为一种正则化机制被应用于图像分类任务与对抗攻击任务。实验结果证实图像层次和特征层次的样本混合均可有效提升模型的识别性能。本文将随机线性加权机制引入到复合错误行为识别任务中，并将融合机制拓展到更宽泛的多输入形式。在 ImagineNet-FC 和 ImagineNet-SA 网络中，视频特征$\mathbf{X}_1$和$\mathbf{X}_2$通过特征聚合操作生成$\mathbf{X}_{12}$。一种最朴素的做法是直接将两个特征进行求和。

为增加融合过程的多样性，本文采用随机线性组合机制对不同种类的视频特征进行聚合。以两种错误样本的特征复合为例，计算过程表示为：

$$\mathbf{X}_{12}=\lambda\mathbf{X}_1+(1-\lambda)\mathbf{X}_2,\ \lambda\sim U(0,1) \tag{3.18}$$

其中λ为随机权重，从均匀分布$U(0,1)$中采样得到。图 3-11 以双手重叠和手臂弯曲对此特征聚合过程进行了展示。

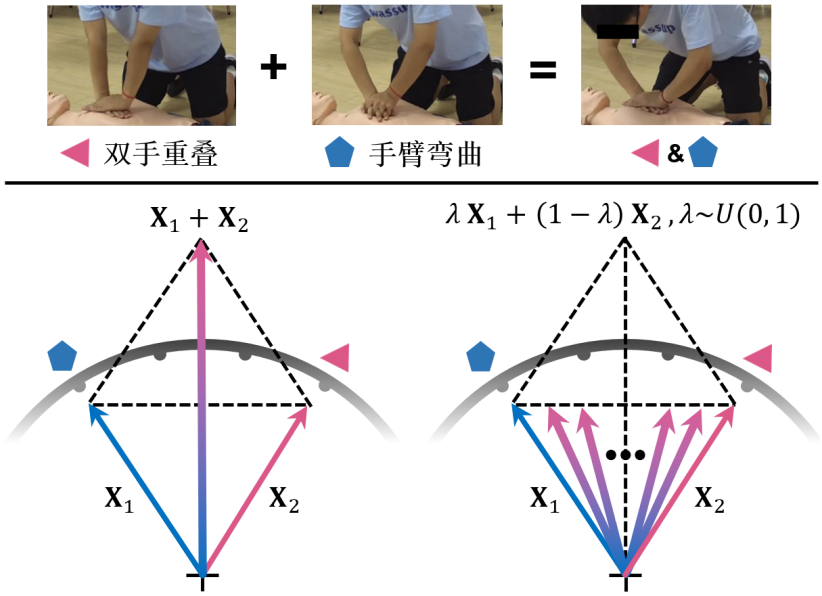


图 3-11　基于随机加权的特征聚合策略示意图

在多种类错误组合的情况下，需要为每个输入特征指派随机权重。以四错误复合为例，聚合特征表示为：

$$\mathbf{X}_{Fuse}=\lambda_1\mathbf{X}_1+\lambda_2\mathbf{X}_2+\lambda_3\mathbf{X}_3+\lambda_4\mathbf{X}_4 \tag{3.19}$$

其中权重参数满足：$\lambda_1+\lambda_2+\lambda_3+\lambda_4=1$。

3.4.4 模型训练与推理

ImagineNet 框架的训练过程

ImagineNet 框架的训练目标是：在不影响模型单错误识别性能的前提下提升复合错误识别性能。为了能够同时实现这两个目标，本文在 ImagineNet 框架的训练过程中采用了四种输入配置以充分发挥组合特征训练机制的作用。由于所有的错误在测试阶段均有可能发生，所以本文使用二进制交叉熵损失函数对 ImagineNet 框架进行训练。

（1）单错误输入案例：前文在双错误的设定下对 ImagineNet 的结构进行了介绍。在单错误输入设定中，公式(3.18)中的随机线性组合机制退化为直连操作。以样本$(\mathbf{X},C)$为例，公式(3.18)中的视频输入特征$\mathbf{X}_1$与$\mathbf{X}_2$被替换为同一个特征，

特征聚合结果表示为：$\mathbf{X}=\lambda\mathbf{X}_1+(1-\lambda)\mathbf{X}_2$。类别标签$C$以独热编码的形式参与 BCE 损失的计算过程。设单类错误输入样本的采样空间为$\mathcal{S}$，训练采样过程可以表示为$s\sim\mathcal{S}$，其中s表示一个视频样本。采样空间$\mathcal{S}$共含有 13 种错误类别。

（2）双错误输入案例：此类案例的构建目的是增强 ImagineNet 框架对双错误复合行为的识别能力。图 3-9 展示了双错误输入设定下的 ImagineNet 模型，公式(3.18)对双错误设定下的特征聚合过程进行了描述。设双错误输入样本的采样空间为$\mathcal{P}$，训练过程中的采样表示为$p\sim\mathcal{P}$，其中p表示满足$C_1\neq C_2$的成对视频样本集合。采样空间$\mathcal{P}$共含有$C_{13}^2=78$种不同的类别组合方式。

（3）三错误输入案例：三种错误类型的设定下特征聚合过程类似于公式(3.19)。设三错误输入样本的采样空间为$\mathcal{T}$，训练过程中的采样表示为$t\sim\mathcal{T}$，其中t表示满足$C_1\neq C_2\neq C_3$的三元视频样本组合。采样空间$\mathcal{T}$共含有$C_{13}^3=286$种不同的类别组合方式。

（4）四错误输入案例：公式(3.19)对四种错误行为特征的聚合过程进行了描述。设四错误输入样本的采样空间为$\mathcal{Q}$，模型训练过程中的采样表示为$q\sim\mathcal{Q}$，其中q表示满足$C_1\neq C_2\neq C_3\neq C_4$的视频样本四元组。采样空间$\mathcal{Q}$共含有$C_{13}^4=715$种不同的类别组合方式。

综合以上四种复合错误输入案例，ImagineNet 框架的最终训练目标是：最小化所有错误组合情况下的总体二进制交叉熵损失：

$$\mathcal{L}_{Total}=\mathop{\mathbb{E}}_{s\sim\mathcal{S}}[\mathcal{L}_{BCE}^{s}]+\mathop{\mathbb{E}}_{p\sim\mathcal{P}}[\mathcal{L}_{BCE}^{p}]+\mathop{\mathbb{E}}_{t\sim\mathcal{T}}[\mathcal{L}_{BCE}^{t}]+\mathop{\mathbb{E}}_{q\sim\mathcal{Q}}[\mathcal{L}_{BCE}^{q}] \tag{3.20}$$

需要注意图 3-10(c)只对 ImagineNet-CA 模型的双错误输入情况进行了展示。在输入为单个错误样本时，ImagineNet-CA 模型的两个输入端口均使用相同的单错误视频特征；在输入为多个错误样本时，ImagineNet-CA 模型将随机选择一个样本作为$\mathbf{Q}$通路的输入信息，其余样本通过对特征取平均作为$\mathbf{K}$和$\mathbf{V}$通路的输入信息。

ImagineNet 框架的推理过程

在 ImagineNet 的推理阶段，模型的输入只有单个测试视频特征$\mathbf{F}\in\mathbb{R}^D$。对于图 3-10 中的 ImagineNet-FC 与 ImagineNet-SA 网络，从训练模式转换到推理模式只需去除特征聚合环节，并使用测试视频特征$\mathbf{F}$替换聚合特征$\mathbf{X}_{12}$即可。对于 ImagineNet-CA 网络，本文通过使用特征复制策略实现两个输入端口的特征填充。虽然特征复制策略会导致 ImagineNet-CA 的交叉注意力机制退化为自注意力机制，但由于训练阶段特征输入机制的差异性，最终会产生不同的性能结果。后续的实验部分中 ImagineNet-CA 模型与 ImagineNet-FC、ImagineNet-SA 的性能对比结果证实了此推断。

3.5 实验分析

3.5.1 单类错误行为识别结果

相较于传统的行为识别数据集，CPR-Coach 数据集更加关注于细微动作的识别。如图 3-1 所示，各单类错误行为之间的视觉差异非常微小。因此本文首先以 Set-1 数据集为基准对现有人体行为识别模型进行了测试。Set-1 数据集中 60%样本用于训练，40%样本用于测试。本文对三种类型的网络进行了单分类性能测试，分别为：基于视频的识别模型（TSN[4]、TSM[126]、TPN[147]、I3D[7]、C3D[34]、TIN[148]、SlowFast[127]、TimeSFormer[38]）、基于 2D 姿势的识别模型（ST-GCN[124]、PoseC3D[149]）和基于多模态信息融合的识别模型（Two-Stream[31]、MMNet[150]）。

不同模型配置信息均列举在表 3-3 中，例如“1x1x8”表示视频主干对视频的采样信息：ClipLen × FrameIntervals × NumClips。除去 SlowFast 使用余弦退火优化器训练 256 轮次、MMNet 使用 SGD 优化器训练 80 轮次外，其他网络均通过 SGD 优化器训练 50 轮次。基于视频的网络输入图像尺寸被设定为 224 × 224；基于关键点的网络输入数据为原始二维骨骼关键点。表 3-3 对 Top-1 Acc 和 Top-3 Acc 进行了记录，每列中的最高数值使用加粗标记、次高数值使用下划线标记。

表 3-3 单类错误行为识别结果

模型	模态	主干	配置	轮次	CE Loss		BCE Loss		Multi-margin Loss	
					Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
TSN[4]	RGB	ResNet-50	1x1x8	50	0.8879	0.9940	0.8829	0.9960	0.8502	0.9901
	RGB	ResNet-50*	1x1x8	50	0.9067	0.9921	0.8919	0.9940	0.8690	0.9901
	Flow	ResNet-50	1x1x8	50	0.7907	0.9603	0.8304	0.9851	0.7073	0.9355
TSM[126]	RGB	ResNet-50	1x1x8	50	0.9067	0.9901	0.9325	0.9950	0.8433	0.9881
I3D[7]	RGB	ResNet-50	32x2x1	50	0.9692	0.9960	0.9117	0.9940	0.8591	0.9861
TPN[147]	RGB	ResNet-50	8x8x1	50	**0.9802**	0.9960	0.9087	**0.9980**	0.8720	0.9901
C3D[34]	RGB	C3D*	16x1x1	50	0.9722	0.9931	0.9702	0.9931	0.8621	0.9802
TIN[148]	RGB	ResNet-50	1x1x8	50	0.8800	0.9901	0.7192	0.9335	0.8393	0.9861
SlowFast[127]	RGB	ResNet-50	4x16x1	256	0.8695	0.9734	0.8719	0.9781	0.8625	0.9688
TimeSFormer[38]	RGB	ViT	8x32x1	50	0.8879	0.9921	0.8998	0.9940	0.8462	0.9762
ST-GCN[124]	Pose	ST-GCN	1x1x300	50	0.9246	**0.9970**	0.9187	0.9881	0.9196	**0.9970**
PoseC3D[149]	Pose	ResNet3D-50	1x1x300	240	0.9208	0.9922	0.9035	0.9715	0.8837	0.9606
Two-Stream[31]	RGB+Flow	TSN+TSN_Flow	Late-Fusion	50	0.9533	0.9891	0.9479	0.9825	0.9296	0.9802
	RGB+Pose	TSN+ST-GCN	Late-Fusion	50	0.9782	0.9962	0.9608	0.9941	**0.9692**	0.9960
MMNet[150]	RGB+Pose+RoI	MS-G3D+Incep.-v3	Late-Fusion	80	0.9756	0.9960	**0.9772**	0.9940	0.9512	0.9876

（注：ResNet-50*表示网络在 Kinetics-400[24]数据集中进行预训练，C3D*在 Sports-1M[8]中进行预训练）

表 3-4 朴素迁移方法的复合错误行为识别性能

模型	配置	模态	预训练	CE Loss		BCE Loss		Multi-Margin Loss	
				mAP	mmit mAP	mAP	mmit mAP	mAP	mmit mAP
TSN[4]	1x1x8	RGB	K-400	0.5598	0.6143	0.4627	0.5629	0.4838	0.5579
TSM[126]	1x1x8	RGB	✗	0.5662	0.6618	0.5721	0.6688	0.5470	0.6255
ST-GCN[124]	1x1x300	Pose	✗	0.5776	0.6692	**0.5868**	0.6865	0.5874	0.6719
PoseC3D[149]	1x1x300	Pose	✗	0.5498	0.6393	0.5556	0.6241	0.5358	0.6142
MMNet[150]	Late-Fusion	RGB+Pose+RoI	✗	**0.5948**	**0.6735**	0.5871	**0.6973**	**0.5894**	**0.6830**

表 3-3 中的实验结果显示，不同的网络与不同的损失函数搭配会影响单错误分类性能。在交叉熵损失函数设定下，以 TSN 为主干的 Two-Stream 网络能够取得最优结果。基于关键点的网络能够在 BCE 损失下取得最优性能。在所有测试模型中，融合多模态信息的网络性能最为稳定，在不同损失函数下都能取得高识别精度。交叉熵损失设定下的网络性能要普遍优于其他两种损失，这主要是因为交叉熵损失具有更加严格的标签依赖设定。一个比较反常的现象是，网络结构更为复杂的 PoseC3D 性能要差于经典的 ST-GCN 框架，这主要是由于 PoseC3D 的输入构造机制是将 2D 关键点堆叠为 3D 热图（Heatmap Volume），而 CPR 胸外按压动作具有重复性和循环性，PoseC3D 的堆叠机制会造成时序信息的丧失，从而导致次优的性能。

图 3-12 使用 t-SNE 算法对行为识别网络的输出特征进行了可视化。结果显示，易混淆的四类行为：按压频率过慢、按压次数不足、按压频率过慢、游走按压在经 TSN，TPN 和 TSM 网络映射后特征之间的距离较近，网络更容易混淆；而经 I3D 和 ST-GCN 这类更加关注时序特征的网络映射之后特征之间的距离相对较远，因此能够更好地区分。

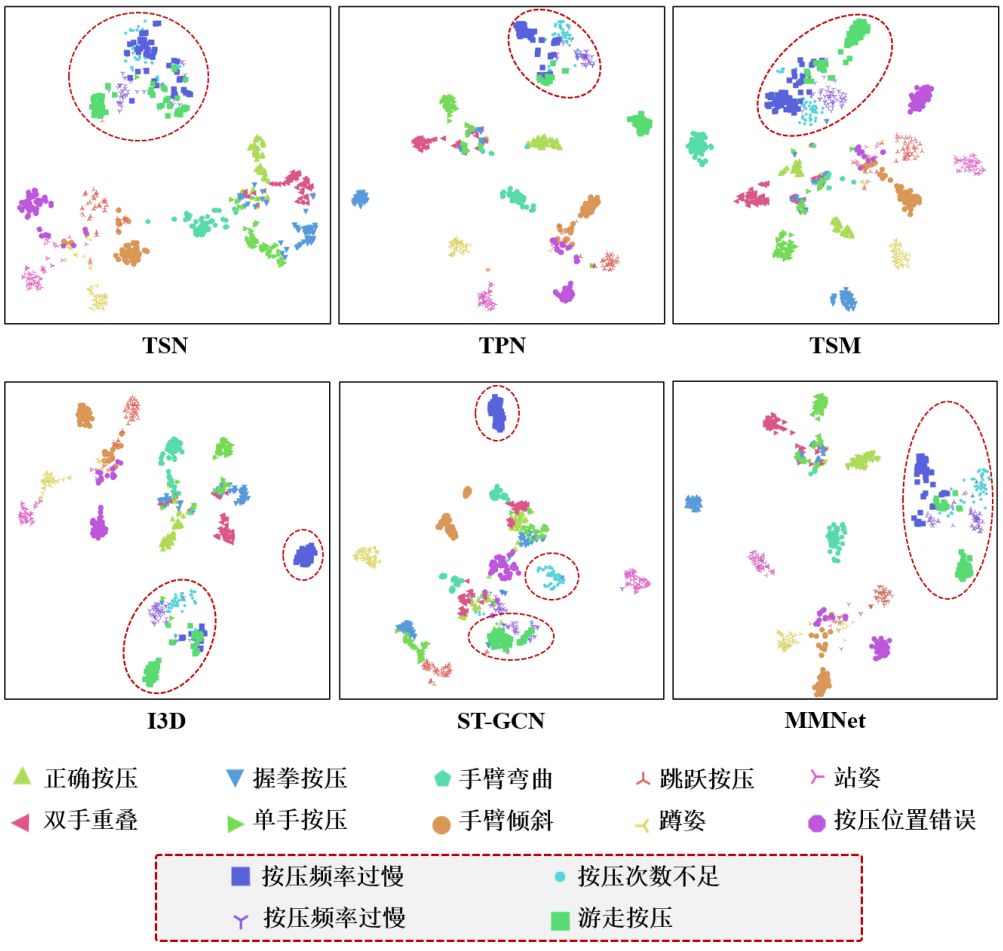


图 3-12　单分类模型 t-SNE 特征可视化结果

表 3-5 朴素迁移方法与 ImagineNet-FC 性能对比

模型	mAP	Δ	mmit mAP	Δ
TSN[4]	0.5598	—	0.6143	—
w/ ImagineNet-FC	**0.6259**	↑6.61%	**0.6893**	↑8.50%
TSM[126]	0.5662	—	0.6618	—
w/ ImagineNet-FC	**0.7053**	↑13.91%	**0.7566**	↑9.48%
ST-GCN[124]	0.5776	—	0.6692	—
w/ ImagineNet-FC	**0.6404**	↑6.28%	**0.7115**	↑4.23%
MMNet[150]	0.5948	—	0.6735	—
w/ ImagineNet-FC	**0.6927**	↑9.79%	**0.7478**	↑7.43%

3.5.2 复合错误行为识别结果

整体而言，现有行为识别模型能够妥善处理 CPR 场景下的单错误识别任务。在后续内容中，本文将着重对复合错误行为识别场景下的模型性能进行探究。通过将 Set-1 作为训练集，Set-2 作为测试集，可以模拟真实的训练－测试模式。一种朴素的做法是将单类训练得到的模型直接迁移到复合错误检测任务中。表 3-4 对在三种损失函数下进行训练的 TSN[4]、TSM[126]、ST-GCN[124]、PoseC3D[149]和 MMNet[150]网络进行了迁移测试。实验结果显示，三种损失函数均无法妥善处理单类错误识别与复合错误识别两个任务之间的巨大差异，MMNet 凭借多模态信息融合机制获得了更好的迁移效果。较低的性能说明复合错误行为识别任务已超出了模型的表示能力。

本文的研究重点并不在于创建一个全新的行为分类模型，而是通过现有的行为识别主干构建性能更优的复合错误行为检测模型。在表 3-5 中，本文选取 TSN[4]、TSM[126]、ST-GCN[124]和 MMNet[150]等各类网络中的经典模型作为视频特征提取主干，对朴素迁移结果和经过 ImagineNet-FC 框架训练后的结果进行了对比。所有模型均使用交叉熵损失函数进行训练。其中 Δ 表示 ImagineNet-FC 所带来的性能提升。结果显示，无论模型的输入是 RGB 信息、Pose 信息还是多模态信息，特征组合训练机制均能够有效处理单类错误识别与复合错误识别之间的领域差异，从而带来复合错误识别性能的提升。以 TSM 为视频主干，ImagineNet-FC 能够带来 15.04%的 mAP 提升与 11.34%的 mmit mAP 提升。

为充分探究 ImagineNet-FC 在最先进的视频主干上的性能，本文使用 Vi-ViT[37]、MViTv2[151]和 Video Swin Transformer[152]三个视频主干模型进行了实验。单分类性能与朴素迁移的复合错误识别结果如表 3-6 所示，所有模型均采用交叉熵损失函数进行训练。结果显示三个主干网络在单分类任务中取得了优异的结果，Top-1 Acc 逼近 100%，而朴素的迁移方式下，三个视频主干的 mAP 均在 56%左右，mmit mAP 在 66%左右。结果说明虽然更先进的视频主干有更强的单分类识别能力，但是朴素迁移策略仍然无法处理两个任务之间的巨大差异，从而造成复

合错误行为识别能力较差。表 3-7 记录了三个视频主干模型通过 ImagineNet-FC 网络训练之后的复合错误识别性能，并且计算了与朴素迁移策略性能的差值。结果显示三个视频主干网络的识别性能出现明显提升，以 Video Swin Transformer 视频主干为例，mAP 性能达到 70.82%，较朴素迁移方式提升 13.86%；mmit mAP 性能达到 76.38%，较朴素迁移方式提升 9.37%。

表 3-6　基于 SOTA 视频主干的单错误分类与复合错误分类性能

模型	配置	预训练	单类识别结果	
			Top-1	Top-3
Vi-ViT[37]	base-16x2	Kinetics-400	0.9814	1.0000
MViTv2[151]	base-32x3x1	Kinetics-400	0.9867	0.9980
Video Swin[152]	base-32x2x1	Kinetics-400	0.9918	1.0000
模型	**配置**	**预训练**	**直接迁移策略结果**	
			mAP	**mmit mAP**
Vi-ViT[37]	base-16x2	Kinetics-400	0.5582	0.6651
MViTv2[151]	base-32x3x1	Kinetics-400	0.5715	0.6740
Video Swin[152]	base-32x2x1	Kinetics-400	0.5696	0.6701

表 3-7　基于 SOTA 视频主干的朴素迁移方法与 ImagineNet-FC 性能对比

模型	mAP	Δ	mmit mAP	Δ
Vi-ViT[37]	0.5582	—	0.6651	—
w/ ImagineNet-FC	**0.6587**	+10.05%	**0.7523**	+8.72%
MViTv2[151]	0.5715	—	0.6740	—
w/ ImagineNet-FC	**0.6869**	+11.54%	**0.7461**	+7.21%
Video Swin[152]	0.5696	—	0.6701	—
w/ ImagineNet-FC	**0.7082**	+13.86%	**0.7638**	+9.37%

为探究不同网络结构对复合错误行为识别性能的影响。表 3-8 与表 3-9 对 ImagineNet 网络的不同结构进行了性能测试，并对自注意力机制不同的堆叠数量进行了改变。其中表 3-8 以 TSN[4]作为视频主干、表 3-9 以 TSM[126]作为视频主干。在两种设定中，ImagineNet-SA 相较于 ImagineNet-FC 与 ImagineNet-CA 均能取得更加优异的复合错误行为识别性能。交叉注意力机制并不能在相同模态输入的设定下发挥作用。过多的自注意力层堆叠会导致过拟合现象发生，从而降低模型的复合错误识别性能。去除位置编码层会导致模型性能降低，这说明时序信息在辨别各类错误时起到重要作用。将表 3-8 与表 3-9 进行整体对比可发现，在复合错误识别任务中 TSM 视频主干的性能要优于 TSN 网络，此结果与表 3-3 中单类错误识别结果保持一致。这是因为 TSM 有更强的特征表示能力。

为探究模型在不同数量复合错误上的识别性能，本文使用 TSN[4]、ST-GCN[124]和 MMNe[150]依次作为视频主干网络，对 ImagineNet-FC 在 Set-2 全集和各 Set-2 子集上的复合错误识别性能进行了探究，实验结果如图 3-13 所示。其中横轴为四种 Set-2 的集合设定，分别为：所有 74 种复合错误行为、59 种双错误复合行为、10 种三错误复合行为和 5 种四错误复合行为；纵轴为 mmit mAP 性能。结

表 3-8 以 TSN 为视频主干的 ImagineNet 复合错误识别性能探究

Model	Variants	GFLOPs	mAP	mmit mAP
ImagineNet-FC	FC	0.001	0.6259	0.6893
ImagineNet-SA	SA	0.068	0.6426	0.7049
	SAx2	0.136	**0.6450**	**0.7131**
	SAx3	0.203	0.6436	0.7086
	w/o PosEmb	0.068	0.6305	0.6906
ImagineNet-CA	CA	0.068	0.6307	0.6933
	CA+SA	0.136	**0.6347**	0.7005
	CA+SAx2	0.203	0.6335	**0.7046**
	w/o PosEmb	0.068	0.6281	0.6953

表 3-9 以 TSM 为视频主干的 ImagineNet 复合错误识别性能探究

Model	Variants	GFLOPs	mAP	mmit mAP
ImagineNet-FC	FC	0.001	0.7053	0.7566
ImagineNet-SA	SA	0.068	**0.7011**	0.7630
	SAx2	0.136	0.7007	**0.7656**
	SAx3	0.203	0.6995	0.7572
	w/o PosEmb	0.068	0.6822	0.7593
ImagineNet-CA	CA	0.068	0.6752	0.7346
	CA+SA	0.136	**0.6766**	**0.7406**
	CA+SAx2	0.203	0.6728	0.7377
	w/o PosEmb	0.068	0.6725	0.7339

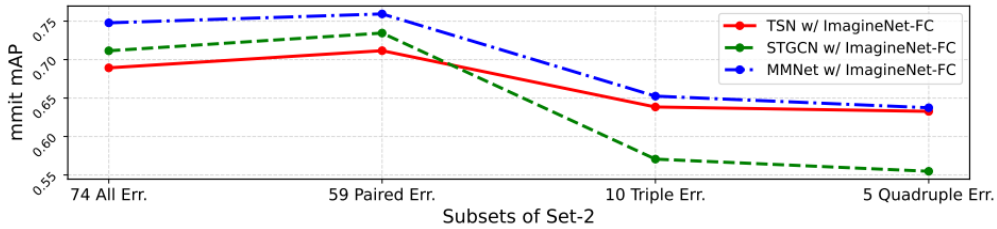


图 3-13 Set-2 中不同子集 mmit mAP 性能对比

果显示模型的性能随着复合错误种类数量的增加而逐渐降低，这是符合直观认知的：越多的复合错误种类意味着越高的辨识难度。在三错误与四错误复合场景中，ST-GCN 主干的性能要明显差于 TSN 主干，这说明 2D 骨骼关键点序列相较于视频序列会丢失掉一部分辨识细节。MMNet 主干网络在所有设定中取得了稳定的性能，这是因为多模态信息之间具备互补性。

在之前实验中，ImagineNet 的两个输入特征均属于同一个模态。由于 ImagineNet 的结构支持多模态数据的融合，所以本文对模态融合的性能进行了对比探究。为探究不同特征聚合方式对复合错识别性能的影响，本文选取 Two-Stream[31]、CBP[153]、BLOCK[154]和 MMNet[150]作为对比方法。基本主干网络使用 TSM[126]和 ST-GCN[124]分别进行 RGB 模态和 2D 关键点模态的特征提取。本文对所有特征融合策略的运行时间进行了统计，计算方法为对运行 1000 次的总耗时取平均。延时与复合错误识别性能对别列举于表 3-10 中。结果显示 ImagineNet-

CA 超越了其他四种多模态信息的融合方法。虽然 BLOCK 达到了和 ImagineNet-CA 相近的性能，但是延时却是 ImagineNet-CA 的近 8 倍。较高的延时主要是双线性模型需要计算两个特征之间的近似外积导致的。Two-Stream 融合策略能够显著降低延时，但是融合性能较差。

表 3-10　RGB 信息与 2D 关键点信息的多模态模型性能对比

Model	Modality	Latency (ms)↓	mAP	mmit mAP
TSM[126]	RGB	---	0.5662	0.6618
ST-GCN[124]	Pose	---	0.5776	0.6692
Two-Stream[31]	RGB+Pose	**0.1501**	0.6003	0.6815
CBP[153]	RGB+Pose	0.3043	0.7089	0.7506
BLOCK[154]	RGB+Pose	1.294	0.7107	**0.7675**
MMNet[150]	RGB+Pose+RoI	0.2479	0.6927	0.7478
w/ ImagineNet-CA	RGB+Pose	0.1642	**0.7110**	0.7515

3.5.3 消融实验

本文以 TSN[4]、TSM[126]和 ST-GCN[124]为主干网络对提出的线性加权特征聚合机制进行了消融实验，并且引入了 CBP[153]方法与 BLOCK[154]方法作为对比策略。消融实验结果显示，相较于对多个特征进行简单求和，基于随机线性加权的特征聚合机制能够生成丰富的特征组合，从而有效地提升模型的复合错误识别性能。对比结果显示，CBP 方法与 BLOCK 方法所带来的性能提升低于线性加权聚合机制。由表 3-11 可知，CBP 与 BLOCK 方法还会引入额外的计算量。结果证实本文所提出的基于随机线性加权的特征聚合机制具有高效性与有效性。

本文在构建 CPR-Coach 数据集时采用了四个摄像头对 CPR 胸外按压视频进行同时采集。多视角的设定能够辅助确定视角数量与模型性能的关系，从而支持模型计算量与识别精度平衡的探究。在实际部署中，使用全部视角视频会造成冗余计算量的引入。图 3-14 选取 TSN 和 ST-GCN 视频网络为主干，对 RGB 模态和 2D 关键点模态下的“视角组合-复合错误识别性能”关系进行了探究。不同视角信息的融合方式为：在 ImagineNet 的最后一个线性层输出进行求和。结果显示在所有模态中，复合错误识别性能均随视角数量的增加而增加，这与人的认知是一致的：越多的视角能够提供越多的互补信息，从而实现更加稳定的复合错误识别。图 3-14 中的单视角实验结果表明，在所有的四个视角中，视角 3 提供了更有价值的分辨信息，而视角 4 则正好相反。此结论对于后续系统的优化与部署具有重要应用价值。图 3-15 对四个视角的识别结果进行了可视化展示。由于视角 3 摄像头摆放位置在被试者的正前方，因此具有最佳视野，最终为错误行为的识别提供了重要特征信息；而视角 4 摄像头的摆放位置在被试者的侧方，会丢失一部分行为细节，从而造成判断信息的遗漏。

表 3-11 随机线性加权特征聚合机制的消融与对比实验

Model	Agg-1	Agg-2	Agg-3	mAP	mmit mAP
TSN[4]	—	—	—	0.5598	0.6143
w/ ImagineNet-FC	×	×	×	0.6198	0.6738
	✓	×	×	**0.6259**	**0.6893**
	×	✓	×	0.6019	0.6775
	×	×	✓	0.6033	0.6725
TSM[126]	—	—	—	0.5662	0.6618
w/ ImagineNet-FC	×	×	×	0.6871	0.7353
	✓	×	×	**0.7053**	**0.7566**
	×	✓	×	0.6434	0.7308
	×	×	✓	0.6569	0.7219
ST-GCN[124]	—	—	—	0.5776	0.6692
w/ ImagineNet-FC	×	×	×	0.6374	0.7089
	✓	×	×	**0.6404**	**0.7115**
	×	✓	×	0.5783	0.6877
	×	×	✓	0.6159	0.6864

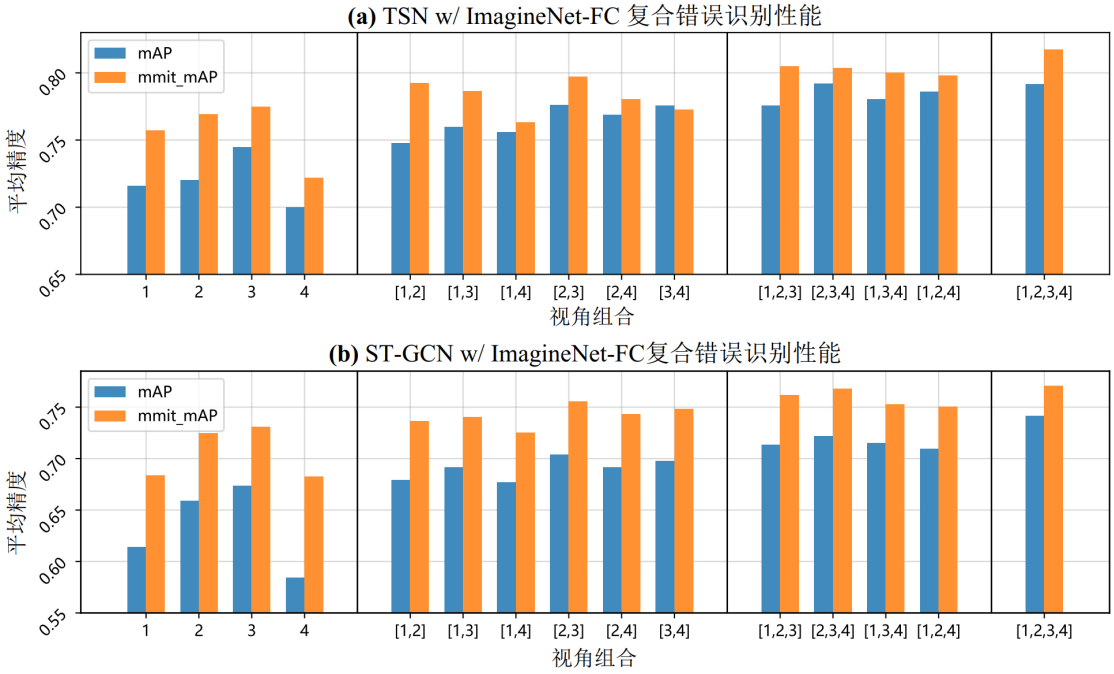
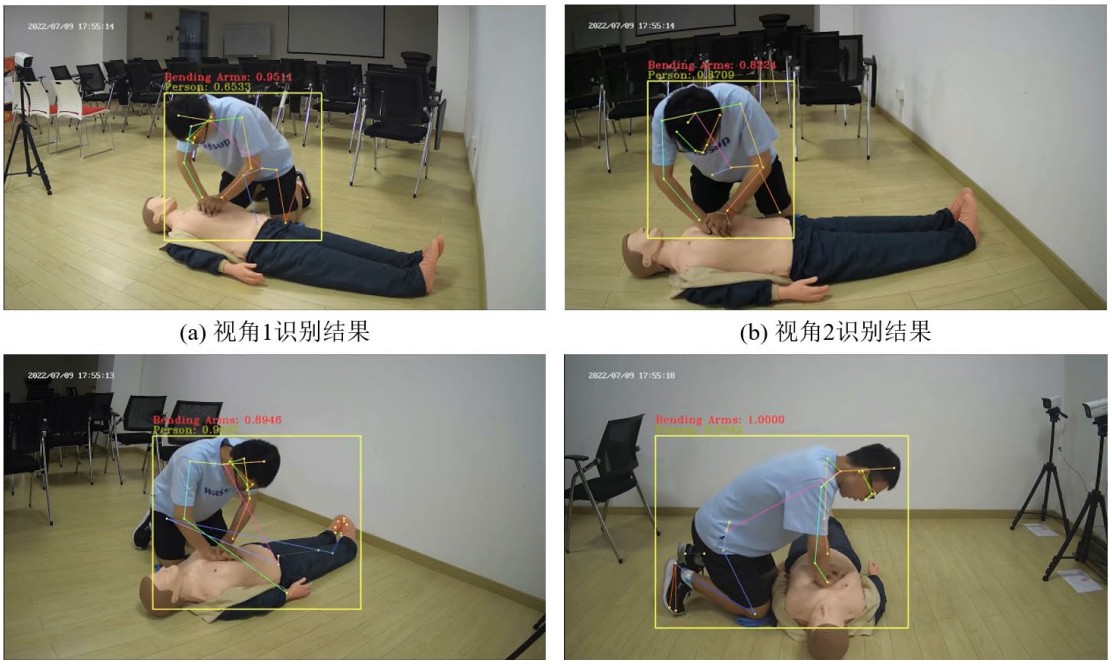


图 3-14 多视角设定下的复合错误识别结果

3.5.4 定性对比与可视化结果

为深入探究 ImagineNet 框架提升复合错误行为识别能力的机制，本文使用 t-SNE 算法对 Set-2 中的视频特征进行了可视化。图 3-17 的第一行分别展示了 TSN、TSM、TPN 在朴素迁移策略下生成的视频特征；第二行展示了这些视频主干在经过 ImagineNet-FC 训练后生成的视频特征。黑色虚线是人为绘制的边界曲线。在朴素的迁移策略下，经过模型映射后的测试集特征分布较为混乱，而 ImagineNet 框架能够有效地减小错误行为的类内间距、扩大类间间距，从而使得特征分布出现明显的分区特性，这充分证实了 ImagineNet 训练策略的有效性。

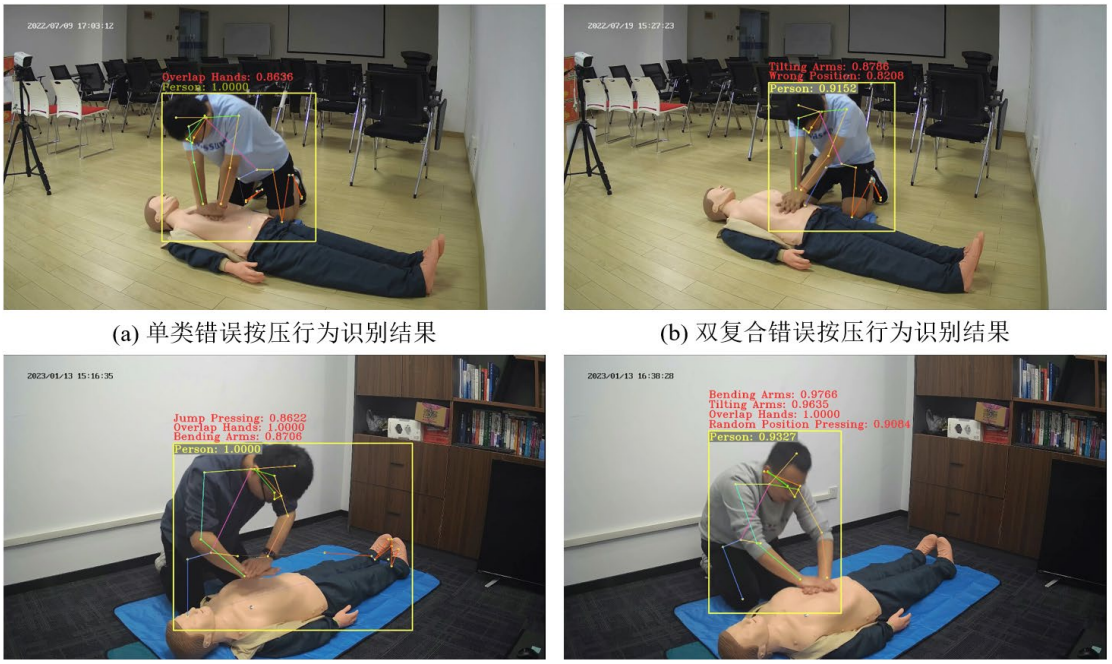
图 3-16 对单错误和三种复合错误案例的识别结果进行了可视化，包括骨架和错误类别的概率等信息。预测模型为 ImagineNet-CA，输入特征分别通过 TSM 和 ST-GCN 进行提取。可视化结果显示，本章所提出的 ImagineNet 能够在各种错误复合的场景中准确地对各类错误进行识别。



(a) 视角1识别结果　　(b) 视角2识别结果

(c) 视角3识别结果　　(d) 视角4识别结果

图 3-15　四个视角下的识别结果展示图



(a) 单类错误按压行为识别结果　　(b) 双复合错误按压行为识别结果

(c) 三错误复合识别结果　　(d) 四错误复合识别结果

图 3-16　单错误与复合错误识别结果

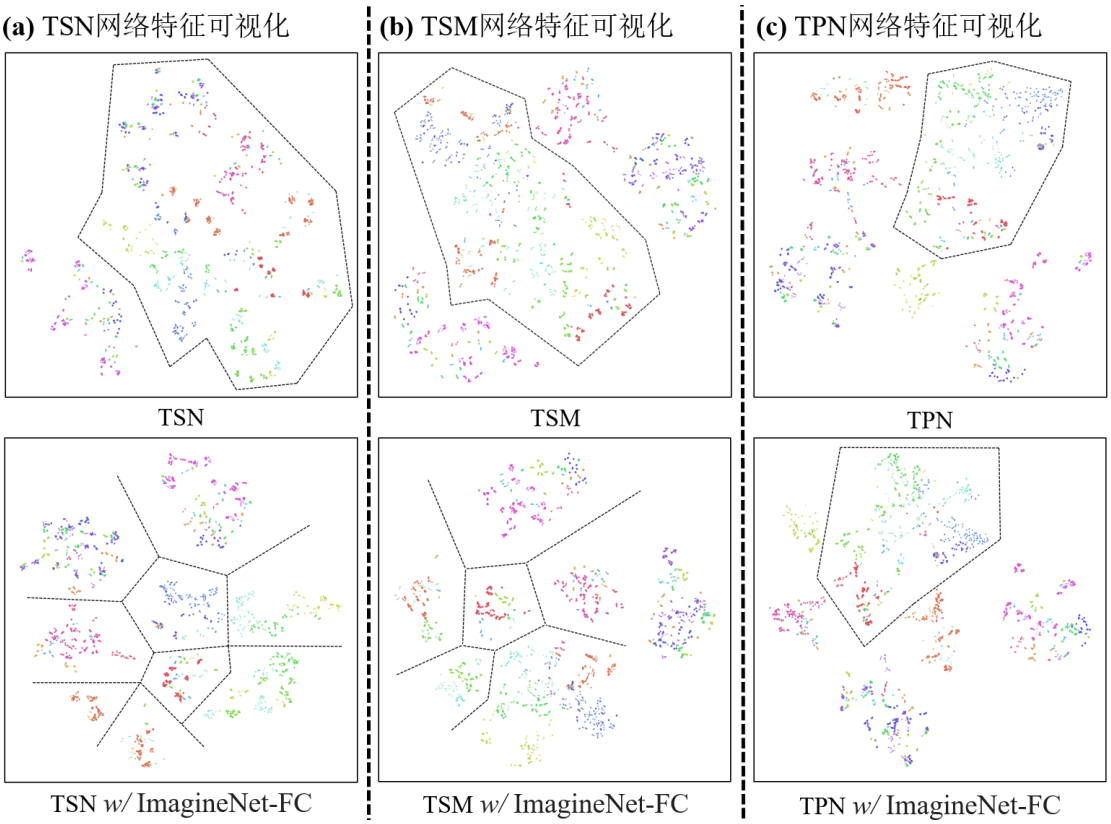


图 3-17 t-SNE 特征可视化对比

3.6 本章小结

本章针对目前医疗技能评估领域中细粒度错误行为识别研究匮乏的问题，首次提出了复合错误行为识别任务范式，并对 CPR 中的胸外按压行为进行了深入探究，构建了细粒度错误行为识别数据集 CPR-Coach。在专业医生的指导下，本章细化了按压过程中的 13 类单错误行为和 74 类复合错误行为。在算法方面，本章针对现实医疗技能评估模型所面临的“单类训练，多类测试”现象，提出了 ImagineNet 框架。该框架能够有效地缓解由训练集与测试集数据分布差异过大所引起的模型识别性能低下问题。在实验部分，本章在 CPR-Coach 数据集上对现有行为识别模型和 ImagineNet 框架进行了性能测试。实验结果充分证实了框架的有效性。

第4章 基于多模态预训练机制的复合错误行为识别算法

4.1 引言

第三章对心肺复苏术中的复合错误行为识别问题进行了定义和探究，尽管提出的 ImagineNet 框架在“单类训练，多类测试”的模式下取得了一定的性能提升，但是现有系统仍然存在两方面的问题：一方面，现有复合错误识别模型的输入模态信息有限，虽然 CPR-Coach 数据集能够提供 RGB 视频、2D 关键点和光流图像三种模态信息，但这些信息均隶属于视觉模态的范畴，并未囊括语言模态信息；另一方面，针对 ImagineNet 框架所开展的系列实验均属于理论层次的识别性能探究，并未在模型的易用性和实用性方面进行改进。这些因素造成了技能评估模型与实际落地应用的较大差距。本章拟通过语言模态信息的引入有效提升医疗技能评估模型在复合错误识别任务中的性能，同时对技能评估系统的人机交互特性进行显著改善。

目前各个国家均面临着医学生培养周期过长、培养成本高昂等难题，几乎所有的医疗服务机构与技能培训中心都十分重视医疗教学中的“降本增效”问题：即如何在保证培训质量的前提下，尽可能地降低人力物力成本。目前学界中已有系列开创性的研究对智能化医疗技能评估问题进行了探究，虽然这些研究在数据集构建和算法设计方面取得了一定成就，但是仍然停留在算法研究的理论层面，并没有在改善系统实用性方面进行探究与创新。为确保性能评估阶段的简便，现有的医疗技能评估系统模型的输入输出通常采用“输入视频—输出结果”的简单映射形式。这种具有硬性输入输出形式的模型通常难以直接应用在真实的医疗技能评估场景中，模型的可扩展性和交互性均有较大的改善空间。总体而言，现有的智能医疗技能评估系统与实际应用之间依然存在较大的鸿沟。

目前多模态学习方法已被广泛应用于语言—图像生成（Language-Image Generation）、视觉问题回答（Visual Question Answering, VQA）、多模态感知融合等各类学习任务。此类研究重点探究了多模态特征对齐、模态间信息关联挖掘和多模态信息融合等问题。对比语言—图像预训练框架[155]（Contrastive Language-Image Pre-Training, CLIP）是多模态预训练对比学习中的开创性工作。通过最小化单个批次内的样本对比损失，预训练过程赋予了 CLIP 框架对齐图像特征与文本特征的能力，从而使得 CLIP 框架在图像识别与零样本学习任务中取得了优异的性能。得益于较强的泛化能力和鲁棒性，CLIP 框架现已逐渐被引入到其他计算机视觉任务中，例如目标检测、语义分割和视频行为识别。多模态学习框架为

改善医疗技能评估系统实用性较差的问题提供了重要解决思路。本文将语言模态信息引入到复合错误行为识别任务中。

提示词工程（Prompt Engineering）是自然语言处理领域中的重要技术。其最初提出目的是解决预训练模型在适配下游任务时需要额外进行数据集标注与模型重新训练的问题。提示词工程在构造语言模态信息、提高模态丰富性和提升对话系统回答质量中起到了重要作用，现已被广泛应用于跨模态学习任务与大语言模型（Large Language Model, LLM）技术中。现有的视觉—语言预训练模型均使用了提示词工程，例如 CLIP 框架、基于多模态信息的文本生成模型与文本—图像生成模型。本文将提示词工程引入到复合错误行为识别任务中，通过提示词工程实现了对医疗技能评估模型的输入模态信息扩充。

本章在第三章的基础上对心肺复苏场景中的复合错误行为识别框架进行了进一步优化，并实现了系统的实际部署测试与随机对照试验开展。受启发于 CLIP 框架，本章提出了一种基于提示词工程的多模态预训练框架 CPR-CLIP，用于提升复合错误行为识别模型的精度，同时改善传统技能评估系统的人机交互能力弱、易用性较差等问题。为丰富模型的语言模态信息，CPR-CLIP 框架首先从错误数量、错误种类、改正建议三个不同的角度完成提示语句构建，实现对复合错误信息的准确描述。其次，CPR-CLIP 框架通过最小化对比预训练损失实现了语言特征与视觉特征的对齐，从而提升了模型的复合错误识别性能。在实验部分，本章首先在第三章构建所的 CPR-Coach 数据集上验证了 CPR-CLIP 框架的有效性，并将 CPR-CLIP 框架封装为一个能够通过自然语言进行智能检索与批量评估的电子助手，并招募了四名医生对系统的实际辅助能力进行评估。实验结果充分证实了语言模态与视觉模态之间的信息互补性和系统的实际效用。

本章的主要贡献概括如下：

1、受启发于多模态预训练框架 CLIP，本章提出了基于提示词工程的多模态预训练框架 CPR-CLIP，用于心肺复苏场景中的复合错误行为识别任务。

2、本章在复合错误行为识别基准 CPR-Coach 上开展了性能对比实验，实验结果证实：视觉模态和语言模态信息之间存在互补特性，多模态预训练过程能够有效提升模型在复合错误识别任务中的性能。

3、本章对 CPR-CLIP 框架进行了实际部署并开展了随机对照试验，结果显示该框架能够在不损失评估精度的前提下带来近 4 倍的评估效率提升，实现了系统在真实医疗技能评估应用中的有效性验证。

4.2 多模态对比预训练与提示词工程

4.2.1 多模态对比预训练框架

CLIP[155]（Contrastive Language-Image Pre-training）是由 OpenAI 公司于 2021 年提出的一种多模态对比预训练框架。如图 4-1(a)所示，CLIP 框架由三个部分构成：图像编码器、文本编码器和对比损失计算。其中图像编码器通常由传统的卷积神经网络或近期的视觉 Transformer 主干模型构成；文本编码器通常由自然语言处理模型中的 Transformer 构成。在对比预训练过程中，首先需要构造一个含有 N 个图像—文本对（Image-Text Pair）的训练批次，通过不同模态样本之间的对应关系构造出 $N \times N$ 的对应矩阵，其中对角线内的 N 个元素为正例，非对角线的 $(N^2 - N)$ 个元素为负例。为保障预训练阶段的有效性，OpenAI 公司从互联网上收集并整理了含有 4 亿个高质量图像—文本对的 WIT（Web Image Text）数据集以保证语料的丰富度。

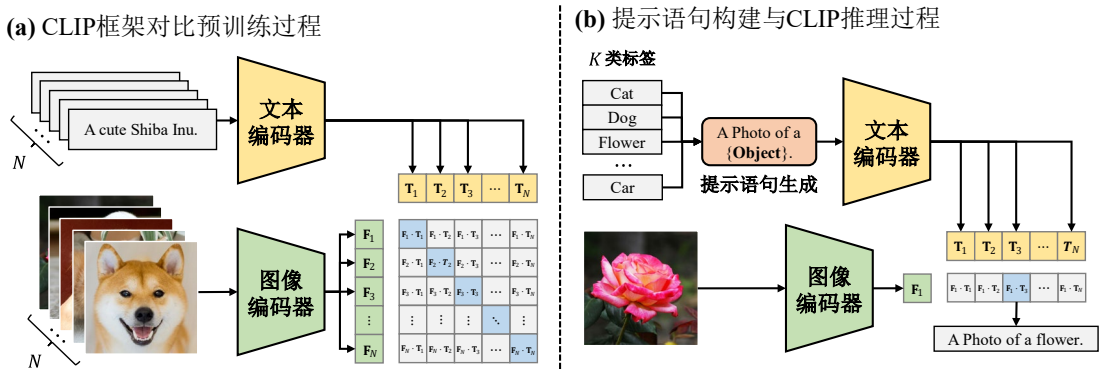


图 4-1 CLIP 框架的对比预训练过程与推理过程

对于图像编码器，预训练过程的目标是最小化损失 $\mathcal{L}_{img}$：

$$\mathcal{L}_{img} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\left(\mathbf{F}_i, \mathbf{T}_j\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(\mathbf{F}_i, \mathbf{T}_j\right)/\tau\right)} \tag{4.1}$$

文本编码器与图像编码器类似，预训练过程的目标是最小化损失 $\mathcal{L}_{txt}$：

$$\mathcal{L}_{txt} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\left(\mathbf{T}_i, \mathbf{F}_j\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(\mathbf{T}_i, \mathbf{F}_j\right)/\tau\right)} \tag{4.2}$$

其中 $\text{sim}(\cdot,\cdot)$ 表示两个不同模态特征之间的余弦相似度，$\mathbf{F}_i \in \mathbb{R}^D$ 表示批次内第 i 个图像的特征，$\mathbf{T}_j \in \mathbb{R}^D$ 表示批次内第 j 个文本的嵌入特征，D 为特征向量的维度，τ 为可学习的温度参数。

最终跨模态对比预训练过程的目标是最小化 $\mathcal{L}_{img}$ 与 $\mathcal{L}_{txt}$ 的均值：

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{img} + \mathcal{L}_{txt}\right) \tag{4.3}$$

CLIP 框架的推理过程如图 4-1(b)所示，测试图像经过图像编码器映射为特征$\mathbf{f}\in\mathbb{R}^D$，给定所有$K$个类别的提示语句特征集合$\{\mathbf{T}_i\}_{i=1}^{K},\mathbf{T}_i\in\mathbb{R}^D$，模型对各个类别的预测概率计算方式为：

$$P(Y=i|\mathbf{f})=\frac{\exp\left(sim\left(\mathbf{f},\mathbf{T}_i\right)/\tau\right)}{\sum_{j=1}^{K}\exp\left(sim\left(\mathbf{f},\mathbf{T}_j\right)/\tau\right)} \tag{4.4}$$

其中K个类别的提示语句形式为“A photo of {class}.”，温度参数τ由预训练阶段学习得到。最终 CLIP 框架对图像的预测类别为：

$$Y=\arg\max_{i}\ P(Y=i|\mathbf{f}) \tag{4.5}$$

CLIP 框架推理过程中的类别数量K与训练过程中的类别数量C可以不一致。因此与传统的闭集（Closed-Set）分类问题不同，CLIP 框架能够在开集（Open-Set）设定下进行类别判断，即在无需任何重新训练或微调的情况下支持零样本检测任务（Zero-shot Detection）。

得益于多模态对比预训练机制与充足的训练语料，CLIP 框架在图像分类和零样本分类任务中均取得了优异且鲁棒的性能。研究者将 CLIP 框架逐渐引入到其他图像任务中，并提出了系列模型：例如目标检测中的 ViLD[156]、GLIP[157]、；图像分割中的 CLIP-S4[158]、GroupViT[159]；图像生成中的 CLIPasso[160]、CLIP Draw[161]。以上模型将 CLIP 框架的跨模态匹配机制利用到不同的视觉模型中，有效弥补了传统单模态模型的不足，从而在各个计算机视觉任务中均取得了优良的性能。

在图像识别任务之外，研究者们还将 CLIP 框架引入到视频行为理解任务中：Wang 等人[162]提出了 ActionCLIP 框架，将视频动作分类问题转化为“视频—文本”匹配问题。ActionCLIP 框架共分为三个阶段：提示词构建阶段、预训练阶段与模型微调阶段。

在提示词构建阶段中，作者提出了丰富的提示语句模板对标签信息进行描述；在预训练过程中，ActionCLIP 模型获得了对齐语言模态特征与视觉模态特征的能力；在微调阶段中，ActionCLIP 框架以端到端的形式在行为识别数据集中进行参数微调。由于引入了丰富的语言模态信息，ActionCLIP 框架在 Kinetics-400[24]行为识别数据集中取得了优良性能。受启发于 ActionCLIP 框架，Li 等人[90]首次将 CLIP 框架和提示词工程技术引入到视频时序行为分割任务中，并提出了 Br-Prompt 框架。此框架旨在弥补传统行为识别模型在时序信息捕获上的缺点。具体而言，传统行为识别模型往往只关注于单个行为，并没有对时序上下文信息进行建模，从而会造成潜在的时序逻辑信息丢失。Br-Prompt 框架首先对指令视频中的连续行为进行截取，并分别构建了“断章取义”信息（Out-of-context Information）

和上下文信息（Contextual Information）。这两类信息会以文本提示的形式与视频片段相互对应，最终通过 CLIP 框架的跨模态预训练机制在语义空间中实现特征对齐。得益于丰富的时序上下文信息和对比预训练过程，Br-Prompt 框架在时序行为分割任务中取得了最优性能。

4.2.2 提示词工程

提示词工程（Prompt Engineering）最初起源于自然语言处理领域，是一种应用在大规模预训练模型（Large Pre-trained Models）领域中的技术。早期的自然语言处理任务相互独立，每个任务均都有专属的数据集与算法，而随着 BERT[105]、GPT[106]和 T5[107]等预训练语言模型的不断发展，各自然语言处理任务之间的壁垒逐渐被打破。这些语言模型在大规模语料库上完成预训练之后，还需要在额外的数据集上进行参数微调以适配不同的下游任务（Downstream Tasks）。这种“预训练—下游适配”的模式面临着两方面的问题：在数据收集方面，需要构建具有一定规模和体量的下游任务数据集以确保模型在迁移后的性能；在模型微调方面，无论是对预训练语言模型进行全部参数微调（Training From Scratch）还是部分参数微调（Parameter-efficient Finetuning），都需要一定的算力支持，整个过程费时费力。因此预训练模型在下游任务中的性能严重依赖于适配数据集的体量和参数微调过程的质量。

针对以上问题，提示词工程应运而生。其目的是完全规避掉模型参数变动环节，充分利用预训练大模型（又称为基础模型）内部的知识完成新的任务。提示词工程的一般做法是在模型的输入中添加与新任务相关的暗示（Hints）内容。以多模态预训练框架 CLIP 在零样本检测任务中的应用为例，预训练过程中 CLIP 框架的自监督信息来源于“图像—文本”的对应关系，而测试过程中 CLIP 框架仅通过更改输入文本“A photo of {Class}.”即可切换模型的检测功能。这些额外添加的提示信息可以由人直接通过自然语言设计得到，也可以通过特征向量的形式由模型自动生成。提示词工程在下游任务迁移中完全规避了传统“预训练—下游适配”模式的缺点：一方面，只需要少量的标注数据就能完成迁移任务，不再需要收集大规模下游任务数据集；另一方面，不再需要对模型参数进行重新训练，从而节省大量算力成本。目前提示词工程已被广泛应用于大语言模型技术中。

Gu 等人[163]将现有的提示词方法按照提示形式划分为两个类别：硬性提示（Hard Prompt）与软性提示（Soft Prompt）。硬性提示又被称为离散提示（Discrete Prompt），是指提示内容以自然语言形式的人工指令或任务案例呈现。目前学界中的硬性提示研究主要包含四类：任务指令（Task Instruction）、上下文学习（In-Context Learning）、检索提示（Retrieval-based Prompting）和思维链提示（Chain-

of-Thought）；软性提示又被称为连续提示（Continuous Prompt），是指提示内容以特征向量的形式呈现，并在模型训练过程中使用梯度下降等算法进行优化。软性提示研究在学界中又被称为提示微调（Prompt-Tuning）。

提示词工程在自然语言处理任务中取得了优异的性能，研究者将其引入到了多模态学习领域，并提出了系列“视觉—语言”预训练模型。如图 4-2 所示，现有多模态预训练模型按照任务形式可划分为三个类别：图像—文本匹配模型、基于多模态的文本生成模型和文本—图像生成模型。图像—文本匹配模型以 CLIP 和 ALIGN[164]为代表，通过“图像—文本对”的对应关系构造对比损失实现模型预训练；基于多模态的文本生成模型囊括多种跨模态任务，例如视觉问题回答（VQA）、视觉常识推理、基于多模态信息的问答系统等。模型需要同时结合图像模态和语言模态的输入进行内容生成；文本—图像生成模型根据人类自然语言描述进行图像生成，例如 DALL-E[165]。

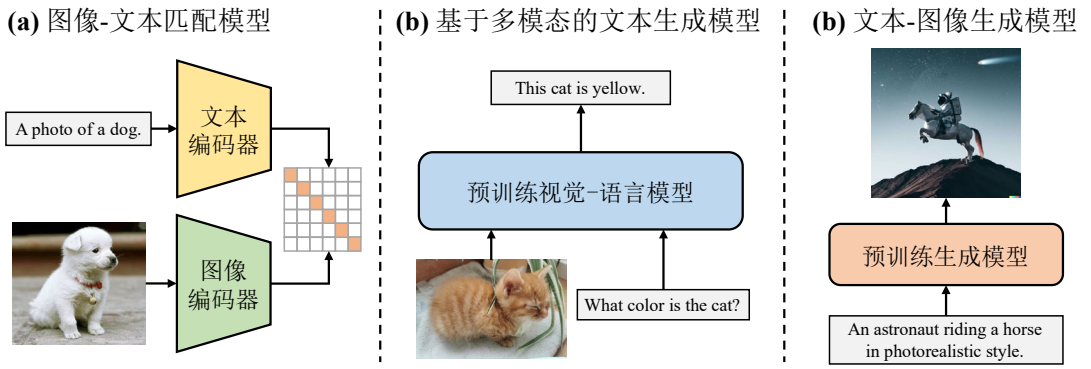


图 4-2　视觉—语言预训练模型分类

本文所提出的 CPR-CLIP 框架隶属于图像—文本匹配模型。CPR-CLIP 框架将提示词工程中的方法引入到心肺复苏复合错误识别任务中，通过构建错误数量、错误种类与改正建议三种类型的提示语句对复合错误信息进行描述，有效扩充了模型的模态信息丰富度。其次，通过跨模态对比预训练过程赋予模型对齐不同来源特征的能力，从而使模型具备更优良的复合错误识别性能与更强的交互应用能力。

4.3 基于多模态预训练机制的复合错误识别算法

本文在第三章中提出了心肺复苏场景下的复合错误识别任务，构建了复合错误行为识别数据集 CPR-Coach，并提出基于特征组合机制的复合错误识别算法 ImagineNet。如图 4-3 所示，复合错误识别任务的形式为：给定一个仅包含单类错误样本的训练集，模型需要对复合错误样本进行种类预测。这种监督信息受限（Restricted Supervision Conditions）的设定在医疗行为分析任务中非常普遍，因

为单独地为每一种复合错误情况进行采集是不可行的。第三章中的朴素迁移实验结果表明，传统的行为识别模型无法在这种监督信息严重受限的设定下取得稳定的性能。第三章所提出的 ImagineNet 算法虽然从一定程度上缓解了训练集与测试集之间的域迁移问题，但是模型的模态信息丰富度与识别性能仍有较大的提升空间。

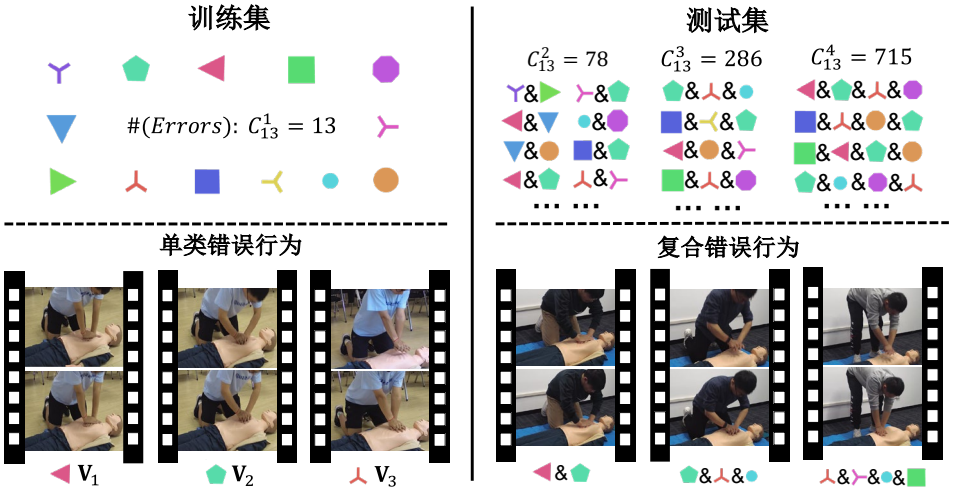


图 4-3 心肺复苏场景下的复合错误行为识别任务形式

为解决 ImagineNet 算法所面临的问题，本章将多模态对比预训练框架 CLIP 与提示词工程引入到复合错误行为识别任务中，提出了多模态预训练框架 CPR-CLIP。此框架的设计动机来自对人类自然语言优势的观察：考虑到语言能够自然地对复合错误进行准确描述，本文设计了三类提示语句作为多模态预训练框架中的文本模态输入，并通过预训练过程将语言特征与增强后的视觉特征在语义空间中进行对齐，最终实现模型泛化能力的提升与交互能力的增强。

此节内容组织如下：4.3.1 小节对 CPR-CLIP 框架进行整体介绍；4.3.2 小节对多模态预训练损失计算的过程进行描述；4.3.3 小节与 4.3.4 小节分别对 CPR-CLIP 框架的两种推理模式进行介绍：单视频预测模式与特定类别视频检索模式。

4.3.1 多模态预训练框架 CPR-CLIP

CPR-CLIP 多模态预训练框架的结构如图 4-4 所示，可划分为三个阶段：视频特征提取阶段、提示语句构造与嵌入阶段、对比损失计算阶段。三个阶段分别对应视觉通路（Visual Pathway）、语言通路（Language Pathway）、损失计算部分。清晰起见，图 4-4 对三个部分使用不同颜色进行区分。

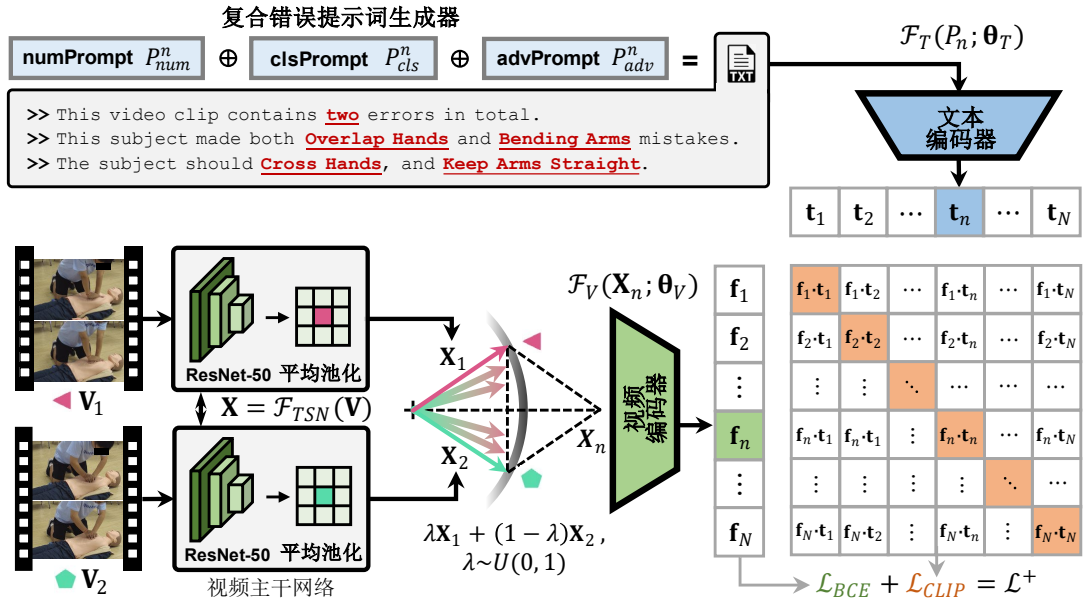


图 4-4　CPR-CLIP 框架的多模态预训练过程

视频特征提取阶段

以双错误复合为案例，首先从单错误样本数据集中采样出两个来自不同类别的错误样本$(\mathbf{V}_1,C_1)$与$(\mathbf{V}_2,C_2)$，满足$C_1\neq C_2$，其中$\mathbf{V}_1=\{I_i\}_{i=1}^{L_1}$，$\mathbf{V}_2=\{I_i\}_{i=1}^{L_2}$，$L_1$与$L_2$分别表示两个视频中含有的帧数。通过视频主干网络可将视频帧映射为特征向量，以 TSN 视频网络为例，映射过程可表示为：

$$\mathbf{X}=\mathcal{F}_{TSN}(\mathbf{V};\boldsymbol{\theta}_{TSN}),\mathbf{X}\in\mathbb{R}^{T\times D} \tag{4.6}$$

其中T表示从初始视频中采样的视频片段数量，D表示特征维度，$\boldsymbol{\theta}_{TSN}$表示视频主干网络的可训练参数。

在特征融合阶段，本章使用与第三章相同的随机线性组合特征增强策略，通过时序平均池化操作（Temporal Average Pooling）获取视频特征：

$$\mathbf{X}_n=\frac{1}{T}\sum_{t=1}^{T}(\lambda\mathbf{X}_1^t+(1-\lambda)\mathbf{X}_2^t),\ \mathbf{X}_n\in\mathbb{R}^D \tag{4.7}$$

其中$\mathbf{X}_1^t$表示视频特征$\mathbf{X}_1$的第t个维度。融合后的特征$\mathbf{X}_n$通过视频特征编码器（Video Feature Encoder）$\mathcal{F}_V(\cdot)$映射为最终的视频特征：

$$\mathbf{f}_n=\mathcal{F}_V(\mathbf{X}_n;\boldsymbol{\theta}_V),\mathbf{f}_n\in\mathbb{R}^D \tag{4.8}$$

其中$\boldsymbol{\theta}_V$表示视频特征编码器中的可训练参数。本文使用多层感知机（MLP）对视频特征编码器进行实例化。

提示语句构造与嵌入阶段

人类的语言能够流利地对各种复合信息进行准确表述。受此启发，本章设计了一套提示语句模板（Prompt Templates）对心肺复苏过程中的复合错误信息进行描述。提示语句模板的具体内容如图 4-4 所示。图中的复合错误提示词生成器展示了双手重叠与手臂弯曲两个错误类别的提示语句构造过程。

提示语句模板分别从错误数量、错误种类和改正建议三个方面对复合错误进行了描述。三种类型的模板分别对应数量提示词P_{num}，种类提示词P_{cls}和建议提示词P_{adv}。以双错误组合为例，提示语句模板的具体定义如下：

$$P_{num}^{n} = \text{“}This\ video\ clip\ contains\ \{\boldsymbol{cnt}\}\ errors\ in\ total.\text{”}$$

$$P_{cls}^{n} = \text{“}This\ subject\ made\ both\ \{\boldsymbol{C}_1\}\ and\ \{\boldsymbol{C}_2\}\ mistakes.\text{”}$$

$$P_{adv}^{n} = \text{“}This\ subject\ should\ \{\boldsymbol{A}_1\}\ and\ \{\boldsymbol{A}_2\}.\text{”}$$

对应中文含义为：

$$P_{num}^{n} = \text{“此视频片段共包含}\{\boldsymbol{cnt}\}\text{种错误。”}$$

$$P_{cls}^{n} = \text{“此被试者犯了}\{\boldsymbol{C}_1\}\text{和}\{\boldsymbol{C}_2\}\text{错误。”}$$

$$P_{adv}^{n} = \text{“此被试者应当}\{\boldsymbol{A}_1\}\text{和}\{\boldsymbol{A}_2\}\text{。”}$$

其中$n \in \{1, 2, \cdots, N\}$表示单个训练批次中的样本序号；$\{cnt\}$表示复合错误的种类数量，会随着错误组合数量的增加而改变；$\{C_1\}$与$\{C_2\}$分别表示错误类别的名称；$\{A_1\}$与$\{A_2\}$分别表示错误操作所对应的改正建议。

三种类型的提示语句通过字符串拼接操作汇聚为最终的提示语句：

$$P_n = P_{num}^{n} \oplus P_{cls}^{n} \oplus P_{adv}^{n} \tag{4.9}$$

与视频通路的处理方式类似，提示语句P_n通过文本编码器（Text Encoder）$\mathcal{F}_T(\cdot)$映射为特征向量：

$$\mathbf{t}_n = \mathcal{F}_T(P_n; \mathbf{t}), \mathbf{t}_n \in \mathbb{R}^D \tag{4.10}$$

其中$\boldsymbol{\theta}_T$为文本编码器$\mathcal{F}_T(\cdot)$的可训练参数。在本文中，文本编码器的结构与 CLIP 框架中的语言通路保持一致，由 12 层 Transformer 层构成，特征维度设定为 512，注意力头数量设定为 8。

4.3.2 损失函数设计

CPR-CLIP 框架预训练过程的目标是：通过自监督预训练机制在语义空间中将视觉信息和语言信息进行对齐，从而获得更高的复合错误行为识别性能。本文采用 CLIP 损失函数实现两个模态之间特征的对齐。在含有N个样本的单个训练批次内部，给定视觉特征集合$\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N\}, \mathbf{f}_n \in \mathbb{R}^D$与文本特征集合$\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_N\}, \mathbf{t}_n \in \mathbb{R}^D$，使用余弦相似度（Cosine Similarity）对跨模态特征

相似度进行度量。两特征之间的余弦相似度计算过程为：

$$sim(\mathbf{f}_n,\mathbf{t}_n)=\frac{\mathbf{f}_n\cdot\mathbf{t}_n}{\|\mathbf{f}_n\|\,\|\mathbf{t}_n\|} \tag{4.11}$$

单个训练批次内部的相似度矩阵$S(\mathbf{F},\mathbf{T})\in\mathbb{R}^{N\times N}$定义为：

$$S(\mathbf{F},\mathbf{T})=\begin{bmatrix} sim(\mathbf{f}_1,\mathbf{t}_1) & \cdots & sim(\mathbf{f}_1,\mathbf{t}_N) \\ \vdots & \ddots & \vdots \\ sim(\mathbf{f}_N,\mathbf{t}_1) & \cdots & sim(\mathbf{f}_N,\mathbf{t}_N) \end{bmatrix} \tag{4.12}$$

对$S(\mathbf{F},\mathbf{T})$行和列分别进行 Softmax 归一化即可获得文本相似度矩阵（Text-wise Similarity Matrix）$S_T(\mathbf{F},\mathbf{T})\in\mathbb{R}^{N\times N}$与视频相似度矩阵（Video-wise Similarity Matrix）$S_V(\mathbf{F},\mathbf{T})\in\mathbb{R}^{N\times N}$。根据单个批次内视觉信息与文本信息标签的一致性，可以构造出跨模态标签矩阵$M_{GT}\in\mathbb{1}^{N\times N},\mathbb{1}=\{0,1\}$。在$M_{GT}$内，视频特征与文本特征属于同一类的元素设定为 1，其余元素设定为 0。

使用 KL 散度（Kullback–Leibler Divergence）对两个相似度矩阵和M_{GT}之间的差异进行度量，并取平均值即可计算出$\mathcal{L}_{CLIP}$：

$$\mathcal{L}_{CLIP}=\frac{1}{2}\left(\mathrm{KL}\left[S_T(\mathbf{F},\mathbf{T})\,\|\,M_{GT}\right]+\mathrm{KL}\left[S_V(\mathbf{F},\mathbf{T})\,\|\,M_{GT}\right]\right) \tag{4.13}$$

以文本相似度矩阵为例，KL 散度的计算方式为：

$$\mathrm{KL}\left[S_T(\mathbf{F},\mathbf{T})\,\|\,M_{GT}\right]=\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}S_T(\mathbf{F},\mathbf{T})\,[i,j]\log\frac{S_T(\mathbf{F},\mathbf{T})\,[i,j]}{M_{GT}[i,j]} \tag{4.14}$$

其中$[i,j]$分别表示$N\times N$矩阵中的第i行第j列元素。优化器的目标是寻找出最优的网络参数组合$(\boldsymbol{\theta}_V,\boldsymbol{\theta}_T)$使得 CLIP 损失最小化：

$$(\boldsymbol{\theta}_V^*,\boldsymbol{\theta}_T^*)=\arg\min_{(\boldsymbol{\theta}_V,\boldsymbol{\theta}_T)}\mathcal{L}_{CLIP} \tag{4.15}$$

虽然对比损失$\mathcal{L}_{CLIP}$能够在预训练阶段为 CPR-CLIP 模型提供监督信息，但是这种建立在单个批次内部的自监督信息仍然是不足的。因此本文继续沿用了第三章 ImagineNet 使用的二进制交叉熵损失$\mathcal{L}_{BCE}$为模型提供额外的强监督信息。具体实现方式如图 4-4 所示，通过在视觉通路的末尾添加线性层实现对错误种类的预测。对于单个测试样本，设网络对所有错误类别的预测分数为S，标准标签的独热编码形式表示为GT，则二进制交叉熵损失$\mathcal{L}_{BCE}$的计算过程为：

$$\mathcal{L}_{BCE}=-\frac{1}{C}\sum_{i=1}^{C}\left(GT[i]\cdot\log S[i]+(1-GT[i])\cdot\log(1-S[i])\right) \tag{4.16}$$

其中$S[i]$表示网络对第i个类别的预测分数，$GT[i]\in\{0,1\}$表示标准标签中

的类别指示信息，标准标签GT由C_1和C_2的独热编码表示形式取并集得到：$GT = \text{Onehot}(C_1) \cup \text{Onehot}(C_2)$。

本文将额外引入$\mathcal{L}_{BCE}$损失的模型命名为 CPR-CLIP+，对应的损失函数为：

$$\mathcal{L}^{+} = \mathcal{L}_{CLIP} + \mathcal{L}_{BCE} \tag{4.17}$$

在后续实验中，本文将会探究单独使用$\mathcal{L}_{CLIP}$与$\mathcal{L}^{+}$训练的模型性能差异。

4.3.3 单视频预测推理

文本所提出的 CPR-CLIP 框架共支持两种不同的推理模式：单视频预测模式与特定类别视频检索模式。两种推理模式适用于不同的使用场景。在单视频预测模式下，CPR-CLIP 框架的功能与第三章所提出的 ImagineNet 框架一致，即对目标视频中所包含的错误种类进行预测；在特定类别视频检索模式下，CPR-CLIP 框架受益于多模态预训练机制，能够支持使用自然语言对特定类别的视频进行检索。本小节将依次介绍这两种推理模式。

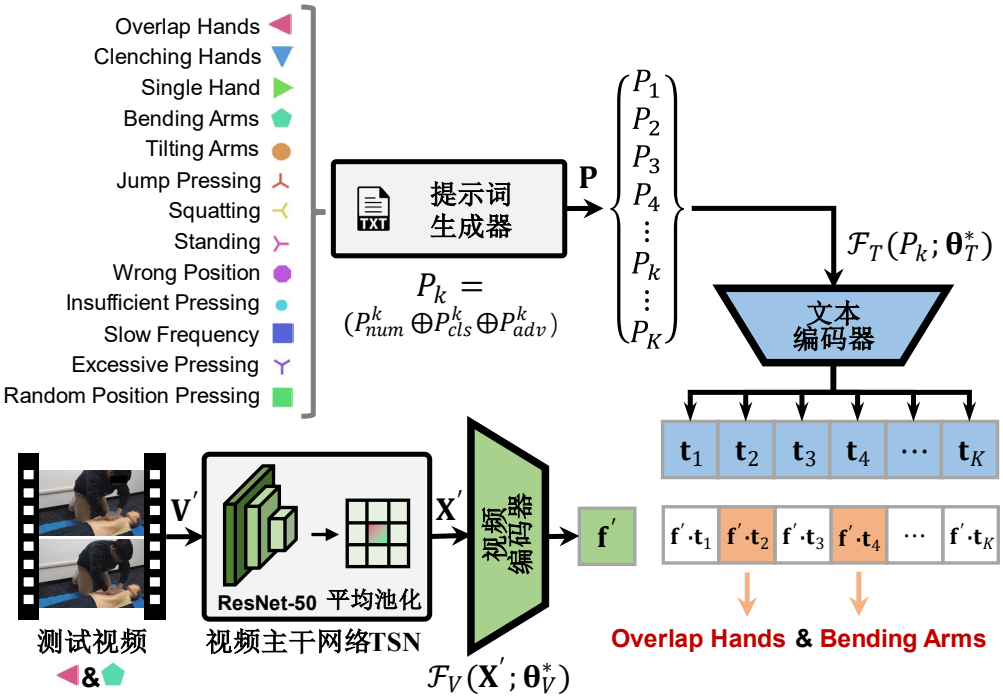


图 4-5 CPR-CLIP 框架的单视频预测模式推理过程

单视频预测模式下的推理过程如图 4-5 所示。在推理过程中，视频特征与文本特征相似度矩阵的计算替代了传统分类过程中对单个类别进行分数预测的过程。在文本通路中，设共有K类独立的错误种类。可以根据提示语句模板构造出单类提示语句集合$\mathbf{P} = \{P_k\}_{k=1}^{K}$。以第$k$个种类为例，提示语句的构造过程为：

$$P_k = P^k_{num} \oplus P^k_{cls} \oplus P^k_{adv} \tag{4.18}$$

通过文本编码器$\mathcal{F}_T(\cdot)$对提示语句集合$\mathbf{P}$进行特征提取，即可获取文本特征集合$\mathbf{T}=\{\mathbf{t}_k\}_{k=1}^{K}$，映射过程为：

$$\mathbf{t}_k=\mathcal{F}_T(P_k;\boldsymbol{\theta}_T^*),\mathbf{t}_k\in\mathbb{R}^D \tag{4.19}$$

在视频通路中，视频特征提取器将测试视频$\mathbf{V}'=\{I_i\}_{i=1}^{L}$映射为特征向量$\mathbf{X}'$，再通过视觉特征编码器$\mathcal{F}_V(\cdot)$映射为视觉特征：

$$\mathbf{f}'=\mathcal{F}_V(\mathbf{X}';\boldsymbol{\theta}_V^*),\mathbf{f}'\in\mathbb{R}^D \tag{4.20}$$

在单视频预测模式下的推理中，多模态预训练阶段所使用的视觉相似度矩阵$S_V(\mathbf{F},\mathbf{T})\in\mathbb{R}^{N\times N}$退化为一个$K$维行向量，表示测试视频$\mathbf{V}'$与提示语句特征集合$\mathbf{T}'$各元素之间的余弦相似度：

$$S_V(\mathbf{f}',\mathbf{T}')=[sim(\mathbf{f}',\mathbf{t}_1),\cdots,sim(\mathbf{f}',\mathbf{t}_K)]^T \tag{4.21}$$

4.3.4 特定类别视频检索推理

在 CPR-CLIP 的预训练过程中，最小化$\mathcal{L}_{CLIP}$损失能够赋予网络对齐文本特征与视频特征的能力。为充分利用这种多模态信息对齐能力，本文将 CPR-CLIP 框架引入到特定类别视频检索应用中，以提升复合错误识别系统的易用性。视频检索模式下的推理过程如图 4-6 所示。

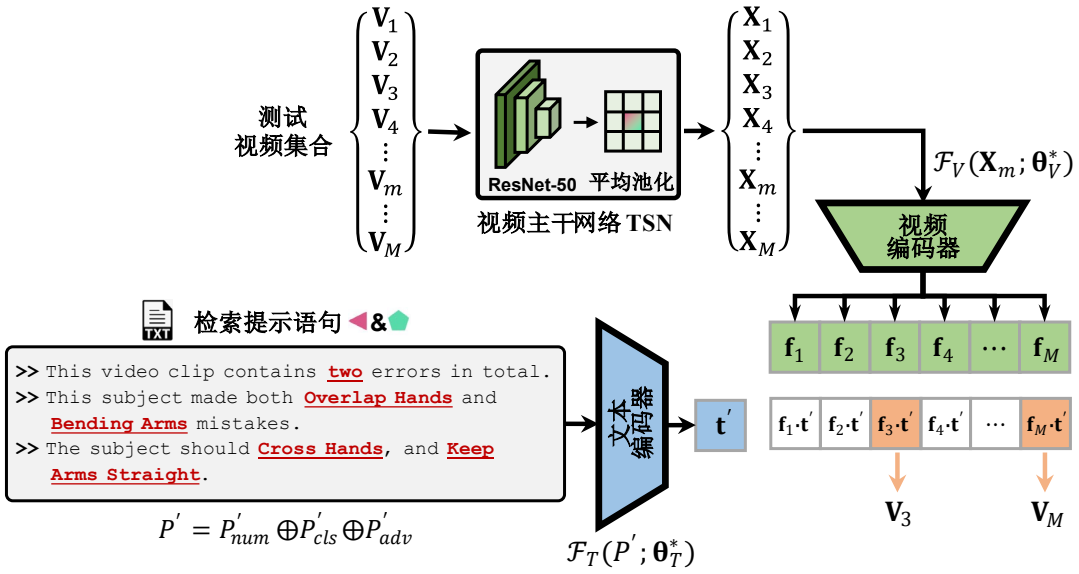


图 4-6　CPR-CLIP 框架的视频检索模式推理过程

给定包含M条视频的集合$\{\mathbf{V}_1,\mathbf{V}_2,\cdots,\mathbf{V}_M\}$，首先通过视频主干网络进行特征提取，生成视频特征集合$\{\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_M\}$。之后通过视觉特征编码器$\mathcal{F}_V(\cdot)$映射为视觉特征集合$\mathbf{F}=\{\mathbf{f}_1,\mathbf{f}_2,\cdots,\mathbf{f}_M\}$，对于特征$\mathbf{X}_m$，映射过程为：

$$\mathbf{f}_m = \mathcal{F}_V(\mathbf{X}_m;\boldsymbol{\theta}_V^*), \mathbf{f}_m \in \mathbb{R}^D \tag{4.22}$$

在语言通路中，首先需要使用者根据想要检索的错误类别构建检索提示语句。图 4-6 展示了双手重叠和手臂弯曲复合错误的提示语句P'构建过程：

$$P' = P'_{num} \oplus P'_{cls} \oplus P'_{adv} \tag{4.23}$$

通过文本编码器$\mathcal{F}_T(\cdot)$将提示语句P'映射为文本特征$\mathbf{t}'$：

$$\mathbf{t}' = \mathcal{F}_T(P';\boldsymbol{\theta}_T^*), \mathbf{t}' \in \mathbb{R}^D \tag{4.24}$$

在视频检索模式下的推理中，多模态预训练阶段所使用的文本相似度矩阵$S_T(\mathbf{F},\mathbf{T}) \in \mathbb{R}^{N\times N}$退化为一个$M$维列向量，表示目标提示语句特征$\mathbf{t}'$与视觉特征集合$\mathbf{F}$之间的相似度：

$$S_V(\mathbf{F},\mathbf{t}') = [sim(\mathbf{f}_1,\mathbf{t}'),\cdots,sim(\mathbf{f}_M,\mathbf{t}')] \tag{4.25}$$

在实际应用中，检索视频集合的M数量通常可达数千甚至数万，如果采用人工评判的方式对每个视频进行错误甄别会消耗巨大的人力成本。而 CPR-CLIP 框架能够支持以自然语言查询的方式对大量视频进行高效检索，将“逐个评估”的评估模式转化为“检索—审核”模式，从而有效节约技能评估过程中高昂的人力成本。

4.4 实验分析

4.4.1 模型与优化器设定

本章中的所有实验均在 AMD EPYC 7742@2.25GHz CPU 和 NVIDIA Tesla A800 GPU 平台上进行。视频主干网络的输入分辨率设定为224×244，参数经过 CPR-Coach Set-1 数据集上的单分类训练得到。在预训练过程中视频主干网络的参数被冻结。多模态对比预训练过程的轮次设定为 60，对应 32k 个训练批次。预训练过程的批次大小设定为$N=32$，使用随机梯度下降（SGD）优化器进行网络参数优化，初始学习率设定为 0.001，分别在第 20 轮和第 40 轮进行学习率衰减，衰减率为 0.1。视频特征编码器$\mathcal{F}_V(\cdot)$和文本特征编码器$\mathcal{F}_T(\cdot)$分别使用线性映射层和 Transformer 模型进行实例化。模型的复合错误行为识别性能使用 mAP 与 mmit mAP 进行度量。

4.4.2 性能对比实验

表 4-1 分别列举了以 TSN[4]、TSM[126]、ST-GCN[124]、ViViT[37]与 Video Swin Transformer[152]为视频主干网络的复合错误行为识别结果。每个部分的首行列举

了朴素迁移策略的识别精度；第二行列举了只使用$\mathcal{L}_{CLIP}$损失进行对比预训练的CPR-CLIP 框架性能；第三行列举了同时使用$\mathcal{L}_{CLIP}$对比损失与二进制交叉熵损失$\mathcal{L}_{BCE}$进行训练的 CPR-CLIP+框架性能。为对比性能差异，表 4-1 以朴素迁移策略为基线对 CPR-CLIP 框架带来的性能提升进行了标注。

结果显示，多模态对比预训练机制的引入能够带来明显的复合错误识别性能提升。以经典的视频网络 TSM 为主干，CPR-CLIP 框架带来了 7.39 百分点 mAP 提升和 4.56 百分点 mmit mAP 性能提升。若使用最先进的视频网络作为主干，则可以带来更高的性能提升。通过使用表示能力更强的 ViViT 或 Video Swin 网络进行视频特征提取，CPR-CLIP 会带来更高的复合错误识别性能提升：例如 CPR-CLIP 在 ViViT 主干上带来 9.21 百分点 mAP 和 8.43 百分点的 mmit mAP 提升；在 Video Swin 主干上带来 9.89 百分点 mAP 和 8.66 百分点的 mmit mAP 提升。而在表示能力较弱的 2D 关键点网络 ST-GCN 中，CPR-CLIP 框架的性能提升作用相对有限。这说明多模态对比预训练机制依赖于优质的视觉特征，采用表示能力更强的主干网络才能够更好地发挥预训练的作用。

将CPR-CLIP+与CPR-CLIP进行性能对比，可以探究二进制交叉熵损失$\mathcal{L}_{BCE}$对复合错误识别性能的作用。对比结果显示，$\mathcal{L}_{BCE}$带来的类别强监督信息能够弥补预训练过程中批次内自监督信息的不足，从而带来更充分的模型训练和更高的复合错误识别精度。在 TSM 视频主干的设定下，CPR-CLIP+相较于朴素迁移策略有 14.14 百分点 mAP 提升与 9.84 百分点 mmit mAP 提升，远高于 CPR-CLIP 带来的性能增益。类似的现象同样发生于 ViViT 主干模型和 Video Swin 主干模型中。特别是在以 Video Swin 为视频主干时，CPR-CLIP+框架的 mAP 和 mmit mAP 指标分别高达 74.39%和 79.24%。

表 4-1　多模态预训练机制的性能对比实验

模型	mAP	Δ	mmit mAP	Δ
TSN[4]	0.5598	—	0.6143	—
CPR-CLIP	0.6034	↑4.36%	0.6727	↑5.84%
CPR-CLIP+	**0.6417**	↑8.19%	**0.7030**	↑8.87%
TSM[126]	0.5662	—	0.6618	—
CPR-CLIP	0.6401	↑7.39%	0.7074	↑4.56%
CPR-CLIP+	**0.7076**	↑14.14%	**0.7602**	↑9.84%
ST-GCN[124]	0.5776	—	0.6692	—
CPR-CLIP	0.6028	↑2.52%	0.6831	↑1.39%
CPR-CLIP+	**0.6358**	↑5.82%	**0.7127**	↑4.35%
ViViT[37]	0.5582	—	0.6651	—
CPR-CLIP	0.6503	↑9.21%	0.7494	↑8.43%
CPR-CLIP+	**0.7251**	↑16.69%	**0.7754**	↑11.03%
Video Swin[152]	0.5696	—	0.6701	—
CPR-CLIP	0.6685	↑9.89%	0.7567	↑8.66%
CPR-CLIP+	**0.7439**	↑17.43%	**0.7924**	↑12.23%

图 4-7(a)(b)分别展示了 CPR-CLIP *w/* TSM 网络在预训练过程中的对比损失 $\mathcal{L}_{CLIP}$ 和平均精度变化情况。注意此训练过程并没有引入二进制交叉熵损失 $\mathcal{L}_{BCE}$，而是只使用了批次内的自监督信息 $\mathcal{L}_{CLIP}$。数据显示，随着训练轮次的加深，$\mathcal{L}_{CLIP}$ 损失不断下降，CPR-CLIP 框架对复合错误的识别精度逐步攀升，这说明批次内的对比预训练能够使模型具备一定的复合错误辨识能力，这也从侧面说明多模态对齐能力有助于复合错误识别性能的提升。

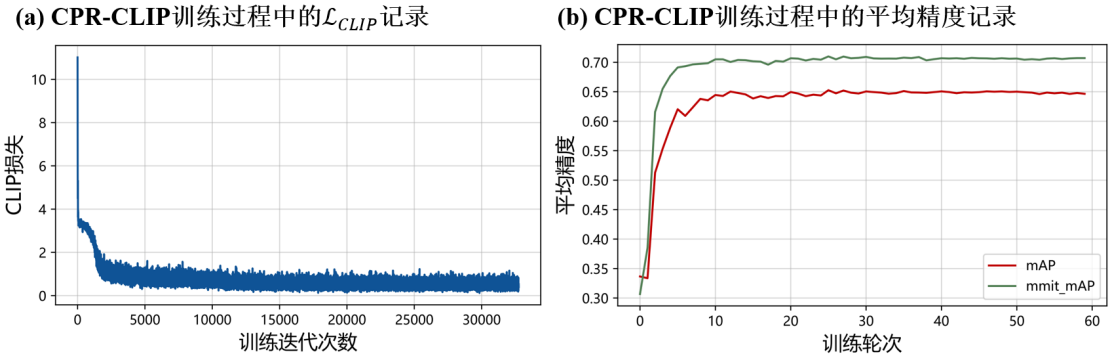


图 4-7 CPR-CLIP *w/* TSM 框架训练过程中的损失与识别精度变化

表 4-2 将 CPR-CLIP+框架与第三章中的复合错误行为识别模型 ImagineNet-FC 进行了性能对比。此部分实验分别采用了 TSN[4]、TSM[126]和 Video Swin Transformer[152]的主干网络设定。作为线性加权特征聚合机制的替代，CBP[153]和 BLOCK[154]特征聚合方法的性能也被记录。对比结果显示，CPR-CLIP+框架的复合错误识别性能明显高于只使用纯视觉模态的模型。在以 Video Swin 为视频主干网络的设定下，CPR-CLIP+框架相较于 ImagineNet-FC 模型获得了 3.57 百分点 mAP 和 2.86 百分点 mmit mAP 的性能增益。这说明通过提示词工程引入的语言模态信息能够与视觉信息形成了互补，模型可以从跨模态对比预训练过程中受益。

表 4-2 CPR-CLIP+框架性能对比实验

模型	视频主干网络	mAP	mmit mAP
CBP[153]	TSN[4]	0.6285	0.6812
BLOCK[154]		0.6225	0.6965
ImagineNet-FC		0.6259	0.6893
CPR-CLIP+		**0.6417**	**0.7030**
CBP[153]	TSM[126]	0.6864	0.7487
BLOCK[154]		0.6651	0.7222
ImagineNet-FC		0.7053	0.7566
CPR-CLIP+		**0.7076**	**0.7602**
CBP[153]	Video Swin[152]	0.6951	0.7524
BLOCK[154]		0.6801	0.7322
ImagineNet-FC		0.7082	0.7638
CPR-CLIP+		**0.7439**	**0.7924**

4.4.3 消融实验

为验证本文所提出的三种提示语句有效性，表 4-3 对P_{num}、P_{cls}和P_{adv}分别进行了提示语句消融实验。实验结果显示，三种提示语句对 CPR-CLIP 最终的复合错误识别性能均有贡献。其中种类提示语句P_{cls}占有最高的权重，这是因为P_{cls}以显式的方式对错误种类信息进行直接描述，这种描述信息所带来的监督信息是不可或缺的；建议提示语句P_{adv}对最终性能也能产生一定的正面影响，因为每种错误的更正建议与错误种类是相互绑定的；数量提示语句P_{num}对最终性能的影响最小，甚至在 CPR-CLIP *w/* TSM 框架的设定下，去除数量提示语句P_{num}反而带来了 0.73 百分点的 mmit mAP 性能提升。这主要是因为在单视频预测推理模式中提示语句中的{*cnt*}只能设定为 1，从而造成了特征偏移（Misalignment）现象。

为探究多模态对比预训练机制与类别监督信息对最终结果的影响关系，表 4-3 补充列举了 CPR-CLIP+框架的性能。性能对比显示，虽然能够通过丰富提示语句种类的方式提高 CPR-CLIP 框架的性能，但是这种跨模态自监督信息（只使用$\mathcal{L}_{CLIP}$损失进行训练）所带来的性能增益依然要弱于明确的类别监督信息（引入$\mathcal{L}_{BCE}$损失进行训练）。只有同时使用批次内的自监督信息与显式的类别监督信息才能取得最优的复合错误行为识别性能。

表 4-3　三种提示语句类型的消融实验结果

主干网络	模型变体	P_{num}	P_{cls}	P_{adv}	mAP	mmit mAP
TSN[4]	CPR-CLIP	✓	✓	✓	**0.6034**	**0.6727**
		×	✓	✓	0.5493	0.6443
		✓	×	✓	0.4364	0.5226
		✓	✓	×	0.5306	0.6604
	CPR-CLIP+	✓	✓	✓	0.6417	0.7030
TSM[126]	CPR-CLIP	✓	✓	✓	**0.6401**	0.7074
		×	✓	✓	0.6298	**0.7147**
		✓	×	✓	0.4498	0.5480
		✓	✓	×	0.5651	0.6870
	CPR-CLIP+	✓	✓	✓	0.7076	0.7602
Video Swin[152]	CPR-CLIP	✓	✓	✓	**0.6685**	**0.7567**
		×	✓	✓	0.6351	0.7328
		✓	×	✓	0.4525	0.5910
		✓	✓	×	0.5591	0.7470
	CPR-CLIP+	✓	✓	✓	0.7439	0.7924

4.4.4 辅助评判系统有效性验证

在 4.3.4 小节介绍的 CPR-CLIP 框架特定类别视频检索推理模式中，可以通过自然语言对所有视频进行快速检索，具体实现方法如公式(4.25)所示。为充分利用此检索功能，本章将 CPR-CLIP 框架封装为一个心肺复苏复合错误判别电子

助手，并招募医师开展了对比实验对此系统的实际效用进行探究。在模型的设定与选型方面，本文充分结合了前文中各模型的性能对比结果，最终使用表征能力较强的 Video Swin Transformer 模型作为视频主干网络。在模型训练方面，本文同时采用了对比预训练损失 $\mathcal{L}_{BCE}$ 与二进制交叉熵损失 $\mathcal{L}_{CLIP}$ 进行框架的训练。通过使用此电子助手，教练医师无需再对单个视频进行逐个判别，而是直接通过自然语言以类别查询的方式从全部视频中进行检索。教练医师最终只需对 CPR-CLIP 框架生成的检索结果进行审核即可，从而大幅节约评估过程所消耗的时间成本。

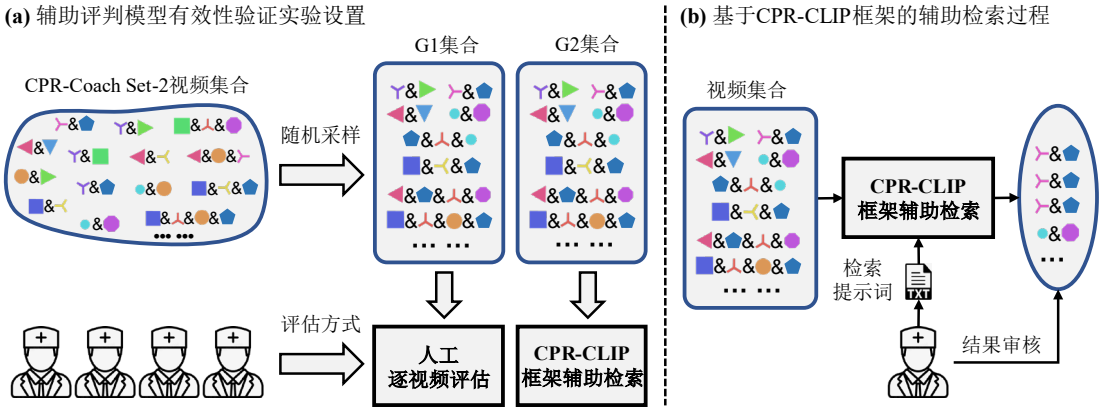


图 4-8　随机对照试验设置与辅助检索过程

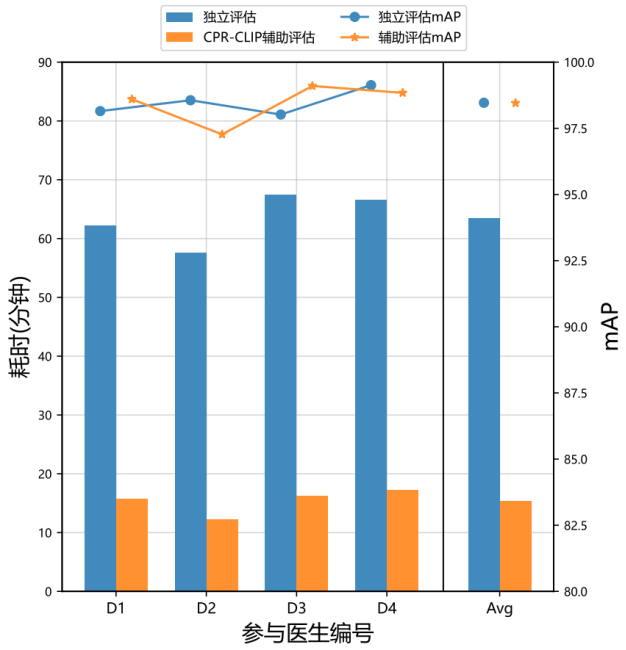


图 4-9　CPR-CLIP 框架的辅助能力对比探究

实验的具体实施过程如图 4-8(a)所示，本章共招募了四名医师对 CPR-CLIP 框架的辅助效用进行测试。首先将 Set-2 数据集随机划分为大小相等的两个子集：

G1 与 G2，其中两个子集中各复合错误种类数量的视频保持一致。其次要求四名医生以传统的判别方式对 G1 中的所有视频进行判别；之后要求四名医生使用 CPR-CLIP 框架的检索模式在 G2 中通过自然语言进行检索与审核。辅助检索过程如图 4-8(b)所示。本文对两种测试模式的耗时与平均均值精度 mAP 进行了记录，所有结果汇总在图 4-9 中，其中耗时数据使用柱状图表示，mAP 数据使用折线图表示。四名医师的实验结果说明基于 CPR-CLIP 框架的智能检索模式能够在不降低评估精度的前提下节约 4 倍左右的耗时。对比结果充分说明，在技能评估系统中引入多模态交互技术能够极大地提升系统的实用性。

4.5 本章小结

本章针对现有医疗技能评估系统人机交互性能差的问题，在第三章基础上对细粒度复合错误识别模型进行了进一步改进。通过多模态对比学习和提示词工程的引入，本文提出了多模态对比预训练框架 CPR-CLIP。该框架首先从错误数量、错误种类和改正建议三个角度实现了提示语句的构建，其次通过最小化语言和视觉模态之间的对比预训练损失实现特征对齐，最终赋予模型更优异的复合错误识别性能。CPR-CLIP 框架的推理过程能够同时支持单视频识别和自然语言检索模式。本章充分利用自然语言检索模式构建了心肺复苏行为电子评估助手，并开展了随机对照试验进行系统的交互能力探究。实验结果充分证实了 CPR-CLIP 框架在实际医疗技能评估应用中的有效性，为后续多模态技术在医疗技能评估系统中的应用探究奠定了基础。

第5章 基于时序聚类注意力机制的扩散时序行为分析算法

5.1 引言

在之前的三个章节中，本文分别针对医疗行为质量评估、复合错误行为识别与评估系统的人机交互能力改善问题进行了探究，所提出的算法均将视频序列视为一个整体，评估任务的形式是将单个视频映射到特定的标签空间中。这类任务的形式较为简单，并未从时序维度对医疗行为序列进行深入分析，例如如何在流程繁琐的外科手术视频中完成某类操作的快速定位、如何对视频中每一帧手术操作类别进行预测、如何实现各个操作环节中的行为合规性检测等。这些任务均已超出前三章探究算法的功能范畴，因此本章将着重对医疗技能评估系统构建中的时序行为分析任务（Temporal Action Analysis, TAA）进行深入探究。

时序行为分割任务（Temporal Action Segmentation, TAS）是时序行为分析任务的重要前置任务，因为对于计算机而言，判断行为正确与否的前提是理解每一帧图像的行为类别。目前学界中已有一些研究者构建了时序行为分割数据集。这些公开数据集根据采集场景可划分为医疗相关和非医疗相关两大类。医疗相关数据集通常对各类外科手术过程的视频进行采集，再对视频进行时序行为划分，例如探究腹腔镜胆囊切除手术的 Cholec80[20]、CholecT50[71]、Hei-Chole[128]，探究白内障手术的 Cataract-101[21]、CATARACTS[73]，探究 *da Vinci* 手术机器人操作的 Nephrec9[22]、ATLAS[74]、RARP45[75]；非医疗相关数据集通常关注日常生活行为，例如早餐制作数据集 Breakfast[68]，沙拉制作数据集 50Salads[11]和日常活动数据集 GTEA[12]。这些数据集为早期的时序行为分割算法提供了测试基准，但依然面临着一些问题：（1）数据集场景相对独立，尚无统一的时序行为划分标准。以上数据集只关注于特定手术或场景中的操作和流程，因此最终构建的数据集只能支持有限种类的行为分割；（2）行为种类划分粒度粗，数据集的复杂度和难度不高。以上数据集的行为种类数量通常在 20 种以下，无法支持复杂操作中的行为分割任务；（3）现有时序行为分割数据集只含有正确行为案例，并未提供错误行为案例，因此现有数据集只能支持时序行为分割任务，无法支持更复杂的时序行为纠错与分析任务。

在算法研究方面，现有的时序行为分割算法根据模型结构可划分为四类：基于循环神经网络（Recurrent Neural Networks, RNN）的框架[77,166]、基于时序卷积网络（Temporal Convolution Networks, TCN）的框架[78,79]、基于图卷积网络（Graph Neural Networks, GNN）的框架[85,86]和基于 Transformer 模型的框架[87,167]。由于

Transformer 模型具有强大的时序信息建模能力，目前时序行为分割的模型基本上都采用了“编码器—解码器”的结构设计。这些算法虽然在现有公开数据集中取得了一定的性能，但是却存在两个方面的缺点：（1）这些算法通常直接对整个时间序列进行无差别建模，忽略了特征信息在时间维度上的差异性；（2）这些算法在处理过长序列时存在延时过高的问题；（3）现有算法只能支持时序行为分割功能，无法支持时序行为纠错与分析功能。针对以上数据集与算法所面临的问题，本章围绕时序行为分析任务分别在四个方面进行了系统性探索探索。图 5-1 对四个研究内容之间的关联进行了阐述。

（1）在**知识图谱构建**方面，本章针对学界中现有时序行为分析数据集无统一构建框架的问题，依据《中国医学生临床技能操作指南》教材构建了时序医疗行为知识图谱。该知识图谱能够同时支持医疗行为的流程知识和文本知识的表示，为后续时序医疗行为分析研究奠定了基础。

（2）在**数据集构建**方面，本章针对现有时序行为分割数据集复杂度低且未提供负面行为案例的问题，以胸腔穿刺术为研究对象设计并构建了 ThoSet 数据集。该数据集依据行为流程知识图谱将标准的胸腔穿刺行为拆解为 39 个子行为，并在专业医生的指导下总结了 22 类遗漏错误和 23 类单流程错误行为。ThoSet 能够支持时序错误行为检测任务。

（3）在**时序行为分割算法**方面，本章在现有的扩散时序行为分割模型（Diffusion Action Segmentation）的基础上，提出了时序聚类注意力模块 *k*M-Att 进行有效的时序特征增强。该模块充分借鉴了人类大脑中逻辑功能分区的结构。在公开基准和 ThoSet 中开展的实验充分证实了此模块的有效性。

（4）在**时序行为合规性评估算法**方面，本章提出了基于动态时间规整算法（Dynamic Time Warping, DTW）的时序行为纠错算法并在 ThoSet 进行了验证。

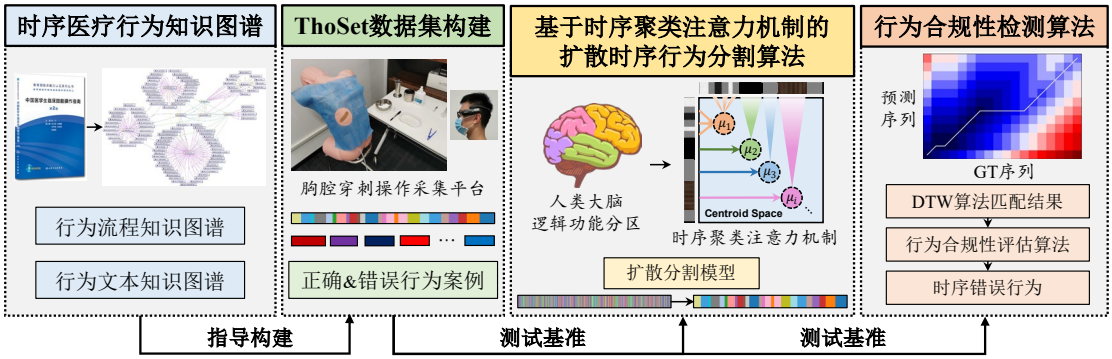


图 5-1　本章各研究内容间的关联

5.2 时序行为分析算法与扩散模型基础

5.2.1 时序行为分析数据集与算法

时序行为分析（TAA）任务旨在让机器理解长时间的人体行为视频，并对行为的正确性、完整性进行分析。时序行为分割（TAS）是时序行为分析的重要前置任务，其目标是为视频中的每一帧指派一个行为标签。作为视频理解任务中的一个重要分支，时序行为分割任务引起了学界的广泛关注。本节对医疗领域和非医疗领域中的行为分割基准数据集进行了简要回顾，并对现有算法进行了分类与总结。

时序行为分析数据集研究现状

在医疗领域中，时序行为分析任务又称为手术流程识别（Surgical Workflow Recognition）任务。现有数据集的研究主题遍布于各类手术场景。国际医学图像处理顶级会议 MICCAI 每年都会举办医疗行为分析任务相关的挑战赛。M2CAI-phase[69]数据集发布于 2016 年外科流程识别和外科器械检测挑战赛，共分为 M2CAI 16-workflow 和 M2CAI 16-tool 两个子数据集，收集了 56 例胆囊切除术的视频并对 8 个手术阶段进行了操作标注；HeiCo[128]数据集发布于 2017 年内窥镜视觉挑战赛（Endoscopic Vision Challenge），收集了 10 例剖宫产手术、10 例直肠切除手术和 10 例乙状结肠切除手术视频，并对手术流程、医疗器械位置框和医疗器械分割信息进行了标注；MISAW[14]数据集发布于 2020 年内窥镜视觉识别挑战赛，收集了 27 例模拟血管缝合手术视频，并从阶段（Phases）、步骤（Steps）和行为（Activities）三个层次划分出 28 种操作；CholecT50[71]数据集公布于 2021 年手术行为三元组识别挑战赛（CholecTriplet-Surgical Action Triplet Recognition），收集了 50 条腹腔镜胆囊切除术视频，并对视频中的 6 类仪器、10 类动作和 15 类目标进行了三元组标注；PETRAW[13]数据集发布于 2021 年内窥镜视觉挑战赛，收集了内窥镜模拟器场景下的 150 条操作视频记录，并提供了视频数据、运动学数据和语义分割数据三种不同模态的信息；SARAS-MESAD[72]发布于 2021 年多领域外科手术行为检测挑战赛（Multi-domain Endoscopic Surgeon Action Detection），收集了在 *da Vinci Xi* 机器人系统上进行的前列腺切除术视频，分为包含真实手术视频的 MESAD-Real 和包含模拟手术的 MESAD-Phantom 两个子数据集。

在 MICCAI 系列挑战赛之外，多所国际医疗机构也针对不同的手术构建了多个手术流程识别基准：

以腹腔镜胆囊切除手术为研究主题，Cholec80[20]数据集收集了斯特拉斯堡大学医院的 13 名外科医生进行的 80 例腹腔镜胆囊切除手术视频，将完整的胆囊切

除手术流程划分为 7 个阶段，并对术中出现的 7 类手术器械进行了标注；Hei-Chole[128]数据集收集了来自德国三个手术中心的 33 个腹腔镜胆囊切除术视频，在手术流程划分上与 Cholec80[20]保持一致，并且额外标注了抓、握持、切割和分层这 4 类行为。Hei-Chole[128]对所有的医疗技能按照腹腔镜技术全球手术评估体系[168]（Global Operative Assessment of Laparoscopic Skills, GOALS）进行了五个维度的评分：深度感知、双手灵活性、效率、组织处理和操作难度。

以白内障手术为研究主题，Cataract-101[21]数据集收集了由奥地利 4 名外科医生主刀的 101 例白内障手术视频，按照白内障手术的准标准化程序将所有视频划分为 10 个手术阶段，并且根据四位外科医生的经验和技能进行了行为质量的排序；CATARACTS[73]数据集收集了在布雷斯特大学医院进行的 50 例白内障手术视频，并对医疗器械和手术的 18 个子流程进行了标注。

以 *da Vinci* 手术机器人系统的操作为研究主题，JIGSAWS[15]数据集收集了 8 名外科医生在 *da Vinci* 机器人系统上执行的三项基本外科技能的数据，包括缝合、打结和传针，并提供了运动学数据和内窥镜相机捕获到的双目视频数据，依据技术技能的客观结构化评估（OSATS）方法对所有操作视频进行全局排序；Nephrec9[22]数据集收集了欧洲癌症中心资深泌尿科医生进行的 9 例 *da Vinci Xi* 机器人辅助肾部分切除手术（RAPN）视频，将肾部分切除术划分为 10 个阶段，并引入状态转移图对手术的阶段进行表示；ATLAS[74]数据集收集了 Roswell Park 癌症学会中 10 名外科医生在 *da Vinci* 机器人上完成的 6 种不同外科手术任务的记录，对每一帧中的医疗器具都进行了标注；RARP45[75]数据集收集了在 *da Vinci Si* 手术机器人系统上进行的 45 例前列腺切除手术的全程记录，将手术流程划分为 1 个背景类和 7 个手术操作类，并提供了视频数据和运动学数据。DESK [76]数据集收集了在 Taurus II 机器人、仿真 Taurus II 机器人、YuMi 机器人三个手术平台上进行的 2,897 条腹腔镜手术基础操作行为记录，训练内容中的手术包迁移行为分为 7 个阶段。

在非医疗领域中，最常见的时序行为分割测评数据集是 Breakfast[68]、50Salads[11]和 GTEA[12]，这些数据集通常以人类的日常活动作为主题。综上所述，虽然目前学界中已经有很多工作探究了医疗场景和非医疗场景下的时序行为分析数据集构建，但是这些数据集的复杂度依然有待提升。现有手术流程分割数据集的行为种类数通常在 10~20 类，有限的数据体量和较少的行为种类可能会导致过拟合现象的发生。本章以胸腔穿刺术为研究对象，构建了时序医疗行为知识图谱，并构建了包含细粒度操作流程的时序分析数据集。

时序行为分析算法研究现状

作为时序行为分析任务的重要前置任务，时序行为分割任务（TAS）是计算

机视觉中一个非常具有挑战性的任务。由于和自然语言处理任务（NLP）有着高度的相似性，早期的时序行为分割算法着重参考了 NLP 领域中的模型。现有时序行为分割算法按照网络结构可分为四类：基于 RNN 的方法[77,166]、基于时序卷积的方法[78,79]、基于图卷积神经网络 GNN 的方法[85,86]和基于 Transformer 模型的方法[87,167]。

早期基于 RNN 的方法直接利用循环神经网络对视频帧序列进行建模，Ding 等人[166]提出了基于时序卷积和循环神经网络的混合方法；Singh 等人[77]使用多通路双向循环神经网络构建细粒度行为检测模型。尽管这些模型在公开数据集上取得了一定的成绩，但是却面临着训练不稳定、精度难以进一步提升的问题。基于时序卷积的方法使用时序卷积（Temporal Convolution）层对视频进行建模，从一定程度上缓解了 RNN 模型所面临的问题。Lea 等人[78]将时序卷积网络引入行为分割和检测中；Lei 等人[79]在时序卷积网络的基础上提出了时序可变形残差网络；Farha 和 Li 等人先后提出了 MS-TCN[80]和 MS-TCN++[82]，一种基于多阶段时序卷积的行为分割框架；Mac 等人[81]提出了关注于时序特征局部一致性的可变形卷积网络检测细粒度行为；Wang 等人[83]提出了门控时序卷积网络对行为分割的结果进行精修；Gao 等人[84]提出一种从整体到局部的高效时序行为分割框架。还有一些工作[85,86]将 GNN 引入到了时序行为分割任务中。

近年来，Transformer[104]模型在自然语言处理和计算机视觉领域中大放异彩。与 RNN 模型不同，Transformer 能够以并行的方式高效处理序列信息。很多时序行为分割的研究者开始使用 Transformer 处理长视频信息：Yi 等人[87]将 Transformer 引入行为分割任务并提出了 ASFormer 模型。此模型由编码器和解码器构成，并且在编码器和解码器层中使用了时序扩散卷积以降低训练阶段模型的显存占用；Aziere 等人[167]将 Depth-wise 卷积和自注意力机制相结合提出了 TCTr 模型，显著改善了自注意力机制的平方级别计算量增长问题；Wang 等人[169]基于编码器－解码器架构提出了交叉增强网络 CETNet，并提出一种新的损失函数对过度分割的错误结果进行惩罚；Du 等人[88]彻底摒弃了时序卷积操作，并提出了一种纯基于 Transformer 架构的模型，称为时序 U-Transformer，并通过边界感知损失函数的设计引入了“相邻帧更有可能属于同一类”的归纳偏置；Du 等人[89]指出 Transformer 模型的引入会导致冗余的计算复杂度和时间复杂度，并提出一种多级可扩展的 MSDTN 框架实现视频中的短期和长期关系进行建模；Tian 等人[170]提出一种同时关注局部和全局信息的 Transformer 网络 LGTNN 对多个尺度的特征进行有效提取，最终通过边界检测网络对行为切换的边缘进行精修；

近期有研究者开始将其他领域中的方法引入到时序行为分割任务中，例如多模态预训练模型、多模态表示学习和扩散生成模型。Li 等人[90]引入了提示词工

程与多模态学习方法并提出 Br-Prompt 框架。此框架能够对时序相邻动作的上下文信息进行建模，通过 CLIP 对编码器和解码器进行有效训练；Amsterdam 等人[91]针对视频数据和传感器数据无法有效融合的问题提出了 ASPnet 框架。此框架在不同模态之间设计了私有和公共有表示通路，并通过注意力瓶颈网络的引入对数据中的长距离依赖性进行建模；Liu 等人[92]将图像生成领域中的扩散模型应用在了时序行为分割模型中并提出了 DiffAct 框架。与传统方法中解码器直接对标签进行预测的方式不同，DiffAct 框架以解码器多次运行的方式逐步生成行为标签。由于引入了这种逐步解码的机制，此框架在时序行为分割基准上取得了最优性能。

虽然现有模型在公开数据集上取得了一定的性能成果，但是依然存在两个方面的问题：一方面，这些模型对输入特征序列一视同仁，忽略了不同行为在时序上的差异，进而导致了行为分割性能无法进一步提升；另一方面，DiffAct 框架中的解码器多次迭代会引发推理过程延时过高问题。针对以上两个问题，本章提出了基于时序聚类注意力机制的 Transformer 模型对时序特征进行差异化增强，并且提出非锁步跳跃解码机制有效削减解码器在去噪过程中的运行轮数，在公开数据集的实验证明了方法的有效性与高效性。

5.2.2 扩散模型基本理论

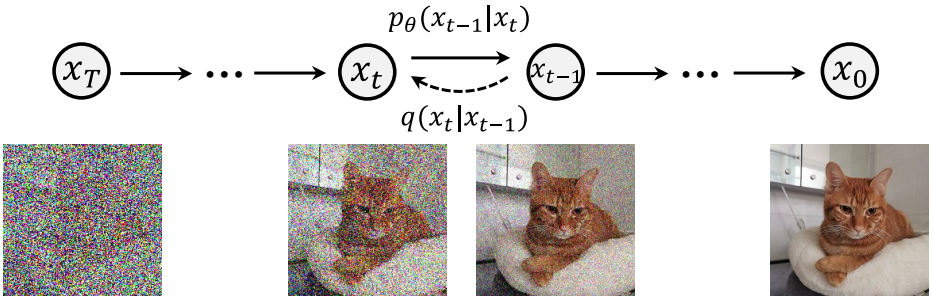


图 5-2　扩散模型的扩散过程与去噪过程示意图

扩散模型（Diffusion Models）最早由 Sohl-Dickstein 等人[171]于 2015 年提出，将统计物理中的非平衡热力学（Non-equilibrium Thermodynamics）原理引入机器学习中的数据加噪与修复问题。扩散模型是一种强大的生成模型，相较于对抗生成模型（GANs）具有训练稳定的特性，目前已经广泛应用于图像生成、自然语言生成、文本—图像生成（Text-to-Image Generation）等任务中。如图 5-2 所示，扩散模型的核心流程是通过前向/扩散过程（Forward Process / Diffusion Process）对原始数据分布进行扰乱，再通过反向/去噪过程（Reverse Process / Denoising Process）对扰乱后的数据分布进行复原。本节对扩散模型基本理论进行回顾，为后续扩散时序行为分割模型构建奠定基础。

给定原始数据分布$x_0 \sim q(x_0)$，前向过程通过添加T次高斯噪声获取到整个加噪路径（Perturbation Path）$\{x_0, x_1, \cdots, x_T\}$。其中单步加噪操作是一个一阶马尔科夫过程，可表示为$q(x_t|x_{t-1})$。通过概率计算的链式法则和一阶马尔可夫性，可以将加噪路径的联合概率分布表示为：

$$q(x_0, x_1, \cdots, x_T) = q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{5.1}$$

单步加噪过程的具体实现方式为：

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\right) \tag{5.2}$$

其中$\beta \in (0,1)$是预定义的超参数，用于实例化单步添加的高斯噪声。为了后续表述方便，定义$\alpha_t := 1 - \beta_t$。由于α_t在模型训练和推理之前已经确定，所以整个加噪过程是确定的。通过正态分布的加法性质对x_0进行单步加噪得到x_t：

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_{t-1}, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{5.3}$$

其中α_t定义为$\{\alpha_0, \alpha_1, \cdots, \alpha_t\}$的累乘：

$$\bar{\alpha}_t := \prod_{i=0}^{t} \alpha_i \tag{5.4}$$

给定初始数据x_0，添加T次高斯噪声后的数据x_T表示为：

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon \tag{5.5}$$

其中$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$为标准正态分布。前向过程的逐步加噪操作将原始数据分布转化为随机噪声分布，后向过程首先从一个随机噪声分布中进行采样$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，并逐步恢复为原始数据。

去噪的单步过程表示为：

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5.6}$$

其中θ表示神经网络参数，均值$\mu_\theta(x_t, t)$和标准差$\Sigma_\theta(x_t, t)$通过神经网络进行估计。以迭代的方式重复此去噪过程，即可将随机噪声x_T复原为初始数据x_0，同时获得完整的去噪路径（Denoising Path）$\{x_T, x_{T-1}, \cdots, x_0\}$。

去噪路径的联合概率分布表示为：

$$p_\theta(x_0, x_1, \cdots, x_T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{5.7}$$

能够从x_T复原为x_0的前提是：加噪路径和去噪路径之间的差异尽可能小。两个路径之间的差异使用 KL 散度表示为：

$$\begin{aligned}&\mathrm{KL}[q(x_0,x_1,\cdots,x_T)|p_\theta(x_0,x_1,\cdots,x_T)]\\&=-\mathbb{E}_{q(x_0,x_1,\cdots,x_T)}[\log p_\theta(x_0,x_1,\cdots,x_T)]+C\end{aligned}\tag{5.8}$$

其中C为不依赖于θ的常数项。将x_0的对数似然变分下界（Variational Lower Bound, VLB）定义为$L_{LVB}(x_0)$：

$$L_{LVB}(x_0):=\mathbb{E}_{q(x_0,x_1,\cdots,x_T)}\left[\log p(x_T)+\sum_{t=1}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right]\tag{5.9}$$

KL 散度可以表示为负变分下界$-L_{LVB}(x_0)$与常数C之和，再根据 Jensen 不等式可获得 KL 散度与$-\log p_\theta(x_0)$的关联：

$$\begin{aligned}&\mathrm{KL}[q(x_0,x_1,\cdots,x_T)|p_\theta(x_0,x_1,\cdots,x_T)]\\&=-L_{VLB}(x_0)+C\\&\geqslant\mathbb{E}[-\log p_\theta(x_0)]+C\end{aligned}\tag{5.10}$$

扩散模型训练的最终目标是最小化$q(x_0,x_1,\cdots,x_T)$和$p_\theta(x_0,x_1,\cdots,x_T)$之间的 KL 散度，从而让两个分布尽可能地近似。KL 散度的最小化等价于最大化变分下界$L_{LVB}(x_0)$，同样等价于最小化$-\log p_\theta(x_0)$，最终实现$p_\theta(x_0)$的最大化。$L_{LVB}(x_0)$的最大化有多种实现方式，DDPM 模型通过神经网络对前向过程中添加的单步噪声进行预测，并将损失函数定义为

$$\mathcal{L}=\mathbb{E}_{t,x_0,\epsilon}[\|\epsilon-\epsilon_\theta(x_t,t)\|^2]\tag{5.11}$$

其中$t\in[1,\cdots,T]$，x_t由公式(5.5)得到，$\epsilon_\theta(x_t,t)$表示神经网络根据x_t和t对噪声的预测值。在完成神经网络的训练之后，通过迭代的方式即可将x_T逐步复原为原始数据x_0。

5.3 时序医疗行为知识图谱构建

知识表示（Knowledge Representation, KR）是指将现实世界中的知识转化为机器可存储、可计算形式的过程，是机器运用先验知识解决实际问题的基础。研究者们先后提出了语义网络、专家系统、语义网和知识图谱（Knowledge Graph, KG）等知识表示技术，以及各种知识表示语言：基于框架的语言、产生式规则、资源描述框架（Resource Description Framework, RDF）、语义资源描述框架（RDF Schema, RDFS）、Web 本体语言（Web Ontology Language, OWL）等。谷歌在 2012 年首次提出了知识图谱的概念，最初用于搜索功能产品命名，之后成为各类结构化知识库的统称。知识图谱通过结构化的语义知识库，以符号形式对世界中的实体（Entity）和关系（Relation）进行描述，基本组成单位是“实体—关系—实体”三元组，又称 SPO（Subject-Predicate-Object）三元组。目前知识图谱技术已被广

泛地应用于搜索引擎、问答系统和推荐系统领域中，是人工智能技术的重要分支。知识图谱技术在医疗、法律、金融等知识密集型的垂直领域有着广阔的应用价值和前景。

在医疗领域中，学界中已经有系列研究构建了多个医学知识图谱。现有的英文医疗知识图谱按细分领域可划分为：药物化学类（Medicinal Chemistry），例如 HetioNet[172]、DrKG[173]、PrimeKG[174]；生物信息类（Bioinformatics），例如 STRING4[175]、Cell Ontology[176]；药物开发类（Drug Development），例如 GEFA[177]、Reaction[178]、DDKG[179]；临床决策支持类（Clinical Decision Support），例如 Disease Ontology[180]、DrugBank[181]、KnowLife[182]。中文领域中的医疗知识图谱发展尚处于早期阶段。表 5-1 列举了目前开源的中文医疗知识图谱：DiseaseKG 由 OpenKG 组织于 2020 年构建，其原始数据来源于各权威医药网站，共含有 8 类实体和 11 类关系；DiaKG[183]由妙健康和阿里云于 2021 年构建，是一个高质量的糖尿病知识图谱中文数据集，包含 18 类实体和 15 类关系；COVID-19[184]学术知识图谱由云南省高校数据科学与智能计算重点实验室于 2020 年构建，收集了和新冠病毒相关的约 2 万篇论文，对作者、论文、标题、研究机构、主要成果等多项内容进行了处理；中文症状库[185]由华东理工大学于 2016 年构建，其原始数据来源于 8 个主流健康咨询网站、3 个中文百科网站和电子病历。

表 5-1　中文医疗领域开源知识图谱

知识图谱名称	知识图谱主题	实体数量	关系数量	提出单位
DiseaseKG[186]	基于 cnSchma 常见疾病信息知识图谱	44,656	312,159	OpenKG
DiaKG[183]	糖尿病知识图谱数据集	22,050	6890	妙健康、阿里云
COVID-19[184]	基于 COVID-19 论文集的学术知识图谱	80 万	120 万	云南省高校数据科学与智能计算重点实验室
中文症状库[185]	包含症状实体和症状相关三元组数据集	135,485	617,499	华东理工大学

构建知识图谱的本质是从非结构化的文本信息中提取结构化的信息，又称为信息提取（Information Extraction, IE）。知识图谱的构建往往是一个庞大而复杂的工程，通常由一系列自然语言处理任务构成：命名实体识别（Named Entity Recognition, NER）、关系提取（Relation Extraction, RE）和事件提取（Event Extraction, EE）任务。命名实体识别的目标是从非结构化文本中提取预先定义的特定类别信息，例如人名、机构名称、地点、时间等；关系提取的目标是从文本中提取实体之间的关系；事件提取的目标是从文本中提取预先定义的事件类别和事件触发词，通常应用于新闻、金融等以事件为中心的领域场景中。在中文领域中目前已经有较多的医疗信息提取数据集和基准，表 5-2 进行了详细列举。

表 5-2　中文医疗领域知识图谱构建数据集与竞赛调研

任务类型	数据集	数据集全称名称	提出单位
信息抽取类任务	CMeEE	中文医学命名实体识别数据集	北京大学计算语言学教育部重点实验室
	CMeEE-V2	中文医学命名实体识别 V2	郑州大学自然语言处理实验室
	CMeIE	中文医学文本实体关系抽取	郑州大学自然语言处理实验室
	CHIP-CDEE	临床发现事件抽取	医渡云
医学术语标准化	CHIP-CDN	临床术语标准化任务	医渡云
医学文本分类任务	CHIP-CTC	临床试验筛选标准短文本分类	同济生命科学学院
	KUAKE-QIC	医疗搜索检索词意图分类	阿里夸克
医学句子语义关系判定任务	KUAKE-QTR	医疗搜索查询词-页面标题相关性	阿里夸克
	KUAKE-QQR	医疗搜索查询词-查询词相关性	阿里夸克
	CHIP-STS	疾病问答迁移学习	平安医疗科技
医疗对话理解与生成任务	CHIP-MDCFNPC	医疗对话临床发现阴阳性判别	阿里夸克
	IMCS21	智能对话诊疗数据集 IMCS-NER、IMCS-IR、IMCS-SR、IMCS-MRG	复旦大学数据智能与社会计算实验室
	MedDG	蕴含实体的中文医疗对话生成	中山大学-腾讯天衍实验室
	IMCS-V2	智能对话诊疗数据集 IMCS-NER、IMCS-DAC、IMCS-SR、IMCS-MRG	复旦大学数据智能与社会计算实验室
医学段落检索	KUAKE-IR	医学段落检索	阿里巴巴搜索事业部

现有的医疗知识图谱关注于医疗知识的表示，并不能支持对医疗行为这种具有时序特性的知识进行表示。时序医疗行为知识图谱的构建对于医疗行为分析研究具有重要意义，因此本节将对原有的医疗知识图谱体系进行扩充，使其具备表示时序医疗行为知识的能力。通过对医疗技能教材内容进行观察发现，医疗行为相关的知识具有以下三个特性：

1、**类别有限**：对于一种特定的医疗行为，教材中通常只包含目的、适应症、禁忌证、操作前准备、操作步骤、并发症与处理、相关知识共七类特定信息；

2、**层次化**：一个完整的医疗行为通常由若干个子行为构成，而一个子行为又由一系列操作构成。这些层次化的隶属关系需要在知识图谱中进行表示；

3、**有序性**：部分医疗行为是严格有序的，违背顺序会造成不合规医疗行为的发生，进而导致无菌环境被破坏、治疗效果受影响等问题。

本研究以《中国医学生临床技能操作指南（第二版）》教材为依据进行医疗行为知识图谱的构建。此教材共含字数 62.4 万，囊括了 60 类常见临床技能操作，细类临床操作数目达 73 种。图 5-3 对本教材所涉及的所有临床操作进行了列举。

图 5-3 《中国医学生临床技能操作指南》中的 73 种技能操作

5.3.1 时序医疗行为知识 Schema 设计

在知识图谱中，Schema 指概念的类型以及类型的属性定义，是构建知识图谱的基础框架。Zhang 等人[187]在医学实体标注规范中根据常用的医学术语分类体系提出了九大类实体标记方案，分别为：疾病（dis）、临床表现（sym）、医疗程序（pro）、医疗设备（equ）、药物（dru）、医学检查项目（ite）、身体（bod）、科室（dep）和微生物类（mic）。其中的“医疗程序”实体可以继续细分为医疗程序（pro-cp）和治疗预防程序（pro-tp）。上述分类框架虽然能够为通用医学信息提取任务提供参考，但是并不能满足医疗行为知识图谱构建这种有更高内容细粒度要求的任务。因此本文首先针对时序医疗行为知识的特点设计了 Schema，之后再完成时序医疗行为知识图谱的构建。

结合上文对临床技能操作知识的三个特性总结以及现有医疗知识表示框架缺点的分析，本研究使用两套独立的 Schema 框架对医疗行为知识进行表示。两套 Schema 框架所表示的内容如下：

1、**医疗行为 Schema**：本文提出了一套全新的 Schema 以解决医疗行为的层次化、有序性表示问题，共包含 12 类实体类别和 11 种关系类别。如表 5-3 所示，本文通过“准备工作类”“二级准备工作类”“操作步骤类”和“二级操作步骤类”等概念的引入来表示医疗行为的层次化信息和有序性信息。

表 5-3　医疗行为 Schema 中的实体类别和关系类别定义

实体类别	关系类别
临床技能操作	——
医疗目的	目的
适用疾病	适应症
禁忌疾病	禁忌证
准备工作类	操作前准备
二级准备工作类	二级准备工作
操作步骤类(有序)	操作步骤
二级操作步骤类(有序)	二级操作步骤
操作步骤类辅助说明类	操作步骤流程说明
二级操作步骤类辅助说明类	操作步骤流程说明
并发症类	并发症
并发症处理方法类	并发症处理方法

表 5-4　CMeIE 数据集中医疗知识的实体类别和关系类别定义

实体类别	关系类别			
疾病	预防	内窥镜检查	病理分型	病史
症状	阶段	筛查	相关(导致)	遗传因素
检查	就诊科室	多发群体	相关(转化)	发病机制
药物	辅助治疗	发病率	相关(症状)	病理生理
部位	化疗	发病年龄	鉴别诊断	药物治疗
手术治疗	放射治疗	多发地区	临床表现	发病部位
其他治疗	手术治疗	发病性别倾向	治疗后症状	转移部位
预后	实验室检查	死亡率	侵及组织症状	外侵部位
流行病学	影像学检查	传播途径	病因	预后状况
社会学	辅助检查	多发季节	高危因素	预后生存率
其他	组织学检查	并发症	风险评估因素	同义词

2、**医疗行为文本知识 Schema：**临床技能操作指南教材中含有大量知识性和说明性文本。如果仅对医疗行为的流程进行表示，则会遗漏丰富的医疗行为知识内容。因此，本文继续沿用中文医学文本实体关系抽取数据集 CMeIE 提出的 Schema 框架对教材中大量的文本内容进行知识提取与表示。CMeIE 共定义了 11 类实体类别和 44 类关系类别。

医疗行为 Schema 侧重于医疗操作相关内容的描述，而行为知识 Schema 侧重于知识内容的表示。二者相互补充，共同构成时序医疗行为知识图谱框架。

5.3.2 医疗行为流程知识图谱构建

如表 5-3 所示，上节参照教材中医疗操作描述框架设计了医疗行为 Schema，并对实体的类型和关系进行了定义。得益于教材文本中清晰的描述框架，本节提出了基于字符串匹配的医疗行为流程知识图谱构建方案。这种朴素且简洁的方法适用于具有清晰目录结构的文本信息提取任务。在精简掉繁杂的编程细节后，

表 5-5 流程知识图谱的实体类别与举例

实体类别	实体举例
临床技能操作	“胸腔穿刺术”
医疗目的	“明确胸腔积液病因” “抽出胸腔内液体促进肺复张” “胸膜腔内给药”
适用疾病	“胸腔积液需要明确诊断” “胸膜腔内给药” “大量胸腔积液产生呼吸困难等压迫症状”
禁忌疾病	“凝血功能障碍” “重症血小板减少者”
准备工作类	“1.患者准备” “2.材料准备” “3.操作者准备”
二级准备工作类	“(1)测量生命体征” “(2)向患者解释胸腔穿刺的目的” “(3)告知需要配合的事项” “(4)签署知情同意书”
操作步骤类 (有序)	“1.体位” “2.穿刺点选择” “3.消毒铺单” “4.麻醉” “5.穿刺” “6.抽液” “7.拔针” “8.穿刺后的观察” “9.标本处理”
二级操作步骤类 (有序)	“(1)穿刺点选择” “(2)标记穿刺点” “(3)叩诊寻找” “(1)准备” “(2)注射皮丘” “(3)间断负压回抽” “(1)准备” “(2)注射皮丘” “(3)间断负压回抽” “(1)准备” “(2)穿刺” “(3)回吸”
操作步骤类辅助说明类	“1.体位: 再次确认病变位于左侧还是右侧。常规取直立坐位，上身略前倾，必要时双前臂合抱……”
二级操作步骤类辅助说明类	“(1)抽取胸腔积液: 当穿刺针回吸到液体后，经穿刺针导管连接 50ml 注射器抽取胸腔积液。”
并发症类	“胸膜反应” “气胸” “复张性肺水肿” “腹腔脏器损伤” “血胸”
并发症处理方法类	“1.胸膜反应:停止操作_平卧_皮下注射 0.1%肾上腺素 0.3~0.5ml”

表 5-6　流程知识图谱的关系类别与三元组举例

关系类别	三元组举例
目的	(胸腔穿刺术，目的，明确胸腔积液病因) (胸腔穿刺术，目的，抽出胸腔内液体促进肺复张) (胸腔穿刺术，目的，胸膜腔内给药)
适应症	(胸腔穿刺术，适应症，胸腔积液需要明确诊断) (胸腔穿刺术，适应症，胸膜腔内给药) (胸腔穿刺术，适应症，大量胸腔积液产生呼吸困难等压迫症状)
禁忌证	(胸腔穿刺术，禁忌证，凝血功能障碍) (胸腔穿刺术，禁忌证，重症血小板减少者)
操作前准备	(胸腔穿刺术，操作前准备，1.患者准备) (胸腔穿刺术，操作前准备，2.材料准备) (胸腔穿刺术，操作前准备，3.操作者准备)
二级准备工作	(1.患者准备，二级准备工作，(1)测量生命体征) (1.患者准备，二级准备工作，(2)向患者解释胸腔穿刺的目的) (1.患者准备，二级准备工作，(3)告知需要配合的事项) (1.患者准备，二级准备工作，(4)签署知情同意书)
并发症	(胸腔穿刺术，并发症，1.胸膜反应) (胸腔穿刺术，并发症，2.气胸) (胸腔穿刺术，并发症，3.复张性肺水肿)
并发症处理方法	(1.胸膜反应，并发症处理方法，停止操作_平卧_皮下注射…) (2.气胸，并发症处理方法，可由以下原因引起_穿刺过深伤及肺…) (3.复张性肺水肿，并发症处理方法，胸腔积液引流速度不能…)
操作步骤(有序)	(胸腔穿刺术，操作步骤，1.体位) (胸腔穿刺术，操作步骤，2.穿刺点选择) (胸腔穿刺术，操作步骤，3.消毒铺单) (胸腔穿刺术，操作步骤，4.麻醉) (胸腔穿刺术，操作步骤，5.穿刺) (胸腔穿刺术，操作步骤，6.抽液) (胸腔穿刺术，操作步骤，7.拔针) (胸腔穿刺术，操作步骤，8.穿刺后的观察) (胸腔穿刺术，操作步骤，9.标本处理)
二级操作步骤(有序)	(2.穿刺点选择，二级操作步骤，(1)穿刺点选择) (2.穿刺点选择，二级操作步骤，(2)标记穿刺点) (2.穿刺点选择，二级操作步骤，(3)叩诊寻找) (3.消毒铺单，二级操作步骤，(1)准备) (3.消毒铺单，二级操作步骤，(2)消毒) (3.消毒铺单，二级操作步骤，(3)铺巾) (4.麻醉，二级操作步骤，(1)准备) (4.麻醉，二级操作步骤，(2)注射皮丘) (4.麻醉，二级操作步骤，(3)间断负压回抽) (5.穿刺，二级操作步骤，(1)准备) (5.穿刺，二级操作步骤，(2)穿刺) (5.穿刺，二级操作步骤，(3)回吸) (6.抽液，二级操作步骤，(1)抽取胸腔积液) (6.抽液，二级操作步骤，(2)注射器排空) (6.抽液，二级操作步骤，(3)诊断性穿刺) (7.拔针，二级操作步骤，(1)拔除动作) (7.拔针，二级操作步骤，(2)嘱咐患者与测量)
操作步骤流程说明	(1.体位，操作步骤流程说明，再次确认病变位于左侧还是右侧…) ((3)间断负压回抽，操作步骤流程说明，如无液体或鲜血吸出…)

图 5-4 对行为信息提取的核心流程进行了展示：首先将教材内容的 OCR 结果进行人工纠正，再以单行形式存储为 TXT 文本；之后依据教材对医疗操作的描述框架设计四个级别的标题正则匹配模板；通过以上模板对全文进行匹配，最终以结构化文本的形式存储为 JSON 文档。

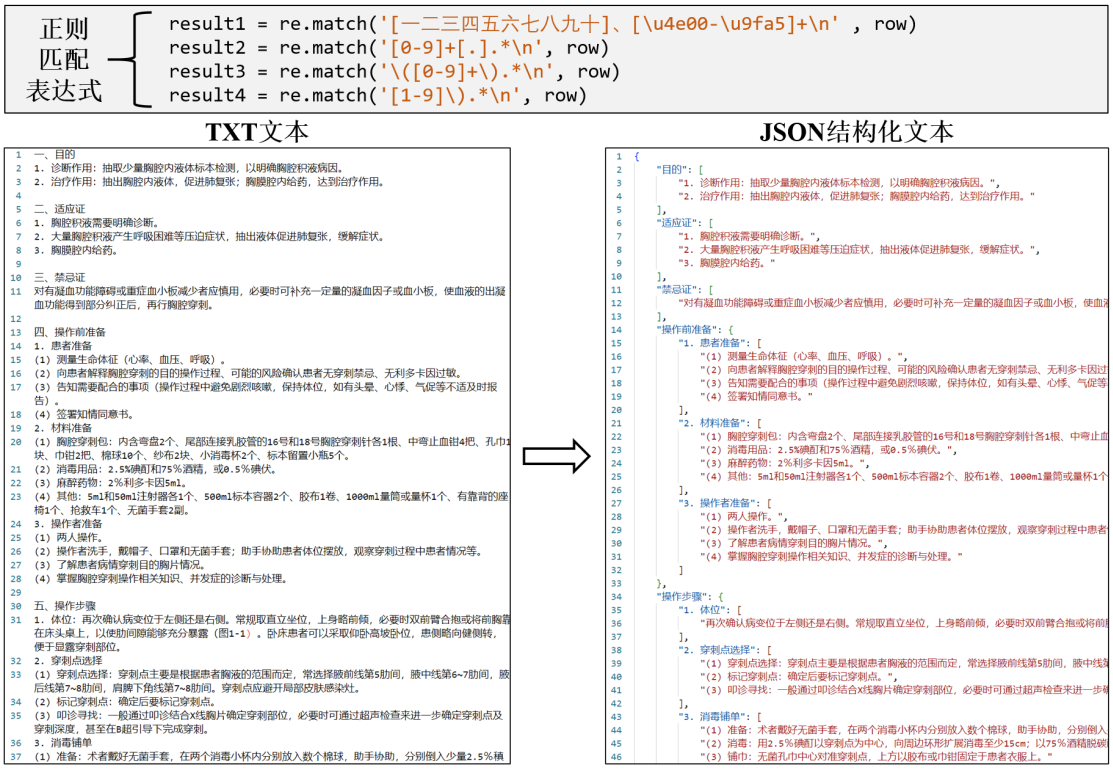


图 5-4　正则匹配表达式设计与提取结果

在完成文本的分级处理之后，本文通过表 5-3 中的医疗行为 Schema 对结构化文本进行二次处理，进一步将文本信息转化为三元组，完成医疗行为流程知识图谱的构建。表 5-5 展示了胸腔穿刺流程知识图谱的实体类别与案例，表 5-6 展示了关系类别与部分三元组案例。以上案例充分说明本节所提出的层次化行为 Schema 能够对行为之间的隶属关系与先后顺序进行准确描述。

5.3.3 医疗行为文本知识图谱构建

上节根据医疗行为 Schema 构建了流程知识图谱，然而还有大量的说明性文本以长句子的形式存储。为了进一步对这些长句子中蕴含的丰富知识进行提取，本节对这类文本信息进行了知识图谱构建探究。目前 NLP 领域中已经有多种实体和关系提取算法，本节使用现有开源模型与数据集完成教材中医疗行为文本知识图谱的构建。在数据集方面，中文医学文本实体关系抽取数据集 CMeIE[188]（Chinese Medical Information Extraction Dataset）发布于 CHIP-2020 会议，是中文医学实体关系抽取的基准数据集。如表 5-4 所示，CMeIE 数据集定义了 11 类

实体类别、44 类关系类别和 53 类 Schema 规则，共包含 17,924 个语句和 41,982 个三元组。语句的 Train-Test-Val 划分比例为 13,817: 522: 3,585，三元组划分比例为 41,982: 1,678: 10,626。

在模型方面，本节使用 PURE[189]（Princeton University Relation Extraction system）算法完成命名实体识别与关系抽取功能。PURE 算法是一种端到端的实体关系提取模型，由两个独立的编码器分别完成命名实体识别和关系抽取功能。和经典方法保持一致，PURE 的命名实体识别模型是一个标准的 BERT[105]预训练语言模型，由编码器完成输入文本的特征提取，最终由线性层完成实体预测；关系提取模型以成对的实体为输入（Subject-Object）进行关系预测。传统的关系提取模型通常直接复用命名实体识别阶段的特征，这些方法认为此阶段的特征已经蕴含了丰富的上下文信息。PURE 算法的核心创新在于：在预测关系之前对每个成对实体进行独立标记，从而实现关系提取过程中的关键词高亮，避免多个实体对之间的特征干扰。PURE 算法的流程如图 5-5 所示。

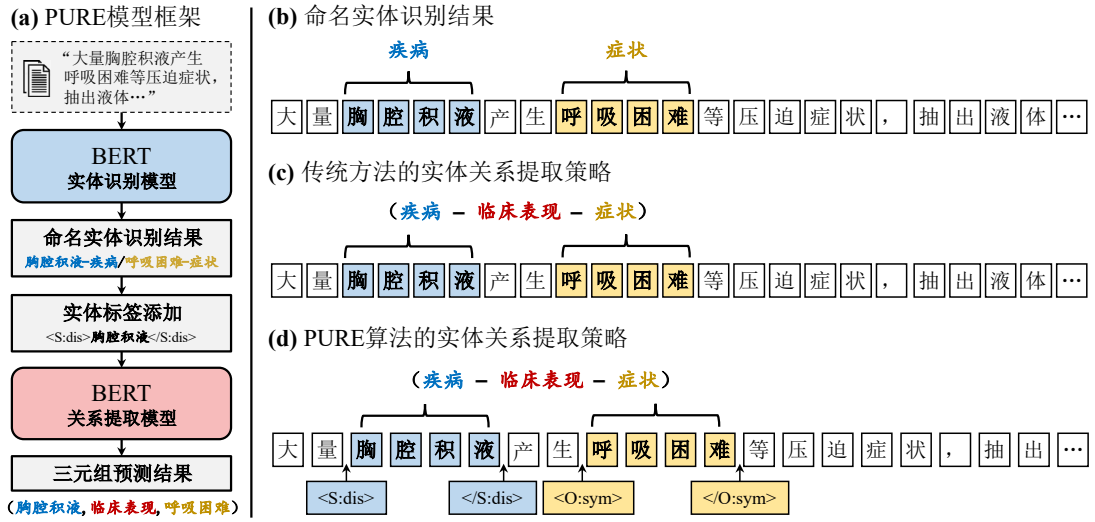


图 5-5　PURE 实体识别与关系提取模型

公开的 CMeIE 数据集提供了丰富的医疗文本、实体与三元组信息，本节以此数据集作为命名实体识别和关系抽取模型的训练语料来源。命名实体模型配置中：学习率设定为 5e-4，训练轮次为 60 轮，批次大小设置为 64，使用 HuggingFace 平台提供的 BERT-Base-Chinese 预训练模型初始化 BERT 模型参数；关系提取模型配置中：学习率设定为 2e-5，训练轮次、批次大小和预训练模型加载方式与命名实体模型保持一致。

在正式完成知识提取之前，本节首先对教材中的全部段落进行语句分割，获取到包含 3,630 条医疗行为相关语句的集合。其次，使用在 CMeIE 数据集上充分训练的 PURE 模型对语句集合进行命名实体识别与实体关系提取。最终获取

到的实体和三元组提取结果统计数量如图 5-6 所示。医疗行为文本知识图谱共含有 5,052 个实体和 2,087 个三元组。图 5-7 对三元组案例进行了展示。

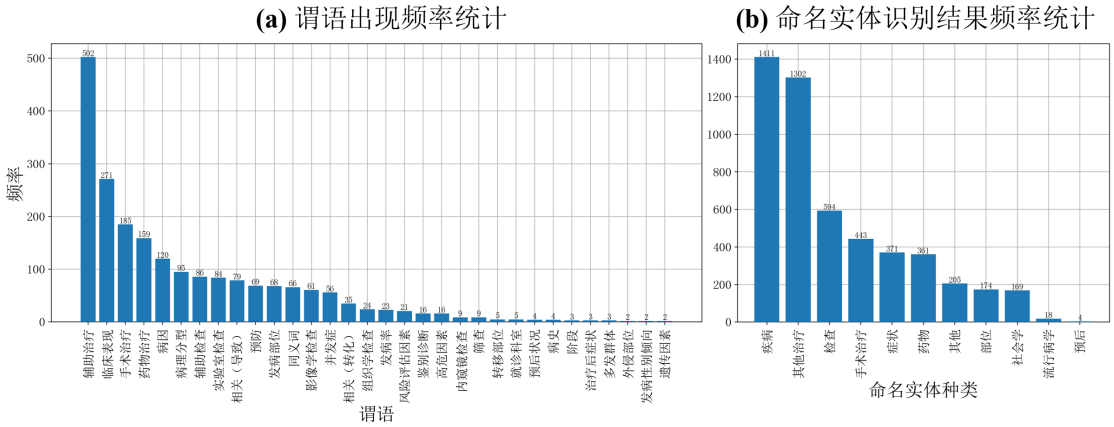


图 5-6 时序医疗行为知识图谱的谓语和实体类型统计

```
{
  "text": "诊断作用：抽取少量胸腔内液体标本检测，以明确胸腔积液病因。",
  "triple_list": [
      ["胸腔积液", "疾病-组织学检查-检查", "抽取少量胸腔内液体标本检测"],
      ["胸腔积液", "疾病-实验室检查-检查", "抽取少量胸腔内液体标本检测"]
  ]
},
{
  "text": "大量胸腔积液产生呼吸困难等压迫症状，抽出液体促进肺复张，缓解症状。",
  "triple_list": [
      ["胸腔积液", "疾病-临床表现-症状", "呼吸困难"],
      ["胸腔积液", "疾病-辅助治疗-其他治疗", "抽出液体"]
  ]
},
{
  "text": "胸膜反应：穿刺中患者出现头晕、气促、心悸、面色苍白、血压下降。停止操作，平
卧，皮下注射 0.1%肾上腺素 0.3~0.5ml。",
  "triple_list": [
      ["胸膜反应", "疾病-临床表现-症状", "面色苍白"],
      ["胸膜反应", "疾病-临床表现-症状", "心悸"],
      ["胸膜反应", "疾病-临床表现-症状", "头晕"],
      ["胸膜反应", "疾病-临床表现-症状", "血压下降"],
      ["胸膜反应", "疾病-临床表现-症状", "气促"],
      ["胸膜反应", "疾病-辅助治疗-其他治疗", "停止操作"]
  ]
},
{
  "text": "其他检查：必要时可行胸部 X 线检查以评价胸腔残余积液量和排除气胸。",
  "triple_list": [
      ["气胸", "疾病-影像学检查-检查", "胸部 X 线检查"]
  ]
}
```

图 5-7 教材语句与三元组案例展示

5.3.4 时序医疗行为知识图谱的细化拓展

以上两个小节分别依据预定义的行为 Schema 和文本知识 Schema 构建了医

疗行为流程知识图谱和**医疗行为文本知识图谱**。虽然医疗行为流程知识图谱已经将医疗行为划分到了“二级操作步骤”的细粒度，但是这仍然无法满足时序医疗行为分析任务的要求。本章的首要目标是构建高质量、高细粒度的医疗时序行为分析数据集，其次是设计高精度、高效率的行为分析算法。因此本节以胸腔穿刺术（Thoracentesis）为研究对象深入探究了医疗行为的拆解，并且通过人工构建的方式提出了高细粒度的时序行为知识图谱，为后续时序行为分析数据集的构建奠定了基础。

胸腔穿刺术简称胸穿，是指对有胸腔积液或气胸的患者通过胸腔穿刺抽取积液或气体，从而达成实施治疗或进一步诊断目的的技术。胸腔穿刺术是可以是诊断性的，例如对原因未明的胸腔积液进行诊断性穿刺，进而作胸腔积液涂片、培养、细胞学和生化学检查以明确病因；也可以是治疗性的，例如通过抽液、抽气操作治疗胸腔大量积液、积气等症状，或向胸腔内注射药物。本章综合考虑了胸腔穿刺术的重要性与复杂性，选取其作为本章的研究对象：一方面，胸腔穿刺术是医学生临床技能操作中最基础、最重要的必备操作技能之一；另一方面，胸腔穿刺术流程复杂、考核点繁多，具备一定的操作难度。后续医疗行为知识图谱的细化拓展、时序医疗行为分析数据集的构建都将围绕胸腔穿刺术展开。

如表 5-5 所示，教材将胸腔穿刺操作划分为 9 个流程。本节以教材划分方式为基础，进一步将胸腔穿刺术划分为 12 个流程和 39 个子行为。并且在医院教培中心老师的指导下总结了 22 种遗漏错误行为、23 种单流程错误行为类和 4 种操作结果。表 5-7 对细化后的胸腔穿刺术拆解情况进行了列举。图 5-8 对拓展后的医疗行为知识图谱进行了可视化。给定所有细化后的行为，本章共总结了五种错误操作类型：单流程遗漏类、单流程错误类、流程遗漏导致后续不合格类、流程错误导致后续不合格类、顺序错误类。表 5-8 对各错误类别进行了举例。单流程遗漏类和单流程错误类是考核场景中最常见的两类错误，操作者往往会因为粗心、紧张出现环节遗漏或错误操作，这类错误并不会对后续的操作结果造成影响，在真实的考核场景中只会被扣除对应的分数。而在流程遗漏导致后续不合格类中，流程的遗漏和后续的错误存在一定的因果关系，例如“48_遗漏手套”会导致“64_无手套镊子夹取”“65_无手套消毒”和“69_无手套洞巾”三个错误的发生，“56_遗漏打开止血夹”会导致“80_没有抽出积液”错误的发生。同理，在流程错误导致后续不合格类中，这种因果关系也存在，例如“63_标记点位置错误”“78_插入太浅&穿刺失败”和“76_未关止血夹”均会引起“80_没有抽出积液”错误的发生。最后，顺序错误类指操作行为的顺序发生了颠倒，可能会导致无菌环境破坏、穿刺失败等后果。

表 5-7 细化后的胸腔穿刺术拆解

序号	流程划分	细分项	遗漏错误行为	单流程错误行为类
1	通知胸穿	0_通知穿刺	39_遗漏通知	
2	洗手	1_七步洗手法	40_遗漏洗手	
3	摆体位	2_通知摆体位&前扶	41_遗漏通知	
4	暴露胸廓	3_通知暴露胸廓&脱衣	42_遗漏通知	
5	叩诊	4_叩诊寻找位置	43_遗漏叩诊	61_叩诊左手错误 62_叩诊单纯右手
		5_记号笔标记位置	44_遗漏标记点	63_标记点位置错误
6	检查穿刺包	6_通知穿刺包良好	45_遗漏检查	
7	消毒	7_通知消毒	46_遗漏通知	
		8_卵圆钳夹棉球	47_遗漏卵圆钳夹取	
		9_戴手套	48_遗漏手套	
		10_镊子夹取棉球		64_无手套镊子夹取
		11_螺旋消毒操作		65_无手套消毒 66_面积过小 67_转圈动作逆序
		12_废弃棉球		68_废弃到白盘
8	铺洞巾	13_拾取&铺洞巾操作	49_遗漏洞巾操作	69_无手套洞巾
9	核对+利多卡因抽取	14_通知麻醉	50_遗漏通知	
		15_核对利多卡因	51_遗漏核对	
		16_拾取小注射器		70_拿大注射器
		17_抽取利多卡因		71_未抽到
10	麻醉	18_麻醉前拿起纱布	52_遗漏纱布拿起	
		19_左手控制穿刺位置		72_未控制麻醉位置
		20_斜向出丘疹	53_遗漏丘疹	
		21_渐进式麻醉		73_未渐进麻醉
		22_末端回抽	54_遗漏回抽	
		23_纱布覆盖&拔针		74_纱布未覆盖点
		24_废弃纱布		75_未废弃纱布
11	穿刺	25_通知穿刺	55_遗漏通知	
		26_胸穿针&关止血夹		76_未关止血夹
		27_左手控制位置		77_未定位穿刺点
		28_完成穿刺		78_插入太浅
		29_接大注射器		79_选错注射器
		30_打开止血夹	56_遗漏打开止血夹	
		31_抽出积液		80_没有抽出积液 81_单手抽取
		32_关止血夹	57_遗漏关闭止血夹	
		33_放下注射器		
		34_拿纱布	58_遗漏纱布	
		35_拔针覆盖纱布		82_直接拔针
		36_废弃纱布		83_未废弃纱布
12	术后	37_贴创可贴	59_遗漏创可贴步骤	
		38_通知结束	60_遗漏结束通知	

表 5-8　错误分类与因果关系总结

错误类型	案例
单流程遗漏类	39_遗漏通知 40_遗漏洗手 ……
单流程错误类	61_叩诊左手错误 73_未渐进麻醉 ……
流程遗漏导致后续不合格类	48_遗漏手套 ⇒ 64_无手套镊子夹取 48_遗漏手套 ⇒ 65_无手套消毒 48_遗漏手套 ⇒ 69_无手套洞巾 52_遗漏纱布拿起 ⇒ 74_纱布未覆盖点 52_遗漏纱布拿起 ⇒ 75_未废弃纱布 56_遗漏打开止血夹 ⇒ 80_没有抽出积液 58_遗漏纱布 ⇒ 82_直接拔针无覆盖动作 58_遗漏纱布 ⇒ 83_未废弃纱布
流程错误导致后续不合格类	63_标记点位置错误 ⇒ 80_没有抽出积液 78_插入太浅&穿刺失败 ⇒ 80_没有抽出积液 76_未关止血夹 ⇒ 80_没有抽出积液
顺序错误类	——

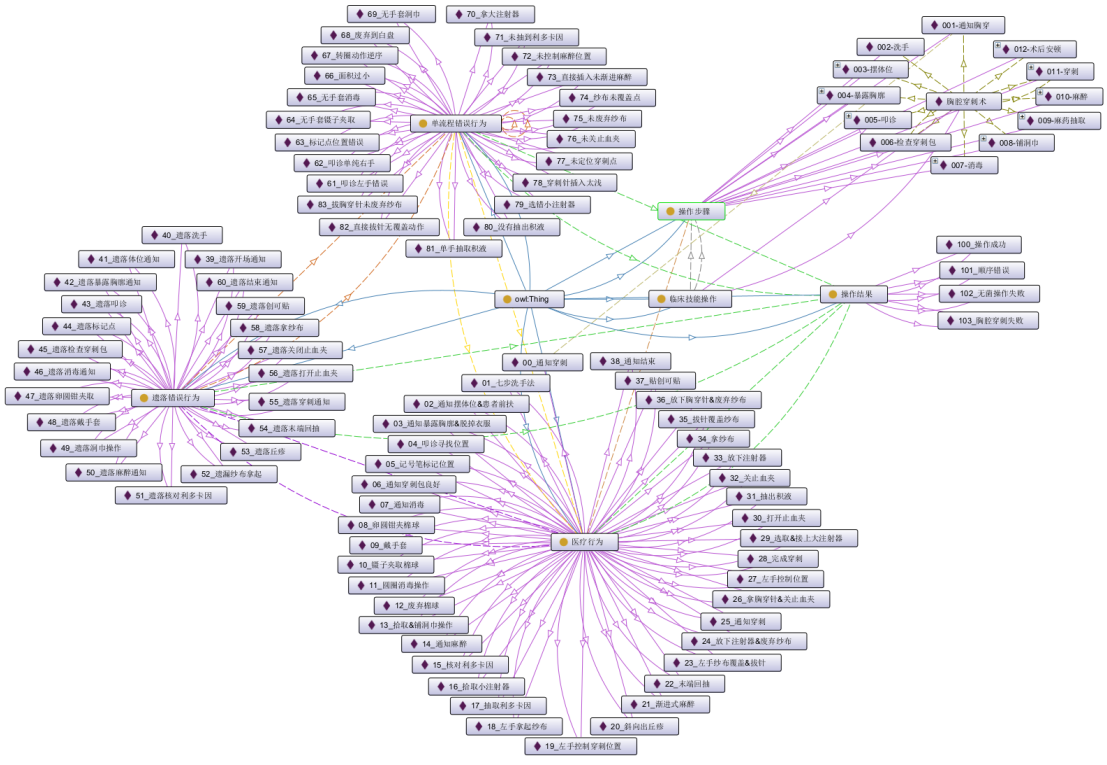


图 5-8　拓展后的胸腔穿刺术时序医疗行为知识图谱

5.4 基于时序聚类注意力机制的扩散时序行为分析算法

5.2 节对现有的时序行为分割数据集和算法进行了回顾。在时序行为分割任务中，数据集中视频的帧数可达数万。而这些算法通常将整个特征序列直接映射到行为标签空间，忽略了时序上的差异性，进而导致行为分割的性能下降。针对此问题，本文受人类大脑功能分区结构启发，提出了基于时序聚类注意力机制的特征增强模块 *k*M-Att 完成视频序列特征的有效增强。

另一方面，模型推理的延时也是评判优劣性的重要指标。本章所提出的模型充分借鉴了 DiffAct 模型的结构，即在解码器生成结果的阶段使用扩散解码策略提升分割准确率。虽然 DiffAct 在公开数据集上取得了可观的性能，但是仍然面临着推理时间长的缺点。实验发现解码器需要迭代 25 次才能生成较好的分割结果，解码器的推理耗时是扩散分割模型推理延时的瓶颈。针对此问题，本文提出了基于非锁步跳跃的扩散去噪机制，有效地削减了解码器迭代次数且不降低模型性能。

此节的内容安排如下：5.4.1 小节对框架整体进行描述；5.4.2 小节详细阐述人类大脑功能分区如何启发本文提出时序聚类注意力特征增强模块 *k*M-Att；5.4.3 小节阐述了基于非锁步跳跃的扩散去噪机制原理；5.4.4 小节阐述医疗行为时序分析下的合规性评估模块设计。

5.4.1 扩散时序行为分割框架

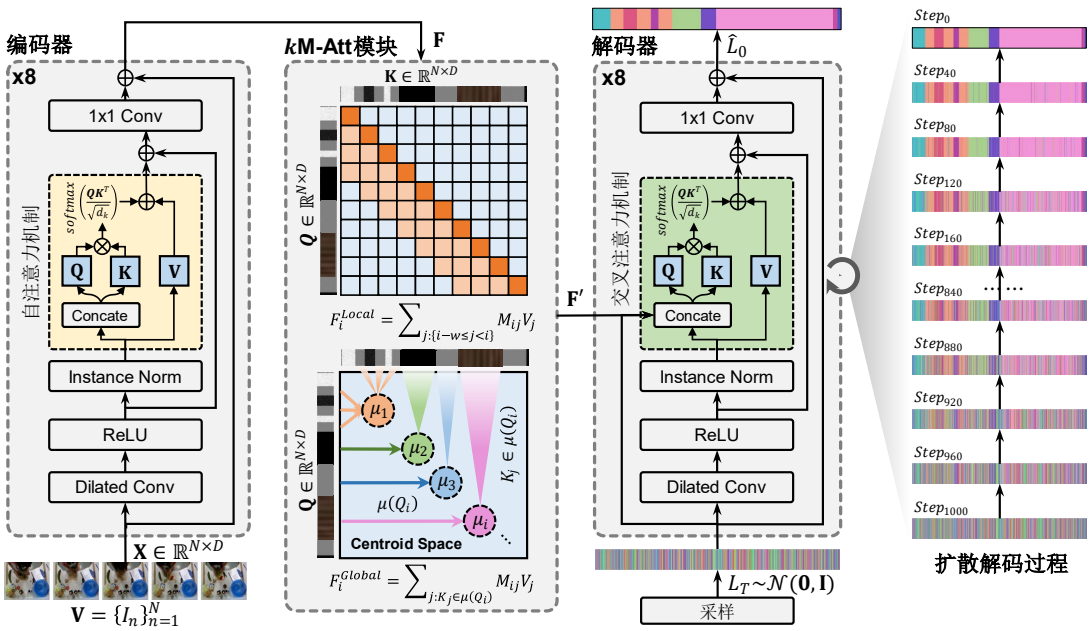


图 5-9 扩散时序行为分割框架示意图

本章所提出的时序行为分割框架如图 5-9 所示，此框架由三部分构成：编码

器（Encoder）、*k*M-Att 特征增强模块和解码器（Decoder）。

给定长度为N的视频特征$\mathbf{V}=\{I_n\}_{n=1}^{N}, I_n\in\mathbb{R}^{H\times W\times 3}$，模型需要为每一帧$I_n$分配一个行为标签$l_n\in\{c_1,c_2,\cdots,c_M\}$，其中$M$表示行为的种类数量。通常视频长度$N$可达$10^3$或$10^4$数量级。一般情况下，时序行为分割模型的输入均是已预先提取好的视频特征$\mathbf{X}\in\mathbb{R}^{N\times D}$，其中$D$为特征的维度。

编码器将视频序列特征$\mathbf{X}$映射为中间特征$\mathbf{F}\in\mathbb{R}^{N\times D}$：

$$\mathbf{F}=\mathcal{F}_{Encoder}(\mathbf{X}) \tag{5.12}$$

如图 5-9 所示，编码器由多个编码器层堆叠而成。每个编码器层由空洞卷积层（Dilated Convolution Layer）、ReLU 激活函数、实例归一化层（Instance Normalization Layer）和自注意力机制构成。其中空洞卷积层使用不同的步长对视频序列进行特征提取。

中间特征$\mathbf{F}$会被送入 *k*M-Att 模块完成时序特征增强。其中 *k*M-Att 模块包含两个支路：局部注意力机制通路和 *k*-means 时序聚类注意力机制通路：

$$\mathbf{F}'=\mathcal{F}_{kMAtt}(\mathbf{F}) \tag{5.13}$$

最终由解码器通过扩散去噪过程（Diffusion Denoising Process）完成初始标签序列$\hat{L}_0$的预测：

$$\hat{L}_0=\mathcal{F}_{Decoder}(\mathbf{F}', L_T, T) \tag{5.14}$$

其中T为扩散解码过程的总步数，L_T为初始随机噪音。解码器层的结构与编码器层类似，最大差异在于解码器使用交叉注意力机制，在解码过程中需要使用到中间特征$\mathbf{F}'$。

5.4.2 基于时序聚类注意力机制的特征增强模块 *k*M-Att

传统的时序行为分割算法通常将所有的帧特征进行无差别建模，例如 ASFormer[87]、UVAST[190]和 DiffAct[92]。这种建模方法忽略了特征在时序上的差异性。在长度达到10^4数量级的序列中，每个行为都会持续一段相当长的时间，从而会导致大量相同标签的帧紧邻在一片区域，最终产生大量的冗余计算。

为了解决以上问题，本文由大量的视频帧联想到了人类大脑的功能分区机制：一个成年人的大脑中含有 850~860 亿个神经元，大脑的不同功能区域将这些巨量神经元进行分区管理，让这些神经元高效地协同完成各种复杂、高级的认知功能：前额叶（Frontal Lobe）负责情绪、认知、决策、判断、社交行为等高级认知类功能；后额叶（Occipital Lobe）负责视觉处理与视觉感知功能；顶叶（Parietal Lobe）负责空间感知、抓握等功能；颞叶（Temporal Lobe）负责听觉、语言、记忆等功能；小脑（Cerebellum）负责身体平衡与运动控制。如图 5-10 所示：

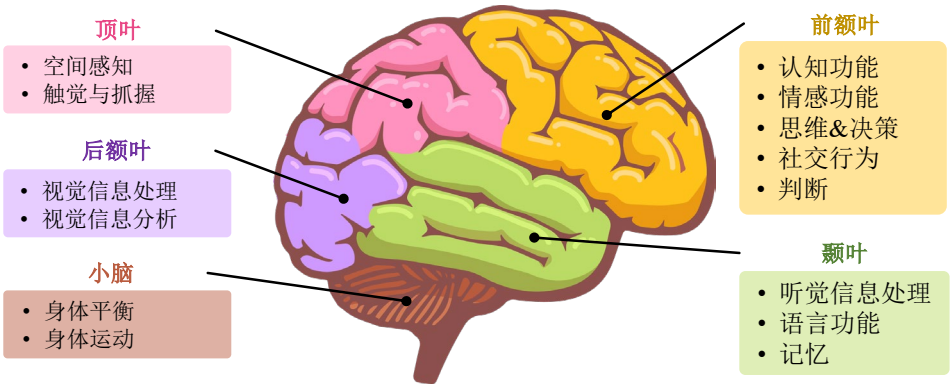


图 5-10 人类大脑功能分区示意图

为了验证这种仿生结构设计的合理性，本文首先从现有时序行为分割数据集中随机挑选了视频，并使用 t-SNE 算法对视频特征进行了可视化。以 Breakfast 数据集中的 Webcam02_P45_pancake.avi 为例，此视频共含有 8,295 帧。可视化结果和 Ground-truth 标签序列如图 5-11(a)所示，本文为不同的行为标签分配了不同的颜色。t-SNE 特征可视化结果和人类大脑的结构颇为相似：时序上临近的特征具有更高的类内相似性，因此类内特征被映射到同一条轨迹上；而不同类之间的特征相互排斥，因此不同种类的行为被映射到不同的区域中。

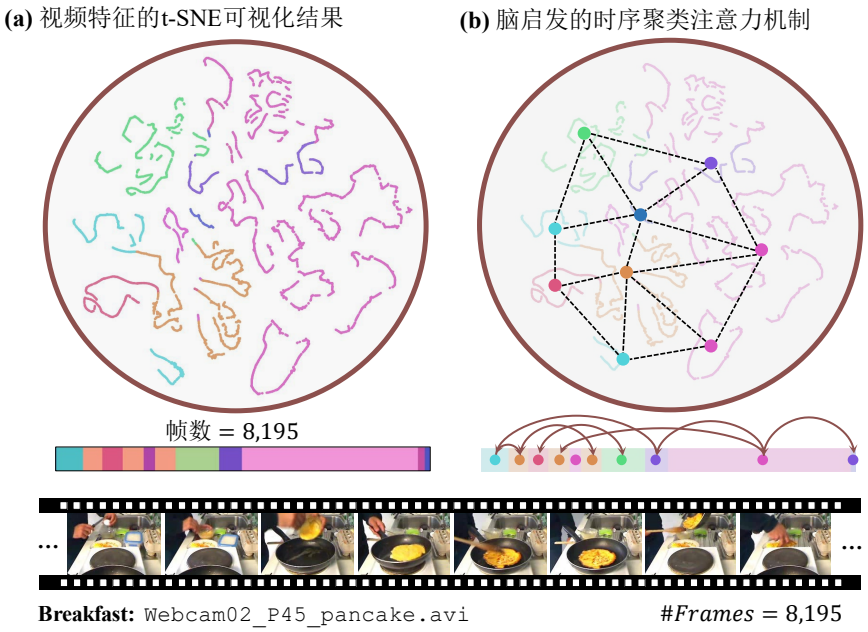


图 5-11 视频特征 t-SNE 可视化结果和时序聚类注意力机制原理

基于以上观察，本文对大脑中神经元的连接与交互方式进行了更加深入地探究。图 5-12(a)(b)分别从微观和宏观的角度绘制了大脑中的神经元连接关系。从局部来看，物理临近交互（Physical Proximity Interactions）行为完成各神经元之间的信息传递；从全局来看，逻辑分区交互（Logical Partitioning Interactions）行

为完成各神经元在功能区域内的信息传递。为了对这两种交互行为进行模拟，本文提出了 *k*M-Att 模块。如图 5-9 所示，*k*M-Att 模块由局部注意力机制和 *k*-means 聚类注意力机制两个分支构成，其中后者的设计充分参考了自然语言处理域中的 Routing Transformer[191]模型的结构。

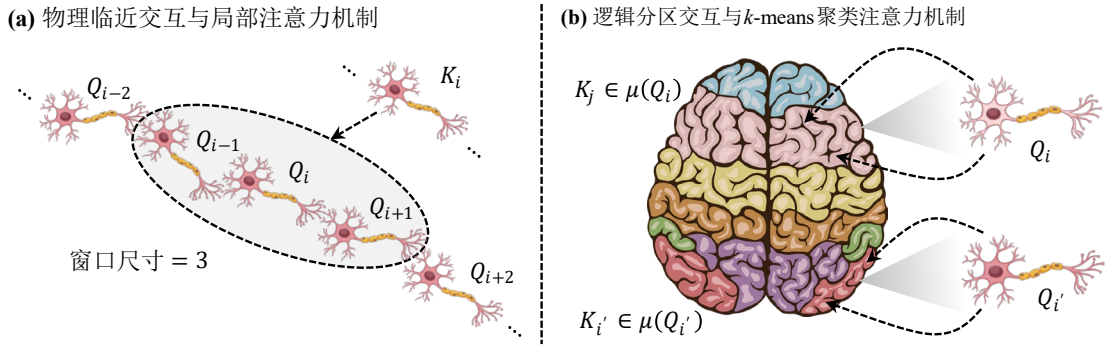


图 5-12　物理临近交互与逻辑分区交互展示图

局部注意力机制设计

编码器输出的特征$\mathbf{F}$通过三个线性层映射为$\mathbf{Q},\mathbf{K},\mathbf{V}$。这些特征参与后续注意力机制的计算。注意力图$\mathbf{M}\in\mathbb{R}^{N\times N}$的计算方法为：

$$\mathbf{M}=\mathrm{softmax}\left(\frac{\mathbf{QK}^{T}}{\sqrt{d_k}}\right) \tag{5.15}$$

通过设定注意力窗口w，可以通过以下公式计算$\mathbf{F}^{Local}$：

$$F_i^{Local}=\sum_{j:\{i-w\le j<i\}} M_{ij}V_j \tag{5.16}$$

图 5-9 中给出了注意力窗口为 4 的注意力图实例，图 5-12 使用神经元形象展示了窗口长度为 3 的交互情况。局部注意力交互机制实现了窗口内物理相邻的神经元之间的交互建模。

时序 *k*-means 聚类注意力机制设计

传统的 *k*-means 聚类算法通过在数据集中随机选择 *k* 个样本进行初始化，而在深度学习模型训练时，数据则是以分批次的形式进行加载。为解决此问题，本文将 *k* 个中心特征定义为神经网络中可以学习的参数：

$$\boldsymbol{\mu}=\{\mu_1,\mu_2,\cdots,\mu_k\},\mu_k\in\mathbb{R}^{D} \tag{5.17}$$

对于$\mathbf{Q}$中的每一个样本Q_i，可以通过$\mathbf{Q}$内的聚类可以找到对应的$\mu(Q_i)\in\boldsymbol{\mu}$。同理可以在$\mathbf{K}$内部进行聚类，为每一个$K_i$找到对应的$\mu(K_i)\in\boldsymbol{\mu}$。$\mathbf{Q}$和$\mathbf{K}$共享特征集合$\boldsymbol{\mu}$，因此$\boldsymbol{\mu}$扮演了$\mathbf{Q}$和$\mathbf{K}$之间的桥梁。通过以下公式实现 *k*-means 聚类注

意力机制，完成$\mathbf{F}^{Global}$计算：

$$F_i^{Global}=\sum_{j:\{\mu(K_j)=\mu(Q_i)\}} M_{ij}V_j \tag{5.18}$$

k-means 聚类注意力机制的本质是在进行最大内积搜索（Maximum Inner Product Search, MIPS），即给定Q_i，寻找一系列Q_k满足：

$$j=\underset{j\in\{1,2,\cdots,n\}}{\operatorname{argmax}} Q_i^T K_j \tag{5.19}$$

获取局部注意力机制通路提取到的特征$\mathbf{F}^{Local}$和时序 *k*-means 聚类注意力机制通路提取到的特征$\mathbf{F}^{Global}$之后，使用特征拼接和线性投影层进行特征映射：

$$\mathbf{F}=\mathcal{F}_{Linear}(\mathbf{F}^{Local}\oplus\mathbf{F}^{Global}) \tag{5.20}$$

其中$\oplus$表示特征拼接操作。在网络的训练过程中，每一个聚类中心点通过指数移动平均（Exponential Moving Average, EMA）进行动态更新：

$$\mu_k\leftarrow\lambda\mu_k+\frac{1-\lambda}{2}\sum_{i:\{\mu(Q_i)=\mu_k\}}Q_i+\frac{1-\lambda}{2}\sum_{j:\{\mu(K_j)=\mu_k\}}K_j \tag{5.21}$$

算法 1 在编程层次对时序 *k*-means 聚类注意力机制的实现进行了详细阐述：

算法 1：时序 *k*-means 聚类注意力机制实现流程

Input：视频特征$\mathbf{F}\in\mathbb{R}^{N\times D}$

Parameter：中心点集合$\boldsymbol{\mu}\in\mathbb{R}^{k\times D}$，衰减参数$\lambda=0.999$

Output：视频特征$\mathbf{F}^{Global}\in\mathbb{R}^{N\times D}$

// 线性层映射&归一化层

$\mathbf{Q}\leftarrow\mathbf{F}\mathbf{W}_Q,\mathbf{K}\leftarrow\mathbf{F}\mathbf{W}_K,\mathbf{V}\leftarrow\mathbf{F}\mathbf{W}_V$

$\mathbf{Q}\leftarrow\mathrm{LayerNorm}(\mathbf{Q}),\mathbf{K}\leftarrow\mathrm{LayerNorm}(\mathbf{K})$

// *k*-means 聚类注意力机制

$\widehat{\mathbf{Q}}\leftarrow\boldsymbol{\mu}\mathbf{Q}^T,\widehat{\mathbf{K}}\leftarrow\boldsymbol{\mu}\mathbf{K}^T$

$Q_{idx}\leftarrow\mathrm{Sort}[\text{Top-k}(\widehat{\mathbf{Q}},N/k)]$

$K_{idx}\leftarrow\mathrm{Sort}[\text{Top-k}(\widehat{\mathbf{K}},N/k)]$

// 收集选中特征

$\mathbf{Q}'\leftarrow\mathbf{Q}[Q_{idx}],\mathbf{K}'\leftarrow\mathbf{K}[K_{idx}],\mathbf{V}'\leftarrow\mathbf{V}[K_{idx}]$

// 生成注意力图

$\mathbf{M}\leftarrow\mathrm{softmax}(\mathbf{Q}'\mathbf{K}'^T/d_{K'})$

$\mathbf{V}'\leftarrow\mathbf{M}\mathbf{V}'$

$\mathbf{F}\leftarrow\mathbf{V}'[K_{idx}]$

// 更新中心点

$\mathbf{Q}_m\leftarrow\text{One-hot}[\mathrm{argmax}(\widehat{\mathbf{Q}}))]$

$\mathbf{K}_m\leftarrow\text{One-hot}[\mathrm{argmax}(\widehat{\mathbf{K}}))]$

$\boldsymbol{\mu}\leftarrow\lambda\boldsymbol{\mu}+(1-\lambda)\mathbf{Q}_m\widehat{\mathbf{Q}}/2+(1-\lambda)\mathbf{K}_m\widehat{\mathbf{K}}/2$

return $\mathbf{F}^{Global}$

其中$\widehat{\mathbf{Q}}$表示中心集合$\boldsymbol{\mu}$和特征$\mathbf{Q}^T$的乘积，$\widehat{\mathbf{K}}$表示$\boldsymbol{\mu}$和特征$\mathbf{K}^T$的乘积。Q_{idx}表示$\mathbf{Q}$中被选中特征的下标索引。在时序 *k*-means 聚类注意力机制的计算过程中，三个部分的计算可能会成为性能瓶颈，分别为：自注意力机制计算的复杂度为$O(N^2D/k)$、聚类操作的计算复杂度为$O(NDk)$、排序算法的计算复杂度为$O(N\log N)$。当聚类中心的数量取$k=\sqrt{N}$时，自注意力机制和聚类操作的复杂度均变为$O(N^{1.5}D)$，此时排序算法$O(N\log N)$的计算复杂度可以被忽略。最终此模块的计算复杂度为$O(N^{1.5}D)$，依然小于自注意力机制计算中$O(N^2D)$的复杂度。因此在处理超长序列时，本文所提出的 *k*M-Att 模块并不会成为计算瓶颈。

5.4.3 基于非锁步跳跃的扩散去噪机制

Liu 等人[92]提出的 DiffAct 框架是第一个将扩散模型引入到时序行为分割任务中的模型，本小节首先对时序行为分割任务中的扩散模型使用流程进行简要回顾，再着重介绍本文所提出的非锁步跳跃扩散去噪机制。

给定一个原始 Ground-truth 标签序列$L_0=\{l_n\}_{n=1}^{N}, L_0\in\mathbb{R}^N$，前向过程以迭代的方式对$L_0\sim q(L_0)$添加高斯噪声。设总添加步数为$T$次，则可以得到加噪序列$\{L_1,L_2,\cdots,L_T\}$。而去噪过程以迭代的方式将纯高斯噪声$L_T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$转化为原始标签$L_0$，从而得到去噪序列$\{L_T,L_{T-1},\cdots,L_0\}$。

具体而言，在前向过程中首先确定每一步的衰减参数$\{\beta_1,\beta_2,\cdots,\beta_T\}$。加噪过程可以表示为：

$$q(L_t|L_{t-1})=\mathcal{N}\left(L_t;\sqrt{1-\beta_t}L_{t-1},\beta_t\mathbf{I}\right) \tag{5.22}$$

使用$\alpha_t=1-\beta_t$进行替换，可以直接由初始标签L_0得到第t步加噪后的标签：

$$q(L_t|L_0)=\mathcal{N}\left(L_s;\sqrt{\bar{\alpha}_t}L_0,(1-\bar{\alpha}_t)\mathbf{I}\right) \tag{5.23}$$

其中衰减参数的累乘项$\bar{\alpha}_t$定义为：

$$\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i \tag{5.24}$$

通过重参数化技巧（Reparameterization Trick），L_t可由L_0通过单步加噪得到：

$$L_t=\sqrt{\bar{\alpha}_t}L_0+\sqrt{1-\bar{\alpha}_t}\cdot\mathcal{N}(\mathbf{0},\mathbf{I}) \tag{5.25}$$

逆向去噪的过程可以表示为：

$$p_\theta(L_{t-1}|L_t)=\mathcal{N}(L_{t-1};\boldsymbol{\mu}_\theta(L_t,t),\sigma_t^2\mathbf{I}) \tag{5.26}$$

其中均值$\boldsymbol{\mu}_\theta(L_t,t)$通过神经网络直接预测得到，标准差$\sigma_t$由噪声衰减系数$\alpha_t$计算得到：

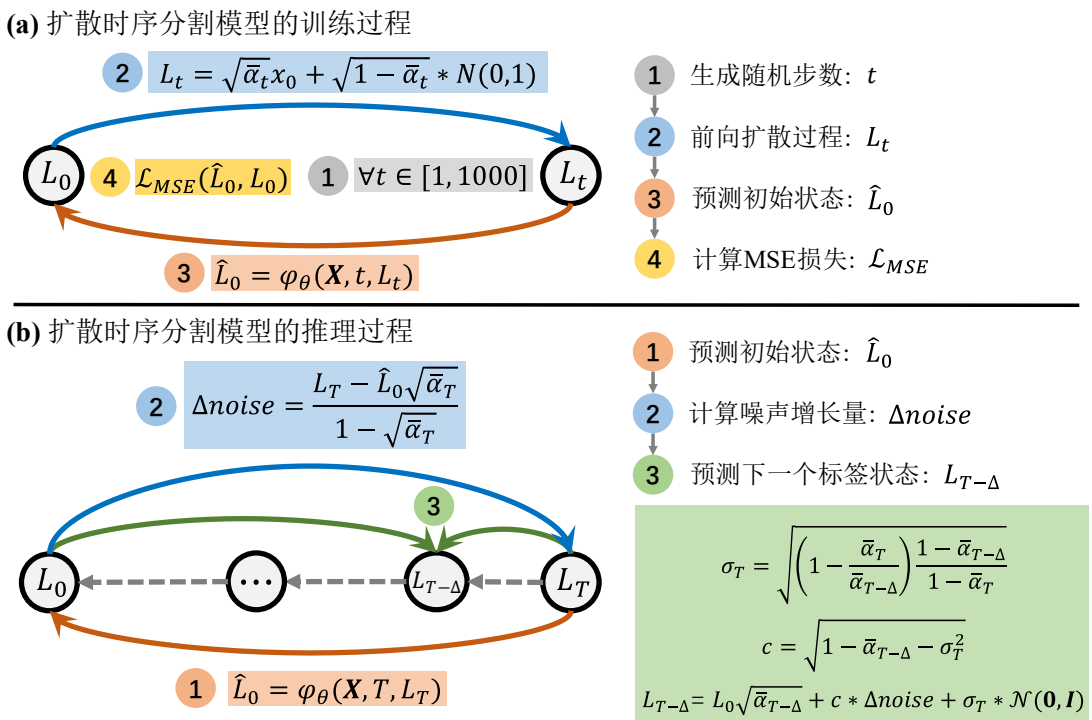


图 5-13 扩散时序分割模型的训练、推理过程与扩散去噪机制对比

$$\sigma_t = \sqrt{\left(1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}\right)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \tag{5.27}$$

与 DDIM[192]模型一致，为了避免预测均值$\boldsymbol{\mu}_\theta(L_t,t)$，本文使用神经网络预测初始标签$L_0$：

$$\hat{L}_0 = \varphi_\theta(L_t, t, \mathbf{X}) \tag{5.28}$$

其中θ为神经网络的参数，$\mathbf{X}$为输入视频特征。给定初始标签预测结果$\hat{L}_0$、含噪声标签L_t和时刻t，可通过如下公式计算L_{t-1}：

$$L_{t-1} = \frac{L_t - \hat{L}_0\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} + \hat{L}_0\sqrt{\bar{\alpha}_{t-1}} + \sigma_t\mathcal{N}(\mathbf{0},\mathbf{I}) \tag{5.29}$$

根据扩散过程的确定性（Determinacy），可以使用$\Delta noise$表示 0 到t时刻的噪声增量，因此公式 5.18 可表示为：

$$\begin{aligned} L_{t-1} &= \Delta noise\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} + \hat{L}_0\sqrt{\bar{\alpha}_{t-1}} + \sigma_t\mathcal{N}(\mathbf{0},\mathbf{I}) \\ \Delta noise &= \frac{L_t - \hat{L}_0\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} \end{aligned} \tag{5.30}$$

以迭代的方式重复此过程，即可将原始高斯噪声$L_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$逐步恢复为预测标签$L_0$。通过均方误差（Mean Squared Error）度量预测值与 Ground-truth 标签的差异性：

$$\mathcal{L}_{MSE} = \left\| \widehat{L}_0, L_0 \right\|^2 \tag{5.31}$$

图 5-13(a)(b)细展示了扩散时序分割模型的训练和推理过程。

传统的锁步跳跃去噪机制通过设定固定步长Δ完成标签序列的解码。图 5-13(c)中的左图展示了$\Delta = 40$的解码过程，在总步数$T = 1000$的设定下，解码器需要运行$1000 \div 40 = 25$次。本文认为这种固定步长的解码机制会忽视早期解码过程（Early Denoising Stage）和晚期解码过程（Late Denoising Stage）之间的差异性：早期解码过程更加关注行为区域的宏观划分，而晚期解码过程则更加关注行为边缘的精修。不恰当的跳跃机制会造成迭代次数的浪费和分割性能的降低。

为了解决上述问题，本文提出了非锁步跳跃解码机制（Unlocked Skip Diffusion Denoising Mechanism）：在早期解码阶段中使用大跳跃步长Δ，在晚期解码阶段中使用小跳跃步长Δ'。最终的解码序列为：

$$\{L_T, L_{T-\Delta}, L_{T-2\Delta}, \cdots, L_{2\Delta'}, L_{\Delta'}, L_0\} \tag{5.32}$$

图 5-13(c)对两种不同的扩散去噪机制进行了对比。后续定性实验和定量实验证实，非锁步跳跃解码机制相较于原始策略能够以更少的迭代次数取得更优的时序分割结果。

5.4.4 损失函数设计

与 DiffAct[92]模型保持一致，本模型在训练过程中使用三种损失函数对预测序列$\hat{L}_0$和标准序列L_0之间的差异进行度量：交叉熵损失$\mathcal{L}_{CE}$、时序平滑损失$\mathcal{L}_{Smooth}$和边界对齐损失$\mathcal{L}_{Align}$。

交叉熵损失$\mathcal{L}_{CE}$对帧级别的预测差异性进行度量，计算方式为：

$$\mathcal{L}_{CE} = \sum_{n=1}^{N-1} \sum_{m=1}^{M} - L_0[n,m] \cdot \log \hat{L}_0[n,m] \tag{5.33}$$

其中N表示视频序列长度，M表示预测标签的总类别数量。$\hat{L}_0[n,m]$表示模型对第n帧第m类的预测概率。

时序平滑损失$\mathcal{L}_{Smooth}$能够对标签的局部相似性进行度量，从而降低网络预测结果中的标签切换频率。时序平滑损失使用均方根误差进行计算：

$$\mathcal{L}_{Smooth} = \frac{1}{(N-1)M} \sum_{n=1}^{N-1} \sum_{m=1}^{M} \left(\log \hat{L}_0[n,m] - \log \hat{L}_0[n+1,m] \right)^2 \tag{5.34}$$

边界对齐损失$\mathcal{L}_{Align}$的功能是度量模型对行为切换位置预测的偏差。给定标准序列L_0，首先依据如下规则生成边界序列$B\in\{0,1\}^{N-1}$，B中所有为 1 的位置指示行为标签切换的发生。

$$B_n=\begin{cases}0,\ L_0[i]=L_0[i+1]\\ 1,\ L_0[i]\neq L_0[i+1]\end{cases} \tag{5.35}$$

其次使用高斯滤波器对边界序列B进行处理，生成平滑后的边界序列$\bar{B}=\lambda(B)$。使用二进制损失函数度量$\hat{L}_0$和$\bar{B}$之间的边界预测误差：

$$\begin{aligned}\mathcal{L}_{Align}=\frac{1}{N-1}\sum_{n=1}^{N-1}\Big[&-\bar{B}_n\log\left(1-\hat{L}_0[n]\cdot\hat{L}_0[n+1]\right)\\&-\left(1-\bar{B}_n\right)\log\left(\hat{L}_0[n]\cdot\hat{L}_0[n+1]\right)\Big]\end{aligned} \tag{5.36}$$

最终损失函数为以上三个损失之和：

$$\mathcal{L}=\mathcal{L}_{CE}+\mathcal{L}_{Smooth}+\mathcal{L}_{Align} \tag{5.37}$$

5.4.5 行为合规性检测算法

本节中的前四个小节实现了时序行为分割模型的构建，该模型能够为输入视频生成逐帧行为标签。但是这距离本章的核心研究目标：“实现时序医疗行为的合规性检测”仍有一定差距。为实现时序维度中的错误行为和遗漏行为检测，本节基于动态时间规整算法[193]（Dynamic Time Warping, DTW）提出了行为合规性检测算法。本文所提出的检测算法共分为两个环节：DTW 算法关联匹配环节，用于关联匹配预测序列$\hat{L}_0$和标准序列L_0，并根据匹配结果高亮出所有歧义位置；行为合规性检测环节，通过本文所设计的算法完成预测序列$\hat{L}_0$向错误行为标签的映射。

DTW 算法关联匹配

DTW 算法是一种利用动态规划方法度量两个序列之间相似度的算法，最初应用于语音识别任务。DTW 算法非常适合本文所面对的行为序列对比任务，因为不同的被试者在进行胸腔穿刺操作时有自己的节奏，会导致每个行为的长度发生变化。本文拟通过 DTW 算法首先完成预测序列$\hat{L}_0$和标准序列L_0之间的匹配。

图 5-14 对 DTW 算法生成的开销矩阵和序列标签进行了可视化。其中(a)图展示了两个正确操作序列（ID_1916.mp4 和 ID_1962.mp4）之间的匹配开销矩阵；(b)展示了错误操作序列（ID_2035.mp4）和正确序列（ID_1961.mp4）之间的匹配开销矩阵，并且对所有错误的行为标签进行了标注。图 5-15 以标准操作序列 ID_1961.mp4 为基准，对错误操作序列 ID_2035.mp4 中的不合规行为进行了标

注。DTW 算法通过构造开销矩阵的方法生成了最佳匹配路径，实现了两个序列在时序维度的对齐。这些对齐信息会被送入下一个环节中进行错误检测。

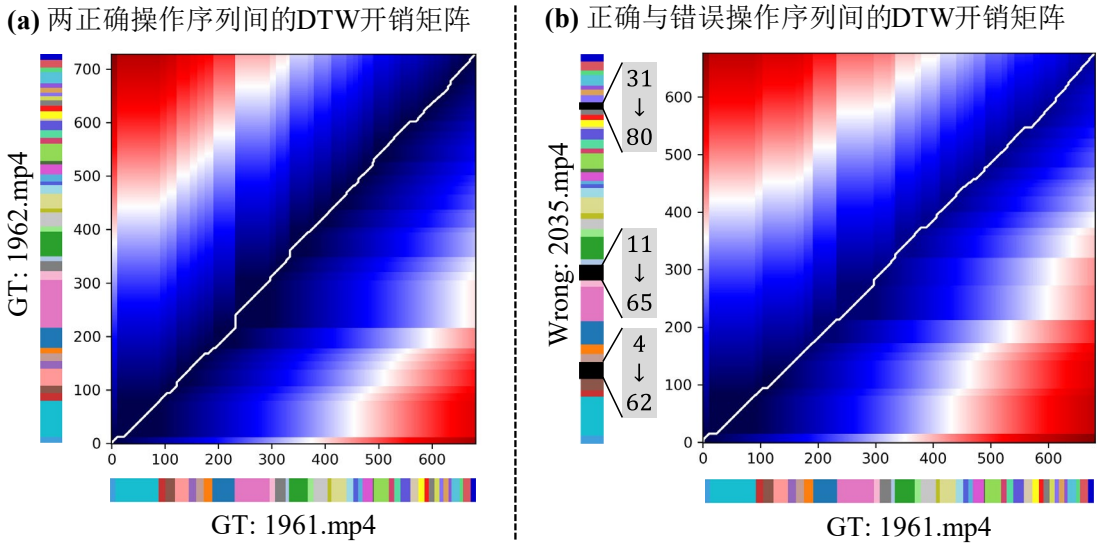


图 5-14　DTW 开销矩阵示例图

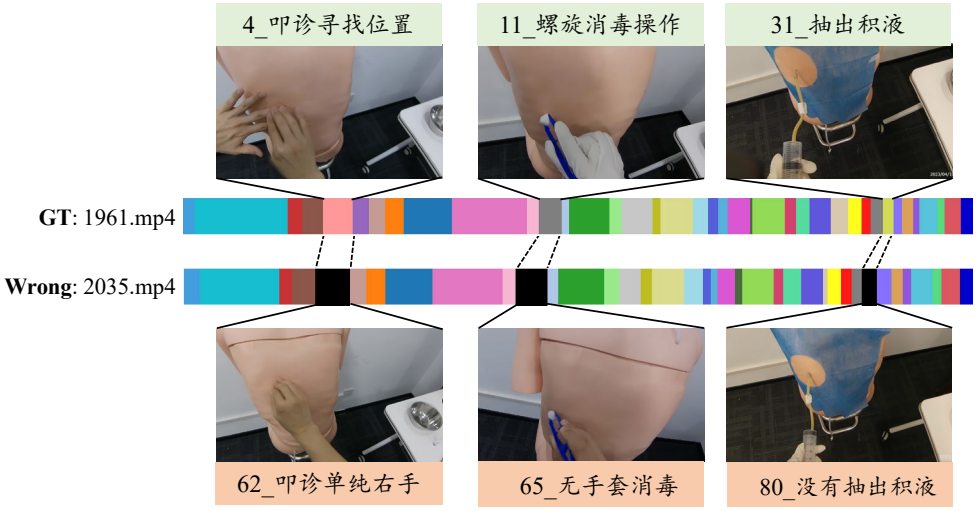


图 5-15　错误序列展示图

行为合规性检测算法

DTW 算法实现了 $\hat{L}_0$ 和 L_0 两个序列在时间维度上的匹配，之后需要从首帧开始进行标签遍历，寻找两序列之间存在差异的序列帧，并最终将这些差异映射为分类信息。算法 2 展示了本文所设计的行为合规性检测算法。各元素含义如下：

M：丢失行为映射表，通过哈希表的形式对表 5-7 进行存储，可以通过正确行为标签获取到对应的遗漏行为标签。例如 $43 = M[4]$ 表示通过“4_叩诊寻找位置”可以获取到对应遗漏行为“43_遗漏叩诊”。

$(\hat{I}, I)$：动态时间规整算法计算得到的两个下标集合，具有相同的元素数量，

通过$\hat{L}_0\left[\hat{I}_i\right]$操作和$L_0[I_i]$操作即可获取完整匹配序列；

S_{Dig}：表示$\hat{L}_0\left[\hat{I}_i\right]$序列与序列$L_0[I_i]$之间所有差异元素的位置集合；

S_{Correct}：表示所有正确行为的标签集合，在本文中$S_{\text{Correct}}=\{0,1,\cdots,38\}$；

为更清晰地展示算法 2 的流程，本文在图 5-16 中结合简单案例进行了两个序列之间的匹配结果展示。在此案例中，-1和-2分别表示行为序列的开始与结束；$\{1,2,3,4,5,6,7,8,9,10\}$为标准操作序列；$\{11,12\}$为错误行为标签，且满足：8 号行为在犯错误时标签为 11，5 号行为在遗漏时标签为 12，即$12=M[5]$。原始序列根据 DTW 算法获得下标序列$\left(\hat{I},I\right)$，最终通过序列索引完成时序标签信息对齐，最终通过逐元素判断获取结果信息。

算法 2：时序行为合规性检测算法

Input：预测序列$\hat{L}_0$，标准序列L_0，丢失行为映射表M

Output：错误行为集合：$[S_{\text{Loss}},S_{\text{Err}}]$

```
// DTW 算法匹配，获取两个下标集合
(Î, I) ← DTW(L̂0, L0), Len(Î) = Len(I)
// 序列遍历&差异检测
for i in Len(I):
    if L̂0[Î_i] ≠ L0[I_i]:
        S_Dig.add(i)  // 获取差异位置集合
// 遗漏行为&错误行为检测
for i in S_Dig:
    if L̂0[Î_i] ∈ S_Correct :
        L_Loss ← M[L0[I_i]]
        S_Loss.add(L_Loss)  // 遗漏行为
    else:
        S_Err.add(L̂0[Î_i])  // 错误行为
return [S_Loss, S_Err]
```

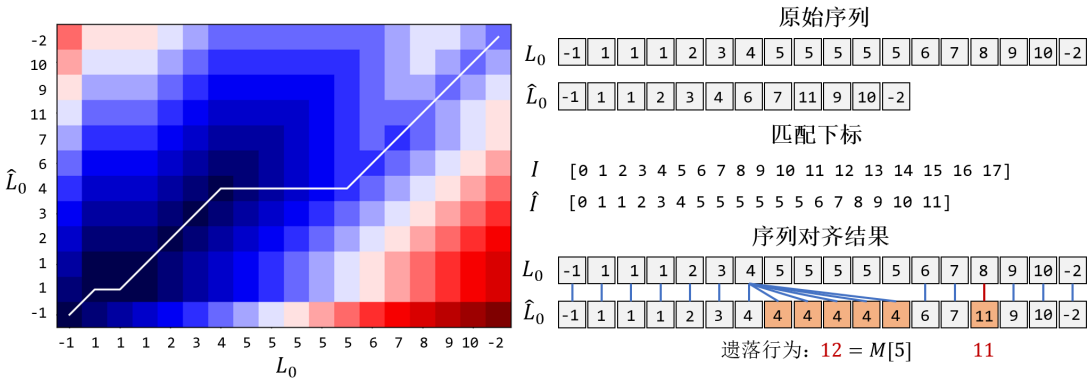


图 5-16　行为合规性检测算法示意图

5.5 实验分析

本节内容组织方式如下：5.5.1 小节对 ThoSet 数据集的构建过程进行了详细阐述；5.5.2 小节对时序行为分割数据集和评估指标进行了介绍；5.5.3 小节对所有模型的设定和训练过程参数配置进行介绍；5.5.4、5.5.5 和 5.5.6 小节分别对模型性能、消融实验和可视化实验结果进行了总结与分析。

5.5.1 ThoSet 数据集构建

现有的手术流程识别数据集例如 MICCAI 系列挑战赛[69]、Cholec80[20]、Hei-Chole[128]、Cataract-101[21]、CATARACTS[73]只对各类手术的整体阶段进行了大致划分，医疗行为种类的数量通常在 10~20 种左右。与 Breakfast[68]、50Salads[11]和 GTEA[12]等日常行为时序分割基准类似，这些数据集都只能支持时序视频行为分割任务。现有医疗时序行为分析数据集依然存在两方面的问题：医疗行为划分的细粒度有待提升、无法支持更深层次的医疗行为合规性评估等任务。

为解决以上问题，本文以胸腔穿刺术为研究对象，构建了具有更高行为细粒度且支持医疗行为合规性评估的 ThoSet 数据集（Thoracocentesis Dataset）。在 5.3.4 小节中，本文对医疗行为流程知识图谱进行了细化拓展，将医疗行为从较低细粒度的“二级操作步骤”进一步细化到粒度更高的子行为层次。拆分后的胸腔穿刺流程如表 5-7 所示，本文在后续的 ThoSet 数据集构建过程中参考了此表的划分方式。本小节从数据采集系统搭建、原始视频采集过程、案例拼接策略和模型性能评估指标共四个方面对 ThoSet 数据集的构建过程进行详细介绍。

数据采集系统搭建

胸腔穿刺数据集采集平台　　第一人称视角摄像头　　摄像头佩戴方式

图 5-17　胸腔穿刺时序行为分析数据集 ThoSet 采集平台

如图 5-17 所示，本文搭建了一套第一人称视角胸腔穿刺行为采集平台。胸腔穿刺操作台上的所有医疗器械摆放与真实考核场景保持一致。本文采用第一人

称视角摄像头完成操作过程中的视频录制，摄像头的具体型号为欧达 ORDRO-EP8-4K。此摄像头能够支持 4K 分辨率录制，但考虑到超高清视频会引起存储空间占用过大问题，最终所有视频均采用1920×1080 分辨率和 30 FPS 进行录制。

视频采集与标注过程

在传统的时序行为分割数据集构建中，只需要对收集到的视频进行时序标签标注即可。与传统的时序行为分割数据集不同，本文所探究的医疗时序行为分析数据集需要同时囊括正确行为案例（即表 5-7 中的 39 类正确子行为）和错误行为案例（即表 5-7 中的 22 类遗漏错误行为和 23 类单流程错误行为），从而确保训练过程中能够为模型提供分辨错误行为的信息。为实现表 5-7 中的正确行为与错误行为的同时采集，本文采用 “正确行为连贯采集+错误行为单独采集”的组合式视频采集策略。图 5-18 从单个被试者的视角展示了此采集策略。阶段一中的数据共分为两个部分：首先需要对标准的胸腔穿刺操作（行为编号 0~39 共 39 种子行为）进行 3 次采集，其次再对每个错误行为（行为编号 61~83 共 23 种单流程错误行为）进行独立的 3 次采集。阶段一结束后，即可获得以 MP4 格式存储的 3 组正确操作视频与错误演示视频。在图 5-18 的阶段一示意图中，正确操作视频使用绿色标记，错误演示视频使用黄色标记，每个行为的 ID 与视频相互绑定。图 5-21 对完整的胸腔穿刺流程中 39 个子行为案例进行了展示。

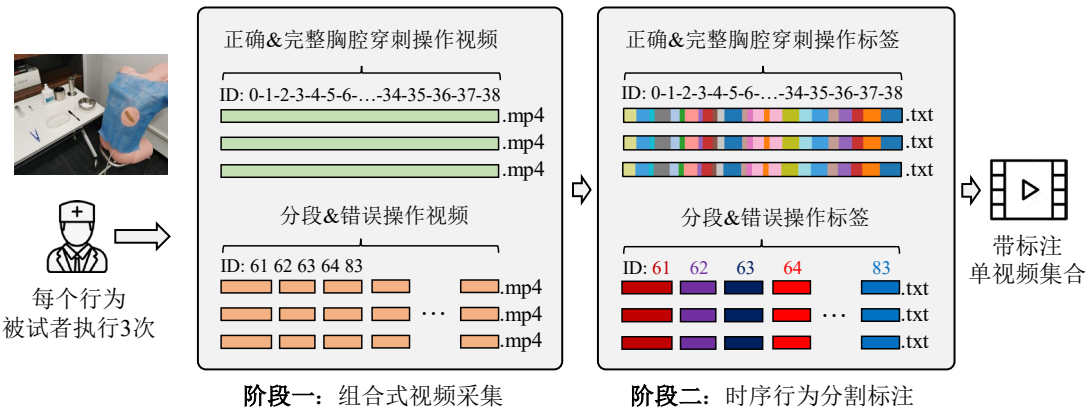


图 5-18 ThoSet 数据集的视频采集与标注过程

在阶段二中，首先对所有的视频进行时序行为分割标注，生成 TXT 格式的标注文件。图 5-18 的阶段二示意图根据不同行为种类对视频标签进行了上色区分。由于标注信息含有逐帧的行为类别信息，因此可以根据标签信息对完整的胸腔穿刺长视频进行剪切，从而获得 39 种正确子行为的剪切视频。本文共招募了 10 名志愿者参与到 ThoSet 数据集的构建中，其中每个志愿者均按照图 5-18 中的流程完成行为采集。考虑到第一视角相机视野的局限性和不稳定性，本文在数据采集过程中使用了“一左一右”的双摄像头配置，实现了数据体量的扩充。在完

成阶段二的标注与长视频分割后，本文共获取到 3,720 个剪辑后的单类别视频。视频数量的具体计算方法为：

$$10\text{人}\times 3\text{次操作}\times 2\text{视角}\times(39\text{子行为}+23\text{错误行为})=3,720 \quad (5.38)$$

经过裁切后的视频集合将以素材库的形式参与到后文中阶段三的序列构建过程。通过汇总与统计，裁切后的单类视频集合含有的总帧数为 667,507，占用存储空间为 46.2 GB，总时长达 6.18h。

基于拼接策略的样例构建方法

图 5-18 中的阶段一和阶段二完成了正确行为和错误行为的视频素材库构建与标注。在阶段三中，本文采用随机拼接策略充分利用以上视频素材库，构建了丰富的胸腔穿刺操作错误行为案例，从而实现 ThoSet 医疗时序行为分析数据集的构建。

图5-19对本文所提出的案例拼接策略与错误行为序列构建流程进行了展示。在“流程错误序列构建”中，使用单流程错误视频对完整操作视频中的对应子行为进行替换，即可获得错误操作序列；在“遗漏错误序列构建”中，只需在完整操作视频中删除对应子行为，即可获得遗漏操作序列。需要注意，由于表 5-8 中的错误行为因果关系存在，需要在阶段三的序列生成过程中引入标签关联机制：在错误案例生成过程中，一旦在早期检测到特定的行为标签，就需要在后续序列中引入关联错误。例如一旦发生“48_遗漏手套”行为，序列构建过程中要自动添加“64_无手套镊子夹取”、“65_无手套消毒”和“69_无手套洞巾”三类行为。

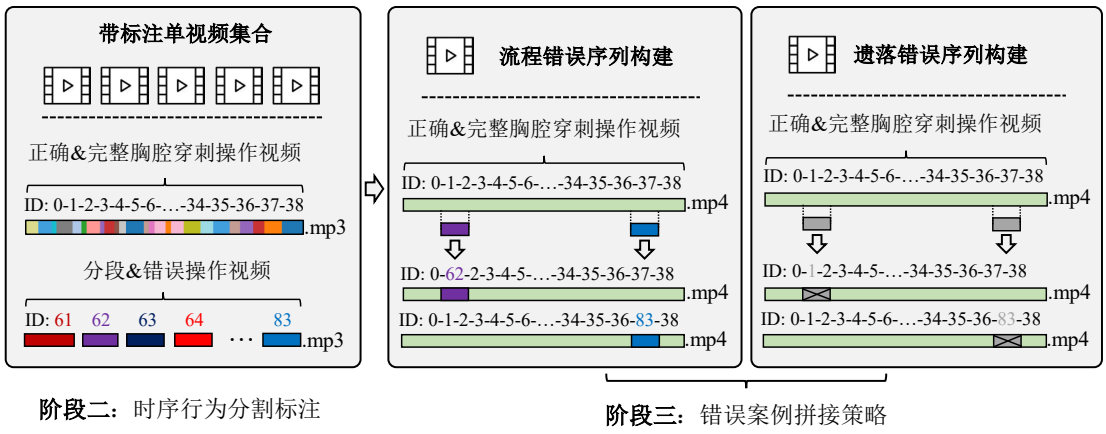


图 5-19　案例拼接策略与错误行为序列构建方法

在实际的 ThoSet 数据集生成与构建阶段中，本文将 7 名被试者的行为划分至训练集，3 名被试者的行为划分至测试集，如图 5-20 所示。本文按照错误出现的频率分别为训练集和测试集构建了四类操作序列：正确序列、较少错误（5%）、中等数量错误（10%）、较多数量错误（20%）。胸腔穿刺术共含有 39 个子行为，

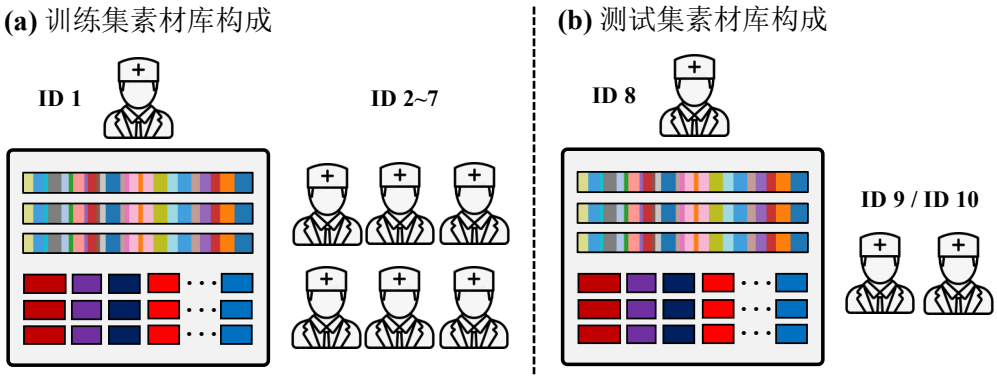


图 5-20 ThoSet 数据集中训练集与测试集素材库划分

在 5%错误率的设定中平均每个序列含有约 2 个错误；在 10%错误率的设定中平均每个序列含有约 4 个错误；在 20%错误率的设定中平均每个序列含有约 8 个错误。

在采样数量方面，本文依据每位被试者采集的三组行为进行序列生成，对正确序列进行 20 次随机采样，对较少错误、中等数量错误和较多数量错误序列分别进行 40 次随机采样。序列中的错误行为数量和种类通过随机数和阈值进行控制，序列生成过程满足表 5-8 中的因果关系约束。单个视角下生成的序列总数量为 1,400 条，双视角下生成序列总数为 2,800 条。表 5-9 对训练集、测试集和平均错误数量信息进行了汇总。数据集视频的总帧数为 647,376，总持续时长近 6h。在数据集的存储环节，本文并未使用视频拼接工具对 2,800 条操作序列进行视频实体构建，而是通过“素材库+视频 ID 索引表”的形式进行存储。换言之，ThoSet 数据集的存储方式不再是传统的物理层次存储，而是采用逻辑链接的方式进行索引存储。这种技巧能够极大地节省数据集的存储空间。

表 5-9 ThoSet 数据集训练集与测试集统计信息

统计信息	训练集	Avg. Len.	Avg. Err.	测试集	Avg. Len.	Avg. Err.
人数	7	—	—	3	—	—
#正确行为序列	280	248.04s	0	120	243.09s	0
#较少错误序列	560	242.69s	2.08	240	233.16s	2.06
#中等数量错误序列	560	232.02s	4.50	240	225.60s	4.07
#较多数量错误序列	560	215.53s	8.39	240	211.42s	8.24
序列总数	1960	233.00s	—	840	227.02s	—

模型性能评估指标

本文所构建的 ThoSet 数据集能够支持时序行为分割任务和错误行为检测任务。在时序行为分割任务中，本文采用与开源基准（Breakfast，50Salads 和 GTEA）相同的评估指标，下一小节对这些指标进行了介绍；在行为合规性评估任务中，本文使用与第三章相同的评估指标：mAP 与 mmit mAP。

图 5-21　胸腔穿刺完整流程展示

5.5.2 时序行为分割数据集与评估指标

Breakfast[68]数据集共含有 1,712 条视频，包含 48 种制作早餐的行为，平均每个视频含有 6 种行为；50Salads[11]数据集共含有 50 条视频，包含 17 种制作沙拉的行为，平均每个视频含有 20 种行为，平均时长为 6.4 分钟；GTEA[12]数据集共含有 28 条视频，包含 11 种日常活动行为，平均每个视频含有 20 种行为，平均时长为 30 秒。其中 Breakfast 数据集具有最大的数据集体量，所有视频均由第三人称视角拍摄。50Salads 和 GTEA 中的视频由第一人称视角拍摄。

与其他时序行为分割模型测试方式保持一致，本章所有实验均使用 I3D 模型提取到的视频特征。所有指标均是取多折交叉验证结果的均值：在 Breakfast 上进行 5 折交叉验证，而在 50Salads 和 GTEA 数据集上进行 4 折交叉验证。评估指标方面，使用帧级别的准确度（Accuracy, Acc.）、编辑得分（Edit Score）、F1 分数在 10%，25%，50%的不同重叠阈值下的设定（F1@{10, 25, 30}）来进行模型性能评估。其中 Acc 表示帧级别的分类精度，Edit 和 F1 分数在行为片段的层次（Segment-Level）进行模型的性能评估。

5.5.3 模型与优化器设定

本章所有实验在 AMD EPYC 7742@2.25GHz CPU 和 NVIDIA A800 GPU 计算平台上进行，所有模型均使用 Pytorch 框架搭建与训练。本文所提出的模型在 Breakfast[68]、50Salads[11]、GTEA[12]和 ThoSet 数据集中的训练配置参数如表 5-10 所示。考虑到 Breakfast 和 ThoSet 数据集具有较大的数据集体量，在编码器中使用 12 层堆叠的配置，在自注意力层中使用 768 的特征维度配置。所有模型均使用 Adam 优化器进行训练。

表 5-10　模型在公开数据集与 ThoSet 中的超参数配置表

数据集	#编码器	#解码器	#特征维度	学习率	训练轮次
Breakfast[68]	12	8	768	1e-4	1k
50Salads[11]	10	8	192	5e-4	5k
GTEA[12]	10	8	192	5e-4	10k
ThoSet	12	8	768	1e-4	100

5.5.4 性能对比实验结果

为充分验证本文所提出的基于时序聚类注意力机制的扩散时序行为分析算法有效性，本节在两类数据集上开展了性能对比实验，分别为：时序行为分割任务基准数据集（Breakfast[68]、50Salads[11]和 GTEA[12]）和自建胸腔穿刺时序行为分析数据集 ThoSet。

时序行为分割公开数据集实验结果

表 5-11 将 *k*M-Att 模型与现有的先进时序行为分割模型进行了性能对比。由于其他模型在报告性能时均只保留了一位小数，为确保可比较性本章在对比实验中保持一致，但在后续的模型性能探究中为确保实验结果准确性，实验结果均保留两位小数。结果显示在数据量充足的 Breakfast 和 50Salads 数据集中，本章提出的模型在帧级别和段级别的分割准确率均显著优于其他算法。在 GTEA 数据集中，本章提出的模型在 F1@{50}和 Avg 性能指标中取得了最优、在 F1@{25}和 Acc 性能指标中取得了次优。三个数据集上的性能差异是由不同的数据集体量引起的。本章提出的 *k*M-Att 模块本质上是插入在编码器与解码器之间的特征增强单元，虽然此模块的引入能够提升模型的特征提取和表示能力，但是会增加模型复杂度，从而导致小型数据集上出现的过拟合问题。三个基准数据集中 Breakfast 和 50Salads 均具备一定的体量，而 GTEA 数据集仅含有 28 条视频。

表 5-11 *k*M-Att 方法与现有算法性能对比

模型	Breakfast				50Salads				GTEA			
	F1@{10,25,50}	Edit	Acc	Avg	F1@{10,25,50}	Edit	Acc	Avg	F1@{10,25,50}	Edit	Acc	Avg
MS-TCN++[82]	64.1/58.6/45.9	65.6	67.6	60.4	80.7/78.5/70.1	74.3	83.7	77.5	88.8/85.7/76.0	83.5	80.1	82.8
SSTDA[194]	75.0/69.1/55.2	73.7	70.2	68.6	83.0/81.5/73.8	75.8	83.2	79.5	90.0/89.1/78.0	86.2	79.8	84.6
GTRM[85]	57.5/54.0/43.3	58.7	65.0	55.7	75.4/72.8/63.9	67.5	82.6	72.4	-/-/-	-	-	-
BCN[195]	68.7/65.5/55.0	66.2	70.4	65.2	82.3/81.3/74.0	74.3	84.4	79.3	88.5/87.1/77.3	84.4	79.8	83.4
MTDA[196]	74.2/68.6/56.5	73.6	71.0	68.8	82.0/80.1/72.5	75.2	83.2	78.6	90.5/88.4/76.2	85.8	80.0	84.2
C2F-TCN[197]	72.2/68.7/57.6	69.6	76.0	68.8	84.3/81.8/72.6	76.4	84.9	80.0	90.3/88.8/77.7	86.4	80.8	84.8
G2L[84]	74.9/69.0/55.2	73.3	70.7	68.6	80.3/78.0/69.8	73.4	82.2	76.7	89.9/87.3/75.8	84.6	78.5	83.2
HASR[198]	74.7/69.5/57.0	71.9	69.4	68.5	86.6/85.7/78.5	81.0	83.9	83.1	90.9/88.6/76.4	87.5	78.7	84.4
ASRF[199]	74.3/68.9/56.1	72.4	67.6	67.9	84.9/83.5/77.3	79.3	84.5	81.9	89.4/87.8/79.8	83.7	77.3	83.6
ASFormer[87]	76.0/70.6/57.4	75.0	73.5	70.5	85.1/83.4/76.0	79.6	85.6	81.9	90.1/88.8/79.2	84.6	79.7	84.5
UARL[98]	65.2/59.4/47.4	66.2	67.8	61.2	85.3/83.5/77.8	78.2	84.1	81.8	92.7/91.5/82.8	88.1	79.6	86.9
DPRN[200]	75.6/70.5/57.6	75.1	71.7	70.1	87.8/86.3/79.4	82.0	87.2	84.5	92.9/92.0/82.9	90.9	**82.0**	88.1
TUT[88]	76.2/71.9/60.0	73.7	76.0	71.6	89.3/88.3/81.7	84.0	87.2	86.1	89.0/86.4/73.3	84.1	76.1	81.8
CETNet[169]	79.3/74.3/61.9	77.8	74.9	73.6	87.6/86.5/80.1	81.7	86.9	84.6	91.8/91.2/81.3	87.9	80.3	86.5
SEDT[201]	-/-/-	-	-	-	89.9/88.7/81.1	84.7	86.5	86.2	93.7/**92.4**/84.0	91.3	81.3	**88.5**
TCTr[167]	76.6/71.1/58.5	76.1	77.5	72.0	87.5/86.1/80.2	83.4	86.6	84.8	91.3/90.1/80.0	87.9	81.1	86.1
DTL[202]	78.8/74.5/62.9	77.7	75.8	73.9	87.1/85.7/78.5	80.5	86.9	83.7	-/-/-	-	-	-
UVAST[190]	76.9/71.5/58.0	77.1	69.7	70.6	89.1/87.6/81.7	83.9	87.4	85.9	92.7/91.3/81.0	**92.1**	80.2	87.5
BrPrompt[90]	-/-/-	-	-	-	89.2/87.8/81.3	83.8	88.1	86.0	**94.1**/92.0/83.0	91.6	81.2	88.4
TST[203]	77.5/72.3/59.5	76.7	73.7	71.9	87.9/86.6/80.5	82.7	86.6	84.9	91.4/90.2/82.1	86.6	80.3	86.1
FAMMSDTN[89]	78.5/72.9/60.2	77.5	74.8	72.8	86.2/84.4/77.9	79.9	86.4	83.0	91.6/90.9/80.9	88.3	80.7	86.5
DiffAct[92]	80.3/75.9/64.6	78.4	76.4	75.1	90.1/89.2/83.7	85.0	88.9	87.4	92.5/91.5/84.7	89.6	80.3	87.7
Ours	**81.1/76.8/65.4**	**79.3**	**76.9**	**75.9**	**91.8/90.7/85.7**	**87.6**	**89.9**	**89.1**	93.3/92.3/**85.7**	89.9	81.5	**88.5**

本章提出的时序分割模型是在 DiffAct 算法上的进一步优化，因此本文在表 5-12 中对两个模型的性能进行了详细对比。结果显示 *k*M-Att 模块均能够对三个数据集上的所有指标产生性能增益：在 Breakfast、50Salads 和 GTEA 数据集上分别产生了 0.8、1.7 和 0.8 百分点的 Avg 指标提升；特别是在具有最长平均时长的 50Salads 中，模型的 F1{50}提高了 2.0 百分点，Edit 指标提高了 2.6 百分点。对比结果充分说明了 *k*M-Att 模块的有效性。在获取到编码器特征之后，时序聚类注意力模块能够对特征进行区域划分，通过局部注意力机制与全局 *k*-means 聚类注意力机制分别完成物理临近交互与逻辑分区交互建模。

表 5-12 *k*M-Att 方法与 DiffAct 模型的性能对比

数据集	模型	F1@{10,25,50}	Edit	Acc	Avg
Breakfast	DiffAct[92]	80.3 / 75.9 / 64.6	78.4	76.4	75.1
	Ours	**81.1** / **76.8** / **65.4**	**79.3**	**76.9**	**75.9**
	Δ	+0.8 / +0.9 / +0.9	+0.9	+0.5	+0.8
50Salads	DiffAct[92]	90.1 / 89.2 / 83.7	85.0	88.9	87.4
	Ours	**91.8** / **90.7** / **85.7**	**87.6**	**89.9**	**89.1**
	Δ	+1.7 / +0.9 / +2.0	+2.6	+1.0	+1.7
GTEA	DiffAct[92]	92.5 / 91.5 / 84.7	89.6	80.3	87.7
	Ours	**93.3** / **92.3** / **85.7**	**89.9**	**81.5**	**88.5**
	Δ	+0.8 / +0.8 / +1.0	+0.3	+1.2	+0.8

为验证本章所提出的非锁步跳跃去噪机制的有效性，本文在表 5-13 中列举了不同去噪机制的性能与延时比较结果。实验在 Breakfast 和 50Salads 中开展，共设定了三种解码模式：锁步跳跃 25 步、锁步跳跃 15 步和非锁步跳跃 15 步。以锁步跳跃 25 步为基准线，表 5-13 对所有性能的变化进行了标注，其中绿色表示性能提升，红色表示性能下降。延时指标通过对数据集中所有交叉验证的耗时取平均计算得到，延时越低表示模型的效率越高。结果显示，相较于传统解码机制，本章提出的非锁步跳跃去噪机制能够以更少的迭代轮次取得更高的分割性能。在 Breakfast 中，将锁步跳跃解码机制的 25 步减少到 15 步后，虽然分割模型的延时降低了 29%，但是模型的整体性能面临全面下降；50Salads 数据集上的实验结果和 Breakfast 基本保持一致。非锁步跳跃解码机制能够在 15 步的设定下在两个数据集上均取得更高的性能，这充分说明了“前期大步长、后期小步长”策略的有效性。由于跳跃解码的每一步均需要重新运行解码器，因此从理论上可以节省$(25-15)\div 25=0.4$的推理时间。延时测试结果与理论值一致，非锁步跳跃解码机制分别在 Breakfast 和 50Salads 数据集上减少了 28.7%和 34.0%的延时。

表 5-13 公开数据集的非锁步跳跃去噪机制对比实验结果

数据集	解码模式	延时↓	F1@{10,25,50}	Edit	Acc	Avg
Breakfast	Locked=25	0.341s	81.06 / 76.80 / 65.36	79.28	76.91	75.88
	Locked=15	0.242s	80.91 / 76.55 / 65.18	78.99	76.91	75.71
	Δ	-29.0%	-0.15 / -0.25 / -0.19	-0.29	0	-0.17
	Unlocked=15	0.243s	81.37 / 77.07 / 65.61	79.32	77.75	76.22
	Δ	-28.7%	+0.31 / +0.27 / +0.25	+0.04	+0.84	+0.34
50Salads	Locked=25	2.12s	91.79 / 90.67 / 85.71	87.57	89.86	89.12
	Locked=15	1.41s	91.76 / 90.15 / 85.73	87.58	89.89	89.02
	Δ	-33.5%	-0.03 / -0.52 / +0.02	+0.01	+0.03	-0.10
	Unlocked=15	1.40s	91.81 / 90.69 / 85.64	87.63	89.87	89.13
	Δ	-34.0%	+0.02 / +0.02 / -0.08	+0.07	+0.01	+0.01

为充分探究 DiffAct 与 *k*M-Att 模型在各公开数据集上的性能，表 5-14 详细列举了模型在每个 Split 下的实验结果，并在最右侧列举了 *k*M-Att 模型对 DiffAct 的相对提升。结果显示在绝大部分情况下本文所提出的 *k*M-Att 模型的时序分割

表 5-14　DiffAct 与 *k*M-Att 在公开基准中的详细实验结果

数据集		DiffAct[92]				*k*M-Att				Δ
		F1@{10,25,50}	Edit	Acc	Avg	F1@{10,25,50}	Edit	Acc	Avg	Avg
Breakfast	S1	80.24/76.98/66.04	77.93	77.49	75.74	80.59/77.07/66.24	79.23	77.54	76.13	+0.39
	S2	80.06/76.02/64.97	78.77	74.72	74.91	82.24/77.70/66.46	80.99	76.10	76.70	+1.79
	S3	82.61/78.31/67.61	81.09	79.64	77.85	82.74/78.88/67.84	81.18	79.79	78.09	+0.24
	S4	77.91/72.88/60.26	75.00	73.86	71.98	78.66/73.55/60.90	75.71	74.20	72.60	+0.62
	All	80.21/76.05/64.72	78.20	76.43	75.12	81.06/76.80/65.36	79.28	76.91	75.88	+0.76
50Salads	S1	89.20/88.24/80.57	84.17	86.18	85.67	89.71/88.24/78.92	83.84	87.43	85.63	-0.04
	S2	91.38/90.86/86.68	86.88	90.60	89.28	93.23/92.71/88.02	89.73	91.88	91.11	+1.83
	S3	91.54/90.54/86.56	86.94	87.89	88.69	92.19/91.69/88.66	87.92	89.12	89.92	+1.23
	S4	89.50/89.00/83.00	84.64	88.08	86.84	90.05/88.56/84.08	87.13	88.30	87.62	+0.78
	S5	90.62/90.10/84.38	83.65	89.39	87.63	93.77/92.14/88.89	89.21	92.57	91.32	+3.69
	All	90.45/89.75/84.24	85.26	88.43	87.62	91.79/90.67/85.71	87.57	89.86	89.12	+1.50
GTEA	S1	92.54/91.04/85.82	89.35	81.54	88.06	92.78/89.73/85.93	90.25	83.22	88.38	+0.32
	S2	90.64/89.89/78.65	87.32	79.03	85.11	89.14/89.14/80.15	83.19	80.10	84.34	-0.77
	S3	93.85/93.08/90.00	91.11	81.10	89.83	94.62/94.62/91.54	92.60	81.50	90.98	+1.15
	S4	95.42/94.66/85.50	91.48	79.54	89.32	96.58/95.82/85.17	93.43	81.08	90.42	+1.10
	All	93.11/92.17/84.99	89.82	80.30	88.08	93.28/92.33/85.70	89.87	81.48	88.53	+0.45

性能均超越了 DiffAct 模型，这充分说明了时序特征增强机制的有效性。

ThoSet 数据集实验结果

表 5-15 列举了 ASFormer[87]、DiffAct[92]以及本文方法在 ThoSet 数据集中的时序分割预测结果。结果显示，*k*M-Att 模型在平均精度 Avg 指标中分别超越 ASFormer 和 DiffAct 模型 1.88 和 0.42 百分点，这充分说明了 *k*M-Att 模型的有效性。整体而言，DiffAct 模型的分割性能优于 ASFormer 模型，只有在帧级别精度 Acc 中，ASFormer 优于 DiffAct。这充分说明了扩散去噪策略的有效性。

表 5-15　ThoSet 数据集上的模型性能对比实验结果

模型	F1@{10,25,50}	Edit	Acc	Avg
ASFormer[87]	93.20 / 92.78 / 91.35	90.94	92.54	92.16
Ours	95.11 / 94.75 / 93.56	94.13	92.67	94.04
Δ	+1.91 / +1.97 / +2.21	+3.19	+0.13	+1.88
DiffAct[92]	94.79 / 94.51 / 93.29	93.83	91.66	93.62
Ours	95.11 / 94.75 / 93.56	94.13	92.67	94.04
Δ	+0.32 / +0.24 / +0.27	+0.30	+1.01	+0.42

表 5-16　ThoSet 数据集的非锁步跳跃去噪机制对比实验结果

数据集	解码模式	延时↓	F1@{10,25,50}	Edit	Acc	Avg
ThoSet	Locked=25	0.516s	95.11 / 94.75 / 93.56	94.13	92.67	94.04
	Locked=15	0.357s	95.00 / 94.62 / 93.37	93.96	92.68	93.93
	Δ	-30.8%	-0.11 / -0.13 / -0.19	-0.17	+0.01	-0.11
	Unlocked=15	0.350s	95.12 / 94.74 / 93.57	94.18	92.61	94.04
	Δ	-32.2%	+0.01 / -0.01 / +0.01	+0.05	-0.06	0

表 5-16 展示了 ThoSet 数据集上不同解码模式设定下的时序分割性能和延时结果。与表 5-13 表示方法一致，锁步跳跃 25 步被设定为基准线，绿色表示性能

提升，红色表示性能下降。实验结果显示，相较于传统的锁步跳跃解码机制，本章所提出的非锁步跳跃去噪机制能以更少的迭代次数实现同等的时序分割性能，推理时间减少近 30%，算法延时能够满足后续实际应用的需求。

行为合规性检测结果

为探究本文提出的行为合规性算法有效性，本文构建了基于 Transformer[104] 架构的模型作为基线方法。此基线模型的输入为时序分割标签结果，输出结果为预测错误类型。基线模型采用传统的编码器—解码器结构设计，其中编码器和解码器均由三个自注意力层构成。两个算法对于错误行为和遗漏行为的识别性能汇总在表 5-17 中对比结果证实了本文行为合规性算法的有效性。

表 5-17　ThoSet 数据集上的行为合规性检测结果

模型	错误行为识别		遗漏行为识别	
	mAP	mmit mAP	mAP	mmit mAP
Transformer[104]	88.52	92.95	90.47	94.09
Ours	89.23	93.64	92.61	95.86

5.5.5 消融实验结果

为充分验证时序 *k*-means 聚类注意力机制和局部注意力机制的有效性，表 5-18 列举了三个公开数据集上的消融实验结果。结果显示两种机制的配合能够在三个数据集上获得整体最优的性能，任何一个分支的缺失都会导致平均性能的降低。这证实了两种机制的互补性：*k*-means 聚类注意力机制对全局信息进行建模，在时序上实现特征的逻辑分区交互；局部注意力机制对相近的位置进行建模，实现物理临近交互。在 Breakfast 和 50Salads 数据集中，去除掉局部注意力分支的性能要优于去除 *k*-means 聚类注意力分支，这说明在大体量的数据集中全局逻辑分区交互具有更重要的作用；然而 GTEA 数据集上的实验结果则相反，局部信息交互在较小体量的数据集上能达到更高的性能。以 DiffAct 在三个数据集上的 Avg 为基准：75.1，87.4，87.7，添加两个机制中的任何一个均可以带来性能的提升，这说明了在编码器和解码器之间添加特征增强模块策略的可行性。

表 5-18　公开数据集中 *k*-means 聚类注意力与局部注意力机制消融实验

数据集	模型	F1@{10,25,50}	Edit	Acc	Avg
Breakfast	*w/o k*-means Att	80.23 / 75.87 / 64.11	78.62	76.79	75.12
	w/o Local Att	80.83 / 76.38 / 65.16	79.02	76.26	75.53
	All	**81.06** / **76.80** / **65.36**	**79.28**	**76.91**	**75.88**
50Salads	*w/o k*-means Att	91.25 / 90.52 / 85.66	86.12	89.63	88.64
	w/o Local Att	91.61 / 90.61 / **85.97**	87.01	89.41	88.92
	All	**91.79** / **90.67** / 85.71	**87.57**	**89.86**	**89.12**
GTEA	*w/o k*-means Att	93.19 / 92.06 / 85.29	90.48	81.56	88.52
	w/o Local Att	93.17 / 91.31 / 84.18	**90.89**	**81.89**	88.29
	All	**93.28** / **92.33** / **85.70**	89.87	81.48	**88.53**

表 5-19　ThoSet 中 *k*-means 聚类注意力与局部注意力机制消融实验

数据集	模型	F1@{10,25,50}	Edit	Acc	Avg
ThoSet	*w/o k*-means Att	94.92 / 94.49 / 92.91	93.70	91.88	93.58
	w/o Local Att	94.01 / 93.64 / 92.37	92.92	91.75	92.94
	All	**95.00** / **94.57** / **93.42**	**93.89**	**92.82**	**93.94**

表 5-19 列举了 ThoSet 数据集上的注意力机制消融实验结果。与公开数据集中的结果一致，两种注意力机制相互配合才能够取得最优的性能结果，这充分说明了 *k*M-Att 模块的有效性。

表 5-20 和表 5-21 对时序 *k*-means 注意力机制的聚类数量的影响进行了探究。结果显示，不同的中心数量会造成性能的波动，但是这种波动并不剧烈。在所有数据集中，聚类中心数量$k=64$可以取得最优性能。在 Breakfast 和 50Salads 数据集中，$k=128$的设定要优于$k=32$，而在 GTEA 数据集中则相反。这主要是由于 Breakfast 和 50Salads 有更大的体量和更复杂的时序上下文信息，聚类中心数的提升不会对性能产生较大的负面影响。更少的聚类中心数量更加契合 GTEA 数据集。实验结果显示超参数k的设定需要结合实际的数据集体量进行讨论。本章所有的实验均使用$k=64$作为默认设定。

表 5-20　公开数据集中 *k*M-Att 模块的聚类数量影响

数据集	#*k*	F1@{10,25,50}	Edit	Acc	Avg
Breakfast	128	**81.09** / 76.54 / 65.60	79.13	76.84	75.84
	64	81.06 / **76.80** / **65.63**	79.28	**76.91**	**75.88**
	32	80.94 / 76.49 / 65.22	**79.37**	76.71	76.75
	16	80.86 / 76.52 / 65.38	79.19	76.75	75.74
50Salads	128	91.53 / **91.12** / 85.68	86.08	89.25	88.73
	64	**91.79** / 90.67 / **85.71**	**87.57**	**89.86**	**89.12**
	32	90.89 / 90.19 / 84.97	85.86	89.58	88.30
	16	91.26 / 90.16 / 84.58	86.38	89.36	88.35
GTEA	128	91.36 / 90.42 / 84.41	88.11	81.27	87.11
	64	**93.28** / **92.33** / **85.70**	89.87	**81.48**	**88.53**
	32	92.59 / 91.85 / 83.95	89.41	81.24	87.81
	16	93.05 / 92.11 / 84.79	**90.30**	80.51	88.15

表 5-21　ThoSet 中 *k*M-Att 模块的聚类数量影响

数据集	#*k*	F1@{10,25,50}	Edit	Acc	Avg
ThoSet	128	94.74 / 94.36 / 93.23	93.56	92.58	93.70
	64	**95.11** / **94.75** / **93.56**	**94.13**	**92.67**	**94.04**
	32	94.94 / 94.74 / 93.48	94.07	91.84	93.81
	16	94.93 / 94.60 / 93.26	93.66	92.51	93.79

5.5.6 定性对比与可视化结果

图 5-22 对非锁步跳跃扩散去噪的过程进行了可视化，并且绘制了 DiffAct 的预测结果与标准标签。视频案例为 50Salads 数据集的 rgb-27-2.avi。黑色和红色

的方框对预测标签中的不同部分进行了高亮。结果显示，相较于 DiffAct 模型，本章所提出的框架能够以更少的迭代次数生成更加优良的结果。例如，DiffAct 在 action_end 标签中为大量无操作帧指派了各种不同的行为，而得益于局部物理临近交互和逻辑分区交互，本模型能够在去噪的过程的早期对不合理的预测结果进行及时调整，最终生成稳定可靠的行为分割结果。在视频前段，DiffAct 在 cut_tomato 和 peel_cucumber 标签之间分配了短暂的 place_tomato_into_bowl 行为。通过对解码过程的可视化可以观察到，基于 *k*M-Att 的模型在解码早期也预测出了此错误行为，但是解码过程末尾的密集迭代过程实现了精修和矫正，最终生成了和标准序列一致的分割结果。相较于 DiffAct，本章所提出的模型能够在节省 40%迭代次数的前提下获得更优的行为顺序和行为切换点预测结果。

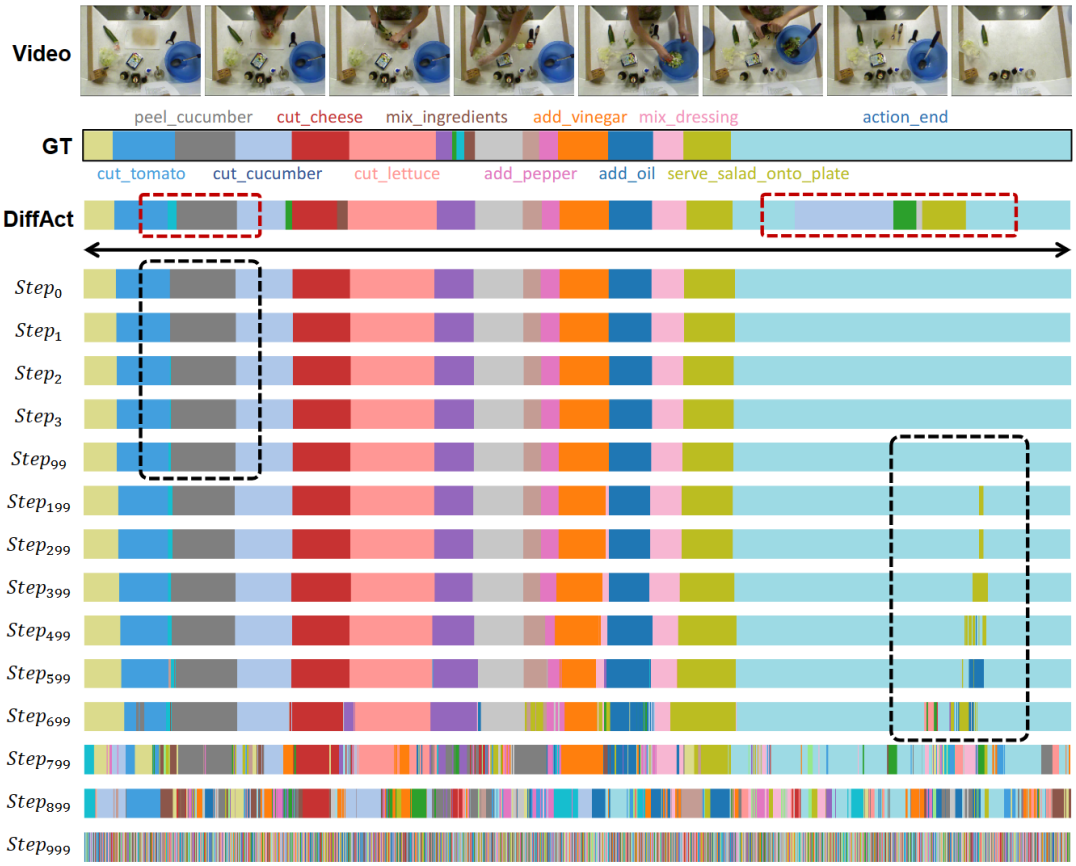


图 5-22　非锁步跳跃扩散去噪过程可视化结果与对比

图 5-23 和图 5-24 分别展示了 Breakfast 和 50Salads 数据集上的模型预测结果对比案例。黑色虚线框对本模型与 DiffAct 预测结果差异较大的部分进行了高亮。可视化结果显示，本模型在时序行为顺序预测与边界定位准确率方面均明显优于 DiffAct 模型。相较于 50Salads 数据集，Breakfast 中的视频具有较低的时序复杂性和预测难度。图 5-23 中的案例显示，本模型对行为顺序预测的能力明显优于 DiffAct 模型，图 5-24 中的 rgb-06-1.avi、rgb-17-2.avi、rgb-20-1.avi、

图 5-23　Breakfast 数据集部分序列可视化结果

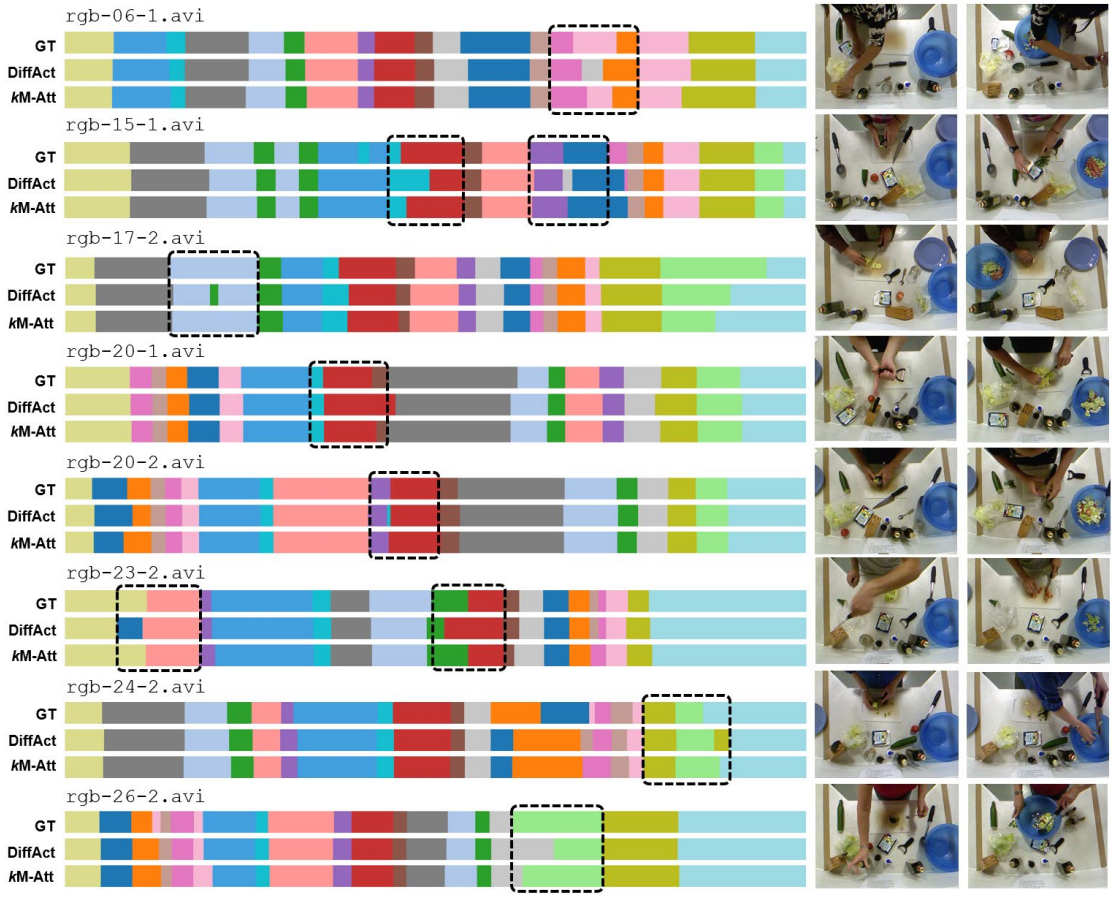


图 5-24　50Salads 数据集部分序列可视化结果

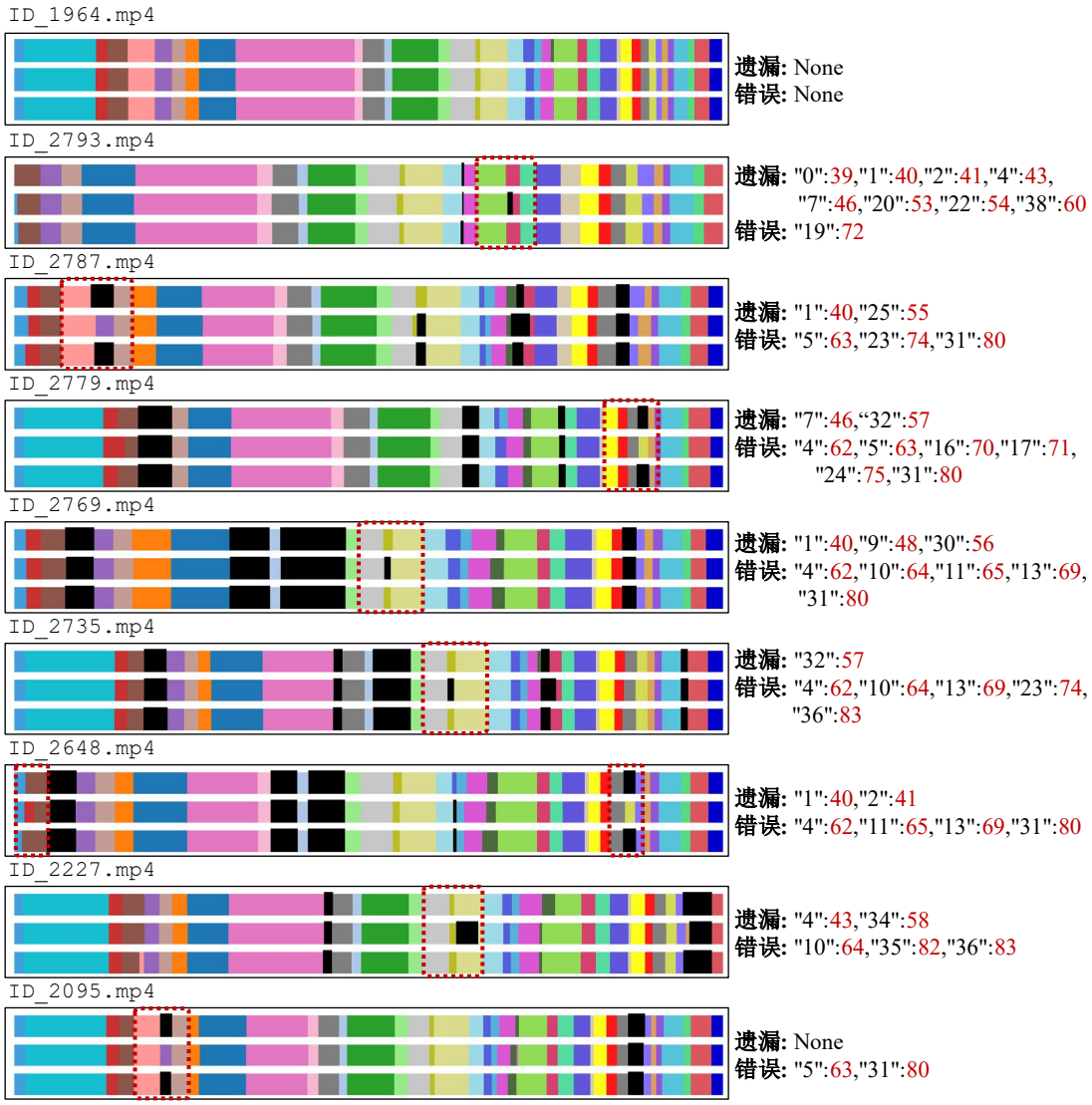


图 5-25 ThoSet 数据集部分序列可视化结果

rgb-20-2.avi、rgb-24-2.avi 同样支撑此结果。图 5-24 中的 rgb-15-1.avi、rgb-23-2.avi、rgb-26-2.avi 案例说明在行为顺序预测精准的前提下，本模型对行为交换边界的定位更加准确。

图 5-25 对 ThoSet 数据集中的部分案例进行了可视化。其中每个视频序列的三个视频分割标签结果分别为：标准序列、DiffAct 模型预测结果、*k*M-Att 模型预测结果。每个序列所对应的遗漏和错误行为标签列举在可视化图的右侧。为展示清晰，所有单流程错误行为标签均使用黑色绘制。对比结果显示，相较于当前最先进的 DiffAct 模型，本文所提出的 *k*M-Att 扩散分割框架能够在各种错误组合情况中有效避免错误检测的问题，从而生成更稳定、更精确的时序行为分割结果。优质的时序行为分割结果为后续的行为合规性判断奠定了坚实基础。

5.6 本章小结

本章针对现有时序医疗技能评估研究中行为划分标准不统一、流程行为粒度粗、算法功能不满足实际需求等问题开展了系列探究。本章首先依据医学生临床技能操作教材设计并构建了时序医疗行为知识图谱。该知识图谱能够同时对医疗行为的时序流程信息和相关知识文本进行表示，为后续时序医疗行为探究奠定了坚实基础。之后，本章以胸腔穿刺术为研究对象，在医疗行为知识图谱的基础上，构建了具有较高行为粒度划分的行为分析数据集 ThoSet。该数据集同时支持时序行为分割和行为合规性检测任务。在算法方面，本章受启发于人类大脑结构提出了时序聚类注意力机制模块 *k*M-Att，并将其和扩散分割模型相结合。公开数据集和 ThoSet 上的结果充分证实了本章方法的有效性。此外，本章在时序行为分割结果的基础上进一步提出了时序行为合规性检测算法。在 ThoSet 数据集中的实验结果充分说明了此检测算法的有效性。

第6章 总结与展望

6.1 本文工作总结

智能医疗技能评估系统在提升医务人员培训效率、节约培训环节人力成本、缓解一线医师工作压力等方面起到至关重要的作用。本文针对现有医疗技能评估研究所面临的细粒度行为识别数据集匮乏、行为标签划分粒度粗、算法性能受限和人机交互能力弱四个方面的问题，分别在任务范式创新、数据集构建、算法设计和集成应用四个层面中开展了系列研究，并将提出的算法应用在对应场景中，分别构建了操作质量评估子系统、复合错误行为识别子系统、多模态复合错误识别子系统和时序医疗行为分析子系统。最终构建了一套具备更广阔应用场景、更精细医疗行为划分粒度、更精准技能评估能力的医疗技能评估系统。本文所构建的系统为科技创新 2030—“新一代人工智能”重大项目中的医疗行为感知系统研究提供了重要支撑。

具体而言，本文在四个核心研究章节中的主要贡献如下：

（1）在医疗技能评估任务中设计了有效的管道自注意力特征增强策略。针对现有技能评估模型直接使用视频主干网络而不进行领域适配问题，本文将单目标跟踪器 SiamMask 引入到技能评估模型中，并设计了一种基于时空管道的稀疏高效特征交互模块 TSA。该模块能够通过跟踪器生成的跟踪框序列实现视频特征图中的元素框选，从而完成时空管道 ST-Tube 构建。最终由 TSA 模块中的自注意力机制完成特征增强。由于单目标跟踪器能够提供位置先验信息，自注意力计算过程能够进行更具针对性的特征增强，因此 TSA 模块具备稀疏高效特性。本文将 TSA 模块嵌入在了 I3D 视频网络主干内部，提出了技能评估框架 TSA-Net。在实验部分，本文将 TSA-Net 应用于了手术机器人操作技能评估数据集 JIGSAWS 和体育运动行为质量评估数据集 MTL-AQA 和 AQA-7。性能对比和计算量对比实验结果证实，本文所提出的 TSA-Net 框架能够以更少的计算量取得更高的评估性能。此章研究内容为医疗技能评估模型的设计提供了新的思路。

（2）提出了细粒度复合错误行为识别任务范式，在构建的 CPR-Coach 数据集基础上提出了 ImagineNet 算法。针对现有医疗技能评估研究中存在的错误行为识别研究匮乏、行为种类划分粒度粗等问题，本文对心肺复苏术中的胸外按压行为进行了深入探究。在专业医师的指导下，本文明确了包含 13 种错误行为的标签空间，以及包含 74 种复合错误行为的标签空间。依据此行为标签结构，本文招募志愿者构建了 CPR-Coach 数据集，该数据集能够同时支持单类错误识别

任务和复合错误识别任务，提供了 RGB、光流信息和 2D 关键点信息共三种模态数据。由于真实的医疗技能评估过程中存在“单类训练，多类测试”现象，本文受启发于人类的想象力机制提出了 ImagineNet 框架，该框架的本质是一种特征组合训练策略。在 CPR-Coach 数据集上的实验结果证实了 ImagineNet 框架在复合错误行为识别任务中的有效性。本章研究内容为医疗技能评估研究提供了新的任务范式，为错误行为检测相关研究提供了重要参考。

（3）将多模态预训练框架和提示词工程引入到医疗技能评估领域，提出了 CPR-CLIP 框架，并开展了医疗技能评估系统的随机对照试验。针对现有医疗技能评估系统的人机交互性差、难以真正落地应用等问题，本文在第三章研究内容的基础上，将多模态对比学习方法与提示词工程引入到了复合错误行为识别任务中。具体而言，本文从错误数量、错误种类和改正建议三个方面构建了错误信息描述提示语句。之后，通过多模态对比预训练框架 CLIP 的引入构造了预训练对比损失。最小化此损失函数能够实现语言模态和视觉模态中的特征对齐。在模型推理阶段，本文共设计了两种模型推理模式：单视频预测推理和特定类别视频检索推理。其中后者能够支持通过自然语言的方式对操作案例库进行快速检索与评估。在实验部分，本文首先在 CPR-Coach 数据集中探究了 CPR-CLIP 框架的单视频预测精度，实验结果证实了语言模态信息引入的必要性；为验证 CPR-CLIP 框架的辅助评估能力，本文招募医生开展了随机对照试验，结果证实在 CPR-CLIP 框架的辅助下能够提升近 4 倍的评估效率。此章研究内容在医疗技能评估系统的交互能力提升方面进行了开创性探索，为后续评估模型的落地应用奠定了一定基础。

（4）创建了时序医疗行为知识图谱，并基于此构建了细粒度时序行为分析数据集 ThoSet，在扩散时序行为分割模型中提出了特征增强模块 kM-Att，并提出了行为合规性评估算法。针对现有时序医疗行为分析研究领域中的行为划分标准不统一、时序错误操作研究匮乏等问题，本文首先以《中国医学生临床技能操作指南》教材为依据，设计并构建了时序医疗行为知识图谱。该知识图谱能够同时支持医疗行为流程和医疗行为知识的表示。之后，本文根据知识图谱对胸腔穿刺术进行了拆分和细化，并通过视频拼接策略构建了 ThoSet 数据集。该数据集具有高细粒度的行为划分标签，能够支持时序行为分割、错误操作识别和遗漏行为检测等任务。在算法创新方面，本文对现有扩散时序行为分割模型的特征增强和推理环节进行了针对性改进，分别提出了基于时序聚类注意力机制的特征增强模块 kM-Att，与基于非锁步跳跃的扩散去噪机制。在多个公开数据集和 ThoSet 数据集上的实验结果验证了算法的有效性。最后，本文基于行为分割结果提出了行为合规性评估算法对错误行为和遗漏行为进行检测。此章研究内容分别在知识

图谱构建、数据集构建和算法设计三个方面对时序医疗行为分析任务进行了探索，为后续时序医疗技能评估研究提供了重要参考。

综上所述，本文通过对四项核心内容的探究，完成了绪论中设定的研究目标：构建一套支持多任务、多场景的医疗技能评估系统。本研究填补了医疗技能评估系统领域中的部分空白，为技能评估研究提供了新的任务范式与数据集，为医疗技能评估算法的设计提供了新的思路，为手术场景下的细粒度医疗行为分析研究提供了重要参考，同时为后续技能评估系统的落地应用奠定了一定基础。本研究有望为医疗技能培训与考核效率的提升、医务人员数量缺口的快速补充和医疗服务系统压力的缓解做出一定贡献。

6.2 未来工作展望

目前医疗技能评估研究尚处于初级阶段，现有研究距离真正落地应用仍有较大的差距。本节分别从数据集构建、算法设计和落地应用三个方面对未来的工作进行展望。

数据集构建是医疗技能评估研究中的基础工作，医疗技能评估数据集的体量、丰富度和数据来源将会是未来研究所关注的重点：

（1）数据集体量与丰富度问题：与日常生活和体育运动场景中的技能评估数据集不同，医疗技能评估数据集的构建需要资深医师的参与，因此具有构建难度大、成本高等问题。在第五章中，本文提出了基于拼接策略的时序行为分析数据集构建方案，该方案能够充分利用有限的视频数据生成大量行为案例。此方案适用于构建第一人称视角操作数据集，而真实的医院有丰富的科室，每个科室内部的医疗操作与技能均有独特之处。因此在数据集的构建过程中，需要研究者在专业医师的指导下根据实际情况设计数据的采集策略。医疗技能评估数据集的体量和丰富度会对技能评估算法的性能和实用性产生重要影响。

（2）医疗行为数据的自动生成：医疗技能评估数据集面临着采集难度大、构建成本高等问题。通过虚拟现实技术实现医疗行为数据的自动生成为数据集构建提供了新的思路。随着数字人与人工智能生成内容（Artificial Intelligence Generative Content, AIGC）技术的不断发展，计算机生成的人、物体和场景越来越逼真。在自动驾驶研究领域中，NVIDA 和 Tesla 等公司已经开始探索通过仿真技术生成虚拟街景，为自动驾驶技术提供充足的训练素材；在文本引导视频生成领域中，OpenAI 公司最新发布的 Sora 模型能够生成单分钟时长的高清视频；在医疗机器人领域中，Ye 等人[204]提出了 RCare World 仿真模拟平台。该平台能够对护理机器人与虚拟人之间的交互进行建模，并且支持高逼真人体建模和多个家庭场景模拟。基于 AIGC 的医疗行为生成技术能够在大幅提升数据集丰富度的同

时有效节约数据获取成本。

医疗技能评估算法的性能会直接关联到系统最终的落地应用效果。由于医疗技能评估研究中数据具有稀缺性和宝贵性，如何通过改进算法实现有限数据的充分利用将会是未来探究的重点：

（1）基于多模态信息的医疗技能评估模型：本文在第四章中通过多模态预训练模型和提示词工程的引入提出了 CPR-CLIP 框架，该框架在提升模型的复合错误识别性能的同时，有效改善了模型的人机交互性能。这些探究结果充分证实：多个模态之间的信息存在互补特性，引入多模态信息能够带来更高的技能评估性能。在手术技能评估模型设计中，同时结合视觉信息和运动学传感器信息能够生成更加精准的技能评估结果。本文研究内容侧重于计算机视觉方法，并未结合运动学传感器数据进行算法构建。未来的医疗技能评估算法研究应当充分利用多模态信息。

（2）超长操作序列的高效建模：本文第五章所探究的胸腔穿刺术持续时间通常在几分钟内，操作序列的视频帧的数量达到了 8,000 甚至 10,000，几乎已经到达了时序行为分析模型的极限。然而，真实医疗场景中一场手术的持续时间动辄数小时，全程视频的帧数会达到惊人的规模。如何对这种具有极端时长的操作视频进行有效识别与评估是一个非常具有挑战性、也非常具有价值的问题。未来研究者可以充分参考 NLP 领域中处理超长输入序列的经验，例如大模型研究中的百万级分词处理、Transformer 模型中的高效自注意力机制实现等技术。对超长序列的实时处理能力是医疗技能评估系统的重要组成部分，在系统的落地应用过程中起到至关重要的作用。

（3）小样本学习（Few-Shot Learning, FSL）在医疗技能评估模型中的应用：人类拥有强大的学习与推理能力，能够依照仅有的少量参考案例对大量未知样本进行精准预测。少样本学习的目标是使人工智能模型在接触较少的训练案例后识别和分类新数据，即让模型拥有人的学习与推理能力。目前医疗技能评估研究面临着样本稀缺、行为采集难度大等困境，如何引入小样本学习策略实现对已有数据的充分利用是一个极具价值的研究方向。

（4）无监督与半监督学习：医疗技能评估数据集的构建过程面临着正例样本多、负例样本少的问题。通常负例样本的收集需要通过志愿者进行特意表演。在未来研究中，有望通过无监督与半监督学习方法在海量的操作视频库中进行智能化负例挖掘，并实现错误行为标签空间和数据集的自动构建，从而大幅节省医疗技能评估数据集构建过程中耗费的人力物力。

在**评估系统的落地应用**方面，构建技能评估系统的最终目标是服务于真实应用，从而缓解一线医师的教学与考核工作压力。现有的评估系统在应用方面仍有

以下方向需要进行探究：

（1）系统的人机交互能力改善：本文在第四章中通过引入多模态预训练框架与提示词工程赋予了 CPR-CLIP 框架多模态检索能力，并开展了随机对照试验验证了 CPR-CLIP 框架对评估效率提升的有效性。第四章只将 CLIP 框架引入到了医疗技能评估领域中，而多模态学习研究领域中还存在着其他形式的学习框架。这些方法均可以被引入到医疗技能评估系统中来，从而有效地改善现有模型交互性差的现状。

（2）模型推理性能的提升与优化：本文在第五章中探究了时序医疗行为分析任务，在构建的 ThoSet 数据集中，视频持续时间通常在 2～3 分钟内，图像帧的数量在 8,000 帧左右。过长的序列会导致时序行为分析模型具有较长的推理时间，从而影响技能评估系统的正常使用。目前学界中已有多种深度学习模型推理加速方案，例如网络剪枝、网络低位宽量化和硬件加速等。在医疗技能评估模型的落地应用过程中，模型的推理加速会为系统的响应速度提供保障。

参考文献

[1] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems. 2012.

[3] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional Two-Stream Network Fusion for Video Action Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 1933-1941.

[4] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks for action recognition in videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2740-2755.

[5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[6] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[C]//Proceedings of the Advances in Neural Information Processing Systems. 2017.

[7] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[8] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[9] PARMAR P, MORRIS B. Action quality assessment across multiple actions[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. IEEE, 2019: 1468-1476.

[10] PARMAR P, GHARAT A, RHODIN H. Domain knowledge-informed self-supervised representations for workout form assessment[C]//Proceedings of the European Conference on Computer Vision. 2022: 105-123.

[11] STEIN S, MCKENNA S J. Combining embedded accelerometers with computer vision for recognizing food preparation activities[C]//Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2013: 729-738.

[12] FATHI A, REN X, REHG J M. Learning to recognize objects in egocentric activities[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2011: 3281-3288.

[13] HUAULMÉ A, HARADA K, NGUYEN Q M, et al. PEg TRAnsfer Workflow recognition challenge report: Does multi-modal data improve recognition?[J]. arXiv preprint arXiv:2202.05821, 2022.

[14] HUAULMÉ A, SARIKAYA D, LE MUT K, et al. Micro-surgical anastomose workflow recognition challenge report[J]. Computer Methods and Programs in Biomedicine, 2021.

[15] GAO Y, VEDULA S S, REILEY C E, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling[C]//Proceedings of the Medical Image Computing and Computer-Assisted Intervention Workshop. 2014.

[16] ZIA A, SHARMA Y, BETTADAPURA V, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools[J]. International Journal of Computer Assisted Radiology and Surgery, 2016, 11: 1623-1636.

[17] ZIA A, SHARMA Y, BETTADAPURA V, et al. Video and accelerometer-based motion analysis for automated surgical skills assessment[J]. International Journal of Computer Assisted Radiology and Surgery, 2018, 13: 443-455.

[18] SHARMA S, KIROS R, SALAKHUTDINOV R. Action recognition using visual attention[J]. arXiv preprint arXiv:1511.04119, 2015.

[19] SHARMA Y, BETTADAPURA V, HAMMERLA N, et al. Video based assessment of OSATS using sequential motion textures[C]//Proceedings of the Workshop on Modeling and Monitoring of Computer Assisted Interventions. Springer, 2014.

[20] NWOYE C I, YU T, GONZALEZ C, et al. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos[J]. Medical Image Analysis, 2022.

[21] SCHOEFFMANN K, TASCHWER M, SARNY S, et al. Cataract-101: video dataset of 101 cataract surgeries[C]//Proceedings of the ACM Multimedia Systems Conference. 2018: 421-425.

[22] NAKAWALA H, BIANCHI R, PESCATORI L E, et al. “Deep-Onto” network for surgical workflow and context recognition[J]. International Journal of Computer Assisted Radiology and Surgery, 2019, 14: 685-696.

[23] SUN Z, KE Q, RAHMANI H, et al. Human action recognition from various data modalities: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[24] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.

[25] CABA HEILBRON F, ESCORCIA V, GHANEM B, et al. Activitynet: A large-scale video benchmark for human activity understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015: 961-970.

[26] LIU C, HU Y, LI Y, et al. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding[J]. arXiv preprint arXiv:1703.07475, 2017.

[27] ELLIS C, MASOOD S Z, TAPPEN M F, et al. Exploring the trade-off between accuracy and observational latency in action recognition[J]. International Journal of Computer Vision, 2013, 101: 420-436.

[28] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 1010-1019.

[29] LIU J, SHAHROUDY A, PEREZ M, et al. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2684-2701.

[30] CALABRESE E, TAVERNI G, AWAI EASTHOPE C, et al. Dhp19: Dynamic vision sensor 3d human pose dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019.

[31] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the Advances in Neural Information Processing Systems. 2014.

[32] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.

[33] LEV G, SADEH G, KLEIN B, et al. Rnn fisher vectors for action recognition and image annotation[C]//Proceedings of the European Conference on Computer Vision. 2016: 833-850.

[34] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2015: 4489-4497.

[35] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[36] CHEN C F, PANDA R, FAN Q. Regionvit: Regional-to-local attention for vision transformers[J]. arXiv preprint arXiv:2106.02689, 2021.

[37] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.

[38] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding?[C]//Proceedings of the International Conference on Machine Learning. 2021.

[39] LIU J, WANG G, HU P, et al. Global context-aware attention lstm networks for 3d action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 1647-1656.

[40] ZHANG P, LAN C, XING J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2017: 2117-2126.

[41] FRIJI R, DRIRA H, CHAIEB F, et al. Geometric deep neural network using rigid and non-rigid transformations for human action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12611-12620.

[42] YANG F, WU Y, SAKTI S, et al. Make skeleton-based action recognition model smaller, faster and better[C]//Proceedings of the ACM Multimedia Asia. 2019.

[43] BANERJEE A, SINGH P K, SARKAR R. Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(6): 2206-2216.

[44] CHI H gun, HA M H, CHI S, et al. Infogcn: Representation learning for human skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20186-20196.

[45] SHI F, LEE C, QIU L, et al. Star: Sparse transformer-based action recognition[J]. arXiv preprint arXiv:2107.07089, 2021.

[46] WANG Q, SHI S, HE J, et al. Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition[C]//Proceedings of the IEEE International Conference on Big Data. 2023: 936-945.

[47] SHAO D, ZHAO Y, DAI B, et al. Finegym: A hierarchical video dataset for fine-grained action understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition. 2020: 2616-2625.

[48] XU J, RAO Y, YU X, et al. Finediving: A fine-grained dataset for procedure-aware action quality assessment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2949-2958.

[49] PARMAR P, MORRIS B T. What and how well you performed? a multitask learning approach to action quality assessment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 304-313.

[50] XU C, FU Y, ZHANG B, et al. Learning to score figure skating sport videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(12): 4578-4590.

[51] BÉJAR HARO B, ZAPPELLA L, VIDAL R. Surgical gesture classification from video data[C]//Proceedings of the Medical Image Computing and Computer-Assisted Intervention. 2012: 34-41.

[52] ZAPPELLA L, BÉJAR B, HAGER G, et al. Surgical gesture classification from video and kinematic data[J]. Medical Image Analysis, 2013, 17(7): 732-745.

[53] MALPANI A, VEDULA S S, CHEN C C G, et al. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task[C]//Proceedings of the Information Processing in Computer-Assisted Interventions. 2014: 138-147.

[54] SHARMA Y, PLÖTZ T, HAMMERLD N, et al. Automated surgical OSATS prediction from videos[C]//Proceedings of the IEEE International Symposium on Biomedical Imaging. IEEE, 2014: 461-464.

[55] ZIA A, SHARMA Y, BETTADAPURA V, et al. Automated assessment of surgical skills using frequency analysis[C]//Proceedings of the Medical Image Computing and Computer-Assisted Intervention. 2015: 430-438.

[56] ZIA A, ZHANG C, XIONG X, et al. Temporal clustering of surgical activities in robot-assisted surgery[J]. International Journal of Computer Assisted Radiology and Surgery, 2017, 12: 1171-1178.

[57] MARTIN J A, REGEHR G, REZNICK R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents[J]. British Journal of Surgery, 1997, 84(2): 273-278.

[58] ISLAM G, KAHOL K, FERRARA J, et al. Development of computer vision algorithm for surgical skill assessment[C]//Proceedings of the Ambient Media and Systems: Second International ICST Conference. 2011: 44-51.

[59] ISLAM G, LI B, KAHOL K. An affordable real-time assessment system for surgical skill training[C]//Proceedings of the International Conference on Intelligent User Interfaces Companion. 2013: 111-112.

[60] CHEN L, ZHANG Q, TIAN Q, et al. Learning Skill-Defining Latent Space in Video-based Analysis of Surgical Expertise–A Multi-Stream Fusion Approach[M]//Medicine Meets Virtual Reality 20. IOS Press, 2013: 66-70.

[61] CHEN L, ZHANG Q, ZHANG P, et al. Instructive video retrieval for surgical skill coaching using attribute learning[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. 2015: 1-6.

[62] ZHANG Q, LI B. Relative hidden markov models for evaluating motion skill[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2013: 548-555.

[63] ZHANG Q, LI B. Relative hidden markov models for video-based evaluation of motion skills in surgical training[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(6): 1206-1218.

[64] VAKANSKI A, JUN H pil, PAUL D, et al. A data set of human body movements for physical rehabilitation exercises[J]. Data, 2018.

[65] DOUGHTY H, DAMEN D, MAYOL-CUEVAS W. Who's better? who's best? pairwise deep ranking for skill determination[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6057-6066.

[66] DOUGHTY H, MAYOL-CUEVAS W, DAMEN D. The pros and cons: Rank-aware temporal attention for skill determination in long videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7862-7871.

[67] LI Z, HUANG Y, CAI M, et al. Manipulation-skill assessment from videos with spatial attention network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 0-0.

[68] KUEHNE H, ARSLAN A, SERRE T. The language of actions: Recovering the syntax and semantics of goal-directed human activities[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2014: 780-787.

[69] TWINANDA A P, SHEHATA S, MUTTER D, et al. Workshop and challenges on modeling and monitoring of computer assisted interventions[EB/OL]. http://camma.u-strasbg.fr/m2cai2016/.

[70] MAIER-HEIN L, WAGNER M, ROSS T, et al. Heidelberg colorectal data set for surgical data science in the sensor operating room[J]. Scientific Data, 2021.

[71] TWINANDA A P, SHEHATA S, MUTTER D, et al. Endonet: a deep architecture for recognition tasks on laparoscopic videos[J]. IEEE Transactions on Medical Imaging, 2016, 36(1): 86-97.

[72] BAWA V S, SINGH G, KAPINGA F, et al. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods[J]. arXiv preprint arXiv:2104.03178, 2021.

[73] AL HAJJ H, LAMARD M, CONZE P H, et al. CATARACTS: Challenge on automatic tool annotation for cataRACT surgery[J]. Medical Image Analysis, 2019, 52: 24-41.

[74] SARIKAYA D, CORSO J J, GURU K A. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection[J]. IEEE Transactions on Medical Imaging, 2017, 36(7): 1542-1549.

[75] VAN AMSTERDAM B, FUNKE I, EDWARDS E, et al. Gesture recognition in robotic surgery with multimodal attention[J]. IEEE Transactions on Medical Imaging, 2022, 41(7): 1677-1687.

[76] MADAPANA N, RAHMAN M M, SANCHEZ-TAMAYO N, et al. Desk: A robotic activity dataset for dexterous surgical skills transfer to medical robots[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2019: 6928-6934.

[77] SINGH B, MARKS T K, JONES M, et al. A multi-stream bi-directional recurrent neural network for fine-grained action detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 1961-1970.

[78] LEA C, FLYNN M D, VIDAL R, et al. Temporal convolutional networks for action

segmentation and detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 156-165.

[79] LEI P, TODOROVIC S. Temporal deformable residual networks for action segmentation in videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6742-6751.

[80] FARHA Y A, GALL J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3575-3584.

[81] MAC K N C, JOSHI D, YEH R A, et al. Learning motion in feature space: Locally-consistent deformable convolution networks for fine-grained action detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6282-6291.

[82] LI S, FARHA Y A, LIU Y, et al. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 45(6): 6647-6658.

[83] WANG D, YUAN Y, WANG Q. Gated forward refinement network for action segmentation[J]. Neurocomputing, 2020, 407: 63-71.

[84] GAO S H, HAN Q, LI Z Y, et al. Global2local: Efficient structure search for video action segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16805-16814.

[85] HUANG Y, SUGANO Y, SATO Y. Improving action segmentation via graph-based temporal reasoning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14024-14034.

[86] ZHANG J, TSAI P H, TSAI M H. Semantic2graph: Graph-based multi-modal feature for action segmentation in videos[J]. arXiv preprint arXiv:2209.05653, 2022.

[87] YI F, WEN H, JIANG T. Asformer: Transformer for action segmentation[J]. arXiv preprint arXiv:2110.08568, 2021.

[88] DU D, SU B, LI Y, et al. Do we really need temporal convolutions in action segmentation?[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. 2023: 1014-1019.

[89] DU Z, WANG Q. Dilated transformer with feature aggregation module for action segmentation[J]. Neural Processing Letters, 2023, 55(5): 6181-6197.

[90] LI M, CHEN L, DUAN Y, et al. Bridge-prompt: Towards ordinal action understanding in instructional videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 19880-19889.

[91] VAN AMSTERDAM B, KADKHODAMOHAMMADI A, LUENGO I, et al. ASPnet: Action Segmentation With Shared-Private Representation of Multiple Data Sources[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 2384-2393.

[92] LIU D, LI Q, DINH A D, et al. Diffusion action segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 10139-10149.

[93] PAN J H, GAO J, ZHENG W S. Action assessment by joint relation graphs[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6331-6340.

[94] WANG Q, ZHANG L, BERTINETTO L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition. 2019: 1328-1338.

[95] WANG S, YANG D, ZHAI P, et al. Tsa-net: Tube self-attention network for action quality assessment[C]//Proceedings of the ACM International Conference on Multimedia. 2021: 4902-4910.

[96] PIRSIAVASH H, VONDRICK C, TORRALBA A. Assessing the quality of actions[C]//Proceedings of the European Conference on Computer Vision. 2014: 556-571.

[97] VENKATARAMAN V, VLACHOS I, TURAGA P K. Dynamical Regularity for Action Analysis.[C]//Proceedings of the British Machine Vision Conference. 2015.

[98] CHEN L, LI M, DUAN Y, et al. Uncertainty-aware representation learning for action segmentation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2022.

[99] TAO L, ELHAMIFAR E, KHUDANPUR S, et al. Sparse hidden markov models for surgical gesture classification and skill evaluation[C]//Proceedings of the Information Processing in Computer-Assisted Interventions. 2012: 167-177.

[100] AHMIDI N, HAGER G D, ISHII L, et al. Robotic path planning for surgeon skill evaluation in minimally-invasive sinus surgery[C]//Proceedings of the Medical Image Computing and Computer-Assisted Intervention. 2012: 471-478.

[101] AHMIDI N, PODDAR P, JONES J D, et al. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty[J]. International Journal of Computer Assisted Radiology and Surgery, 2015, 10: 981-991.

[102] DIPIETRO R, LEA C, MALPANI A, et al. Recognizing surgical activities with recurrent neural networks[C]//Proceedings of the Medical Image Computing and Computer-Assisted Intervention. 2016: 551-558.

[103] ZHANG Q, LI B. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model[C]//Proceedings of the International ACM Workshop on Medical Multimedia Analysis and Retrieval. 2011: 19-24.

[104] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems. 2017.

[105] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[106] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. OpenAI, 2018.

[107] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[108] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the Advances in Neural Information Processing Systems. 2022: 27730-27744.

[109] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[110] ZENG A, LIU X, DU Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[111] HAN K, GUO J, ZHANG C, et al. Attribute-aware attention model for fine-grained

representation learning[C]//Proceedings of the ACM International Conference on Multimedia. 2018: 2040-2048.

[112] FAN Q, ZHUO W, TANG C K, et al. Few-shot object detection with attention-RPN and multi-relation detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4013-4022.

[113] ZHAO H, ZHANG Y, LIU S, et al. Psanet: Point-wise spatial attention network for scene parsing[C]//Proceedings of the European Conference on Computer Vision. 2018: 267-283.

[114] BUADES A, COLL B, MOREL J M. A non-local algorithm for image denoising[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2005: 60-65.

[115] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 603-612.

[116] CHEN L C, YANG Y, WANG J, et al. Attention to scale: Scale-aware semantic image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 3640-3649.

[117] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.

[118] TANG Y, NI Z, ZHOU J, et al. Uncertainty-aware score distribution learning for action quality assessment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9839-9848.

[119] PARMAR P, TRAN MORRIS B. Learning to score olympic events[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2017: 20-28.

[120] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Proceedings of the European Conference on Computer Vision. 2014: 740-755.

[121] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Proceedings of the Advances in Neural Information Processing Systems. 2015.

[122] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 724-732.

[123] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[124] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[125] ABU-EL-HAIJA S, KOTHARI N, LEE J, et al. Youtube-8m: A large-scale video classification benchmark[J]. arXiv preprint arXiv:1609.08675, 2016.

[126] LIN J, GAN C, HAN S. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7083-7093.

[127] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.

2019: 6202-6211.

[128] WAGNER M, MÜLLER-STICH B P, KISILENKO A, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark[J]. Medical Image Analysis, 2023, 86: 102770.

[129] NIEBLES J C, CHEN C W, FEI-FEI L. Modeling temporal structure of decomposable motion segments for activity classification[C]//Proceedings of the European Conference on Computer Vision. 2010: 392-405.

[130] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[131] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2011: 2556-2563.

[132] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(8): 1819-1837.

[133] XU D, SHI Y, TSANG I W, et al. Survey on multi-output learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(7): 2409-2429.

[134] WANG J, YANG Y, MAO J, et al. Cnn-rnn: A unified framework for multi-label image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 2285-2294.

[135] JOHNSON R, ZHANG T. Effective use of word order for text categorization with convolutional neural networks[J]. arXiv preprint arXiv:1412.1058, 2014.

[136] COLE E, MAC AODHA O, LORIEUL T, et al. Multi-label learning from single positive labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 933-942.

[137] DMITRIEV K, KAUFMAN A E. Learning multi-class segmentations from single-class datasets[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9501-9511.

[138] ZACH C, POCK T, BISCHOF H. A duality based approach for realtime tv-l 1 optical flow[C]//Proceedings of the Pattern Recognition: 29th DAGM Symposium. 2007: 214-223.

[139] FANG H S, LI J, TANG H, et al. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[140] BETTADAPURA V, SCHINDLER G, PLÖTZ T, et al. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2013: 2619-2626.

[141] ZHANG B, ABBING J, GHANEM A, et al. Towards accurate surgical workflow recognition with convolutional networks and transformers[J]. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2022, 10(4): 349-356.

[142] ZHANG B, GHANEM A, SIMES A, et al. Swnet: Surgical workflow recognition with deep convolutional network[C]//Proceedings of the Medical Imaging with Deep Learning. PMLR, 2021: 855-869.

[143] KANNAN S, YENGERA G, MUTTER D, et al. Future-state predicting LSTM for early

surgery type recognition[J]. IEEE Transactions on Medical Imaging, 2019, 39(3): 556-566.

[144] MONFORT M, ANDONIAN A, ZHOU B, et al. Moments in time dataset: one million videos for event understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(2): 502-508.

[145] ZHANG H, CISSE M, DAUPHIN Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

[146] VERMA V, LAMB A, BECKHAM C, et al. Manifold mixup: Better representations by interpolating hidden states[C]//Proceedings of the International Conference on Machine Learning. 2019: 6438-6447.

[147] YANG C, XU Y, SHI J, et al. Temporal pyramid network for action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 591-600.

[148] SHAO H, QIAN S, LIU Y. Temporal interlacing network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 11966-11973.

[149] DUAN H, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2969-2978.

[150] BRUCE X B, LIU Y, ZHANG X, et al. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3522-3538.

[151] LI Y, WU C Y, FAN H, et al. Mvitv2: Improved multiscale vision transformers for classification and detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4804-4814.

[152] LIU Z, NING J, CAO Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3202-3211.

[153] GAO Y, BEIJBOM O, ZHANG N, et al. Compact bilinear pooling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 317-326.

[154] BEN-YOUNES H, CADENE R, THOME N, et al. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8102-8109.

[155] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[156] GU X, LIN T Y, KUO W, et al. Open-vocabulary object detection via vision and language knowledge distillation[J]. arXiv preprint arXiv:2104.13921, 2021.

[157] LI L H, ZHANG P, ZHANG H, et al. Grounded language-image pre-training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10965-10975.

[158] HE W, JAMONNAK S, GOU L, et al. CLIP-S4: Language-Guided Self-Supervised Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11207-11216.

[159] XU J, DE MELLO S, LIU S, et al. Groupvit: Semantic segmentation emerges from text supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. 2022: 18134-18144.

[160] VINKER Y, PAJOUHESHGAR E, BO J Y, et al. Clipasso: Semantically-aware object sketching[J]. ACM Transactions on Graphics, 2022, 41(4): 1-11.

[161] FRANS K, SOROS L, WITKOWSKI O. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders[C]//Proceedings of the Advances in Neural Information Processing Systems. 2022: 5207-5218.

[162] WANG M, XING J, LIU Y. Actionclip: A new paradigm for video action recognition[J]. arXiv preprint arXiv:2109.08472, 2021.

[163] GU J, HAN Z, CHEN S, et al. A systematic survey of prompt engineering on vision-language foundation models[J]. arXiv preprint arXiv:2307.12980, 2023.

[164] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the International Conference on Machine Learning. 2021: 4904-4916.

[165] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//Proceedings of the International Conference on Machine Learning. 2021: 8821-8831.

[166] DING L, XU C. Tricornet: A hybrid temporal convolutional and recurrent network for video action segmentation[J]. arXiv preprint arXiv:1705.07818, 2017.

[167] AZIERE N, TODOROVIC S. Multistage temporal convolution transformer for action segmentation[J]. Image and Vision Computing, 2022, 128: 104567.

[168] CHANG L, HOGLE N J, MOORE B B, et al. Reliable assessment of laparoscopic performance in the operating room using videotape analysis[J]. Surgical Innovation, 2007, 14(2): 122-126.

[169] WANG J, WANG Z, ZHUANG S, et al. Cross-enhancement transformer for action segmentation[J]. Multimedia Tools and Applications, 2023: 1-14.

[170] TIAN X, JIN Y, TANG X. Local–Global Transformer Neural Network for temporal action segmentation[J]. Multimedia Systems, 2023, 29(2): 615-626.

[171] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the International Conference on Machine Learning. 2015: 2256-2265.

[172] HIMMELSTEIN D S, LIZEE A, HESSLER C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing[J]. Elife, 2017.

[173] IOANNIDIS V N, SONG X, MANCHANDA S, et al. Drkg-drug repurposing knowledge graph for covid-19[J]. arXiv preprint arXiv:2010.09600, 2020.

[174] CHANDAK P, HUANG K, ZITNIK M. Building a knowledge graph to enable precision medicine[J]. Scientific Data, 2023, 10(1): 67.

[175] Downloads - STRING functional protein association networks[EB/OL]. https://string-db.org/cgi/download?sessionId=bZu2zuQkYSr2.

[176] DIEHL A D, MEEHAN T F, BRADFORD Y M, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability[J]. Journal of Biomedical Semantics, 2016, 7: 1-10.

[177] NGUYEN T M, NGUYEN T, LE T M, et al. Gefa: early fusion approach in drug-target affinity prediction[J]. IEEE/ACM Transactions on Computational Biology and

Bioinformatics, 2021, 19(2): 718-728.

[178] KEARNES S M, MASER M R, WLEKLINSKI M, et al. The open reaction database[J]. Journal of the American Chemical Society, 2021, 143(45): 18820-18826.

[179] GHORBANALI Z, ZARE-MIRAKABAD F, AKBARI M, et al. DrugRep-KG: Toward Learning a Unified Latent Space for Drug Repurposing Using Knowledge Graphs[J]. Journal of Chemical Information and Modeling, 2023, 63(8): 2532-2545.

[180] SCHRIML L M, ARZE C, NADENDLA S, et al. Disease Ontology: a backbone for disease semantic integration[J]. Nucleic Acids Research, 2012, 40(D1): D940-D946.

[181] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: a major update to the DrugBank database for 2018[J]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082.

[182] ERNST P, SIU A, WEIKUM G. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences[J]. BMC Bioinformatics, 2015, 16: 1-13.

[183] CHANG D, CHEN M, LIU C, et al. Diakg: An annotated diabetes dataset for medical knowledge graph construction[C]//Proceedings of the Knowledge Graph and Semantic Computing. 2021: 308-314.

[184] COVID-19 Knowledge Graph[EB/OL]. http://old.openkg.cn/dataset/covid-19.

[185] Chinese symptom set[EB/OL]. http://old.openkg.cn/dataset/symptom-in-chinese.

[186] DiseaseKG[EB/OL]. http://old.openkg.cn/dataset/disease-information.

[187] ZHANG H, ZONG Y, CHANG B, et al. 面向医学文本处理的医学实体标注规范 (Medical entity annotation standard for medical text processing)[C]//Proceedings of the Chinese National Conference on Computational Linguistics. 2020: 561-571.

[188] GUAN T, ZAN H, ZHOU X, et al. CMeIE: Construction and evaluation of Chinese medical information extraction dataset[C]//Proceedings of the Natural Language Processing and Chinese Computing. 2020: 270-282.

[189] ZHONG Z, CHEN D. A frustratingly easy approach for entity and relation extraction[J]. arXiv preprint arXiv:2010.12812, 2020.

[190] BEHRMANN N, GOLESTANEH S A, KOLTER Z, et al. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation[C]//Proceedings of the European Conference on Computer Vision. 2022: 52-68.

[191] ROY A, SAFFAR M, VASWANI A, et al. Efficient content-based sparse attention with routing transformers[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 53-68.

[192] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.

[193] RAKTHANMANON T, CAMPANA B, MUEEN A, et al. Searching and mining trillions of time series subsequences under dynamic time warping[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 262-270.

[194] CHEN M H, LI B, BAO Y, et al. Action segmentation with joint self-supervised temporal domain adaptation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9454-9463.

[195] WANG Z, GAO Z, WANG L, et al. Boundary-aware cascade networks for temporal action segmentation[C]//Proceedings of the European Conference on Computer Vision. 2020: 34-51.

[196] CHEN M H, LI B, BAO Y, et al. Action segmentation with mixed temporal domain adaptation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 605-614.

[197] SINGHANIA D, RAHAMAN R, YAO A. Coarse to fine multi-resolution temporal convolutional network[J]. arXiv preprint arXiv:2105.10859, 2021.

[198] AHN H, LEE D. Refining action segmentation with hierarchical video representations[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16302-16310.

[199] ISHIKAWA Y, KASAI S, AOKI Y, et al. Alleviating over-segmentation errors by detecting action boundaries[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 2322-2331.

[200] PARK J, KIM D, HUH S, et al. Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction[J]. Pattern Recognition, 2022, 129: 108764.

[201] KIM G hyeon, KIM E. Stacked encoder–decoder transformer with boundary smoothing for action segmentation[J]. Electronics Letters, 2022, 58(25): 972-974.

[202] XU Z, RAWAT Y, WONG Y, et al. Don't Pour Cereal into Coffee: Differentiable Temporal Logic for Temporal Action Segmentation[C]//Proceedings of the Advances in Neural Information Processing Systems. 2022: 14890-14903.

[203] LIU Z, WANG L, ZHOU D, et al. Temporal Segment Transformer for Action Segmentation[J]. arXiv preprint arXiv:2302.13074, 2023.

[204] YE R, XU W, FU H, et al. RCare World: A Human-centric Simulation World for Caregiving Robots[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2022: 33-40.

攻读博士学位期间取得的学术成果

已发表论文：

[1] **Shunli Wang**, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, Lihua Zhang*. A Survey of Video-based Action Quality Assessment[C]. *In Proceedings of the International Conference on Networking Systems of AI* (INSAI 2021, EI 会议综述) , 对应正文第一章

[2] **Shunli Wang**, Dingkang Yang, Peng Zhai, Chixiao Chen, Lihua Zhang*. TSA-Net: Tube Self-Attention Network for Action Quality Assessment[C]. *In Proceedings of the ACM International Conference on Multimedia* (ACM MM 2021, CCF-A), 对应正文第二章

[3] **Shunli Wang**, Shuaibing Wang, Dingkang Yang, Mingcheng Li, Haopeng Kuang, Xiao Zhao, Liuzhen Su, Peng Zhai, Lihua Zhang*. CPR-Coach: Recognizing Composite Error Actions based on Single-class Training[C]. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR 2024, CCF-A), 对应正文第三章

[4] **Shunli Wang**, Dingkang Yang, Peng Zhai, Lihua Zhang*. CPR-CLIP: Multimodal Pre-training for Composite Error Recognition in CPR Training[J]. *IEEE Signal Processing Letters* (IEEE SPL 2023, SCI 二区, IF=3.9), 对应正文第四章

[5] **Shunli Wang**, Shuaibing Wang, Bo Jiao, Dingkang Yang, Liuzhen Su, Peng Zhai, Chixiao Chen, Lihua Zhang*. CA-SpaceNet: Counterfactual Analysis for 6D Pose Estimation in Space[C]. *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS 2022, CCF-C)

[6] Shuaibing Wang, **Shunli Wang**, Dingkang Yang, Mingcheng Li, Ziyun Qian, Liuzhen Su, Lihua Zhang*. HandGCAT: Occlusion-Robust 3D Hand Mesh Reconstruction from Monocular Images[C]. *In Proceedings of the IEEE International Conference on Multimedia & Expo* (ICME 2023, CCF-B)

[7] Dingkang Yang, Shuai Huang, **Shunli Wang**, Yang Liu, PengZhai, Liuzhen Su, Mingcheng Li, Lihua Zhang*. Emotion Recognition for Multiple Context Awareness[C]. *In Proceedings of the European Conference on Computer Vision* (ECCV 2022, CCF-B)

[8] Bo Jiao, Haozhe Zhu, Jinshan Zhang, **Shunli Wang**, Xiaoyang Kang, Lihua Zhang, Mingyu Wang and Chixiao Chen*. Computing Utilization Enhancement for Chiplet-based Homogeneous Processing-in-Memory Deep Learning Processors[C]. *In Proceedings of the 26th Asia and South Pacific Design Automation Conference* (ASP-DAC 2021, CCF-C)

实审中发明专利：

[1] 张立华, **王顺利**，杨鼎康, 邝昊鹏. 一种基于深度学习的运动员行为质量评估方法, CN202111193385.4

[2] 张立华, **王顺利**, 王帅兵, 焦博, 翟鹏, 杨鼎康, 苏柳桢. 一种太空目标6D 位姿估计系统, CN202211159308.1

[3] 张立华, 钟楚轶, **王顺利**, 杨鼎康, 黄帅. 一种基于卷积神经网络和迁移学习的心脏病发作检测系统, CN202210948571.2

[4] 张立华, 杨鼎康, **王顺利**, 邝昊鹏, 黄帅. 一种基于情境感知的多模态情感识别方法和系统, CN202111080047.x

[5] 张立华, 黄帅, 杨鼎康, **王顺利**, 邝昊鹏. 一种用于陪伴机器人的多模态情感识别方法和系统, CN202111079583.8

[6] 张立华, 陈嘉伟, 王帅兵, 苏柳桢, **王顺利**, 余豪文, 李明程, 杨晓东. 一种基于增强现实的儿科气管插管培训系统及方法, CN202211511923.4

软件著作权：

[1] 基于机器学习的表情识别系统 v1.0, 2020SR1661701, 排名第二

[2] 基于深度学习的行为识别系统 v1.0, 2020SR1661702, 排名第二

竞赛获奖与 Demo 收录：

[1] ACM-MM 2021 3rd VRU Challenge 国际视频关系理解挑战赛, 二等奖

[2] MICCAI 2022 P2ILF Challenge 国际医学图像分割团队挑战赛, 排名 6/55

[3] 2021 年“华为杯”第 18 届中国研究生数学建模竞赛, 三等奖

[4] ICCV 2023 Demo: Composite Error Action Recognition System for Cardio-pulmonary Resuscitation (CPR), 第一作者

[5] ICCV 2023 Demo: CA-SpaceNet: Counterfactual Analysis for 6D Pose Estimation in Space, 第一作者

参与科研项目：

[1] 科技创新 2030—“新一代人工智能”重大项目, 标准化儿童患者模型关键技术与应用 (项目编号: 2021ZD0113500)

[2] 上海市人工智能科技重大专项, 人工智能前沿基础理论与关键技术 (项目编号: 2021SHZDZX0103)

致谢

时光荏苒，不觉间我已在复旦度过了五年的光阴，如今步入了博士生涯的尾声。在这紧张而充实的五年中，我经历了诸多考验：从开题时的迷茫，到撰写第一篇学术论文；从论文的数次修改，到最终被接收；从毕业论文的撰写，到找到理想的工作……一路走来，尽管曲折不断，但我也算幸运地一路前行。读博不仅是对心智的磨砺，更是对心性的淬炼。五年的学习生活使我更加成熟稳重，在面对工作与生活中常态化的挫折和失意时也愈加从容坚韧。回首这五年的时光，我的每一步成长都离不开导师、家人和同学们的支持与帮助。

首先，最重要的，是要感谢我的恩师张立华教授。是张老师给了我能够在复旦大学继续求学和深造的宝贵机会。张老师严谨治学的风范和精益求精的科研态度一直深深地熏陶着我。在张老师的身边，我深刻体会到了清华大学“自强不息，厚德载物”的校训。张老师一直教导我们，做科研一定要以落地应用为导向，不能仅仅止步于学术论文的发表。在博士课题的选题阶段，张老师给予了我诸多宝贵的意见，为我早期的研究指明了正确的方向。在研究过程中，张老师为我们营造了一片创新的沃土：充足的经费支持、优质的软硬件平台资源，以及丰富的外部交流与合作机会，这为我课题的正常开展奠定了坚实的基础。在科研素养培养方面，张老师为我提供了宝贵的锻炼机会，让我有幸参与到国家级项目中。在我们出现错误时，张老师也会耐心、及时地进行批评与指正。在生活方面，张老师对我们的关怀无微不至，特别是在疫情期间，他定期与实验室中的每一位同学进行线上沟通，想尽一切办法为同学们排忧解难。我曾无数次感慨，能成为张老师的学生是我人生中的一大幸事。这段求学之旅，正是因为有了张老师的引领和支持才变得如此丰富和充实。再次深深地感谢张立华教授，您是我人生道路上的导师和榜样。

感谢课题组中的康晓洋老师、陈迟晓老师、董志岩老师、曹凯老师、王哲老师在这五年中对我学业上的指导。感谢翟鹏师兄一路上对我的帮助与提携，在和师兄相处的过程中我收获颇丰。感谢鼎康一直以来在科研方面给予的重要支持和建议。感谢实验室中的苏柳桢、邝昊鹏、焦博、赵肖、张绪坤、杜洋涛、弓佩弦、龙威帆同学，和王帅兵、李明程、孙铭阳、钱子赟、韩铭浩师弟，我们一起齐心协力高效地完成了多项科研任务。感谢郑龙澍师兄、王紫烟师姐、吴秉慧师兄、胡坤师兄、韩暑师兄在科研路上对我的帮助。感谢博立电子科技集团的张沛轩和林野老师在工程细节上给予的充分指导。感谢工研院的阮舰老师、赵琳琦老师、姚明远老师、李悦老师在科研事务和学生工作方面的辛勤付出。特别感谢潘从元老师对我学业和生活上的指导与鼓励。

感谢在百忙之中出席答辩的谢少荣教授、王晓颖教授、袁建军教授、张荣君教授、金城教授、林燕丹教授。

感谢我的父母对我的养育之恩，在我成长和多年求学路上给了我支持与鼓励。家庭是我最温暖的港湾和最坚实的后盾。感谢丹怡在我漫漫求学路上给予的陪伴、支持与包容，你的鼓励为我注入了源源不断前行的动力。是你把我无数个失意的日子变成了无数个灿烂的日子。感谢我们的小猫小福一直以来的陪伴，熟睡中的小福总是能让我感叹这个世界是多么的美好。

回首五年的求学之路，压力与挫折是常态，收获成果的喜悦犹如绚丽但短暂的烟花。感谢这些压力和挫折能让我有所长进，那些没有让我们倒下的，只会让我们变得更强大。感谢所有收获成果的喜悦让我一步步地树立起自信。研究生阶段的学习让我领悟到，一个人的成功不应当被定义为比别人多走了多远，而应当被定义为比当初站在起点的自己多走了多远。博士毕业是学校求学阶段的终点，却是工作求知阶段的起点。在未来的工作和学习中，我期许自己能够成为一个从不抱怨、勇于担当、踏实肯干、听话出活的人。

最后，衷心感谢党和国家为我们创造了一个安定和谐的社会环境，让我们能够安心地完成学业。目前的世界正经历着百年未有之大变局，西方发达国家在高科技领域对我国开展实施围追堵截。作为新时代的科学研究人员，我们有责任为国家科技的自立自强贡献力量，早日突破西方发达国家的科技封锁，助力中华民族伟大复兴的实现。正如《亮剑》主题曲《中国军魂》歌词所歌颂的：从来是狭路相逢勇者胜，向前进，向前进！