分类号：TP391　　学校代码：10363

密　级：公开　　学　　号：2210910105

安徽工程大学

Anhui Polytechnic University

硕士学位论文

题目　基于视频的学生课堂状况感知与识别研究

论 文 作 者　王岱嵘

指 导 教 师　陈新泉 教授 刘进军 教授

学科（专业）　软件工程

研 究 方 向　领域软件工程

论文提交日期：2024 年 5 月 23 日

分类号：TP391　　单位代码：10363

密　级：公开　　学号：2210910105

题　　目 基于视频的学生课堂状况感知与识别研究

英文并列

题　　目 Research on Perception and Recognition of Students' Classroom Situation Based on Video

学　生　姓　名：王岱嵘

指　导　教　师：陈新泉　教授/刘进军　教授

专　　业：软件工程

研　究　方　向：领域软件工程

论文答辩日期：2024 年 5 月 24 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：王斌鹏

日期：　2024 年 5 月 23 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在_____年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：

日期： 2024 年 5 月 23 日

指导教师签名：

日期： 2024 年 5 月 23 日

基于视频的学生课堂状况感知与识别研究

摘 要

教育的内在要求是增强学生对知识的学习效果。对学生学习效果最好的评价指标就是学生在课堂上的投入程度。所以，对学生课堂状况的评估在教育领域起着至关重要的作用。在传统课堂上，对学生课堂状况的评估依赖于督导工作组的督导老师。但是，这种方式不仅耗费人力，而且缺乏数据支持，还容易使被监督课堂弄虚作假。近年来，随着深度学习的发展，研究者们在学生课堂状况感知与识别领域取得了一定的成果，但依然存在许多问题，如缺少公开的学生行为数据集、划分学生行为类别不合理、算法精度低、缺乏可靠的课堂质量评估算法等。针对上述问题，本文的主要工作内容如下：

第一，针对数据的稀缺及划分学生行为类别不合理的问题，本文构建了双粒度标注的学生行为（Student Action，SA）数据集。

本文通过采集学生志愿者们模拟课堂教学实验的视频数据及收集网络上单人、多人的真实课堂行为数据，通过数据处理及数据增强方法将 SA 数据集使用两种标注方式，其一为细粒度的 9 种：写，站，睡觉，读，举手，玩手机，正常听课，低头，左顾右盼；其二为粗粒度的 3 种：未参与（包含睡觉、玩手机、左顾右盼、低头），参与（包含正常听课、读、写），互动（包含站、举手）。

第二，针对算法精度低的问题，本文提出了 SGE-BI（Spatial Group

wise Enhance-BiFormer）双注意力机制模块并将其成功应用在YOLOv5 算法中。

本文介绍了卷积神经网络及目标检测领域的相关研究成果，简要分析了目前基于深度学习模型的缺点，对注意力机制的相关理论进行了叙述，并设计了目标检测领域的多种流行算法在 SA 数据集上的对比实验，确定了采用 YOLOv5 算法为本文的主干网络。在上述实验的基础上，本文设计了 SGE-BI 双注意力机制模块，并将其添加在YOLOv5 算法中，通过设计消融实验，证明了该模块能够明显提高原算法的性能。此外，本文设计了一组对比实验，通过在 YOLOv5 上堆叠该模块，证明了在主干网络上添加两个 SGE-BI 模块较为合适。

第三，针对缺乏可靠课堂质量评估算法的问题，本文结合双粒度标注的学生行为数据集对传统层次分析法（Analytic Hierarchy Process，AHP）进行改进，提出了双粒度层次分析加权法（Analytic Hierarchy Process-Double Granularity weighted，AHP-DG）。

本文进行了数据分析并验证了使用传统 AHP 法会对学生课堂的评价有倾向性，进而产生不客观的结论。针对传统 AHP 法的上述缺陷，本文利用 SA 数据集的双粒度标注，将其改进为加权平均，以拟合在现实学生课堂评价中互动性及其他课堂行为占比多少较合适的线性函数。最后，本文采集了多个真实课堂的视频，并邀请了多名从事教育工作的专家对这些课堂进行评分，通过对比专家和改进后的算法评分，证明了改进后的算法可以调整评价的主观性，从而较之传统AHP 法更合理。

关键词：课堂状况评估，目标检测，深度学习，卷积神经网络，注意力机制，层次分析法

RESEARCH ON PERCEPTION AND RECOGNITION OF STUDENTS' CLASSROOM SITUATION BASED ON VIDEO

ABSTRACT

The inherent requirement of education is to enhance the learning effectiveness of students towards knowledge. The best evaluation indicator for student learning effectiveness is their level of engagement in the classroom. Therefore, the evaluation of student classroom conditions plays a crucial role in the field of education. In traditional classrooms, the evaluation of student classroom conditions relies on the supervising teachers of the supervisory working group. This approach not only consumes manpower, but also lacks data support, making it easy for supervised classrooms to engage in fraud. In recent years, with the development of deep learning, researchers have achieved certain results in the field of student classroom situation perception and recognition. However, there are still many problems, such as the lack of publicly available student behavior datasets, unreasonable classification of student behavior categories, low algorithm accuracy, and a lack of reliable classroom situation evaluation algorithms. In response to the above issues, the main work of this thesis is as follows:

First, addressing the scarcity of data and the unreasonable

categorization of student behavior, constructed a dual granularity annotated Student Action dataset.

We collected video data of student volunteers simulating classroom teaching, collected real classroom behavior data of single and multiple people on the internet, and used data processing and data augmentation methods to annotate the SA dataset in two ways. One is the fine-grained nine methods: write, stand, sleep, read, raise hands, play with mobile phones, listen normally, bow your head, and look left and right; The second consists of three coarse-grained categories: non participation (including sleeping, playing with the phone, looking left and right, and bowing), participation (including normal listening, reading, and writing), and interaction (including standing and raising hands).

Second, addressing the issue of low algorithm accuracy, proposed the SGE-BI (Spatial Group wise Enhance BiFormer) dual attention mechanism module and successfully applied it in the YOLOv5 algorithm.

This thesis introduces the relevant research achievements in the field of convolutional neural networks and object detection, briefly analyzes the shortcomings of current deep learning models, and describes the relevant theories of attention mechanisms. We have designed comparative experiments on the SA dataset using various popular algorithms that include the field of object detection and apply the Transformers method, and determined that the YOLOv5 algorithm is the backbone network of

this paper. On the basis of the above experiment, we designed the SGE-BI dual attention mechanism module and added it to the YOLOv5 network. By designing ablation experiments, we have demonstrated that this module can significantly improve the performance of the original algorithm. In addition, we designed a comparative experiment to demonstrate that adding two SGE-BI modules on the backbone network is more appropriate by stacking the module on YOLOv5.

Third, addressing the issue of lacking reliable classroom quality assessment algorithms, combining the dual granularity annotated student behavior dataset, the traditional Analytic Hierarchy Process (AHP) is improved, and the Analytic Hierarchy Process Double Granularity Weighted (AHP-DG) is proposed.

We conducted data analysis and verified that the use of traditional AHP method tends to bias the evaluation of students in the classroom, leading to non objective conclusions. In response to the aforementioned shortcomings of the traditional AHP method, we utilize the dual granularity annotation of the SA dataset and improve it to a weighted average to fit a linear function that is more appropriate for the proportion of interactivity and other classroom behaviors in real-life student classroom evaluations. Finally, we collected videos of multiple real classrooms and invited multiple experts in education to rate these classrooms. By comparing the ratings of the experts and the improved algorithm, we proved that the

improved algorithm can adjust the subjectivity of the evaluation, making it more reasonable than the traditional AHP method.

KEY WORDS: Classroom condition evaluation, Object detection, Deep learning, Convolutional neural network, Attention mechanism, Analytic Hierarchy Process

目录

第1章　绪论

1.1 研究背景及意义

对学生课堂状况的监督及评价是教学工作的重要一环。R. Taylor[1]在 20 世纪 30 年代就引入了“任务时间”（time on task）这一理念，主张学生在学习上花费的时间越长，获得的知识也越多。继 Taylor 之后，Pace 等人[2]进一步提出了“努力质量”（Quality of effort）这个观点，该研究显示，评价学生学习效果时，不仅要考虑学习的时长，还应当考虑他们学习时的专注程度。George[3]的研究结果指出，衡量学校教学质量的优劣，学生在课堂上的专注度是一个关键指标。在这些研究的基础上，Tinto、Chickering 等学者[4][5]还指出，影响学生学习效果的要素中，还包括了师生间的互动、教师反馈等因素。对学生课堂状况进行监督及评价不仅可以及时掌握学生的学习状态，了解学生对教师教授知识的接收程度，同时也能够督促教师针对课堂状况及时调整教学方法。

在传统课堂上，对学生课堂状况的监督及评价依赖教学督导工作组的督导老师。但是，普遍情况下，督导老师对学生课堂状况的监督及评价存在如下难以克服的缺陷：首先，课堂和学生的数量众多，督导老师的人力极其有限，从而严重削弱了督导老师对传统课堂教学的监督效果；其次，督导老师在监督课堂教学时，一般只选择一节课的某时间段监督老师和学生的互动情况及学生的听课质量；此外，督导老师无法长期跟踪该课程连续课堂的教学效果，从而导致评价教学的数据样本不足；最后，由于上述因素，被监督课堂的老师或学生很容易在被监督时为了学校或教育部门的评价而弄虚作假，在无人监督时恢复原状，从而难以对被监督课堂进行客观的评价。

除了上述因素，在传统课堂教学模式中，还有其他影响教学质量的缺陷：一是教师对学生在课堂中的表现缺乏直观认知；二是教师无法精确掌握学生们长时间段的课堂表现变化。这些缺陷使得教师难以对教学方法进行精准调整。

近年来， 人工智能（Artificial Intelligence，AI）技术在社会各领域的广泛应用，方便了人们的生活。在机场、车站，人们直接刷脸进行身份鉴别；在购物中心，智能机器人通过“眼睛”识别并分析顾客表情，以改进服务质量；在刑侦方

向，警察通过仪器检测嫌疑人的表现并分析其心理，判断嫌疑人口供是否真实，以侦破案件。基于上述论述，本文可以使用 AI 技术对采集到的视频帧进行感知与识别，通过将学生的课堂行为进行分类并设计评价算法，完成对学生课堂状况的感知、识别与评价。

1.2 国内外研究现状

在课堂状况分析领域，涉及的所有关键技术包括面部表情识别、人体行为检测、语音处理、文本分析和生理信号解析等。面部表情识别技术的流程是：首先捕捉面部图像，接着进行图像的预处理，然后提取相关特征，并对表情进行分类。而人体行为识别技术则通过获取图像或视频资料，对其进行预处理，紧接着提取人体特征并分析行为。语音分析则从预处理采集到的语音数据开始，继而提取特征、降低特征维度并消除冗余信息，最终实现对语音的分类和课堂状况的评估。文本分析利用自然语言处理技术，可以进一步分为基于情感词典、传统机器学习以及深度学习的方法。生理信号分析则涉及使用设备采集脑电、心电等生理信号，并对这些信号进行后续处理和分析，以评估受测课堂的学生表现。

目前，课堂状况分析领域的主流研究对于学生行为的分类主要借鉴了面部表情识别技术的分类方法。著名心理学家 Mehrabiadu[6]在其研究中提到，人们在表达情绪时，面部表情和肢体语言所传达的信息量占到了 55%，而其他表达形式的总和不到这个数值的一半。面部表情，作为人们表达内在情绪的一种方式，涉及眼睛、眉毛、嘴巴、鼻子和脸部肌肉的运动。面部表情识别（Facial expressing recognition，FER）技术基于这些表情，旨在定义人类的内在情感。如图 1-1[7]所示，1975 年，心理学家 Boucher 和 Ekman[7]首次提出了人类六种基本情绪的分类：愤怒（Angry）、悲伤（Sadness）、快乐（Happiness）、惊讶（Surprise）、厌恶（Disgust）和恐惧（Fear），并创建了相应的人脸表情数据库。传统的表情识别算法通常使用机器学习的方法，通过特征提取（如 PCA[8]、LBP[9]、Haar[10]等）和分类器（例如 Adaboost[11]、SVM[12]等）来分类面部特征。行为识别主要针对图像或视频中的个体进行行为分类。然而，这些传统算法在速度和精度上往往不能满足实际应用需求。

过去十年里，随着 GPU 等计算机硬件的快速进步[13]，基于神经网络的深度学习技术引起了研究者们的广泛兴趣，并已被应用于表情和行为识别领域[14][15]。

深度学习技术突破了传统算法在速度和精度上的限制，在实际应用中展现出更优越的性能和鲁棒性。在课堂状况分析领域最新的研究成果中，深度学习技术也成为主要的方法。

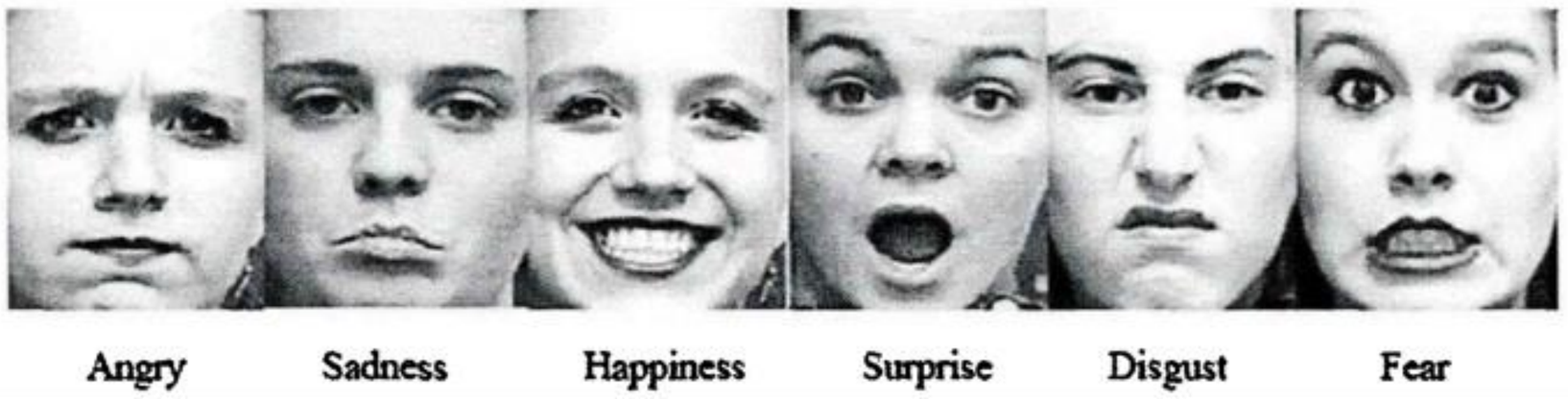


图 1-1 六种基本情绪

Fig.1-1 Six basic emotions

在课堂状况分析领域，公开可用的数据集相对匮乏，因此数据集的构建和标注选择成了研究的关键环节。一些研究者通过收集学生在课堂上的行为数据来创建数据集以供实验之用，而其他研究者则依赖现有的公开表情或行为识别数据集来验证他们的算法，两种方法都已经展示出了实验效果。本节将对近年来国内外学者使用深度学习技术进行学生课堂行为分析的研究成果进行综述，分别从使用公开数据集和自建数据集两个角度进行梳理。

1. 使用公开表情、行为识别数据集的研究

吴慧婷[16]开发了一个综合了学习情绪、疲劳度、专注度和课堂接受度四个维度的在线课堂评估系统，旨在衡量学生的课堂投入度。该工作采用了 GENKI-4K（4000 张图像）、CelebA（10046 张图像）和 SMILEsmileD（13465 张图像）三个公开数据集，分别含有笑脸和非笑脸的图像，对这些图像进行预处理，并通过一个含有四个卷积和四个池化层以及一个全连接层的神经网络进行训练，从而评价学生的情绪状态。为了评估疲劳度，该工作运用了一个类似结构的网络以及 PERCLOS 方法。同时，该工作还利用脸部姿态估计算法和问卷调查来测定学生的专注度及课堂接受度，并通过加权法整合这四个维度的信息，以计算学生的课堂投入度。

张璟[17]的研究将课堂专注度的分析流程划分为两个主要阶段。在第一阶段，该研究将 WIDER FACE 数据集（包含 32,203 张图像和 393,703 个标记的人脸）输入基于 ResNet101 的神经网络。该网络集成了多尺度特征融合以及作者自创的通道和空间注意力机制，目的是增强模型对人脸关键特征的捕捉能力并减少不相

关信息的影响。在第二阶段，张璟采用了 FER2013 数据集。该数据集由 ICML 比赛提供，包括 28,709 张训练图像、3,589 张验证图像和 3,589 张测试图像，涵盖生气（Angry）4953 张、厌恶（Disgust）547 张、害怕（Fear）5121 张、高兴（Happy）8989 张、悲伤（Sad）6077 张、惊讶（Surprise）4002 张和中性（Neutral）6198 张，共七种不同的表情类别的图像。面对数据类别分布的不均衡性，该研究采取了图像增强技术，如随机翻转、旋转和缩放，以及利用生成对抗网络 SRGAN 处理图像，以便进行放大和尺寸统一。接着，该研究将这些表情类别在课堂环境下重新定义为正面或负面情绪，并设计了一个轻量级网络，该网络包含了 3×3 深度可分离卷积模块和改进的 Inception 结构。该研究还采用了结合中心损失和 Softmax 损失的联合损失函数，并引入了残差网络以实现特征的跨层连接。最终，这一算法被用于分析实际收集的学生课堂行为图像。

Jisi 等人[18]利用了 UCF-101 和 HMDB-51 两个公开的人体运动数据集，其分别包含了 101 种和 51 种人体动作的视频。该研究提出了一种结合空间变换网络（Spatial Transformer Network，STN）和卷积神经网络（Convolutional Neural Network，CNN）的新方法来提取视频中的时空特征，并通过赋予不同权重来融合这些特征，最后使用 Softmax 分类器对学生的行为进行分类和预测。

2. 使用自建学生行为数据集的研究

Yang 等人[19]收集了 282,240 条包括面部表情、文本、声音等多模态信息，并将线上学习者的学习情绪划分为六类：高兴（happiness）、吃惊（surprise）、中性（neutrality）、生气（anger）、疲劳（fatigue）和困惑（confusion）。他们开发了一种结合多模态感知和逻辑函数的模型，通过为不同模态信息在逻辑函数中赋予特定权重，来评估学习者的情绪状态。这个情绪评估框架包括了数据收集、处理、分析、可视化以及情绪分类和反馈等多个环节。

Huang 等人[20]在实际课堂环境中收集了数据，并手动筛除了遮挡严重的图像。他们通过裁剪、调整灰度、去噪等预处理技术优化了图像，随后进行边缘检测和分割，以区分每位学生与其他人和背景。基于这些处理，该工作构建了一个学生课堂行为数据集，定义了六种行为：抬头（Raising head）、举手（Raising hands）、站立（Standing）、低头（Lowering head）、趴桌（Bending over desk）和左顾右盼（Looking around）。在此定义中，抬头被视为专注（Attentive），举手和站立表示

需要反馈（Feedback required），而其余三种行为被归类为不专注（Inattentive）。该团队还提出了一种结合卷积神经网络（CNN）和级联网络的面部特征点定位技术，并结合了头部姿态估计和面部表情识别来判断学生的课堂行为。该研究对比了传统的 PCA、AdaBoost 和 LBP 方法，并证明了自己的方法在效果上的优越性。

Ashwin 等人[21]收集了大量数据，对学生在课堂上的行为进行了细致的分类和标注，创建了一个学生课堂行为的数据集。该工作依据两种不同的检测方法构建了两套数据集：第一套是通过分析每位学生的多模态图像数据来单独分类其学习状态，并对所有学生的状态进行集体平均，创建了一个包含 50 名学生、24,000 张图像的数据集，其每张图像仅包含单个学生，情绪状态被标注为参与（engaged）、无聊（boredom）和中性（neutral），每种状态有 8,000 张图像。第二套数据集考虑了包含所有学生的完整图像帧中的所有多模态特征，并对图像帧中的专注度进行分类，构建了一个包含 2,400 名学生、36,000 张图像的数据集，每张图像包括了多名学生，他们的情绪表达被认为是一致的，每个情绪状态有 12,000 张图像。该研究团队还开发了两个基于 Inceptionv3 架构的卷积神经网络（CNN）模型，分别命名为 CNN-1 和 CNN-2，用于分别训练两种方法构建的数据集。然后，该工作将 CNN-1 和 CNN-2 结合起来，提出了一个混合的深度学习模型。这个模型整合了面部表情、手势和身体姿态等图像中的多模态信息，用以检测真实课堂环境中学生的状态。在测试集上，该模型的准确率超过了 95%。

潘仙张等人[22]利用了公开的学习情感数据集 BNU（目前该数据集难以获取），该数据集含有 144 名学习者的 22,708 张表情图像、1,792 组图像序列和 243 个视频片段。此外，该研究中也包含了自行采集的数据，这部分数据来自于男女比例为 2:1 的 90 名本科生的课堂采样，样本中课程难度和学生成绩保持相对平衡。收集到的数据经过了详细的标注，标签包括掌握、认真听课、中性、困惑和不想听课。研究对比了 CNN、CNN+AdaBoost、CNN+决策树、CNN+SVM 等多种方法，发现 CNN+SVM 的准确率和召回率更高，并据此开发了一个基于面部表情识别的课堂教学反馈系统，以分析学生的课堂专注度。

史雨等人[23]从收集的数据中手动筛选出了 1,000 张分布均匀的学生样本，并进行了两轮标注：第一轮标注人体位置，第二轮标注学生的课堂学习状态。该工作将学生的学习状态划分为正常听课(Normal)，阅读和记笔记(Read)，玩手机

(Phone)，睡觉(Sleep)，左顾右盼等其他未正常听课行为(Abnormal)五个类别。该工作使用 K-means++聚类算法进行目标候选框的聚类分析，并构建了一个双重 YOLOv3 网络系统来检测人体位置和学生的课堂学习状态，有效地减少了 YOLOv3 算法的漏检率。

王泽杰等人[24]使用高清摄像头在 6 个大学教室内收集了 300 人次的 200GB 视频数据，并从中筛选出有特定分类动作的视频进行抽帧和标注，构建了两个数据集。第一个是姿态数据集，含有 5,500 张图片，分类为正坐（1600 张）、侧身（1400 张）、低头（1400 张）和举手（1100 张）四种学生课堂行为。第二个是手部动作数据集，含有 4,000 张图片，进一步细分为常规动作（2400 张）、玩手机（800 张）、书写（800 张）三个类别。研究中使用 OpenPose 算法[25]提取人体姿态特征，并通过一个 6 层的小型 CNN 网络对这些特征图进行分类，随后对分类为低头的图像使用经过剪枝的 YOLOv3 网络进行手部动作识别，以此来综合分析学生的课堂行为。

Min 等人[26]开发了一个检测系统，该系统通过分析在线考试和网络课堂的数据，来评估学生在新冠疫情期间的在线学习状态。这个系统由三个主要模块组成，包含八种不同的子模型：人脸分析（Face Analysis）、视频分析（Video Analysis）和动作识别（Action Recognition）模块。人脸分析模块包括五个子模型：人脸检测模型（Face Detection Model）用于定位人脸区域，人脸识别模型（Face Recognition Model）用于确认身份，人脸欺诈检测模型（Face Spoofing Detection Model）用于识别非真实人脸，凝视估计模型（Gaze Estimation Model）用于判断视线方向，以及人脸表情识别模型（Facial Expression Recognition Model）用于识别表情。这些模型运用不同的特征提取技术来分析线上的人脸信息。视频分析模块由两个子模型构成：视频分析模型（Video Analysis Model）运用 EfficiencyNet-B0 网络[27]来判断学生的屏幕是显示的考试内容还是在线课程资料，而不当物品检测模型（Improper Object Detection Model）则采用基于 SSD（Single Shot Multibox Detector）算法[28]的 MobileNetv2 网络[29]来识别画面中的非法物品。动作识别模块只包含动作识别模型（Action Recognition Model），这个模型也基于 MobileNetv2 网络，用来识别在线考试中的六种行为：正常（normal）、笔记本电脑（laptop）、智能手机（smartphone）、平板电脑（tablet）、翻书（flipping pages）和垂下手臂

（arm down）和在线课堂中的七种行为：正常（normal）、笔记本电脑（laptop）、智能手机（smartphone）、平板电脑（tablet）、小睡（nap）、打盹（doze off）、打哈欠（yawn）。这三个模块分别对学生在在线考试和网课中的表现进行统计和输出。

尽管在学生课堂状况分析领域，研究者们已经取得了上述进展，但是仍存在以下问题。首先，这一领域缺乏公开的学生行为数据集。由于学生行为与教学质量之间的关系极为复杂，目前依赖公开的表情识别数据集的研究通常不能合理地分类学生的行为。其次，课堂图像常常受到遮挡且后排学生目标尺寸小，于是要求算法具有更高的精度，然而，许多现有算法的精度仍不够高。最后，即便学生的课堂行为能被网络正确识别并分类，如何建立一个可靠的课堂质量评估算法也是一个紧迫待解的问题。目前，大多数研究仅仅停留在对学生行为的简单分类上，而忽视了课堂质量评估算法的设计。

1.3 研究内容及创新点

本文的工作主要是构建一个能够精准感知与识别学生课堂视频中的各种学生行为，并通过可信的预设算法对该课堂质量进行评估的教辅系统。首先，回顾了神经网络、目标检测和注意力机制等算法的发展现状以及相关的基础知识，针对目前学生课堂状况分析领域缺乏公开数据集且学生行为分类不合理、算法精度低、缺少课堂质量评估算法设计等问题展开研究，最后基于目标检测算法、注意力机制、赋权法等构建了该教辅系统，研究框架如图 1-2 所示。

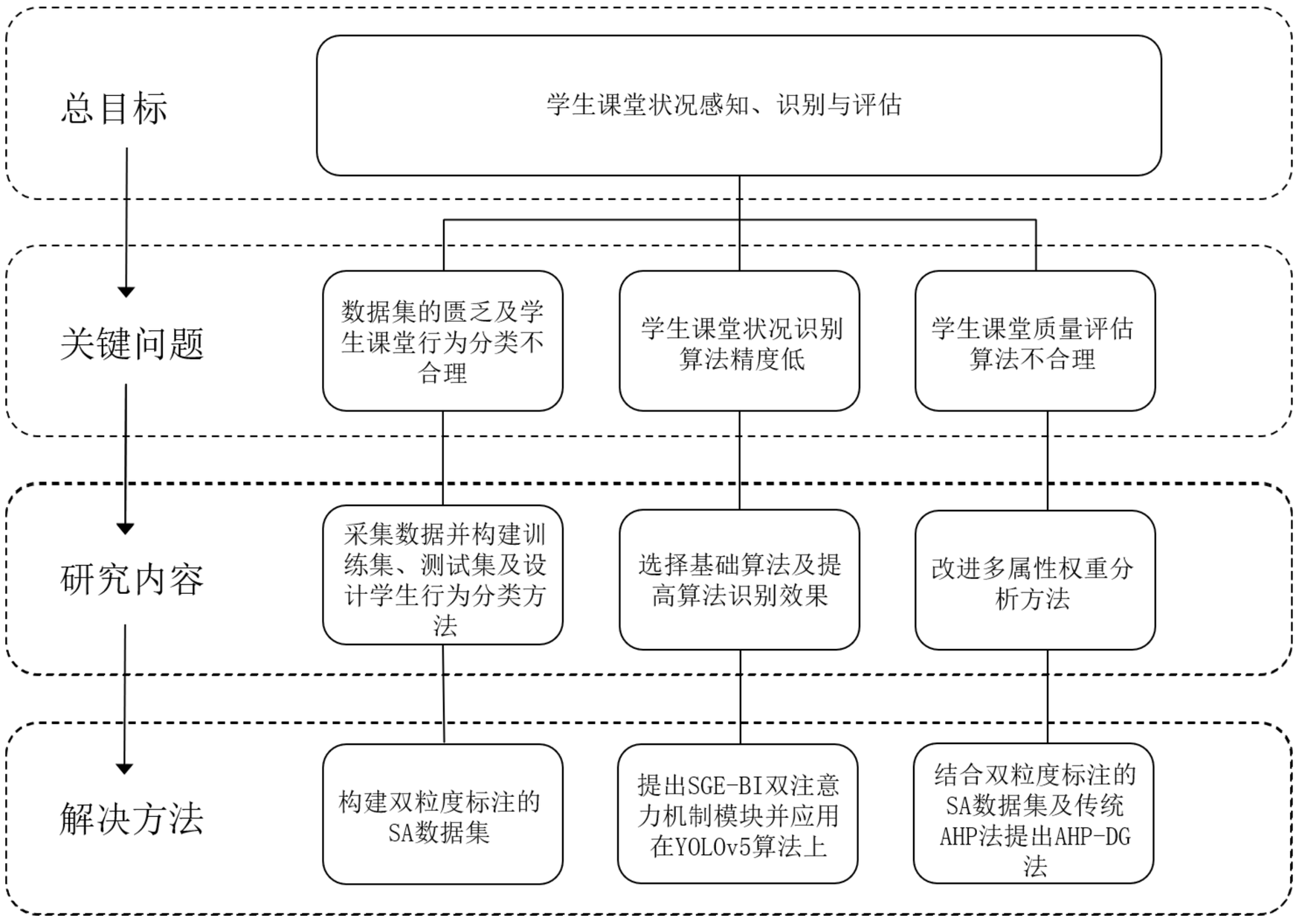


图 1-2 研究框架图

Fig.1-2 Research framework diagram

研究内容及创新点总结如下：

1. 构建了双粒度标注的学生行为（Student Action，SA）数据集

本文采集了教室的学生课堂行为视频以及收集了网络上单人、多人课堂行为图像和视频，通过数据处理构建 SA 数据集，将学生行为划分为细粒度的 9 种：写，站，睡觉，读，举手，玩手机，正常听课，低头，左顾右盼；以及粗粒度的 3 种：未参与（包含睡觉、玩手机、左顾右盼、低头），参与（包含正常听课、读、写），互动（包含站、举手）。

2. 提出了 SGE-BI（Spatial Group wise Enhance-BiFormer）双注意力机制模块并将其成功应用在 YOLOv5m 算法中

在实验评估了多种业界和学术界流行的目标检测算法后，本文选定了性能最优的 YOLOv5m 作为基础算法，并引入了 SGE（Spatial Group wise Enhance）和 BiFormer 注意力机制模块构建的 SGE-BI 双注意力机制模块，将其集成到 YOLOv5m 的 neck 网络部分，从而在原有算法的基础上将算法的精度、mAP@0.5 和 mAP@0.5:0.95 分别提高了 1.9%、0.5%和 0.8%，明显提高了算法识别的性能。

3. 结合双粒度标注的学生行为数据集对传统层次分析法（Analytic

Hierarchy Process，AHP）进行改进，提出了双粒度层次分析加权法（Analytic Hierarchy Process-Double Granularity weighted，AHP-DG）

在课堂质量评估算法方面，对传统的 AHP 法进行了改进，提出了 AHP-DG 法：细粒度 AHP 法通过赋予明显认真听讲的学生（尤其是站和举手状态）更高的权重，以突出互动性较好的课堂；而粗粒度 AHP 法则给予参与和互动状态的学生相似的权重，旨在为表现平均（如参与较多但互动一般）的课堂提供更公正的评价，并将这两种评分进行加权。最终，通过使用传统 AHP 法和 AHP-DG 法对采集的 10 个真实课堂教学视频的评估，结合教育领域专家的评价，验证了 AHP-DG 法相较于传统 AHP 法更具有合理性。

1.4 论文组织架构

本文分为 6 个章节，其中第 3 章至第 5 章为论文主体。

第 1 章是绪论，首先梳理了学生课堂状况分析领域的背景及发展现状，之后，介绍了本文的主要研究内容和组织架构。

第 2 章为相关技术与预备知识介绍，包括神经网络及在其基础上发展起来的深度学习的分类和目标检测算法、注意力机制。

第 3 章介绍了本文构建的双粒度标注的 SA 数据集，并具体说明了该数据集的数据采集、处理方法。

第 4 章介绍了本文提出的基于 YOLOv5 的学生课堂行为识别模型。本章针对目前该领域中识别网络精度普遍较低的问题，提出了 SGE-BI 双注意力机制模块，并将其添加在 YOLOv5 网络中，大大提高了模型的识别能力。在实验部分，本文首先通过对比多个流行的目标检测模型的性能，说明了选择 YOLOv5 的理由，然后通过对比本文方法与原算法和仅添加单注意力机制算法的性能，证明了本文算法的有效性，最后通过在原算法中堆叠 SGE-BI 模块，说明了在原算法中添加两个 SGE-BI 模块是最合适的。

第 5 章介绍了本文提出的学生课堂质量评估算法。本章首先简要介绍了主观赋权法和客观赋权法，然后基于层次分析法（AHP 法），结合本文构建的双粒度标注的 SA 数据集，提出了双粒度层次分析加权法（AHP-DG 法）。最后，本章通过仿真实验，以领域内专家对多个课堂的评分为基准，证明了该方法能够调整课堂质量评价的主观性，从而比原方法更合理。

第 6 章是本文的总结与展望。

第2章　相关理论与预备知识介绍

本章对论文中涉及的相关技术，如神经网络、卷积神经网络、基于卷积神经网络的分类和目标检测算法、注意力机制等进行了介绍。这些相关技术是本文后续实验研究、方法改进的基础。

2.1 神经网络

自从计算机问世以来，人们一直梦想着让它能够模仿人脑的工作机制。在1943年，心理学家McCulloch和逻辑学家Pitts共同提出了MP模型[30]，这是一种早期尝试，用以模仿大脑神经元的工作原理。大约20年后，Rosenblatt进一步提出了感知机模型[31]，这标志着神经网络理念的具体化。然而，由于当时计算机硬件的限制，神经网络的进展一度陷入了瓶颈。神经网络的本质是模拟神经元的结构，通过将众多神经元相互连接，并组织成多层次的深层网络，从而赋予计算机学习和逻辑推理的能力。一个单隐藏层感知机包含输入层、一层隐藏层和输出层，如图2-1所示。

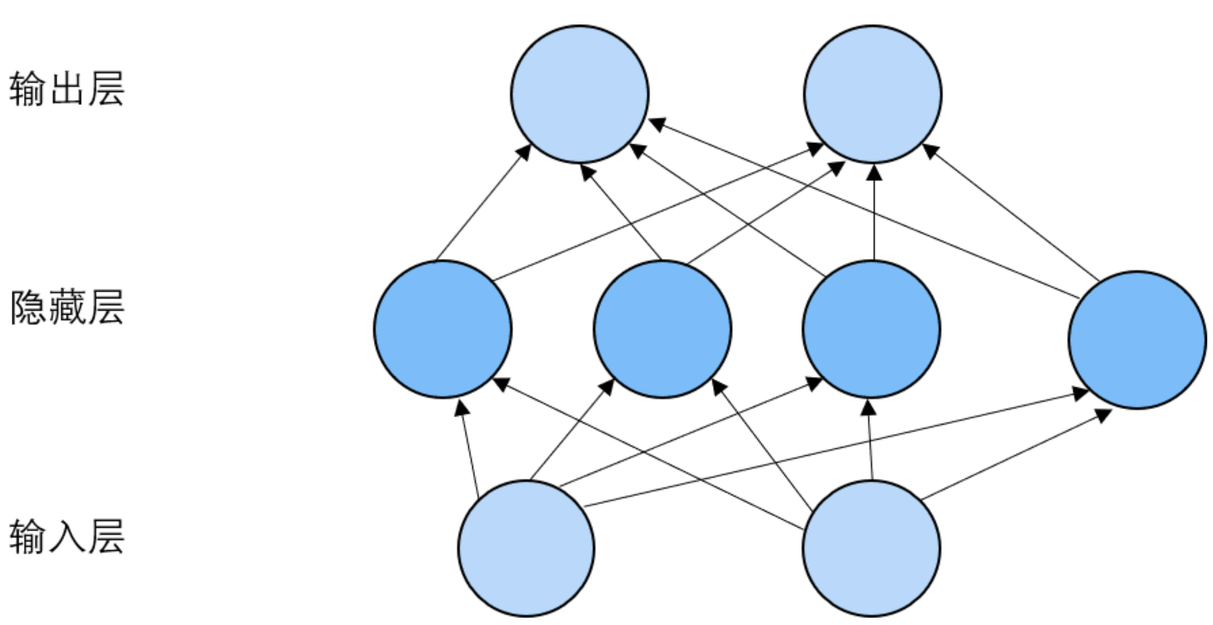


图 2-1　单隐藏层感知机

Fig.2-1 Single hidden Layer Perceptron

2.1.1 神经元

在一个神经网络中，神经元（Neuron）是其最小的构成单位。图2-2显示了一个线性的神经元模型。

假如$x = \{x_1, x_2, \cdots, x_n\}$是该神经元的$n$个输入，$w = \{w_1, w_2, \cdots, w_n\}$是每个输入对应的权重，则可得到其输出：

$$u = w^T x + b \tag{2-1}$$

其中，b为偏置项，u为该神经元的输出。

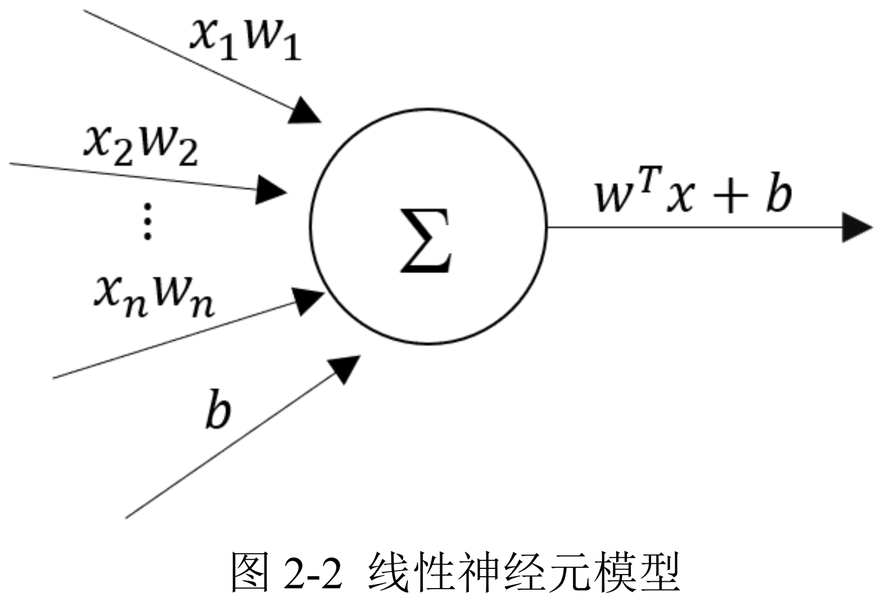


图 2-2 线性神经元模型

Fig.2-2 Linear neuron model

2.1.2 前向传播与反向传播

前向传播（Forward Propagation）和反向传播（Back Propagation）是神经网络中的两个基本概念。

1. 前向传播

前向传播是指神经网络从输入层到输出层的传播过程。在前向传播的过程中，在每一层节点中进行加权，之后通过激活函数（Activation Function）的计算引入非线性性，得到当前层的输出结果，并将其传递到下一层。这个过程一直持续到神经网络的输出层，从而得到神经网络的最终预测结果。在 2.1.4 节中将对激活函数进行更详细的介绍。

将公式(2-1)得到的输出结果u通过一个激活函数：

$$o = f(u) \tag{2-2}$$

其中，f是激活函数，o是加入激活函数后的神经元输出结果。

2. 反向传播

反向传播是一种计算损失函数（Loss Function）对每个权重和偏差的梯度的方法。在神经网络的训练过程中，通过反向传播计算损失函数对每个权重和偏差的梯度，并使用梯度下降法来调整这些参数值，以使损失函数最小化。

具体而言，反向传播将损失函数对神经网络输出的偏导数作为初始误差的信号，并利用链式法则将这个误差信号一层层传递回神经网络中的每个神经元的输入端，计算每个神经元输入对其输出的偏导，再将其与误差信号相乘，得到每个神经元的误差信号。通过这个过程，反向传播算法逐步计算出所有权重和偏差的梯度，以更新网络参数。

下面论述输出层在反向传播过程中的权重更新：

假如一个全连接网络有n个输出$output_1, output_2, \cdots, output_n$和$m$个该层的输入$input_1, input_2, \cdots, input_m$，则根据前向传播：

$$output_i = f\left(\sum w_j^o input_j + b^o\right) \tag{2-3}$$

其中，w_j^o为该层输入的权重，b^o为该层的偏置项，$f(\cdot)$为激活函数。

令实际值为$target_1, target_2, \cdots, target_n$，损失函数为$loss(\cdot)$，我们计算该层的总误差：

$$E_{total} = \sum loss(target_i - output_i) \tag{2-4}$$

以权重参数w_1^o为例，利用链式法则计算总误差对其偏导：

$$\frac{\partial E_{total}}{\partial w_1^o} = \frac{\partial E_{total}}{\partial\ output_1} \cdot \frac{\partial\ output_1}{\partial\ input_1} \cdot \frac{\partial\ input_1}{\partial w_1^o} \tag{2-5}$$

现在对w_1^o值进行更新：

$$(w_1^o)' = w_1^o - \eta \cdot \frac{\partial E_{total}}{\partial w_1^o} \tag{2-6}$$

其中，η是学习率，即每次网络训练权重更新的幅度。

更新隐藏层权重与上述计算方法类似，但由于隐藏层权重在更新时会接收到后一层所有神经元传来的误差，所以在计算时将其考虑到即可。

2.1.3 欠拟合、过拟合及其处理方法

在网络的训练过程中，人们追求的是网络不仅在训练数据上有出色的表现，同时也能在新的、未见过的测试数据上展现同等的性能。这种网络对新输入数据的适应能力被称为泛化（Generalization）。网络在处理未见过数据时的误差被称作测试误差（Test Error）。在尝试减少训练误差和测试误差的时候，网络可能会遇到问题：一方面，网络无法在训练数据上达到足够低的误差，这种现象称为欠拟合（Underfitting）；另一方面，如果网络在训练数据上的误差远远低于在测试数据上的误差，这种情况则被称为过拟合（Overfitting）。

处理欠拟合的方法通常是增加网络的复杂度，避免其过于简单；而处理过拟合通常有正则化（Regularization）的方法，如权重衰退，或使用 Dropout 方法等。

1. 权重衰退

权重衰退是一种正则化的方法，即通过限制参数值的选择范围来控制模型容量。假如网络的损失函数为$loss(w, b)$，可以强制限制权重参数，令：

$$\| w \|^2 \leq \theta \tag{2-7}$$

其中θ为任意数，则更小的θ意味着更强的正则项。

更一般地，通常加入均方范数进行限制：

$$\min \ \ loss(w,b) + \frac{\lambda}{2} \| w \|^2 \tag{2-8}$$

其中，λ是提前设定的值，以控制正则项的重要程度。

之后计算梯度并更新参数：

$$\frac{\partial}{\partial w}(loss(w,b) + \frac{\lambda}{2} \| w \|^2) = \frac{\partial \ell(w,b)}{\partial w} + \lambda w \tag{2-9}$$

$$w' = (1 - \eta\lambda)w - \eta \frac{\partial loss(w,b)}{\partial w} \tag{2-10}$$

其中，η为学习率。

通常，$\eta\lambda < 1$，在深度学习中叫做权重衰退。对比公式(2-6)，可以发现，公式(2-10)的权重参数在更新之前进行了一次减小运算。权重衰退通过$L2$正则项控制网络参数，从而控制网络复杂度，以减小过拟合。

2. Dropout 方法

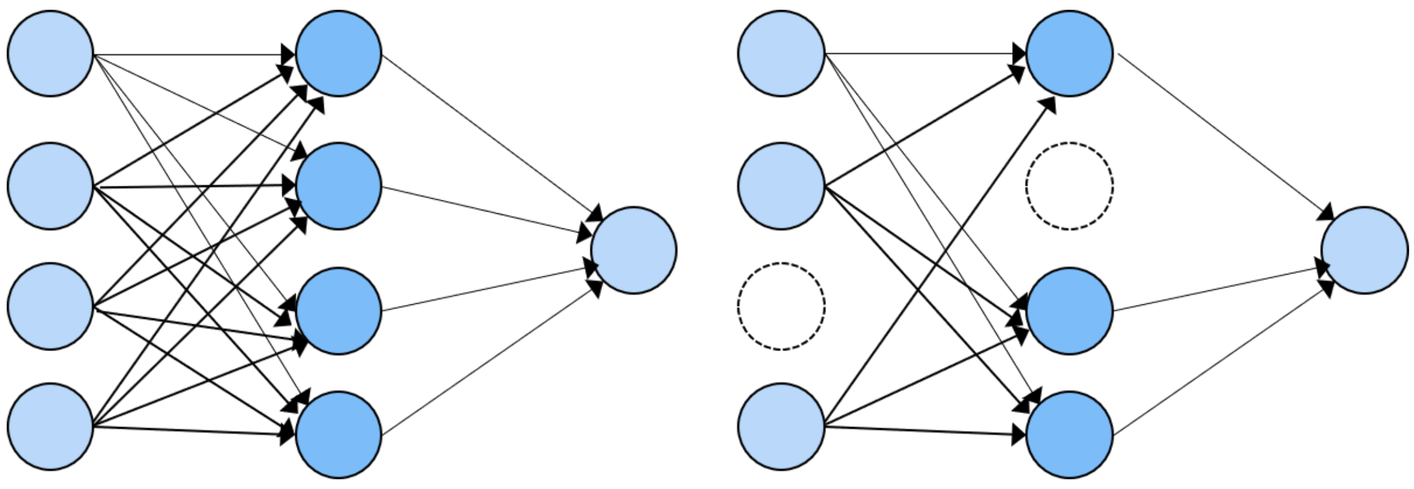
Dropout 的方法主要的思想是一个好的网络模型应该对随机噪音鲁棒。对x加入噪音得到x'，我们希望其期望不变，即：

$$E[x'] = x \tag{2-11}$$

对每个输入元素添加如下扰动：

$$x_i' = \begin{cases} 0, & p \\ \dfrac{x_i}{1-p}, & other \end{cases} \tag{2-12}$$

其中p为使$x_i' = 0$的概率。之后，网络再执行前向传播、反向传播和参数更新。图 2-3 是未使用 Dropout 和使用 Dropout 方法的网络结构。



(a) 未使用 Dropout　　(b) 使用 Dropout

图 2-3 Dropout 方法

Fig.2-3 Dropout method

2.1.4 数值稳定性、模型初始化与激活函数

1. 数值稳定性

当网络模型的深度较大时，研究者通常要考虑数值稳定性的问题。假如有如下网络：网络的深度为d层，t表示当前层，h表示输入或输出，计算损失$loss$关于权重参数w^t的梯度：

$$\frac{\partial loss}{\partial w^t} = \frac{\partial loss}{\partial h^d}\frac{\partial h^d}{\partial h^{d-1}}\cdots\frac{\partial h^{t+1}}{\partial h^t}\frac{\partial h^t}{\partial w^t} \tag{2-13}$$

在上式中，需要计算$d-t$次矩阵的乘法，多次的乘法运算会使梯度在反向传播的过程中呈指数级递增或递减，从而导致网络权重极大幅或极小幅更新，即梯度爆炸和梯度消失现象。所以，深度学习中，研究者需要利用有效的模型初始化方法和激活函数提高网络模型的稳定性。

2. 模型初始化

目前，一种有效保证网络模型稳定性的方法是将每层的输入和梯度都看做随机变量，使它们的均值和方差都保持一致。一般地，对前向传播：

$$E[h_i^t] = 0, Var[h_i^t] = a \quad \forall i,t \tag{2-14}$$

对反向传播：

$$E\left[\frac{\partial loss}{\partial h_i^t}\right] = 0, Var\left[\frac{\partial loss}{\partial h_i^t}\right] = b \quad \forall i,t \tag{2-15}$$

其中，a,b为常数。

假设在上述网络中$W^t \in \mathbb{R}^{n_t \times n_{t-1}}$，$w_{ij}^t$独立同分布，且$h_i^{t-1}$独立于$w_{ij}^t$，于是：

$$E\left[w_{ij}^t\right] = 0, Var\left[w_{ij}^t\right] = \gamma_t \tag{2-16}$$

其中，γ_t为常数。

假如暂时不考虑激活函数和偏置项，则对前向传播，某层输入与输出的均值：

$$\begin{aligned} E[h_i^t] &= E\left[\sum_j w_{ij}^t h_j^{t-1}\right] \\ &= \sum_j E\left[w_{ij}^t\right]E\left[h_j^{t-1}\right] \\ &= 0 \end{aligned} \tag{2-17}$$

方差：

$$Var[h_i^t] = E[(h_i^t)^2] - E[h_i^t]^2$$

$$= E\left[\left(\sum_j w_{ij}^t h_j^{t-1}\right)^2\right]$$

$$= E\left[\sum_j \left(w_{ij}^t\right)^2 \left(h_j^{t-1}\right)^2 + \sum_{j\neq k} w_{ij}^t w_{ik}^t h_j^{t-1} h_k^{t-1}\right] \tag{2-18}$$

$$= \sum_j E\left[\left(w_{ij}^t\right)^2\right] E\left[\left(h_j^{t-1}\right)^2\right]$$

$$= \sum_j Var\left[w_{ij}^t\right] Var\left[h_j^{t-1}\right]$$

$$= n_{t-1}\gamma_t Var\left[h_j^{t-1}\right]$$

由上式可知，若保持输入和输出的方差一致，则：

$$n_{t-1}\gamma_t = 1 \tag{2-19}$$

同理可知，在反向传播中：

$$E\left[\frac{\partial loss}{\partial h_i^{t-1}}\right] = 0 \tag{2-20}$$

$$n_t\gamma_t = 1 \tag{2-21}$$

但实际网络模型中，公式(2-19)和公式(2-21)两个条件几乎不会同时满足，现有的主流方法主要是对网络的输入进行初始化，限制其均值与方差。下面简要介绍两种网络初始化的方法。

（1） Xavier 法

初始化：

$$\gamma_t(n_{t-1} + n_t)/2 = 1 \tag{2-22}$$

即对权重的方差加以限制：

$$\gamma_t = 2/(n_{t-1} + n_t) \tag{2-23}$$

（2） 批量归一化（Batch Normalization，BN）

初始化：

$$h_i' = \alpha\frac{h_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{2-24}$$

其中，h_i和h_i'分别表示 batch 的输入和更新后的输入，α和β用来调整数值分布的方差大小和均值位置，它们在反向传播中进行学习，默认值是 1 和 0。μ和

σ^2即输入的均值与方差。ε为随机扰动。

目前，几乎所有的主流网络都使用了批量归一化，其在本文算法中也是基础性方法之一。它不仅固定小批量中的均值和方差，学习出适合的偏移和缩放，加快收敛速度，使得学习率可以调得较大，并且有研究表明，它通过在每个小批量中加入噪音控制了模型的复杂度，从而提高了模型的鲁棒性。

3. 激活函数

在实际应用中，数据样本几乎不可能是线性可分的。于是，激活函数作为非线性函数，被研究者们引入神经网络以增加网络拟合数据的非线性性。激活函数的选择同样可以保证网络模型的数值稳定性。

由于非线性函数在零点附近可以用泰勒公式展开成线性函数，我们可以用线性函数进行简化分析。

假设激活函数是线性函数：

$$f(x) = ax + b \tag{2-25}$$

那么，t层输入的期望为：

$$E[h_i^t] = E[ax_i^t + b] = b \tag{2-26}$$

方差为：

$$\begin{aligned} Var[h_i^t] &= E[(h_i^t)^2] - E[h_i^t]^2 \\ &= E[(ax_i^t + b)^2] - b^2 \\ &= E[\alpha^2(x_i^t)^2 + 2abx_i^t + b^2] - b^2 \\ &= a^2 Var[x_i^t] \end{aligned} \tag{2-27}$$

由公式(2.26)和公式(2.27)，当$a = 1, b = 0$时，输入数据具有数值稳定性。

表 2-1 三种激活函数的定义及泰勒展开

Table.2-1 The definitions of three activation functions and Taylor expansion

名称及定义	泰勒展开
$Sigmoid(x) = \frac{1}{1+e^{-x}}$	$sigmoid(x) = \frac{1}{2} + \frac{x}{4} - \frac{x^3}{48} + O(x^5)$
$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$tanh(x) = 0 + x - \frac{x^3}{3} + O(x^5)$
$ReLU(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$	$relu(x) = 0 + x \quad for \ x \geq 0$

表 2-1 显示了常用的三种激活函数 Sigmoid、Tanh、ReLU 的定义及在零点

处的泰勒展开式。表中可以看出，Sigmoid 函数用泰勒公式展开后，与上述稳定性分析不符合。这可以解释在深度学习中，随着网络层数的加深，使用 Sigmoid 函数作为激活函数的模型训练效果相较于选择其他激活函数较差的原因之一。

当然，激活函数的种类远不止上述三种，选取方式也不只有数值稳定性一项标准，通常在网络模型中，激活函数的选择还需要具体分析、实验验证。

2.2 卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）是一类包含利用卷积层提取特征的多层感知机的变种，本文识别算法的骨架基于 CNN。一般而言，CNN 由卷积层、池化层、全连接层组成。

2.2.1 卷积层

严格来说，CNN 中的卷积层是一个错误的叫法，因为它所表达的实际上是互相关运算（cross-correlation），而不是数学上的卷积运算。在卷积层中，输入张量和核张量通过互相关运算产生输出张量。

在 CNN 中，核张量也称卷积核（convolution kernel）或者滤波器（filter），简单而言，它是该卷积层的权重，通常是可学习的参数。图 2-4 是一个二维图像的输入经过卷积核的互相关运算得到输出的过程。

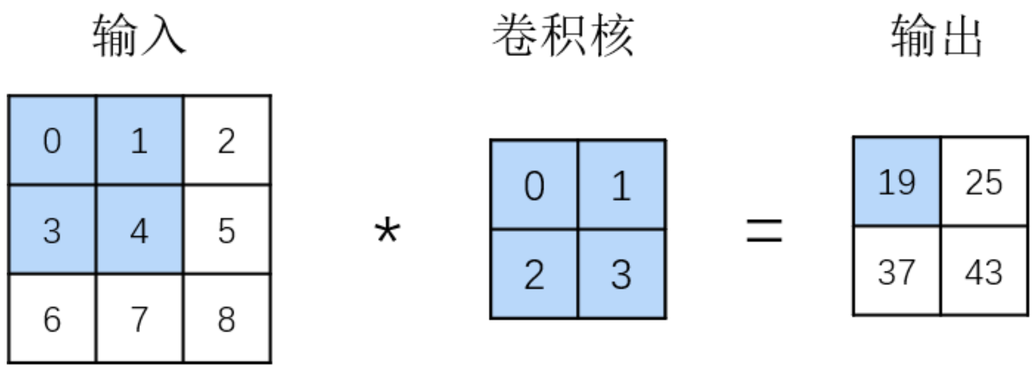


图 2-4 卷积操作

Fig.2-4 Convolution operation

上图中，有一个输入张量，其大小为 3×3（高度为 3，宽度也为 3），使用的卷积核或滤波器具有 2×2 的尺寸，该尺寸即卷积窗口。在执行二维卷积操作时，该卷积窗口从输入张量的左上角开始滑动，按照从左至右、由上至下的顺序移动。每当卷积窗口移至新的位置时，它覆盖的输入张量区域会与卷积核进行逐元素的乘法操作，随后将乘积求和以产生一个单独的标量值。这个标量值构成输出张量在该特定位置的值。在这个例子中，通过这种二维卷积操作，得到了一个 2×2 维度的输出张量，以图中标深色的运算为例，运算结果为：

$$0 \times 0 + 1 \times 1 + 3 \times 2 + 4 \times 3 = 19 \tag{2-28}$$

需要注意的是，输出张量的尺寸通常比输入张量小。这是因为卷积核的尺寸大于 1，并且卷积操作只在卷积核能够完全覆盖的输入图像区域执行。因此，输出张量的尺寸可以通过输入张量尺寸减去卷积核尺寸来计算得出：

$$(n_h - k_h + 1) \times (n_w - k_w + 1) \tag{2-29}$$

其中，$n_h \times n_w$即输入张量的尺寸，$k_h \times k_w$即输出张量的尺寸。

2.2.2 池化层

在处理图像时，人们通常希望逐渐降低隐藏表示的空间分辨率和汇集信息，这样，随着神经网络的深度加深，每个神经元对其敏感的感受野就越大。深度学习的任务通常关注图像的全局而非部分，所以在网络模型中，最后一层神经元应该对整个输入图像具有全局敏感性。此外，当检测较低层的特征时，研究者通常希望保持这些特征的平移不变性，比如，当拍摄者拍摄一个物体时，可能由于相机的震动等原因，使图像整体有少量像素的位移。

池化（pooling）层的作用就在于此。首先，它通过逐渐聚合信息，生成越来越粗糙的映射，最终实现学习全局表示的目标。其次，它能够降低网络对局部空间位置的敏感性。

池化层的操作与卷积层相似，它们都使用一个预设形状的窗口进行滑动遍历，这个窗口按照指定的步长在输入数据的全部区域上移动，对于遍历过的每个窗口位置，池化层计算出一个输出值。与卷积层的权重相关的计算不同，池化层不涉及任何学习参数。池化操作是一种确定性的过程，通常会计算窗口内所有元素的最大值或者平均值，这两种操作分别称作最大池化（maximum pooling）和平均池化（average pooling）。

无论是最大池化还是平均池化，池化窗口都从输入张量的左上角开始，按照从左到右、从上到下的顺序在输入张量上滑动。对于池化窗口覆盖的每个区域，它会计算出该区域内的最大值或平均值，具体取决于是采用的最大池化还是平均池化。在图 2-5 中，显示了一个最大池化的操作。

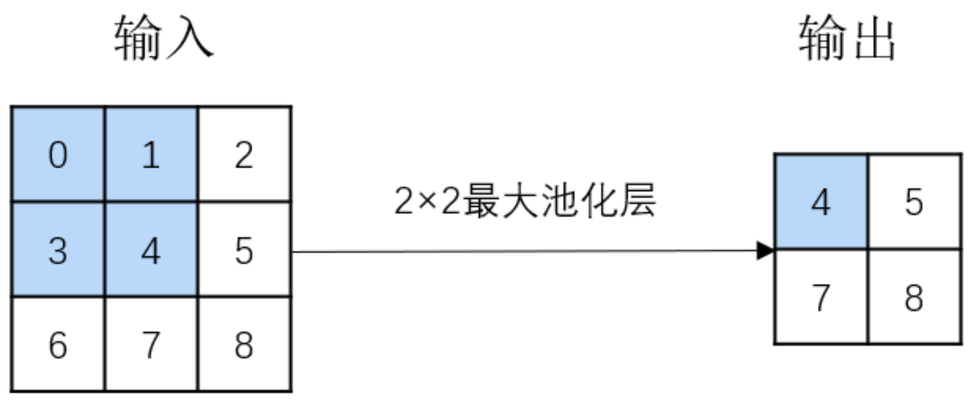


图 2-5 池化操作

Fig.2-5 Pooling operation

上图显示的输出张量的尺寸为 2×2，其包含的四个元素代表了每个池化窗口内的最大值。如，按照图中标深色的最大池化运算为例：

$$\max(0,1,3,4) = 4 \tag{2-30}$$

2.3 基于卷积神经网络的分类算法

分类算法是将图像中的对象根据其特征进行分别归类的方法。基于卷积神经网络的算法，最初被开发出来的主要目的就是为了处理图像的分类问题。

2012 年，Alex Krizhevsky 和 Hinton 共同设计的一个 8 层的卷积神经网络 AlexNet[32]，在那年的 ImageNet Large Scale Visual Recognition Challenge（ILSVRC）竞赛中取得了划时代的成绩，将之前的 top-5 最低分类错误率从 26%大幅降低到 15.3%，在学术和工业界引起了极大的轰动，引发了对深度学习的广泛热情。AlexNet 采用了 Dropout 方法，通过随机丢弃网络中的神经元来减少过拟合，采用了 ReLU 激活函数取代了此前更常用的 Sigmoid 函数，并证明了 ReLU 在更深层网络中的效果更佳。AlexNet 还将 CNN 中通常使用的平均池化层换成了最大池化层，这一改进有效地提升了网络性能。AlexNet 网络结构如图 2-6 所示。

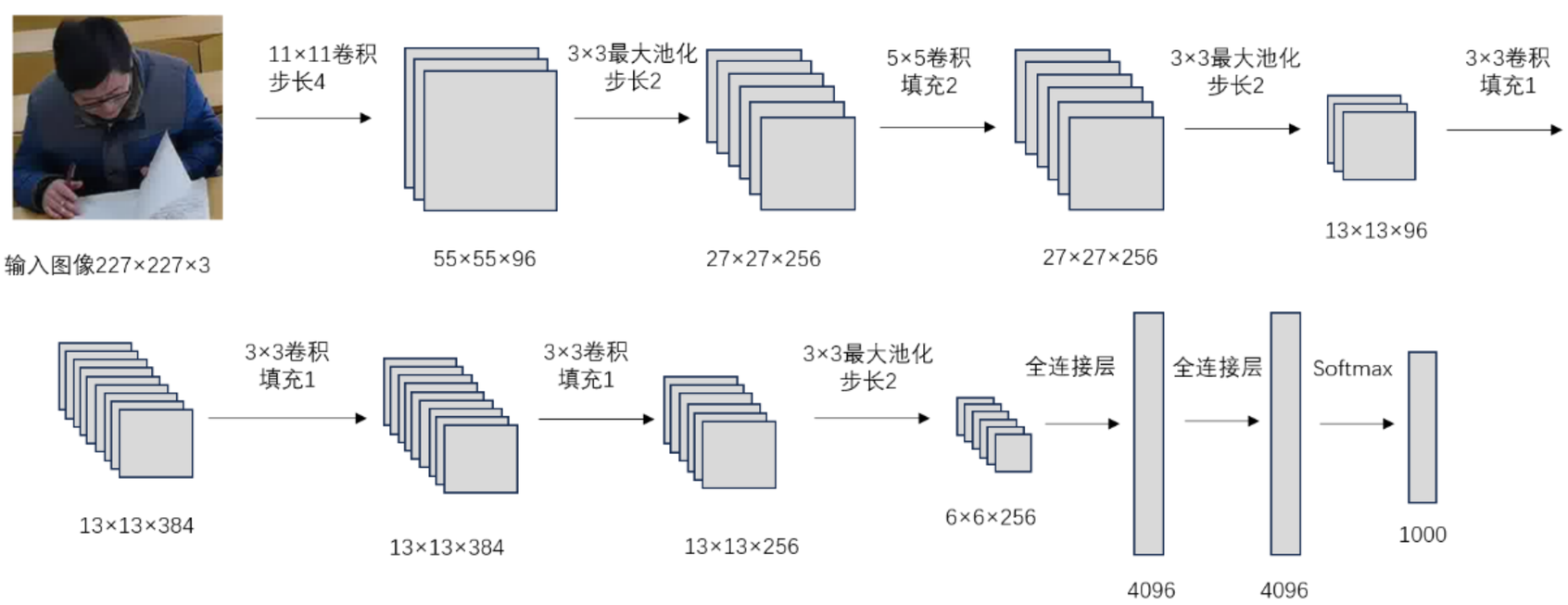


图 2-6 AlexNet 网络结构

Fig.2-6 AlexNet Network Architecture

2014 年，Simonyan 等人提出了层数更深的 VGG 模型[33]，并指出，在相同的计算成本下，3×3 的卷积层能够取得更好的效果。VGG 模型完全由 3×3 的卷积层构成，并且设计了多个这样的卷积层以及一个最大池化层结合的 VGG 块，从而进一步提高了模型的分类准确性。

尽管 AlexNet 和 VGG 通过增加网络层数来提升性能，但随着网络层数的增加，它们也会遇到过拟合、梯度消失或梯度爆炸等问题。此外，这些网络中的全连接层参数众多，在训练中，会消耗大量内存和计算资源。2013 年，Lin 等人推出了 NiN 模型[34]，该模型构建了包含两个 1×1 卷积层的 NiN 块，并用全局平均池化层取代了全连接层。1×1 的卷积层为每个像素引入了额外的非线性性，而全局平均池化层的使用极大地提升了计算效率。2015 年，Szegedy 等人结合了 NiN 中串联网络的思想提出 GoogLeNet[35]，引入了 Inception 块，该模块合并了经过多尺寸卷积核运算的多通道图像，使得网络在加深的同时保持了合理的计算成本。

同年，何恺明等人提出了在深度学习领域具有划时代意义的残差网络 ResNet[36]，它通过引入残差块解决了深度网络中的性能退化问题，使网络中的某些层跳过特征提取效果较差的下一层，这使得网络深度得以大幅度增加而不受限制。残差块的设计对后续深度学习网络架构的发展产生了深刻的影响。该结构细节如图 2-7 所示。

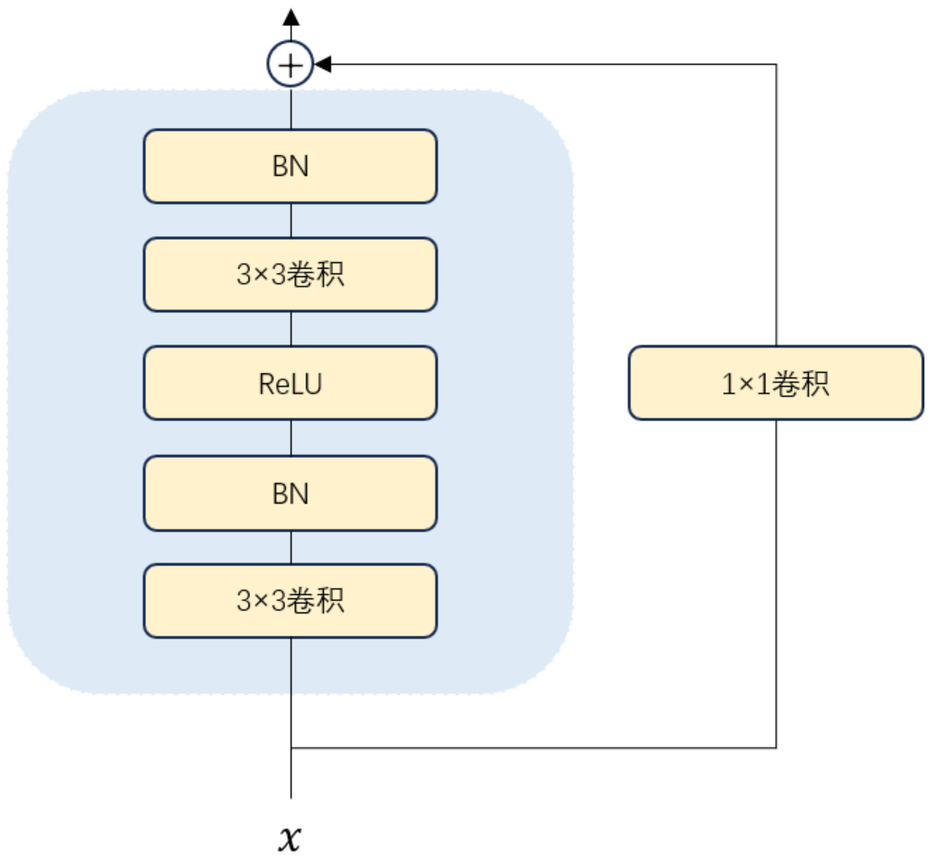


图 2-7 残差块细节

Fig.2-7 Residual block details

到了 2017 年，黄高和刘壮提出了 DenseNet[37]，该网络设计了 Dense Block，通过级联的方式让每一层都接收到前面所有层的输入，这不仅有效缓解了梯度消失问题，加强了特征传递，还减少了模型的参数量。

2.4 目标检测算法

目标检测是一种用于在图像中识别和分割出不同种类物体的方法，学生课堂状况识别是目标检测的一种应用。随着上述网络和相关技术的不断进步，目标检测领域也不断涌现出新算法。这些算法大致分为两类：一类是基于候选区域的算法，另一类是基于回归的算法[38]。

1. 基于候选区域的算法

基于候选区域的目标检测算法采用二阶段（two-stage）的方法：首先，生成锚框（Anchor）以框定图像中可能包含物体的区域；接着，对这些区域进行分类，并预测它们属于特定物体的概率。这类算法的代表是基于区域的卷积神经网络（Region with Convolutional Neural Network，R-CNN）系列算法。

2014 年，Girshick 等人提出了 R-CNN 算法[39]，这是卷积神经网络在目标检测领域的首次成功应用。该研究结合了 AlexNet 和选择性搜索（Selective search）算法[40]，后者首先从输入图像中分割出 2000 个候选区域，然后将这些区域统一调整到相同尺寸，并使用 CNN 提取特征，最终将这些特征传递给一个 SVM 分类器进行物体分类和区域回归。图 2-8 显示了 R-CNN 算法的结构。

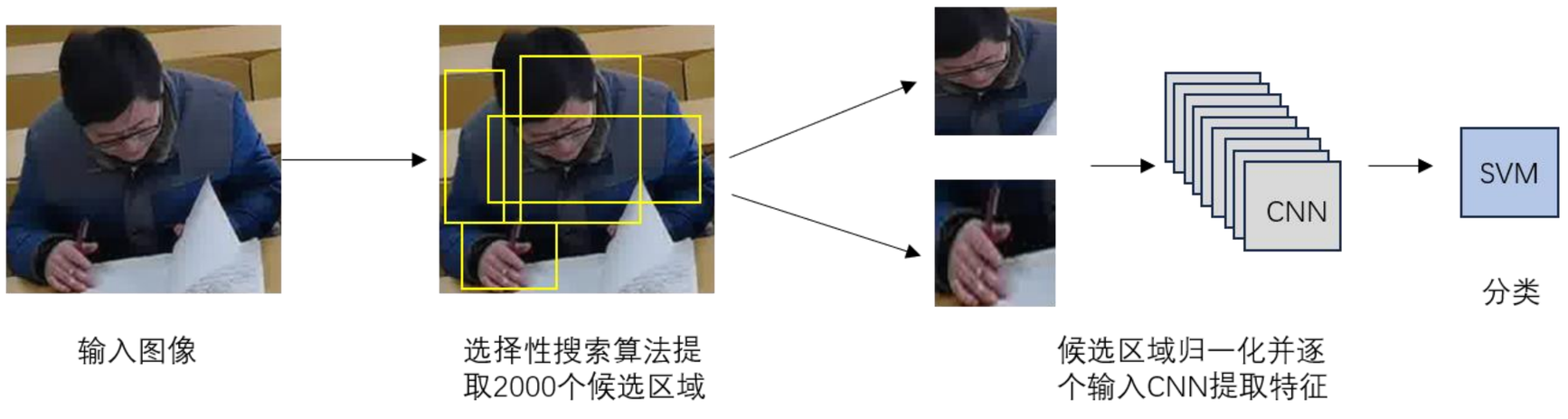


图 2-8 R-CNN 算法结构

Fig.2-8 R-CNN algorithm structure

2015 年，Girshick 等人进一步提出了 Fast R-CNN 算法[41]，该算法借鉴了 SPP-Net[42]的思路，将 SPP 层改进为 ROI 池化层，这样可以保证不同大小的候选区域经过池化后输出尺寸一致，从而允许一次性将所有候选区域输入 CNN，显著减少了计算成本。同时，Fast R-CNN 通过改进损失函数的计算方式，增强了算法的

检测性能。

2015 年，Ren 等人提出了 Faster R-CNN 算法[43]，如图 2-9 所示。该算法引入了区域生成网络（Region Proposal Network，RPN）来取代原有的选择性搜索算法。RPN 利用卷积神经网络来提取特征并产生大量的候选区域。接着，Faster R-CNN 通过非极大抑制（Non-maximum suppression，NMS）算法[44]来筛选这些候选区域，精确定位目标。Faster R-CNN 显著减少了选择性搜索算法的计算负担，实现了端到端的目标检测。

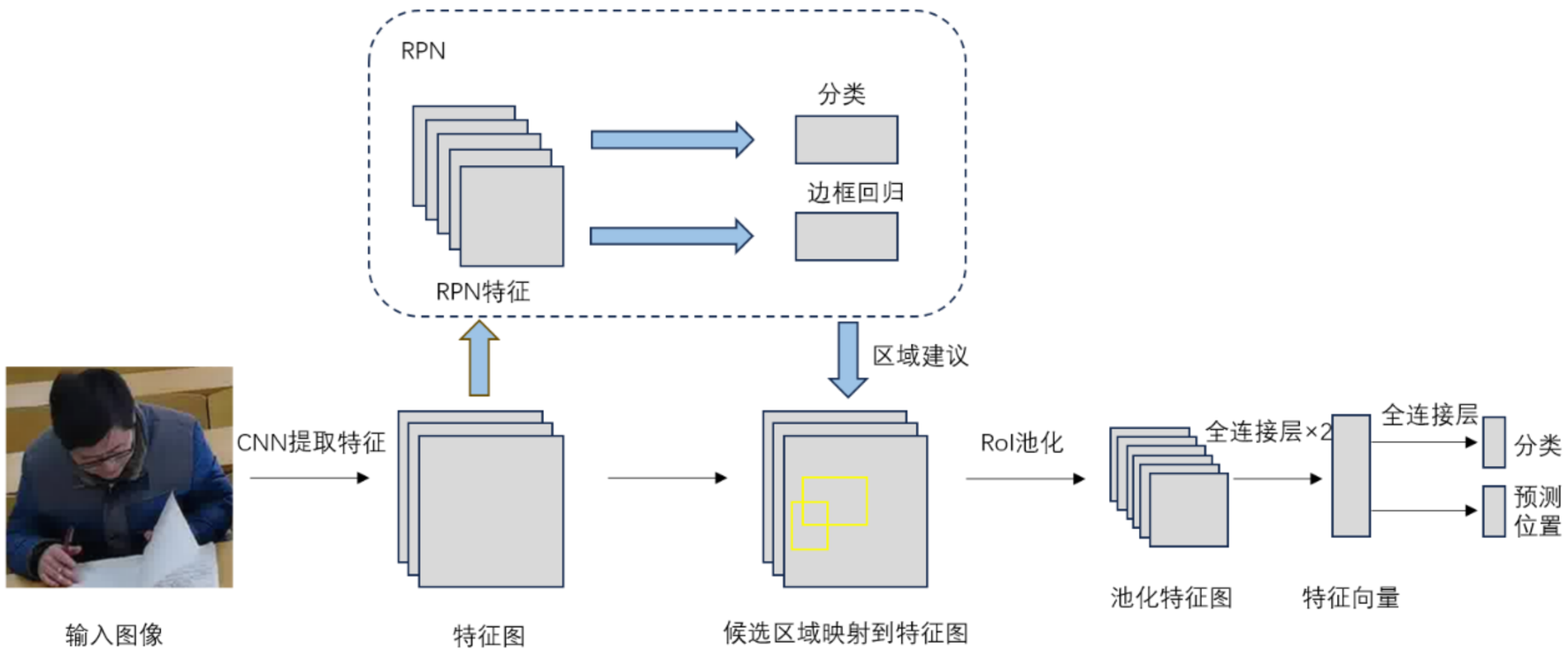


图 2-9 Faster R-CNN 算法结构

Fig.2-9 Faster R-CNN algorithm structure

基于 Faster R-CNN 的四个核心部分：基础特征提取网络、ROI 池化层、RPN 和 NMS，后续的研究者们做出了一系列改进。具体包括：针对基础特征提取网络，Lin 等人在 2017 年提出了特征金字塔网络（Feature Pyramid Network，FPN）[45]，它能够融合高层的语义信息丰富的特征图与低层的定位信息丰富的特征图；在 ROI 池化层方面，何恺明等人在同年提出了 Mask R-CNN 算法[46]，将 ROI 池化层升级为 ROI Align 层，使得算法能够利用像素级标注；在 RPN 的改进上，2018 年 Cai 等人提出了 Cascade R-CNN 算法[47]，优化了候选区域的生成过程，增强了定位精度；对于 NMS 算法，2019 年 Rezatofighi 等人[48]改进了边框损失函数，从而提高了网络的检测性能。

2. 基于回归的算法

一阶段（one-stage）算法，也称为基于回归的算法，旨在直接对物体的位置和类别进行预测。这种算法在图像中密集地采样各种尺寸的锚框，用以定位物体

并进行分类。其中，SSD（Single Shot Multibox Detector）和 YOLO（You Only Look Once）系列算法是这一领域的主流。

（1） SSD 系列算法

2016 年，Liu 等人推出了 SSD 算法。该算法以 VGG-16 为骨干网络，去除了全连接层，转而加入了多个卷积层，并引入了多尺度特征图来捕获不同大小的物体——大尺寸特征图侧重于小物体检测，而小尺寸特征图则侧重于大物体检测。通过 NMS 算法的处理，SSD 算法能够确定物体的位置和类别。SSD 算法结构如图 2-10 所示。

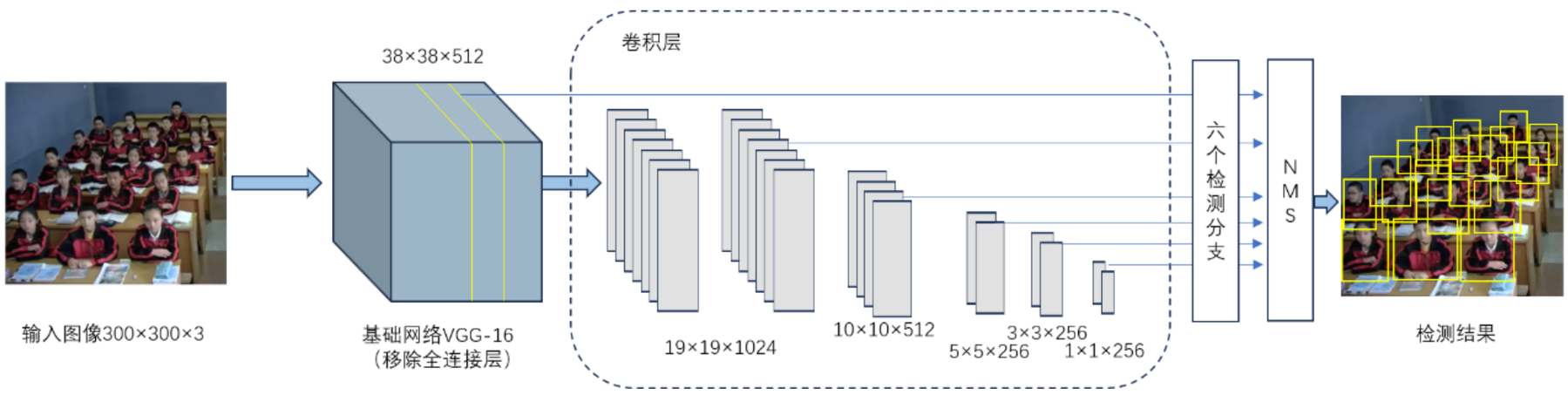


图 2-10 SSD 算法结构

Fig.2-10 SSD algorithm structure

2017 年，Fu 等人提出了 DSSD（Deconvolutional single shot detector）算法[49]，该算法以 ResNet101 替代了 SSD 中的 VGG-16 基础网络，并且融入了特征金字塔的概念，通过添加反卷积层来融合不同尺寸的特征图信息，这样做使得网络融合后的特征在语义和定位信息上更为丰富。尽管这种方法牺牲了一定的速度，但算法的精度得到了显著提升。同年，Lin 等人推出了 RetinaNet 算法[50]，该算法优化了交叉熵损失函数，首次实现了一阶段算法在速度和精度方面超越二阶段算法。

上述算法采用基于 Anchor-based 的方式来生成锚框，这种方法在输入图像中均匀地选取锚点（通常每个像素点都作为锚点），并以这些锚点为中心，生成各种尺寸和比例的锚框。而另一种方法是基于 Anchor-free 的方式，该方法通过关键点检测来预测物体的中心点或边界点，从而生成锚框。对 SSD 算法的后续改进主要集中在基础网络、FPN 以及 Anchor-free 方式上。

在 2019 年，Zhao 等人提出了 MLFPN 结构[51]，将基础网络提取的特征输入到一个级联的多个小的特征融合模块中，使得每个特征图包含来自多个层的特征，从而增强了对小目标的检测能力。同年，Duan 等人在一阶段关键点检测算法

CornerNet 的基础上，提出了 CenterNet 算法[52]，该算法将原本检测物体的一对关键点增加到了三个，这一改进提升了模型的精度和召回率。到了 2020 年，Zhai 等人提出了 DF-SSD 算法[53]，该算法使用 DenseNet-S-32-1 替换了原本的 VGG-16 网络，并优化了多尺度特征融合机制，进一步提升了模型性能。

（2） YOLO 系列算法

2015 年，Redmon 等人提出了 YOLOv1 算法[54]，仅使用一个 CNN 模型就能够同时定位和检测物体，实现了端到端的目标检测。YOLOv1 算法将输入图像固定为 448×448 的尺寸，并通过一系列卷积层、池化层以及最终的两个全连接层处理，输出一个 7×7×30 的张量。这意味着原始图像被分割成 49 个格子，每个格子生成两个锚框，30 维向量中包含 20 个对象的分类概率、两个边界框的位置信息以及它们的置信度，最终，通过 NMS 算法筛选出最终的物体位置，如图 2-11 所示。

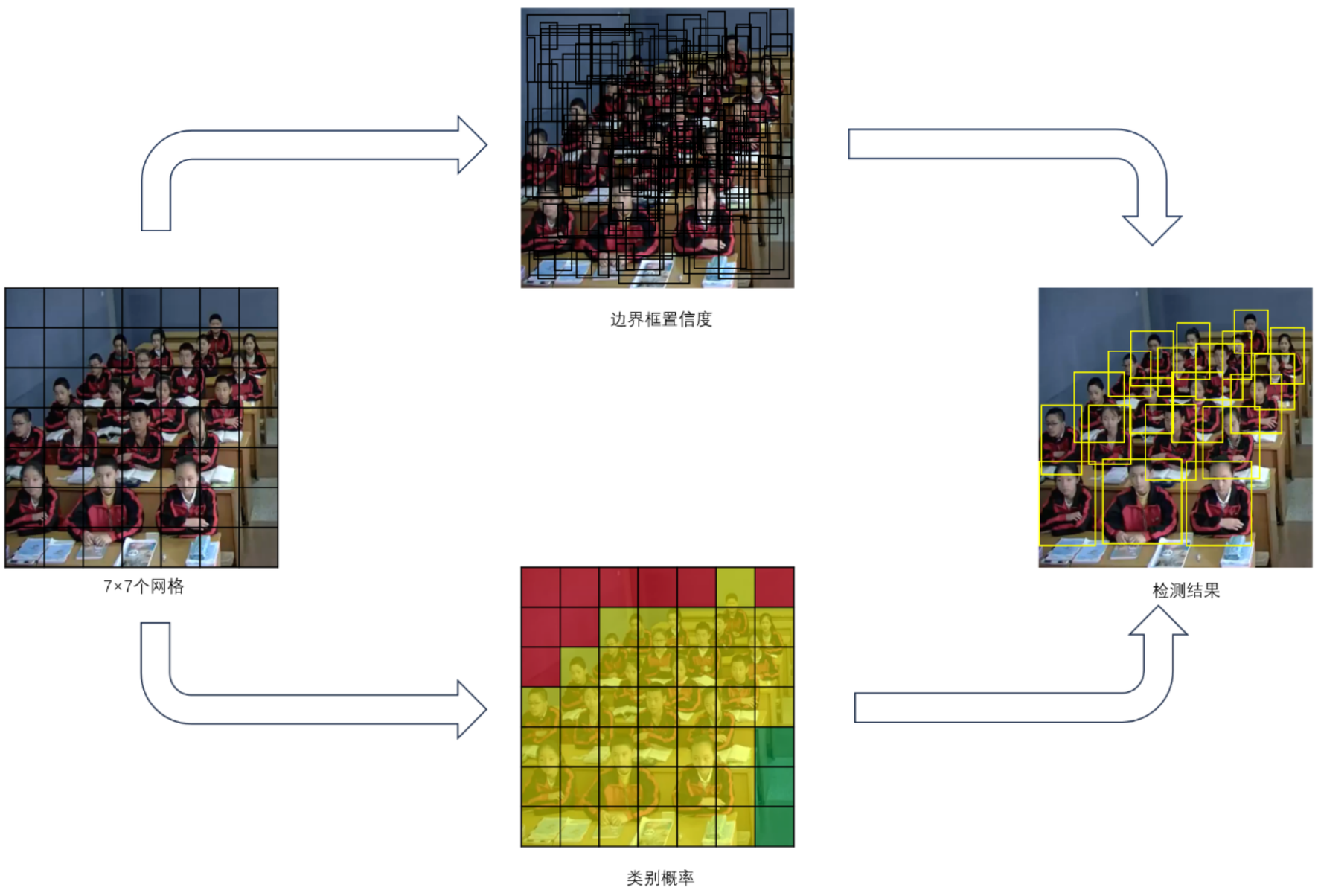


图 2-11 YOLOv1 算法结构

Fig.2-11 YOLOv1 algorithm structure

到了 2017 年，Redmon 团队又提出了 YOLOv2 算法[55]，对 YOLOv1 进行了大幅的改进。YOLOv2 在每个卷积层后引入了批量归一化（Batch Normalization，BN），采用了基于锚点的方法，并使用 K-Means 聚类算法来提取边界框的尺寸。

此外，YOLOv2 去除了网络的全连接层，引入了 pass-through 层来结合浅层和深层特征，同时采用了不同尺寸的输入图像进行训练，这些改进大幅提升了算法的精度和速度，并扩展了可识别的物体种类。

2018 年，Redmon 及其团队再次改进了原算法，提出了 YOLOv3 算法[56]。这一版本吸收了 ResNet 的残差结构，加深了网络的层数，以提高分类的准确性。YOLOv3 引入了多尺度输出，改善了对小目标的检测能力。在分类方法上，YOLOv3 从原来的 Softmax 方法转变为适用于多标签分类任务的 Logistic 回归。

2020 年，Bochkovskiy 等人提出了 YOLOv4 算法[57]，该算法使用了 Mosaic 数据增强技术。这种技术对提升小目标检测效果特别有效。YOLOv4 在其主干网络中借鉴了跨阶段局部网络（Cross Stage Partial Networks，CSPNet）的思想，并设计了 CSP 结构，这一结构旨在保留低层特征信息，从而增大检测精度。

同年，Ultralytics 公司推出了 YOLOv5 算法[58]，该算法选用了 SiLU 作为激活函数，增强了模型的稳定性。此外，YOLOv5 还对 CSP 结构进行了优化，设计出两种新的 CSP 结构，分别集成在用于特征提取的 Backbone 网络和随后的 Neck 网络中。YOLOv5 的 Neck 网络采用了 FPN 与像素聚合网络（Pixel Aggregation Network，PAN）的结合，该设计的目的是将高层特征向下传递以丰富低层语义信息，同时将低层的定位特征向上传递。这种特征金字塔的构建不仅融合了丰富的语义信息，还保留了精确的定位信息，显著提升了对多尺度目标的检测能力。

2.5 注意力机制及其发展史简述

注意力机制是近年来在深度学习目标检测领域应用越来越广泛的算法，其最初的设计灵感来自于灵长类动物视觉系统在复杂场景中迅速聚焦于关键点的能力[59]。一般而言，该方法通过构建查询（query）、线索（key）及输入值（value）之间的相互关系，通过注意力池化层有偏向性地选择某些输入。至今，计算机视觉中的注意力机制已经演化出三种主要方法：空间注意力机制、通道注意力机制和自注意力机制。

2014 年，DeepMind 团队首次提出空间注意力机制模型 RAM（Residual Attention Module）[60]，并开创性地将其与深度神经网络相结合，能够自适应地选择图像空间中的重要区域。2017 年，Hu 等人提出了 SE-Net（Squeeze-and-Excitation

Network）[61]，这是一种通道注意力机制，它通过自适应地调整通道间规模捕获通道间的相关性，显著提升了模型性能。同年，自注意力机制最初在自然语言处理（Natural Language Processing，NLP）领域中被成功应用[62]，随后，发展出了基于自注意力的 Transformers 架构[63]。Transformers 在 NLP 领域取得显著成果之后，Dosovitskiy 等人在 2020 年提出了 ViT[64]，创新性地将 Transformers 应用到计算机视觉任务中，这标志着注意力机制在计算机视觉领域的重大突破。

2.6 本章小结

本章首先介绍了神经网络的起源、理论基础和网络模型的数值稳定性理论及其优化方法，其次介绍了卷积神经网络的理论基础，重点介绍了基于深度学习理论的分类、目标检测算法，最后简要介绍了注意力机制的基本原理及其发展历程。这些内容将为后续章节实验部分所用到的方法提供坚实的理论支持。

第 3 章 学生行为数据集

学生课堂状况分析的方法源于面部表情识别及行为分析，依赖于丰富的学生课堂行为数据集的支持。部分研究者选择使用公开的表情数据集来验证和分析他们的算法。然而，随着对这些算法研究的不断深入，这些公开数据集已不足以满足针对学生课堂行为的实验需求。遗憾的是，针对教学环境中的学生课堂行为数据集相对较少，在公开数据集中几乎难以找到该领域的相关数据集。

为了解决上述问题，本文对学生课堂行为的分类方式进行了相关研究，并构建了包含 12500 张图像、49 个单人学生课堂行为视频片段、101 个密集教室学生课堂视频片段和 17 个学生课堂行为长视频的学生行为（Student Action，SA）数据集。该数据集对本文后续的研究具有重要的支撑作用。

3.1 学生课堂行为数据集研究现状

目前，在学生课堂状况分析领域，有一部分研究者搭建了数据集（库）。

刘永娜等人[65]构建的北京师范大学学习情绪数据库（Beijing Normal University Learning Affect Database，BNU LAD）包含了 144 名学习者的情绪数据，涵盖了愉快、好奇、困惑、厌烦、专注、走神和疲惫七种主要的学习相关情绪，这些情绪都进行了多标签和多强度的标注。

Ramón 等人[66]收集了 25 位参与者的面部图像和脑电波数据构建了数据集，这些参与者均为大学生，且年龄分布在 18 至 47 岁之间，包括 18 位男性和 7 位女性。实验将这些学生分成 A、B 两组进行。A 组包括 18 名学生，他们参与了与 Java 编程相关的任务，以诱发学习情绪，如兴趣、参与、兴奋与专注；B 组包括 7 名学生，他们观看了关于野生动物、自然景观等不同主题的视频，以激发学生的无聊和放松等情绪。

Bian 等人[67]构建的在线自发学习情绪数据库（Online Learning Spontaneous Facial Expression Database，OL-SFED）记录了 82 名年龄在 17 至 26 岁之间的学生的数据。这些学生通过观看不同类型的视频，激发了愉悦、困惑、疲劳、分心和中性五种关键的学习情绪，并对这些情绪进行了采集。

孔玺[68]独立创建了一个面向学习者的面部表情数据集。该工作将学习情绪

分类为常态、高兴、愤怒、悲伤、恐惧、专注和走神七种类型。该数据库以山东某高校数字媒体专业的 70 名学生为实验对象，让他们观看不同类型的数字学习内容，以此来构建一个智慧学习环境下的学习情绪数据集。

3.2 SA 数据集的采集

3.2.1 数据采集来源及环境

SA 数据集的数据来源有两种，其一是网络上可收集的课堂视频，如图 3-1 所示；其二是本文使用高清摄像头采集的普通教室学生课堂行为视频，如图 3-2 所示。前者包含了大、中、小学生（小学生占比较少）的日常课堂学习视频，后者为本文招募的在校大学生志愿者参与的采集实验视频。采集视频教室的环境与真实的课堂环境基本相同，光照为普通教室日光灯与窗外正常日光。

图 3-1 网络收集的视频帧

Fig.3-1 Video frames collected from the internet

图 3-2 自采视频帧

Fig.3-2 Self collected video frames

3.2.2 志愿者、实验材料及数据采集过程

本实验共招募了 29 名 20～30 岁的在校大学生志愿者参与实验。实验准备了分辨率为 1920×1080 的高清摄像头、学生日常使用的笔、笔记本、每人一部手机及“学生课堂行为实验指导手册”。其中“学生课堂行为实验指导手册”用于培训志愿者。数据采集过程如表 3-1 所示。

表 3-1 数据采集过程

Table.3-1 Data collection process

采集过程	采集细节
实验准备	向志愿者分发“学生课堂行为指导手册”及其他实验材料，并介绍实验流程；工作人员架设摄像头
数据采集	开始实验并使用摄像头采集视频
数据确认	志愿者观看视频并确认自己在实验中的课堂行为
学生行为标注	标注人员对采集到数据的学生课堂行为进行标注

3.3 SA 数据集的构建方法

本文对上述采集到的视频数据首先进行了抽帧，之后滤除了相似的图像，然后进行了数据增强及数据标注，数据标注采用了两种粒度的学生行为分类方法。

3.3.1 学生行为分类方法

在数据采集前，本文规划了 12 种典型的学生课堂行为：写，站，睡觉，读，举手，玩手机，正常听课，低头，左、右、左后、右后交头接耳，并且指导志愿者保持这些行为出现的频率和持续时间的相对均衡，以确保数据平衡性。

图 3-3 SA 数据集双粒度划分的学生课堂行为

Fig.3-3 Student classroom behavior based on double granularity partition of SA dataset

最终，本文将左、右、左后、右后交头接耳 4 种学生行为合并成左顾右盼。经过数据处理，最终将 SA 数据集分类为细粒度的 9 种：写，站，睡觉，读，举手，玩手机，正常听课，低头，左顾右盼和粗粒度的 3 种：未参与（包含睡觉、

玩手机、左顾右盼、低头），参与（包含正常听课、读、写），互动（包含站、举手），如图 3-3 所示。

3.3.2 数据增强与标注

在对采集到的所有视频数据进行抽帧并滤除相似的图像之后，为了增强数据集的鲁棒性，本文采用了如下数据增强方法：裁剪、镜像及调整图像的亮度、色调、对比度。在 SA 数据集中，使用上述每种增强方法的图像数量基本相同，而原始图像数据与增强后的图像数据比例约 1:1。

对网络上采集到的数据，学生课堂行为的标注对照了 3.3.1 节的标注方法，由经过系统训练的人员完成；对自采数据，经过志愿者本人对自己在某时间段的课堂行为进行确认并记录后，由标注人员完成标注。

本文使用的数据增强和标注工具为 EasyData[1]平台。

3.4 SA 数据集的数据统计

构建完成的 SA 数据集包含 12500 张双粒度标注的图像、49 个单人学生课堂行为视频片段、101 个密集教室学生课堂视频片段和 17 个学生课堂行为长视频。在标注的图像中，训练集图像共 11160 张，测试集图像共 1340 张，为避免训练集和测试集图像场景重复，它们分别使用不同场景下的视频处理所得。双粒度标注的图像统计得到细粒度分类：写 6110 人，站 9112 人，睡觉 2298 人，读 11702 人，举手 3739 人，玩手机 5626 人，正常听课 24138 人，低头 5357 人，左顾右盼 11586 人；粗粒度分类：未参与 24867 人，参与 41950 人，互动 12851 人，基本保证了数据均衡。SA 数据集统计数据如图 3-4 所示，其中，粗、细粒度分别用橙色和蓝色表示。此外，SA 数据集中包含的 10 个长视频为采集的真实学生课堂教学视频，均剪辑为 20 分钟。它们不参与数据集中图像的制作，即不参与模型训练，以验证最终的算法在实际教学课堂中的效果。

[1] https://ai.baidu.com/easydata/

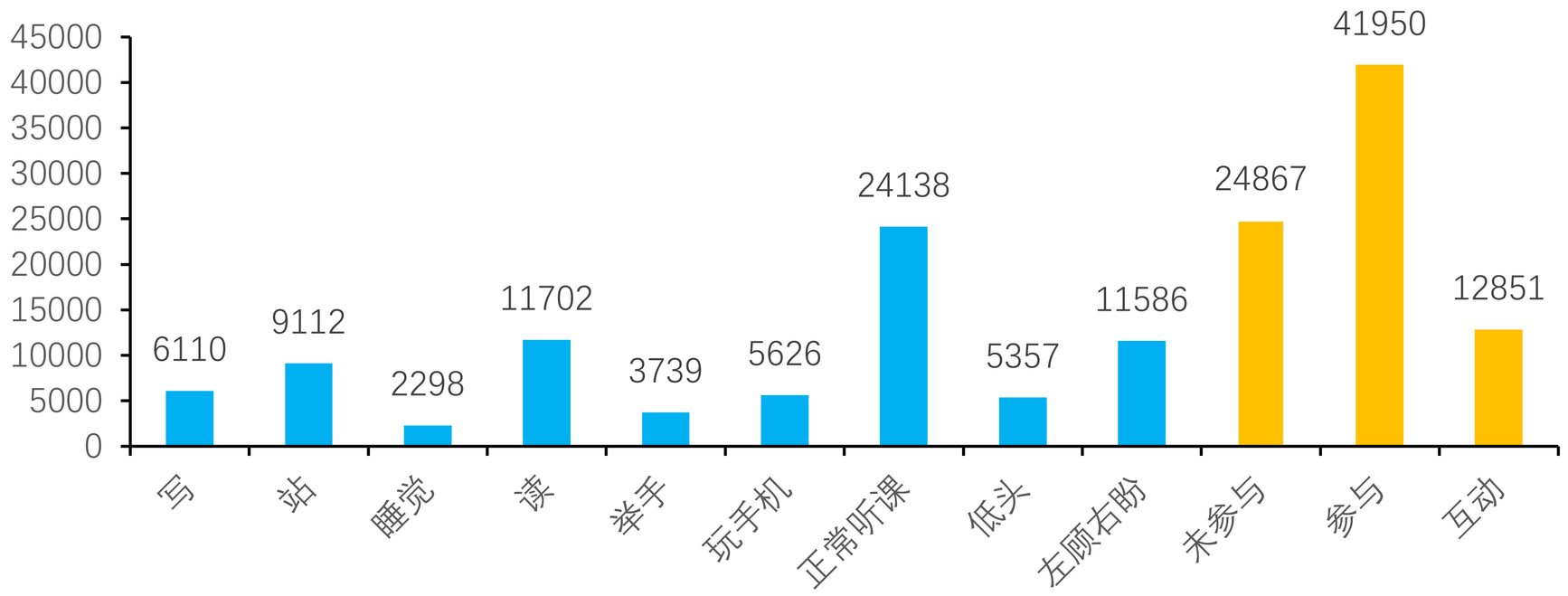


图 3-4 SA 数据集学生课堂行为统计图

Fig.3-4 Student classroom behavior statistics bar chart of SA dataset

此外，SA 数据集将单人图像定义为简单难度，2-20 人图像定义为中等难度，20 人以上图像定义为困难难度，如图 3-5 所示。SA 数据集中简单、中等、困难图片的比例约 2：6.5：1.5。

图 3-5 数据集的三种难度

Fig.3-5 Three degrees of difficulty of SA datasets

3.5 本章小结

构建学生课堂行为数据集是学生课堂状况分析领域研究中的一项关键任务。本章探讨了在课堂环境中设计和搭建学生课堂行为数据集的过程，并针对传统数据标注方法的局限性进行改进，设计了新的数据标注方式，即使用了两种粒度划分的学生课堂行为，该方法的作用将在第 5 章进行论述。

SA 数据集有三类数据：学生课堂行为图像、学生课堂行为视频片段和学生课堂行为长视频。该数据集中的学生课堂行为图像为本文算法的研究提供了训练和测试样本，学生课堂行为长视频则为模型验证提供了重要的数据支持。

第 4 章 基于改进 YOLOv5 的学生课堂行为识别算法

由于学生课堂上往往存在遮挡、后排目标小的情况，学生课堂行为识别在实际应用中具有相当的挑战性。目前，研究者们已经提出了一些算法解决上述问题，但它们的精度往往较低。本文在构建好的 SA 数据集上对比了多个工业界和学术界比较流行的目标检测算法，选择了其中精度最高，大小较合适的 YOLOv5m 算法作为基础网络，提出了 SGE-BI 双注意力机制模块，并将该模块嵌入 YOLOv5m 算法的 neck 网络中，将原算法的精度、mAP@0.5 和 mAP@0.5:0.95 值分别提高了 1.9%、0.5%和 0.8%，大大提高了学生课堂行为识别的准确性。

4.1 SGE-BI 双注意力机制模块

4.1.1 理论基础

尽管基于卷积神经网络（CNN）的方法在目标检测任务中已经取得了显著的效果，但这些算法在结构上仍有一定的局限性，从而影响它们的最终检测性能。首先，图像中的完整特征（比如一个学生）是由众多子特征（比如手部的姿态）构成的。CNN 通过下采样提取这些子特征，并以分组的形式将它们分布在每层的特征中，这样的处理方式可能会因为图案相似性或背景干扰而导致错误的定位。其次，CNN 在捕捉图像中长距离的上下文联系方面存在困难。近年来不断发展的注意力机制提供了一种有效的补充，它能够帮助改善 CNN 在上述方面的不足。基于以上原因，本文提出 SGE-BI 双注意力机制模块并添加在 CNN 网络中，取得了很好的效果。

4.1.2 SGE-BI 模块

SGE-BI 模块是一个双注意力机制模块，主要分为两个部分：SGE 模块和 BiFormer 模块。SGE 模块对输入图像的每个语义组的每个空间位置生成一个关注因子调整每个子特征的重要性，以增强网络模型的学习能力并抑制噪声[69]；BiFormer 模块采用了多头自注意力机制[70]，可以捕捉图像中长距离的上下文依赖，提高模型的识别精度[71]。SGE-BI 模块结构如图 4-1 所示，其中 Patch Embedding[64]的作用是将原始的二维图像转换成一系列的一维向量，类似于 CNN 中的卷积操作。Patch Merging[72]的作用是将多个小特征图合成一个大特征图，以

增大感受野，之后通过进行特征提取，模拟出类似 CNN 中池化的操作。

在 SGE 模块中，特征图被分割为 8 个语义组，由于每个语义组仅包含两个参数，因此 SGE 模块的参数量极其有限，对模型整体运行速度的影响微乎其微；而在 BiFormer 模块中，参数量大约为 26M 个，其使用的计算资源大致与 YOLOv5m 相当。

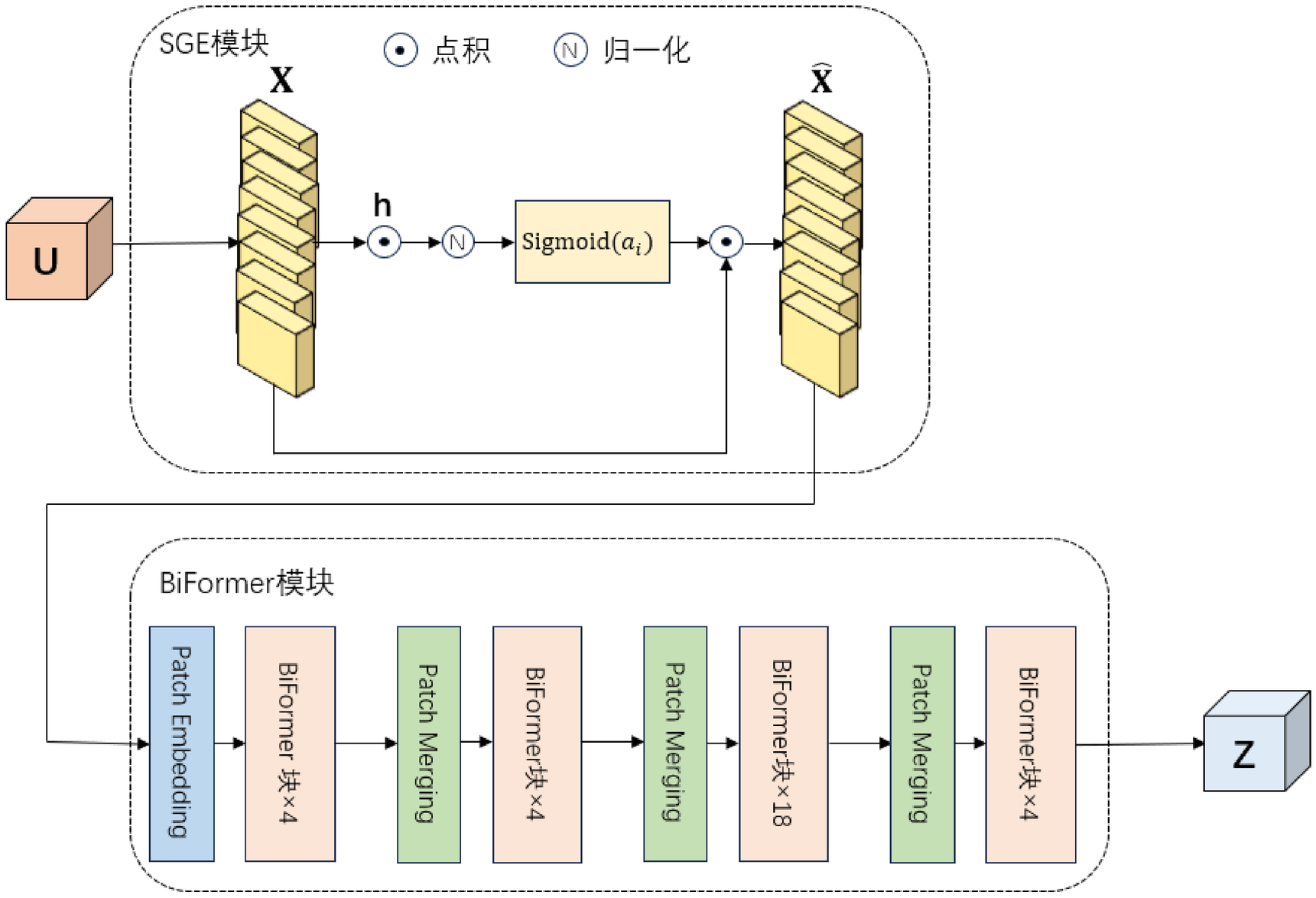


图 4-1 SGE-BI 结构

Fig.4-1 Structure of SGE-BI

如图 4-1 所示，假如模块的输入为U，它有C个通道，大小为$H \times W \times C$，Z是输出的特征图，通过 SEG-BI 模块的算法（算法 1）可以分为以下两个步骤：

步骤一：将U输入 SGE 模块。首先沿通道维，将U分割成e组（本文e设置为 8），其中某一个组在每个位置的向量表示为：$X = \{x_1, x_2 \dots x_m\}, x_i \in \mathbb{R}^{C/e}, m = H \times W$。定义一个组内平均函数$\mathrm{F_{avg}}(\cdot)$，用以表示该组空间的全局语义特征：

$$h = \mathrm{F_{avg}}(x_i) = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{4-1}$$

之后，将局部特征x_i与全局特征h进行点积：

$$s_i = h \cdot x_i = \| h \| \| x_i \| \cos\theta_i \tag{4-2}$$

其中，s_i表示每个局部特征x_i与全局特征h的相似度。易知，当x_i的大小和方向更接近全局特征h时，s_i会取得更大的初始值。为避免误差，对s_i进行归一

1. SGE-BI 算法流程

算法 1： SGE-BI.

输入： 特征图$U(H,W,C)$,分组数e;

循环控制$t = 0$.

过程：

1: 将U分割为e组;
2: **repeat**
3: 选取一个组 $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2} \ldots \boldsymbol{x_m}\}$;
4: 根据式（4-1）计算全局语义特征向量$\boldsymbol{h}$;
5: 将局部特征x_i与全局特征$\boldsymbol{h}$点积得到相似度s_i;
6: 归一化s_i得到$\hat{s}_i$;
7: 初始化一对参数(γ, β)用以网络学习;
8: 根据式（4-3）得到增强系数a_i;
9: 通过 Sigmoid 函数优化a_i;
10: 将增强系数a_i与局部特征x_i点积得到增强向量组$\hat{X}_1$;
11: **until** 执行 e 次
12: 得到特征图$L = \{\hat{X}_1, \hat{X}_2 \ldots \hat{X}_e\}$;
13: 将L执行 Patch Embedding 操作得到L_u;
14: 对L_u执行 4 次算法 2,更新为特征图L_p;
15: **while** $t < 3$ **do** # 使用如下循环对特征图L_p进行迭代
16: 执行 Patch Merging 操作;
17: 执行 4 次算法 2;
18: $t = t + 1$;
19: **if** $t = 1$ **then**
20: 执行 Patch Merging;
21: 执行 18 次算法 2;
22: $t = t + 1$;
23: **end if**
24: **end while**

输出： Z

化，得到$\hat{s}_i$，并引入一对参数γ，β用以网络学习：

$$a_i = \gamma \hat{s}_i + \beta \tag{4-3}$$

将$\hat{s}_i$运算得到的增强系数a_i经过一个 Sigmoid 函数进行优化，再与原始的局部特征x_i点积，得到利用该组内全局特征增强之后的初始向量组$\hat{X}_1$。所有组计算完成之后，得到增强后的特征图$L = \{\hat{X}_1, \hat{X}_2 \ldots \hat{X}_e\}$。

2. BiFormer 块算法流程

算法 2：BiFormer 块.

输入：特征图L_u;

深度可分离卷积操作DWConv(·);

层归一化函数LN(·);

矩阵拼接函数gather(·);

多层感知机 MLP(·).

过程：

1: $M = \mathrm{DWConv}(L_u) + L_u$;

2: $Y = \mathrm{LN}(M),\ Y \in \mathbb{R}^{H_r,W_r,C_r}, H_r == W_r$;

3: 将Y分割为S^2块;

4: 根据式（4-5）对Y进行线性变换得到Q, K, V;

5: 根据式（4-6）得到邻接矩阵A_r;

6: 根据式（4-7）得到路线索引矩阵I;

7: 根据式（4-8）拼接Q的关联矩阵,得到K_g, V_g;

8: 根据式（4-9）得到 BRA 模块的输出O;

输出：$O' = \mathrm{MLP}(\mathrm{LN}(O)) + O$

步骤二：将L输入 BiFormer 模块。先进行一次 Patch Embedding 操作，得到L_u，并将其输入一个 BiFormer 块（算法 2），结构如图 4-2 所示。

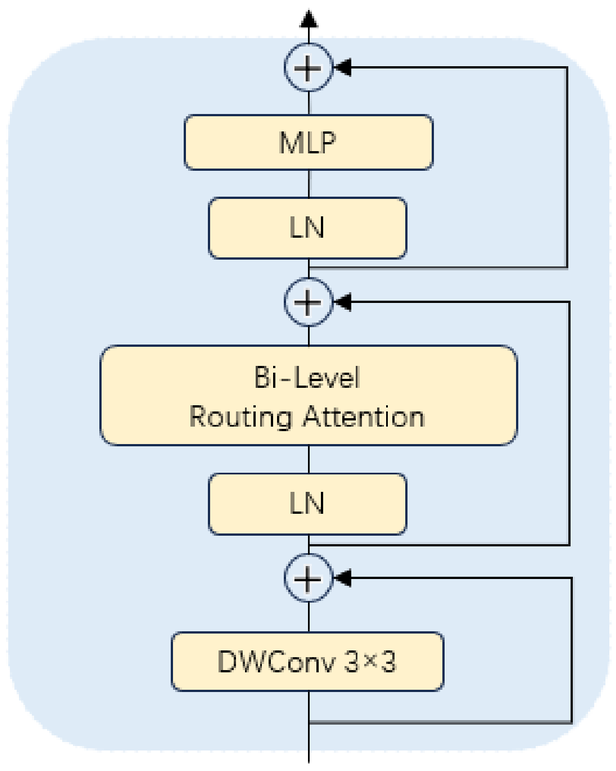


图 4-2 BiFormer 块细节

Fig.4-2 BiFormer block details

将图 4-2 中的深度可分离卷积（Depthwise Convolution）定义为DWConv(·)，多层感知机（Multilayer Perceptron）定义为MLP(·)。

定义 LN(·)为层归一化（Layer Norm，LN）函数：

$$\mathrm{LN}(y_i) = \alpha \frac{y_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \theta \tag{4-4}$$

其中，μ和σ分别表示神经元输出y_i的均值和标准差，ϵ是一个较小的常数，用于防止分母为零，α和θ是两个可学习的参数。

图 4-2 中 Bi-Level Routing Attention（BRA）模块操作如下：

假如输入特征图为Y，为避免字母重复，取r为重塑符号，令其大小为$H_r \times W_r \times C_r$。将特征图重塑为$Y_r$，大小为$S^2 \times \frac{H_r W_r}{S^2} \times C_r$。其中，$H_r = W_r$，$S^2$为$Y_r$被切割的块数。然后，将$Y_r$进行线性变换得到查询（$Q$），键（$K$），值（$V$），$Q$，$K$，$V \in \mathbb{R}^{S^2 \times \frac{H_r W_r}{S^2} \times C_r}$，即：

$$Q = Y_r W_q, K = Y_r W_k, V = Y_r W_v \tag{4-5}$$

其中，W_q，W_k，W_v是将Q、K、V线性变换的权重向量。之后，得到区域之间的关联度邻接矩阵A_r：

$$A_r = Q_r {K_r}^{\mathrm{T}} \tag{4-6}$$

其中，Q_r和K_r是区域级的查询和键，$Q_r, K_r \in \mathbb{R}^{S^2 \times C}$。再导出路线索引矩阵$I$：

$$I = \mathrm{topkIndex}(A_r) \tag{4-7}$$

其中，topkIndex(·)为建立I的第i行包含第i个区域最相关的k个区域的索引函数，$I \in \mathbb{N}^{S^2 \times k}$。本文中，$k$设置为 4。

为提高内存利用率，定义一个拼接矩阵的函数gather(·)，将这些与Q相关的K，V拼接起来：

$$K_g = \mathrm{gather}(K, I), V_g = \mathrm{gather}(V, I) \tag{4-8}$$

其中，$K_g, V_g \in \mathbb{R}^{S^2 \times \frac{kH_r W_r}{S^2} \times C_r}$。之后，应用多头自注意力机制在键值对集合上：

$$O = \mathrm{Attention}\left(Q, K_g, V_g\right) + \mathrm{LCE}(V) \tag{4-9}$$

其中，函数LCE(·)为参数化的深度可分离卷积，如文献[73]中所述。

如图 4-1 所示，BiFormer 模块遵循大多数计算机视觉中的 Transformers 架构，即四级金字塔结构。特征图在四个 BiFormer 块间进行 Patch Merging 操作，

最后得到输出Z。在本文中，每级金字塔结构包含的 BiFormer 块的数量分别是 4，4，18，4。

4.2 基于 YOLOv5 的改进模型

本文使用 SA 数据集训练并测试了目前在目标检测领域的多种流行网络模型之后，选择了识别准确度最高、参数量较合适的 YOLOv5m 构建学生课堂状况分析系统，对比实验结果在 4.3 节。

Ultralytics 公司在 2021 年推出了 YOLOv5 模型，其因高精度和快速处理能力而在工业和学术领域得到广泛应用。YOLOv5 的架构包括输入端、backbone 网络、neck 网络以及输出端。输入端利用 Mosaic 数据增强技术通过裁剪、混合和拼接等方式来扩充训练数据集；backbone 网络采用 Focus 模块进行快速下采样并增强特征提取，接着利用融合特征的 Bottleneck 结构和 C3 结构来进一步优化特征提取能力，并通过 SPPF 模块实现局部与全局特征的融合；neck 网络结合了特征金字塔（FPN）和像素聚合网络（PAN）的设计，促进浅层和深层语义特征的融合；输出端采用三个不同尺度的检测头来提升检测的准确性。目前，YOLOv5 共有五个版本：YOLOv5n、YOLOv5s、YOLOv5m、YOLOv5l 和 YOLOv5x，它们的参数量分别约为 190 万、720 万、2100 万、4600 万和 8650 万。这些模型深度（即网络层数）和宽度（即通道数）依次递增，随之检测精度越高，速度越慢。本文选择了参数量和性能均衡的 YOLOv5m 版本。为了增强 YOLOv5m 模型在学生课堂行为识别任务上的性能，本文对该模型进行了以下优化：首先，添加了一个 SGE-BI 模块到模型的 neck 网络，并验证了该模块能显著提升 YOLOv5m 的识别能力；接着，逐步增加了 neck 网络中的 SGE-BI 模块数量。实验结果表明，当 neck 网络中包含两个 SGE-BI 模块时，模型的检测效果达到最优。第 4.3 节将展示这些优化方法的对比实验结果。图 4-3 展示了改进的 YOLOv5 模型结构。

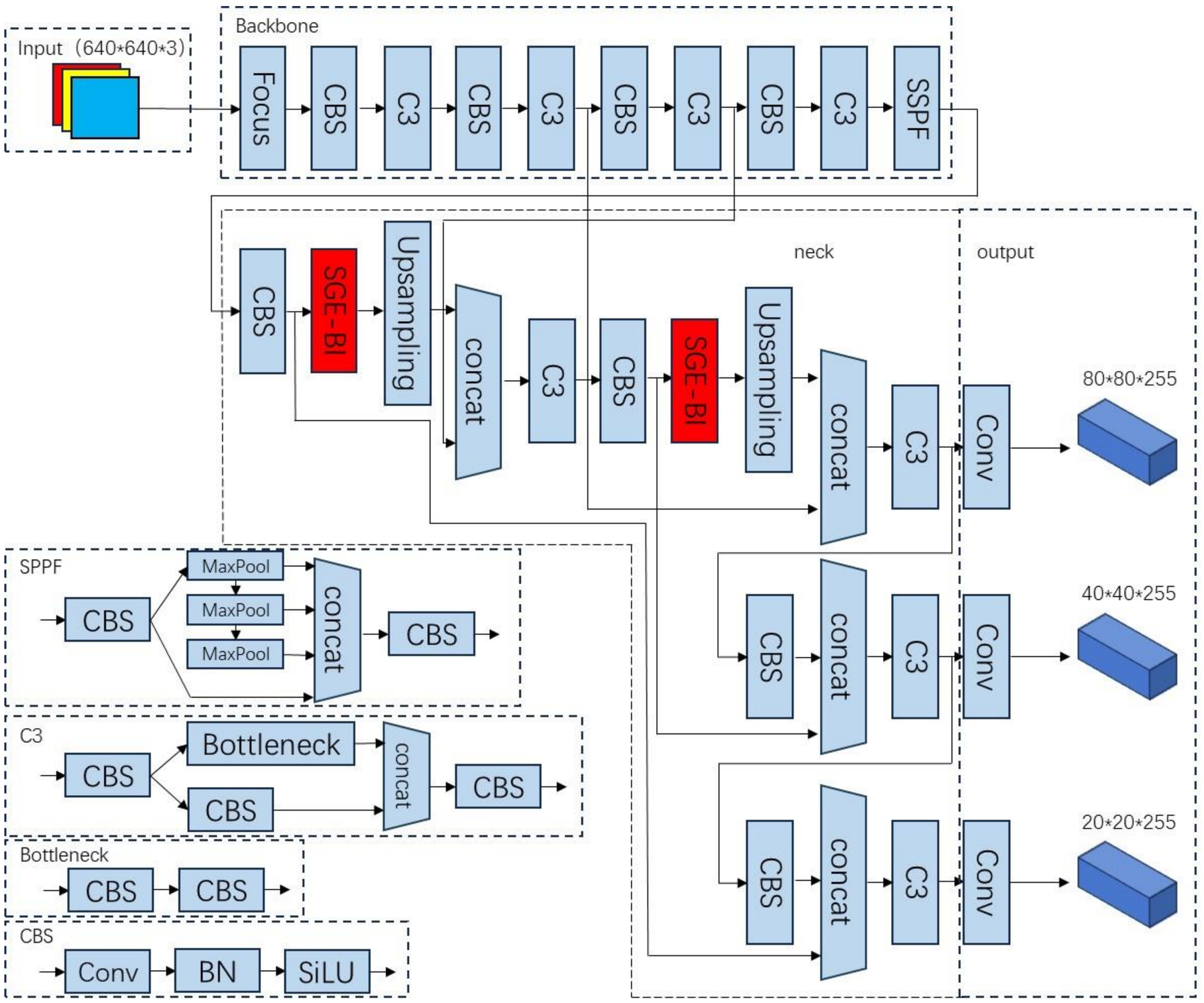


图 4-3 改进的 YOLOv5 结构

Fig.4-3 Improved YOLOv5 structure

4.3 实验及分析

4.3.1 评价指标、实验环境与设计方法

在目标检测算法中，评价一个模型好坏的指标通常有精度（Precision，P），召回率（Recall，R）及平均精度均值（Mean Average Precision，mAP），公式如下：

$$P = \frac{TP}{TP + FP} \tag{4-10}$$

$$R = \frac{TP}{TP + FN} \tag{4-11}$$

$$mAP = \frac{\sum AP}{N} \tag{4-12}$$

其中，TP、FP、FN分别表示正确识别的正样本数（真正例）、错误识别的正样本数（假正例）、错误识别的负样本数（假负例），N表示类别数。mAP 结合了

P 和 R，是 P-R 曲线下的面积，故成为目标检测算法性能的重要指标。

本文使用 mAP@0.5 和 mAP@0.5:0.95 两种 mAP 指标，其中，mAP@0.5 指在 IoU（候选框与标记框重合率）阈值为 0.5 时的 mAP 值，mAP@0.5:0.95 指 IoU 阈值从 0.5 到 0.95 每间隔 0.5 计算一次，一共 10 个 IoU 阈值上的平均 mAP 值。后者比前者的标准更加严格。

本文使用目前在目标检测领域的一系列主流算法训练 SA 数据集，通过对比各算法的 mAP 值，选择了检测效果最好的 YOLOv5m，并将 SGE-BI 模块添加在其 neck 网络中，再进行消融实验证明了算法改进的有效性，之后，通过在 neck 网络中堆叠 SGE-BI 模块，验证了在该网络中添加两个 SGE-BI 模块效果最好。

本文使用 Ubuntu20.04 操作系统，Python3.9 编程，PyTorch 机器学习库（版本 1.13.1，CUDA 版本 11.6）；使用的 CPU 型号为 Intel Core i9-10900X，GPU 型号为 NVIDIA GeForce RTX4090。图像在训练前，统一被降采样至 640×640 尺寸，初始学习率设置为 0.01，batch size 设置为 30。

4.3.2 多种目标检测模型的对比实验

为了挑选出适合学生课堂行为识别的目标检测模型，本研究在 SA 数据集上训练并测试了九种不同的流行算法，包括本文第 2 章介绍的 SSD、R-CNN 系列、YOLOv5s 以及采用了 Transformers 方法的 Deforamble-DETR 等。这些算法既包括基于候选区域的方法，也包括基于回归的方法，并且对比了近几年较流行的集成了注意力机制的最新算法，代表了当前学术界和工业界认可的最高精度和最佳性能的目标检测模型。表 4-1 展示了上述算法与本文提出方法的对比实验结果。

表 4-1 多种模型在 SA 数据集上的对比实验

Table 4-1. Comparative experiments of different modules on the SA dataset

算法	mAP@0.5	mAP@0.5:0.95
RTMDet[74]	0.918	0.752
SSD	0.919	0.769
Deformable-DETR[75]	0.935	0.780
Two-Stage Deformable-DETR[75]	0.942	0.791
Cascade R-CNN	0.923	0.797
RetinaNet[76]	0.936	0.798
Faster R-CNN	0.933	0.800
YOLOv5s	0.946	0.813
TOOD[77]	0.948	0.815
YOLOv5m	0.948	0.828
本文方法	**0.953**	**0.836**

从表 4-1 的对比实验数据可以看出，在 SA 数据集上，YOLOv5m 在识别能

力上超过了参数更少的 YOLOv5s，而且相比于其他目标检测算法，YOLOv5m 的性能也更为突出。通过与原算法的比较，本研究提出的方法在 mAP@0.5 上提升了 0.5 个百分点，在 mAP@0.5:0.95 上提升了 0.8 个百分点。

4.3.3 消融实验

本文为了评估 SGE-BI 双注意力机制模块在学生课堂行为检测中的效果，在 SA 数据集上进行了消融实验。这些实验对比了原算法、增加了 SGE 模块的改进算法、增加了 BiFormer 模块的改进算法、增加了单个 SGE-BI 模块的改进算法以及本文提出的方法。实验成果展示在表 4-2 中。

根据表 4-2 的数据可以观察到，尽管在原算法中添加 SGE 或 BiFormer 单注意力机制模块会导致召回率有所降低，但模型的精度、mAP@0.5 和 mAP@0.5:0.95 指标均有显著提升。这表明在原始算法中加入 SGE 或 BiFormer 单注意力机制模块能有效提高模型的检测能力。当添加单个 SGE-BI 模块后，与原算法相比，召回率虽降低了 0.7%，但精度、mAP@0.5 和 mAP@0.5:0.95 分别提升了 1.8%、0.4%和 0.7%，这证明了 SGE-BI 双重注意力机制模块的加入能显著增强模型识别能力。对比单独添加 SGE 或 BiFormer 模块的改进算法，采用添加 SGE-BI 双注意力机制模块的算法在召回率上相当，而在精度、mAP@0.5 和 mAP@0.5:0.95 上则实现了显著的提高，这验证了本文提出的 SGE-BI 模块在 YOLOv5m 网络上的优越检测效果。此外，使用本文方法对 YOLOv5m 进行改进后，与原算法比较，可以发现在召回率基本持平的前提下，精度、mAP@0.5 和 mAP@0.5:0.95 分别提高了 1.9%、0.5%和 0.8%，改进后的算法明显提升了原算法的检测性能。

表 4-2 消融实验

Table 4-2. Ablation experiments

算法	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv5m	0.935	**0.883**	0.948	0.828
YOLOv5m+SGE	0.945	0.878	0.949	0.833
YOLOv5m+BiFormer	0.943	0.878	0.949	0.832
YOLOv5m+SGE-BI	0.953	0.876	0.952	0.835
本文方法	**0.954**	0.881	**0.953**	**0.836**

4.3.4 添加多个 SGE-BI 模块的对比实验

本文进一步探讨了在 YOLOv5m 模型中堆叠更多 SGE-BI 模块是否能够持续提升算法性能。在已经集成了两个 SGE-BI 模块的 YOLOv5m 模型上，本研究尝试继续添加该模块。表 4-3 展示了在本文方法基础上继续添加一个和两个 SGE-BI 模块（使得模型中共有三个和四个 SGE-BI 模块）后，在 SA 数据集上进行训练的实验结果对比。

表 4-3　YOLOv5 添加更多 SGE-BI 模块对比实验

Table 4-3. Comparative experiments on adding more SGE-BI modules in YOLOv5

算法	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv5m+3 SGE-BI	0.938	0.891	**0.954**	0.836
YOLOv5m+4 SGE-BI	0.939	**0.892**	**0.954**	0.836
本文方法	**0.954**	0.881	0.953	**0.836**

根据表 4-3 的数据，我们注意到在本文方法基础上额外添加一个或两个 SGE-BI 模块后，与原算法相比，召回率确实有所增加，但精度大致以相似的幅度降低，同时 mAP@0.5 和 mAP@0.5:0.95 的值几乎未变。当考虑到算法运行速度时，整体性能甚至有所降低。这一对比实验结果表明，YOLOv5m 模型在集成两个 SGE-BI 模块后性能达到最优。

4.4 本章小结

本文提出了一种基于两种注意力机制的 SGE-BI 双注意力机制模块，并将其添加在 YOLOv5 网络中，用于识别学生课堂行为。该模块通过对输入图像的每个语义组的每个空间位置生成一个关注因子调整每个子特征的重要性，以增强网络模型的学习能力和抑制噪声，并通过使用多头自注意力机制的方法增加了网络捕捉图像中长距离上下文依赖的能力，提高了模型的识别效果。通过对比实验，在 SA 数据集上将精度、mAP@0.5 和 mAP@0.5:0.95 值分别提高了 1.9%、0.5% 和 0.8%，明显提高了原算法的性能，满足了在实际课堂教学的复杂环境下精确识别学生课堂行为的要求，为学生课堂行为识别提供了一种有效的算法。在下一章，本文将基于使用该模型对真实学生课堂行为识别的数据构建一种课堂评价方法。

第 5 章 学生课堂质量评估算法

在学生课堂中，以何种方式评估课堂质量是一个重要的问题。当前，大多数研究仅对学生课堂行为进行简单的识别，而忽略了课堂质量评估算法的设计。为了更客观地评估学生课堂质量，本章在第 4 章学生课堂状况识别模型研究的基础上进一步设计了课堂质量评估算法。

5.1 赋权法

设计课堂质量评估算法，对学生课堂状况识别模型结果的量化至关重要。为了更准确地衡量各种课堂行为的重要性，权重的设定是其中的关键因素，它直接影响评估算法的判断能力。目前，关于权重设定的研究已有诸多成果，通常将权重设定方法分为两大类：主观赋权法和客观赋权法。

5.1.1 主观赋权法

主观赋权法是较早采用的确定属性权重的方法，依据的是领域专家基于个人经验和具体问题来合理分配各个权重，并据此进行综合评估以确定最终权重。该方法的优点是，由于参照了专家意见，避免了权重与实际属性重要性相差甚远的情况，但也存在明显缺陷，即过于主观，缺乏客观性。常用的主观赋权法包括专家调查法（Delphi 法）[78]、层次分析法（AHP 法）[79]等。

1. 专家调查法

Delphi 法是在 20 世纪 50 年代由美国兰德公司与道格拉斯公司合作开发的一种专家调查法。这种方法是主观赋权法的一种，它高度依赖于专家的经验，并且已经在多个领域受到了广泛的应用，越来越多的研究人员开始深入探究 Delphi 法。

实施 Delphi 法的过程包括以下几个步骤：（1）明确调查的问题和范围；（2）确定相关领域的专家并向他们提供问题的背景资料；（3）专家依据个人经验对问题作出判断，并阐述其理由；（4）调查者汇总专家意见，进行对比，并将结果反馈给专家以便进行第二轮咨询；(5)专家根据上一轮的汇总意见调整自己的观点，这一过程重复进行，直至专家意见不再变动；（6）对专家的最终反馈提出处理意见，并据此确定最终的 Delphi 法属性权重，公式如下：

$$Q = \frac{\sum_{i=1}^{n} q_i \cdot w_i}{\sum_{i=1}^{n} w_i} \tag{5-1}$$

其中，n表示进行判断的专家数量，Q为最终权值，q_i为第i位专家对该属性判断的权值，w_i为该专家打分所占的权值。

2. 层次分析法

假设给定的准则层属性集为$Q = \{q_1, q_2, \cdots, q_n\}$，将$q_i$和$q_j$对目标层结果影响的重要性比值表示为$d_{ij}$，$i, j \in 1,2, \cdots, n$。

对该比值进行量化，设定当属性q_i和q_j对结果同样重要时，$d_{ij} = 1$；当d_{ij}为介于1到9之间的正整数时，q_i相对q_j的重要程度与d_{ij}的大小成正比；当q_i相比q_j对结果的重要程度最大时，$d_{ij} = 9$。此外：

$$d_{ij} = \frac{1}{m}, m = 1,2, \ldots 9 \text{ 当且仅当 } d_{ji} = m \tag{5-2}$$

令$d_{ij} = a_{ij} / b_{ij} (i, j = 1,2, \cdots, n)$，通过分析$Q$中$n$个属性相互之间的重要性，对$a_{ij} / b_{ij}$逐个进行赋值，得到判断矩阵$D$，也称正互反矩阵：

$$\begin{aligned} D = (d_{ij})_{n \times n} &= \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \\ &= \begin{bmatrix} \frac{a_{11}}{b_{11}} & \frac{a_{12}}{b_{12}} & \cdots & \frac{a_{1n}}{b_{1n}} \\ \frac{a_{21}}{b_{21}} & \frac{a_{22}}{b_{22}} & \cdots & \frac{a_{2n}}{b_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{a_{n1}}{b_{n1}} & \frac{a_{n2}}{b_{n2}} & \cdots & \frac{a_{nn}}{b_{nn}} \end{bmatrix} \end{aligned} \tag{5-3}$$

接下来，对判断矩阵D进行一致性检验：

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{5-4}$$

其中，$\lambda_{\max}$是判断矩阵D的最大特征根，n是矩阵的阶数，CI是一致性指标。之后，我们对一致性比率指标CR进行判断：

$$CR = \frac{CI}{RI} \tag{5-5}$$

其中，RI为随机一致性指标，由统计学给出，如表 5-1 所示。

表 5-1 随机一致性指标

Table 5-1. RI indicators

矩阵阶数	1	2	3	4	5	6	7	8	9
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46

若$CR < 0.1$，则可认为判断矩阵的一致性可以被接受，否则需要对判断矩阵进行修正，并再次进行一致性比率指标的判断。

若CR满足条件，通过特征根法求解出Q的权重向量，即将$\lambda_{\max}$对应的特征向量进行归一化，得到属性$q_1, q_2, \cdots, q_n$对应的权重向量：

$$w = (w_1, w_2, \cdots, w_n)^{\mathrm{T}} \tag{5-6}$$

5.1.2 客观赋权法

客观赋权法则基于属性的真实数据来确定权重，权重的确定依据各属性所提供信息量的大小。如果某个属性对模型决策没有贡献，则该属性的权重可以忽略。相反，如果某属性对模型决策有着至关重要的影响，则应赋予较大的权重。客观赋权法的数据来源于客观实际，计算基于数学理论，因而具有较强的客观性。但此方法的不足在于，它没有考虑决策者的主观意愿，有时会导致计算出的权重与专家评估的权重存在较大差异，可能会低估重要属性的权重。常用的客观赋权方法包括熵值法[80]等。

1. 熵值法

熵被用作衡量事件不确定性的指标。如果一个事件的信息量越丰富，其不确定性越低，相应的熵值也越低，反之则熵值越高。因此，熵可以用来表示事件未来的不确定性程度。同时，熵也可以用来量化一个属性的离散程度，熵值较高意味着该属性在评价结果中的影响较大。熵值法是一种利用决策矩阵来确定各属性权重的方法。

假如一个问题中的决策矩阵为：

$$D = \left(d_{ij}\right)_{n\times m} \tag{5-7}$$

其中，D经过归一化处理，n为待测对象的个数，m为属性个数。

将第j个属性的熵作如下定义：

$$E_j = -\frac{1}{\ln n}\sum_{j=1}^{n} b_{ij}\ln b_{ij} \tag{5-8}$$

其中，$i = 1,2,\cdots,n$，$j = 1,2,\cdots,m$，$b_{ij} = \frac{d_{ij}}{\sum_{j=1}^{n} d_{ij}}$；于是，$0 \le E_j \le 1$。

定义第j个属性的熵权a_j：

$$a_j = \frac{1 - E_j}{m - \sum_{i=1}^{m} E_j} \tag{5-9}$$

根据公式(5-9)和熵函数的性质，我们可以观察到当测试对象在某一属性上的值一致时，熵值达到最大，即 1，而相应的熵权值则为 0。这表明该属性没有提供任何有效信息，可以考虑剔除该属性或将其权重设置为 0。相反，当测试对象在某一属性上的值差异显著时，熵值会减小，而熵权值会增大，这表明该属性提供了关键信息，应当赋予较高的权重。熵与熵权是相互反向的指标：熵值越高，熵权值越低。熵权值的大小直接反映了属性提供的有用信息量，并与测试对象有直接关联。

在熵值法中，确定各属性权重系数的步骤首先是将数据转换为非负值，以避免在计算熵值时因数据为负而无法进行对数运算。这种非负化一般通过对数据进行平移实现。随后，需要计算第j个属性下第i个待测对象占该属性的比重b_{ij}。最后，利用公式(5-8)和(5-9)计算该属性的熵值和熵权值，以此来确定第j个属性的权重。

5.2 双粒度层次分析加权法

由上述对 AHP 法的介绍可以看出，该方法是一种基于专家主观判断的权重分配方法。在这种方法中，专家依据其实践经验来判断准则层不同属性间的相对重要性，并据此确定它们对目标层结果的影响程度。然后，通过数学模型来检验所赋予权重的合理性，并计算出准则层属性的权重向量。例如，在学生课堂行为识别和评估的场景中，需要对学生行为的相对重要性进行赋权，这些权值代表了准则层的学生行为对目标层课堂表现的影响程度。

本文构建了双粒度的 SA 数据集，将学生行为划分为细粒度的 9 种，分别是写，站，睡觉，读，举手，玩手机，正常听课，低头，左顾右盼；以及粗粒度的 3 种：未参与（包含睡觉、玩手机、左顾右盼、低头），参与（包含正常听课、读、

写），互动（包含站、举手）。目标层和准则层如图 5-1 所示。

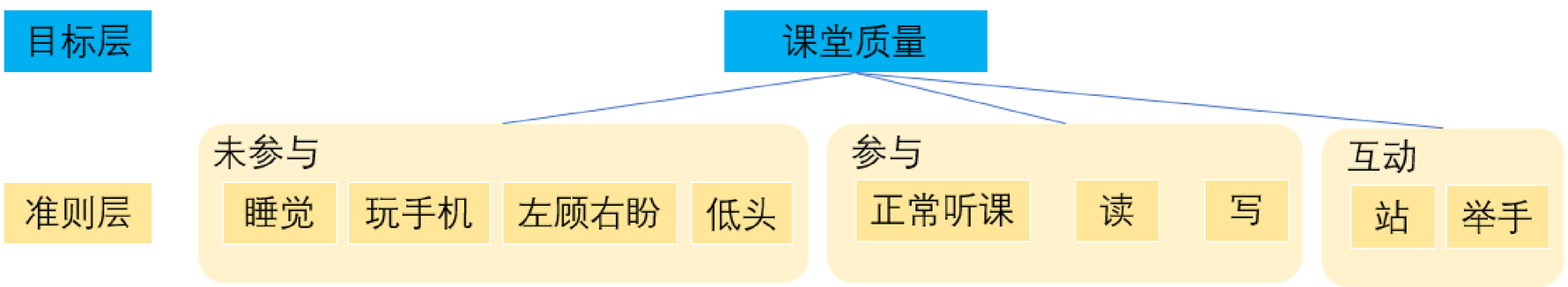


图 5-1　学生课堂状况评估的目标层和准则层

Fig.5-1　The target layer and criterion layer for evaluating students' classroom conditions

在进行数据分析后，本文认为传统的 AHP 法由于其主观性，可能会过分强调某些属性，导致得到的权重向量在进行计算后会使评估结果出现偏差，进而得出不够客观的结论。这一数据分析结果详见第 5.3 节。因此，结合本文构建的双粒度标注的数据集，本研究对传统的 AHP 法进行了优化，提出了改进的双粒度层次分析加权法（Analytic Hierarchy Process-Double Granularity weighted，AHP-DG），步骤如下：

以学生课堂行为样本集$Q = \{q_1, q_2, \cdots, q_n\}$为例，使用 5.1 节介绍的传统 AHP 法中的特征根法计算得出样本集Q的权重向量，从而得到各个学生课堂行为$q_1, q_2, \cdots, q_n$对应的权重向量$w = (w_1, w_2, \cdots, w_n)^{\mathrm{T}}$。

按照上述流程分别计算两种不同粒度下学生课堂行为的权重向量。在细粒度划分中，互动行为（如站、举手）被赋予较高的重要性；而在粗粒度划分中，参与（如正常听课、读、写）和互动行为被赋予近似的重要性值。然后，使用本文第 4 章提出的算法识别得到实际课堂视频中学生的各种行为数量，计算这些行为占总行为数的比例并将其分别与相应的权重相乘，再计算出细粒度和粗粒度下的课堂得分u和v，之后对两者进行加权：

$$S = \mu \cdot u + (1 - \mu) \cdot v. \tag{5-10}$$

其中，μ为细粒度课堂得分在总分S中所占比重的大小，取值范围在0到1之间。当μ 值较大时，互动性强的课堂会得到比μ值较小时更高的总分S。为选择评价的倾向性，μ的大小可以根据实际需求进行调整（本文根据 5.3 节论述的专家评分，μ取 0.4）。

5.3 试验及分析

为了验证提出的学生课堂状况评估方法的合理性，本文利用改进后的模型对 SA 数据集中采集的 10 个真实的课堂长视频进行识别。表 5-2 展示了应用本文第

4 章方法对 10 个视频中细粒度与粗粒度两种划分下检测的学生行为数量，其中粗粒度的三种划分以不同深浅的颜色加以区分。行为检测的频率设置为每秒一帧。

表 5-2 使用本文方法检测的 10 个真实课堂学生行为统计

Table 5-2. Ten real classroom student behaviors detected using the method of this thesis

学生行为	课堂 1	课堂 2	课堂 3	课堂 4	课堂 5	课堂 6	课堂 7	课堂 8	课堂 9	课堂 10
睡觉	0	482	246	152	166	191	16	1	48	65
玩手机	330	612	744	896	244	425	2	414	9	817
低头	5	1286	449	417	294	294	50	9	91	288
左顾右盼	1725	2181	1230	929	656	1093	504	649	2513	653
正常听课	8695	2760	817	753	610	1768	6128	8412	15141	37003
读	3106	1609	1012	764	805	4343	625	1408	680	2753
写	388	975	569	498	495	968	368	164	233	390
站	4506	862	665	349	384	922	2779	3256	1817	1801
举手	850	579	247	190	261	207	1587	936	1664	3119
合计	19605	11346	5979	4948	3915	10211	12059	15249	22196	46889

专家评分由九位教育领域的专家在观看这 10 个视频后给出的主观评价，评价等级由高至低分别为优、中、良。本文采用这种相对简单的 3 级评价有如下原因：一是简单的评价等级在实际应用中能够一目了然；二是该方法方便进行量化，以匹配系统评分。本文在确定每个课堂的专家评分时，采用了多数公决原则。

表 5-3 领域内专家对 10 个课堂的评分

Table 5-3. Rating of 10 classrooms by experts in the field

	课堂 1	课堂 2	课堂 3	课堂 4	课堂 5	课堂 6	课堂 7	课堂 8	课堂 9	课堂 10
专家一	优	良	良	良	中	中	优	优	优	中
专家二	优	良	良	良	良	中	优	优	中	中
专家三	优	良	良	良	良	中	优	优	中	中
专家四	优	中	中	良	中	优	优	优	优	优
专家五	优	良	良	良	良	中	优	优	中	优
专家六	优	良	中	良	中	优	优	优	优	优
专家七	优	良	良	良	良	中	优	优	中	优
专家八	优	良	良	良	良	中	优	优	中	优
专家九	优	良	良	良	良	良	优	中	优	中
专家评分	**优**	**良**	**良**	**良**	**良**	**中**	**优**	**优**	**中**	**优**

为对应专家评分，本文根据百分制设置了系统评分，具体的量化方法见表 5-4。

表 5-4 系统评分量化方法

Table 5-4. Quantitative Method for System Scoring

系统评分	范围
良	<50
中	50-60
优	>60

根据第 5.2 节介绍的方法，本文分别计算了 10 个课堂的细粒度 AHP 法、粗粒度 AHP 法以及 AHP-DG 法的百分制得分和系统评分，如表 5-5 所示。

表 5-5 细粒度 AHP 法、粗粒度 AHP 法、本文方法及专家对 10 个课堂的评分

Table 5-5. Fine grained AHP method, coarse grained AHP method, this thesis 's method and expert evaluation of 10 classrooms

课堂评分	课堂 1	课堂 2	课堂 3	课堂 4	**课堂 5**	课堂 6	课堂 7	课堂 8	**课堂 9**	**课堂 10**
细粒度 AHP 法百分制得分	60.58	45.26	44.4	40.31	**51.35**	55.23	68.21	60.83	**53.59**	**53.54**
细粒度 AHP 法系统评分	优	良	良	良	**中**	中	优	优	**中**	**中**
粗粒度 AHP 法百分制得分	68.77	44.1	41.99	38.04	**49**	57.27	75.57	71.14	**63.9**	**67.57**
粗粒度 AHP 法系统评分	优	良	良	良	**良**	中	优	优	**优**	**优**
本文方法百分制得分	65.49	44.56	42.95	38.95	**49.94**	56.46	72.63	67.02	**59.78**	**61.96**
本文方法系统评分	**优**	**良**	**良**	**良**	**良**	**中**	**优**	**优**	**中**	**优**
专家评分	**优**	**良**	**良**	**良**	**良**	**中**	**优**	**优**	**中**	**优**

从表 5-5 的数据可以观察到，以 5、9 和 10 三个典型课堂为例，细粒度 AHP 法倾向于给予互动性强的课堂较高评分，而对于正常听课、读和写的学生比例较高的课堂评分则会相对压低。相反，粗粒度 AHP 法过分重视了学生的参与状态（即正常听课、读和写），导致这类课堂的评分高于专家评分。而采用本文提出的方法后，通过调整权重变量μ的值，平衡了两种方法的极端状况，实现了系统评分与专家评分的一致性。这一结果验证了本文改进的 AHP 法能够调节评价的主观性，使其比传统 AHP 法更为合理。

5.4 本章小结

为了解决学生课堂状况分析领域评价算法过于简单的问题，本章提出了基于一种主观赋权法——AHP 法改进的学生课堂质量评估算法，即 AHP-DG 法。该方法利用双粒度标注的 SA 数据集，通过引入不同粒度的权重变量μ，实现了对互动性和其他课堂行为侧重程度的调整，即该方法并非完全的主观赋权法，而是一种能够调节专家赋权主观性的评价法，从而相较于传统主观赋权法更合理。

第 6 章 总结与展望

6.1 工作总结

本文首先介绍了学生课堂状况分析领域的研究背景及意义，并回顾了该领域相关技术在国内外的研究历程和现状。同时，本文还详细介绍了神经网络和卷积神经网络的发展，从基本的感知机模型到近年来兴起的卷积神经网络，再到基于卷积神经网络的分类、目标检测算法都进行了深入论述。随后，本文简单介绍了基于注意力机制的相关算法的发展历程。之后，本文介绍了国内外在学生课堂行为数据集方面的研究现状，并论述了 SA 数据集的构建过程，及提出了一个基于改进 YOLOv5 的学生课堂行为识别算法，通过添加 SGE-BI 双注意力机制模块，大大提高了算法在实际应用中的识别能力。最后，本文提出了一个基于层次分析法的学生课堂质量评估算法，通过对学生课堂行为两种粒度划分的不同赋权，综合评价学生课堂的质量。

以下是本文的工作总结：

第一，构建了具有双粒度标注的学生行为（SA）数据集。本文在创建 SA 数据集的过程中，收集了模拟课堂教学场景的学生志愿者视频资料，详细说明了收集过程，并整理了来自网络的单人和多人真实课堂行为资料。之后，本文介绍了数据增强与标注过程，并进行了数据统计。为了使课堂质量评估算法更合理，SA 数据集创造性地通过两种粒度进行了分类：一种是细粒度的，包括九种学生课堂行为：写、站、睡觉、读、举手、玩手机、正常听课、低头和左顾右盼；另一种是粗粒度的，包括三种行为：未参与（包括睡觉、玩手机、左顾右盼和低头）、参与（包括正常听课、读和写）以及互动（包括站和举手）。

第二，设计了 SGE-BI 双注意力机制模块，成功集成到 YOLOv5 算法中。本文深入探讨了基于卷积神经网络的分类和目标检测领域的最新研究成果，并简要评述了当前基于卷积神经网络模型的局限性，同时对注意力机制理论进行了阐述。本文在 SA 数据集上进行了多个目标检测算法的对比实验，这些算法涵盖了一阶段、二阶段及近年来流行的 Transformers 方法，选择了其中识别效果最好的 YOLOv5m 作为主干网络。本文提出了 SGE-BI 双注意力机制模块，并将其融入

YOLOv5m 算法中，通过消融实验，验证了该模块能够显著提升原算法的识别能力。此外，本文还进行了一系列对比实验，结果表明在基础网络中添加两个 SGE-BI 模块最为有效。

第三，结合双层标注的学生行为数据集，对传统层次分析法（AHP）进行了改进，提出了双粒度层次分析加权法（AHP-DG）。层次分析法（AHP）是一种基于专家主观判断的权重分析方法，其通过专家的经验判断来评估不同准则层属性对目标层结果的影响程度。本文的数据分析表明，使用传统的 AHP 法在评估学生课堂行为时存在偏差，从而导致不客观的结论。为了克服这一问题，本文利用 SA 数据集中的双粒度标注，将传统的 AHP 法改进为加权平均法，以此来适应现实中课堂互动性及其他行为的合理权重分配。为验证提出方法的合理性，本文收集了 10 个真实课堂视频并邀请教育专家进行评分，在比较了专家评分、原方法和改进方法的评分后，证明了改进的方法能够有效调整评价的主观性，使之比传统的 AHP 法更为合理。

6.2 未来展望

在传统课堂教学中，学生在课堂上的表现往往是教育工作者最为关心的问题。本研究主要探讨了传统教学环境中监督学生学习状态的问题，以及如何利用新技术解决这个问题。研究涉及收集和构建学生行为数据集、提高识别算法精度、改进学生课堂质量评估算法，并据此代替人力监督学生课堂状况。然而，由于时间和资源的限制，研究还存在一些需要进一步优化和解决的问题：

第一，在数据集方面，可以考虑收集更多学生课堂行为数据进行扩充，以使网络模型能够得到更充分的训练。第二，目前，YOLO 系列算法已经更新至 YOLOv9，且其他识别效果更好的算法也不断被应用于目标检测领域，本文所采用的识别模型尚有提升空间，后续可以尝试将新算法引入，以增强本文模型识别能力。第三，在学生行为分类上，可以进一步研究更加细致的分类，以更准确地拟合学生行为与课堂质量之间的函数关系。第四，针对每个学生在课堂中的表现，未来可以考虑进行 ELO 积分系统等算法的研究，以追踪每个学生的课堂表现，从而更全面地评估课堂质量，以辅助教师教学。第五，使用声音、文字及生理信号等多维度信息对学生课堂质量进行更精确的判断。

参考文献

[1] 陈琼琼. 大学生参与度评价:高教质量评估的新视角——美国“全国学生参与度调查”的解析[J]. 高教发展与评估, 2009, 25(01): 24-30+121.

[2] Pace C R. Achievement and the Quality of Student Effort[C]. Paper presented at a Meeting of the National Commission on Excellence in Education, 1982.

[3] Kuh George D. Assessing what really matters to student learning inside the national survey of student engagement [J]. Change, 2001, 33(3): 10-17.

[4] Tinto V. Leaving College:Rethinking the Causes and Cures of Student Attrition [M]. Chicago: University of Chicago Press, 1987.

[5] Chickering A W, Gamson Z F. Seven Principles for Good Practice in Undergraduate Education[J]. AAHE Bulletin, 1987, 39(7): 3-7.

[6] Mehrabian A, Russell J A. An approach to environmental psychology[M]. the MIT Press, 1974.

[7] Boucher J D, Ekman P. Facial Areas and Emotional Information[J]. Journal of Communication, 1975, 25(2): 21-29.

[8] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection[J]. IEEE Transactions on pattern analysis and machine intelligence, 1997, 19(7): 711-720.

[9] Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: Application to face recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(12): 2037-2041.

[10] Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection[C]//Proceedings. international conference on image processing. IEEE, 2002, 1: I-I.

[11] Nakamura M, Nomiya H, Uehara K. Improvement of boosting algorithm by modifying the weighting rule[J]. Annals of Mathematics and Artificial Intelligence, 2004, 41(1): 95-109.

[12] 王宏漫, 欧宗瑛. 采用 PCA/ICA 特征和 SVM 分类的人脸识别[J]. 计算机辅助设计与图形学学报, 2003(04): 416-420+431.

[13] Suita S, Nishimura T, Tokura H, et al. Efficient convolution pooling on the GPU[J]. Journal of Parallel and Distributed Computing, 2020, 138: 222-229.

[14] Lopes A T, De Aguiar E, De Souza A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order[J]. Pattern recognition, 2017, 61: 610-628.

[15] 罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(06): 169-180.

[16] 吴慧婷. 基于多维度信息融合的学生在线学习投入度研究[D]. 武汉: 华中师范大学, 2020.

[17] 张璟. 基于表情识别的课堂专注度分析的研究[D]. 太原: 山西大学, 2020.

[18] Jisi A, Yin S. A new feature fusion network for student behavior recognition in education[J]. Journal of Applied Science and Engineering, 2021, 24(2): 133-140.

[19] Yang J, Xue Y, Zeng Z, et al. Research on multimodal affective computing oriented to online collaborative learning[C]//2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT). IEEE, 2019, 2161: 137-139.
[20] Huang W, Li N, Qiu Z, et al. An Automatic Recognition Method for Students' Classroom Behaviors Based on Image Processing[J]. Traitement du Signal, 2020, 37(3): 503-509.
[21] TS A, Guddeti R M R. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks[J]. Education and information technologies, 2020, 25(2): 1387-1415.
[22] 潘仙张, 陈坚, 马仁利. 基于面部表情识别的课堂教学反馈系统[J]. 计算机系统应用, 2021, 30(10): 102-108.
[23] 史雨, 辛宇, 袁静, 等. 基于深度学习的学生课堂状态检测算法与应用[J]. Artificial Intelligence and Robotics Research, 2021, 10: 123.
[24] 王泽杰, 沈超敏, 赵春, 等. 融合人体姿态估计和目标检测的学生课堂行为识别[J]. 华东师范大学学报(自然科学版), 2022, (02): 55-66.
[25] Kuh G D. Assessing what really matters to student learning: Inside the National Survey of Student Engagement [J]. Change: The Magazine of Higher Learning, 2001, 33(3): 10-66.
[26] Ko M S, Joo H, Song H. Exam Fraud Prevention and Class Concentration Improvement System for an Online Environment[C]//2022 International Conference on Electronics, Information, and Communication (ICEIC). IEEE, 2022: 1-2.
[27] Tan M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. California: IMLS, 2019: 6105-6114.
[28] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
[29] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510-4520.
[30] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133.
[31] Rosenblatt F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms[R]. Cornell Aeronautical Lab Inc Buffalo NY, 1961.
[32] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
[33] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[34] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.
[35] 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.

[36] 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.
[37] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[38] 罗会兰, 陈鸿坤. 基于深度学习的目标检测研究综述[J]. 电子学报, 2020, 48(06): 1230-1239.
[39] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[40] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.
[41] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[42] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
[43] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137-1149.
[44] Neubeck A, Van Gool L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, 3: 850-855.
[45] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
[46] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
[47] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.
[48] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 658-666.
[49] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.
[50] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[51] Zhao Q, Sheng T, Wang Y, et al. M2det: A single-shot object detector based on multi-level feature pyramid network[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 9259-9266.
[52] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object

detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.
[53] Zhai S, Shang D, Wang S, et al. DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion[J]. IEEE access, 2020, 8: 24344-24357.
[54] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
[55] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[56] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
[57] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
[58] 谈世磊, 别雄波, 卢功林, 等. 基于 YOLOv5 网络模型的人员口罩佩戴实时检测[J]. 激光杂志, 2021, 42(02): 147-150.
[59] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.
[60] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[J]. Advances in Neural Information Processing Systems, 2014, abs/1406.6247.
[61] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.
[62] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach. California: Curran Associates Inc., 2017: 6000-6010.
[63] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019: 4171-4186.
[64] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. ArXiv, 2020, abs/2010.11929.
[65] 刘永娜, 孙波, 陈玖冰, 等. BNU 学习情感数据库的设计与实现[J]. 现代教育技术, 2015, 25(10): 99-105.
[66] Ramón Z C, Lucia B E M, Daniel L H, et al. Creation of a Facial Expression Corpus from EEG Signals for Learning Centered Emotions[C] // 2017 IEEE 17th International Conference on Advanced Learning Technologies(ICALT).Los Angeles: IEEE Computer Society, 2017:387 - 390.
[67] Bian C L, Zhang Y, Yang F, et al. A Spontaneous Facial Expression Database for Academic Emotion Inference in Online Learning [J].IET Computer Vision,2018,13 (3): 329-337.

[68]孔玺. 智慧学习环境下数字学习画面的情感研究[D]. 济南: 山东师范大学, 2020.
[69]LI X, HU X, YANG J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks[J]. ArXiv, 2019, abs/1905.09646.
[70]He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans. LA: IEEE, 2022: 16000-16009.
[71]ZHU L, WANG X, KE Z, et al. BiFormer: Vision Transformer with Bi-Level Routing Attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver. Canada: IEEE, 2023: 10323-10333.
[72]LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual: IEEE, 2021: 10012-10022.
[73]REN S, ZHOU D, He S, et al. Shunted self-attention via multi-scale token aggregation[C]-//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10853-10862.
[74]LYU C, ZHANG W, HUANG H, et al. Rtmdet: An empirical study of designing real-time object detectors[J]. ArXiv, 2022, abs/2212.07784.
[75]ZHU X, SU W, LU L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. ArXiv, 2020, abs/2010.04159.
[76]LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice. Italy: IEEE, 2017: 2980-2988.
[77]FENG C, ZHONG Y, GAO Y, et al. Tood: Task-aligned one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual: IEEE, 2021: 3490-3499.
[78]镇常青. 多目标决策中的权重调查确定方法[J]. 系统工程理论与实践, 1987, 7(2): 16-24.
[79]Lin C C, Wang W C, Yu W D. Improving AHP for construction with an adaptive AHP approach (A3)[J]. Automation in construction, 2008, 17(2): 180-187.
[80]杨宇. 多指标综合评价中赋权方法评析[J]. 统计与决策, 2006(13): 17-19.

攻读学位期间发表的学术论文目录

[1] 王岱嵘, 陈新泉, 丁达之. 基于改进 YOLOv5 的对课堂教学的评估算法与应用[J]. 安徽工程大学学报, 2024(第一作者, 录用)

致谢

时光匆匆，转眼三年的研究生生涯就要结束了。在与母校临别之际，回忆起我三年的校园生活和这篇论文从选题到完成的过程，在此要感谢多位同学、老师的悉心帮助。

首先，我要感谢我的导师陈新泉老师。他是一位知识渊博，态度严谨的学者，也是一位待人真诚、不慕名利的长者。从论文的选题到最终完成，他都不断给予我耐心的指导和支持，在此向他致以真诚的谢意。

其次，我要感谢安徽师范大学物理与电子信息学院的王桂丽老师。她的小组课题与我的课题研究方向相近，三年来，我持续参与王老师小组的课题讨论，从生活到学习上，她都给予我很多帮助和指导，借此也向王老师致以衷心的感谢。

此外，我要感谢我的三位室友——韩龙、马志远、汪邦荣。他们不仅在日常生活中给予我很大的帮助，而且在我课题遇到困难的时候，通过与他们的讨论，获得了不少灵感。

最后，我要感谢从本论文开题到答辩提出宝贵意见的老师们，感谢您们在百忙之中抽出时间对我进行有益的指导；也感谢参与本课题的志愿者们，感谢你们的无私帮助。