计算机科学
COMPUTER SCIENCE

融合跨模态注意力与角色交互的学生课堂专注度研究

卓铁农, 英迪, 赵晖

引用本文

卓铁农, 英迪, 赵晖. 融合跨模态注意力与角色交互的学生课堂专注度研究[J]. 计算机科学, 2026, 53(2): 67-77.

ZHUO Tienong, YING Di, ZHAO Hui. Research on Student Classroom Concentration Integrating Cross-modal Attention and Role Interaction [J]. Computer Science, 2026, 53(2): 67-77.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250300026

融合跨模态注意力与角色交互的学生课堂专注度研究

卓铁农[1]　英　迪[2]　赵　晖[2]
1 新疆大学软件学院　乌鲁木齐 830046
2 新疆大学计算机科学与技术学院　乌鲁木齐 830046
(shaoyang1906@163.com)

摘　要　随着智慧教育的不断发展,学校可以通过检测学生课堂的专注度对学生的学习情况与教师的教学质量进行评估,从而优化教学体系。以往的研究多侧重于单模态、单角色的特征提取,但教学课堂是一个多模态、多角色且角色之间相互影响的复杂场景,因此从多模态多角色角度去探讨学生课堂的专注度具有重大意义。然而,多模态之间如何有效建模时间相关性与语义交互性,以及多角色之间如何相互影响是实现学生课堂专注度评判的重大挑战。针对以上问题,构建了一个包含教师音频和学生视频的学生课堂专注度数据集,并提出了基于多模态多角色的长短时上下文学生课堂专注度评估模型(Long-Short Context Model,LSCM)。其中多模态是指学生的视频与教师的音频,多角色是指学生与学生、学生与教师。该模型主要包含长时上下文模块和短时上下文模块两个模块。长时上下文模块通过音频自注意机制和视觉自注意机制提取单一学生的长时行为特征,并利用视听交叉注意机制增强音频与视觉信息的关联性;短时上下文模块则聚焦于局部时间片段,以刻画课堂环境中多个学生专注度的动态变化。最后,模型输出视频中各个学生的专注度类别。实验表明,该方法通过有效挖掘多模态数据的互补性及角色间的关联性,使专注度检测准确率较现有方法显著提高,验证了多模态融合与角色交互建模的有效性。

关键词:多模态;学生专注度;教学课堂;角色交互;注意力机制
中图分类号　TP391.1

Research on Student Classroom Concentration Integrating Cross-modal Attention and Role Interaction

ZHUO Tienong[1],YING Di[2] and ZHAO Hui[2]
1 School of Software,Xinjiang University,Urumqi 830046,China
2 School of Computer Science and Technology,Xinjiang University,Urumqi 830046,China

Abstract　With the continuous development of innovative education,schools can assess students' learning and teachers' teaching quality by detecting students' concentration in the classroom to optimize the teaching system. Previous studies have focused on single-modality and single-role feature extraction. However,the teaching classroom is a complex scene with multimodal,multiple roles,and interactions between the roles,so it is of great significance to explore students' classroom attentiveness from the perspective of multimodal and multiple roles. However,how to effectively model the temporal relevance and semantic interaction between multimodal and how the multiple roles interact is a significant challenge in realizing the judgment of students' classroom concentration. To address the above problems,a student classroom concentration dataset containing teacher's audio and student's video is constructed,and a Long-Short Context Model(LSCM) based on multimodal and multi-role assessment of students' classroom concentration is proposed,in which multimodal refers to the student's video and the teacher's audio. Multi-role refers to the student-to-student and student-to-teacher. The model contains two main modules:the long-term context module and the short-term context module. Specifically,the long-term context module extracts the long-time behavioral characteristics of a single student through the audio self-attention mechanism and the visual self-attention mechanism. The audio-visual cross-attention mechanism enhances the correlation between the audio and visual information. In contrast,the short-term context module focuses on localized time segments to portray the dynamic changes in the attentiveness of multiple students in the classroom environment. Finally,the model outputs the concentration categories of each student in the video. Experiments show that this method signifi-

到稿日期:2025-03-05　返修日期:2025-08-26
基金项目:新疆维吾尔自治区重点研发计划(2023B01032);国家自然科学基金(62166041)
This work was supported by the Key R&D Program of Xinjiang Uygur Autonomous Region(2023B01032)and National Natural Science Foundation of China(62166041).
通信作者:赵晖(zhaohui@xju.edu.cn)

cantly improves concentration detection accuracy compared with existing methods by effectively exploiting the complementary nature of multimodal data and the correlation between roles. It also verifies the effectiveness of multimodal fusion and role interaction modeling.

Keywords Multimodal, Student concentration, Teaching classroom, Role interaction, Attention mechanism

1 引言

信息技术的不断发展和人工智能的进步为高等教育质量的提升提供了有力支持，而课堂教学质量则在人才培养过程中扮演着至关重要的角色。目前，许多大学课堂面临学生专注度低、互动不足等问题，这直接影响了教学效果及学生的全面发展。因此，如何有效评估学生课堂专注度，并据此优化教学体系，已成为亟待解决的关键课题。

目前，许多研究者提出了不同方法来评估学生的课堂专注度。Zhong 等[1]将模糊综合评价算法应用于专注度分析，通过整合姿态检测、表情识别和疲劳度评估等多维度数据，并运用运筹学方法进行指标融合，实现了对学生专注度的综合评估。Zaletelj 等[2]基于 Kinect 信号，深入探究了姿态特征与行为线索之间的关联性，分析姿态与行为之间的关系，以确定学生的不同专注度等级。Duan[3]设计并实现了一种基于机器视觉的课堂专注度分析系统，该系统结合人脸识别技术、抬头检测算法和侧脸检测算法，以评估学生的课堂专注度。Zuo 等[4]提出了一种客观量化的课堂专注度算法，通过精确定位人眼并计算眼睛的张合度来评估学生的专注情况。He 等[5]提出了一种融合特征提取方法，结合深度学习模型的自学习能力进行表情识别，并通过情绪分类评估学生的专注度。同时，Wang 等[6]提出一种用于线上学习的专注度评价模型 CE-HPE，其利用头部姿态与面部表情两种特征进行专注度估计。研究者利用计算机视觉技术分析学生的眼动轨迹、头部姿态以及面部表情，并基于视频信息中的多特征进行课堂专注状态评估。然而，单模态的视频信息有限，受周围环境、光照、遮挡等因素的影响较大。

以往的研究主要集中在单一模态和单一角色的特征提取上，然而，教学课堂是一个充满互动的多模态、多角色的复杂环境，各个角色之间相互影响，彼此之间的动态关系对课堂专注度的影响不容忽视。因此，从多模态和多角色的角度来探讨学生课堂专注度具有重要的理论和实践意义。然而，现有研究面临两个主要挑战：一方面，缺乏包含音频、视频等多模态信息的高质量公开数据集；另一方面，如何有效地建模多模态之间的时间相关性和语义交互性，以及如何揭示不同角色之间的相互作用，仍是实现准确评估学生课堂专注度的核心难题。为解决上述问题，本文构建了一个包含教师音频和学生视频的学生课堂专注度数据集，并提出了一个基于多模态多角色的长短时上下文学生课堂专注度评估模型——LSCM，该模型包括长时上下文模块和短时上下文模块两个核心模块。长时上下文模块通过音频自注意机制和视觉自注意机制提取单一学生的长时行为特征，同时考虑教师音频与学生视频之间的互动关系，利用视听交叉注意机制增强音频与视觉信息的关联性，进而捕捉教师与学生之间专注度的相互影响。短时上下文模块则专注于局部时间片段，考虑学生间专注度的相互作用，捕捉课堂中多个学生专注度的动态变化。最终，模型输出每个学生在视频中的专注度类别，从而实现对课堂专注度的全面评估。本文的主要贡献如下：

1）构建了一个包含音频与视频信息的课堂专注度数据集，并对学生课堂行为进行精细标注，划分为专注、不专注和极不专注 3 个等级，为多模态专注度检测研究提供了关键数据支持。

2）提出了一个基于音频和视频的多模态多角色学生课堂专注度检测模型 LSCM，结合学生视频与教师音频，以及学生与学生、学生与教师的角色互动。模型包括长时与短时上下文模块，长时模块通过捕捉学生长期行为特征提供全局视角，短时模块则聚焦课堂内短期变化，确保精准评估学生专注度波动。

3）实验结果表明，该模型通过充分挖掘多模态数据的互补性和角色间的关联性，显著提升了专注度检测的准确率，验证了多模态融合与角色交互建模的有效性。

2 相关工作

2.1 学生课堂专注度研究现状

专注度又称注意力，通常被定义为一个人在集中精力专心完成特定任务或活动时的情绪反应，处于专注状态下的人能够持续地把自己的时间、精力集中到处理当前的事情[7]。早在 20 世纪 30 年代，Tyler[8]就提出时间任务的概念，即学习时间越久，就能获得更多的知识。而 Pace[9]提出，要想获得最佳学习效果，不但要考虑时间因素，还要关注学习过程中的专注程度，也就是学习的质量。因此，学习投入的质量和时间都是要考虑的重要因素。

2.1.1 传统学生专注度识别

问卷调查是一种常见的专注度识别方法，通过问卷问答的方式来收集专注度相关的信息。传统问卷调查的专注度识别属于非自动化的专注度检测，通过分析收集问卷中的学生课后的数据以及上课时的观察性评分量表来判断学生的专注度。在传统问卷调查中，通常会选择一些与专注度相关的特征作为调查问题，如注意力集中时间、注意力分散情况、专注度障碍、最近一段时间是否容易分心等。被调查者需要回答问题，并选择相应的答案。

在 1970 年代后期，人们对促进学生成功的兴趣日益浓厚，这导致研究人员、政策制定者和高等教育管理人员强调学生的参与。为了应对这一趋势，2000 年代初引入了国家级的专注度调查，例如大学生体验问卷（College Student Experiences Questionnaire，CSEQ）和美国全国学生专注度调查（National Survey of Student Engagement，NSSE）[10]。其他国家也实施了类似的调查，包括英国全国学生调查、南非学生参与调查、加拿大全国学生参与调查、爱尔兰学生参与调查和澳大利亚学生参与调查。这些调查主要是 NSSE 的改编版本，用

于跨机构数据收集。尽管在过去的二十年里,学生参与研究和调查在许多国家越来越受欢迎,但是这种方式存在明显的局限和不足之处,因为问卷收集得到的问题答案大多具有主观性色彩,学习者可能在不同的时间对问题有不一样的看法。所以,采用这种方法既耗费大量的时间和精力,效率也不高,有其明显的缺陷,无法大范围地推广。

2.1.2 基于图像的专注度识别

专注度检测通常分为自动化检测和非自动化检测。其中,基于图像的专注度识别属于自动化检测方式,其通过计算机视觉技术对图像进行分析,从而评估个体的专注程度。

基于图像的研究主要通过捕捉和分析静态图像中的学生行为来评估专注度[11-13]。Whitehill 等[14]对单张图像进行学生专注度识别,分析图片中人脸的面部像素特征,并依据这些特征训练了一个用于预测专注度等级的 SVM 分类器,从而对图像数据集中的学生专注度进行预测输出。Sukumaran 等[15]提出了一种专注度识别模型,将学生的面部表情特征与头部特征进行融合,构建了专注度识别系统。该系统在每节课结束后,通过学生的随堂测验分数进行了性能验证。此模型可用于学生情感状态的实时视频处理。教师可以在课程结束时在电子表格上获得专注度统计数据的详细分析,从而进行必要的后续行动。Zhong 等[1]针对在线教育中学生专注度评估问题,分别对图像中的人脸和头部姿态、疲劳度打分,利用模糊综合评判的方式计算专注度评分。

2.1.3 基于视频的专注度识别

近年来,教育领域的学习模型发生了很大改变,从线下学习转向了在线学习。虽然这种转变提供了一些优势,例如将学习过程从时间和空间的限制中解放出来,并使教育能够随时随地进行,但由于互动有限,在线学习期间检测学生的专注度成为一个挑战。Santoni 等[16]采用深度学习集成方法,通过在线学习中的视频记录来识别学生的专注度。

Chen 等[17]通过计算机视觉预测了学生对协作学习的专注度,引入了一种深度神经网络(MDNN),它将面部表情和注视方向集成为两个关键组成部分,以预测学生在协作学习中的专注度。同时,Buono 等[18]提出了一种使用 LSTM 的深度学习方法,该方法融合了眼睛凝视、头部姿势和面部动作单位等特征,从而预测了学生在线学习的专注度。与此同时,Ikram 等[19]使用 32 个面对面课堂视频,利用 VGG16 的深度学习方法预测学生的专注度。

借鉴情感计算领域的研究,Kaur 等[11]提出了一个新的学生专注度检测方法,探索了受试者行为线索与专注度之间的关联性。该研究采用了一种基于深度多实例学习的框架,成功识别出学生的专注度。近年来,随着深度学习技术的发展,多模态融合方法逐渐受到关注。Lai 等[20]利用交叉验证提供可靠的学习专注度标签,提出了基于多模态生理信号的学习专注度识别的方法。Deng 等[21]提出了一种名为 Tg-GTM(文本引导的图时间建模)的多模态模型,用于小样本视频分类。为了实现多模态时间对齐,将查询文本描述作为附加模态集成到时间对齐过程中,使模型能够利用时间上下文依赖性和多模态信息,提高小样本视频分类的准确性。

相比之下,基于视频的方法则通过捕捉和分析视频中的学生行为来评估专注度。通过捕捉学生专注度随时间的动态演变,可以更全面地了解学生在学习过程中的专注度是如何波动的。这不仅可以识别学生的静态特征,如面部表情和身体姿势,还可以跟踪学生的动态行为,如头部运动目光交流等。这种方法能够更全面地评估学生的专注度,因为它能够捕捉到更多的行为细节和连续的学习过程。这种基于视频的方法可以进一步分为端到端和基于特征的专注度检测[22]。基于视频的端到端方法直接处理原始输入数据以预测互动标签,然而,学生的专注度会随着时间的推移而变化,注意力和兴趣水平会有所波动。Das 等[23]利用知识迁移的概念,提出了一种新的帧级别框架学生专注度分类方法。该方法侧重于在细粒度级别上识别学生的专注度,使教师能够确定不专注或专注的特定时刻,从而进行有针对性的干预。

2.2 数据集的局限性

在专注度识别领域,已有多个数据集被提出用于研究学生的专注度,特别是在课堂环境中的应用,限制了其在现实世界中的有效应用。

现有数据集大多未公开,导致模型训练和验证的可重复性较差。Hernandez 等[24]首次尝试识别用户参与问题,将确定电视观众参与度的问题建模为二元分类任务,使用面部几何特征和 SVM 进行分类,并创建了一个非常小的自定义数据集(未公开)。同样,Whitehill 等[14]构建了一个未公开的小规模数据集,但其仅采用单编码者标注的二元分类(专注/不专注),难以覆盖多样化的专注状态。此外,该数据集是在受限设置下捕获的,没有捕捉到现实世界用户参与识别的“野外”需求。Gupta 等[25]提出的 DAiSEE 数据集虽然包含了更大规模的样本(112 个用户和 9068 个视频片段),并结合了面部表情和情感标注数据,但同样存在类似的问题。该数据集主要是在受控环境下收集的,即针对的是复杂的现实课堂环境,无法全面反映学生在实际课堂中专注度的变化。因此,尽管这些数据集为参与度研究提供了基础,但其应用范围和准确性仍然受到限制。

3 数据集

现有的专注度识别数据集在应用于实际课堂环境时存在诸多局限性,主要体现在数据规模、采集环境、标注方式和模态信息等方面,导致其难以满足现实世界对学生专注度检测的需求。

首先,大多数已有的数据集规模较小,限制了模型的泛化能力。这类数据集的受试者较少,覆盖的个体特征、学习情境和行为模式有限,导致训练出的模型难以推广到大规模的学生群体。

其次,数据采集环境过于受控,缺乏现实课堂的动态性。在现实的课堂中,学生的专注度受到讲授方式、课堂互动、外部干扰等多种因素的影响,现有数据集未能有效捕捉这些动态变化。

此外,当前的数据集主要依赖视觉信息(如面部表情、头部姿态)进行专注度评估,而缺乏其他模态信息的补充。单纯

依靠视觉特征容易受到光照变化、遮挡等因素的干扰，导致模型的鲁棒性下降。另一方面，课堂教学中的音频信息对学生的专注度具有重要影响，然而，现有数据集并未整合语音特征，忽略了音频模态对专注度识别的贡献。

基于以上问题，本文提出构建一个包含教师语音的学生专注度多模态数据集。相比现有数据集，该数据集能够捕捉学生在不同教学方式下的自然专注度变化。同时，除了视觉信息外，还引入教师语音特征，使得模型能够利用多模态数据融合技术，提高专注度检测的准确性和鲁棒性。该数据集的构建将为课堂专注度研究提供更全面的支持，并推动智能教育技术的发展。

3.1 数据的收集

为了更准确地研究学生的专注度，本文的数据集探索了多模态信息，尤其是引入了教师音频。教师的音频信息对学生的专注度有重要影响。教师的声音能够反映课堂内容的情感变化，从而影响学生的注意力水平。教师的音频信息作为一种情感线索，可以帮助研究人员理解学生在特定情境下的专注度变化，并进一步提升模型的准确性。

为了收集学生在课堂上的数据，首先，本文开展了一次对所有年级、性别和年龄不设限的志愿者招募，数据集中有 50 名志愿者参与，年龄从 23 至 30 岁。有 25 名女性和 25 名男性。总共收集到 25 个视频，总数据量约为 48 GB。

该数据集通过整合教师音频与学生视频特征，为研究课堂专注度机制提供了全面的数据支持，能够促进多模态深度学习模型在教育领域的应用，同时为优化教学设计和提高课堂教学效果提供了有力的参考。

3.2 数据的处理

视频处理：为了保证每段视频内容的连贯性和逻辑性，本文采用了基于内容的分割方法，而非传统的固定时间窗口切分方式。具体的策略是按照教学内容的完整性进行分割。研究发现，不同课程中，教师在表达一句完整话语时所需的时间会有所不同。因此，视频的切分依据是教师完成一句话的时间，而不是按统一的时间间隔进行切割。这种方式确保了每个视频片段具有完整的教学内容，同时也能更好地反映学生在课堂上对具体内容的专注度变化。对于 25 个录制的视频，分割后的视频片段长度在 9 秒到 60 秒，最终得到 1 356 个专注度相关的视频片段。本文在视频处理阶段使用了 YOLOv5（You Only Look Once v5）算法[26]对视频中的每一帧进行人物检测，从而获取学生的位置坐标信息。本文通过 YOLOv5 仅进行人物框提取而非精细分类，并返回人物位置（边界框坐标）和置信度，在模型训练和测试阶段没有使用 YOLOv5 模型。具体过程如下。

1）视频处理与帧提取：在视频分割后，首先将每个视频片段拆分为单独的帧。每一帧图像将作为 YOLOv5 模型的输入，以进行人物检测。对于每个视频片段中的每一帧，对其进行处理，检测出该帧中的学生。

2）人物检测：使用 YOLOv5 模型对每一帧进行实时人物检测。YOLOv5 会返回每个检测到的学生的边界框坐标（即框住学生的矩形区域），以及每个边界框的置信度。边界框坐标通常包括 4 个值：左上角和右下角的坐标（x_min，y_min，x_max，y_max）。这些坐标表示学生在图像中的位置。

3）学生位置提取：在每帧的检测过程中，YOLOv5 会提供学生的位置坐标。如果帧中有多个学生，YOLOv5 会为每个学生生成一个单独的边界框。这些位置坐标为后续分析提供了重要的空间信息，能够帮助跟踪学生在视频中的位置变化。

为了进一步分析，除了保存视频帧中的学生图像外，还需要保存学生的位置坐标信息，以便后续的专注度分析。具体处理流程如下：对于每一帧图像中的每个学生，通过 YOLOv5 提取出的边界框坐标，将该学生的图像区域裁剪出来，并保存为单独的图像文件。除了图像之外，YOLOv5 还会为每个学生提供位置信息（边界框坐标）。将这些坐标信息保存为 CSV 文件，其中包含每个学生在每一帧中的位置。

4）音频处理：对于每个视频片段，使用 ffmpeg 工具，根据视频的起始和结束时间戳提取相应的音频部分，并保存为 WAV 格式。音频采样率设定为 16 000 Hz。

3.3 数据标注

为精确描述学生的专注度状态，本文将学生的专注度划分为以下 3 个类别：

1）专注：学生表现出良好的专注度，目光注视老师、黑板或其他课堂内容。

2）一般专注：无法明确判断学生的专注状态，例如目光偏离但未表现出明显的分心行为。

3）极不专注：学生专注度明显分散，例如目光游离、低头玩手机或做与课堂无关的行为。

按照时间段式标注：根据学生在视频中的具体表现，对每位学生的专注度状态进行动态标注。例如，一个 20 秒视频中，若编号为 1_1 的学生从第 3 秒至第 10 秒表现出“不专注”，则该时间段标注为 1；若编号为 2_2 的学生从第 15 秒至第 20 秒表现出“专注”，则标注为 2。

3.4 数据集格式

数据集的格式如表 1 所列，包含视频名称、时间戳、学生编号、时间段以及对应的专注度状态。

数据集的每一行记录了一个学生在某一时间点的位置信息和专注状态。具体信息如下。

video_name：视频的名称，表示数据所属的视频文件。timestamp：时间戳，标明当前视频帧的时间点。person_id：学生的编号，标识不同的学生。（x_min，y_min，x_max，y_max）：表示学生在当前时间点的位置的边界框坐标，用于确定学生的空间位置。label：专注标签，表示学生在该时间点是否专注。start_time：表示在切割的视频中，从开始专注或不专注发生的起始时间。end_time：表示在切割的视频中从结束专注或不专注结束的最后时间。例如第一条数据，表示的是在 xinlixue1_2 这个视频片段中的人物（person_id 为 1_45）。该学生从 0 秒开始，到 18 秒结束时呈现出专注的状态，则该段视频的数据标签 label 为 2。在视频的第 0.04 秒时，通过 YOLOv5 得到该同学的图像位置信息，左上角的坐标为（0.023 438，0.671 296），右下角的坐标为（0.101 562，0.520 37）。

表 1 学生专注度数据集

Table 1 Student concentration datasets

video_name	timestamp	person_id	x_min	y_min	x_max	y_max	label	start_time	end_time
xinlixue1_2	0.04	1_45	0.023438	0.520370	0.101562	0.671296	2	0	18
xinlixue1_2	0.21	1_45	0.023438	0.520370	0.101562	0.671296	2	0	18
xinlixue1_2	0.37	1_45	0.023438	0.520370	0.101562	0.671296	2	0	18
xinlixue1_2	0.54	1_45	0.023438	0.519444	0.101562	0.671296	2	0	18
xinlixue1_2	0.71	1_45	0.023438	0.519444	0.101562	0.671296	2	0	18
xinlixue1_15	0.04	1_45	0.000521	0.525000	0.09375	0.677778	2	0	22
xinlixue1_15	0.21	1_45	0.000521	0.525000	0.094271	0.677778	2	0	22
xinlixue1_15	0.37	1_45	0.000521	0.525000	0.094271	0.677778	2	0	22
xinlixue1_15	0.54	1_45	0.000521	0.525926	0.094271	0.677778	2	0	22
xinlixue1_22	0.04	1_45	0.000521	0.525926	0.097396	0.677778	2	0	26
xinlixue1_22	0.21	1_45	0.000521	0.525926	0.097396	0.677778	2	0	26
xinlixue1_22	0.37	1_45	0	0.525926	0.097396	0.677778	2	0	26
xinlixue1_22	0.54	1_45	0.000521	0.525926	0.097396	0.677778	2	0	26
xinlixue1_22	0.71	1_45	0.000521	0.525926	0.097396	0.677778	2	0	26

4 多模态多角色学生课堂专注度检测

本文提出的多模态多角色的长短时上下文学生专注度评估模型(Long-Short Context Modeling，LSCM)旨在通过整合教师音频和学生视觉信息来检测学生的课堂专注度状态。如图 1 所示，模型整体架构由长时上下文和短时上下文两个核心模块构成。其中长时上下文建模主要是学生单人与教师音频之间的关系，短时上下文建模主要是局部时间中多个学生专注度的变化。

在音视频编码模块中，教师音频和学生视频数据分别通过独立的编码器进行处理，生成教师的音频编码和学生的视觉编码。教师音频经过编码器后产生音频自注意力特征，而学生视频则生成视觉自注意力特征。

长时上下文模块专注于全局时间尺度上的特征提取，通过音频自注意机制提取教师的音频特征，视觉自注意机制提取学生在课堂中的长时行为特征。视听交叉注意机制结合教师的音频信息和学生的视觉信息，捕捉教师对学生专注度的影响。

短时上下文模块则聚焦于局部时间片段，利用内部卷积操作提取多人交互特征，以刻画课堂环境中多个学生专注度的动态变化。通过分析多个学生在短时间内的行为特征，捕捉学生间专注度的相互影响。最终，模型通过整合长时和短时的多模态特征，输出学生的专注度预测结果。

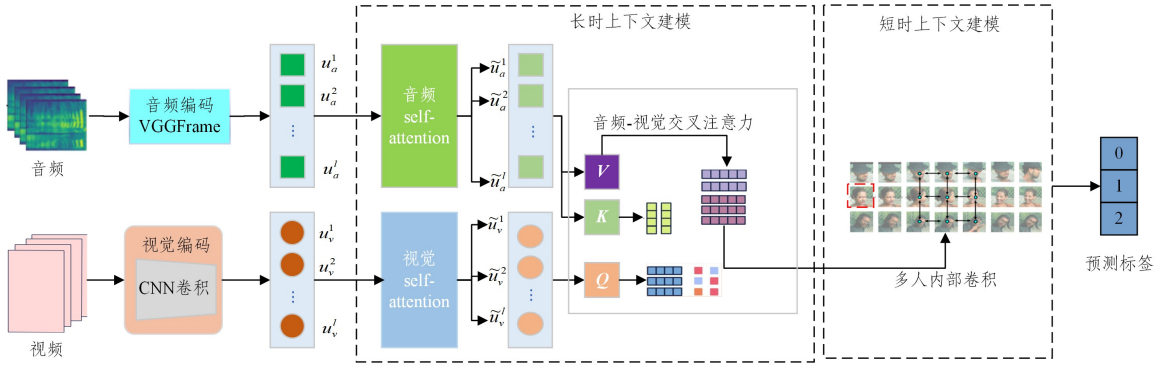

图 1 多模态多角色的长短时上下文学生课堂专注度评估模型

Fig. 1 Multimodal and multi-role model for assessing student classroom attention in long-short context modeling

4.1 音频和视觉编码

1)音频编码器

音频编码器以音频梅尔谱图 $\boldsymbol{A}\in\boldsymbol{R}^{4T\times M}$ 作为输入，其中 $4T$ 表示时间维度的长度，M 表示梅尔谱图的频率维度。该编码器的目标是提取帧级音频特征并实现时间维度的精细分类，尤其针对时间维度进行了高度下采样。为了更好地捕捉时间维度的层次结构特征，本文采用了一种改进的 VGGFrame 网络架构，如图 2 所示。标准的 VGGFrame 在经过多层卷积后，通常通过时间下采样层对时间维度进行压缩。然而，为了保留更多的时间分辨率信息并增强模型的时间敏感性，改进的 VGGFrame 做出了以下调整。

首先，移除 Block-4 后的时间下采样层，避免时间维度被过度压缩，从而保留音频特征的时间连续性；其次，引入反卷积层进行时间维度的上采样操作，通过反卷积操作对时间维度进行上采样，使得时间维度恢复到较高分辨率；最后，通过特征融合机制将上采样后的特征与中间层特征进行连接，实现了多层次特征的有机融合，增强了特征的层次化表示能力。

音频编码器最终输出一个时间序列的音频嵌入 $\boldsymbol{f}_a\in\boldsymbol{R}^{T\times C}$，其中 T 为上采样后的时间步数，C 为嵌入维度。

2)视觉编码器

视觉编码器的输入为每个学生的裁剪轨迹 $\boldsymbol{V}_i\in\boldsymbol{R}^{T\times H\times W\times 1}$，其中 $H\times W$ 是每帧裁剪图像的空间分辨率，T 是时间维度。视觉编码器的任务是将这些图像序列转换为时间序列的视觉嵌入。

通过卷积神经网络(CNN)对学生的裁剪图像序列进行

处理，提取出高维的视觉特征。每一帧图像的处理结果通过时间维度的堆叠形成了一个视觉嵌入时间序列 $\boldsymbol{f}_{vi} \in \boldsymbol{R}^{T\times C}$，其中 T 表示时间帧数，C 表示每一帧的特征维度。对于学生 i，最终的视觉上下文嵌入可以表示为 $\boldsymbol{f}_v \in \boldsymbol{R}^{T\times C}$。这一过程能够有效捕捉到学生在时间序列中的视觉变化，进而形成对该学生的时空特征的深入理解。然而，这种视觉编码器通常是独立于音频编码器工作的，导致视觉嵌入 $\boldsymbol{f}_v$ 和音频嵌入 $\boldsymbol{f}_a$ 之间缺乏信息交互。

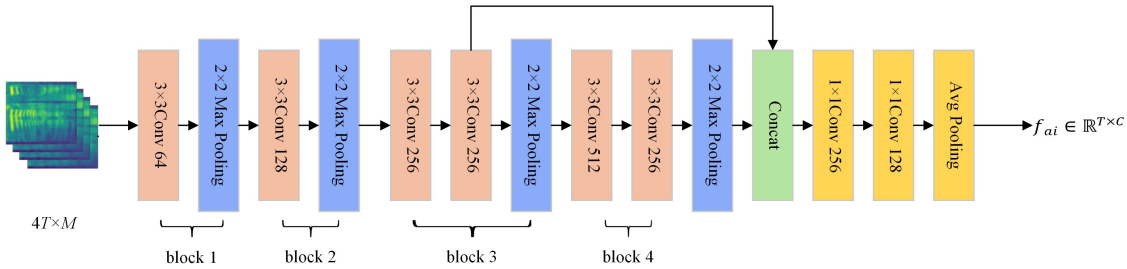

图 2 音频编码器 VGGFrame 网络结构

Fig. 2 Structure of audio encoder VGGFrame network

4.2 长短时上下文模型

为了解决教师音频和学生视觉嵌入之间的语义鸿沟，本文引入了多模态多角色的长短时上下文学生课堂专注度评估模型，用于捕捉教师音频和学生视觉信息之间的时间相关性和跨模态的语义交互。该模型由 N 个块组成，每个块连续包含一个长时上下文模块和短时上下文模块。通过交替输入音频嵌入 $\boldsymbol{f}_a$ 和视觉嵌入 $\boldsymbol{f}_v$，实现了音频与视觉信息的融合，进而生成上下文感知的嵌入表示。

具体而言，LSCM 模型首先接收教师音频和学生视觉的初始嵌入表示 $\boldsymbol{f}_a$ 和 $\boldsymbol{f}_v$，这两个嵌入分别代表了音频和视觉在时间序列上的特征。接着，LSCM 通过交替输入音频和视觉信息，利用其内在的长短时上下文结构来捕捉它们在时间维度上的相关性，并通过这种方式促进了模态之间的信息交换。通过多层次的交替建模，LSCM 模型不仅能够保留音频和视觉信息的时序性，还能够学习到它们之间潜在的跨模态语义联系。

在此过程中，音频和视觉嵌入分别通过 LSCM 模块生成上下文感知的音频嵌入 $\boldsymbol{u}_a^N \in \boldsymbol{R}^{S\times T\times C}$ 和视觉嵌入 $\boldsymbol{u}_v^N \in \boldsymbol{R}^{S\times T\times C}$，其中 S 表示批次大小，T 表示时间步长，C 表示特征维度。最终，这两个上下文感知的嵌入通过连接操作结合成一个完整的多模态嵌入 $\boldsymbol{u}^N = concat(\boldsymbol{u}_a^N, \boldsymbol{u}_v^N)$，作为多模态模型的输入。

通过 LSCM 模块中音频自注意机制和视觉自注意力机制以及视听交叉自注意力机制，不但解决了全局时间上的建模，而且解决了跨模态之间的语义交互存在鸿沟的问题，从而增强了模型对不同模态语义交互的理解。该方法对于多模态任务中，特别是需要同时处理音频和视觉信息的任务，提供了一种有效的解决方案。

4.2.1 长时上下文建模

为了有效捕捉学生的长期行为特征，并探索教师与学生之间的影响，本文设计了长时上下文模块。该模块由两个核心子模块组成：1）音频自注意力机制与视觉自注意力机制，用于在时间维度上对单个学生的行为进行精细化建模；2）视听交叉注意力机制，通过融合教师音频特征与学生视觉特征，学习两者之间的交互关系，从而捕捉师生互动对学生专注度的影响。

1）音频和视觉自注意力机制

为了捕捉个体行为的长期依赖关系，在时间维度上引入了 Transformer 自注意力机制。自注意力机制的核心思想是通过计算输入序列中各个位置之间的相关性，让每个时间步的特征能够直接关注到全局时间范围内的其他时间步，从而突破了传统 RNN 或 LSTM 的局部视野限制。每个时间步的特征会根据其与其他时间步的相似度进行加权，捕捉到全局范围内的重要信息，这使得模型能够理解和处理较长时间跨度内的行为模式。这种全局视野的特性使得 Transformer 在面对大容量、长时间序列时，能够显著提升建模效率。

$$\widetilde{\boldsymbol{u}}_v^l = LN(MHA(\boldsymbol{u}_v^{l-1}, \boldsymbol{u}_v^{l-1}, \boldsymbol{u}_v^{l-1}) + \boldsymbol{u}_v^{l-1}) \tag{1}$$

$$\widetilde{\boldsymbol{u}}_v^l = LN(MLP(\widetilde{\boldsymbol{u}}_v^l) + \widetilde{\boldsymbol{u}}_v^l) \tag{2}$$

其中，$\boldsymbol{u}_v^0 = \boldsymbol{f}_v$，$\boldsymbol{u}_v^{l-1} \in \boldsymbol{R}^{S\times T\times C}$ 是前一个 LSCM 模块的输出视觉嵌入；$LN(\cdot)$ 表示层归一化；$MHA(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v})$ 是多头专注度，$\boldsymbol{q}$ 是查询，$\boldsymbol{k}$ 是键，$\boldsymbol{v}$ 是值；MLP 是多层感知机。

$$\widetilde{\boldsymbol{u}}_a^l = LN(MHA(\boldsymbol{u}_a^{l-1}, \boldsymbol{u}_a^{l-1}, \boldsymbol{u}_a^{l-1}) + \boldsymbol{u}_a^{l-1}) \tag{3}$$

$$\widetilde{\boldsymbol{u}}_a^l = LN(MLP(\widetilde{\boldsymbol{u}}_a^l) + \widetilde{\boldsymbol{u}}_a^l) \tag{4}$$

其中，$\boldsymbol{u}_a^0 = \boldsymbol{f}_a$，$\boldsymbol{u}_a^{l-1} \in \boldsymbol{R}^{S\times T\times C}$ 是前一个 LSCM 模块的输出音频嵌入；$LN(\cdot)$ 表示层归一化；$MHA(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v})$ 是多头专注度，$\boldsymbol{q}$ 是查询，$\boldsymbol{k}$ 是键，$\boldsymbol{v}$ 是值；MLP 是多层感知机。

时间维度上的 Transformer 层直接作用于多模态模型的最终嵌入 $\boldsymbol{u}^N$，生成一个优化后的时间嵌入序列。这一优化后的时间序列能够更好地捕捉学生行为的时间特征，进而提升后续分类任务的性能。

2）视听交叉注意力

尽管自注意力机制在时间维度上能够有效建模长期依赖关系，但教师音频与学生视觉模态之间的交互仍然是多模态任务中的一个关键挑战。为了进一步优化音频与视觉信息的融合，本文引入了视听交叉注意力机制，加强教师音频和学生视觉嵌入之间的语义融合。这种机制的核心思想是通过互相作用的专注度机制，将音频和视觉模态的特征进行有意义的结合，使模型能够在理解语音语义的同时，关注到与之对应的视觉行为特征，从而增强多模态之间的交互。

在视听交叉注意力机制中，教师音频和学生视觉嵌入通过交替的注意力计算实现信息的深度交互。具体而言，对于音频嵌入序列中的每一个时间步，模型将利用视觉嵌入的时

间特征来计算专注度权重。这意味着音频特征不仅仅依赖于它自己的信息，还能够动态地关注到视觉模态中的相关特征。音频的语音信号与视觉信号通常具有紧密的语义关联。通过这种跨模态的专注度机制，音频的每一个时间步都可以根据视觉信息进行加权调整，从而在保留音频特征细节的同时，增强其与视觉模态的语义一致性。

视听交叉注意力机制主要用于融合不同模态的信息。首先，以视频特征作为查询(Query)，音频特征作为键(Key)和值(Value)，通过多头注意力机制计算视频特征对音频特征的注意力映射。然后，将注意力加权后的信息与原始视频特征进行残差连接并进行层归一化，以稳定训练。接着，融合后的特征通过前馈神经网络进行进一步处理，并再次经过残差连接和层归一化，增强特征表达能力。这样，视频模态特征$\boldsymbol{u}_v^{l-1}$通过交叉注意力机制关注音频模态特征$\boldsymbol{u}_a^{l-1}$，以获取有助于视频理解的关键信息。最终，计算得到的新视频特征$\boldsymbol{u}_v^l$会结合音频信息，实现多模态的深度融合。

$$\hat{\boldsymbol{u}}_v^l = LN(MHA(\tilde{\boldsymbol{u}}_v^l, \tilde{\boldsymbol{u}}_a^l, \tilde{\boldsymbol{u}}_a^l) + \tilde{\boldsymbol{u}}_v^l) \tag{5}$$

$$\hat{\boldsymbol{u}}_v^l = LN(MLP(\hat{\boldsymbol{u}}_v^l) + \hat{\boldsymbol{u}}_v^l) \tag{6}$$

其中，$\hat{\boldsymbol{u}}_v^l$是音频增强的视觉嵌入。通过视听交叉注意力机制，模型能够在音频信号与视觉信号之间实现深度的信息交互，极大地增强了模型对语音和行为的理解能力。在教师的语音和学生的行为响应之间，语音信号传递着语义信息，而视觉信号则能够提供与语音信号相对应的非语言信息。视听交叉注意力机制能够有效地将这两种模态的特征融合，从而帮助模型更准确地理解和预测情感、意图等复杂的多模态任务。

此外，这种机制通过交替地关注音频和视觉信号，提高了模型的理解能力。教师的音频和学生的视觉反应之间的联系可以通过交叉注意力机制进行有效建模，使得模型在对学生行为的预测中，能够结合教师的语音语义和学生的行为等视觉信息，从而提升对教学互动过程的全面理解。

4.2.2 短时上下文建模

对于视频中的给定时刻，不同学生之间的专注度状态在较近帧内通常具有更强的关联性，因此模型需要有效地捕捉局部时间范围内的学生与学生之间专注度的关系。这种关系反映了学生在短时间内的交互模式，如共同注视某一目标、同步的行为反应或交替的专注度转移。为此，本文引入了一种跨学生卷积网络，负责捕捉短时间内不同学生的专注度状态。

$$\boldsymbol{u}_v^l = MLP(LN(Conv_{s\times k}(\hat{\boldsymbol{u}}_v^l))) + \hat{\boldsymbol{u}}_v^l \tag{7}$$

$$\boldsymbol{u}_a^l = MLP(LN(Conv_{s\times k}(\hat{\boldsymbol{u}}_a^l))) + \hat{\boldsymbol{u}}_a^l \tag{8}$$

视觉嵌入$\boldsymbol{u}_v^l \in \boldsymbol{R}^{S\times T\times C}$和音频嵌入$\boldsymbol{u}_a^l \in \boldsymbol{R}^{S\times T\times C}$是第$l$个LSCM模块的输出，并将传递到下一个模块。其中，k表示感受野的时间长度，s表示视频中学生的数量。短时上下文模块具有较短的时间感受野，能够捕捉交互中的局部动态模式。在整体框架中，$\boldsymbol{u}_v^l$视觉特征和$\boldsymbol{u}_a^l$音频特征会继续传递到下一个LSCM模块进行进一步建模，经过N层LSCM处理后，得到最终的音频和视觉嵌入$\boldsymbol{u}_N^v$和$\boldsymbol{u}_N^a$。在分类阶段，LSCM采用基于音视频联合特征的分类策略。最终$\boldsymbol{u}_N^v$和$\boldsymbol{u}_N^a$进行拼接融合，形成最终的特征表示 $\boldsymbol{outsAV}$。

该网络的设计是通过捕捉相邻帧之间的动态变化，提取短时间窗口内的关键特征，从而增强对交互模式的理解。与全局建模不同，这种方法聚焦于细粒度的局部时序特性，能够更准确地反映学生之间的实时专注度波动。此外，网络还能够根据时间上下文动态调整特征权重，以突出具有代表性的局部模式。

5 实验设计

5.1 实验参数

在模型的超参数设置中，首先考虑设置 batch 的大小。在此，将 batch 的大小设置为 2，表示每次更新模型参数时，计算的是 2 个样本的平均梯度。这在保证训练稳定性的同时，能够有效减少内存消耗。

同时，本文使用 Adam 优化器对模型进行训练，设置批量大小为 2，学习率为 5×10^{-5}，并进行了 30 个 epoch 的训练。在训练过程中，学习率每个 epoch 减少 5%，以帮助模型在训练的后期更精细地调整参数。Adam 优化器是一个常用的自适应学习率优化器，通过结合动量和自适应学习率调整，能够加速收敛并稳定训练过程。

在训练过程中，随着 epoch 的增加，学习率逐渐减小，这样可以帮助模型在接近最优解时减少过大的参数更新，从而避免训练过程中的振荡，进一步提高模型的精度和泛化能力。学习率逐步减少的方法，有助于在训练初期快速收敛，而在训练后期使模型实现更加细致地调整，以获得更精确的预测结果。

总体来说，结合 Adam 优化器和逐步降低学习率的策略，可以有效提高模型的训练效率与性能，避免过拟合，并使模型在处理复杂任务时具有更好的鲁棒性。通过 30 个 epoch 的训练，模型能够逐步学习到数据中的深层次特征，并最终在测试集上取得较好的表现。

5.2 评价指标

本文以平均精度均值(Mean Average Precision，mAP)作为核心评价指标，全面衡量模型在多类别分类任务中的预测性能，体现不同类别下的精确率与召回率的平衡性，并得到$p(r)$。给定真实标签，可以计算 TP(真阳性)、FP(假阳性)、FN(假阴性)。mAP 通过计算各类别平均精度的均值，反映模型对正例样本的识别能力。平均精度通过比较模型的分类预测与实际标签，计算出 AP，以全面评估模型的分类准确性。这一综合指标能够有效评估模型在不同场景下的分类性能表现。

平均精度均值用于多类别目标检测或分类任务，是每个类别的平均精度(AP)的算术平均。

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$AP = \int_0^1 p(r)\mathrm{d}r \tag{11}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{12}$$

其中，$p(r)$为精确率与召回率的函数关系曲线，N是总类别数。

5.3 实验结果与分析

5.3.1 对比实验

本节对 LSCM 模型在多模态专注度分析任务中的实验结果进行详细分析。通过对比各模型的 mAP 指标，验证 LSCM 在多模态特征融合、长时上下文建模和短时上下文建模方面的优势。实验选择了 C3D，R(2+1)D，R3D，C3D+TCN，LRCN，VTN 等作为基线模型。

C3D(3D Convolutional Network)[27]是一种处理单模态的经典 3D 卷积网络，用于从视频数据中提取时空特征。C3D 通过 3D 卷积核同时提取空间和时间特征，能够有效处理视频数据中的视觉信息，保留视频中的动态变化。通过沿着时间轴的卷积运算，捕获视频帧之间的时序特征，从而保留视频中的动态信息。其多层 3D 卷积和池化结构使得网络能够从原始视频数据中学习到丰富的时空特征，最终通过全连接层进行分类。

LRCN(Long-term Recurrent Convolutional Network)[28]是一种基于卷积神经网络(CNN)和长短期记忆网络(LSTM)的单模态混合架构。LRCN 的核心思想是将视频帧视为句子中的单词，通过 CNN 提取视频帧的空间视觉特征，然后将这些特征输入 LSTM 中，以学习帧之间的时序关系，从而同时捕捉视频中的静态空间信息和动态时序信息。

R(2+1)D(Residual(2+1)D Network)[29]是一种基于视频单模态改进的 3D-CNN 结构。每个 3D 卷积操作被分解为一个用于捕获视频中视觉空间特征的 2D 卷积和一个用于提取时间特征，并通过残差连接来增强特征流通的 1D 卷积。这种分解降低了计算复杂度，同时保持了模型的时空特征提取能力。

R3D(Residual 3D Convolutional Network)[30]是一种基于视频单模态的 3D 残差卷积网络，通过 3D 卷积操作同时提取视频中的空间特征和时间特征。R3D 由多个残差块组成，每个块包含两个 3D 卷积层和一个短路连接。短路连接使得每个块的输入能够直接流入后续层，增强了信息流动。

C3D+TCN[31]是一种基于视频的单模态模型，C3D 模块作为前端，主要负责处理视频帧序列中的视觉信息，通过对视频帧序列进行 3D 卷积操作，能够同时提取空间和时间维度上的局部特征。提取的这些局部时空特征随后被传递到 TCN(Temporal Convolutional Network)模块。TCN 模块通过一维卷积操作在时间维度上对特征进行全局建模，能够捕捉视频中较长时间跨度内的上下文信息。

VTN(Vision Transformer Network)[32]是一种基于 Transformer 架构的单模态深度学习模型，用于处理视觉信息，尤其是视频数据。VTN 利用 Transformer 的自注意力机制(Self-Attention Mechanism)将视频帧序列视为一个时间序列，同时捕捉视频中的空间信息和时间信息，并通过自注意力机制捕捉帧与帧之间的全局依赖关系，进一步增强长距离的时间依赖关系，从而增强模型对视频内容的理解能力。

MViTV2(Multiscale Vision Transformers)[33]是一种单模态优化的多尺度视觉 Transformer 模型，它通过引入多尺度结构，使得模型能够在不同层次上提取视觉特征，从而同时捕捉到图像的局部细节和全局上下文信息。MViTv2 采用了多尺度的注意力机制，其中每个尺度的卷积特征映射通过特定的网络模块进行融合。这种设计使得模型能够在不同的分辨率下处理特征，从而更加高效地学习到不同层次的语义信息。

Swin-T(Swin Transformer)[34]也是一种单模态处理视觉信息的 Transformer 架构。它采用了窗口化自注意力机制(Window-based Attention)，将图像分割成局部窗口进行自注意力计算，通过窗口的平移操作(Shifted Windows)来捕捉跨窗口的信息，并采用了层级化的结构，使得低层的特征图较为粗糙，而高层则捕捉更为抽象的高维语义信息。

EfficientNetB7-BiLSTM[35]是一种结合了卷积神经网络和双向长短期记忆网络的单模态混合模型。EfficientNetB7 通过其深层卷积层提取图像的多层次视觉特征，之后，提取的特征序列被送入 BiLSTM 网络，BiLSTM 通过其双向结构捕捉特征序列中的时序依赖关系，因此其在处理序列数据，如视频帧序列时，具有优势。

TgGTM(Text-guided Graph Temporal Modeling)[21]是一种新的用于小样本视频分类的多模态深度学习模型，旨在更好地利用相邻视频帧中隐含的多模态知识和时间线索。其核心思想是通过将类别标签作为文本语义引导与时空图的深度融合，解决视频数据在有限标注样本下的时空特征建模难题。它促进了任务内视频特征之间多模态语义知识的交互式传递，增强了未知视频的先验知识。此外，通过 Temporal masking layer 关注相邻帧之间的时间关系，以实现全面的小样本视频分类。

AVSlowFast(Audiovisual SlowFast)模型[36]是一种多模态时空特征学习框架，通过协同整合视觉与听觉模态信息来增强视频内容理解能力。该架构在经典 SlowFast 双路径视觉网络的基础上，引入了音频处理分支，构建起三元异构特征学习体系。模型通过跨模态融合机制实现多源信息互补，包括层级特征拼接、三维时空卷积融合等方法。通过视听模态的协同学习，模型在复杂场景下的语义理解鲁棒性得到了显著提升。

LSCM(长短时上下文模型)是本文提出的多模态模型，其通过捕捉音频和视觉信息之间的时间相关性和跨模态的语义交互，解决教师音频和学生视觉嵌入之间的语义鸿沟。模型包括长时与短时上下文模块，长时模块通过捕捉学生长期行为特征提供全局视角，短时模块则聚焦短期变化，确保精准评估学生专注度状态。

LSCM 模型通过交叉注意力机制融合了教师音频和学生视频模态的特征，从而捕捉模态间的交互关系。如表 1 所列，从实验结果可以看出，LSCM 的 mAP 达到了 87.33%，显著高于其他基线模型。相比之下，C3D，R(2+1)D 和 R3D 等仅依赖单一模态的模型，其 mAP 分别为 50.23%，51.82%和 52.53%。这表明单一模态的特征提取虽然能够捕捉到一定的行为模式，但在复杂的专注度分析任务中，其表征能力有限。

进一步分析，C3D+TCN 通过引入时间卷积网络(TCN)

对时间序列进行建模，mAP 提升至 56.48%，说明时间序列建模能够在一定程度上增强模型对时序信息的捕捉能力，但仍未突破单一模态的局限。EfficientNetB7-BiLSTM 通过双向 LSTM 进一步捕捉时序依赖，mAP 达到 65.19%，显著高于其他单模态模型，表明时序建模方法对性能提升有积极作用，同时也表明了单纯的时间序列建模虽然能够捕捉到一定的时序信息，但仍无法与多模态模型相比。

表 1 对比实验结果

Table 1 Results of comparative experiments

模型类别	模型	*mAP*/%
单模态	C3D	50.23
	LRCN	62.91
	R(2+1)D	51.82
	R3D	52.53
	C3D+TCN	56.48
	VTN	68.73
	MViTV2	70.45
	Swin-T	73.28
	EfficientNetB7-BiLSTM	65.19
多模态(文本+视频	TgGTM	79.89
多模态(音频+视频)	AVSlowFast	75.92
多模态(音频+视频	LSCM	87.33

从实验结果可以看出，LSCM 的 mAP 达到了 87.33%，显著高于 EfficientNetB7-BiLSTM 的 65.19%。VTN，MViTV2 和 Swin-T 都是通过引入视觉 Transformer 来进行捕捉，其 mAP 分别为 68.73%，70.45%，73.28%。LSCM 中的长时上下文模块，通过自注意力和时间序列建模方法捕捉长时间范围内的视觉特征，比其他模型有更好的表现。MViTV2 利用 Transformer 的自注意力机制，能够捕捉视频中的长时依赖关系，而其主要依赖视频模态，缺乏对音频等多模态信息的融合，这可能是其 mAP 低于 LSCM 的主要原因。然而，这些模型仍然仅依赖视频模态，缺乏对音频等多模态信息的融合，因此其存在性能上限。LSCM 通过交叉注意力机制，能够有效捕捉音频和视频之间的依赖关系，从而在专注度分析任务中取得了更好的表现。

从实验结果可以看出，虽然 Swin-T 通过分层的窗口注意力机制，能够高效地捕捉视频中的局部和全局时空特征。但是在 mAP 评价指标中，LSCM 比 Swin-T 的 mAP 高出了 14.05个百分点，表明了在短时上下文建模和细化局部时间片段内学生的专注度状态方面，LSCM 比 Swin-T 更好。

实验数据明确证实了多模态融合策略在视频理解任务中的有效性。TgGTM 模型通过视频-文本标签实现跨模态融合，实现了 79.89%的 mAP，较基准单模态模型提升显著。由于 LSCM 采用交叉注意力机制与专注度强相关的视听特征，而 TgGTM 的文本引导仅能提供静态标签，缺乏细粒度模态交互。同时，LSCM 的短时上下文模块聚焦瞬时状态变化，避免 TgGTM 因过度依赖文本标签(固定行为分类)而忽略动态细节。LSCM 的 mAP 较 TgGTM 提高了 7.44 个百分点。

LSCM 中音频与视觉交叉注意力将教师音频的关键时间与学生视觉动态对齐，而 AVSlowFast 的简单特征拼接无法处理模态间的时间延迟问题。LSCM 比 AVSlowFast 在 mAP 上高出 11.41 个百分点，表明了 LSCM 中的交叉注意力机制中深度融合教师音频和学生视频模态特征的有效性。同时，LSCM 的 mAP 达到 87.33%，是所有模型中表现最优的。这不仅证明了多模态融合的重要性，还表明了 LSCM 在长时和短时上下文建模方面的创新设计能够更有效地捕捉时间依赖性和局部细节。

5.3.2 消融实验

表 2 列出了针对音频编码器(ResNet-34 或 VGG-Frame)、长时上下文建模(LIM)和短时上下文建模(SIM)的消融实验结果。通过逐步去除或替换模型的不同组件，分析了各个模块对模型性能的贡献。实验结果表明，各个组件的组合对模型的性能提升具有显著影响。

表 2 消融实验结果

Table 2 Results of ablation experiments

R-34	VGGFrame	LIM	SIM	*mAP*/%
√	×	×	×	72.76
×	√	×	×	74.58
×	√	×	√	86.45
×	√	√	×	86.61
×	√	√	√	87.33

首先，ResNet-34(R-34)作为音频编码器之一，当独立使用时，模型的 mAP 达到了 72.76%。这表明，仅使用 ResNet-34 作为音频特征提取器能够为模型提供基础的音频特征，但显然仍有进一步提升的空间。

其次，采用 VGGFrame 替代 ResNet-34 后，模型的 mAP 提升至 74.58%。这表明 VGGFrame 在捕捉音频信息方面的表现优于 ResNet-34，这可能是因为其在视觉和音频特征的提取上具有更强的全局建模能力，尤其是在音频特征的空间结构上。尽管如此，仅有 VGGFrame 编码器的情况下，模型的时序建模仍未得到充分利用，导致模型性能未达到最佳水平。

当 VGGFrame 与短时上下文建模(SIM) 结合时，模型的 mAP 达到了 86.45%，较前一设置进一步提升。短时上下文建模主要捕捉短时相邻帧之间的动态变化，实验结果表明，短时建模对专注度评估至关重要。

当 VGGFrame 与长时上下文建模(LIM)结合时，模型的 mAP 再次提升，达到了 86.61%。这表明，长时上下文建模能够帮助模型捕捉更长时间跨度内的依赖关系，且解决了音频与视觉之间的语义鸿沟，同时也表明了通过自注意力和时间序列建模方法捕捉长时间范围内的特征的有效性。

结合了 VGGFrame，LIM 和 SIM 的完整模型表现最佳，mAP 达到了 87.33%。此结果表明，长短时上下文建模的结合对教师音频和学生视觉信息的有效融合至关重要。通过短时建模捕捉细粒度的时序变化，同时通过长时建模捕捉长期依赖，模型能够更全面地理解音频和视觉信息之间的时空关联。

5.3.3 学生专注度识别结果

为了评估 LSCM 模型在不同专注度水平上的分类性能，在测试集的数据中进行了验证。测试集共有 1781 个，其中标签为 0 的有 492 个，标签为 1 的有 95 个，标签为 2 的有 1194 个。分析基于多模态融合的专注度预测模型在各类别上的表现差异，以验证模型的有效性。对专注度的 3 个类别

进行了识别，识别结果如表3所列。

表3 学生专注度识别结果

Table 3 Results of student concentration recognition

(%)

专注度水平	P	R	F1
0	87.52	50.53	64.07
1	70.12	44.54	63.28
2	92.54	84.78	88.49

实验结果表明，模型在不同类别上的表现存在显著差异，其中 Label 2($F1=88.49\%$)的预测效果最优，而 Label 0($F1=64.07\%$)和 Label 1($F1=63.28\%$)的表现相对较差。这一差异可能与数据集的类别分布不均衡以及不同类别特征的表征难度有关。具体而言，Label 2 的样本数量远多于 Label 0 和 Label 1，导致模型在训练过程中更倾向于学习 Label 2 的特征，从而在该类别上表现出较高的精确率(92.54%)和召回率(84.78%)。相比之下，Label 0 和 Label 1 的召回率较低(分别为 50.53%和 44.54%)，表明模型在这两类样本中存在较多的漏判现象。不专注的特征与专注的特征有较高的相似性，因为在不专注视频中，学生的身体或头部频繁转动，以及出现身体静止时打瞌睡的现象，与专注的特征在身体姿态上很难区分，导致模型识别时混淆。

从多模态融合的角度来看，模型通过音频和视频的自注意力机制以及交叉注意力机制，有效地捕捉了两种模态的互补信息，从而提升了整体的预测性能。特别是在 Label 2 上，音频和视频特征的协同作用显著增强了模型对专注度状态的动态评估能力。然而，对于 Label 0 和 Label 1，模型的表现相对较弱，可能是由于这两类样本的专注度特征在音频和视频模态上的表现不够显著，或者交叉注意力机制在融合过程中未能充分提取关键信息。此外，长时上下文模块虽然能够捕捉时间序列上的依赖关系，但在处理少数类样本时，其效果受到数据量的限制；而短时上下文模块虽然关注局部时间片段内的特征，但对于 Label 0 和 Label 1 这类专注度变化较为细微的类别，其细化能力较弱。

结束语 本文提出的基于多模态多角色建模的长短时上下文学生课堂专注度评估模型(LSCM)，针对教学课堂中多模态和多角色相互影响的复杂场景，提供了有效的解决方案。通过构建包含教师音频和学生视频的专注度数据集，LSCM 能够有效融合学生的视觉信息与教师的音频信息，并挖掘学生与学生、学生与教师之间的相互影响。在模块设计中，长时上下文模块通过捕捉学生长期行为特征提供全局视角，短时上下文模块则专注于局部时间片段的专注度动态变化。实验结果表明，LSCM 通过多模态数据的互补性和角色间关联性的挖掘，显著提升了课堂专注度检测的准确性。综上所述，本文模型为智慧教育中的课堂专注度评估提供了新的方法，不仅能够为教师和学生提供及时反馈，优化教学效果，还能为教育管理者提供有力的数据支持，助力提升整体教育质量。

然而，本文的研究仍存在一些局限性，需要在未来的工作中进一步改进和探索。

1)数据规模与多样性：当前构建的数据集规模有限，未来可以扩展数据集的规模和多样性，涵盖更多教学场景和不同学科，以提升模型的泛化能力。

2)多模态信息融合的深度：尽管通过音频-视觉交叉注意力机制实现了多模态信息的融合，但在更深层次的语义理解和跨模态对齐方面仍有提升空间。未来可以探索更先进的多模态融合技术，如图像-文本-音频的联合建模，以进一步提升模型性能。

参 考 文 献

[1] ZHONG M C, ZHANG J L, LAN Y B, et al. Study on Online Education Focus Degree Based on Face Detection and Fuzzy Comprehensive Evaluation[J]. Computer Science, 2020, 47(S2): 196-203.

[2] ZALETELJ J, KOSIR A. Predicting Students' Attention in the Classroom from Kinect Facial and Body Features[J]. EURASIP Journal on Image and Video Processing, 2017, 2017: 80.

[3] DUAN J L. Evaluation and Evaluation System of Students' Attentiveness Based on Machine Vision[D]. Hangzhou: Zhejiang Gongshang University, 2018.

[4] ZUO G C, WANG H D, CHEN L S, et al. Evaluation of Modern Apprenticeship Learning Effect Based on Face Recognition Technology[J]. Intelligent Computer and Applications, 2019, 9(2): 116-118.

[5] HE X L, GAO Q, LI Y Y, et al. Spontaneous Learning Facial Expression Recognition Based on Deep Learning[J]. Computer Applications and Software, 2019, 36(3): 180-186.

[6] WANG Y K, SUN Y J, PU D B, et al. Multi modal based online learning focus evaluation [J]. Journal of Changchun Normal University, 2024, 43(8): 59-66.

[7] SINATRA G M, HEDDY B C, LOMBARDI D. The challenges of defining and measuring student engagement in science [J]. Educational psychologist, 2015, 50(1): 1-13.

[8] TYLER R W. Basic Principles of Curriculum and Instruction [M]. Chicago: University of Chicago Press, 1949: 1-128.

[9] PACE C R. Measuring the Quality of Student Effort [J]. Current Issues in Higher Education, 1980, 2(3): 10-16.

[10] NSSE. Nsse: Evidence-based improvement in highereducation [EB/OL]. https://nsse.indiana.edu/nsse/about-nsse/index.html.

[11] KAUR A, MUSTAFA A, MEHTA L, et al. Prediction and localization of student engagement in the wild[C]// 2018 Digital Image Computing: Techniques and Applications(DICTA). 2018: 1-8.

[12] MOHAMAD N O, DRAS M, HAMEY L, et al. Automatic recognition of student engagementusing deep learning and facial expression[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2019: 273-289.

[13] BATRA S, WANG H, NAG A, et al. Dmcnet: Diversified model combination network for understanding engagement from video screengrabs[J]. Systems and Soft Computing, 2022, 4: 200039.

[14] WHITEHILL J, SERPELL Z, LIN Y C, et al. The faces of engagement: Automatic recognition of student engagementfrom facial expressions[J]. IEEE Transactions on Affective Compu-

ting,2014,5(1):86-98.
[15] SUKUMARAN A,MANOHARAN A. Multimodal engagement recognition from image traits using deep learning techniques[J]. IEEE Access,2024,12:25228-25244.
[16] SANTONI M M,BASARUDDIN T,JUNNS K,et al. Automatic detection of students'engagement during online learning:A bagging ensemble deep learning approach[J]. IEEE Access,2024,12:96063-96073.
[17] CHEN Y,ZHOU J,GAO Q,et al. Mdnn:Predicting student engagement via gaze direction and facial expression in collaborative learning[J]. Computer Modeling in Engineering & Sciences,2023,136(1):381-401.
[18] BUONO P,DE C B,D'ERRICO F,et al. Assessing student engagement from eacialbehavior in on-line learning[J]. Multimedia Tools and Applications,2023,82(9):12859-12877.
[19] IKRAM S,AHMAD H,MAHMOOD N,et al. Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm[J]. Applied Sciences,2023,13(15):8637.
[20] LAI S,WU F T. Recognition of Learning Concentration Based on Multimodal Physiological Signals[J]. Modern Educational Technology,2023,33(6):101-108.
[21] DENG F Q,ZHONG J M,LI N N,et al. Text-guided Graph Temporal Modeling for few-shot video classification[J]. Engineering Applications of Artificial Intelligence,2024,137:109076.
[22] ABEDI A,KHAN S S. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network[C]// 2021 18th Conference on Robots and Vision(CRV). 2021:151-157.
[23] DAS R,DEV S. Enhancing frame-level student engagement classification through knowledge trans fer techniques[J]. Applied Intelligence,2024,54(2):2261-2276.
[24] HERNANDEZ J,LIU Z,HULTEN G,et al. Measuring the engagement level of tv viewers[C]// 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition(FG). IEEE,2013:1-7.
[25] GUPTA A,D'CUNHA A,AWASTHI K,et al. Daisee:Towards user engagement recognition in the wild[J]. arXiv:1609.01885,2016.
[26] ZHU X,LYU S,WANG X,et al. Tph-yolov5:Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:2778-2788.
[27] TAND 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.
[28] DONAHUE J,ANNE H L,GUADARRAMA S,et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:2625-2634.
[29] QIU Z,YAO T,MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5533-5541.
[30] XU H,DAS A,SAENKO K. R-c3d:Region convolutional 3d network for temporal activity detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5783-5792.
[31] ABEDI A,KHAN S S. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network[C]// 2021 18th Conference on Robots and Vision(CRV). IEEE,2021:151-157.
[32] NEIMARK D,BAR O,ZOHAR M,et al. Video transformer network[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:3163-3172.
[33] 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.
[34] 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.
[35] YOUSAF K,NAWAZ T,HABIB A. Using two-stream efficientnet-bilstm network for multiclass classification of disturbing youtube videos[J]. Multimedia Tools and Applications,2024,83(12):36519-36546.
[36] XIAO F,LEEY J,GRAUMAN K,et al. Audiovisual slowfast networks for video recognition[J]. arXiv:2001.08740,2020.

ZHUO Tienong, born in 1995, master. His main research interest is digital image processing.

ZHUO Hui, born in 1972, Ph.D, professor, Ph.D supervisor, is a member of CCF(No. 25440S). Her main research interests include artificial intelligence, natural language processing, emotion computing, speech and digital image processing.

(责任编辑:喻藜)