中图分类号：

论文编号：10357E20301289

安徽大學
Anhui University

专业硕士学位论文

基于步骤分割和关键动作打分的手卫生评估研究

作者姓名　　朱启文

专业学位类别　电子信息硕士

专业学位领域　软件工程

指导教师　　李成龙

Research on Hand Hygiene Assessment based on Step Segmentation and Key Action Scorer

A Dissertation Submitted for the Degree of Master

Candidate：ZHU Qiwen

Supervisor：Prof. LI Chenglong

中图分类号：

论文编号：10357E20301289

硕 士 学 位 论 文

基于步骤分割和关键动作打分的手卫生评估研究

作者姓名	朱启文	申请学位级别	硕士
指导教师姓名	李成龙	职　　称	教授
学科专业	电子信息 软件工程	研究方向	计算机视觉
学习时间自	2020 年 9 月 1 日	起至	2023 年 6 月 30 日止
论文提交日期	2023 年 5 月 23 日	论文答辩日期	2023 年 5 月 17 日

摘　要

人工智能技术的快速发展给人们的日常生活带来了巨大的变化，其中计算机视觉技术为人们带来了极大的便利，动作质量评估作为计算机视觉中的一个热点话题，近些年受到了越来越多研究者的关注。动作质量评估是一种监测和评估行为质量的人工智能技术，这种技术通过利用计算机对视频中人的行为进行质量评价。在医疗、体育比赛、技能培训等领域有非常重要的实际应用价值。受新冠肺炎疫情影响，动作质量评估在医疗方面的应用被越来越多地提出。其中，手卫生（Hand hygiene）评估是动作质量评估在医疗方面重要的应用之一，手卫生是由世界卫生组织（World Health Organization，WHO）提出的一种六步洗手动作，目的是规范医务人员的洗手行为，良好的手卫生习惯可以预防绝大多数传染病，但是目前还没有很好的办法来监督医务人员做好手卫生，这就导致了疾病传播的潜在风险。现有的动作评估方法通常对整个动作视频进行整体质量评估，然而手卫生动作的细粒度知识以及不同步骤之间的关系在手卫生评估中是非常重要的。本文针对手卫生评估方法展开研究，并取得了如下成果：

第一，为了提供一个统一的手卫生评估平台，本文创建了一个统一的高质量视频数据集，称为 HHA300。HHA300 包含 300 个不同的洗手视频序列，共计超过 310000 帧，这些视频采集自不同的人员、场景、视角，包含有很多现实世界中手卫生行为可能发生的情况，保障了数据集的多样性和挑战性。为了同时完成步骤分割和动作评估的任务，本文在医务人员的监督下按照设计好的规则为 HHA300 提供了两种不同形式的标注，包括细粒度的帧级别标注和粗粒度的视频分数标注。高质量的标注使得该数据集可以很好用于该领域相关算法研究。

第二，提出了一种新的细粒度学习框架，以联合学习的方式来进行步骤分割和关键动作打分，从而使模型进行准确的手卫生评估。现有的步骤分割方法通常采用多阶段卷积网络缺乏长距离帧之间的相关性，容易导致错误的分割，为了解决这个问题，本文设计了一个多阶段卷积-Transformer 网络用于步骤分割。同时，本文观察到每个洗手步骤都包含几个决定洗手动作完成质量的关键动作，因此设计了一套基于可学习 Sigmoid 的关键动作打分器来评估每个步骤中关键动作的质量。在相关数据集上的大量实验表明，本文的方法很好地评估了手卫生视频的动作质量，验证了算法的有效性。

关键词：手卫生，动作评估，步骤分割，关键动作

Abstract

The rapid advancement of Artificial Intelligence technology has drastically altered the lives of people, particularly computer vision technology, which has made their daily lives much more straightforward. As a hot topic in computer vision, action quality assessment has attracted more and more researchers' attention in recent years. Action quality assessment is an artificial intelligence technology to monitor and assess the behavior quality. This technology assesses the quality of people's behavior in videos by using computers. It is very valuable for practical applications in medical treatment, sports competition, skill training and other fields. The application of action quality assessment in medical treatment has been increasingly suggested due to the epidemic in COVID-19. Among them, Hand hygiene assessment is one of the important applications of action quality assessment in medical treatment. Hand hygiene is a six-step hand washing action proposed by the World Health Organization (WHO), which aims to standardize the hand washing behavior of medical staff. Good hand hygiene habits can prevent most infectious diseases, but there is no good way to supervise medical staff to do hand hygiene well, which leads to the potential risk of disease transmission. The existing action assessment methods usually assess the overall quality of the whole action video. However, the fine-grained knowledge of hand hygiene actions and the relationship between different steps are very important in hand hygiene assessment. In this thesis, the methods of hand hygiene assessment are studied, and the following results are obtained:

Firstly, in order to provide a unified hand hygiene assessment platform, this thesis created a unified high-quality video dataset called HHA300. HHA300 contains 300 different hand-washing video sequences with a total of 310000 frames. These videos are collected from different people, scenes and perspectives, including many situations in which hand hygiene behaviors may occur in the real world, which ensures the diversity and challenge of the dataset. In order to complete the tasks of step segmentation and action assessment at the same time, this thesis provide HHA300 with two different forms of annotation according to the designed rules under the supervision of medical staff, including fine-grained frame-level annotation and

coarse-grained video score annotation. High-quality annotation makes this data set can be well used in the research on related algorithms in this field.

Secondly, this thesis proposes a new fine-grained learning framework, which uses joint learning to steps segmentation and key actions scorer, so that the model can accurately assess hand hygiene. To address the issue of wrong segmentation caused by the lack of long-distance frame correlation in existing multi-stage convolutional networks for step segmentation, this thesis proposes a multi-stage convolution-Transformer network for this task. At the same time, this thesis observes that each hand washing step contains several key actions that determine the quality of hand washing, so a set of key action scorers based on learnable Sigmoid is designed to assess the quality of key actions in each step. The effectiveness of the algorithm and its ability to assess the action quality of hand hygiene video are verified by a large number of experiments on related datasets using the method in this thesis.

Key words: Hand hygiene, Action assessment, Step segmentation, Key action

目 录

图 目 录

Content of Figures

表 目 录

Content of Tables

第一章　绪论

1.1　研究背景与意义

随着计算机技术的不断发展，人类的生产生活方式也受到了计算机技术的影响，其在人类活动中有了越来越多的应用。人工智能是计算机技术领域非常重要的一个研究方向，从这个名词提出至今，研究人员一直在致力于智能机器的研发，这种智能机器能够模拟人类的思维，并且拥有比人类更加强大的能力。得益于计算机软硬件的进步，人工智能在近年快速发展。机器学习作为人工智能研究中最活跃的分支，是一种利用数据自动构建分析模型的方法，能够使系统从数据中学习，并做出相应决策。随着机器学习算法的不断创新和计算机软硬件的高速发展，更加逼真的智能机器将在更多领域走进人们的生活。

让机器去理解人类的行为一直是研究者想要实现的目标，作为理解人体动作的关键技术之一，动作质量评估受到了越来越多的关注，并在近几年成为了人体行为理解研究中新的热门课题。动作质量评估的应用场景有很多，例如医疗领域的技能培训[1]-[2]、体育比赛的 AI 打分[3]-[10]、特定技能的等级评估任务[11]-[13]等。目前，动作评估技术在许多领域都有成功的案例，例如在体育运动领域中，2020 年东京奥运会上，国际体操联合会和国外技术公司合作，把 AI 评分系统引进到东京奥运会中让 AI 评分系统来承担比赛计时和运动员分数评定等方面的工作。AI 评分系统不仅能对运动员的表现进行打分，更重要的是提高分数的公平性，还能通过反馈动作质量来提高运动员的竞技水平。在动作质量评估任务中，对于一个含有待评估动作的视频，动作评估模型往往对其动作质量分数进行回归预测，使用的真值标签是待评估动作的真实分数。随着深度学习领域的发展，大多数动作评估模型的处理流程是首先使用卷积神经网络进行视频特征提取，然后通过对视频片段之间的序列关系进行建模，形成视频级别的特征表示。最后使用特征表示结合给出的真值标签训练出评分模型，通过训练好的模型得到视频中动作的评估分数。

近年来，受到新冠肺炎疫情影响，越来越多的研究将动作评估模型应用在医疗卫生领域。手卫生评估就是动作质量评估在医疗方面重要的应用之一，手卫生作为预防传染病的重要途径之一，对于其的规范性培训以及医务人员行为的监督都需要动作评估模型进行支撑。世界卫生组织将每年的 10 月 15 日定为“世界洗手日”，目的是呼吁全球通过规范的洗手行为，加强卫生安全意识，防止传染病的传播。世界卫生组织在 2013 年制

定了世界卫生组织多模式手卫生改进战略实施指南并被实验研究证明可行[14]。中国国家卫生健康委员会于 2019 年发布的医务人员手卫生规范[15]中提出了医务人员洗手方法规范如下：在流动水下，打湿双手；取适量消毒洗手液（肥皂），涂抹至整个双手；认真搓洗双手至少 15s，清洗双手所有皮肤，具体揉搓步骤如图 1.1 示（步骤不分先后）；在流动水下冲洗干净，使用纸巾擦干双手，取适量护手液护肤。

图 1. 1　手卫生步骤示例

Fig. 1. 1　Example of hand hygiene steps

然而，生活中很多人没有养成良好的洗手习惯，甚至存在一些误区：有的人不经常洗手，有的用纸巾或毛巾代替清水和肥皂，有的不使用具有消毒功能的洗手液，有的洗手时间过短，这些做法都不能达到科学洗手的标准。科学证据表明，规范的手卫生行为可以有效地预防多种传染病，在预防传染病的过程中，除了控制呼吸道传播途径外，还需要注意手部清洁和消毒。以新型冠状病毒为例，自爆发以来，全球已经累计确诊超 6 亿人，新冠病人将带有病毒的飞沫传播到空气中，这些飞沫可以黏附在外界的物体上，存活一定的时间[16]，如果人们的手在这时触碰到了这些带有病毒的物体，那么就会成为这些病毒传播疾病的重要媒介。据估计，人类一只手上平均携带 40 多万个细菌[17]，而手每小时至少会三次触碰到自己的鼻子等上呼吸道部位，这都增加了疾病传播的风险。为了避免这样的传播，人们应该进行规范的洗手行为，因此对洗手过程的标准程度有一个正确的评估就显得尤为重要。得到正确的评估后，人们可以利用评估结果对自己的洗手动作进行修正，进而达到标准的手卫生要求，减少传染病的传播。但是，目前针对手卫生评估的研究[18]-[23]都是基于图像分类的，不但大大限制了其应用场景，而且由于手卫生步骤之间存在一定的相似性，导致其准确度不高。因此如何利用手卫生视频的特性进

行鲁棒的手卫生视频评估的研究非常必要。

1.2 国内外研究现状

视频数据在各种场景下的快速增长，对视频管理、分析和处理提出了新的挑战和需求。视频理解旨在通过智能分析技术，自动化地对视频中的内容进行识别和解析，从而实现对视频数据的高效利用。视频理解算法顺应了这个时代的需求，近年来受到了广泛关注，取得了快速发展。视频理解涉及生活的多个方面，目前视频理解已经发展成一个十分广阔的学术研究和产业应用方向。本节将介绍视频理解中的两大基础领域：视频动作分割和视频动作质量评估。

1.2.1 视频动作分割

近年来，深度学习技术的快速发展以及存储设备性能的不断提高，利用视频数据的视频分析方法取得了很大的进步。从经典的滑窗方法到马尔可夫模型再到深度学习方法，相关动作分割算法的性能已经得到显著提高。与此同时，为了促进该领域的发展，研究人员构建了多个大规模的动作分割数据集，为相关算法提供了有效的评测平台，并加快了算法性能的迭代。视频动作分割可以看作是一个视频逐帧分类问题，将视频中的每一帧对应于一种动作。早期的动作分割方法[24]-[25]尝试将适用于修剪视频的方法与滑窗结合在一起，具体来说，这些方法采用不同尺度的时间窗口来检测和分类动作片段。但是这些方法需要大量的计算资源，很难应用于长视频。除此之外，还有一些方法[26]-[28]在逐帧分类器上使用马尔可夫模型建模粗粒度的时序关系，但是由于需要在长序列上求解最大值问题，这些方法通常运行非常缓慢。大多数最近的方法[29]-[31]使用时序卷积网络来关注帧间相关性。2019 年提出的 MS-TCN[32]使用时序卷积网络来聚合视频帧之间的依赖信息，并使用多阶段时序卷积网络来更好地调整分类结果。Li 等人[33]在 MS-TCN 的基础上设计了新的结构，解决了某些局部信息无法提取的问题。Wang 等人[34]用自适应级联网络来区分困难样本，大大提高了困难样本的分类精度，结合动作边界信息的时间正则化方法减少了过分割。该模型的时间感受野在动作分割中起着重要作用，大的感受野有助于建模长距离帧之间的依赖性，而小的则更加关注局部的细节信息。近年来，神经架构搜索（Neural Architecture Search，NAS）受到了广泛的关注[35]-[38]，它旨在实现网络结构设计的自动化，从而减轻研究者们的手工设计负担，并能够在图像分类等任务的准确度上达到或者超过手工设计的水平。受到这一启发，Gao 等人[39]提出通过从全局到局

部的搜索方案来寻找感受野的更好组合，先利用全局搜索来找到粗略的组合，再利用局部搜索来进一步获得精炼的感受野组合模式，在全局搜索的基础上，还提出了一种期望指导的迭代局部搜索方案，以有效地优化组合。自从 Transformer 模型在自然语言处理任务中取得优越的性能后[40]-[42]，其在计算机视觉方面的应用也引起了众多研究者的关注，并在 2020 年首次在图像分类任务上得到了比卷积神经网络模型更好的结果。此后，越来越多的研究[43]-[45]开始将 Transformer 模型迁移到计算机视觉领域，并利用其强大的建模能力取得很好的效果。Yi 等人[46]利用额外的时序卷积和预定义的层次表示模式解决了将 Transformer 引入动作分割任务时遇到的归纳偏置等问题，设计了一个基于 Transformer 的高效动作分割模型。为了解决边界模糊和过分割问题，Li 等人[47]利用高效时序金字塔网络捕获局部和全局信息提高了分割性能，并且提出了一种新的后处理无监督方法 LBS（Local Burr Suppression）减少了过分割错误。

1.2.2　视频动作质量评估

在手卫生评估方面，Zhong 等人[20]应用迭代方法设计了一个手卫生行为检测系统，利用行为分类的结果来实现手卫生评估。手卫生的研究[48]-[49]更多的是关于姿势估计和检测任务，这和本文要完成的手卫生评估是不同的。关于一般动作质量评估的研究很多。现有方法通常将其视为一项回归预测任务，最终目标是缩小回归预测得分与专家给出的真值得分之间的差距。Pirsiavash 等人[50]使用离散余弦变换将关节轨迹编码为输入特征，并使用支持向量回归来构建从特征到最终得分的映射，并且可以提供可解释的反馈，说明如何提高动作质量。基于人类使用的注意机制，在评估视频时，Li 等人[13]提出了一种基于递归神经网络的空间注意模型，该模型考虑了来自先前帧的累积注意力状态和关于正在进行的任务的高级知识。Paret 等人[10]证明 C3D[51]能够有效地保存视频中的时空信息，有助于提高评估任务的性能，并且提出了一种增量 LSTM 训练策略，用于在样本有限的情况下进行有效的训练，提高了预测质量。Pan 等人[52]以 I3D[53]为骨干网络提取时空特征，建立可训练的关节图，分析其关节运动，提出了新颖的学习关节动作细节的框架来评估动作表现。Xu 等人[9]提出了一个包含两个 LSTM 的深度架构来学习视频的不同尺度特征分别称为 S-LSTM 和 M-LSTM，S-LSTM 主要使用自注意力策略用来选择重要的片段特征学习表示局部信息，并直接用于回归任务；M-LSTM 的功能是在多个尺度上对全局和局部信息建模，并通过跳跃 LSTM 来节省计算代价。这两个子网络可以直接用作预测的模型，也可以集成到一个框架中进行最终的回归任务，除此之外还给出了一

个花样滑冰运动视频数据集。最近，Zeng 等人[54]不仅利用了视频中的动态信息，而且还注意到了静态的姿态信息，通过在基于图的上下文感知注意模块中结合静态和动态信息，可以通过良好的视频表示实现更加准确的动作评估。Yu 等人[55]利用对比学习的思想，将动作评估问题转化为另一个视频的回归相对分数问题，通过参考具有相同属性的另一个视频的分数，利用两个视频的相对信息，突出视频之间的差异性，来学习待评估视频的分数。类似的，Jain 等人[56]提出了一种基于深度度量学习的动作评分系统，该系统可以学习两个视频之间的相似性，并通过对比进行评估。Xu 等人[57]构造了一个新的细粒度数据集，所有视频都在两个级别上进行语义标注，即动作类型和子动作类型，其中，不同的动作类型由不同的子动作类型组合生成。同时提出了一种过程感知的方法，以细粒度的方式量化查询和样本之间的质量差异用于动作评估。

1.3 论文结构安排

为了推动动作评估领域的发展，本文针对手卫生评估相关技术开展了一系列研究，主要工作如下：

（1）提出了一个大规模手卫生评估数据集（Hand Hygiene Assessment 300，HHA300）。为了充分开展视频分析和手卫生评估技术相关研究，本文提出了一个手卫生数据集用于建立一个统一的研究平台。本文收集了 300 个手卫生视频序列，所有视频的总帧数超过了 310000，视频序列的平均长度和最大长度分别为 1048 帧和 1579 帧。该数据集包含了手卫生评估任务中的大部分现实场景挑战。为了提高数据集的多样性和挑战性，HHA300 包含各种各样的场景以及不同的拍摄角度和人员。本数据集在采集过程中统一采用万宝泽监控摄像头，这是一种被广泛应用于家居，监控等领域的监控摄像。

（2）提出了一种基于步骤分割和关键动作打分器的手卫生评估算法。由于现有动作评估方法大多都是针对短动作并对整个视频进行回归，这样的设计无法针对像手卫生这样含有很多步骤的长视频动作。在面对手卫生数据的时候，由于手卫生动作的长时性以及步骤之间的相似性，现有的模型无法对每个步骤的动作质量进行准确的评估。除此之外，本文在构建数据集时发现，对于每一个手卫生步骤，都有一些规定的关键动作来决定该步骤的完成质量。因此本文设计了一种基于步骤分割和关键动作打分器的手卫生评估算法，一方面，本文设计了一个基于多阶段卷积-Transformer 的步骤分割网络，另一方面，提出了一个基于可学习 Sigmoid 的关键动作打分器以适应不同步骤的不同关键

动作。本文在 HHA300 数据集和相关数据集上进行了充分的实验和分析，验证了该方法在手卫生评估以及步骤分割方面的有效性。

基于以上所述内容，本文的组织结构安排如下：

第一章绪论：主要介绍手卫生视频动作评估的背景和研究意义，还简要地概述了视频理解领域相关技术的现状，最后本文将汇总本文的研究内容，并对整篇文章的结构加以说明。

第二章视频动作分析技术：主要介绍和本文相关的视频动作分割技术，视频动作质量评估技术在计算机视觉当中的应用，最后给出小结。

第三章手卫生视频数据集构建：详细介绍了 HHA300 数据集的构建方式，通过一系列分析，展示了 HHA300 手卫生视频数据集的规模和多样性。

第四章基于步骤分割和关键动作打分器的手卫生评估：首先介绍手卫生评估方法的研究动机，然后介绍了本文提出方法的网络结构以及相应细节。最后，通过一系列实验以及相应分析证明了该方法的有效性。

总结和展望：总结了本文所有的工作内容以及贡献，提出其中存在的不足地方，并对未来的研究方向做出规划，以期达到更高的水平。

第二章　视频动作分析相关技术介绍

本章节将重点介绍视频动作分析相关技术，包含了两部分内容，分别为基于时序卷积和边界信息的动作分割技术和基于动静态结合的动作质量评估方法的详细分析和讨论。

2.1　基于时序卷积和边界信息的动作分割技术

本小节主要介绍基于时序信息和边界信息的动作分割技术，包括基于时序卷积的动作分割和基于边界信息的动作分割两个方面。

2.1.1　基于时序卷积的动作分割算法

多阶段时序卷积网络[32]（Multi-Stage Temporal Convolutional Network， MS-TCN）提出了一个多阶段时序卷积架构用于长时的动作分割任务，作者首先提出了一个单阶段模型，为了避免降低时间分辨率和增加参数数量，它只由时序卷积层构成，不包含池化层和全连接层。这样，该模型可以处理任意长度的输入序列，并且保持较低的计算复杂度。作者将这个模型命名为单阶段时序卷积网络（Single-Stage Temporal Convolutional Network，SS-TCN）。为了调整输入特征的维度，SS-TCN 的第一层采用 1×1 卷积层，使其与网络中的特征映射数量一致，在此之后包含多层空洞卷积[58]，每一层的空洞率是前一层的两倍，即 1，2，4，……，512 等。这种设计使得 SS-TCN 能够在参数较少的情况下获得较大的感受野。所有空洞卷积层都使用相同数量和大小为 3 的卷积核，并对每层输出应用 ReLU 激活函数[59]。受到 ResNet[60]的启发，作者进一步使用残差连接来促进梯度传递。最后通过 Softmax 函数对每一帧进行概率分类。如图 2.1 所示：

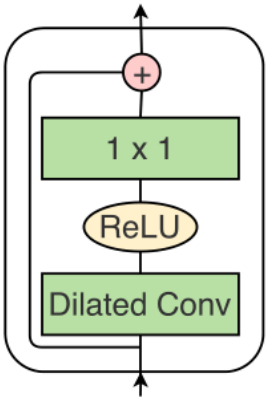

图 2. 1　单阶段时序卷积网络

Fig. 2. 1　Single-Stage Temporal Convolutional Network

在 SS-TCN 的基础上，为了提高时序动作分割的性能，作者从人体姿态估计任务中借鉴了一种多阶段网络结构[61]-[63]，该结构将若干个预测器按顺序串联，使得后续的预测器可以利用前面预测器的输出进行进一步的优化。这种多阶段或堆叠式的架构已经在许多任务中证明了其有效性。基于此，作者提出了一种多阶段时序卷积网络即 MS-TCN 用于时序动作分割任务。在这个多阶段模型中，每个阶段都从前一阶段获得一个初始预测，并对其进行细化。第一阶段的输入是视频的帧级特征，如公式(2.1)，(2.2)所示 ：

$$Y^0 = x_{1:T} \tag{2.1}$$

$$Y^s = F(Y^{s-1}) \tag{2.2}$$

其中Y^s为s阶段的输出，F为上文讨论的 SS-TCN。使用这种多阶段体系结构有助于提供更多上下文信息来预测每一帧的类标签。此外，由于每个阶段的输出都是结果预测，因此网络能够捕获动作类之间的依赖关系并学习合理的动作序列，这有助于减少过分割（Over-Segmentation）错误。

尽管 MS-TCN 取得了很好的效果，仍有一些问题制约了它的进一步提升。首先在 MS-TCN 中，第一阶段是用来生成初始预测，后面的几个阶段是用来调整这一预测。很明显这是两个不同的任务，但在 MS-TCN 中他们拥有相同的网络结构。第二，尽管在 MS-TCN 中，高层有较大的感受野，但低层的感受野依旧很小。针对第一个问题，作者将第一个阶段和其余的阶段进行了解耦设计，使不同的阶段能适应不同的任务。因为生成初始预测比调整预测更加复杂，因而第一阶段需要更为复杂的设计。该研究发现，由于后几个阶段的任务是相同的，这些阶段的参数也可以共享，进而进一步降低了模型的参数量。而对于第二个问题，作者设计了一个双重空洞层（Dual dilated layer， DDL），使得在不同层和不同尺度的信息得以融合。双重空洞层的结构如图 2.2 所示。

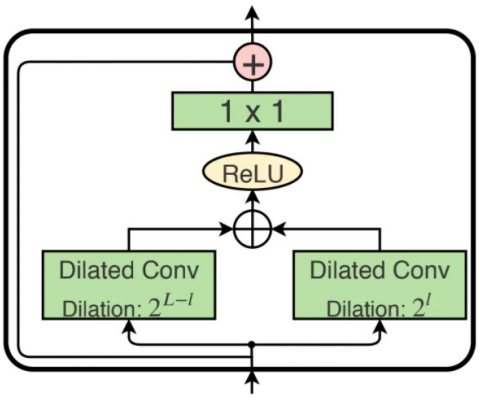

图 2. 2　双重空洞层

Fig. 2. 2　Dual dilated layer

MS-TCN++是在 MS-TCN 中引入 DDL 结构。与 MS-TCN 一样，MS-TCN++第一个阶段输出的是原始的预测结果，之后每个阶段都是对前一个阶段的输出结果进行修正操作。作者经过试验得到的最终的 MS-TCN++ 的结构图如图 2.3 所示，其中预测生成阶段（Prediction Generation Stage）的结构是 DDL，修正阶段（Refinement Stage）的结构是 SS-TCN。

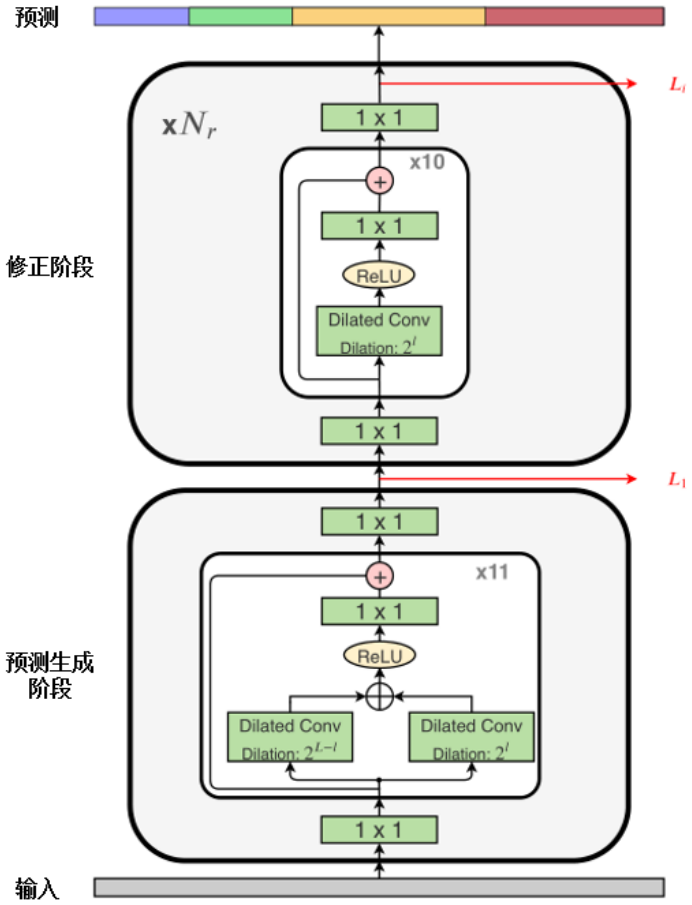

图 2. 3　MS-TCN++框架

Fig. 2. 3　Framework of MS-TCN++

2.1.2　基于边界信息的动作分割算法

动作边界信息对于视频分析任务具有重要意义。Wang 等人[34]提出了一种基于动作边界信息和级联网络的动作分割方法 BCN。由于动作分割需要处理未经修剪的长视频，因此如何建模长距离帧间的相关性是一个关键问题。以往的动作分割模型通常采用 LSTM[64]，编解码结构、空洞卷积等技术来增大网络感受野，从而提高预测性能。然而，这些模型也存在一些缺陷：一方面，由于动作分割要求输出结果与输入视频具有相同的时间分辨率，并且评价指标（如准确率）对于单帧预测错误非常敏感，因此模型容易过拟合训练数据，导致过分割现象；另一方面，由于计算资源的限制，无法无限制地增加

模型参数和深度来处理难样本。作者通过可视化实验发现，如图 2.4 所示，当前最先进的方法 MS-TCN[32]和 MS-TCN++[33]在两个方面存在问题：首先是在动作边界附近或者具有歧义性的帧上分类精度低；其次是在长动作区间内出现过分割问题。

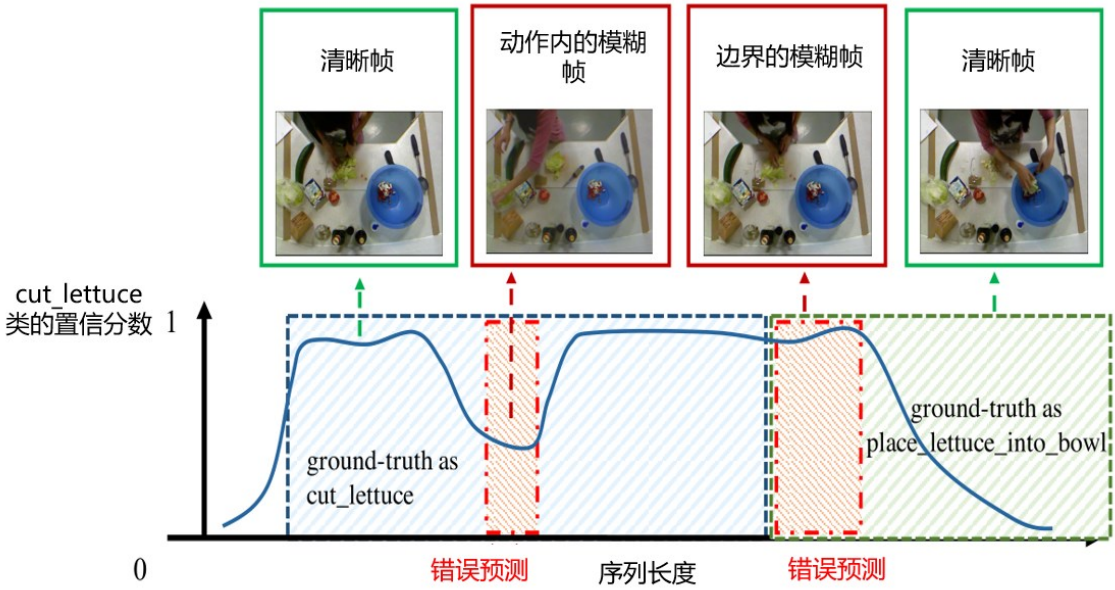

图 2. 4　置信分数可视化图

Fig. 2. 4　Schematic diagram of epipolar rectification

为了解决视频中存在的两种挑战，作者分别进行了相应的分析。一方面，视频中存在大量的模糊帧，如动作边界、视角转换、遮挡等，这些帧难以提取有效的特征，而简单地增加模型复杂度会导致过拟合现象。另一方面，视频中同时包含了简单帧和困难帧，它们具有不同的语义信息和歧义程度。如果使用一个统一的模型来处理这些帧，会由于训练数据的分布不一致和不均衡而导致对困难帧的预测性能下降（表现为低置信度或错误的预测）。为了克服视频中样本难度的差异，作者提出了一种动态的建模方法，该方法根据样本的困难程度，自适应地选择不同网络进行学习：对于简单样本，使用浅层子网络学习；对于困难样本，使用深层子网络学习。同时，作者通过自适应的调整损失函数和聚合函数中各个子网络的权重，实现了对不同难度样本的动态建模。该方法有效地提高了对困难样本的分类精度。然而，动态建模的方法也带来了一个问题，即过分割现象更加严重，这需要一个有效的时序正则化方法来约束模型。作者提出了一种基于邻域信息池化的时序正则化方法，与常用的基于相邻结果差值惩罚的损失函数不同，该方法利用邻域信息来增强预测结果的光滑程度。单单利用简单的池化肯定会导致分类精度的下降，或者平滑效果不佳的情况，因此作者引入了时序动作检测中的“动作边界”概念，

将子网络预测的动作边界作为参数，传入一种新提出的带参数的池化方法：局部分界池化（Local Barrier Pooling，LBP）。该方法利用动作边界来保持内部语义一致性，隔离某一动作外部信息的干扰，有效地改善了过分割现象。作者提出的框架如图 2.5 所示，处理流程是，给定一个长度为 T 的视频帧序列，首先提取每帧的 I3D[53]特征作为输入，提取规则为以每帧为中心，在局部区间提取特征作为此帧的特征得到长度为 T 的特征序列，将其分别输入到阶段级联分支和局部分界池化分支中。对于阶段级联分支，作者使用 n 个阶段的 MS-TCN 为骨干网络，前 n-1 个阶段为用于处理不同难度分布的级联阶段，最后一个阶段为融合前面 n-1 个阶段结果的融合阶段。对于级联阶段，作者首先把第一个阶段的权重初始化为全 1 向量，然后每个阶段都根据上一个阶段对于每一帧的分类置信度自适应地利用一个权重因子调整当前阶段对于这一帧的权重，具体为：对于上一个阶段预测结果中置信度高的帧，降低权重；对于上一个阶段的预测结果中置信度低的帧，提高权重，从而最终得到一个权重矩阵。得到这样一个权重矩阵以后在各个阶段之间进行权重矩阵聚合，并在损失函数中加入权重矩阵。在局部分界池化分支中，作者首先使用边界生成模块预测动作边界，并通过阈值筛选出一些有效边界。然后，将这些边界和阶段级联分支预测的分类置信度输入到 LBP 模块中，得到最终的预测结果。

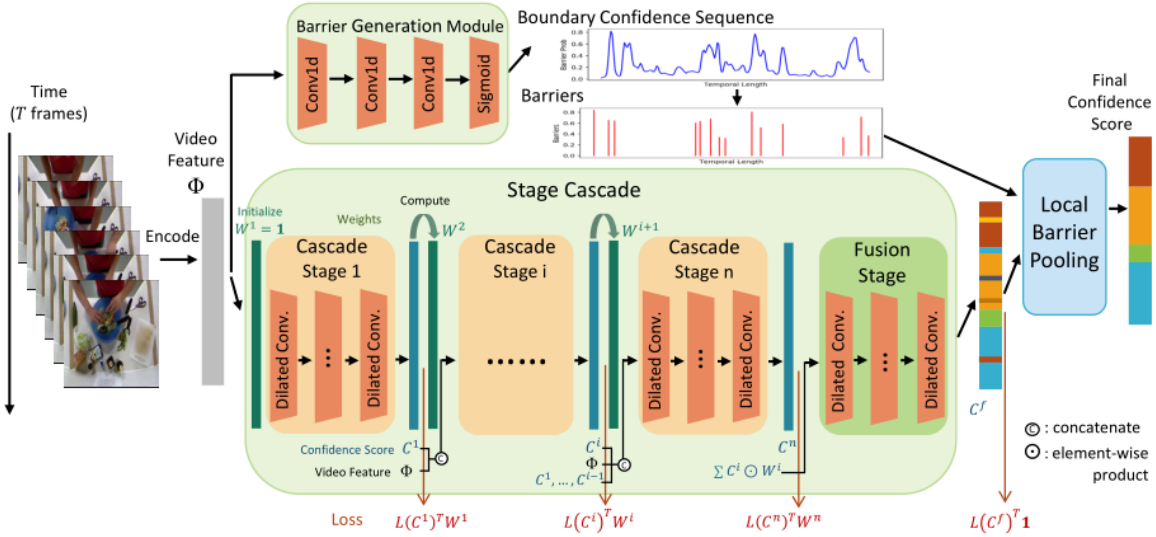

图 2. 5　BCN 框架

Fig. 2. 5　Framework of BCN

2.2　基于动静态结合的动作质量评估方法

动作质量评价的目的是对体育视频进行评分。然而，现有的研究大多只关注视频动

态信息（即运动信息），而忽略了运动员在视频中所表现的具体姿势，这对于长视频中的动作评估是很重要的。因此，Zeng 等人[54]提出了一种新的混合动态-静态上下文感知注意力网络（Action - net）用于长视频中的动作评估。该网络不仅学习视频的动态信息，也关注特定帧中运动员的静态姿态，反映了特定时刻的动作质量。该网络由两部分组成：视频动静态特征提取和视频动静态特征联合。如图 2.6 所示，在视频动静态特征提取部分，作者使用两个流分别提取视频中的动态信息和静态信息；在视频动静态特征联合部分，作者使用一个由时序切片图卷积单元（Temporal Clipwise GCN Unit，TCG-U）和两个注意力单元（Attention Unit ，ATT-U）组成的上下文感知注意力模块（Context-aware Attention），来融合两个流的特征，并回归最终的视频评分。TGC-U 用于探索实例之间的关系，注意力单元用于为每个实例分配适当的权重。该方法通过结合动静态信息和上下文感知注意力机制，有效地提高了长视频中的动作质量评估性能。

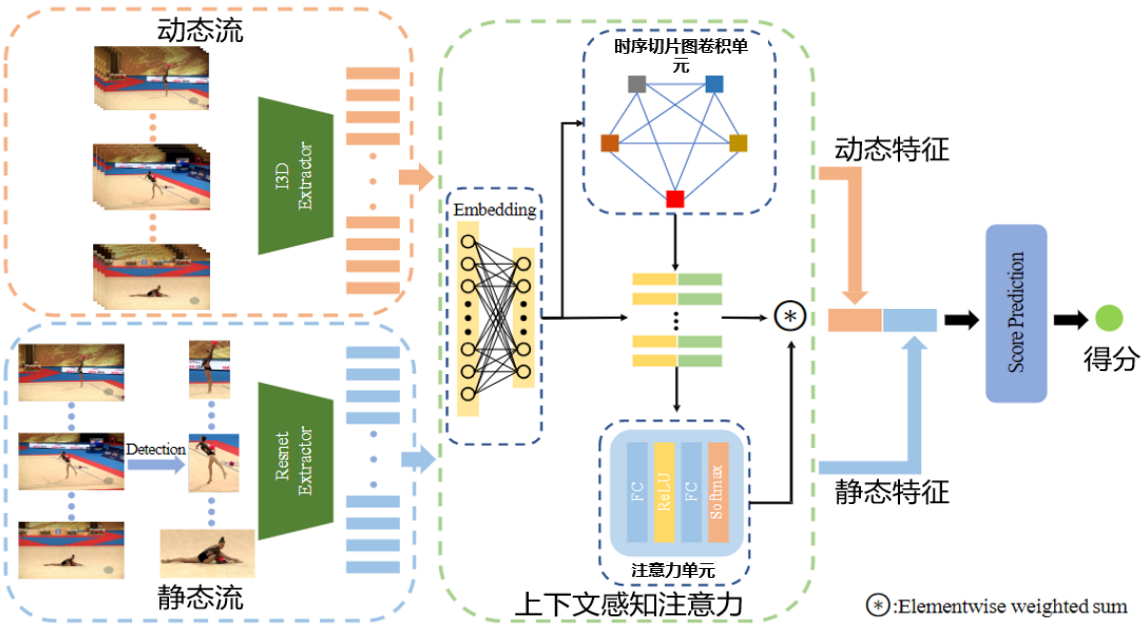

图 2. 6　Action-Net 框架

Fig. 2. 6　Framework of Action-Net

2.2.1　视频动静态特征提取

视频特征提取是视频理解领域非常重要的一个环节。目前仍然有大量的研究者在研究如何有效地提取视频的特征信息。作者在提取视频静态特征时，使用的是经过 Kinetics 数据集[53]预训练的 I3D 模型。I3D 模型是 Carreira 等人提出的用于视频行为识别的一种方法。他们认为经过深度学习多年以来的探索，已经开发了许多非常成功的图像分类网络结构，这些有效的结构大多是通过反复的试验和修改。因此作者提出，为了避免重复

这些试错的过程，作者建议简单地将成功的 2D 图像分类模型转换为 3D 卷积网络。目前已经有很多现成的优秀 2D 卷积神经网络模型（Inception[65]-[67]、VGG[68]、ResNet[60]等），在 2D 卷积神经网络模型的基础上，将二维的卷积核加上时间维度拓展成三维卷积核，从而将 2D 卷积神经网络拓展为适用于视频任务的 3D 卷积神经网络。与此同时，2D 卷积神经网络拥有在 ImageNet[69]上预训练好的模型参数，这些参数还可以拓展到 3D 卷积神经网络中。为了利用 2D 卷积网络的参数初始化 3D 卷积网络，作者采用了一种简单而有效的策略：将 2D 卷积核沿时间维度复制 N 次，形成一个 3D 卷积核，并对其进行归一化处理，即除以 N，以防止梯度过大。这种策略避免了从头训练 3D 网络的耗时过程，提高了研究效率。I3D 中使用的 2D 卷积神经网络为 Inception-V1，并在其每一层后引入批处理归一化和 ReLU 激活函数。对于 3D 卷积，时间维度步长的确定十分重要，它确定了时间尺度上感受野的大小。如果时间步长太大，那么则容易混淆不同物体的空间信息；如果时间步长太小，则无法捕捉到物体的运动信息，而捕捉运动信息对于行为识别来说是至关重要的。因此作者建议在开始的两次卷积过程中不要在时间尺度上进行卷积。通过这样的方法，作者提出的 I3D 网络相对于其他 3D 网络减少了很多计算量，并且网络能够更加快速地达到收敛。Action-Net 的作者利用从原始视频中采样的视频片段序列，将其输入到与训练后的 I3D 网络中提取 1024 维的动态特征。

视频片段可以产生动态信息，但稍微不正确的姿势在大量的视频帧中很容易被忽略，因为它们转瞬即逝。为了解决这一问题，作者还在动态流的基础上提取了静态流，其网络结构与动态流相同，以提供被检测运动员的补充姿态和外观信息。对于静态流，作者从原始视频中按照一定采样率对视频中的所有帧进行采样来提取姿态信息。由于运动视频中的运动员往往只占据图像的很小一部分，因此以整个图像为输入，除了计算量非常大，而且还很难提取出高质量的运动员空间特征。因此作者通过对视频每秒采样 1 帧，并使用在 COCO[70]上预训练的 YOLOv3[71]对采样帧中检测到的运动员的区域进行裁剪。为了过滤掉与动作无关的观众，作者只选择了每一帧中最大的人体边界框，并丢弃了所有没有检测到运动员的帧。之后，将裁剪后的图片送入到 ResNet 网络中进行特征提取，这样，就获得了高质量的静态特征。

2.2.2 视频动静态特征联合

为了将提取到的动静态特征有效地联合起来，作者提出了上下文感知注意模块，上下文感知注意模块由两个单元组成：首先作者引入了图神经网络[72]（Graph Convolutional

Network， GCN），作者将视频中的每个切片都认为是图的一个节点，图中的每个节点之间都有相互依赖关系。因此作者提出了建模片段/帧之间的关系的 TCG-U 和估计这些片段/帧的注意权重的注意力单元（Attention Unit ， ATT-U）。在获得每个实例的特征$\{f_I^i\}_{i=1}^N$后，作者使用GCN将所有实例特征聚合，输出每个实例的上下文相关特征$\{f_H^i\}_{i=1}^N$。为了迭代的学习所有实例之间的关系，作者构造了一个以所有实例为顶点的图 G。然后对图 G 采用指数核方式计算邻接矩阵$A \in R^{N\times N}$，如公式(2.3)所示：

$$A_{(i,j)} = exp^{\frac{-\left\|f_I^i - f_I^j\right\|}{K}} \tag{2.3}$$

其中K是一个正超参数，用于调整两个顶点之间距离的尺度，邻接矩阵A的元素$A_{(i,j)}$表示第 i 个和第 j 个实例之间的时序关系。在信息传递过程中，为了保持矩阵的原始分布，作者对邻接矩阵进行归一化。为了从由所有实例组成的图 G 中学习更多的上下文知识，作者使用多个 GCN 层迭代更新每个顶点的表示，以生成更健壮的上下文相关特征。在得到上下文相关的特征$\{f_H^i\}_{i=1}^N$后，作者将对应的实例特征进行组合，得到融合的局部上下文特征$\{f_C^i\}_{i=1}^N$，其中既包含实例信息，也包含全局上下文信息，更能代表视频中的每个实例，如公式(2.4)所示:

$$f_C^i = concatenate\left(f_I^i, f_R^i\right) \tag{2.4}$$

然后，作者利用简单的注意力单元进行注意力权重的估计，注意力单元主要包括两个全连接层，一个 ReLU 激活函数层和一个 Softmax 层，计算所有融合上下文特征的加权和，得到相应的动态特征和静态特征并进行连接。通过这种方式，作者利用视频中的所有实例构建融合的上下文特征，该特征不仅捕获相应的实例属性，还捕获上下文信息。最终并对输入视频的最终得分进行回归，如公式(2.5)所示：

$$s = f_r(f_D; f_S) \tag{2.5}$$

其中f_r表示两个具有 Sigmoid 激活函数的完全连接层。使用 Sigmoid 激活函数将估计分数归一化到[0,1]的范围内得到最终的得分。通过这样的方法作者从采样的视频帧中学习视频动态信息和特定的静态信息，并且通过利用所提出的时序切片 GCN 单元学习实例之间的关系来学习对于整个视频有效的表示来完成准确的动作质量评估。

2.3　本章小结

本章主要介绍了两种视频动作分析的算法，第一部分介绍了基于不同信息的视频动

作分割算法的：包括了基于时序卷积的动作分割网络和基于边界信息的动作分割网络，前者利用时序卷积来建模视频帧之间的关系，实现了视频中的动作分割；后者分析了现有方法存在的问题，并提出了一种引入边界信息来提高动作分割精度的解决方案。这些算法通过关注视频中的不同信息来进行准确的视频动作分割，是该领域最具代表性的方法。第二部分介绍了基于动静态结合的动作质量评估方法，这个方法除了关注传统视频分析中比较重要的动态信息，还把注意力放在了视频中的静态信息上，即视频中某一帧中人物姿态是否正确。该方法使用检测技术将静态信息提取出来，并且利用图神经网络中节点之间的相互关系，设计了一个时序切片 GCN 单元，结合注意力单元将动态信息和静态信息进行有效的上下文提取以及信息结合，从而实现更加准确地评估视频中人物动作的质量。

第三章　大规模手卫生评估数据集构建

随着深度学习的发展以及大型存储设备的诞生，越来越多的优秀数据集被研究者们创建，为人工智能的快速迭代提供了良好的基石。然而手卫生评估这一研究方向却缺少一个大规模的有挑战性的数据集，现有的关于手卫生的数据集[18][19][73]大多是在设计好的实验室场景拍摄的，无法反映现实场景中的问题，一个优秀的数据集应该尽可能包括更多的现实场景的挑战，这样才可以促使算法在训练过程中学会如何克服这些现实场景的挑战，更加准确地完成相应的任务。除此之外，这些数据集只提供了帧级别的步骤类别标注用于将视频帧图片分类，这样的标注也导致了这些数据集的使用场景受限，无法进行真正意义上的手卫生视频动作质量评估。因此为了促进手卫生评估的研究和发展，本文建立了一个统一的手卫生评估视频数据集HHA300。本章将从数据采集、数据处理、数据标注、数据分析等 4 个方面对 HHA300 手卫生评估视频数据集进行全面分析，阐述如何在专业人员的帮助下构建一个大规模高质量的手卫生评估基准数据集。

3.1　数据采集

HHA300 数据集采集的所有数据均为视频数据，视频内容为某一位志愿者在某一场景的洗手视频，采集设备使用的是万宝泽无线电池摄像头 M639，这种摄像头可以很容易地布置在洗手池周围以便采集数据。为了构建一个适合手卫生评估的视频数据集，本文采用了一种专业化的数据采集方法。本文邀请了多名医务人员参与数据拍摄，并在不同的场景下进行手卫生行为。部分数据样本如图 3.3 所示。在数据采集过程中，本文对拍摄场景进行了筛选和调整，以保证数据的质量和多样性。同时，为了模拟真实场景下的手卫生行为，本文将摄像头固定在洗手池上方进行录制。

为了在覆盖多种场景和人群，以反映手卫生数据集的多样性，本文在校园、医院、实验室等不同环境下录制了 80 余名志愿者的手卫生视频，其中包括 50 名医护专业人士和 30 名非医护人士。和现有的公开手卫生数据集相比，本文的数据集具有更高的场景和样本丰富度。此外，为了保持数据集中不同类别视频的均衡性，本文按照一定比例安排不同类型的人员参与视频录制，使得各类视频数量基本相等。

为了保证本文的数据集和真实场景更加接近，以增加数据集的真实性和难度。因此，本文在夜晚、白天、模糊等不同光照下录制了手卫生视频数据，并且在视频中设置了缺失步骤、错误步骤、角度变换等可能出现的场景挑战，要求志愿者按照预定规则执行手

卫生行为。这样做既可以丰富挑战类型，又可以避免数据采集风格过于单一。通过上述方式，本文共收集了 311 个手卫生视频片段，平均时长为 39s，最长时长为 61s。

图 3. 1　HHA300 中的部分数据样本

Fig. 3. 1　Data Samples from HHA300 Dataset

3.2　数据处理

通过前一阶段的数据采集，本文一共获取到 311 个手卫生视频序列，为了确保每个视频序列的质量和研究价值，本文过滤掉了一些有瑕疵的序列，例如分辨率过低、摄像头失焦等视频序列。经过过滤后本文确定了 300 个视频序列，并且将每个视频序列经过 FFmpeg 工具库的编解码处理，以 30 FPS 的速率解码为图片帧。本文共获得了 310000 张图片帧，平均每个视频序列有 1048 帧，对应约 35s 的时长。除此之外，为了提供更多的研究价值，本文还使用 Desenflow 方法对每个手卫生视频提取了光流数据。众所周知，光流可以反映某一时刻物体在成像平面上每个像素点运动的速度，是通过寻找对应像素点在相邻帧上的变化计算出相邻帧之间物体的运动信息的一种方法。因此，为了辅助完成视频分析任务，本文将 300 个视频序列的光流信息全部提取出来供研究者使用，这样可以更好地让该数据集在更多任务上的实现应用。

3.3　数据标注

动作分割和动作评估数据集需要高质量的帧类别标注和动作分数标注，这对于训练健壮的模型和确保性能评估的公平性是必不可少的。在获得了 300 个手卫生视频序列和超过 310000 张手卫生视频图片后，本文需要对这些手卫生视频以及图片进行详细的标

注，不同于其他手卫生数据集的标注方式，本文对 HHA300 数据集提供了两种形式的标注分别为细粒度的帧级别标注和粗粒度的视频分数标注。本文对数据集中的每个视频样本进行了手工标注，以获取手卫生行为的类别和质量评分。为了保证标注的准确性和一致性，本文邀请了两名具有手卫生知识的研究生作为标注人员，并由专业医务人员对他们进行了培训和指导。同时，本文制定了一套详细的标注规则，包括标注工具、标注格式、标注步骤等，以规范和统一标注过程。首先本文为手卫生评估任务提供了视频级别的标注，本文通过制定好的规则，例如：未完成某一动作扣除 0.5 分，某一动作不标准扣除 0.25 分等，对每一个采集到的手卫生视频进行打分，使得每一个视频都拥有一个评估分数。为了减少主观上的误差，在标注过程中，本文让两名标注人员标注同一视频，当两位标注人员给出的分数差的绝对值低于阈值时，最终的分数取两位标注人员给出分数的平均值，如果分数差绝对值高于阈值，则请求医务人员帮助，让医务人员选择更加合适的分数。此外，针对手卫生数据集在不同任务上的应用，除了视频级别的标注，为了能够实现步骤分割，本文对 HHA300 数据集进行了帧级别的类别标注，本文将视频中的每一帧与世界卫生组织提出的六步洗手法进行对应，将所有的帧分为 7 类，分别为：掌心相对，手指并拢相互揉搓；手心对手背沿指缝相互揉搓；掌心相对,手指交叉指缝相互揉搓；弯曲手指关节在掌心旋转揉搓；大拇指在掌心旋转揉搓；五指并拢,指尖在掌心旋转揉搓和背景动作。本文还对手卫生视频中每个步骤的属性和关键动作进行了定义，包含 3 种属性和 12 种关键动作，如表 3.1，3.2 所示。

表 3. 1　数据集的步骤属性定义

Table 3. 1　The step attribute definition criteria for the dataset

属性定义	描述
不存在(NE)	Not Exist－表明该步骤在该视频序列中不存在
存在但不标准(EN)	Exists but Nonstandard－表明该步骤在该视频序列中存在但动作不标准
存在且标准(ES)	Exists and Standard－表明该步骤在该视频序列中存在且动作标准

这些属性和关键动作是本文在医务人员的指导监督下定义的，并且在标注的时候，本文同样参考两位标注人员的意见，如果标注人员对属性的定义达成一致，则该属性被确定。如果标注人员对属性的定义存在分歧，则以医务人员的意见为准。最终本文总共获取到了 300 个标注文件。值得注意的是，本文的对于每一个视频序列的帧类别标注和分数标注都整合到了一起，最终，每个视频序列都会对应一个标注文件。除最后一行为评估分数外，其他均为该视频序列每一帧的类别序号。

表 3. 2　数据集的步骤关键动作定义

Table 3. 2　The key action of step definition criteria for the dataset

步骤	关键动作一	关键动作二
步骤一	手心相对	手指并拢
步骤二	手心对手背	手指交叉
步骤三	手心相对	手指交叉
步骤四	手指弯曲	掌心摩擦
步骤五	握住拇指	旋转摩擦
步骤六	掌心向上	指尖在掌心摩擦

3.4　数据分析

3.4.1　数据集规模

HHA300 是一个具有一定规模性的手卫生视频数据集，它包含 300 个手卫生行为视频序列，每个序列包含视频数据，帧数据以及光流数据，总帧数达到了 310000，其中最长的视频序列帧数达到了 1579 帧。之前的手卫生图片数据集例如，HWQA 数据集[19]总帧数为 8408，SIH 数据集[73]总帧数为 83000，AQA 数据集[18]总帧数为 309315。但是这些数据集大部分都是非现实场景拍摄的数据，绝大多数数据来源于实验室布置的场景。本文的 HHA300 手卫生数据集一共包含了 300 个视频序列，并将其分为了 225 个训练数据集合和 75 个测试数据集合。训练集中，视频序列的帧数范围为最少 375 帧到最多 1579 帧，平均为 1048 帧。测试集中，视频序列的总帧数为 77000 帧，平均每个序列为 1026 帧。本文的数据在医务人员的监督下录制，涵盖了多种真实场景和拍摄角度，能够反映现实世界中可能出现的各种情况。数据统计结果如表 3.3 所示。

众所周知，数据集中的类别不均衡问题也可能会导致模型性能较差，因此本文还对数据集中的除背景类外的六个类别进行了帧数数据统计，统计结果如图 3.2 所示，由图可以观察到，HHA300 的数据集中每个类别的数量都较为均衡，没有存在太大的差距，因此不存在类别不均衡问题。

3.4.2　数据集多样性

场景的复杂性和类型是增强数据集多样性的关键因素。为此，本文在 HHA300 数据集中从大量的拍摄人员、变换的摄像机角度、场景复杂性和其他环境因素中收集手卫生视频。为了阐明 HHA300 的优势，本文从以下几个方面分析它的多样性。

表 3.3　HHA300 数据集帧数统计分析

Table 3.3　Statistical analysis of the frame number in the HHA300 dataset

基准	视频数量	最小帧数	最大帧数	平均帧数	总帧数
HHA300(train)	225	373	1579	1048	236000
HHA300(test)	75	406	1436	1026	77000

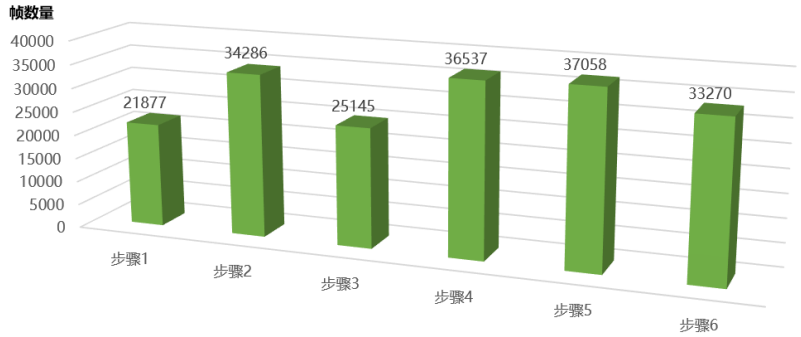

图 3.2　HHA300 的步骤帧数统计图

Fig. 3.2　Statistical diagram of steps frame in HHA300

首先，与现有的其他数据集不同，本文的 HHA300 数据集的拍摄场景非常丰富，包含有医务室、公厕、宿舍等。其次，本文还拥有大量的拍摄人员，包括几十名医务人员和学生。此外，为了更加接近真实的生产环境，本文采用的拍摄角度也不是固定的，本文的数据集在拍摄过程中共有六个不同的摄像机视点，如顶部、左侧和右侧。这些环境因素具有不同的复杂性，因此给步骤分割和手卫生评估带来一些困难，对于算法的要求也更高。除了以上外部因素带来的挑战，本文也在考虑内部因素带来的挑战。在手卫生评估任务中，每个步骤的动作质量对评估模型的结果有一定的影响。因此，本文在数据采集时定义了一系列规则用来模拟更真实的情况。表 3.4 显示了训练集中所有视频序列中步骤属性的视频分布。

表 3.4　HHA300 训练集中步骤属性的分布

Table 3.4　Distribution on Different Attributes in HHA300 Training Set

步骤 / 属性	步骤一	步骤二	步骤三	步骤四	步骤五	步骤六
不存在(NE)	37	46	48	49	42	53
存在但不标准(EN)	27	25	32	34	21	38
存在且标准(ES)	161	154	145	142	162	134

本文要求拍摄人员在进行手卫生行为时进行不同规范程度的动作，以考虑每个步骤中不同程度的标准化。这样，本文的 HHA300 包含了几乎所有可能的现实世界挑战情况。

3.4.3 分析讨论

本节主要讨论和分析与先前工作以及任务的不同之处。本文提出的数据集 HHA300 可用于两个任务，分别为动作分割和手卫生动作质量评估。在动作分割方面，目前现有的数据集为 50Salads[76]：50Salads 数据集包含 50 个视频，正如数据集的名称所示，这些视频描述了做沙拉时的动作，视角主要为俯视视角；GTEA[77]：GTEA 数据集包含 28 个视频，对应于 7 种不同的活动，如准备咖啡或奶酪三明治等，并由 4 名志愿者拍摄，这些视频的帧被标注为包括背景在内的 11 个动作类别，视角主要为第一人称视角；Breakfast[78]： Breakfast 数据集是三个数据集中最大的，有 1712 个视频，主要展示了早餐准备的相关活动，该数据集视角为第三人称视角。HHA300 与上述生活领域的数据集对比可以发现，HHA300 将医疗卫生领域的手卫生行为引入到了动作分割领域中，并且拥有一定的规模。同时，HHA300 也拥有多视角和多场景的优点，更加具有挑战性且接近现实场景。在手卫生方面，已有的数据集例如 HWQA，只拥有 8408 帧图片，并且这些视角都是同一个角度拍摄的；SIH 数据集拥有 83000 帧图像，这些图片是使用深度红外相机拍摄的，但也只有一个角度。目前已有的数据集只提供了帧类别的标签用于完成图像分类任务，无法完成真正意义上的手卫生质量评估任务。与这些数据集相比，HHA300 数据集拥有更大的规模，更丰富的场景以及视角，最重要的是 HHA300 数据集除了拥有逐帧的类别标签外，还拥有在医务人员指导下标注的质量分数用于手卫生质量评估，这是以往数据集无法提供的。因此 HHA300 手卫生数据集拥有更加广泛的用途和研究价值，可以作为该领域的基准数据集。

3.5 本章小结

本文在本章详细介绍了 HHA300 数据集的采集、处理和标注过程，以及数据的来源、规模和属性。本章还分析了 HHA300 数据集与其他动作分割和手卫生数据集的差异和特点，以及数据集的应用价值。本文认为，HHA300 数据集将促进手卫生行为视频理解的研究，如手卫生动作分割和评估等。在第四章，本文将针对手卫生评估任务提出一个基准算法，并进行实验分析。

第四章　基于步骤分割和关键动作打分的手卫生评估算法

4.1　引言

近年来，旨在评价动作质量的动作评定引起了广泛关注。它在现实世界的许多领域都有重要的应用，比如体育和医疗。但是以往的动作评估模型[2][9][10][54][56]大多直接对一个视频进行评估，输出一个评估分数，这些方法会忽略一个长动作中的许多细节，而一个长动作通常涉及几个较短的步骤，这些模型无法关注到这些较短的步骤，因此降低了长动作评估的性能。世界卫生组织规定的手卫生是一个标准的长动作，其中包括有六个步骤。因此，手卫生评估任务中有两个主要问题，首先每个手卫生视频中都包含六个步骤中的一部分或者全部，现有的动作评估模型无法知道哪些帧属于哪个步骤，这给准确评估整个手卫生视频带来了很大的挑战。第二，在每个步骤中，有几个关键的动作决定步骤的完成质量，并不是所有的动作对手卫生评估都是有用的。为了解决这两个问题，本文提出了一个细粒度的学习框架，该框架基于一种联合步骤分割和关键动作打分的方法，用于鲁棒的手卫生评估。现有的评估短动作(如跳水)的方法大多是整体预测，不能准确评价长动作，因为长动作的质量取决于每一步的完成质量。不同于现有动作评估方法对整个视频的整体评估，本文将输入视频划分为基于步骤的视频片段进行细粒度评分。为此，本文设计了一个多阶段卷积-Transformer 网络，用于手卫生视频中的精确步骤分割。手卫生视频中的动作具有连续性和相似性，现有的多阶段卷积模型由于缺乏长距离帧之间的相关性，容易导致过分割错误。为了解决这个问题，本文在多阶段卷积网络中嵌入了线性 Transformer[74]，形成了多阶段卷积-Transformer 网络，在不增加太多计算量的情况下，建立了有效的长距离帧之间的相关性。准确评估每个步骤是手卫生评估中最关键的问题。本文观察到，手卫生中的每一步都涉及两个关键动作，这两个动作决定了这一步的质量。因此，本文提出了一个基于可学习 Sigmoid 的关键动作打分器，每个关键动作打分器对应一个步骤，内部由两个结构相同的分支组成对应两个关键动作。具体来说，每个分支包括全连接层和可学习的 Sigmoid 层，并且不同的分支具有独立的参数来对不同关键动作的特征进行建模。本文在提出的 HHA300 数据集以及传统的动作分割数据集上进行了充分的实验，证明了方法的可行性。

4.2 模型介绍

为了完成鲁棒准确的手卫生动作评估任务，本文设计了基于步骤分割和关键动作打分器的手卫生评估方法，本节将详细阐述该方法的实现细节。

4.2.1 网络结构概述

基于步骤分割和关键动作打分器的手卫生评估模型如图 4.1 所示。整个模型由两个主要模块组成，分别是基于多阶段卷积-Transformer 的步骤分割模块、基于关键动作打分器的手卫生评估模块。首先，将视频数据与光流数据同时输入网络。使用 I3D 特征提取器[53]将 RGB 和光流信息作为输入来计算相应的外观和运动特征。然后，这两个特征被连接并输入到多阶段卷积-Transformer 网络中。在网络中，每个阶段产生相应的特征，本文使用卷积来建立帧间的短距离相关性，使用线性 Transformer 来建立长距离相关性，从而增强这些特征。然后，本文获得与每个步骤相对应的时间片段，并通过关键动作打分器分别预测质量分数来评估这些时间片段。最后，本文将每一步的分数相加，得出最终的评估分数。

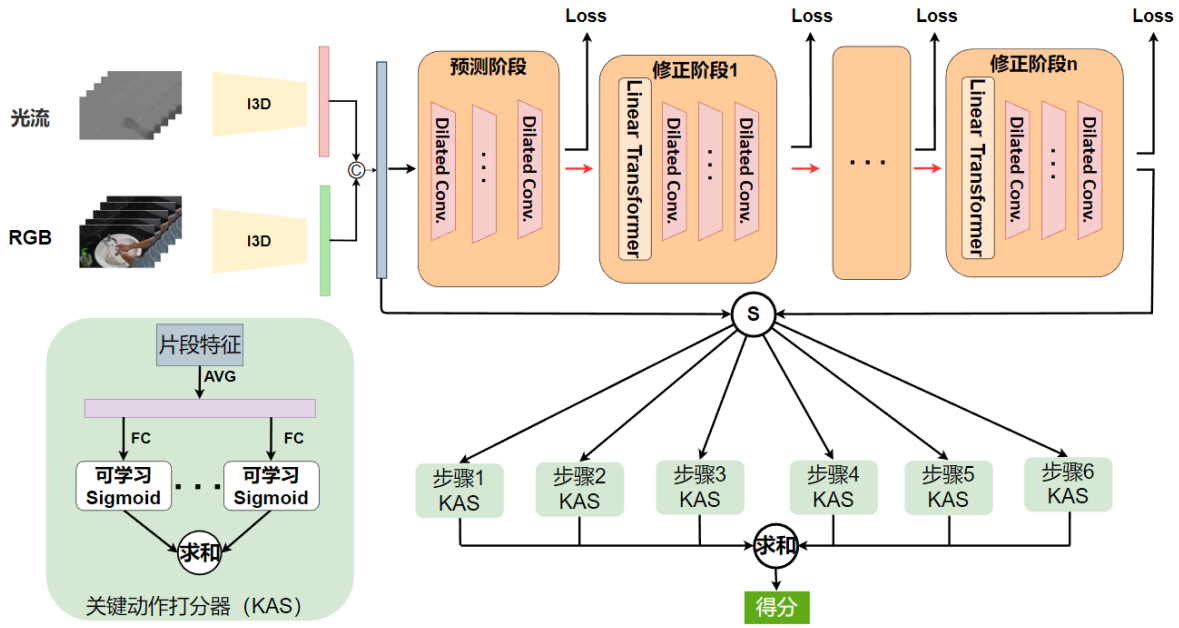

图 4. 1　基于步骤分割和关键动作打分器的手卫生评估网络

Fig. 4. 1　Hand hygiene assessment network based on step segmentation and key action scorer

4.2.2 基于多阶段卷积-Transformer 的步骤分割

现有的大部分方法评估的是跳水之类的短视频，其输入是包含动作的一整段视频，输出为该动作的评分。具体来说，给定输入视频特征$\Phi = \{\varphi(x_1), \varphi(x_2), \cdots, \varphi(x_T)\} \in R^{T\times D}$，这些方法把动作评估任务表述为一个回归任务，如公式(4.1)所示，即预测出动作

的得分S：

$$S = F(\Phi) \tag{4.1}$$

其中F是回归模型。然而，这些方法不能对长动作进行准确的评估，因为长动作的质量取决于每个步骤的完成质量。因此，本文建议在评估动作质量之前对步骤进行分割，这样就可以根据每一步骤的完成质量来确定整体动作的完成质量，实现更加可解释的准确的动作质量评估。目前，研究者通常使用动作分割方法获取长动作中每一步骤。现有的动作分割方法[31]-[34]大多采用多阶段卷积模型。这些方法在第一阶段预测出一个粗略的分割结果，前一阶段的结果在随后的每个阶段被逐渐细化和修正。然而，手卫生视频中的动作是连续且相似的，多阶段卷积模型中只存在短距离的帧间相关性，无法准确地区分这些连续且相似的动作，很容易导致分割错误(即过分割)。

为了获得精确的步骤分割结果，本文提出在多阶段卷积模型中嵌入线性 Transformer 来形成多阶段卷积-Transformer 网络，与多阶段卷积模型相比，该网络可以在不增加太多计算量的情况下建模长距离帧之间的相关性。Transformer[40]首先应用于自然语言处理的研究，内部结构如图 4.2 所示，左侧为编码块，右侧为解码块。它使用自注意力机制代替 RNN 的序列结构，从而使模型可以并行训练并获得全局信息。由于其强大的建模能力，目前已经广泛应用在计算机视觉领域中。

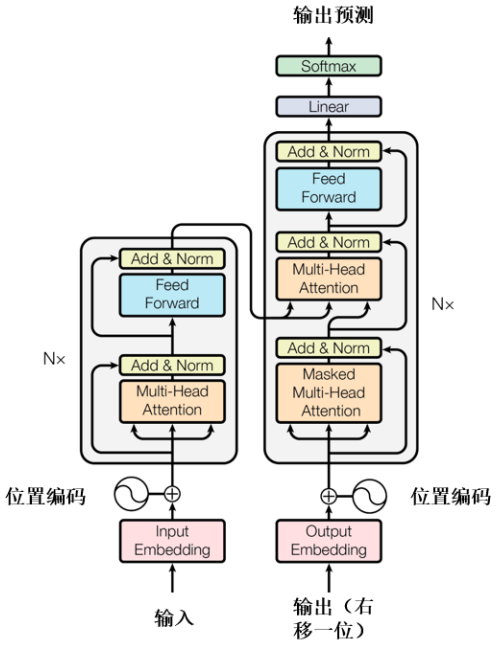

图 4. 2　Transformer 的编解码架构

Fig. 4. 2　The Encoder-Decoder architecture of Transformer

多头注意力是 Transformer 结构的核心，它能够实现全局信息的建模，由多个自注

意力组成，自注意力机制的输入向量通常被命名为查询（Query）、键（Key）和值（Value）。计算方式如公式(4.2)，(4.3)，(4.4)所示：

$$Q = xW_Q \tag{4.2}$$

$$K = xW_K \tag{4.3}$$

$$V = xW_V \tag{4.4}$$

其中，x为输入特征，W_Q，W_K，W_V为三个可学习的矩阵。值向量的权重分布由查询向量和键向量之间的相似度来确定。形式上，注意层表示如公式(4.5)所示:

$$V' = soft\,max(\frac{QK^T}{\sqrt{d_k}})V \tag{4.5}$$

其中Q，K和V分别表示查询、键和值，而d_k是向量维度。自注意力机制通过这样的计算方法可以捕获长距离的依赖关系，因此为了降低多阶段卷积网络中由于缺乏长距离帧间相关性而导致的过分割风险，本文引入了 Transformer 来进一步建模长距离相关性。具体来说，本文设计了一种基于多阶段卷积-Transformer 网络的步骤分割，每一阶段包括若干个具有不同核大小和步长的空洞卷积和普通卷积以及 Transformer，以建立短距离和长距离帧间的相关性。如图 4.1 上半部分所示，本文首先使用 I3D 骨干网络提取视频 RGB 和光流信息的特征，并将两种特征沿通道维度拼接起来得到初始特征。之后本文参考了 MS-TCN++[33]的做法，对于预测阶段和后续修正阶段使用了不同的网络结构。初始特征在预测阶段中生成初始的预测结果和阶段特征，然后进行进一步的修正。预测阶段的详细结构如图 4.3 所示。

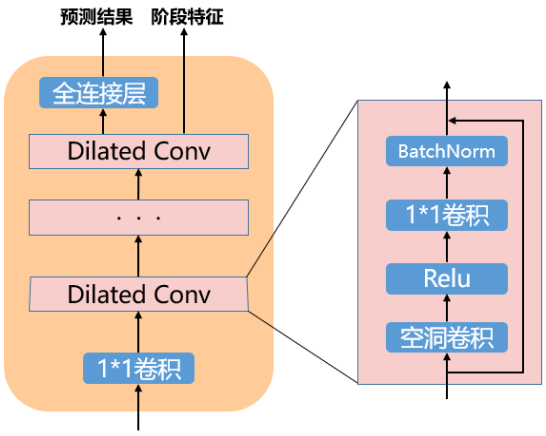

图 4. 3　预测阶段网络结构

Fig. 4. 3　Network structure of prediction stage

预测阶段包括一个 1×1 卷积，m 个空洞卷积块（Dilated Conv），一个全连接层。这

个 m 为超参数，可以根据需要自行设置。首先，初始特征会经过一个 1×1 卷积进行特征的降维方便后续计算，然后会经过若干个空洞卷积块，空洞卷积块内部包含了 1×1 卷积，空洞卷积，ReLU 激活层，和 BatchNorm 归一化层以及残差连接，这些结构保证了短距离帧之间的相关性得以建模。除此之外，m 个空洞卷积块各不相同，每一层的空洞率都是前一层的 2 倍，这样设计可以尽可能地扩大模型的感受野。最后，视频特征首先会输入到全连接层预测出一个粗略的动作分割结果，然后这些特征会被送入后续的修正阶段中。在一系列的修正阶段中，每个阶段中使用前一阶段的预测结果和阶段特征来计算细化后的预测结果和阶段特征。修正阶段的网络结构与预测阶段基本相同，但是即便是使用了空洞卷积，长距离的帧之间仍然缺乏相关性，因此本文在预测阶段的基础上加入了 Transformer 结构来作为修正阶段，具体来说就是使用阶段特征作为键向量、值向量和查询向量，并通过自注意力机制来增强前一阶段的特征，以建立所有视频帧之间的长距离相关性，但是原始的自注意力机制是一种特殊形式的自注意力，称为 Softmax Attention，相似度是Q和K之间的矩阵乘法计算的，随着序列长度T的增加，矩阵乘法计算代价也变得极其昂贵，这会大大影响算法的训练和实际使用的速度。因此，本文引入了一种 Transformer 的变体，即线性 Transformer，它不像原始的 Transformer 一样使用矩阵乘法，而是使用另一种核函数来度量相似度来减少 Transformer 的计算量，并且具有相似的性能。具体来说，假设矩阵下标为i，将返回矩阵V'的第i行作为一个向量，可以针对任意相似函数$sim(\cdot)$写出广义自注意力方程，如公式(4.6)所示:

$$V' = \frac{\sum_{j=1}^{T} sim(Q_i, K_j)V_j}{\sum_{j=1}^{T} sim(Q_i, K_j)} \tag{4.5}$$

这里的$sim(Q_i, K_j)$表示查询向量和键向量的相似度，并且$sim(\cdot)$只需要非负即可。为了完成这个要求，可以利用任意值域非负的激活函数$\theta(\cdot)$，当针对Q_i, K_j的激活函数相同时，就相当于引入了一个核函数。如公式(4.3)所示：

$$sim(Q_i, K_j) = \theta(Q_i)^T \cdot \theta(K_j) \tag{4.6}$$

经过这样的变换，公式(4.5)就可以改写为公式(4.7)：

$$V' = \frac{\sum_{j=1}^{T} \theta(Q_i)^T \cdot \theta(K_j)V_j}{\sum_{j=1}^{T} \theta(Q_i)^T \cdot \theta(K_j)} \tag{4.7}$$

然后可以利用矩阵乘法结合律继续将公式(4.7)简化为公式(4.8)：

$$V' = \frac{\theta(Q_i)^T \sum_{j=1}^{T} \theta(K_j) V_j^T}{\theta(Q_i)^T \sum_{j=1}^{T} \theta(K_j)} \tag{4.8}$$

由公式(4.5)可知 Softmax Attention 的计算量为$O(T^2)$，相比之下，由公式(4.8)提出的线性 Transformer 的计算量近似为$O(T)$，因为可以只计算一次$\sum_{j=1}^{T} \theta(K_j) V_j^T$和$\sum_{j=1}^{T} \theta(K_j)$，在每次查询时复用它们。本文参考线性 Transformer 定义$\theta(x) = ELU(x) + 1$，$ELU()$函数[75]是$ReLU()$函数的改进版本，避免了在x为负时将梯度设置为 0。本文利用上面的变换代替原始 Transformer，可以大大的减少计算量。此外，多头注意力机制不包含位置信息，因此，Transformer 向编码器部分和解码器部分的输入添加了名为位置编码的额外矢量。但是本文认为，因为视频帧本身具有时序信息，所以 Transformer 中的位置编码将干扰时序信息，因此在本文的方法中不使用位置编码。本文将线性 Transformer 加入到除了预测阶段以外的每一个修正阶段中，在保留原始特征信息的基础上利用 Transformer 强大的全局建模能力来增强每一阶段特征的长距离相关性，从而解决过分割问题。综上，本文提出的基于多阶段卷积-Transformer 的步骤分割方法具体流程为，首先通过 I3D 特征提取网络提取视频特征和光流特征，然后将两种特征连接起来，输入到预测阶段的纯卷积网络中，输出一个初始的步骤分割结果和经过卷积计算后的特征并将其输入下一阶段。后续的每一修正阶段都在预测阶段的基础上加入了线性 Transformer，在建模全局信息的同时，修正上一阶段的预测结果，最后，通过最后一个阶段得到步骤分割的最终结果。所有的预测结果都会利用真值进行监督，以确保每一阶段都能学习到最合适的参数。

4.2.3 基于可学习 Sigmoid 的关键动作打分器

在得到步骤分割的结果之后，对于每一个步骤，本文选择连续且最长的部分作为代表段用于后续的评估。这样做主要出于两方面考虑，一方面，由于可能存在的错误分割，一个步骤可能在视频序列中出现多次分散在不同的时间点零散帧，因此本文选择连续的帧来过滤掉一些分割错误的帧。另一方面，在训练了步骤分割模型一段时间后之后，即使出现一些错误分割，最长的部分仍然是正确分割的。因此，本文使用这些连续且最长的部分作为手卫生评估模块的输入。在获得每个步骤片段后，准确评估每个步骤是最关键的问题。本文在数据集制作的过程中观察到每一步都涉及两个关键动作，它们决定了这一步的质量。针对这一特性，为了能够准确评估手部卫生的每个步骤，本文设计了一个关键行动打分器(Key action scorer，KAS)来评估每个步骤中的关键动作，如图 4.1 下半部分所示，每个步骤片段会输入到对应的关键行动打分器，以计算相应的评估分数。

因此，准确评估这些关键行动的质量是关键动作打分器的目标。经过本文的研究，本文发现，大多数动作评估任务都采用 Sigmoid 函数作为最终的激活函数预测得分，它是一种非常普遍的激活函数，可以将向量映射到 0 到 1 之间，并且应用于各种不同的任务中。因此，本文也参考前人的设计，使用 Sigmoid 作为最后映射得分的函数。但是，由于不同的关键动作之间存在一些差异，如果使用统一的 Sigmoid 激活函数则无法应对这样的任务。因此，对于每一个关键动作，都应该使用一个特定的结构来评估它，本文发现 Sigmoid 函数可以通过引入一个可学习的参数来控制陡度，具有可学习参数的 Sigmoid 公式如下公式(4.4)所示:

$$LS = \frac{1}{1 + e^{-\lambda x}} \tag{4.9}$$

其中x可学习 Sigmoid 函数的输入，λ是控制 Sigmoid 函数陡度的可学习参数，原始的 Sigmoid 函数的λ为 1。本文基于上述可学习的 Sigmoid 函数的特点，设计了一个全新的手卫生评估模块，包括六个基于可学习 Sigmoid 的关键动作打分器。每个关键动作打分器对应一个手卫生步骤，由两个结构相同的分支组成，每个分支包括若全连接层和可学习的 Sigmoid 函数层，不同的分支具有独立的参数来模拟不同关键动作的特征，并使用可学习的 Sigmoid 来更准确地评估不同的关键动作。如图 4.1 左下角所示，本文首先对步骤特征进行全局平均池化，然后将它们输入到由全连接层和可学习 Sigmoid 组成的不同关键动作评估分支中，关键行动评估分支的数量由每个步骤中关键行动的数量决定，在本文的工作中，本文设置分支的数量为两个，如表 3.2 中定义所示。本文希望每个打分器及其分支都能够学习一组特有参数，以准确评价对应的关键动作，具体的计算方式如公式(4.5)所示:

$$s_i = \frac{1}{k}\sum_{j=1}^{k} LS_j\left(FC_j\left(AVG(\varphi_i)\right)\right) \tag{4.10}$$

其中s_i，φ_i分别是第i个手卫生步骤的分数和对应的步骤特征。k是关键动作的数量。LS、FC和AVG分别表示可学习的 Sigmoid、全连接层和全局平均池化层的操作。每一个关键动作打分器都会得到一个步骤得分，最终总分是所有步骤得分的总和，如公式(4.6)所示:

$$S = \sum_{i=1}^{6} s_i \tag{4.11}$$

这里的S代表手卫生视频的总得分。

4.3　学习算法

本文的损失函数由两部分组成。第一部分是步骤分割模型中的损失，在这一部分，本文参考了 MS-TCN[32]提出的损失函数，它包括逐帧分类的交叉熵损失和截尾均方误差的对数概率平滑，交叉熵损失如公式(4.7)所示：

$$L_{cls} = \frac{1}{T}\sum_{t} - log(y_{t,c}) \tag{4.12}$$

其中，$y_{t,c}$是t时刻真值标签的预测概率。为了抑制过度分割的现象，提高预测结果的连续性和准确性，本文还使用了一种平滑损失函数，该损失函数基于截断均方误差和帧级对数概率之比进行计算，如公式(4.8)，(4.9)，(4.10)所示：

$$L_{T-MSE} = \frac{1}{TC}\sum_{t,c} \tilde{\Delta}_{t,c}^2 \tag{4.13}$$

$$\tilde{\Delta}_{t,c} = \begin{cases} \Delta_{t,c}: \Delta_{t,c} < \tau \\ \tau: otherwise \end{cases} \tag{4.14}$$

$$\Delta_{t,c} = \left| log\, y_{t,c} - log\, y_{t-1,c} \right| \tag{4.15}$$

其中T是视频长度，C是类的数量，$y_{t,c}$是t时刻c类的概率。最终的损失函数是上述损失的组合，模型的总损失为每一个阶段预测结果与真值的损失总和，如公式(4.11)，(4.12)所示：

$$L_s = L_{cls} + \gamma L_{T-MSE} \tag{4.16}$$

$$L_{SEG} = \sum_{s} L_s \tag{4.17}$$

其中γ是模型的超参数，用于确定不同损失权重。第二部分损失函数是手卫生评估任务的损失，本文使用均方误差来测量预测得分和实际得分之间的差异，如公式(4.13)所示:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{4.18}$$

其中$\hat{y}$和y分别表示预测得分和实际得分。最终损失函数是上述损失的组合，如公式(4.14)所示：

$$L = L_{\text{SEG}} + L_{\text{MSE}} \tag{4.19}$$

4.4 实验与分析

本小节将介绍模型训练所需数据集和训练细节，并系统地对提出的基于步骤分割和

关键动作打分的手卫生评估模型进行系统性实验分析。在 4.4.1 小节介绍了进行实验相关的数据集与评价指标；在 4.4.2 小节详细描述了相关的实验设置；在 4.4.3 详细地列出了实验结果并对所得数据进行了分析；在 4.4.4 小节本文对模型的进行了消融实验，验证了所提出的各个模块的有效性。

4.4.1　数据集与评价指标

本文在提出的 HHA300 手卫生数据集上评估了所提出的基于步骤分割和关键动作打分器的手卫生评估方法，并且为了验证基于多阶段卷积-Transformer 的步骤分割方法，本文还在三个公共动作分割数据集上进行了实验，这三个数据集分别为：50Salads[76] ：包含 50 个视频，描述了做沙拉时的 17 种动作类别。本文参考之前的工作[32]采用五重交叉验证法进行评估，并报告平均结果。GTEA[77] ：包含 28 个视频，展示了 7 种不同活动（如准备咖啡或奶酪三明治），每个视频平均有 20 个动作实例，由 4 名志愿者拍摄，视频帧标注为 11 个动作类别（包括背景），本文使用交叉验证法进行评估，并报告平均结果。Breakfast[78]数据集：包含 1712 个视频，在 18 个不同厨房录制，主要涉及早餐准备相关活动。本文使用了 Kuehne[26] 中提出的标准 4 分割法并报告了平均结果。对于所有数据集，本文提取视频帧的 I3D 特征，并使用这些特征作为动作分割模块的输入。

在评价指标方面，对于步骤分割的评估，本文采用逐帧准确率 Acc、分段编辑距 Edit 和重叠阈值 10%、25%和 50%的分段 F1 分数，用 F1@{10，25，50}表示。众所周知，逐帧准确率是动作分割最常见的评价指标之一。但是由于逐帧准确率与帧数有关，长动作类由于帧数较多对该度量的影响大于短动作类，使得该度量无法反映过分割的误差。为此，Lea 等人[29]引入了两个新的度量，包括分段编辑距离和分段 F1 分数，这两种度量可以有效的反应模型出现的过分割错误。具体来说，编辑距离是用来衡量预测的结果序列和真值序列之间的差异程度，通过动作的发生的先后顺序来反映过分割错误，而 F1 分数类似于检测任务中的 IoU，用来计算预测结果和真值结果的重叠程度。对于手卫生评估的评价，参考现有的方法[54][55]，本文使用一种称为 Spearman 秩相关系数ρ的标准评价度量。它的计算公式如下：

$$\rho = \frac{\sum_i (p_i - \overline{p})(q_i - \overline{q})}{\sqrt{\sum_i (p_i - \overline{p})(q_i - \overline{q})}} \tag{4.20}$$

其中p和q分别代表两个预测的分数序列和真值分数序列中的每个样本的排名。例如在一系列预测结果中选取一批样本，根据这批样本的预测分数和真值分数分别进行排名，这

批样本的预测排名和真值排名越接近，那么 Spearman 秩相关系数就越大。本文还使用相对 L2 距离$(R-l_2)$[55]来作为度量，其定义为:

$$R-l_2=\frac{1}{K}\sum_{k=1}^{K}\left(\frac{|s_k-\hat{s}_k|}{s_{\max}-s_{s\min}}\right)^2 \tag{4.21}$$

其中s_{max}和s_{min}是该动作的最高分和最低分，s_k和$\hat{s}_k$分别表示第k个样本的真值分数和预测分数。相对 L2 距离反映了预测序列和真值序列的数值差距，差距越小，相对 L2 距离越小。Spearman 秩相关系数更侧重于预测分数的排名，取值范围为[0,1]，而$R-l_2$则侧重于预测分数的数值，取值范围为[0,+∞]。

4.4.2 实验设置

本章所有实验采用的软硬件信息如表 4.1 所示。本文的模型选择 Adam 算法作为模型优化器，共计训练 50 个 epoch，在前 30 个 epoch 学习率设置为 0.001，后 20 个 epoch 设置为 0.0005。

为了便于比较，本文在 HHA300 数据集上评估了一些步骤分割模型。此外，按照 Xu 等人[9]给出的对比设置，本文还对一些动作评估方法进行了评估。本文考虑以下模型组件的不同组合。

表 4.1 计算机软件和硬件环境

Table 4.1 Computer software and hardware environments

项目	参数
CPU	Intel i7-6700K 4.0GHz
内存	16G
显卡	Nvidia RTX1080Ti
操作系统	Ubuntu 18.04 LTS
Python 版本	3.7
PyTorch 版本	1.1
CUDA	10.1
CUDNN	7.6.5

1.特征提取器：本文使用了三种预训练过的特征提取器，分别是 I3D，C3D[51]，ResNet[60]。I3D 和 C3D 是视频理解领域常见的特征提取器，I3D 在第二章已有详细介绍，

C3D 是作为一个通用网络被提出的，主要将其用于行为识别等领域。而 ResNet 是非常经典的图像特征提取器，它首次提出了残差结构，能够有效地处理深层神经网络中的退化问题。对于 I3D 和 C3D，本文使用这两个模型分别提取 RGB 特征和光流特征并结合起来输入到后续组件中。对于 ResNet，本文采用 ResNet50 处理 RGB 图像和光流信息，然后进行平均池化以获得相同维度的特征，最后同样将这两种特征沿通道维度拼接进行后续操作。

2.多层感知机（Multi-Layer Perception，MLP）：MLP 是一种前馈神经网络，广泛应用于深度学习的任务中。本文在提取特征后使用平均或最大池化进行视频的特征描述，然后通过含有两个隐藏层的 MLP 预测得分，损失函数方面本文使用预测结果和真值标签之间的均方误差进行优化。

3.长短期记忆网络[64]（Long short-term memory，LSTM）：LSTM 是一种改进的循环神经网络，它通过引入三个门结构（输入门、遗忘门和输出门），能够有效地控制信息的存储和遗忘，避免了传统循环神经网络在处理长序列任务时遇到的梯度消失和梯度爆炸等问题。与 Parmar 等人[10]类似，本文通过 LSTM 架构生成视频级描述。在本文的设置中，LSTM 隐藏层的维度为 256，同时还包括一个全连接的回归层。

除了结合特征提取器，MLP 和 LSTM 这些组件之外，为了全面的比较本文提出的方法与现有的方法，本文还结合将先进的动作分割模型与 MLP 和 LSTM 等相结合，例如 MS-TCN -MLP，MS-TCN -LSTM。

4.4.3　实验结果与分析

本文在 HHA300 手卫生数据集上针对一些方法评估了本文的所提出的框架，如表 4.2 所示。表中 Avg 代表平均池化操作，Max 代表最大池化操作，MLP 指的是具有两个隐藏层的多层感知器，LSTM 是长短期记忆层。从结果可以看出，对于不同特征提取器，MLP 和 LSTM 的组合，基于 LSTM 的两种方法优于基于 MLP 的其他方法。LSTM 可以通过增强长距离帧之间的相关性来实现更好的特征描述。然而这些方法的实验结果都不理想，表中可以看到这些方法的 Spearman 秩相关系数都很小，相对 L2 距离很大。主要原因是这些方法将视频的所有特征作为评估模型的输入来回归评估分数，而没有考虑细粒度的手卫生动作例如不同的步骤完成质量如何以及无法过滤掉不相关的背景动作。对于基于动作分割模型和 MLP 的方法，评价结果优于前面几种方法。这主要是因为动作分割模型可以将手卫生划分为不同的步骤以进行细粒度的评估。在这些方法中，除了本

文提出的方法外，BCN[34]-MLP 的效果最好，这主要是因为具有边界感知级联的模型能进行更精确的步骤分割。因此可以发现，步骤分割的性能对于手卫生评估至关重要。从结果中可以看到本文提出的基于步骤分割和关键动作打分器的手卫生评估方法在这两个指标上明显优于其他方法。具体来说，本文的方法实现了 0.822 的 Spearman 秩相关系数和 0.96 的相对 L2 距离。与其他方法相比，本文的框架包括用于步骤分割的阶段卷积-Transformer 模型和基于关键动作打分的手卫生评估模型用来评估每个步骤的完成质量。它不仅将手卫生视频中的长动作分割成基于步骤的片段，而且基于每个步骤的关键动作做出准确的评估。

表 4. 2　在 HHA300 数据集上进行步骤分割和手卫生评估的结果

Table 4. 2　Results of step segmentation and hand hygiene assessment on HHA300 dataset.

方法	F1{10，25，50}↑			Edit↑	Acc↑	ρ↑	$R-l_2$ (×100)↓
ResNet50-Avg-MLP	-			-	-	0.245	39.92
ResNet50-Max-MLP	-			-	-	0.281	37.22
ResNet50-LSTM	-			-	-	0.311	36.97
C3D-Avg-MLP	-			-	-	0.274	37.97
C3D-Max-MLP	-			-	-	0.286	38.54
C3D-LSTM	-			-	-	0.350	36.81
I3D-Avg-MLP	-			-	-	0.378	36.50
I3D-Max-MLP	-			-	-	0.389	36.13
I3D-LSTM	-			-	-	0.406	34.60
MS-TCN[32]-MLP	82.0	81.7	75.6	74.6	88.7	0.704	2.52
MS-TCN++[33]-MLP	83.3	83.3	75.9	77.9	89.0	0.711	2.01
ASFR[79]-MLP	88.2	88.7	82.4	82.0	**89.9**	0.728	1.80
BCN[34]-MLP	87.4	87.1	81.1	81.3	89.1	0.774	1.58
Ours	**89.7**	**89.2**	**83.0**	**83.3**	89.1	**0.852**	**1.07**

对于基于多阶段卷积-Transformer 的步骤分割，本文在 HHA300 数据集上对其进行了评估，并将其与四个多阶段网络进行了比较。从结果可以看出，本文的方法在三个指标上优于其他方法。相比于 BCN，虽然本文在准确率上与其相同，但在片段编辑距离上本文高 2.0%，在分段 F1 分数上在阈值为 10%，25%，50%的情况下高 2.3%，2.1%，1.9%。由于准确率与类别的帧数有关，这使得该度量不能反映过分割的误差，而分段编

辑距离和分段 F1 分数可以很好地反映模型减少过分割的能力。因此，实验结果表明本文的多级卷积-Transformer 模型与其他多级卷积模型相比，能够在保证精度的同时有效缓解过分割现象。HHA300 数据集的定性结果如图 4.4 所示，可以发现 BCN 模型的预测有一些过分割误差，但本文可以通过多阶段卷积-Transformer 网络来减少这些误差。

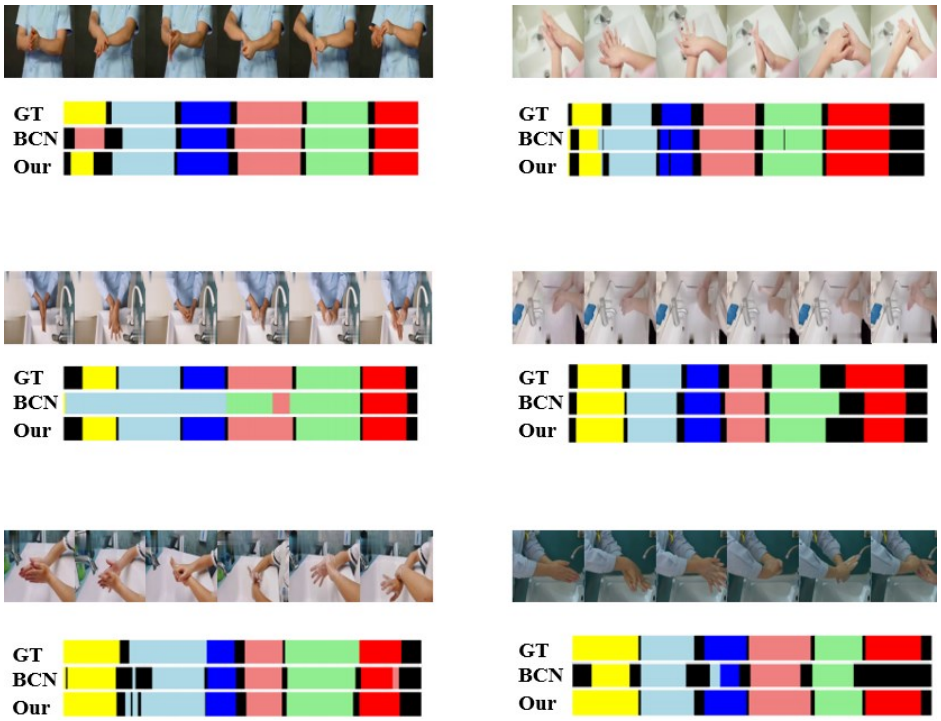

图 4. 4　HHA300 数据集上部分样本的步骤分割定性结果

Fig. 4. 4　Qualitative results of step segmentation on some samples from HHA300 dataset

此外，本文在三个具有挑战性的数据集上评估了本文的方法，包括 50Salads、GTEA 和 Breakfast 数据集。结果见表 4.3，可以看出，本文在所有三个公开数据集上的逐帧准确率都达到了最高，实现了最先进的性能。同时，在分段编辑距离和分段 F1 分数上也取得了有竞争力的结果。与基准模型 BCN 相比，本文在 50Salads、GTEA 和 Breakfast 数据集上分别实现了 0.8%、1.5%和 1.1%的逐帧准确率提升同时还提高了分段编辑距离和分段 F1 分数。在 Breakfast 数据集上，本文的提出的多阶段卷积-Transformer 模型和 BCN 在分段编辑距离方面具有相似的性能，但本文的框架在 F1@{10，25，50}中以{1.2%，1.6%，2.0%}的优于它。这些结果表明，本文的框架可以识别与真实片段和地面真相重叠的动作片段。然而，与 ASRF 相比，尽管本文的框架在逐帧准确率方面优于它，但本文的框架的片段编辑距离和片段 F1 分数不如 ASRF。ASRF 通过检测动作边界来校正预测结果可以提高片段编辑距离和片段 F1 分数，但 ASRF 预测动作边界有一定的偏差，因此影响了逐帧的准确率。本文的框架可以在三个度量中实现良好的平衡，并在不损失

逐帧准确率的情况下改进其他度量。

表 4. 3　在公开数据集上进行动作分割的结果

Table 4. 3　Results of action segmentation on public datasets

	50Salads			GTEA			Breakfast		
	Acc	Edit	F1{10,25,50}	Acc	Edit	F1{10,25,50}	Acc	Edit	F1{10,25,50}
IDT+LM[80]	48.7	45.8	44.4 38.9 27.8	-	-	-	-	-	-
ST-CNN[81]	59.4	45.9	55.9 45.6 37.1	60.6	-	58.7 54.4 41.9	-	-	-
Bi-LSTM[82]	55.7	55.6	62.6 58.3 47.0	55.5	-	66.5 59.0 43.6	-	-	-
ED-TCN[29]	64.7	59.8	58.0 63.9 52.6	64.0	-	72.2 69.3 56.0	43.3	-	-
TDRN[30]	68.1	66.0	72.9 68.5 57.2	70.1	74.1	79.2 74.4 62.7	-	-	-
SSA-GAN[83]	73.3	69.8	74.9 71.7 67.0	74.4	76.0	80.6 79.1 74.2	-	-	-
MS-TCN	80.7	67.9	76.3 74.0 64.5	76.3	79.0	85.8 83.4 69.8	66.3	61.7	52.6 48.1 37.9
MS-TCN++	83.7	74.3	80.7 78.5 70.1	80.1	83.5	88.8 85.7 76.0	67.6	65.6	64.1 58.6 45.9
ASRF	84.5	**79.3**	**84.9 83.5 77.3**	77.3	83.7	**89.4** 87.8 **79.8**	67.6	**72.4**	**74.3 68.9** 56.1
BCN	84.4	74.3	82.3 81.3 74.0	79.8	84.4	88.5 87.1 77.3	70.4	66.2	68.7 65.5 55.0
Ours	**85.2**	76.5	83.7 82.6 76.3	**81.3**	**84.7**	89.0 **88.2** 78.0	**71.5**	67.7	69.9 67.1 **57.0**

4.4.4　消融实验

为了验证基于步骤分割和关键动作打分的手卫生评估方法中所提出的每个模块的有效性，本文采用以下几种方案进行消融实验。

1.本文设计了一个基于多阶段卷积-Transformer 的步骤分割网络。为了验证在多阶段卷积中嵌入 Transformer 的有效性，本文进行了相关实验。表 4.4 显示了多阶段卷积网络和多阶段卷积-Transformer 网络在 HHA300 数据集上的比较结果。如表所示，嵌入线性 Transformer 后，虽然在逐帧准确率上的提升较小，但其他指标有很大的提升，这些指标通常用来反映模型降低过分割的能力。因此，提出的基于卷积-Transformer 的多阶段模型可以有效缓解过分割问题。

表 4. 4　在 HHA300 数据集上本方法和多阶段纯卷积模型的结果

Table 4. 4　Results of our framework and multi-stage convolution model on HHA300

方法	Acc	Edit	F1{10,25,50}
多阶段纯卷积模型	89.0	80.9	87.2 86.1 81.9
Ours	**89.1**	**83.3**	**89.7 89.2 83.0**

此外，本文还比较了传统 Transformer 和线性 Transformer 的结果和计算成本，如表 4.5 所示。基于这些结果可以发现，基于线性 Transformer 的模型可以实现与传统 Transformer 相似的性能，但具有更低的计算成本。

表 4. 5　在 HHA300 数据集上本方法和原始 Transformer 的结果

Table 4. 5　Results of our framework and the original Transformer on HHA300

方法	Acc	Edit	F1{10,25,50}	Params(M)	FLOPs(G)	GPU Mem.
Ours-原始 Transformer	**89.8**	82.7	89.4 89.1 **83.1**	12.46	6.14	~2.63G
Ours	89.1	**83.3**	**89.7 89.2** 83.0	12.46	5.77	~2.09G

2.为了验证本文提出的基于步骤对长动作进行评估的方案的有效性，本文将其与用整个视频回归分数的传统方法进行比较。为了公平起见，本文在两种方法中都使用了 MLP 来回归得分，结果如表 4.6 所示。从表中可以看出，本文基于步骤的评估方法远远优于用整个视频回归分数的传统方法。这是因为传统方法很难准确评估基于步骤的任务，其质量通常由每个步骤中的关键行为决定。而本文提出的基于步骤的评估模型可以很好地解决这个问题，从而获得更好的结果。

表 4. 6　在 HHA300 数据集上基于步骤的评估与传统方法的比较。

Table 4. 6　Comparison of the step based assessment with the traditional method on HHA300

方法	Spearman's Correlation↑	$R-l_2(\times 100)\downarrow$
完整视频-MLP	0.390	35.05
步骤分割-MLP	**0.852**	**1.07**

3.为了验证关键动作打分的有效性，本文评估了表 4.7 中的两种方法。分别是对手卫生视频进行步骤分割之后结合 MLP 的评估方法和关键动作打分器的评估方法。结果验证了本文设计的有效性。

表 4. 7　在 HHA300 上用于手卫生评估的关键动作打分的结果

Table 4. 7　Results of key action scorer of the hand hygiene assessment on HHA300

方法	Spearman's Correlation↑	$R-l_2(\times 100)\downarrow$
Ours-MLP	0.814	1.49
Ours-KAS	**0.852**	**1.07**

除此之外，本文还尝试将每个步骤对应的关键动作打分器的输出显示出来，这样使用者就可以根据每个步骤的得分对自己的手卫生行为进行更正。如图 4.5 所示，步骤分割模型将手卫生视频分成基于步骤的视频片段，手卫生评估模型评估每个步骤的关键

动作，并最终计算整个视频的最终评估分数。可以发现，步骤一，二，五，六这四个步骤在数据集中标注的属性为 ES，即存在且标准，因此关键动作打分器给出的质量分数较高，都在 0.9 分以上；而步骤四在数据集中标注的属性为 EN，即存在但不标准，关键动作打分器给出的得分低于前面四个步骤，只有 0.73 分；步骤三在该手卫生视频中不存在，因此关键动作打分器给步骤三的分数为 0。根据这个结果，本文的模型给出的建议是该志愿者应当在手卫生行为中加入步骤三，并且规范步骤四。

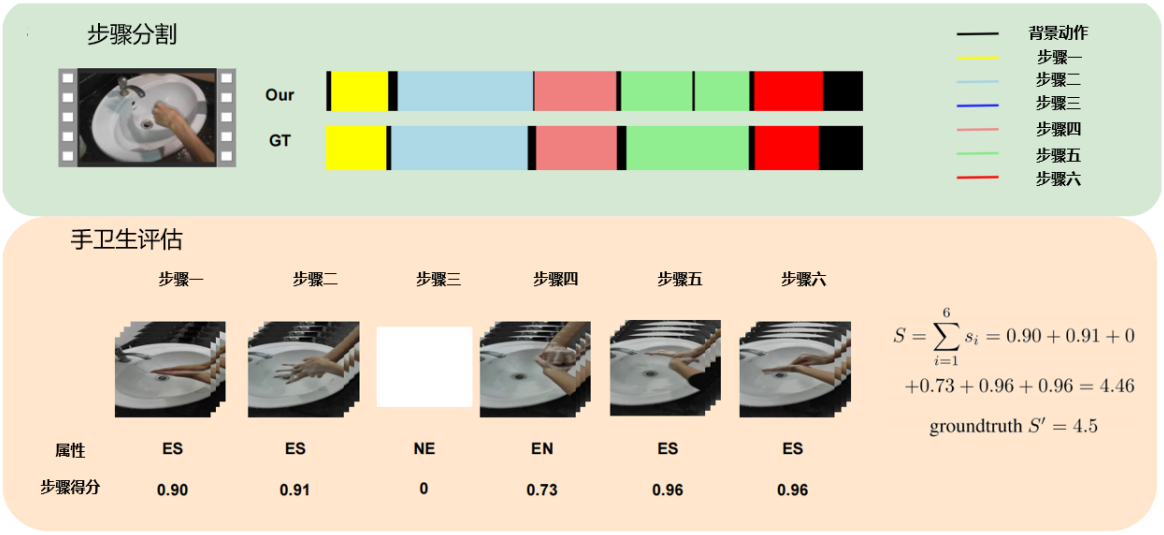

图 4. 5　本文模型进行步骤分割和评估的示例

Fig. 4. 5　An example of step segmentation and assessment by our method.

4.在本文设计的关键动作打分器中，使用了可学习的 Sigmoid 函数代替原始的 Sigmoid 函数。为了验证可学习的 Sigmoid 对关键动作打分器的影响，本文将它与原始的 Sigmoid 进行了比较。实验结果如表 4.8 所示，实验结果表明本文选择的可学习 Sigmoid 可以达到最好的性能，这证明了可学习 Sigmoid 在关键动作打分器中的有效性。

表 4. 8　在 HHA300 数据集上本方法和原始 Sigmoid 的结果

Table 4. 8　Results of our framework and without original Sigmoid on HHA300

方法	Spearman's Correlation↑	$R-l_2(\times 100)\downarrow$
KAS-原始 Sigmoid	0.821	1.42
KAS-可学习 Sigmoid	**0.852**	**1.07**

4.5　本章小结

现有的大多方法利用整个视频来回归长动作的质量得分，无法分析其中的细粒度知识，这就导致了用于长动作例如手卫生的动作质量评估准确度较低，针对这些问题，本

章提出一种基于步骤分割和关键动作打分的手卫生评估算法，与之前的方法相比，它的优势主要体现在拥有能够将手卫生视频精准分割为步骤片段的多阶段卷积-Transformer网络，以及可以针对关键动作的基于可学习 Sigmoid 的关键动作打分器。对于多阶段卷积-Transformer 网络，本文在多阶段卷积模型的基础上加入了线性 Transformer，在不增加太多计算量的情况下加强了长距离帧之间的相关性，增强了模型抑制过分割的能力，提高了步骤分割的性能。而对于基于可学习 Sigmoid 的关键动作打分器，本文利用可学习的 Sigmoid 函数，通过不同的参数去更好的评估不同的关键动作，提高了模型的可解释性。通过本文提出的基于步骤分割和关键动作打分器的手卫生评估方法，可以对手卫生行为实现鲁棒的质量评估，并且在实验部分，本文在提出的 HHA300 数据集以及其他公开数据集上验证了有效性，与其他方法相比，本文提出的方法具有较好的性能表现。

总结与展望

人工智能技术的快速发展给人们的日常生活带来了巨大的变化，其中，计算机视觉更是为人类提供了巨大的帮助，动作质量评估作为计算机视觉中的一个热点话题，近些年受到了越来越多研究者的关注。动作质量评估是一种监测和评估行为质量的人工智能技术，这种技术通过利用计算机对视频中人的行为进行质量评价。在医疗、体育比赛、技能培训等领域有非常重要的实际应用价值。受新冠肺炎疫情影响，动作质量评估在医疗方面的应用被越来越多的提出。其中，手卫生评估是动作质量评估在医疗方面重要的应用之一，手卫生是由世界卫生组织提出的一种六步洗手动作，目的是规范医务人员的洗手行为，良好的手卫生习惯可以预防绝大多数传染病，但是目前还没有很好的办法来监督医务人员做好手卫生，这就导致了疾病传播的潜在风险。现有的动作评估方法通常对整个动作视频进行整体质量评估。然而，手卫生动作的细粒度知识以及不同步骤之间的关系在手卫生评估中是非常重要的。本文针对手卫生评估方法展开研究，并取得了如下成果：

本文创建了一个统一的高质量视频数据集用于手卫生动作分析，包括手卫生步骤分割和手卫生评估，称为 HHA300。HHA300 包含不同人的 300 个洗手视频，来自各种各样的视角和复杂的背景，包括很多现实世界中手卫生行为可能发生的情况，保障了数据集的多样性和挑战性。为了提供高质量的细粒度注释，除了对每个视频标注质量分数，本文还提供了帧级标签，包括世卫组织规定的六个步骤和背景动作。除了提供了一个高质量的手卫生评估数据集，本文还提出了一个基于步骤分割和关键动作打分的手卫生评估模型以进行准确的手卫生评估，首先，不同于现有的动作评价方法直接对整个视频进行评价，本文将输入视频分割成基于步骤的视频片段进行细粒度评分。对于步骤分割，由于手卫生视频中的动作是连续且相似的，多阶段的纯卷积网络容易导致过分割。为了解决这个问题，本文利用线性 Transformer 来建立长距离帧之间的相关性，提出了一个基于多阶段卷积-Transformer 的步骤分割模型。在制作数据集的过程中本文还观察到手卫生的每一个步骤中都包含一些决定步骤质量的关键动作。为了准确评估每个步骤的质量，同时，本文观察到每个洗手步骤都包含几个决定洗手动作完成质量的关键动作，因此设计了一套基于可学习 Sigmoid 的关键动作打分器来评估每个步骤中关键动作的质量。实验结果表明，本文的框架可以准确地评估手卫生视频。

本文的相关贡献和讨论虽然与之前的方法相比有了一定的提升，但仍有很大的进步

空间，需要该领域众多研究者一起努力，希望本文能够作为一个铺垫，以促进手卫生视频分析的发展。此外，经过反思和讨论，对本文工作出现中的不足，进行未来展望：

1.尽管本文提出了手卫生视频数据集，但是对于视频理解领域来说，数据集的规模还有待扩充。同时，本文希望通过在后续该数据集的实际使用中，研究者们能发现数据集存在的不足。最后能够提出更大规模，更加完善的手卫生视频数据集，以促进手卫生视频理解领域的发展。

2.虽然本文提出了有效的手卫生评估方法，但是它的实时性还达不到生产生活的需要，并且无法做到在线对手卫生行为进行分析，因此在未来，本文希望对本文提出的基于步骤分割和关键动作打分的手卫生评估模型进行精简以及改进，在保持性能的同时，提高模型的运行速度，将其扩展为手卫生评估的在线版本，以提高实用性。

3. 本文使用的评价指标是参考体育比赛的动作评估的，但手卫生评估任务有其特殊性，可能需要不同的评价指标。未来工作可以探索更适合手卫生评估任务的评价指标，完善评价方法。

参考文献

[1] Antunes M, Baptista R, Demisse G, et al. Visual and human-interpretable feedback for assisting physical activity[C]. Proceedings of the European Conference on Computer Vision Workshops. 2016: 115-129.

[2] Paiement A, Tao L, Hannuna S, et al. Online quality assessment of human movement from skeleton data[C]. British Machine Vision Conference. 2014: 153-166.

[3] Li Y, Chai X, Chen X. End-to-end learning for action quality assessment[C]. Pacific-Rim Conference on Multimedia. 2018: 125-134.

[4] Li Y, Chai X, Chen X. Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports[C]. Asian Conference on Computer Vision. 2019: 149-164.

[5] Parmar P, Morris B. Action quality assessment across multiple actions[C]. IEEE Winter conference on Applications of Computer Vision. 2019: 1468-1476.

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

[7] Wang Z, Majewicz Fey A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery[J]. International journal of computer assisted radiology and surgery, 2018, 13: 1959-1970.

[8] Xiang X, Tian Y, Reiter A, et al. S3d: Stacking segmental p3d for action quality assessment[C]. 2018 25th IEEE International Conference on Image Processing. 2018: 928-932.

[9] 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.

[10] Parmar P, Tran Morris B. Learning to score olympic events[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 20-28.

[11] Doughty H, Damen D, Mayol-Cuevas W. Who's better? who's best? pairwise deep ranking for skill determination[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6057-6066.

[12] 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 Conference on Computer Vision and Pattern Recognition. 2019: 7862-7871.

[13] Li Z, Huang Y, Cai M, et al. Manipulation-skill assessment from videos with spatial attention network[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019: 0-0.

[14] Allegranzi B, Gayet-Ageron A, Damani N, et al. Global implementation of WHO's multimodal strategy for improvement of hand hygiene: a quasi-experimental study[J]. The Lancet infectious diseases, 2013, 13(10): 843-851.

[15] 国家卫生健康委员会.医务人员手卫生规范 WS/T313—2019 [J].中华医院感染学杂志,2020,30(5):796.

[16] Wiktorczyk-Kapischke N, Grudlewska-Buda K, Wałecka-Zacharska E, et al. SARS-CoV-2 in the environment—non-droplet spreading routes[J]. Science of the total environment, 2021, 770: 145260.

[17] Fierer N, Hamady M, Lauber C L, et al. The influence of sex, handedness, and washing on the diversity of hand surface bacteria[J]. Proceedings of the National Academy of Sciences, 2008, 105(46): 17994-17999.

[18] Ivanovs M, Kadikis R, Lulla M, et al. Automated quality assessment of hand washing using deep learning[J]. arXiv preprint arXiv:2011.11383, 2020.

[19] Llorca D F, Parra I, Sotelo M Á, et al. A vision-based system for automatic hand washing quality assessment[J]. Machine Vision and Applications, 2011, 22: 219-234.

[20] Zhong C, Reibman A R, Mina H A, et al. Designing a Computer-Vision Application: A Case Study for Hand-Hygiene Assessment in an Open-Room Environment[J]. Journal of Imaging, 2021, 7(9): 170.

[21] Xie T, Tian J, Ma L. A vision-based hand hygiene monitoring approach using self-attention convolutional neural network[J]. Biomedical Signal Processing and Control, 2022, 76: 103651.

[22] Zhong C, Reibman A R, Mina H A, et al. Multi-view hand-hygiene recognition for food safety[J]. Journal of Imaging, 2020, 6(11): 120.

[23] Vo H Q, Do T, Pham V C, et al. Fine-grained hand gesture recognition in multi-viewpoint

hand hygiene[C]. IEEE International Conference on Systems, Man, and Cybernetics. 2021: 1443-1448.

[24] Rohrbach M, Amin S, Andriluka M, et al. A database for fine grained activity detection of cooking activities[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012: 1194-1201.

[25] Karaman S, Seidenari L, Del Bimbo A. Fast saliency based pooling of fisher encoded dense trajectories[C]. Proceedings of the European Conference on Computer Vision Workshops. 2014, 1(2): 5.

[26] Kuehne H, Gall J, Serre T. An end-to-end generative framework for video segmentation and recognition[C]. IEEE Winter Conference on Applications of Computer Vision. 2016: 1-8.

[27] Lea C, Reiter A, Vidal R, et al. Segmental spatiotemporal cnns for fine-grained action segmentation[C]. Proceedings of the European Conference on Computer Vision. 2016: 36-52.

[28] Richard A, Kuehne H, Gall J. Weakly supervised action learning with rnn based fine-to-coarse modeling[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 754-763.

[29] Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 156-165.

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

[31] Ding L, Xu C. Weakly-supervised action segmentation with iterative soft boundary assignment[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6508-6516.

[32] Farha Y A, Gall J. MS-TCN: Multi-stage temporal convolutional network for action segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3575-3584.

[33] 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, 1-1.

[34] 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.

[35] Xie L, Chen X, Bi K, et al. Weight-sharing neural architecture search: A battle to shrink the optimization gap[J]. ACM Computing Surveys, 2021, 54(9): 1-37.

[36] Zoph B, Le QV. Neural architecture search with reinforcement learning[C]. International Conference on Learning Representations. 2017.

[37] Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable i mage recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8697–8710.

[38] Xu H, Yao L, Zhang W, et al. Auto-FPN: Automatic network architecture adaptation for object detection beyond classification[C]. Proceedings of the IEEE International Conference on Computer Vision. 2019: 6649-6658.

[39] Gao S H, Han Q, Li Z Y, et al. Global2local: Efficient structure search for video action segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 16805-16814.

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

[41] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[42] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]. The North American Chapter of the Association for Computational Linguistics. 2018,(1).

[43] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. Proceedings of the European Conference on Computer Vision. 2020: 213-229.

[44] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. International Conference on Learning Representations. 2021.

[45] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]. International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[46] Yi F, Wen H, Jiang T. Asformer: Transformer for action segmentation[C]. Proceedings of British Machine Vision Conference, 2021.

[47] Li Y, Dong Z, Liu K, et al. Efficient two-step networks for temporal action segmentation[J]. Neurocomputing, 2021, 454: 373-381.

[48] Ward M A, Schweizer M L, Polgreen P M, et al. Automated and electronically assisted hand hygiene monitoring systems: a systematic review[J]. American journal of infection control, 2014, 42(5): 472-478.

[49] Yamamoto K, Yoshii M, Kinoshita F, et al. Classification vs regression by cnn for handwashing skills evaluations in nursing education[C]. International Conference on Artificial Intelligence in Information and Communication. 2020: 590-593.

[50] Pirsiavash H, Vondrick C, Torralba A. Assessing the quality of actions[C]. Proceedings of the European Conference on Computer Vision. 2014: 556-571.

[51] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.

[52] Pan J H, Gao J, Zheng W S. Action assessment by joint relation graphs[C]. Proceedings of the IEEE International Conference on Computer Vision. 2019: 6331-6340.

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

[54] Zeng L A, Hong F T, Zheng W S, et al. Hybrid dynamic-static context-aware attention network for action assessment in long videos[C]. Proceedings of the 28th ACM International Conference on Multimedia. 2020: 2526-2534.

[55] Yu X, Rao Y, Zhao W, et al. Group-aware contrastive regression for action quality assessment[C]. Proceedings of the IEEE International Conference on Computer Vision. 2021: 7919-7928.

[56] Jain H, Harit G, Sharma A. Action quality assessment using siamese network-based deep

metric learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(6): 2260-2273.

[57] Xu J, Rao Y, Yu X, et al. Finediving: A fine-grained dataset for procedure-aware action quality assessment[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 2949-2958.

[58] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]. International Conference on Learning Representations. 2015.

[59] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]. Proceedings of the International Conference on Machine Learning. 2010: 807-814.

[60] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[61] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[62] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. Proceedings of the European Conference on Computer Vision. 2016: 483-499.

[63] Dantone M, Gall J, Leistner C, et al. Body parts dependent joint regressors for human pose estimation in still images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2131-2143.

[64] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[65] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[66] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. International Conference on Machine Learning. PMLR, 2015: 448-456.

[67] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[68] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations. 2015.

[69] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.

[70] 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.

[71] Redmon J, Farhadi A. Yolov3: An incremental improvement[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[72] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]. International Conference on Learning Representations. 2017.

[73] Dietz A, Pösch A, Reithmeier E. Hand hygiene monitoring based on segmentation of interacting hands with convolutional networks[C]. Imaging Informatics for Healthcare, Research, and Applications. 2018, 10579: 273-278.

[74] Katharopoulos A, Vyas A, Pappas N, et al. Transformers are rnns: Fast autoregressive transformers with linear attention[C]. International Conference on Machine Learning. PMLR, 2020: 5156-5165.

[75] Clevert, Djork-Arné, Unterthiner T , Hochreiter S . Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)[C]. International Conference on Learning Representations. 2016.

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

[77] Fathi A, Ren X, Rehg J M. Learning to recognize objects in egocentric activities[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2011: 3281-3288.

[78] 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 Conference on Computer Vision and Pattern Recognition. 2014: 780-787.

[79] Ishikawa Y, Kasai S, Aoki Y, et al. Alleviating over-segmentation errors by detecting action boundaries[C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision. 2021: 2322-2331.

[80] Richard A, Gall J. Temporal action detection using a statistical language model[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3131-3140.

[81] Lea C, Reiter A, Vidal R, et al. Segmental spatiotemporal cnns for fine-grained action segmentation[C]. Proceedings of the European Conference on Computer Vision. 2016: 36-52.

[82] 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 Conference on Computer Vision and Pattern Recognition. 2016: 1961-1970.

[83] Gammulle H, Denman S, Sridharan S, et al. Fine-grained action segmentation using the semi-supervised action GAN[J]. Pattern Recognition, 2020, 98: 107039.